nature human behaviour

Supplementary information

https://doi.org/10.1038/s41562-026-02566-6

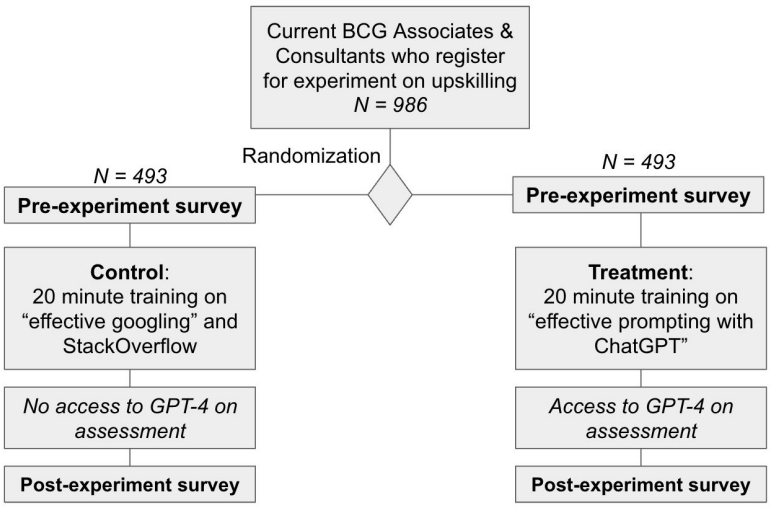
Generative AI and the temporary upskilling of knowledge workers

In the format provided by the authors and unedited

Supplementary 1 Experimental design and task overview

Supplementary 1.1 Experimental design

Supplementary Figure 1 Overview of the experimental design



Notes: The registration survey is in Supplementary Section 5.1. The pre-experiment survey and training information can be found in Supplementary Section 5.2. The text of all three tasks is summarized in Supplementary Section 1.2 with exact survey in Supplementary Section 3. The post-experiment survey can be found in Supplementary Section 5.3.

Supplementary 1.2 Task description

For the purpose of this study, we identify three types of tasks to test nontechnical workers' ability to perform data science-related work. The three tasks are independent of each other and test different skills, including the ability to 1) write Python code to perform data cleaning; 2) identify and correct modeling errors, apply statistical metrics, interpret statistical plots, and calculate probabilities; and 3) build and assess a predictive model. The details of each task are as follows:

Coding task

The coding task is designed to test the ability of a worker to write functioning Python code. Workers are asked to clean, merge, and filter two datasets. This type of task is required in the early stages of any data science exercise. We provide participants with detailed guidelines and tips to help them accomplish the task (see Section 1.3 for details) to isolate the skill of writing functioning code from that of knowing the appropriate data cleaning steps. Specifically, participants are prompted to write Python code to read two data files that contain information about customers, order numbers, and products. They are then provided clear data quality and cleaning guidelines and tips for handling missing, junk, and duplicate data and asked to identify the five customers with the highest average by cost in a certain period. This exercise involved the following steps:

- Reading the data/creating a data frame
- Splitting combined string features into separate columns
- Handling feature types for filtering
- Detecting junk values
- Removing null and junk values
- Joining data
- Calculating basic metrics

Before beginning the task, we provide participants with written and video instructions on how to download and install Colab, Google's integrated development environment (IDE), to ensure that it is functioning properly (see Supplementary Section 5 for details). We choose Colab as the IDE because it is relatively simple to set up for nontechnical workers and allows us to access and trace the code for evaluation after task completion.

We divide the task into ten distinct steps, as described in the coding task rubric (Supplementary Section 4.2). For phase analysis, we combine steps 1-3 as the "data cleaning" phase, steps 4-5 as the "data correction" phase and steps 6-10 as the "data merging and filtering" phase.

Heterogeneous treatment effects on the coding task, the robustness of the results and the details of the phase analysis can be found in Supplementary Sections 2.2.1, 2.2.2 and 2.2.3, respectively.

Statistical understanding task

In the statistical knowledge task, we assess participants' ability to understand and apply statistical knowledge. Workers are guided through a series of potential statistical analysis recommendations and other erroneous GenAI outputs including 1)

an erroneous analysis plan for creating a model that predicts whether a couple will take out a mortgage based on historical data, 2) analyzing and interpreting residual plots, 3) understanding the application of statistical metrics such as empirical risk to graphical outputs and 4) calculating situational probability. The responses to these questions include a mix of open-ended and multiple-choice answers. The questions are designed to encourage engagement and interpretation rather than simply finding and copying answers. Participants are explicitly instructed not to copy and paste the questions or images into ChatGPT or Google.

The questions are framed around the following three scenarios:

1. Given a data sample that contains financial and demographic information about partners who purchase a home, we ask participants to identify the steps needed to explore and preprocess the data, split the data, train and evaluate the model and then interpret various residual plots.
2. Given a certain distribution of data points, participants are asked how to classify additional data points, identify empirical risks and name potential classifiers that can create certain boundaries.
3. Participants are given a scenario involving a delivery truck that went missing, where the manager has to decide which route to choose based on situational probability.

In addition to the overall average treatment effects, we address the following questions:

- the type of statistical knowledge they test (i.e., debugging a series of steps in a proposed model by ChatGPT, analyzing and interpreting plots, and calculating situational probability)
- whether the responses are open ended or multiple choice, and
- whether the questions are visual or text based.

Supplementary Section 4.3 provides the grading rubric for the statistical knowledge task.

Heterogeneous treatment effects on the statistical knowledge task, the robustness of the results and the details of treatment effects based on the type of questions can be found in Sections 2.2.1, 2.2.2 and 2.2.4, respectively.

Prediction task

The prediction task aims to test participants' ability to conduct quantitative analysis and build a predictive model based on historical data. We provide a dataset containing information about soccer matches, including the date, home-team name, away-team name, tournament type, whether the game is at a neutral location, and the number of goals scored by the home and away teams. Participants are asked to quantify the predictability of each match in the dataset. Finally, they are asked to identify the most surprising match based on their predictability metric.

The deliverables of the problem-solving task include the participants' textual description of the methodology they employ to solve the problem and a tabular format of their results for the quantified predictability of each match. Unlike the coding and statistical knowledge tasks, this task is more "open ended", requiring participants to choose their method to build the model and identify an appropriate outcome

to measure predictability. The rubric for grading the prediction task is detailed in Supplementary Sections 1.3 and 4.

Heterogeneous treatment effects on the prediction task, the robustness of the results and the details on the phase analysis can be found in Sections 2.2.1, 2.2.2 and 2.2.5, respectively.

Supplementary 1.3 Task grading methodology

Each task is graded with quantifiable measures of the correctness of the answers and approach, depending on the hypothesis. Each task is graded on both the correctness of the answer and the steps the participant uses to solve the problem. Below, we describe the main outcomes for the correctness of the answer and for the process scores.

Coding task

There is one distinct correct answer for the coding assignment. Correctness is a binary measure where 0 is incorrect and 1 is correct. Second, we compare the output from the workers and data scientists to a rubric we create with 10 steps, where each step is necessary to obtain the correct score. As with the statistics task, we grade this against the rubric (shown in Supplementary Section 4). The rubric scores are weighted correctness scores such that the final score is determined by a weighted sum across all answers as follows:

$$\text{Total correctness} = \sum_{i=1}^{n} (\text{Correctness of answer}_i \times \text{Complexity weight}_i) \qquad (1)$$

where n is the total number of distinct questions and the correctness of the answer, and the complexity weighting is defined as the level of complexity of the question. Similarly, this score is between $(0, 1)$, where 1 means that the worker takes all of the correct steps.

Statistical understanding task

Each question in the statistics task is graded against the rubric (shown in Supplementary Section 4). The rubric scores are weighted correctness scores such that the final score is determined by a weighted sum across all answers as follows:

$$\text{Total correctness} = \sum_{i=1}^{n} (\text{Correctness of answer}_i \times \text{Complexity weight}_i) \qquad (2)$$

where n is the total number of distinct questions and the correctness of the answer, and the complexity weighting is defined as the level of complexity of the question. The complexity weightings are determined by asking several lead data scientists with more than 5 years of experience to rank the complexity of each question and average it across their answers. This correctness measure is bounded by $(0, 1)$, where 1 is a perfect score.

Prediction task

The problem-solving task is designed to have numerous possible answers, some of which are better than others. We use the answers submitted by the data scientists as the baseline/benchmark to grade the workers' results. Specifically, participants submit a predictability score for each match. We normalize participants' predictability scores for each match, score_i, and calculate a loss score for the answers submitted by the workers compared to the data science benchmarks, DS score_i. For each participant, we create a loss score defined as follows:

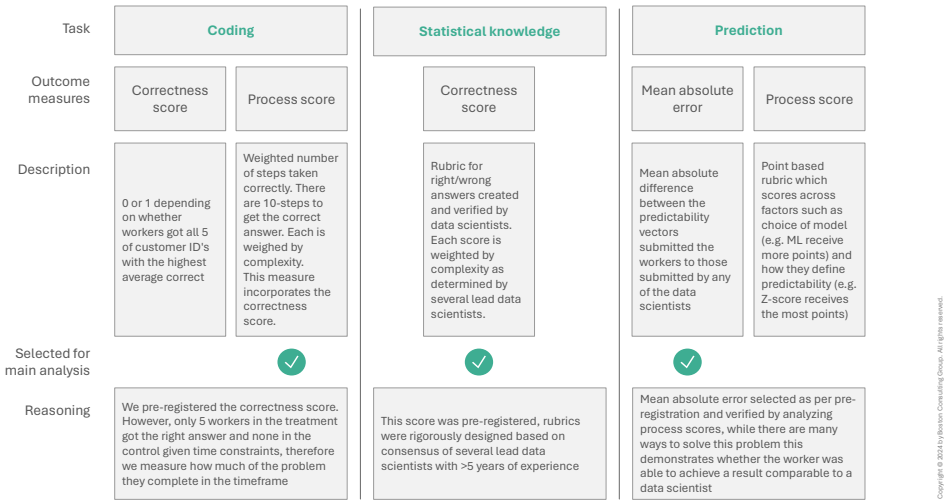
$$\text{Loss Score} = \frac{1}{n}\sum_{i=0}^{n} |\text{score}_i - \text{DS score}_i| \tag{3}$$

where n is the number of soccer matches in the dataset.

The final score we give workers on the prediction problem is -1 times the loss score so that the score is between $(-1, 0)$, where 0 is a perfect score.

Outcome selection

Supplementary Figure 2 Outcome description and selection for the main analysis



Notes: This figure illustrates the different outcomes for each task, those selected for our main analysis and the reasoning behind the selection.

Supplementary 1.3.1 Grading details

The validation process across all tasks was designed to ensure technical rigor while addressing task-specific challenges. We implemented a rigorous validation process (measuring accuracy and consistency with both human manual validation and pre-defined metrics) designed to ensure both accuracy and consistency:

- **Randomized Batch Grading:** Participant responses were graded in randomized batches of five to reduce the risk of contextual bias. Statistical variability within batches was closely monitored and informed improvements to grading prompts.
- **Detailed Rubrics:** Grading was guided by comprehensive, task-specific rubrics created and validated by Data Scientists and Lead Data Scientists (Data Scientists with 4+ years of experience at BCG, likely more in the industry). These rubrics outlined clear expectations for each task and were iteratively refined to handle diverse approaches while maintaining fairness.
- **Preprocessing to Mitigate Bias:** Free-text responses were paraphrased and summarized to prevent GPT-4 Turbo from favoring verbose or overly polished submissions. This preprocessing step ensured a level playing field for all participants.
- **Human Benchmarks:** Between 25% and 100% of responses, depending on the task, were manually reviewed and graded by Data Scientists to establish a ground truth. These manual grades were compared to the automated scores to validate the accuracy of the system and refine grading strategies. Prompt Optimization: Grading prompts were refined iteratively to reduce inconsistencies and improve alignment with ground truth scores. Metrics like mean squared error were used to measure and address discrepancies between human and automated evaluations.
- **Multi-Pass Grading and Manual Review:** Each submission was graded multiple times, and consensus or averaged scores were used when the results were consistent. Submissions with high variability across evaluations were flagged for manual review.

Please refer to the detailed grading explanation in Supplementary Section 4.1 for further explanation on the LLM grading strategies, and the grading validation.

Supplementary 1.4 Estimation of treatment effects

Across each of these metrics, we employ Equation (4) to estimate the average treatment effects based on ordinary least squares regression, where y_i is the dependent variable (e.g., representing a quantifiable measure of output quality in the coding task and efficiency of code), and T_{GPT} is the ChatGPT treatment dummy. Finally, X_i is a set of covariates collected in the survey—office location, gender, tenure at BCG, and whether the worker is a native English speaker.

$$y_i = \beta_0 + \beta_{\text{GPT}} T_{\text{GPT}} + \gamma X_i + \varepsilon_i \tag{4}$$

Supplementary 2 Supplementary tables and figures

Supplementary 2.1 Balance check

Supplementary Table 1 Comparison of worker covariates by treatment assignment

Flow from initial allocation into analysis sample				
	Total (N)	*Treatment (N)*	*Control (N)*	*P-value*
Total workers allocated	986	493	493	
↪ began survey	573	298	275	0.336
↪ completed first task	511	270	241	0.199
↪ completed second task	487	260	227	0.135

Pre-allocation attributes of final sample: N = 487			
	Treatment mean: $\bar{X}_{TRT}$	Control mean: $\bar{X}_{CTL}$	P-value
Female	0.369	0.37	0.985
Bachelors Degree	0.238	0.291	0.192
Masters Degree	0.677	0.604	0.092
Doctorate	0.085	0.106	0.428
Consultant	0.515	0.493	0.361
Low Tenure	0.535	0.542	0.893
Native English	0.471	0.423	0.343
Office in Africa	0.019	0.018	0.896
Office in Asia Pacific	0.135	0.115	0.505
Office in Central or South America	0.019	0.004	0.14
Office in Europe or Middle East	0.492	0.52	0.546
Office in North America	0.335	0.344	0.835
Code at Work	0.277	0.285	0.875
Some Python Familiarity	0.335	0.331	0.922
Familiar with ChatGPT for Coding	0.413	0.381	0.518
Use ChatGPT Daily for Work	0.394	0.438	0.371

Notes: This table reports the means and standard errors of various pretreatment covariates for the treatment and control groups. The first panel describes the flow of the sample from the allocation to the sample we use for our main experimental analysis. The complete allocated sample is described in the first line, with each following line defined cumulatively. Each worker is assigned two tasks, and the following lines compare the number of workers who submit any work for each of the tasks. Those who complete both tasks make up the main experimental sample. The second panel looks at the pre-allocation characteristics of the job seekers in the sample we use for our analysis, N = 487. We report the fraction of workers on their self-reported i) gender, ii) highest degree achieved, and iii) office location. The reported p values are for two-sided t-tests of the null hypothesis of no difference in means across groups.

Supplementary Table 2 Comparing those who submit something for both tasks (primary analysis sample) to attritors

		Sample Mean	Attritor Mean	P-value
Control	Female	0.38	0.48	0.18
Treatment		0.37	0.43	0.42
Control	Ofice in Europe or Middle East	0.52	0.43	0.19
Treatment		0.50	0.57	0.38
Control	Native English speaker	0.50	0.49	0.86
Treatment		0.41	0.48	0.41
Control	New hire (<1 year)	0.51	0.40	0.11
Treatment		0.54	0.55	0.98
Control	Proficient coder or better	0.31	0.22	0.11
Treatment		0.32	0.20	0.09
Control	Never coded	0.30	0.37	0.35
Treatment		0.31	0.32	0.93
Control	At most 1 coding language	0.93	0.95	0.75
Treatment		0.96	1.00	0.00
Control	PhD	0.11	0.00	0.00
Treatment		0.09	0.05	0.26
Control	Uses ChatGPT daily for work	0.38	0.48	0.16
Treatment		0.45	0.50	0.52
Control	Familiar with prompt engineering	0.68	0.58	0.15
Treatment		0.67	0.68	0.91
Control	Never code for work	0.59	0.68	0.28
Treatment		0.61	0.69	0.42

Notes: This table reports the mean of various pre-experiment covariates amongst the primary analysis sample with those who attrit. The primary analysis sample is made up of workers who submitted anything to be graded for both of their two tasks. The sample here is made up of all workers who started the pre-experiment survey, N = 573. We run a two-sided Welch Two Sample t-test on each covariate for attritors and non-attritors, within each treatment group.

Supplementary 2.2 Treatment effects of AI

Supplementary Table 3 Effects of AI to workers performance on data science tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.490***	0.201***	0.172***
	(0.036)	(0.026)	(0.042)
Mean Y in Control Group	-0.63	-0.32	-0.43
P-value (Trt)	0.000	0.000	0.000
Observations	300	330	298
R^2	0.360	0.151	0.056

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions. Each outcome is normalized relative to the performance of BCG data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The third outcome is the score they got on the prediction problem, which is -1 times their mean absolute error from the correct answer. Exact definition of grading for each problem can be found in Sup Section 1. All reported standard errors are two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 4 Effects of AI to whether or not they submit any answer on each task

	Dependent variable:		
	Stats Submitted	Prediction Submitted	Coding Submitted
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.062	0.018	0.069
	(0.033)	(0.035)	(0.040)
Mean Y in Control Group	0.85	0.87	0.78
P-value (Trt)	0.066	0.616	0.089
Observations	369	364	369
R^2	0.024	0.006	0.015

Notes: This table analyzes the effect of the treatment on the consultants submitting any answer to each question. Text of problems can be found in Supplementary Section 1. All regressions include controls for gender, location, native english status, and low tenure. Reported entries are coefficient estimates with two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 5 Effects of AI to whether or not they submit any answer on each task

	Dependent variable:	
	Task 1 Submitted	Task 2 Submitted
	(1)	(2)
GenAI Treatment Assigned (Trt)	0.026	0.043
	(0.026)	(0.030)
Mean Y in Control Group	0.88	0.83
P-value (Trt)	0.315	0.151
Observations	573	573
R^2	0.014	0.016

Notes: This table analyzes the effect of the treatment on the consultants submitting any answer to each question. Text of problems can be found in Supplementary Section 1. All regressions include controls for gender, location, native english status, and low tenure. Reported entries are coefficient estimates with two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 6 Effects of AI on number of minutes to complete each task

	Dependent variable:		
	Mins on Stats	Mins on Prediction	Mins on Coding
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	2.312	-14.450^{***}	-8.884^{***}
	(2.407)	(2.946)	(2.521)
Mean Y in Control Group	63.61	68.36	78.01
P-value (Trt)	0.340	0.000	0.000
Observations	327	318	303
R^2	0.021	0.094	0.058

Notes: This table analyzes the effect of the treatment on the length of time it took for consultants to finish each task, conditional on completion. The outcome in Column (1) is the number of minutes they spent on the Statistics task. The outcome in Column (2) is the number of minutes spent on the Problem Solving and Prediction task. And the outcome in Column (3) is the number of minutes spent on the Coding task. Consultants were randomly assigned two of the three tasks, and given 90 minutes maximum to complete each. All regressions include controls for gender, location, native english status, and low tenure. All reported standard errors are two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 7 Effects of AI treatment on post-experiment data science knowledge (no AI use)

	Dependent variable:				
	Data science or coding question				
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	0.008	0.027	0.048	−0.006	−0.077
	(0.038)	(0.027)	(0.027)	(0.050)	(0.040)
Mean Y in Control Group	0.35	0.91	0.73	0.80	0.78
P-value (Trt)	0.827	0.319	0.075	0.902	0.057
Observations	573	399	418	253	408
R^2	0.0001	0.003	0.008	0.0001	0.009

Notes: This table analyzes the effect of the treatment on the consultants ability to answer data science and coding questions, after the conclusion of the experiment. Full text of questions can be found in Supplementary Section 5.2. Reported entries are coefficient estimates with two-sided 95% confidence intervals computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 8 Effects of AI treatment to post experiment questions about GPT-4's capabilities

	Dependent variable:						
	"Can GPT-4 answer [this question] correctly?"						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
GenAI Treatment Assigned (Trt)	1.390	6.006*	5.539*	6.214**	0.302	9.654***	5.340*
	(2.002)	(2.351)	(2.428)	(2.287)	(2.348)	(2.481)	(2.577)
Mean Y in Control Group	75.82	73.66	74.75	69.08	77.93	67.14	64.93
P-value (Trt)	0.488	0.012	0.025	0.007	0.898	0.000	0.040
Observations	454	475	473	464	465	431	451
R^2	0.042	0.024	0.027	0.025	0.005	0.086	0.013

Notes: This table analyzes the effect of the treatment on the consultants confidence in GPT-4's ability to get the right answer on various questions, after the conclusion of the experiment. For each question, the consultant gave a percentage confidence in GPT-4's ability to answer the question correctly. The question in Columns 1 and 4, and 5 GPT-4 usually get correct. Questions 2,3,6 and7, GPT-4 almost never get correct. Text of questions can be found in Supplementary Section 1. All regressions include controls for gender, location, native english status, and low tenure. All reported standard errors are two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 9 Effects of AI to workers performance on tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
Assigned Control	-0.629^{***}	-0.323^{***}	-0.432^{***}
	(0.081)	(0.044)	(0.034)
GenAI Treatment Assigned (Trt)	-0.140	-0.123^{**}	-0.260^{***}
	(0.083)	(0.044)	(0.024)
P-value (Assigned Control)	0.000	0.000	0.000
P-value (GenAI Treatment)	0.093	0.005	0.000
Observations	331	357	342
R^2	0.363	0.185	0.152

Notes: This table illustrates the performance of workers in the control and treatment arms compared to data scientists'. The first outcome is the percentage of correct steps they took on the coding problem, weighted by difficulty. The second outcome is the points the worker got on the statistics question, divided by the maximum points. The third outcome is their mean absolute error of the distance from their answer to the data scientist's answer. Exact definition of grading for each problem can be found in Supplementary Section 1. All reported standard errors are two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001.

Supplementary 2.2.1 Heterogeneous treatment effects of AI

Supplementary Table 10 Effects of AI to workers performance on data science tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.547^{***}	0.311^{***}	0.091
	(0.052)	(0.048)	(0.087)
Coding basics	0.071	0.095^{*}	0.002
	(0.043)	(0.042)	(0.082)
Competent coder	0.151^{**}	0.238^{***}	-0.037
	(0.049)	(0.041)	(0.092)
Coding basics x Trt	-0.131	-0.143^{*}	0.052
	(0.080)	(0.063)	(0.109)
Competent coder x Trt	-0.038	-0.154^{*}	0.175
	(0.085)	(0.062)	(0.112)
Constant	-0.708^{***}	-0.444^{***}	-0.422^{***}
	(0.026)	(0.032)	(0.064)
Mean Y in Control Group	-0.63	-0.32	-0.43
P-value (Trt)	0.000	0.000	0.296
P-value (Trt x Coding basics)	0.101	0.023	0.629
P-value (Trt x Competent coder)	0.652	0.014	0.118
Observations	300	330	298
R^2	0.387	0.243	0.070

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, by their pre-experiment coding knowledge. The omitted variable is "No prior coding experience." Workers who report having basic coding ability are classified as "Coding basics" and those who report knowing how to code are classified as "competent coders." Each outcome is normalized relative to the performance of BCG data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The third outcome is the score they got on the prediction problem, which is 1- their mean absolute error. Exact definition of grading for each problem can be found in Supplementary Section 1. All standard errors are huber white robust. * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 11 Effects of AI to workers performance on data science tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.503^{***}	0.183^{***}	0.128^{**}
	(0.041)	(0.030)	(0.047)
Knowledge of GPT's strengths	0.017	0.003	−0.128
	(0.052)	(0.051)	(0.085)
Knowledge of GPT's strengths x Trt	−0.062	0.080	0.195^{*}
	(0.088)	(0.063)	(0.098)
Constant	-0.633^{***}	-0.324^{***}	-0.404^{***}
	(0.024)	(0.020)	(0.038)
P-value (Trt)	0.000	0.000	0.007
P-value (Trt x Knowledge of GPT's strengths)	0.483	0.204	0.048
Observations	300	330	298
R^2	0.361	0.160	0.069

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, by their pre-experiment coding knowledge. "Knowledge of GPT's strengths" is 1 if the consultant got 4 out of 7, or more of questions correct on the pre-experiment survey asking about their guesses of whether or not GPT-4 can correctly answer a question. The omitted variable is consultants who got fewer tha 4 out of 7 correct. Each outcome is normalized relative to the performance of BCG data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The third outcome is the score they got on the prediction problem, which is -1 times their mean absolute error. Exact definition of grading for each problem can be found in Supplementary Section 1. All standard errors are huber white robust. * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 12 Effects of AI to workers performance on data science tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.512^{***}	0.250^{***}	0.214^{***}
	(0.048)	(0.038)	(0.061)
Overconfident	0.026	0.041	0.088
	(0.042)	(0.037)	(0.068)
Overconfident x Trt	−0.046	−0.094	−0.087
	(0.072)	(0.052)	(0.083)
Constant	-0.642^{***}	-0.345^{***}	-0.475^{***}
	(0.029)	(0.028)	(0.050)
P-value (Trt)	0.000	0.000	0.001
P-value (Trt x Overconfident)	0.525	0.072	0.298
Observations	300	330	298
R^2	0.361	0.159	0.063

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, by their pre-experiment estimates of GPT-4's current capabilities. "Overconfident" is 1 if the consultant believed that GPT-4 could correctly answer at least four of the five questions which GPT-4 could not do. Each outcome is normalized relative to the performance of BCG data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The third outcome is the score they got on the prediction problem, which is -1 times their mean absolute error. Exact definition of grading for each problem can be found in Supplementary Section 1. All standard errors are huber white robust. * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 13 Effects of AI to workers performance on data science tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.468^{***}	0.184^{***}	0.193^{***}
	(0.041)	(0.030)	(0.049)
Quant Degree	0.015	0.019	0.061
	(0.052)	(0.042)	(0.075)
Quant Degree x Trt	0.090	0.089	−0.070
	(0.081)	(0.057)	(0.093)
Constant	-0.634^{***}	-0.329^{***}	-0.450^{***}
	(0.023)	(0.021)	(0.041)
P-value (Trt)	0.000	0.000	0.000
P-value (Quant Degree x Trt)	0.269	0.120	0.452
Observations	300	330	298
R^2	0.367	0.167	0.059

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, by their pre-experiment estimates of GPT-4's current capabilities. "Quant Degree" is 1 if the worker reports having any formal degree in Statistics, Economics, Mathematics or Data Science. Each outcome is normalized relative to the performance of BCG data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The third outcome is the score they got on the prediction problem, which is -1 times their mean absolute error. Exact definition of grading for each problem can be found in Supplementary Section 1. All standard errors are huber white robust. * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 14 Effects of AI treatment to post experiment data science knowledge without use of AI

	Dependent variable:				
2-6					
	Data science or coding question				
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	0.0004	0.023	0.079^{*}	-0.067	-0.148^{**}
	(0.052)	(0.032)	(0.034)	(0.068)	(0.049)
Assigned Coding Task	0.391^{***}	0.914^{***}	0.724^{***}	0.817^{***}	0.825^{***}
	(0.038)	(0.023)	(0.025)	(0.049)	(0.036)
P-value (Trt)	0.994	0.483	0.020	0.318	0.002
Observations	323	270	270	151	270
R^2	0.00000	0.002	0.020	0.007	0.033

Notes: This table analyzes the effect of the treatment on the consultants ability to answer data science and coding questions, after the conclusion of the experiment. The first problem is about coding, the second is about statistics, and the final three are about machine learning and prediction. For each problem the treatment is interacted with an indicator for whichever task gave the worker experience in that topic. Full text of questions can be found in Supplementary Section 5.2. Reported entries are coefficient estimates with **95% confidence intervals computed using Huber–White (HC0) robust variance**. Significance stars and **p-values are based on the same HC0 robust variance**. * p <0.05, ** p <0.01, *** p <0.001.

Supplementary 2.2.2 Robustness of the results

Supplementary Table 15 Lee Bounds on Treatment Effects for Main Results

	Treatment effect	**Lee Lower Bound**	**Lee Upper Bound**
Statistics Task Score	0.204***	0.197***	0.201***
P-value	< 0.001	< 0.001	< 0.001
Prediction Task Score	0.182***	0.164***	0.199**
P-value	< 0.001	< 0.001	0.005
Coding Task Score	0.491***	0.471***	0.555***
P-value	< 0.001	< 0.001	< 0.001

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions. The first outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The second outcome is the percentage of correct steps they take in answering the coding question. The third outcome is the score they got on the prediction problem, which is their mean absolute error multiplied by -1 for ease of interpretation. Exact definition of grading can be found in Supplementary Section 1. All regressions include no controls. $N = 505$. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 16 Effects of AI to workers performance on coding task, relative to data scientists

	Dependent variable:				
	Coding Score				
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	0.490***	0.491***	0.487***	0.559***	0.564***
	(0.038)	(0.038)	(0.038)	(0.069)	(0.069)
Coding basics				0.102	0.074
				(0.071)	(0.072)
Competent coder				0.179*	0.129
				(0.072)	(0.080)
Europe		0.053	0.072	0.075	0.088*
		(0.041)	(0.041)	(0.041)	(0.042)
Female		0.027	0.041	0.052	0.042
		(0.040)	(0.041)	(0.040)	(0.041)
Low tenure		−0.042	−0.063	−0.056	−0.080
		(0.039)	(0.040)	(0.038)	(0.041)
Native English speaker		0.051	0.023	0.055	0.025
		(0.041)	(0.042)	(0.041)	(0.042)
Masters degree			−0.072		−0.074
			(0.046)		(0.046)
PhD			−0.125		−0.117
			(0.072)		(0.072)
Degree in quantitative field			0.048		0.040
			(0.045)		(0.045)
STEM			0.063		0.045
			(0.040)		(0.043)
Minimal experience with prediction			−0.102*		−0.106*
			(0.051)		(0.051)
Some experience with prediction			0.046		0.0002
			(0.050)		(0.056)
Data visualization experience			0.034		0.023
			(0.049)		(0.049)
Coding basics x Trt				−0.162	−0.160
				(0.094)	(0.095)
Competent coder x Trt				−0.042	−0.056
				(0.094)	(0.094)
Constant	−0.629***	−0.668***	−0.653***	−0.781***	−0.690***
	(0.028)	(0.050)	(0.082)	(0.071)	(0.093)
P-value (Trt)	0.000	0.000	0.000	0.000	0.000
Observations	300	300	300	300	300
R^2	0.360	0.369	0.409	0.403	0.425

Notes: This table analyzes the effect of the treatment on the worker's score on the coding task relative to the data scientists benchmark, where 0 is a perfect score. The first specification is the effect of treatment on their score. The second specification adds controls for gender, office location, native english status, and prior coding ability. The third specification is the same as column (2) but with additional controls for different measures of the workers prior technical experience. The fourth specification is the same as column (2), where the treatment is interacted with the workers prior coding experience. The fifth specification is the same as column (3), where the treatment is interacted with the workers prior coding experience.All reported standard errors are two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 17 Effects of AI to workers performance on statistics task, relative to data scientists

	Dependent variable:				
	Statistics Score				
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	0.201^{***}	0.201^{***}	0.206^{***}	0.308^{***}	0.309^{***}
	(0.026)	(0.026)	(0.026)	(0.045)	(0.045)
Coding basics				0.094^{*}	0.088
				(0.047)	(0.048)
Competent coder				0.234^{***}	0.210^{***}
				(0.045)	(0.050)
Europe		0.028	0.027	0.035	0.030
		(0.029)	(0.029)	(0.028)	(0.029)
Female		−0.033	−0.010	−0.014	−0.011
		(0.027)	(0.028)	(0.026)	(0.027)
Low tenure		0.036	0.051	0.032	0.042
		(0.027)	(0.028)	(0.026)	(0.028)
Native English speaker		0.038	0.033	0.038	0.027
		(0.029)	(0.030)	(0.028)	(0.030)
Masters degree			−0.027		−0.033
			(0.032)		(0.032)
PhD			0.004		−0.015
			(0.052)		(0.051)
Degree in quantitative field			0.041		0.032
			(0.032)		(0.031)
STEM			0.0003		−0.024
			(0.028)		(0.029)
Minimal experience with prediction			0.048		0.036
			(0.036)		(0.035)
Some experience with prediction			0.121^{***}		0.066
			(0.034)		(0.037)
Data visualization experience			0.044		0.020
			(0.035)		(0.035)
Coding basics x Trt				-0.137^{*}	-0.130^{*}
				(0.063)	(0.063)
Competent coder x Trt				-0.147^{*}	-0.153^{*}
				(0.061)	(0.062)
Constant	-0.323^{***}	-0.362^{***}	-0.461^{***}	-0.491^{***}	-0.499^{***}
	(0.019)	(0.036)	(0.061)	(0.046)	(0.065)
P-value (Trt)	0.000	0.000	0.000	0.000	0.000
Observations	330	330	330	330	330
R^2	0.151	0.165	0.224	0.254	0.273

Notes: This table analyzes the effect of the treatment on the worker's score on the statistcis task relative to the data scientists benchmark, where 0 is a perfect score. The first specification is the effect of treatment on their score. The second specification adds controls for gender, office location, native english status, and prior coding ability. The third specification is the same as column (2) but with additional controls for different measures of the workers prior technical experience. The fourth specification is the same as column (2), where the treatment is interacted with the workers prior coding experience. The fifth specification is the same as column (3), where the treatment is interacted with the workers prior coding experience. All reported standard errors are two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance * p <0.05, ** p <0.01, *** p <0.001.

Supplementary Table 18 Effects of AI to workers performance on a prediction task, relative to data scientists

	Dependent variable:				
	Mean Absolute Error				
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	0.172***	0.182***	0.168***	0.103	0.086
	(0.041)	(0.041)	(0.042)	(0.078)	(0.078)
Coding basics				0.014	0.043
				(0.073)	(0.076)
Competent coder				−0.026	0.023
				(0.080)	(0.096)
Europe		0.020	0.016	0.025	0.019
		(0.046)	(0.046)	(0.046)	(0.047)
Female		0.023	0.011	0.028	0.008
		(0.043)	(0.044)	(0.043)	(0.044)
Low tenure		−0.008	−0.026	−0.011	−0.028
		(0.042)	(0.044)	(0.042)	(0.044)
Native English speaker		0.084	0.101*	0.088	0.104*
		(0.047)	(0.049)	(0.047)	(0.049)
Masters degree			0.088		0.101*
			(0.050)		(0.050)
PhD			−0.084		−0.076
			(0.089)		(0.089)
Degree in quantitative field			0.002		0.012
			(0.048)		(0.049)
STEM			−0.016		−0.046
			(0.045)		(0.048)
Minimal experience with prediction			−0.099		−0.094
			(0.054)		(0.054)
Some experience with prediction			−0.037		−0.079
			(0.057)		(0.065)
Data visualization experience			0.050		0.030
			(0.052)		(0.053)
Coding basics x Trt				0.051	0.058
				(0.101)	(0.101)
Competent coder x Trt				0.173	0.176
				(0.107)	(0.107)
Constant	−0.432***	−0.491***	−0.516***	−0.494***	−0.515***
	(0.030)	(0.055)	(0.087)	(0.077)	(0.098)
P-value (Trt)	0.000	0.000	0.000	0.240	0.311
Observations	298	298	298	298	298
R^2	0.056	0.068	0.101	0.083	0.120

Notes: This table analyzes the effect of the treatment on the worker's score on the prediction problem, -1 times their mean absolute error from the correct answer, where 0 is a perfect score. The first specification is the effect of treatment on their score. The second specification adds controls for gender, office location, native english status, and prior coding ability. The third specification is the same as column (2) but with additional controls for different measures of the workers prior technical experience. The fourth specification is the same as column (2), where the treatment is interacted with the workers prior coding experience. The fifth specification is the same as column (3), where the treatment is interacted with the workers prior coding experience. All reported standard errors are two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001.

Supplementary 2.2.3 Treatment effects on the coding task – Details

Supplementary Table 19 Effect of AI treatment on workers' coding performance, relative to data scientists

	Dependent variable:		
	Data Cleaning	Data Correct	Merge & Filter
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.204***	0.440***	0.738***
	(0.032)	(0.059)	(0.061)
Mean Y in Control Group	-0.26	-0.64	-0.82
P-value (Trt)	0.000	0.000	0.000
Observations	300	300	300
R^2	0.129	0.172	0.343

Notes: This table reports the effect of the treatment on the consultants performance on the coding task. The y-axis is the worker's binary grade on each set of actions in python, where 0 is a perfect score and -1 is the lowest possible score. The first outcome is whether or not the worker correctly cleaned the data. The second outcome is whether or not they correctly handle nulls and duplicates. The third outcome is whether or not they correctly merge and then filter a dataset. The sample includes all experimental participants who submitted something for grading on the prediction task. Text of problems can be found in Supplementary Section 1. Exact definition of grading for each problem can be found in Supplementary Section 1. All regressions include controls for gender, location, native english status, and low tenure. Reported entries are coefficient estimates with two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001.

Supplementary 2.2.4 Treatment effects on the statistics task – Details

Supplementary Table 20 Effect of AI treatment on workers' statistics performance, relative to data scientists

	Dependent variable:		
	Identifying Errors	Visual ML Questions	Probability Questions
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.308***	0.068*	0.302***
	(0.038)	(0.032)	(0.050)
Mean Y in Control Group	-0.48	-0.22	-0.26
p-value (Trt)	0.000	0.033	0.000
Observations	330	330	330
R^2	0.183	0.035	0.109

Notes: This table reports the effect of the treatment on the consultants grades on each step of the statistics task. Each outcome is the worker's grade on each set of problems, where 0 is a perfect score and -1 is the lowest possible score. The first outcome is the workers' score on a problem where they apply statistical metrics and interpret their graphs. The second outcome is their score on questions where they look at data plotted and decide what types of classifier to use on it. The third outcome is their score on probability questions. The sample includes all experimental participants who submitted something for grading on the statistics task. Text of problems can be found in Supplementary Section 1. Exact definition of grading for each problem can be found in Supplementary Section 1. Reported entries are coefficient estimates with two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001. All regressions include controls for gender, location, native english status, and low tenure.

Supplementary 2.2.5 Treatment effects on the prediction task – Details

Supplementary Table 21 Effect of AI treatment on workers' prediction scores, relative to data scientists

	Dependent variable:	
	Mean Absolute Error	Process Score
	(1)	(2)
GenAI Treatment Assigned (Trt)	0.182***	0.129**
	(0.041)	(0.045)
Mean Y in Control Group	-0.43	-0.37
p-value (Trt)	0.000	0.004
Observations	298	281
R^2	0.068	0.036

Notes: This table reports the effect of the treatment on the consultants performance on the prediction task. Each outcome is the worker's grade on each set of problems, where 0 is a perfect score and -1 is the lowest possible score. The first outcome is the score they got on the prediction problem, which is -1 times their mean absolute error. The second outcome is the percentage of "correct" steps they took. The sample includes all experimental participants who submitted something for grading on the prediction task. Text of problems can be found in Supplementary Section 1. Exact definition of grading for each problem can be found in Supplementary Section 1. All regressions include controls for gender, location, native english status, and low tenure. Reported entries are coefficient estimates with two-sided 95% standard errors computed using Huber–White (HC0) robust variance. Significance stars and p-values indicate two-sided tests using Huber–White (HC0) robust variance: * p <0.05, ** p <0.01, *** p <0.001. All regressions include controls for gender, location, native english status, and low tenure.

Supplementary Table 22 Effects of GenAI on Self-Reported Capabilities and Professional Confidence, Before and After Experiment

	Dependent variable:		
	Δ in Beliefs in Data Analytics Abilities	Δ in Confidence in Data Science Skills	Δ in Confidence with GenAI
GenAI Treatment Assigned (Trt)	-0.305	0.651***	-0.087
	(1.673)	(0.114)	(0.199)
Constant	-0.451	-0.343***	0.309*
	(1.170)	(0.086)	(0.143)
Observations	421	463	477
p-value (Trt)	0.855	0.000	0.662
R^2	0.000	0.067	0.001

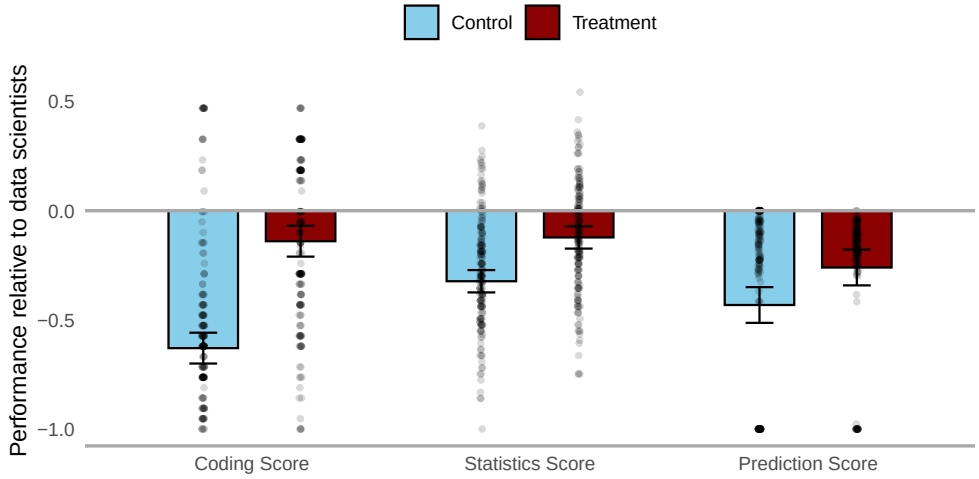
Notes: This table analyzes the effect of treatment on three key outcomes. Column (1) measures the change in respondents' agreement with the statement "Generative AI enables my ability to execute data analytics tasks and reach desired outcomes" on a 1-5 scale from strongly disagree to strongly agree, as measured before and after the experiment. Column (2) measures the change in response to "How confident are you in your ability to contribute to data science projects?" on a 1-7 scale, as measured before and after the experiment. Column (3) measures the change in a composite measure of professional confidence with GenAI, combining responses to three statements: "Generative AI helps me feel more competent in my role", "Generative AI enables my ability to execute tasks and reach desired outcomes", and "Generative AI helps me be more confident that I will meet my project managers expectations". Changes are measured as post-experiment score on the survey question minus the pre-experiment score. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Supplementary Table 23 Effects of GenAI on Self-Reported Beliefs in ChatGPT's Impact on Career Development, Before and After Experiment

	Dependent variable:	
	Δ in Beliefs about ChatGPT's Usefullness for Problem Solving	Δ in Beliefs about ChatGPT's Usefullness for Practicality & Effectiveness
GenAI Treatment Assigned (Trt)	-0.215*	-0.201*
	(0.086)	(0.082)
Constant	0.242***	0.128**
	(0.062)	(0.056)
Observations	478	478
p-value (Trt)	0.013	0.015
R^2	0.013	0.012

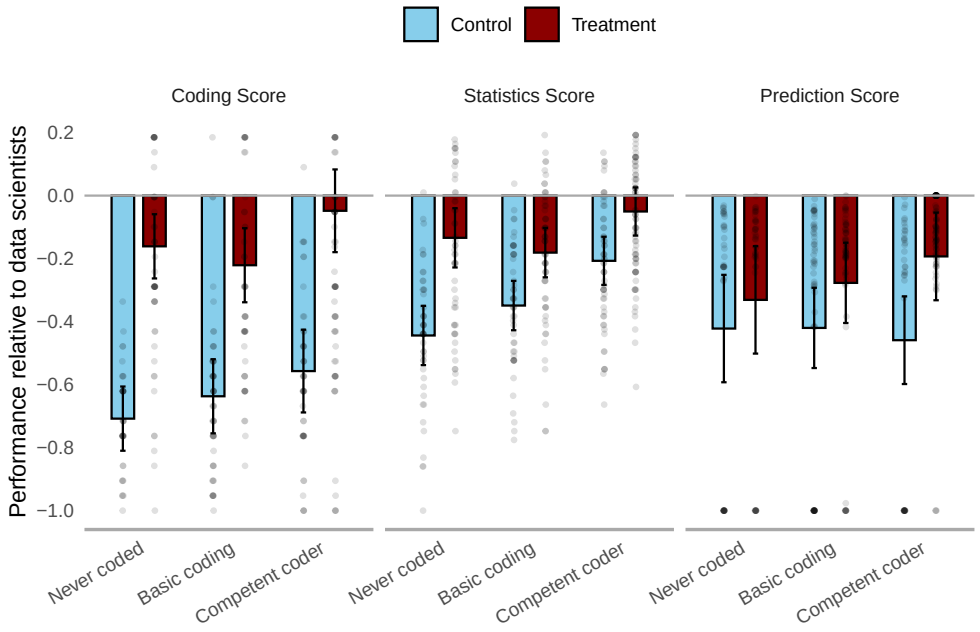
Notes: This table analyzes the effect of treatment on beliefs about how sustained use of ChatGPT for data science would impact career development. Column (1) measures the change in agreement with the statement "Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Problem solving and insights' dimension". Column (2) measures the change in agreement with an analogous statement about the 'Practicality and Effectiveness' dimension. Both questions used a 5-point scale from strongly disagree (1) to strongly agree (5). Changes are measured as post-experiment score on the survey question minus the pre-experiment score. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Supplementary Figure 3 Effect of AI treatment on workers' ability to solve data science problems



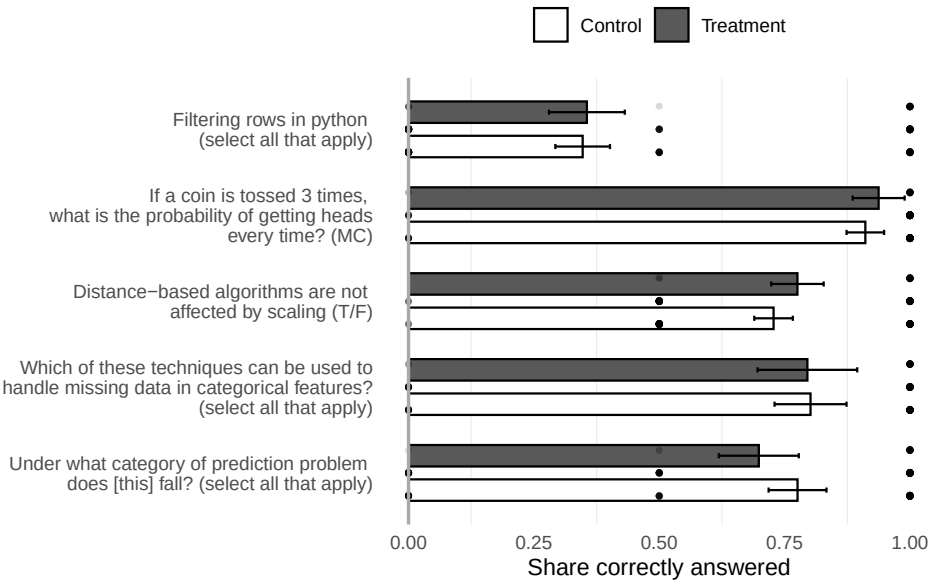
Notes: This plot reports the effect of the treatment on the workers' performance across the 3 tasks. The x-axis is the mean score for each treatment group on each set of problems, where 0 is the average for data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the workers' scores on each statistics question, divided by the total number of possible points. The third outcome is the score they received on the prediction problem, which is -1 times mean absolute error. The error bars indicate the 95% CIs of the mean of the treatment group relative to the mean in the control group. All standard errors are Huber–White robust. The sample includes all experimental participants who submitted a portion of each task for grading. The tasks are described in Supplementary Section 1.2 and the full text is in Supplementary Section 3. Grading for each problem can be found in Supplementary Section 4. Regression details for comparing treatment and control can be found in Supplementary Table 3.

Supplementary Figure 4 Effect of AI treatment on workers' ability to solve data science problems according to their prior coding experience



Notes: This plot reports the effect of the treatment on the workers' scores on each data science task according to their self-reported prior experience with coding. The x-axis is the mean score for each treatment group on each set of problems, where 0 is the benchmark set by the data scientists and -1 means that the worker received no points on the task. The first outcome is the percentage of correct steps that they take in answering the coding question. The second outcome is the sum of the workers' scores on each statistics question, divided by the total number of possible points. The third outcome is the score they received on the prediction problem, which is -1 times their mean absolute error. Huber–White robust standard errors are based on the difference between the means of the treatment and control groups, and we plot two-sided 95% CIs around each estimate. Because the data scientists are all at least competent coders, we cannot compare the performance of each worker in the treatment group to that of the data scientists interacted with coding experience. The sample includes all experimental participants who submitted a portion of each task for grading. N per cell: Never coded (Control: N = 67; Treatment: N = 79), Basic coding (Control: N = 82; Treatment: N = 90), Competent coder (Control: N = 76; Treatment: N = 92). The tasks are described in Supplementary Section 1.2 and the full text is in Supplementary Section 3. Grading for each problem can be found in Supplementary Section 4. Regression details can be found in Supplementary Table 10.

Supplementary Figure 5 Effects of AI treatment on post-experiment data science knowledge without the use of GenAI



Notes: This analysis looks at the effect of treatment on workers' ability to correctly answer data science questions after the conclusion of the experiment. The x-axis is the mean probability of getting the correct answer for each treatment group. The y-axis has the text of each question, with the format of the answer. We report Huber–White (HC0) robust standard errors and two-sided 95% confidence intervals. The text of the questions can be found in Supplementary Section 5.2. Regression details can be found in Supplementary Table 7.

Supplementary 3 Details of tasks

(a) Statistics task

(i) Instructions – Control

Please read through the instructions to complete the next task !

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.

For this task, **DO NOT USE ChatGPT** but you can use Google searches or other tools as you normally do for casework. **However, do not Google image search and refrain from copying and pasting the exact question.** Do not spend more than 90 minutes on this task.

We highly encourage you to do your best to complete this task and while it might be challenging sometimes, we truly appreciate the effort. **Don't forget that top 50th percentile,** as judged by quality of answers, with access to similar resources, will be noted as such to their CDA.

(ii) Instructions – ChatGPT

Please read through the instructions to complete the next task !

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.

For this task, **USE ChatGPT** enterprise version by accessing https://chat.openai.com/ using your BCG login to perform the task below and answer all the questions. **However, do not send any images to GPT and refrain from copying and pasting the exact question.** Do not spend more than 90 minutes on this task.

We highly encourage you to do your best to complete this task and while it might be challenging sometimes, we truly appreciate the effort. **Don't forget that top 50th percentile,** as judged by quality of answers, with access to similar resources, will be noted as such to their CDA.

(iii) Task

Instructions :

For this task, **USE ChatGPT** enterprise version by accessing https://chat.openai.com/ using your BCG login to perform the task below and answer all the questions. **However, do not send any images to GPT and refrain from copying and pasting the exact question**. [For this task, **DO NOT USE ChatGPT** but you can use Google searches or other tools as you normally do for casework. However, **do not Google image search and refrain from copying and pasting the exact question**.]

Question 1: The following is the first five rows of data containing financial and demographic information about domestic partners who have co-purchased a home in the last several years.

Age 1	Age 2	Income 1	Income 2	Borough	ZIP Code	Date	Price	Mortgage
39	37	270000	180000	Manhattan	10076	1 January 2016	1,125,000	Yes
NULL	38	445000	670000	Manhattan	10025	1 January 2016	2,249,000	Yes
27	29	145000	225000	Queens	11106	2 January 2016	900,000	Yes
33	NULL	90000	76000	Brooklyn	11203	2 January 2016	415,000	Yes
68	55	78000	450000	Bronx	10474	2 January 2016	3,399,000	No

1. You have been tasked with predicting based on demographics and price whether a mortgage was taken out to by the house. You prompt ChatGPT for detailed instructions on how to do this, and ChatGPT recommend using a logistic regression model. It recommends the following steps.

1. Understand Your Dataset

- **Explore and Preprocess**: Start by exploring your dataset to understand the features available and their types (numerical, categorical). Clean the data by handling outliers and possibly irrelevant features. Preprocessing steps like encoding techniques (e.g., one-hot encoding) might be necessary for categorical data. Ensure that your dataset does not have missing values. You can either fill them in with a strategy (mean, median, mode) or remove the rows/columns with missing values, depending on the situation.
    a. Which of the following are among the steps you could take to address this point? Select all that apply.
        i. Plot the distribution of each of the numerical variables and remove rows with outliers from this dataset
        ii. One-hot encode the 'Borough' variable
        iii. Investigate relationships between variables
        iv. Convert date to a numerical variable
        v. One-hot encode the ZIP code variable
        vi. One-hot encode the age variables
        vii. Bin the ZIP codes by neighborhoods and do not process further
        viii. Bin the ZIP codes by neighborhoods and one-hot encode
        ix. Check columns with null values and remove those with >80% missing values
        x. Impute NULL values by using a summary statistics or by developing a simple model that predicts those values based

2. Split the Data

- **Train-Test Split**: Divide your dataset into a training set and a testing set (commonly a 70-30 or 80-20 split) to evaluate the model's performance on unseen data.

3. Train the Model

- **Training**: Use the training dataset to train your model, adjusting parameters as needed. For complex models, consider using cross-validation to fine-tune hyperparameters and prevent overfitting.

4. Evaluate the Model

- **Performance Metrics**: Evaluate your model on the test set using appropriate metrics such as accuracy, precision, recall, F1 score, and the ROC-AUC curve. These metrics will help you understand how well your model is performing in terms of both its ability to predict mortgages correctly and its robustness against false positives or negatives.

b. What issue necessitates using all these metrics? Which of the above steps is affected by this issue and how? (Answer in 100 words or less – bullet points ok)
c. Would you change the order of any of the above steps? Why or why not? (Answer in 100 words or less – bullet points ok)

2. You want to try a k-Nearest Neighbors model. Which of the following are not required (although recommended) for logistic regression, but absolutely necessary for k-Nearest Neighbors? Select all that apply.
    a. Transform numerical variables (e.g. log)
    b. Make sure there are only two classes to predict
    c. Convert Mortgage column from string to binary
    d. Standardize numerical variables
    e. Impute the missing age with the other age in the same row
    f. One-hot encode the appropriate variables
3. You also try a decision tree model for the same classification problem, to compare performance. You realize your model is performing quite poorly on both training and validation sets. You double-check the code and there are no bugs. What could be causing this problem? Select all that apply.
    a. Your model is underfit
    b. Your mode is overfit
    c. The learning rate hyperparameter is too small
    d. The learning rate hyperparameter is too large
    e. The decision tree is too shallow
    f. The decision tree is too deep
    g. None of the above
4. Next, you have been instructed to predict the price based on the other variables, and this time you have been instructed to use linear regression. Following instructions from ChatGPT, you perform a basic linear regression. You notice that your R^2 value is too low. You prompt ChatGPT for suggestions on how to diagnose the problem, and it is recommended that you check the residual plots. You notice that the residual plot does not appear random. What could this mean? Select all that apply.
    a. The observed values of your dependent variable are independent from each other
    b. Your model is missing an important variable
    c. There is some interaction between your variables
    d. A higher order term might be required in your regression
    e. Variance of the residual is the same for any value of X
5. For the following residual plots, what could be the characteristics of or issues with the data or model that are corresponding with these results (choose from the list provided for each image)? It is possible that more than one characteristic or issue applies to any given image, and it is possible that a characteristic or issue may apply to more than one image.

<u>Image A</u>

197

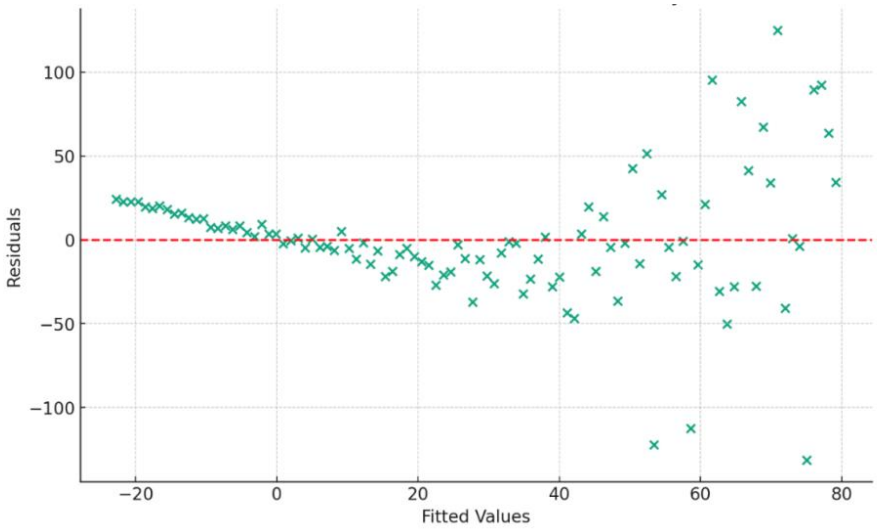


Image B

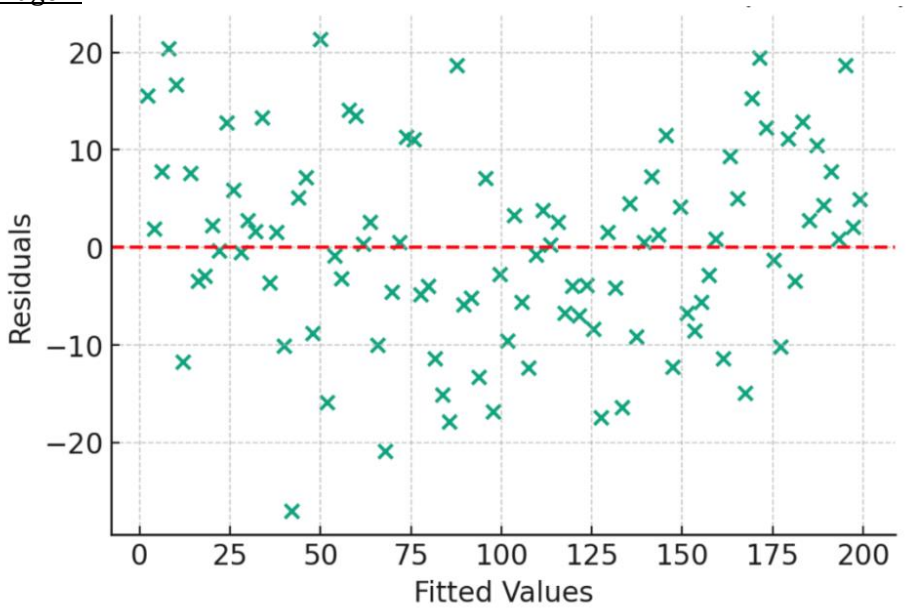


Image C

198

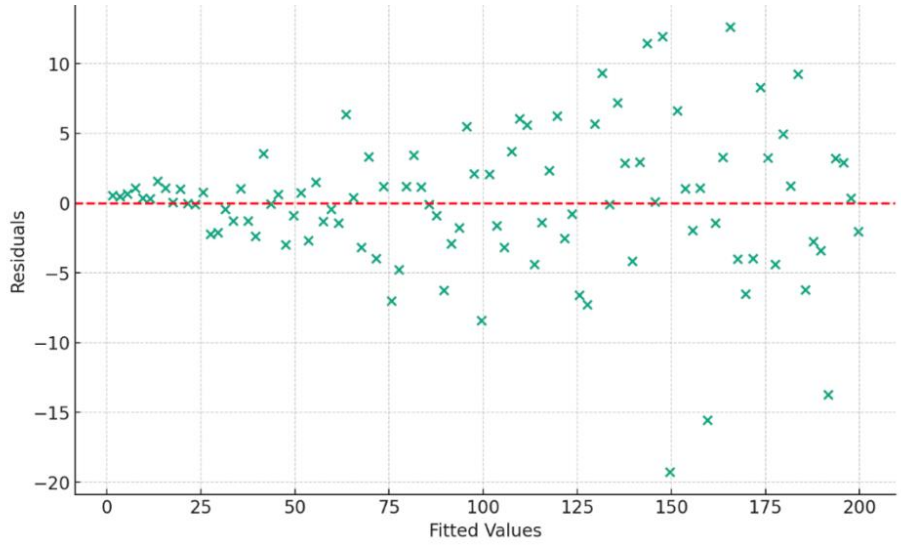


Image D

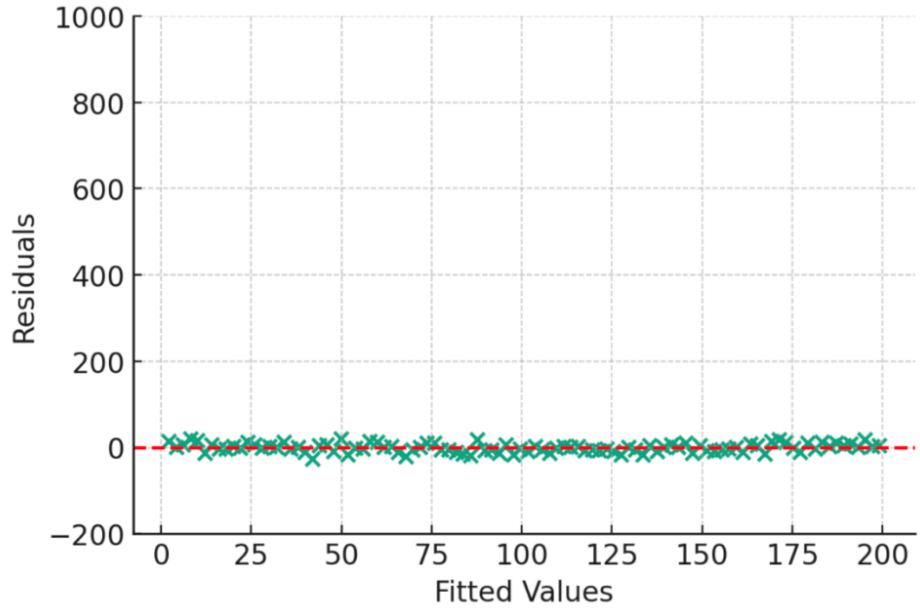


Image E

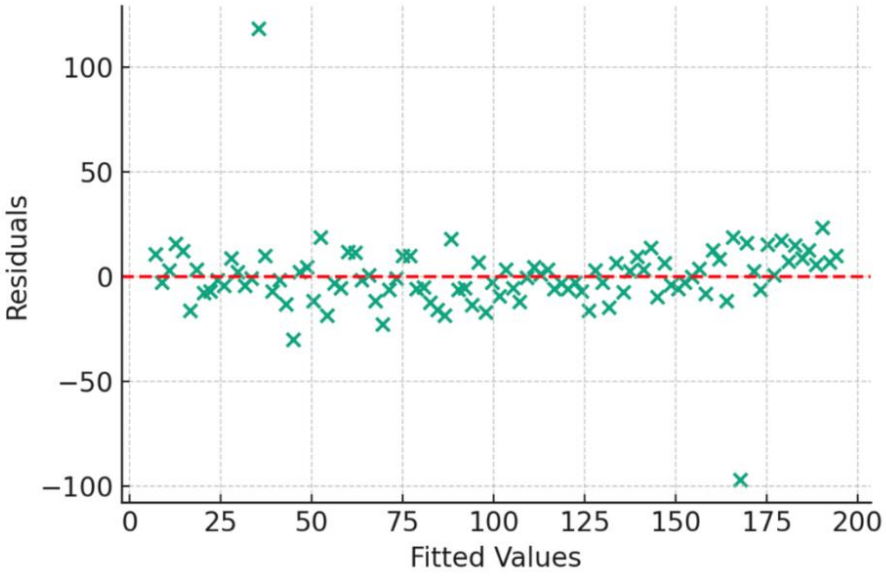


Image F

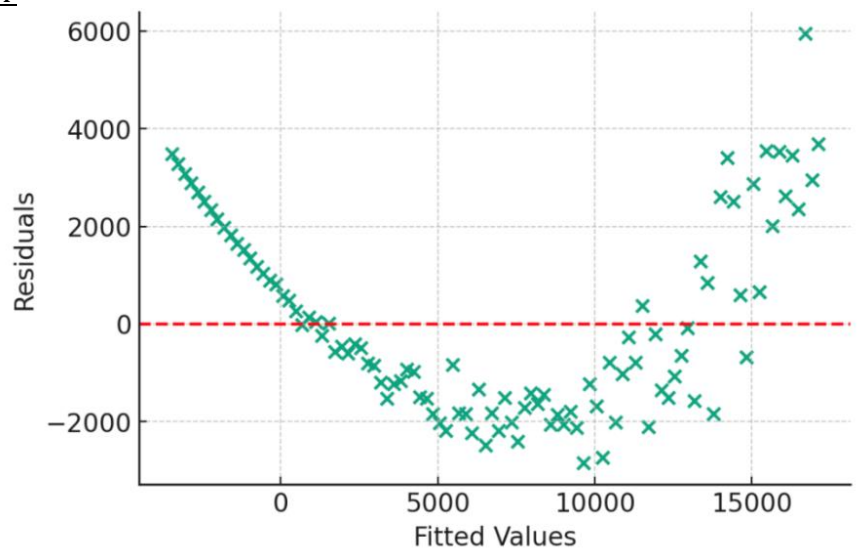
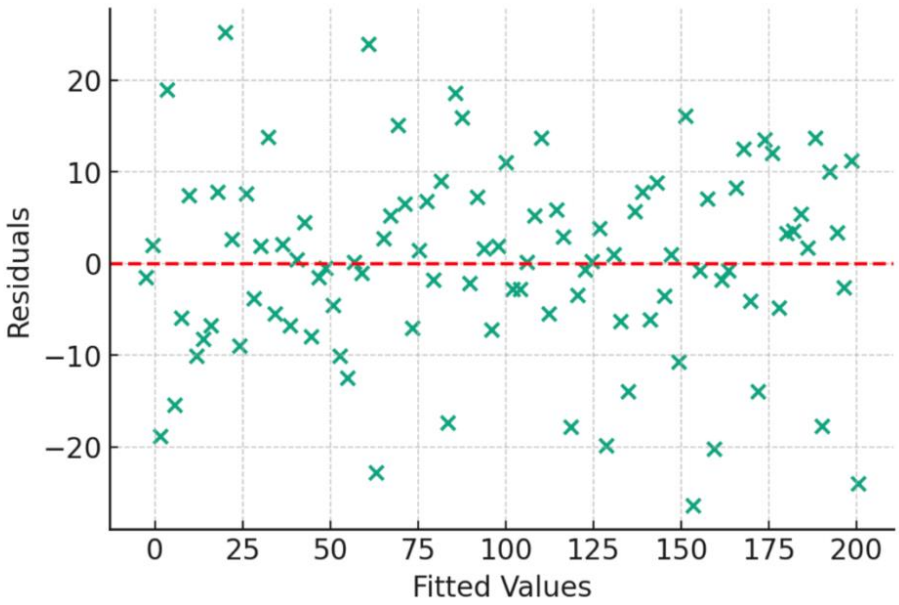


Image G

200



Characteristic or issues choices:

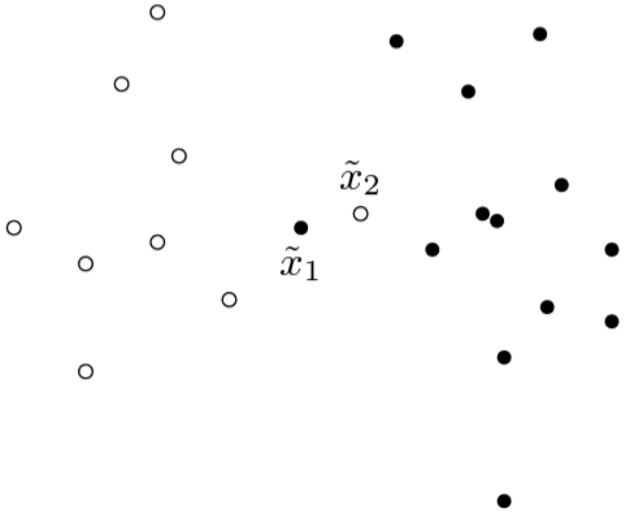
a) No characteristic or issue is apparent
b) Heteroscedastic data
c) Outliers
d) Response variable requires transformation
e) A higher order variable might be required

6. You are asked to train a new model to predict price on the newest version of the dataset. In this version, there are several more fields collected with demographic information and financial information of the couples. However, this data is only from the last month. Which of the following steps recommended by ChatGPT could be beneficial to take to address some of the issues that are likely to arise because of this? Select all that apply.
    a. Perform PCA
    b. Use a neural network instead of linear regression
    c. Use a regularized model instead of linear regression
    d. None of the above

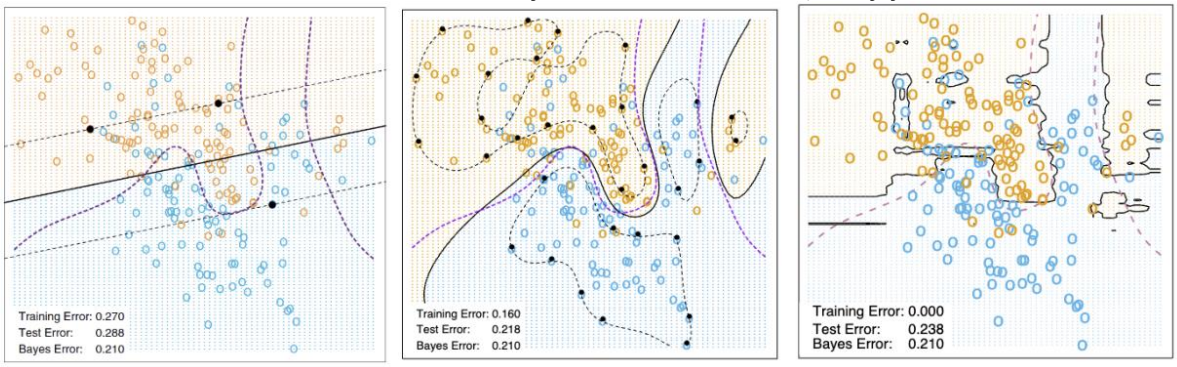
Question 2:

1. You are asked to prepare a simple linear model to classify the following points into class 1 (black dots) and class 2 (white dots). What is the best empirical risk of this model that you can achieve with 0-1 loss? Justify your answer and show your working steps.

201



2. ChatGPT has run 3 classifiers on your data and provided a visual output, but not specified which models yielded which output. For each of the three images, name a classifier that could create the boundary represented by the solid black line, and one that could not (class 1 is the orange dots, and class 2 is the blue dots). You can ignore the dashed line, you can use the metrics on the bottom-left but you do not need them. Justify your answer.



(a) (b) (c)

Question 3: Imagine that you are a farmer with two potential fields to plant your crops, Field A and Field B. You estimate that the chances of achieving a high yield are 65% for Field A and 35% for Field B, based on the soil quality. Given the size and weather conditions of the fields, if you choose to plant in Field A, you have a 45% of achieving the yield in a season. If you choose to plant in Field B, you have a 75% chance of achieving a high yield in a season due to its smaller size and more manageable conditions.

1. If you must choose one field to plant this season to maximize your yield, which field should you choose? Justify your answer and show your steps.
2. Assume you made the rational decision for the first planting season, but the yield was not as expected. You have another chance to plant – has your initial idea of which field is more likely to yield a high output changed? Where should you plant now? Justify your answer and show your steps.

Optional break
Please feel free to take an optional break if you desire. The next section (once you advance), will be a timed 60-90 minute section focusing on a data science task.

Please keep breaks to under 10 minutes if possible to not disturb the flow of your engagement. Thank you!

1. **Are you ready to proceed to the next task which might take 60-90 minutes**? **If so, please select "Yes" and press next**
    - Yes

(b) Problem-solving task

(i) Instructions – Control

Please read through the instructions to complete the next task!

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.

You CANNOT USE ChatGPT or GPT-4 or any other Generative AI tool for the set of questions below. However, you may choose to use Google searches or other tools as you normally do for casework, if you would like to.

We highly encourage you to do your best to complete this task and while it might be challenging sometimes, we truly appreciate the effort. **Don't forget that top 50th percentile**, as judged by quality of answers, with access to similar resources, will be noted as such to their CDA.

(ii) Instructions – ChatGPT

Please read through the instructions to complete the next task!

*You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.*

For this task, USE ChatGPT** enterprise version by accessing https://chat.openai.com/ using your BCG login. **Feel free to work with ChatGPT in whatever way you like (uploading files, copying and pasting the question or results)

***We highly encourage you to do your best to complete this task** and while it might be challenging sometimes, we truly appreciate the effort. **Don't forget that top 50th percentile**, as judged by quality of answers, with access to similar resources, will be noted as such to their CDATask*

(iii) Task

Instructions:

For this task, **USE ChatGPT** enterprise version by accessing https://chat.openai.com/ using your BCG login. Feel free to work with ChatGPT in whatever way you like (uploading files, copying and pasting the question or results). [For this task, you **CANNOT USE ChatGPT** or GPT-4 or any other Generative AI tool for the set of questions below. However, you may choose to use Google searches or other tools as you normally do for casework, if you would like to.]

Question

Imagine you are staffed on a case where you must implement a data-driven strategy for sports investing. You are given a dataset containing records of 45,360 international football matches, spanning from the inaugural official match in 1872 through to the year 2024. The competitions range from the FIFA World Cup, FIFI Wild Cup, to ordinary friendly games. All matches are men's senior internationals, excluding Olympic Games, matches involving B-teams, U-23, or league select teams.

Your task (make sure to describe and document your approach and your findings):

1) Develop and implement a method for quantifying how predictable each match result was. You can solve this problem however you like, using any analytics platforms at your disposal (e.g. Excel, Alteryx, Python).
    a. Explain in detail each step you took for your approach and justify.
    b. What was the most surprising match result in this dataset, based on your method?
    c. Return a .csv or Excel file containing four columns – the match date, the home team, the away team, and your numerically determined match result predictability using the above method for each match in the dataset

Keep in mind you have 90 minutes to complete this task. Time box and make sure you return a final answer.

DATASET INFORMATION

The `results.csv` file encompasses columns for:

- `date` - the match date
- `home_team` - the home team's name
- `away_team` - the away team's name
- `home_score` - home team's score at the end of the match, including extra time but excluding penalties
- `away_score` - away team's score at the end of the match, including extra time but excluding penalties
- `tournament` - tournament name
- `city` - the city or locality of the match
- `country` - the country hosting the match
- `neutral` - a TRUE/FALSE indicator of whether the venue was neutral

Assume that the result as shown in this dataset (win or tie) is the entire result – there are some cases of penalty shootouts and goals scored from penalties, but for complexity, we will ignore those for this exercise.

For clarity, current names are used for both teams and countries in historical matches. For example, an 1882 match featuring the team then known as Ireland against England is listed under Northern Ireland, reflecting the modern successor of the 1882 team. Country names are recorded as they were at the time of the match, but discrepancies between team and country names (e.g., Ghana vs. Gold Coast) are accounted for, with the `neutral` column marking such matches as non-neutral to clarify they were played at home.

Optional break
Please feel free to take an optional break if you desire. The next section (once you advance), will be a timed 90 minute section focusing on a data science task.

Please keep breaks to under 10 minutes if possible to not disturb the flow of your engagement. Thank you!

2. **Are you ready to proceed to the next task which might take 60-90 minutes? If so, please select "Yes" and press next**
    - Yes

(c) Coding task

Before we get started, you need to setup Colab as the coding environment. Here's a link to the documentation (Colab_Setup_Insturctions).

(i) Instructions – Control

Please read through the instructions to complete the next task!

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically**

submit and move you to the next portion of the study, regardless of your progress. You may choose to advance prior to the 90 minutes if you have fully completed the task.

You CANNOT use ChatGPT or GPT-4 or any other Generative AI tool for the set of questions below. However, you may choose to use Google searches or other tools as you normally do for casework, if you would like to.

We highly encourage you to do your best to complete this task and while it might be challenging sometimes, we truly appreciate the effort. **Don't forget that top 50th percentile**, as judged by quality of answers, with access to similar resources, will be noted as such to their CDA.

(ii) Instructions – ChatGPT

Please read through the instructions to complete the next task !

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.

USE ChatGPT enterprise version by accessing https://chat.openai.com/ using your BCG login to perform this task **in whatever way you like (uploading files, copying and pasting the questions or errors and results).**

We highly encourage you to do your best to complete this task and while it might be challenging sometimes, we truly appreciate the effort. **Don't forget that top 50th percentile,** as judged by quality of answers, with access to similar resources, will be noted as such to their CDA.

(iii) Task

Assignment (Qualtrics)

You will find your shared folder located at the following link : ${e://Field/google_drive_link}

USE ChatGPT enterprise version by accessing https://chat.openai.com/ using your BCG login to perform this task in whatever way you like (uploading files, copying and pasting the questions or errors and results). [**DO NOT USE ChatGPT** or any other Generative AI tool for the set of questions below. However, you may choose to use Google searches or other tools as you normally do for casework, if you would like to.]

Use the CoLab document found in this link to complete the Data Science Task. You will first need to follow the instructions contained in the "Colab_Setup_Instructions.docx" file located within this shared folder.

Make sure to write and run the Python code needed to solve the question in this Google Colab notebook so that we may review your code. **Your grade will be determined by the code present within the Colab notebook. Code stored in any other location will not be considered for evaluation. Therefore, please ensure that all your work is fully contained within Colab.** Return to this survey when you are finished with the task in Colab. For instructions on how to set up and use Colab, view the videos below:

If you have never worked in a Colab or Jupyter notebook before, you may find the following link helpful for some instructions on how to create, delete, and run cells :
https://www.tutorialspoint.com/google_colab/your_first_colab_notebook

Assignment (Instructions file)

Make sure to write and run the **Python code** needed to solve the question in this **Google Colab** notebook so that we may review your code.
Overview of Datasets
Dataset 1: Orders Data (orders.csv)

- **File Type**: CSV
- **Contents**:
    - customer_id: The unique identifier for each customer.
    - order_info: Information about each order in the format order number ; date and time. The order number in this dataset (once decoupled from the date and time) corresponds with that in the next.

Dataset 2: Products Data (products.csv)

- **File Type**: CSV
- **Contents**:
    - customer_id: The unique identifier for each customer associated with an order.
    - order_id: ID of each order in the format order number. The order number in this dataset corresponds with that in the previous once decoupled from the date and time.
    - order_products: Details about the products in each order in the following format: {product_id: [product_cost, product_quantity], ...}. Each product is sold either at its original price or a 20% discount.

Commonalities

- Both datasets share the customer_id field and order_id information with the order number.
- Each order ID is unique, whereas customer IDs may be repeated across multiple orders. However, each combination of order ID and customer ID is unique.
- Whenever you have information about either one of these, it is correct.

Data Quality and Cleaning Guidelines

- **Order and Customer IDs**: Entries are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere.
- **Date and Time Fields**: Entries with incorrect values should be removed.

- **Product Quantities and Product IDs**: Always correct unless marked as NULL, which indicates missing values.
- **Product Costs**: Each product ID is associated with a unique cost. Sometimes, the original unique cost is discounted at 20% so that is what is shown in a certain order. However, when not discounted, sometimes there are junk or NULL values instead of the correct cost. Junk or NULL values in the product costs should be replaced with the original cost (the discounted cost should be left as is wherever it is shown but not added in elsewhere).

Tips for Handling Junk and NULL Values; duplicates

- Examine common values in each column to identify patterns and potential corrections.
- Attempt to fix junk or NULL values using information elsewhere in the data before considering row deletion.
- Date time fields can have incorrect fields that are not correctable, discard the affected rows and values to maintain data integrity.
- Check for duplicates at every stage

QUESTION:
Use the above datasets to answer the question: **Which 5 customer IDs had the highest average order by cost in May 2022?**

- **Step 1**: Start with loading and viewing the data
- **Step 2**: Split and clean the orders data
  Remember that entries for customer IDs and order IDs are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere. Date and time fields entries with incorrect or junk values should be removed. And duplicates should be dropped.
- **Step 3**: Clean the products data
  Remember that entries for customer IDs and order IDs are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere.
  Product Quantities and Product IDs: Always correct unless marked as NULL, which indicates missing values.
  Product Costs: Each product ID is associated with a unique cost. Sometimes, the original unique cost is discounted at 20% so that is what is shown in a certain order. However, when not discounted, sometimes there are junk or NULL values instead of the correct cost. Junk or NULL values in the product costs should be replaced with the original cost (the discounted cost should be left as is wherever it is shown but not added in elsewhere).
  Don't forget to drop duplicates!
- **Step 4**: Merge the data sets to answer the question: Which 5 customer IDs had the highest average order by cost in May 2022?

Optional break
Please feel free to take an optional break if you desire. The next section (once you advance), will be a timed 60-90 minute section focusing on a data science task.

Please keep breaks to under 10 minutes if possible to not disturb the flow of your engagement. Thank you!

1. **Are you ready to proceed to the next task which might take ~90 minutes? If so, please select "Yes" and press next**
    - Yes

(d) ChatGPT (or other search) conversation

1. **Did you use ChatGPT to help complete your task?**
    - Yes
    - No

2. **If you answered "Yes" above, please "Share" your ChatGPT conversations with us, so we can better understand how ChatGPT assisted you with the tasks. Only include conversations that you used for the task you just completed. If you used multiple ChatGPT conversations for the task, please share a link to each one.**

    NOTE : *Please do not delete your conversations for a week or so, so we can make sure to collect the data we need to from them. If you delete the conversation on your side, we will no longer be able to view your shared links.*

    The screenshot below shows how you can Share your conversation. Select the 3 dots on the Individual Conversation tab, and select "Share", then copy the link and paste it below.

3. **Did you use Google to help complete your task?**
    - Yes
    - No

4. **Please explain any other tools you used to complete your tasks. Include the name of the tool used, and how you used it to assist you.**
    Be as specific as you can.

    Ex :
    "Excel - I used excel to create a model and analyze the data.
    Stack Overflow - I used Stack Overflow to help with coding questions."

Supplementary 4 Grading

Supplementary 4.1 Grading Overview

211

Task Grading

Use of LLMs in grading

1. LLMs were used in the grading of all three tasks. Specifically, the API endpoint of the model 'gpt-4-turbo' was used to grade (which, at the time of grading, pointed to `pt-4-turbo-2024-04-09`).

To mitigate, measure and manage the well-known issues of inconsistency and inaccuracy with LLMs, we designed the following grading architecture

- Processing responses to mitigate bias
    - First, participant answers were preprocessed to minimize GPT-4's known bias to prefer its own answers and answers of a certain length. For detailed text-based answers (more details on the task-specific descriptions below), the LLM was asked to paraphrase and summarize the student answers, so that there would be no obvious difference between the group assisted by ChatGPT and the group without access to ChatGPT. Human validation was done to verify that this step significantly minimized GPT's bias by confirming that the grading discrepancy between the treated and control groups did shrink after this processing step.
- All grading was run using randomized batching
    - Randomized, overlapping batches of participants' answers were prepared, with each participants' answers appearing in a batch at least 5 times. This allows for each participants' answer to appear in batches with different answers for each of the 5 iterations of being graded for a single answer. This mitigates bias that may be caused by comparison to the answers of the other participants in the batch. For example, the appearance of an average answer in a batch with alongside a very poor answer might overinflate the grade by comparison (due to the LLM perceiving one answer as a lot better than the other). In addition, the appearance of an average answer in a batch with an exceptionally good answer might deflate the grade by comparison. Randomizing batches, having the answer graded 5 times in comparison with different answers, and comparing the scores across batches allows any variability to be caught and studied. Studying variability across batches helped make decisions about which prompts or prompting strategies yield most consistent results. For any individual answer by a participant, we kept answers for which at least 3 answers were identical, the other 2 were within an acceptable range, and averaged the 5 once these conditions were met.
- Prompting strategies were separately defined for each question
    - For each task, and for each underlying question or question type, several prompting strategies were tested. Some prompting strategies included providing rubrics with how many points should be assigned for the completion of certain logical steps with all steps in one query. Alternatively, each step in a query of its own, asking yes/no questions about whether a student had provided a specific

answer or a shown a specific step of work towards calculating a specific answer, yes/no questions about the manner in which the student answered the question, one-shot or few-shot examples for how the student may have solved each question or a whole or sub-step of each question, and even allowing the LLM room to exercise judgment and deviate from the rubric when an answer was not reflected exactly in what the rubric planned for. The accuracy of the prompting strategies was determined by randomly selecting answers for human grading and comparing the human-assigned grade with the LLM-assigned grade. The final prompting strategies were chosen to minimize variability across aforementioned batches and maximize accuracy.

- All steps and final grading were human-validated
    - For each task, a ground truth set of grades was created by a Data Scientist for each substep of each question. Having a ground truth made it possible to validate each batch of grades output by the LLM with an objective reality of what the grades should be, and batches of LLM outputs of grades of participant responses were able to be judged by common metrics such as mean squared error from the ground truth. This mean squared error metric was used in gauging the accuracy of prompt engineering and when deciding what prompting strategy should be used for each question.
    - The language of the rubrics was adjusted throughout grading to accommodate the semantic requirements of LLM-based grading. After each round of testing of participants' answers on each task, the rubric was adjusted as needed when the following issues arose
        - LLM was instructed in most cases to abstain from grading a participants' response if the response was not explicitly reflected in the rubric with instructions on how to score or manage. When answers came up that were not reflected explicitly in the rubric, the answers were flagged for human review and subsequently added to the rubric with the appropriate treatment of points assigned based on human (Data Scientist) judgment.
        - The original rubric submitted for pre-registration was insufficient for an LLM to provide accurate scores. To ensure accuracy of LLM-based grading, the language and grading-logic was simplified. Specifically, the process of grading each question was broken down into smaller steps. Sub-steps had to be defined with sub-points assigned to them based on human (Data Scientist) judgment, with tasks broken down to the smallest assessable pieces. However, the cumulative point distributions and rankings of step-difficulty were kept constant against the pre-registered rubric.

More specific examples of how this framework was deployed across tasks will follow in descriptions of each task below.

Statistics Task

Grading

An answer key was created for the Statistics task detailing the correct answers to each question and the correct steps it would take to arrive at the correct answer. A grading rubric was assigned to each correct answer in the answer key by assigning working steps with point values based on how difficult they were to execute correctly, and correct answers a point value based on how difficult they were to arrive at (where applicable, correct answers without correct work were not awarded full points). The answer key and the rubric were reviewed and validated for correctness and validity in difficulty assessment and point assignment respectively by a few Lead Data Scientists (or Data Scientists with 4+ years of experience at BCG, likely longer in the industry).

Some of the questions in the Statistics task allowed for free response answers. These answers were graded with the help of LLMs.

<u>LLM Grading for Statistics</u>

For the statistics task, the LLM grading framework was deployed in the following way

- Preprocessing
    - Most answers were not preprocessed and were provided to the LLM for grading as is. However, all answers to the Bayesian probability question were summarized and paraphrased by the LLM before grading. This is because extensive testing and validation with the help of the ground truth set showed that without this step, students assisted by ChatGPT were scored fairly while students not assisted by ChatGPT were consistently underscored on this question. After investigation, this was identified to be because when answering the questions, the ChatGPT group had more clear and verbose explanations for the answers and calculations, probably copied directly from ChatGPT output. These more clear and verbose explanations made these answers much easier for the LLM to process and ‘understand’. By contrast, participants not assisted by ChatGPT suffered due to their relative inability to explain in as precise or technical terms, or in clear language (especially given time constraints), why they why they took the steps they were taking, even if those steps were correct. Therefore, to avoid a bias towards LLM answers, all answers were at first paraphrased by the LLM in context of the question (summarization prompts were very clear that the output should not be augmented or changed from the original answer in any way, and the outputs were validated by human review extensively). This increased the accuracy and consistency of results.

- Randomized batching
    - Each answer in the statistics task was batched with four other random answers to the same question and graded in a single query by the LLM. This was done to minimize the bias arising from comparison as described in the overall LLM grading architecture. Each answer was graded at least 5 times in order for the variability and consistency of grades to be quantified. Most prompting strategies for the statistics task (described below) asked the LLM to answer yes/no questions about the answers provided by the participants – yes or no questions have an implied consensus and 3 yes or no answers across 5 runs provides a clear tie-breaker. The questions about understanding and correcting ChatGPT's proposed machine learning methodology were not scored by the yes/no strategy because it did not perform well on these questions, which were much too open-ended. For this, a clear rubric was provided of possible answers and sub-answers and how many points should be provided for each. 2% of the answers in the dataset were flagged as 'high variability' using this approach – meaning that the LLM failed to provide the same exact score 3 or more times for the same answer. These were validated and graded manually. For any answers where the LLM provided the same score 3 or more times, a consensus score was established and the score was established as long as it had been validated by the ground truth set.
- Prompting
    - As touched on in the batching section, two primary prompting strategies were employed for the answers in the statistics task. The first, performing better on more open-ended questions, providing a rubric of potential answers and how many points they should be awarded. Here is an example of the same:

```
 You are a grader for a batch of students on the following
exercise. You must assign points to each of the students'
answers based on the rubric provided. DO NOT stray from the
rubric and provide points for anything that is not mentioned in
the rubric explicitly.

  EXERCISE QUESTION:

  Performance Metrics: Evaluate your model on the test set using
appropriate metrics such as accuracy, precision, recall, F1
score, and the ROC-AUC curve. These metrics will help you
understand how well your model is performing in terms of both
its ability to predict mortgages correctly and its robustness
against false positives or negatives.
  What issue necessitates using all these metrics?
```

```
  ANSWER RUBRIC:
  This answer carries a total of 3 points.
  If the student correctly identifies that the issue with the
data is imbalanced or unbalanced, or that the classes of
positive and negative are imbalanced or unbalanced, assign 3
points.
  OR
  If the student mentions stratefied sampling or undersampling,
assign 2 points.
  OR
  If the student recognizes that overfitting might be an issue,
assign 1 point.

  Return your score assignment in the following format
  STUDENT 1 SCORE: points
  STUDENT 1 EXPLANATION: explanation
  STUDENT 2 SCORE: points
  STUDENT 2 EXPLANATION: explanation
  ...
```

- The second, performing better on questions expecting a specific set of answers or answers with justification through calculations (less open-ended), yes or no questions were asked about the answers and steps taken to answer the question, with each answer broken up into several substeps asked about separately. Point values were assigned based on a logic defined for each group of yes/no questions. Here is an example of the same for answering one question:

```
StatsQ2.A.1: |-
  You are a grader for a batch of students on the following
exercise. You must grade the answers of the students to the
following exercise question according to the rubric.

  EXERCISE QUESTION:
  You are asked to prepare a simple linear model to classify the
following points into class 1 (black dots) and class 2 (white
dots). What is the best empirical risk of this model that you
can achieve with 0-1 loss? Justify your answer and show your
working steps.

  ANSWER RUBRIC:
```

Did the student identify that there will be only 1 misclassification for the model with best empirical risk?

Return your answer in the following format
STUDENT 1 RESULT: yes or no
STUDENT 2 RESULT: yes or no
...

StatsQ2.A.2: |-
You are a grader for a batch of students on the following exercise. You must grade the answers of the students to the following exercise question according to the rubric.

EXERCISE QUESTION:
You are asked to prepare a simple linear model to classify the following points into class 1 (black dots) and class 2 (white dots). What is the best empirical risk of this model that you can achieve with 0-1 loss? Justify your answer and show your working steps.

ANSWER RUBRIC:
Did the student explicitly return the final answer as 2/22 (0.0909) or 20/22 (0.9090)?

Return your answer in the following format
STUDENT 1 RESULT: yes or no
STUDENT 2 RESULT: yes or no
...

StatsQ2.A.3: |-
You are a grader for a batch of students on the following exercise. You must grade the answers of the students to the following exercise question according to the rubric.

EXERCISE QUESTION:
You are asked to prepare a simple linear model to classify the following points into class 1 (black dots) and class 2 (white dots). What is the best empirical risk of this model that you can achieve with 0-1 loss? Justify your answer and show your working steps.

```
  ANSWER RUBRIC:
  Did the student explicitly return the final answer as 21/22
(0.9545)?

  Return your answer in the following format
  STUDENT 1 RESULT: yes or no
  STUDENT 2 RESULT: yes or no
  ...
StatsQ2.A.4: |-
  You are a grader for a batch of students on the following
exercise. You must grade the answers of the students to the
following exercise question according to the rubric.

  EXERCISE QUESTION:
  You are asked to prepare a simple linear model to classify the
following points into class 1 (black dots) and class 2 (white
dots). What is the best empirical risk of this model that you
can achieve with 0-1 loss? Justify your answer and show your
working steps.

  ANSWER RUBRIC:
  Did the student explicitly return the final answer as 21/22
(0.0454)?

  Return your answer in the following format
  STUDENT 1 RESULT: yes or no
  STUDENT 2 RESULT: yes or no
  ...
```

    - All prompting strategies were selected for maximal accuracy and minimal variability across answers in batches of answers (maximal consistency).
- Manual validation
    - For each free response answer in the statistics task, ground truth label sets of over 40% of the total corpus of answers were graded by a Data Scientist. Each answer was validated to have maximum accuracy for this 40% with the assumption that the accuracy extended to the rest of the data.
- Rubric semantic adjustment
    - For the questions answered with the scoring rubric strategy, the LLM was initially asked to abstain from answering questions not explicitly mentioned in the rubric. For the yes/no strategy, the LLM initially returned some answers that all

contained no responses. The abstentions and no answers were studied in detail, and where an answer or answer type was missing from the rubric, it was added in along with an assignment of points decided by a Data Scientist. This ultimately ensured that all answers were scored according to a rubric as intended by the rubric designer, rather than arbitrarily through LLM 'judgment'.

Coding Task

Process Grading

The coding task is designed such that there is a deterministic set of logical steps that needs to be taken to arrive at the correct answer. The correct answer could be identified by executing 10 steps with the provided dataset. Each step was assigned a numerical score based on how difficult that step was to execute. This rubric was validated by a Lead Data Scientist (4+ years of experience at BCG, likely more in the industry). The following is an overarching view of the rubric steps (there are 11 steps in the following rubric, but Step 4 was discarded in post-processing due to it not having a material impact on correctness, i.e. you could solve the problem correctly without executing that step).

RUBRIC STEP 1: Load the products and orders data frames (1 point).
RUBRIC STEP 2: Breaking up the order_info column into two columns with string manipulation - 4 POINTS
RUBRIC STEP 3: Converting the order_date column to datetime WITHOUT coercing errors, with allowance for mixed formatting - 2 points
RUBRIC STEP 4: Checking the values of order_date column using display or value_counts somehow, and removing null and junk values - 2 points
RUBRIC STEP 5: Deleting duplicates in order_id or a set of customer_id and order_id - 3 points
RUBRIC STEP 6: Impute NULL values in products dataframe in customer_id column using the references from the orders dataframe where the information is available - 5 points
RUBRIC STEP 7: Replace values of price for each product in the dataframe with the correct price. This involves studying all the prices and replacing all null and junk ones with the correct price - 10 points
RUBRIC STEP 8: Merge the data frames correctly - 2 points
RUBRIC STEP 9: Filter the data correctly for the correct date range - 1 point
RUBRIC STEP 10: Get the total cost for each order by multiplying the prices and quantities of each product in the order correctly - 2 points
RUBRIC STEP 11: Sort the data by total order cost to get the top 5 values (it does not matter if they filter for the top 5 as long as they sort) - 1 point

The code was graded with the help of LLMs.

LLM Grading for Coding

For the coding task, the aforementioned LLM grading framework was deployed in the following way:

- Preprocessing
    - The coding was task was performed by participants on Google Colab, in iPython notebooks. The iPython notebooks were converted to python files, with markdown blocks being converted into quoted comments. The text from the derived python files was the input text for the LLMs to grade.
- Randomized batching
    - Batching was not performed for the coding task. Since the size of the input text was quite large, this was done to avoid issues with the model's context window that might arise if multiple users' answers were graded at once. Each answer was still graded independently no less than 5 times. The prompting strategies were adjusted appropriately to ensure maximal consistency and accuracy.
- Prompting
    - After extensive testing and validation, it was determined that the prompting strategy that should be used is asking, step by step with the past answers in memory, whether each step had been taken, and to identify which lines of code define it. Along with this, a rubric was provided with various ways of performing those steps, with examples in code of those ways. With this information, the model was asked to identify how the step was executed if executed, and to assign points based on how well as defined by the rubric. This means that at first, the model was asked whether step 1 of the rubric was implemented correctly (with examples of correct implementation), to identify the lines of code in the input where this implementation took place, and to assign a score based on where the execution fell in the rubric. Then, with the question and answer about step 1 added to the query context and therefore model memory of the conversation, the model was asked the same question about step 2. Having the information about former steps in memory allowed the model to avoid confusing and conflating steps, especially when they had logical dependency on one another. This method was evaluated on the ground truth set to show 100% accuracy and 0% variability across 5 runs. Here are the prompts for the model up to three steps of execution.

System Prompt (Persona setting)

```
You will be given code used to solve a problem and an
accompanying rubric for each step expected to solve the problem.
Assign points for the completion of each step based on the
```

rubric. The code does not have to follow the rubric exactly, but it should follow the same logic.

For each step you are asked about, return your answer in this format:
SCORE: 1
REASONING: reasoning without code
LINES EXECUTING THIS STEP: code in the original that is executing that step

User Prompt 1

```
RUBRIC STEP 1: Load the products and orders data frames (1 point).
Example:
orders = pd.read_csv('orders.csv')
products = pd.read_csv('products.csv')

HOW MANY POINTS WOULD YOU ASSIGN FOR THIS STEP? WHY?
```

User Prompt 2

```
RUBRIC STEP 2: Breaking up the order_info column into two columns with string manipulation - 4 POINTS
EXAMPLE:
orders['order_id'] = orders['order_info'].apply(lambda x: x.split(';')[0].strip())
orders['order_date'] = orders['order_info'].apply(lambda x: x.split(';')[-1].strip())

HOW MANY POINTS WOULD YOU ASSIGN FOR THIS STEP? WHY?
```

User Prompt 3

```
RUBRIC STEP 3: Converting the order_date column to datetime WITHOUT coercing errors, with allowance for mixed formatting - 2 points
EXAMPLE:
orders['order_date'] = pd.to_datetime(orders['order_date'], format='mixed')
OR
Converting the order_date column to datetime WITH coercing errors - 1 point
```

221

```
  EXAMPLE:
  orders['order_date'] = pd.to_datetime(orders['order_date'],
errors='coerce')

  HOW MANY POINTS WOULD YOU ASSIGN FOR THIS STEP? WHY?
```

    - Alternative prompting strategies involved asking whether a certain step had been conducted by the participant as a yes/no question, and then asking whether the step had been conducted with a specific coding strategy or with certain parameters as a yes/no question. The yes/no strategy proved unsuccessful due to the model often conflating and confusing steps, and assuming overlaps between logically interconnected steps. Similarly, asking independently about steps without the other steps in context or memory caused confusion, which caused the decision to be made not to ask parallel questions about each step but to ask consecutive questions about each step. Another strategy employed was providing all steps at once and asking for the output in a certain format to be able to parse results for each step. This performed better than other alternative methods but still showed inconsistency and inaccuracy, likely due to being too complicated and asking for too much complicated logic work from the LLM from one query.
- Validation
    - For the coding task, ground truth label sets of over 25% of the total corpus of answers was graded by a Data Scientist. Each answer was validated to have maximum accuracy for this 25% with the assumption that the accuracy extended to the rest of the data.
- Rubric semantic adjustment
    - The LLM was initially asked to abstain from answering questions not explicitly mentioned in the rubric. The abstentions were studied in detail, and where an answer or answer type was missing from the rubric, it was added in along with an assignment of points decided by a Data Scientist. This ultimately ensured that all answers were scored according to a rubric as intended by the rubric designer, rather than arbitrarily through LLM ‘judgment’.

Problem Solving

The problem-solving task provides a dataset consisting of information about soccer matches in history – the date of the match, the home team name, the away team name, the tournament type, whether the game took place at a neutral location (neither home nor away grounds), the number of goals scored by the home team, and the number of goals scored by the away team. Participants were asked to quantify the predictability of each match in the dataset. Finally, as their last step, they were asked to identify which match, based on their predictability metric, was the most surprising match in the dataset. The deliverables of the problem-solving task include the

participants' textual description of the methodology they employed to solve the problem, and their results for the quantified predictability of each match in the dataset in a tabular format.

Methodology Score

Rubric Development

The methodology score for the problem-solving task is based solely on the participants' description of the methods they each applied to perform the task. The Data Scientists' descriptions were studied by a human reviewer (Data Scientist) to understand the major themes of methods employed and the key decision points differentiating methods among participants. Broad categories were formed based on these to represent the rubric, and all participants' data was reviewed manually (by a Data Scientist) and programmatically (with LLM support) to confirm that all methodologies fell within the categorical spectrum defined by the rubric.

A key limitation of this methodology is that we are limited by the description provided by the participant. For example, if a participant took several steps to validate their analysis but failed to mention this in their description of their methodology, this will be flagged as a participant who did not validate their analysis, and they would not get credit for the same. After careful consideration and several passes through the data, the categories were finalized to be generalizable to the maximum amount of the data while limiting bias on how detailed the descriptions were. This was based on and validated by subjective judgment of a few Lead Data Scientists (or Data Scientists with 4+ years of experience at BCG, likely longer in the industry) and BCG consultants supporting the study, who studied iterations of the rubric along with the range of detail in participant answers.

The categories defined by the rubric to score methodology are as follows:

1. Predictive methodology

In order to solve the problem posed by the task, each participant has to define a quantitative, classifying, or gradable outcome related to each soccer match in the dataset, and perform a prediction of that outcome for each match. The predictive methodology category refers to the mechanism employed by the participant to arrive at that prediction. This is the very basis of the methodological evaluation and other categories are best understood in relation to the predictive methodology. You will recall that these methodology categories were created based on approaches of the Data Scientist participants, and validated extensively (with manual human (Data Scientist) review and LLM validation) to be mutually exclusive and collectively exhaustive. Within this category are the following methodological subcategories

- Machine Learning
    - Regression

        - Linear regression on difference of goals scored by each team: The participants define the goal difference between home and away teams (e.g. negative goals are permitted based on the losing team in the definition) in each match as the outcome variable to predict and perform linear regression on a combination of feature variables to predict this.
        - Ensemble regression on goal difference: The participants define the goal difference between home and away teams (e.g. negative goals are permitted based on the losing team in the definition) in each match as the outcome variable to predict and perform regression using ensemble methods such as random forest on a combination of feature variables to predict this.
        - Linear regression on numerical definitions of classes: The participants define a categorical outcome in each match as the outcome variable and define custom classes as a specific numeric (for example, +1 for home team victory, 0 for draw, -1 for home team loss). Then, the participants perform linear regression and transform the result to predict the custom class outcome (based on whether it is closest to 1, 0, or -1, in the example described). This method was scored lower than the above two because regression methods are incorrect for classification, due to assuming continuous and unbounded outputs. Regression also optimizes over error metrics that are incorrect for classification problems (for example, mean squared error), and leads to decision boundary creation with faulty statistical logic.
    - Classification: Perform classification to predict match outcomes as victory or loss, using one of the following methods
        - Logistic regression: The participants define a binary outcome (for example, 1 if the home team wins, and 0 if the home team loses) in each match as the outcome variable and perform logistic regression on a combination of feature variables to predict this.
        - Random forest: The participants define a binary outcome (for example, 1 if the home team wins, and 0 if the home team loses) in each match as the outcome variable and perform random forest classification on a combination of feature variables to predict this.
        - Gradient boosting: The participants define a binary outcome (for example, 1 if the home team wins, and 0 if the home team loses) in each match as the outcome variable and perform gradient boosted classification on a combination of feature variables to predict this.
- Probabilistic Analysis
    - Fitting historical outcomes to a Poisson distribution: Participants fit historical victory or goal statistics per team to a Poisson distribution and use those calculations to calculate the likelihood of the observed match outcomes (for

example, the likelihood of the exact number of goals scored by each team), or used those calculations to calculate the most probable outcome.
    - Using Elo ratings: Participants use established methodologies in sports or competition data science such as the calculation of Elo ratings, which measure the relative strengths of teams based on past performances. These ratings were then incorporated into a probabilistic model to estimate the likelihood of different match outcomes.
- Summary statistics: Participants developed custom summary statistics by analyzing combinations, slices, and subsets of historical data. These statistics may include team-level metrics (for example, average goals scored by each team, win rates, or head-to-head performance) and/or match-level metrics (for example, average rate of home team wins or average goal difference on the match level). These statistics were then used as inputs for a probabilistic calculation to estimate the likelihood of match outcomes.

2. Definition of predictability

The problem statement for the problem-solving task explicitly asks the participants to provide a 'predictability score' for each match in the dataset. Depending on the predictive methodology selected by the participant, the definition of predictability definition used makes logical sense or does not. The following are the categories for predictability definition

- Outcome difference: Participants defined the predictability of a match result as the difference between their prediction of the outcome and the actual outcome. This can work for a number of definitions of the outcome variable itself, for example, the difference between the predicted goal difference between the home and away teams and the actual goal difference, or the binary difference between whether a home team win was predicted and whether one occurred, or the difference in the probability of a predicted home team win as extracted from the model and the observed result of whether the home team won. This last example is to be distinguished from the following method of defining predictability.
- Probability: Participants do not solidly define predictability beyond the raw output of their predictive models. For example, when using classification, participants return the probability of a home team victory as the predictability of the match result (a 0% forecasted chance of a home team victory signals a highly predictable loss if the home team did lose, but when this raw number is returned as predictability, this is logically and contextually unsound and signals a lack of understanding of the problem or of the predictive model functionality). For another example, when using summary statistics, participants return the raw probability of victory for a certain team based on historical probability of victory. This does not automatically have any meaning when it comes to the predictability of a match when not compared to any outcome.
- Processed probability: Participants define predictability of a match result as a function of the predictive power or predictive results of their model for that match. They

meaningfully process the probability outcomes or outputs of their models to arrive at a well-defined quantitative notion of predictability.

- Team difference: Participants define their own predictability metric relative to the data or model they use relating to the difference between predicted outcomes for each team, where there is relatively subjective logic. For example, those that choose to regress on goal difference between teams can define predictability such that a larger predicted goal difference implies a more predictable match. Similarly, those that choose to create metrics for strengths of teams could define the predictability of the match as the difference between team strengths (with a higher difference implying a more predictable result if the stronger team did win, and an unpredictable result if not). This method does not make sense for classifiers due to the fact that you would be predicting the outcome of one team, and if you predicted the outcomes of both, they are mutually exclusive and interdependent outcomes.
- Z-score: Participants that use historical data to arrive at likelihoods of match outcomes define predictability of the match as the likelihood of the outcome occurring, and use z-scores to quantify that likelihood (for example, with the Poisson method).

3. Categorical feature management
    - Match-based analysis: Participants use features as they appear on a match basis as part of their predictions (such as home and away advantage, neutral field or match type)
    - Team-based analysis: Participants create features representing team strengths or statistics based on historical data and incorporate those into predictions for each match
    - Both: Participants create features representing team strengths or statistics based on historical data and incorporate those into predictions for each match, and also use match-based features (such as home and away advantage, neutral field or match type)
4. Temporal feature management (used date)
    - Yes / No: Participants did or did not make use of the temporal features. For example, participants did or did not extract years, months, or decades from the temporal data to use as predictive features, or participants did or did not use moving or rolling windows of dates in their summarized statistical analyses.
5. Method of validation (if any)
    - Accuracy validation: Participants conducted simple metric measurement of accuracy to assess the performance of their methodology and made changes accordingly, to improve performance
    - Cross validation: Participants conducted cross-validated metric measurement on rotating training and testing sets to assess the performance of their methodology and made changes accordingly, to improve performance.

    - Model selection: Participants attempted various different methods or models as their predictive methodology and evaluated each method using one or more metrics to select the highest performing methodology.
6. Completion of final step (return of what the participant identifies at the most surprising match):
    - Yes / No: Participants did or did not reach the conclusion of their predictive and predictability analysis to return a final answer, whatever that may be based on their individual methodology.

The final scoring rubric, determined on the basis of the above categories, is as follows.

First, a basic score is arrived at based on a combination of predictive methodology and predictability definition. The reasoning behind the scoring is evident in the method descriptions above.

Classification Scoring

For those that employed classification, the following points were awarded according to their classification methodology and predictability definition

- Outcome difference for predictability definition
    - Logistic regression employed: 7 points
    - Random forest employed: 8 points
    - XGBoost employed: 8 points
    - LightGBM employed: 9 points

These methodologies are ranked in ascending order of robustness and success when applied to the problem at hand based on Data Scientist testing (this ranking of the models is specific to this use case and cannot be generalized to all problem types)

- Processed probability for predictability definition
    - Logistic regression employed: 7 points
    - Random forest employed: 8 points
    - XGBoost employed: 8 points
    - LightGBM employed: 9 points
- Team difference for predictability definition:
    - Logistic regression employed: 5 points

This method does not make sense for classification.

- Probability for predictability definition
    - Logistic regression employed: 5 points
    - Random forest employed: 5 points
    - XGBoost employed: 5 points
    - LightGBM employed: 5 points

As described above, the probability for predictability definition is a relatively poor and illogical definition. No additional points are awarded for model robustness when the outcome is illogical, but points are awarded for model execution.

Regression Scoring

For those that employed regression, the following points were awarded according to their regression methodology and predictability definition

- Outcome difference for predictability definition
    - Linear regression on difference of goals scored by each team: 8 points
    - Random forest regression on goal difference: 9 points
    - XGBoost regression on goal difference: 9 points
    - Linear regression on numerical definitions of classes: 6 points

Regression is a generally more informative and robust way to solve this problem due to the inherent ability to incorporate more information (e.g. the difference in goals rather than just categorical feature) into the dependent variable. Therefore, the regression models generally score higher, even when regression is performed on classes, which is a poor way to use regression.

- Processed probability for predictability definition
    - Linear regression on difference of goals scored by each team: 8 points
    - Linear regression on numerical definitions of classes: 6 points
- Probability for predictability definition
    - Linear regression on difference of goals scored by each team: 5 points
    - Ensemble regression on goal difference: 5.5 points
    - Linear regression on numerical definitions of classes: 4 points

Summary Statistics Scoring

- Outcome difference for predictability definition: 6 points
- Processed probability for predictability definition: 6 points
- Team difference for predictability definition: 5 points
- Probability for predictability definition: 4 points
- Z-Score scoring: 8 points

Poisson Scoring

- Outcome difference for predictability definition: 8 points
- Probability for predictability definition: 6 points

Elo Scoring

- Outcome difference for predictability definition: 8 points
- Processed probability for predictability definition: 8 points
- Team difference for predictability definition: 7 points
- Probability for predictability definition: 5 points

Categorical Feature Management Scoring

The following points were awarded based on the categorical feature management

- If a team-based analysis was used, an additional 3 points were awarded due to the difficulty of incorporating team-based features and the additional analysis this would entail
- If both team-based and match-based analysis were used, an additional 5 points were awarded due to the difficulty of incorporating both types of features and the additional analysis this would entail

Temporal Feature Management Scoring

If temporal features were used in the analysis, an additional 5 points were rewarded due to the difficulty of incorporating those features and the additional analysis this would entail.

Method of Validation Scoring

The following points were awarded based on the method of validation employed

- If accuracy was measured to validate the model, an additional 4 points were awarded due to the effort to validate the model, which is essential in predictive modeling
- If cross validation was performed or extensive model selection, an additional 7 points were awarded due to the effort to validate, and the robustness of the methodology chosen to do so

Completion of Final Step Scoring

If the final step was completed, an additional 3 points were rewarded to signal end-to-end completion of the exercise.

LLM Grading for Problem-Solving

For the problem-solving task, the LLM grading did follow the aforementioned architecture, but was slightly different because each and every LLM output was manually reviewed by a Data Scientist and corrected as needed due to the complexity of the problem at hand. Therefore, the adoption of the LLM grading architecture for problem solving is described as follows:

- Preprocessing
    - Each methodology was summarized along different dimensions by the LLM through the asking of specific questions related to the categories identified above. This process was akin to a ‘field extraction’ where the methodologies, initially all formatted differently and largely unstructured, were organized using the LLM into a more structured format, easier to parse and understand for grading purposes, making the downstream grading by the LLM more accurate. Each

summary was then manually reviewed and validated. The prompts used to summarize are as follows.

- Randomized batching
    - Each summarized category answer in the problem-solving task was batched with four other random answers in the same category and graded in a single query by the LLM. This was done to minimize the bias arising from comparison as described in the overall LLM grading architecture. Each answer was graded at least 5 times in order to arrive at a consensus similar to the statistics task
- Prompting
    - The actual grading was performed by prompting involving two tiers of classification – the first using a yes/no approach, and the next using a simple classification. At first, the LLM was asked as a yes/no prompt whether each kind of predictive methodology was employed (whether classification was employed, regression was employed, Poisson distributions were fitted to, and so on), and if classification or regression were chosen, it was given choices of the sub-methodologies to choose from to classify which was used (for example, if classification was employed, which kind of model from the list below was used to classify?). For the categories without sub-categories, a yes/no answer sufficed.
    - Many other strategies were tested, but the prompting strategies and results informed the rubric as well, so for the above rubric, this was the only tested method that was relevant.
- Validation
    - 100% of the methodology data was manually graded and reviewed by a Data Scientist. Every LLM output was validated for accuracy, and with under 5% inaccuracy, the answers were manually changed to the correct answers.
- Rubric semantic adjustment
    - The rubric was informed by how much detail of information could be accurately parsed out by an LLM, and how much information was present consistently across methodologies. Therefore, the rubric was constantly adjusted in conjunction with changing prompts till the above described rubric was arrived at.

Correctness

We define correctness on the problem-solving task as closeness to the answers that the Data Scientists arrive at for the same exercise within the given time. Each Data Scientists' predictability result is standardized and then regarded as a baseline. We collect two measurements of correctness for each participant.

1. The standardized predictability result of each participant in the control or treatment group is then compared to each Data Scientist baseline in the data. The mean absolute error of the participants' result is calculated in comparison to each baseline. For each participant,

the lowest recorded mean absolute error against each baseline is recorded as one of their correctness scores.

2. For each Data Scientist, and for each participant, we have a breakdown of their predictive methodology (needed for the predictive methodology score described above). For the second score, we calculate each participants' predictability result only in comparison with the Data Scientists baselines for those Data Scientist who employed comparative predictive methodology to them. For each participant, the lowest recorded mean absolute error against the applicable baselines is recorded as the second correctness score.

Three Data Scientists were excluded from the baselines. Of these three, one returned their match dates as years, having discarded other data information. This made their dataset impossible to merge with others. Another Data Scientist used ChatGPT for assistance and was therefore excluded from analysis and baselines. The last one that was excluded only returned predictability scores for 5 matches, leaving the rest blank. It is impossible to get a good correctness score from 5 entries given the initial size of the dataset (40k+ entries).

Supplementary 4.2 Coding Problem Rubric

Upskill Experiment - Coding Test[232]

Assignment

Make sure to write and run the **Python code** needed to solve the question in this **Google Colab** notebook so that we may review your code.

Overview of Datasets

Dataset 1: Orders Data (`orders.csv`)

- **File Type**: CSV
- **Contents**:
  - `customer_id`: The unique identifier for each customer.
  - `order_info`: Information about each order in the format `order number ; date and time`. The order number in this dataset (once decoupled from the date and time) corresponds with that in the next.

Dataset 2: Products Data (`products.csv`)

- **File Type**: CSV
- **Contents**:
  - `customer_id`: The unique identifier for each customer associated with an order.
  - `order_id`: ID of each order in the format `order number`. The order number in this dataset corresponds with that in the previous once decoupled from the date and time.
  - `order_products`: Details about the products in each order in the following format: `{product_id: [product_cost, product_quantity], ...}`. Each product is sold either at its original price or a 20% discount.

Commonalities

- Both datasets share the `customer_id` field and `order_id` information with the order number.
- Each order ID is unique, whereas customer IDs may be repeated across multiple orders. However, each combination of order ID and customer ID is unique.
- Whenever you have information about either one of these, it is correct.

Data Quality and Cleaning Guidelines

- **Order and Customer IDs**: Entries are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere.
- **Date and Time Fields**: Entries with incorrect values should be removed.
- **Product Quantities and Product IDs**: Always correct unless marked as NULL, which indicates missing values.
- **Product Costs**: Each product ID is associated with a unique cost. Sometimes, the original unique cost is discounted at 20% so that is what is shown in a certain order. However, when not discounted,

sometimes there are junk or NULL values instead of the correct cost. Junk or NULL values in the product costs should be replaced with the original cost (the discounted cost should be left as is wherever it is shown but not added in elsewhere).

233

Tips for Handling Junk and NULL Values; duplicates

- Examine common values in each column to identify patterns and potential corrections.
- Attempt to fix junk or NULL values using information elsewhere in the data before considering row deletion.
- Date time fields can have incorrect fields that are not correctable, discard the affected rows and values to maintain data integrity.
- Check for duplicates at every stage

QUESTION:

Use the above datasets to answer the question: **Which 5 customer IDs had the highest average order by cost in May 2022?**

SOLUTION:

from google.colab import drive drive.mount('/content/drive')

Step 1: Start with loading and viewing the data

Imports and Reading in the File

1 point

In [1]:
```
import pandas as pd
```

In [2]:
```
orders = pd.read_csv('orders.csv')
```

In [3]:
```
products = pd.read_csv('products.csv')
```

Studying the Files

[OPTIONAL]

In [4]:
```
orders.head()
```

Out[4]:

	customer_id	order_info
0	107654909	15659 ; 2023-08-04 01:33:45.202520256
1	251079598	14859 ; 2023-11-23 14:53:10.639063904
2	344439380	17194 ; 2023-03-23 04:04:23.546354640
3	752668623	15776 ; 2020-10-09 19:55:36.453645364
4	705264936	17338 ; 2021-10-18 08:03:18.739873984

In [5]:
```
products.head()
```

Out[5]:

	index	customer_id	order_id	order_products

0	0	913891745.0	13702	{'129': [360, 178], '986': [391, 34], '317': [...
1	1	634096553.0	10795	{'317': [81, 175], '910': [0, 102], '129': [28...
2	2	189473854.0	12927	{'910': [313, 103]}
3	3	774788031.0	13557	{'129': [NaT, 88], '722': [316, 17], '910': [3...
4	4	850303382.0	14520	{'129': [0, 125], '910': [313, 57], '986': [10...

Step 2: Split and clean the orders data

Remember that entries for customer IDs and order IDs are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere. Date and time fields entries with incorrect or junk values should be removed. And duplicates should be dropped.

Break up the order info

4 points

In [6]:
```
orders['order_id'] = orders['order_info'].apply(lambda x: x.split(';')[0].strip())
```

In [7]:
```
orders['order_date'] = orders['order_info'].apply(lambda x: x.split(';')[-1].strip())
```

Order date to datetime

2 points

In [8]:
```
orders['order_date'] = pd.to_datetime(orders['order_date'], format='mixed')
```

Check order date for junk values

2 points for deleting and identifying junk values - THEY WILL NOT MAKE A DIFFERENCE TO THE CORRECTNESS

In [9]:
```
orders.order_date.value_counts()
```

Out[9]:
```
order_date
2099-12-31 23:59:59.000000000    558
2001-01-01 00:00:00.000000000    228
2022-05-03 07:25:26.192619264      5
2022-12-30 09:06:23.438343824      5
2021-05-29 23:50:38.343834384      5
                                ...
2022-07-31 19:35:26.732673264      1
2021-01-28 22:44:14.905490548      1
2021-03-30 09:32:18.793879384      1
2022-01-17 14:16:27.218721872      1
2020-01-06 06:09:23.816381638      1
Name: count, Length: 10002, dtype: int64
```

Delete duplicates of Customer and Order ID together

3 points

In [10]:
```
orders = orders[~orders[['customer_id', 'order_id']].apply(frozenset, axis=1).duplicated
```

Drop order_info column

[OPTIONAL]

235

In [11]:
```
orders = orders[['customer_id', 'order_id', 'order_date']]
```

In [12]:
```
orders.order_id.value_counts()
```

Out[12]:
```
order_id
15659    1
18684    1
11383    1
19951    1
14804    1
        ..
11967    1
17991    1
11791    1
16555    1
13725    1
Name: count, Length: 10000, dtype: int64
```

Step 3: Clean the products data

Remember that entries for customer IDs and order IDs are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere.

Product Quantities and Product IDs: Always correct unless marked as NULL, which indicates missing values.

Product Costs: Each product ID is associated with a unique cost. Sometimes, the original unique cost is discounted at 20% so that is what is shown in a certain order. However, when not discounted, sometimes there are junk or NULL values instead of the correct cost. Junk or NULL values in the product costs should be replaced with the original cost (the discounted cost should be left as is wherever it is shown but not added in elsewhere).

Don't forget to drop duplicates!

Impute NULL values in products for customer_id in products dataframe

5 points

In [13]:
```
products.describe()
```

Out[13]:

	index	customer_id	order_id
count	10000.00000	9.704000e+03	10000.00000
mean	4999.50000	5.465848e+08	15000.50000
std	2886.89568	2.632115e+08	2886.89568
min	0.00000	1.010432e+08	10001.00000
25%	2499.75000	3.076407e+08	12500.75000
50%	4999.50000	5.484692e+08	15000.50000
75%	7499.25000	7.769422e+08	17500.25000
max	9999.00000	9.987171e+08	20000.00000

In [14]:
```
customers_dict = dict(zip(orders.order_id, orders.customer_id))
products['customer_id'] = products['order_id'].apply(lambda x: customers_dict[str(x)])
```

236

Delete junk values from order date in orders

[OPTIONAL]

In [15]:
```
orders.order_date.value_counts()
```

Out[15]:
```
order_date
2099-12-31 23:59:59.000000000    199
2001-01-01 00:00:00.000000000     74
2023-08-04 01:33:45.202520256      1
2021-03-20 07:44:18.145814584      1
2022-03-21 02:08:10.369036904      1
                                ...
2022-03-24 17:44:41.908190816      1
2023-12-15 12:32:19.873987392      1
2020-05-08 11:49:42.178217822      1
2023-07-01 04:33:11.719171920      1
2020-01-06 06:09:23.816381638      1
Name: count, Length: 9665, dtype: int64
```

In [16]:
```
orders = orders[~(orders.order_date.isin([pd.to_datetime('2099-12-31 23:59:59.000000000'
```

In [17]:
```
orders.head()
```

Out[17]:

	customer_id	order_id	order_date
0	107654909	15659	2023-08-04 01:33:45.202520256
1	251079598	14859	2023-11-23 14:53:10.639063904
2	344439380	17194	2023-03-23 04:04:23.546354640
3	752668623	15776	2020-10-09 19:55:36.453645364
4	705264936	17338	2021-10-18 08:03:18.739873984

Read Products DF

[OPTIONAL]

In [18]:
```
products.head()
```

Out[18]:

	index	customer_id	order_id	order_products
0	0	913891745	13702	{'129': [360, 178], '986': [391, 34], '317': [...
1	1	634096553	10795	{'317': [81, 175], '910': [0, 102], '129': [28...
2	2	189473854	12927	{'910': [313, 103]}
3	3	774788031	13557	{'129': [NaT, 88], '722': [316, 17], '910': [3...
4	4	850303382	14520	{'129': [0, 125], '910': [313, 57], '986': [10...

Replace all prices with correct price

10 points

In [19]:
```
import json
```

In [20]:
```
products['order_products'] = products['order_products'].apply(lambda x: json.loads(x.rep
```

237

In [21]:
```
from collections import defaultdict
from collections import Counter
```

In [22]:
```
def product_search(x):
    for key in list(x.keys()):
        product_dict[key].append(x[key][0])
```

In [23]:
```
product_dict = defaultdict(list)
products['order_products'].apply(lambda x: product_search(x))
```

Out[23]:
```
0       None
1       None
2       None
3       None
4       None
        ...
9995    None
9996    None
9997    None
9998    None
9999    None
Name: order_products, Length: 10000, dtype: object
```

In [24]:
```
from statistics import mode
```

In [25]:
```
for key in product_dict.keys():
    print(key)
    print(Counter(product_dict[key]))
```

```
129
Counter({360: 4587, 288.0: 748, -100: 315, 0: 246, 100: 166})
986
Counter({391: 4542, 312.8: 727, -100: 284, 0: 227, 100: 190})
317
Counter({81: 4573, 64.8: 738, -100: 266, 0: 242, 100: 187})
722
Counter({316: 4617, 252.8: 694, -100: 310, 0: 251, 100: 197})
910
Counter({313: 4586, 250.4: 716, -100: 304, 0: 255, 100: 167})
```

In [26]:
```
for key in product_dict.keys():
    product_dict[key] = mode(product_dict[key])
```

In [27]:
```
def correct_cost(x):
    for key in x.keys():
        if x[key][0] == 100:
            x[key][0] = product_dict[key]
        elif x[key][0] == -100:
            x[key][0] = product_dict[key]
        elif x[key][0] == 0:
            x[key][0] = product_dict[key]
    return x
```

In [28]:
```
products['order_products'] = products['order_products'].apply(lambda x: correct_cost(x))
```

Step 4: Merge the data sets to answer the question: Which 5 customer IDs had the highest average order by cost in May 2022?

Merge the data and get final answer

5 points

238

In [29]:
```
products.order_id = products.order_id.apply(str)
```

In [30]:
```
final = orders.merge(products, right_on=['customer_id', 'order_id'], left_on=['customer_
```

In [31]:
```
final['order_month'] = final['order_date'].apply(lambda x: str(x.year) + ' ' + str(x.mon
```

In [32]:
```
final['total_order_cost'] = final['order_products'].apply(lambda x: sum([y[0]*y[1] for y
```

In [33]:
```
may_2022 = final[final['order_month'] == '2022 5']
```

In [34]:
```
pd.DataFrame(may_2022.groupby('customer_id')['total_order_cost'].mean()).sort_values('to
```

Out[34]:

customer_id	total_order_cost
585775494	224539.0
613911991	204480.0
349369215	202248.0
272552610	199774.0
723415497	197863.0

Supplementary 4.3 Statistics Problem Rubric

240

INSTRUCTIONS

- *Do not Google image search or send any images to GPT. Refrain from copying and pasting the exact question into Google or GPT unless completely stuck. Do not spend more than 1.5 hours on this task.*

Question 1: The following is the first five rows of data containing financial and demographic information about domestic partners who have co-purchased a home in the last several years. Please note that the following table is illustrative and represents a snapshot sample of the data to solve this problem. All the information you need to solve the problem is contained within this snapshot.

Age 1	Age 2	Income 1	Income 2	Borough	ZIP Code	Date	Price	Mortgage
39	37	270000	180000	Manhattan	10076	1 January 2016	1,125,000	Yes
NULL	38	445000	670000	Manhattan	10025	1 January 2016	2,249,000	Yes
27	29	145000	225000	Queens	11106	2 January 2016	900,000	Yes
33	NULL	90000	76000	Brooklyn	11203	2 January 2016	415,000	Yes
68	55	78000	450000	Bronx	10474	2 January 2016	3,399,000	No

1. You have been tasked with predicting based on demographics and price whether a mortgage was taken out to by the house. You prompt ChatGPT for detailed instructions on how to do this, and ChatGPT recommend using a logistic regression model. It recommends the following steps.

1. Understand Your Dataset

- **Explore and Preprocess**: Start by exploring your dataset to understand the features available and their types (numerical, categorical). Clean the data by handling outliers and possibly irrelevant features. Preprocessing steps like encoding techniques (e.g., one-hot encoding) might be necessary for categorical data. Ensure that your dataset does not have missing values. You can either fill them in with a strategy (mean, median, mode) or remove the rows/columns with missing values, depending on the situation.

a. Which of the following are among the steps you could take to address this point? Select all that apply.

i. Plot the distribution of each of the numerical variables and remove rows with outliers from this dataset **+0.5 points**

ii. One-hot encode the 'Borough' variable **+0.5 points**

iii. Investigate relationships between variables **+0.5 points**

iv. Convert date to a numerical variable **0 points**

241

v. One-hot encode the ZIP code variable **-1 point only if vii. not selected**
vi. One-hot encode the age variables **-0.5 points**
vii. Bin the ZIP codes by neighborhoods and do not process further **-1 point**
viii. Bin the ZIP codes by neighborhoods and one-hot encode **+1 point only if v. or vii. not selected**
ix. Check columns with null values and remove those with >80% missing values **+0.5 points**
x. Impute NULL values by using a summary statistic or by developing a simple model that predicts those values based on other features **+0.5 points**

-1 if more than one about ZIP code selected
Maximum: 3.5 points
Minimum: 0 points

Explanation of point assignment: There is only one reasonable handling of the ZIP code variable. There are several issues with one-hot encoding the raw ZIP code variable, including the curse of dimensionality and sparse resultant data. No reasonable data scientist would make that choice and therefore it is wrong. You also cannot bin without one-hot encoding because binned data is categorical. Choosing either one is a subtraction of a whole point but not an immediate 0. Given that this choice is more difficult there is more positive credit for getting this right than for getting other correct answers.
Date to numeric is contentious, therefore 0 penalty or reward (date is ordinal but sometimes represented as a number, although never treated like a numeric in the sense that you would never take a summary statistic such as mean or median of the date column (e.g. if you represent months or years numerically you would never take a mean of those), you would only take summary statistics such as mode because it is essentially categorical).
No one would ever one-hot encode age unless binned.

2. Split the Data

- **Train-Test Split**: Divide your dataset into a training set and a testing set (commonly a 70-30 or 80-20 split) to evaluate the model's performance on unseen data.

3. Train the Model

- **Training**: Use the training dataset to train your model, adjusting parameters as needed. For complex models, consider using cross-validation to fine-tune hyperparameters and prevent overfitting.

4. Evaluate the Model

- **Performance Metrics**: Evaluate your model on the test set using appropriate metrics such as accuracy, precision, recall, F1 score, and the ROC-AUC curve. These metrics will help you understand how well your model is performing in terms of both its ability to predict mortgages correctly and its robustness against false positives or negatives.

b. What issue necessitates using all these metrics? Which of the above steps is affected by this issue and how? (Answer in 100 words or less – bullet points ok)
(3 points – imbalanced data is the issue
OR
1 point for overfitting being recognized as the issue without reference to imbalance)
AND
2 points – affects step 2 (train-test split), as stratified sampling would be a fix
(full credit for mention of stratified sampling even if step 2 not mentioned)
Total 5 points

c. Would you change the order of any of the above steps? Why or why not? (Answer in 100 words or less – bullet points ok)
2 points for identifying that steps 1 and 2 should be switched.
3 points for identifying that the issue is data leakage.
Total 5 points

2. You want to try a k-Nearest Neighbors model. Which of the following are not required (although recommended) for logistic regression, but absolutely necessary for k-Nearest Neighbors? Select all that apply.

a. Transform numerical variables (e.g. log) +2 points
b. Make sure there are only two classes to predict -2 points
c. Convert Mortgage column from string to binary -1 points
d. Standardize numerical variables +2 points
e. Impute the missing age with the other age in the same row -1 points
f. One-hot encode the appropriate variables -1 points

Maximum 4, minimum 0

Explanation of point assignment: The key here is what is absolutely necessary for kNN but not required for logistic regression, although recommended. Option (b) is not required at all for kNN so it has the biggest point subtraction. Option (c) , (e), and (f) would need to be done for either model so they are not correct but do not warrant as much subtraction because they are not as egregious. The correct answers are (a) and (d) which are the numerical transformations, which kNN is particularly sensitive to.

3. You also try a decision tree model for the same classification problem, to compare performance. You realize your model is performing quite poorly on both training and validation sets. You double-check the code and there are no bugs. What could be causing this problem? Select all that apply.

a. Your model is underfit
b. Your model is overfit
c. The learning rate hyperparameter is too small

d. The learning rate hyperparameter is too large
e. The decision tree is too shallow
f. The decision tree is too deep
g. None of the above

h. Award the following points for the selection of the options, with a minimum of 0 and a maximum of 4:

i. a = +2 points
j. b = -4 points
k. c = -2 points
l. d = -2 points
m. e = +2 points
n. f = -4 points
o. g = 0 points

Explanation of point assignment: Wrong answers are egregiously wrong (opposite of what is correct) and each result in an immediate 0. Two middle answers subtract only half of total points because they are not egregiously wrong but indicate a fundamental misunderstanding of decision tree model (which is one of the most simple models to understand). Therefore, if you pick the 2 correct answers and one of the hyperparameter answers, you get half of the total credit. Selecting one of the hyperparameter answers indicates that you might be guessing / selecting one of each and hoping for the best.

4. Next, you have been instructed to predict the price based on the other variables, and this time you have been instructed to use linear regression. Following instructions from ChatGPT, you perform a basic linear regression. You notice that your R^2 value is too low. You prompt ChatGPT for suggestions on how to diagnose the problem, and it is recommended that you check the residual plots. You notice that the residual plot does not appear random. What could this mean? Select all that apply.

a. The observed values of your dependent variable are independent from each other -2 points
b. Your model is missing an important variable +1 points
c. There is some interaction between your variables +1 points
d. A higher order term might be required in your regression +1 points
e. Variance of the residual is the same for any value of X -2 points

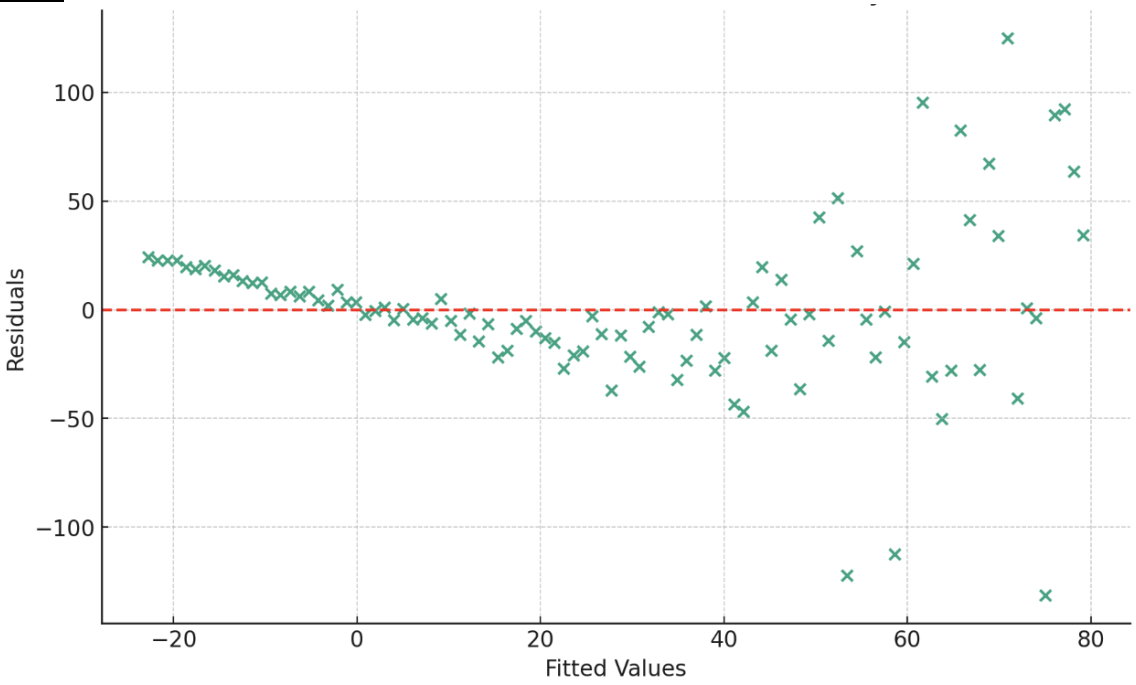
Explanation of point assignment: Wrong answers are unequivocally wrong as to things that would cause a low R^2 value (these are assumptions needed in order for regression to work, they would only be correct in the opposite version of those statements), they each subtract 2 but but are not an immediate 0 unless both incorrect answers picked or 3 correct answers not picked

Max 3 min 0

5. For the following residual plots, what could be the characteristics of or issues with the data or model that are corresponding with these results (choose from the list provided for each image)? It is possible that more than one characteristic or issue applies to any given image, and it is possible that a characteristic or issue may apply to more than one image.

Explanation of point assignment: Offer no credit or no penalty for understandable mistakes, offer penalty for egregious mistakes, and offer reward for correctness, weighted by difficulty of getting the correct answers

Image A



a. No characteristic or issue is apparent **-2 points**

b. Heteroscedastic data **1 points**

c. Outliers **0 points**

d. Response variable requires transformation **-1 points**

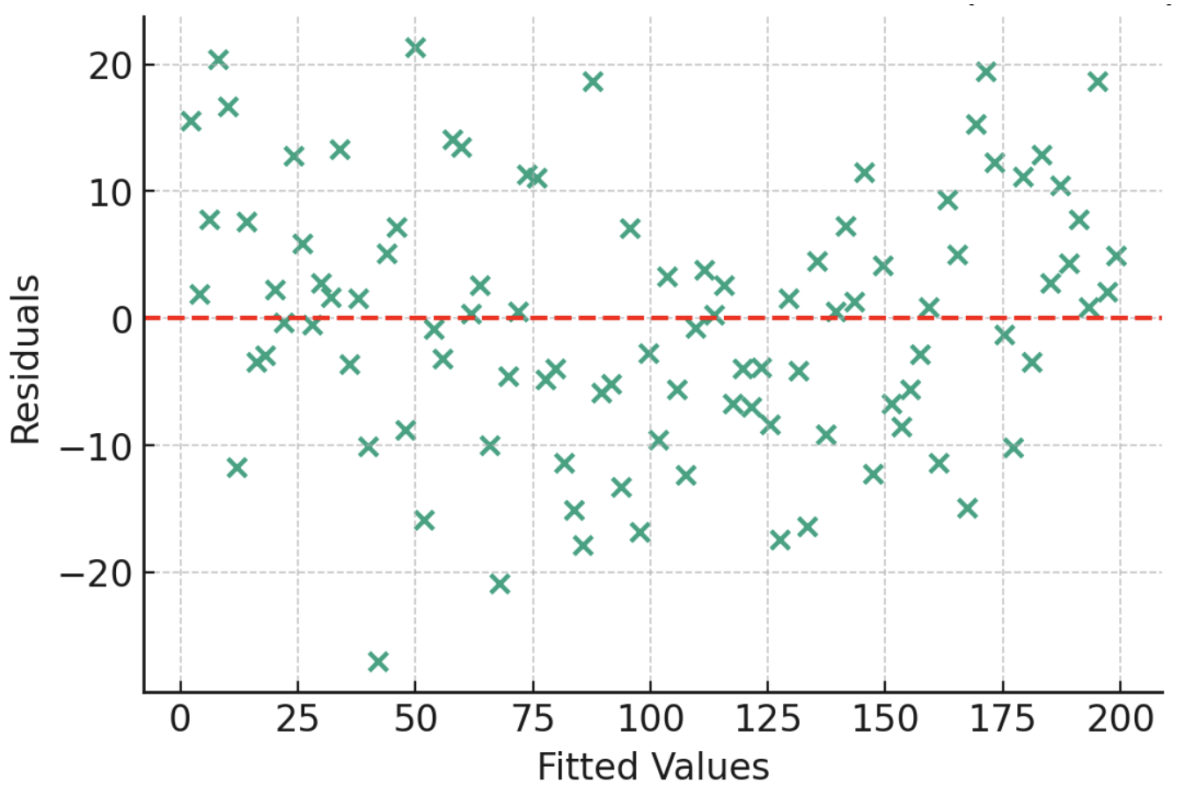
e. A higher order variable might be required **2 points**

Max 3 min 0

Explanation of point assignment: Clearly there is a characteristic/issue apparent, outliers might be confusing, there's no apparent transformation for the response variable but there is evidence of heteroscedasticity (obvious) and non-linearity and higher order variables (not as obvious but still obvious)

Image B

245



a. No characteristic or issue is apparent **2 points**

c. Heteroscedastic data **-2 points**

d. Outliers **0 points**

e. Response variable requires transformation **-2 points**

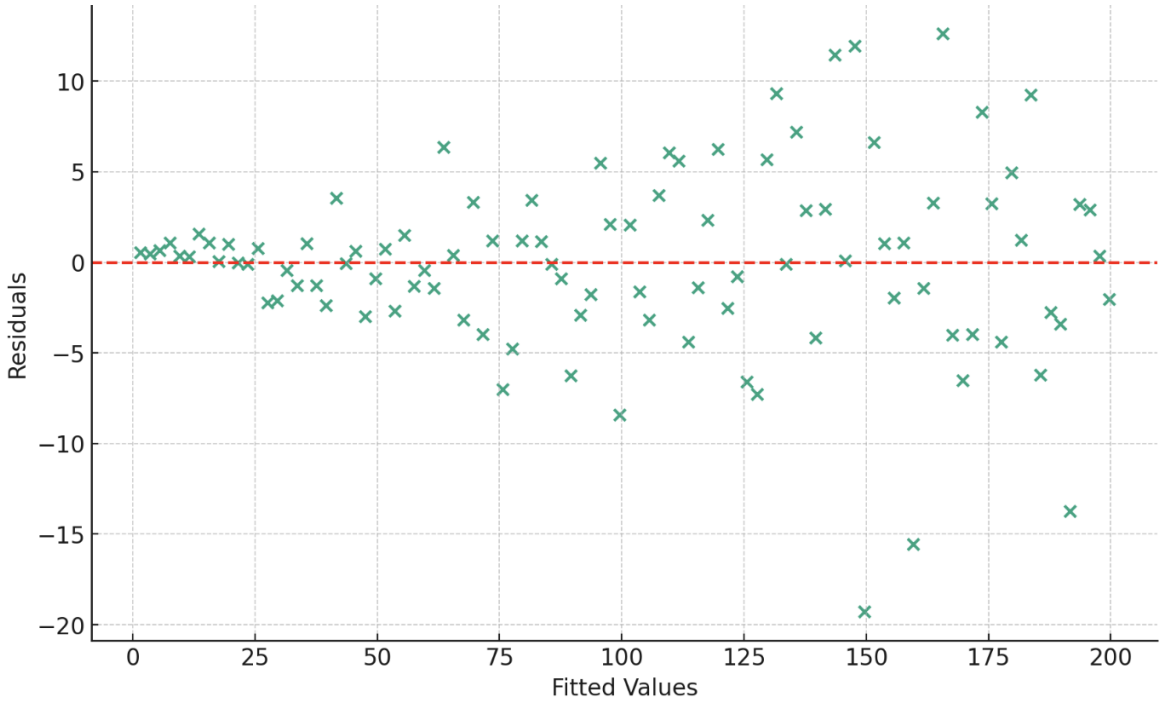
f. A higher order variable might be required **-2 points**

Max 2 min 0

Explanation of point assignment: Clearly there is a characteristic/issue apparent, this should be identifiable as a normal residual plot, no penalty for outliers because inevitable confusion

Image C

246



a. No characteristic or issue is apparent **-2 points**

d. Heteroscedastic data **2 points**

e. Outliers **0 points**

f. Response variable requires transformation **-2 points**

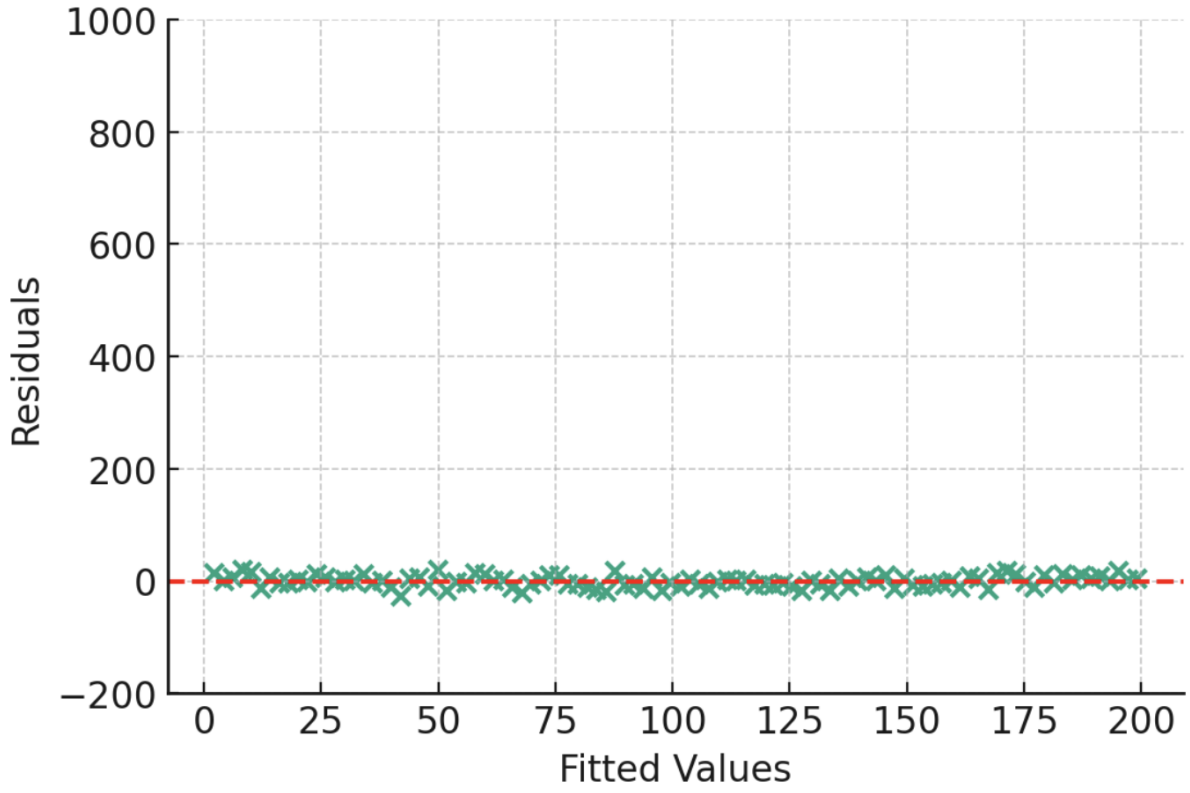
g. A higher order variable might be required **-2 points**

Max 2 min 0

Explanation of point assignment: Clearly there is a characteristic/issue apparent, outliers might be confusing, there's no apparent transformation for the response variable, and there is a clear linear center making general linearity apparent

Image D

247



a. No characteristic or issue is apparent **0 points**

e. Heteroscedastic data **-2 points**

f. Outliers **-2 points**

g. Response variable requires transformation **2 points**

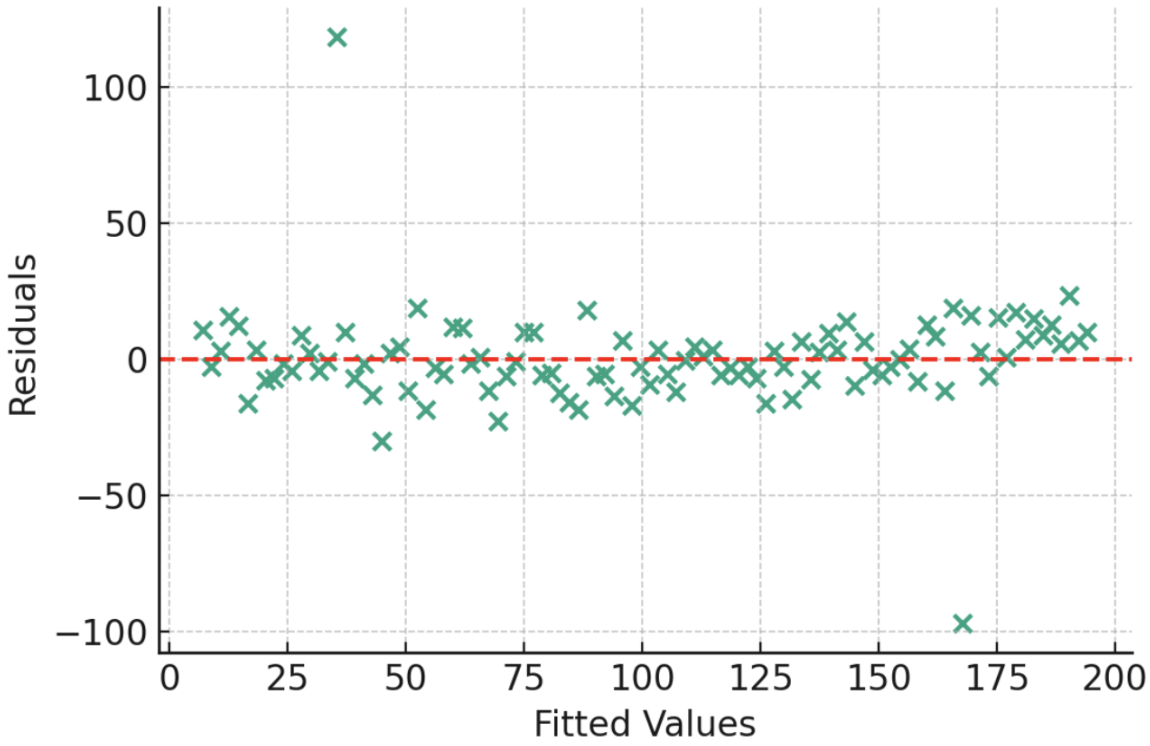
h. A higher order variable might be required **-2 points**

Max 2 min 0

Explanation of point assignment: Some might think that this is an example of a normal residual plot zoomed out, since this is confusing there is no penalty. There is no visible heteroscedasticity or outlier, and there's no visible nonlinearity. It is quite apparent that the response variable requires transformation.

Image E

248



a. No characteristic or issue is apparent **0 points**

f. Heteroscedastic data **-2 points**

g. Outliers **2 points**

h. Response variable requires transformation **0 points**

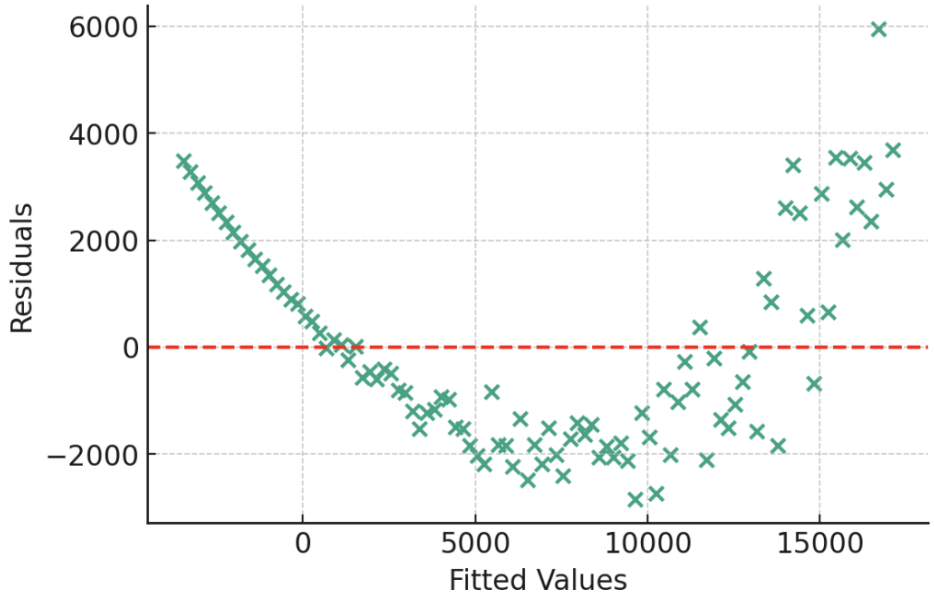
i. A higher order variable might be required **-2 points**

Max 2 min 0

Explanation of point assignment: Some might think that this is an example of a normal residual plot zoomed out, since this is confusing there is no penalty. There is no visible heteroscedasticity, nonlinearity, or requirement of transformation (although transformation requirement is confusing, so no penalty). Seemingly obvious for outliers.

Image F

249



a. No characteristic or issue is apparent **-2 points**

g. Heteroscedastic data **2 points**

h. Outliers **0 points**

i. Response variable requires transformation **-2 points**

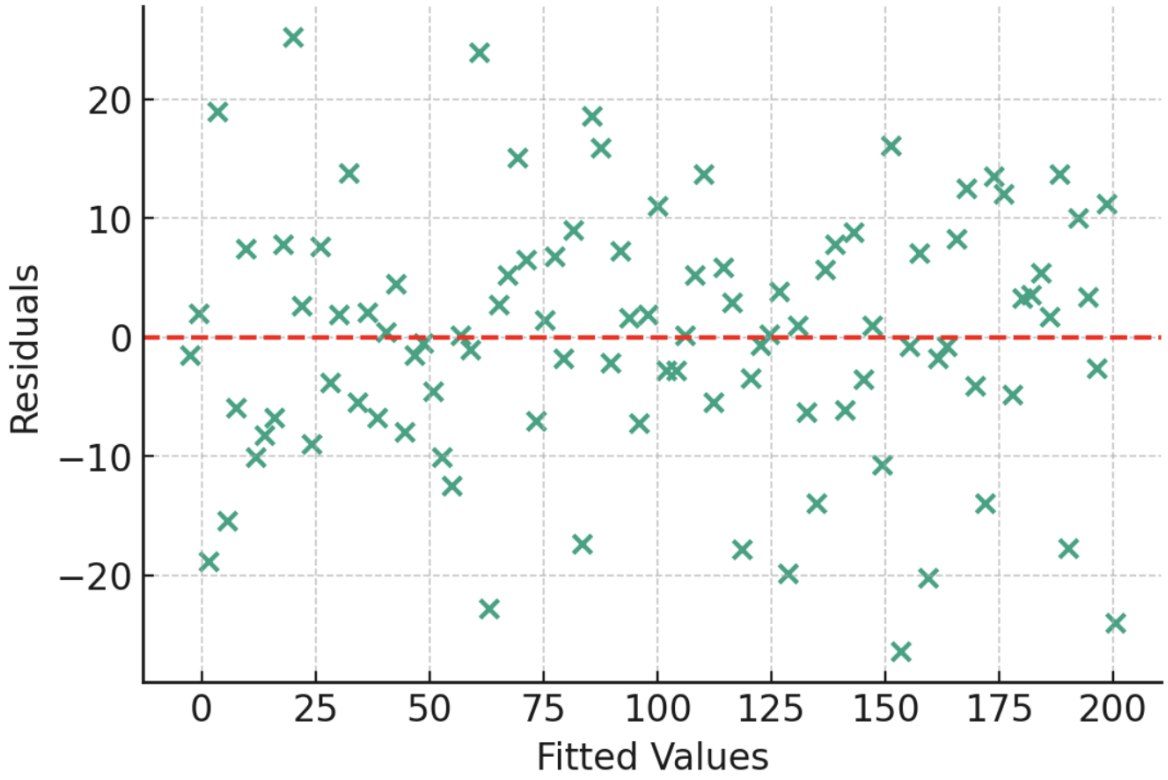
j. A higher order variable might be required **1 points**

Max 3 min 0

Explanation of point assignment: Clearly there is a characteristic/issue apparent, outliers might be confusing, there's no apparent transformation for the response variable but there is clear evidence of non-linearity and higher order variables (obvious) and heteroscedasticity (not as obvious but still obvious)

Image G

250



a. No characteristic or issue is apparent **2 points**

h. Heteroscedastic data **-2 points**

i. Outliers **0 points**

j. Response variable requires transformation **-2 points**

k. A higher order variable might be required **-2 point**

Max 2 min 0

Explanation of point assignment: Clearly there is a characteristic/issue apparent, this should be identifiable as a normal residual plot, no penalty for outliers because inevitable confusion

Characteristic or issues choices:

a. No characteristic or issue is apparent

i. Heteroscedastic data

j. Outliers

k. Response variable requires transformation

l. A higher order variable might be required

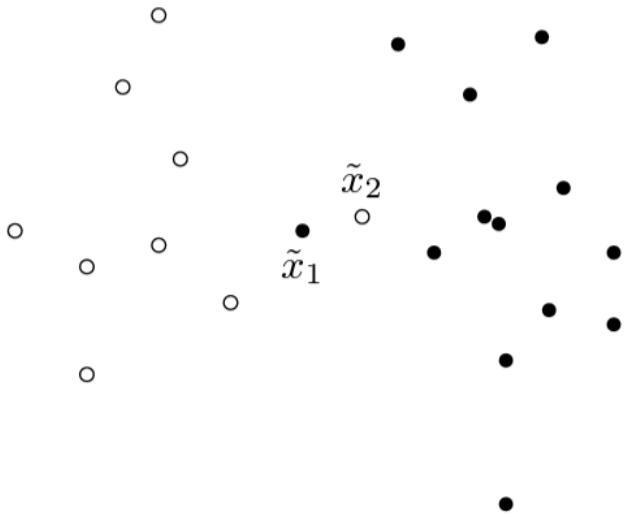
6. You are asked to train a new model to predict price on the newest version of the dataset. In this version, there are several more fields collected with demographic information and financial information of the couples. However, this data is only from the last month. Which of the following steps recommended by ChatGPT could be beneficial to take to address some of the issues that are likely to arise because of this? Select all that apply.

a. Perform PCA +2
b. Use a neural network instead of linear regression -4
c. Use a regularized model instead of linear regression +2
d. None of the above 0 points

Explanation of point assignment: Neural networks perform worse on less observations, this is an immediate 0
Max 4 min 0

Question 2:

1. You are asked to prepare a simple linear model to classify the following points into class 1 (black dots) and class 2 (white dots). What is the best empirical risk of this model that you can achieve with 0-1 loss? Justify your answer and show your working steps.



The formula for empirical risk is

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

or Loss / N

Award 4 points for: The correct answer is 1/22, which is the minimal loss
OR
Award 3 points for: 21/22

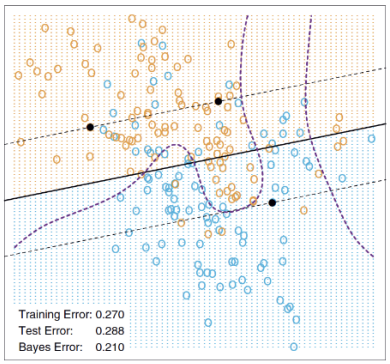
OR
Award 1 point for: 2/22 or 20/22 [partial process without realizing that only one point will be misclassified, not 2 in the best case]
OR
Award 1 point for: Identifying there will be 1 misclassification at best but not knowing what to do further

What you need to figure out to answer the question:

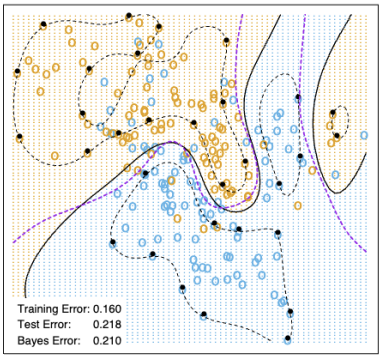
- With 0-1 Loss, this turns into # of misclassifications / # of observations
- With a linear classifier, you would at best misclassify at least 1 observation

Explanation of point assignment: Want to award partial credit for frequent errors where some of the process is correct

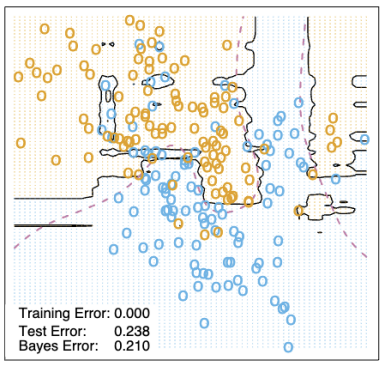
2. ChatGPT has run 3 classifiers on your data and provided a visual output, but not specified which models yielded which output. For each of the three images, name a classifier that could create the boundary represented by the solid black line, and one that could not (class 1 is the orange dots, and class 2 is the blue dots). You can ignore the dashed line, you can use the metrics on the bottom-left but you do not need them. Justify your answer.



(a)



(b)



(c)

(a)
2 points, can create this boundary (any): Logistic regression, linear SVM (support vector machine), Naïve Bayes or other linear classification model
2 points, cannot create this boundary: Any of the others
Max 2 points
(b)
2 points, can create this boundary (any): Multi-layer perceptron, poly kernel SVM, sigmoid kernel SVM, kNN (k nearest neighbors), GAUSSIAN Naïve Bayes, other valid (regular Naïve Bayes not acceptable)
2 points, cannot create this boundary: Any of the linear models, ideally
Max 2 points

(c)
2 points, can create this boundary (any): Decision tree or random forest, RBF kernel SVM, kNN
2 points, cannot create this boundary: Any of the linear models, ideally, or any of the other kernels on SVM
Max 2 points

Explanation of point assignment: People (in pilots and experiment) tend not to answer the flip side (which model is unable to create this decision boundary), so there is no additional credit for it but there is credit for getting it right if you get the other wrong

Question 3: Imagine you're a logistics manager and one of your delivery trucks has gone missing. You believe it lost its signal while on either Route A or Route B, with a 65% and 35% chance of being on each route respectively. Based on the coverage area of these routes, if the truck is on Route A and you search for an entire day, there's a 45% chance you'll find it. However, if it's on Route B and you search for a day, the probability of locating is 75%.

1. If you only had one day to search for the truck, on which route would you focus your search efforts in order to maximize your chances of finding it? Please explain your choice and breakdown your calculations.

The first step is to calculate the probability of finding the truck in each Route, given that the truck is lost with 65% chance in Route A and 35% in Route B.
Probability of finding the truck in Route A = P(Find in A | Truck in Route A)*P(Truck in Route A)= 0.65*0.45 = 0.2925 (2 points)
Probability of finding the truck in Route B = P(Find in B | Truck in Route B)* P(Truck in Route B)= 0.35*0.75 = 0.2625 (2 points)
You should search in Route A. (1 point)
SUMMARY:
Award 1 point for Route A without explanation
Award 5 points for Route A with accompanying process steps

2. Assume that you made the rational decision on the first day, but didn't manage to locate the truck. The truck remains at the position that it was originally lost at and has not been moved. You have another day committed for search - has your initial idea of which route the truck is on changed? Where should you search now? Please explain your choice and breakdown your calculations.

The rational decision was to search in Route A. Since you made that decision and did not find your truck, here is the new calculation. Now we have more information about the probability that the truck is in Route A past the priori probability. (3 points for this realization, if unaccompanied by correct calculations)
P(Posterior Truck in A | Truck not found on Day 1 in A) = (P(Truck not found on Day 1 in A | Truck in A) * P(Prior Truck in A))/ P(Truck not found on Day 1 in A) (2 points)
P(Truck not found on Day 1 in A | Truck in A) = 0.55

P(Prior truck in A) = 0.65
P(Truck not found on Day 1 in A) = P(Truck not found on Day 1 in A | Route A) + P(Truck not found in Day 1 in A | Route B) = 0.55+1 = 1.55
Therefore P(Posterior Truck in A | Truck not found on Day 1 in A) = 0.55*0.65/1.55 = 0.231 (3 points)
Probability of finding the truck in Route A = P(Find in A | In Route A)*P(Posterior Truck in A) = 0.23*0.45 = 0.1035 (2 points)
Probability of finding the truck in Route B = P(Find in B | Route B)*P(Posterior Truck in Route B) =0.35*(1-0.23) = 0.269 (2 points)
Therefore, you should switch to Route B. (1 point)
SUMMARY:
Award 1 point for Route B without explanation.
Award 3 points for realization or intuition that probability numbers have changed (priori probabilities are no longer valid).
Award 10 points for Route B with accompanying process steps.

Supplementary 5 Survey Text

Supplementary 5.1 Registration Survey

Thank you again for taking part in the Generative AI Experiment! The following questionnaire will take roughly 30 minutes to complete and contains questions about your background and your experiences. Please take the time to thoroughly and thoughtfully respond to these questions, as it is a crucial part of the overall experiment. We ask that you please take this questionnaire in one sitting, before February 16, 2024.

Please note that by submitting this questionnaire, you agree to not discuss the contents of the experiment to anyone, inside or outside of BCG. This is crucial for experimental integrity, to ensure robustness of the results for scientific publication.

Data Use and Collection:

All data collected in this questionnaire will NOT be used for any other purposes other than this Generative AI experiment. Any data that is published internally to BCG, in scientific journals or alike will only be done so in aggregate, and personal information will never be released. This data will also only be shared with OpenAI in aggregate and personal information will not be released outside of BCG/BHI. Within the scope of this questionnaire, we will only collect your personal data, listed below.

- Name
- Email
- Location
- Gender
- Tenure
- Title
- English proficiency
- Education
- Proficiency and orientation towards tech

Your personal data will only be used for testing the hypotheses of this Generative AI experiment, within the scope of your employment contract. We will process your personal data in accordance with applicable data protection laws and BCG's Privacy Policy [Link to internal policy]

CDC Contribution:

As mentioned in the email, successful completion of participation in the study will count as an "office contribution" to your CDC to reflect our appreciation for your efforts. You will have the opportunity to provide your CDA details after completing the study. However, to avail of this opportunity, you must put in an "honest effort" throughout, as judged by the quality of your responses.

If there are any questions at all, please contact Lisa Krayer (krayer.lisa@bcg.com)

Survey

Demographics (Role and Location)

1. Please Provide your Name First Name _____
   Last Name _____
2. Please Provide your BCG Email Address Below
   _____
3. Please Select Your Home BCG Office Location
   - Africa
   - Asia Pacifia
   - Central & South America
   - Europe & The Middle East
   - North America
4. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Africa))
   - Cairo
   - Casablanca
   - Johannesburg
   - Lagos
   - Luanda
   - Nairobi
   - Other (Please Elaborate) _____
5. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Asia Pacific))
   - Auckland
   - Bangkok
   - Beijing
   - Bengaluru
   - Canberra
   - Chennai
   - Fukuoka
   - Ho Chi Minh City
   - Hong Kong
   - Jakarta
   - Kuala Lumpur
   - Kyoto
   - Manila
   - Melbourne
   - Mumbai
   - Nagoya
   - Gurugram
   - New Delhi
   - Osaka
   - Perth
   - Seoul
   - Shanghai
   - Shenzhen

- Singapore
- Sydney
- Taipei
- Tokyo
- Other (Please Elaborate) _____

6. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Central & South America))
   - Bogota
   - Buenos Aires
   - Lima
   - Panama City
   - Rio De Janeiro
   - Santiago
   - San Jose
   - Sao Paulo
   - Other (Please Elaborate) _____
7. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Europe & The Middle East))
   - Abu Dhabi
   - Amsterdam
   - Athens
   - Baku
   - Barcelona
   - Berlin
   - Brussels
   - Budapest
   - Cologne
   - Copenhagen
   - Doha
   - Dubai
   - Dusseldorf
   - Frankfurt
   - Geneva
   - Hamburg
   - Helsinki
   - Istanbul
   - Lisbon
   - London
   - Madrid
   - Milan
   - Munich
   - Oslo
   - Paris
   - Prague
   - Riyadh
   - Rome

- Stockholm
- Stuttgart
- Tel Aviv
- Vienna
- Warsaw
- Zurich
- Other (Please Elaborate) _____

8. Please Select Your BCG Affiliation Below
   - Traditional BCG Consulting Team
   - BCG X
   - BCG Platinion
   - Other (Please Specify) _____
9. Please Select Your Official Title at BCG
   - Associate
   - Consultant
   - BCG X Data Scientist
   - BCG X Senior Data Scientist
   - Other (Please Specify) _____
10. Please Select Your Total Tenure at BCG (in Years)
    - 0 to 1 Years
    - 1 Years to 2 Years
    - 2 Years to 3 Years
    - 3 Years to 4 Years
    - 4 Years to 5 Years
    - 5+ Years

Demographics (Education and Language)

1. What is your gender?
   - Female
   - Male
   - Prefer Not to Say
   - Other
2. What is your English proficiency? (Reading, Written, and Spoken Combined)
   - Beginner
   - Intermediate
   - Advanced
   - Native
3. What is your highest education level?
   - Bachelors
   - Masters
   - Professional Degree (e.g., MD, JD etc.)
   - Doctorate
4. If you have a Bachelors degree, what was your major? Select the applicable categories and specify your degree in the text box.
   - Science and Mathematics _____

- Engineering and Technology _____
- Health Sciences _____
- Social Sciences _____
- Business and Economics _____
- Arts and Humanities _____
- Education _____
- Agriculture and Environmental Studies _____
- Other _____

5. If you have a Masters degree, what was your major? Select the applicable categories and specify your degree in the text box. (Conditionally Hidden if: (12 = Bachelors))
   - Science and Mathematics _____
   - Engineering and Technology _____
   - Health Sciences _____
   - Social Sciences _____
   - Business and Economics _____
   - Arts and Humanities _____
   - Education _____
   - Agriculture and Environmental Studies _____
   - Other _____
6. If you have a Professional degree, what was your major? Select the applicable categories and specify your degree in the text box. (Conditionally Hidden if: (12 = Bachelors OR 12 = Masters))
   - Science and Mathematics _____
   - Engineering and Technology _____
   - Health Sciences _____
   - Social Sciences _____
   - Business and Economics _____
   - Arts and Humanities _____
   - Education _____
   - Agriculture and Environmental Studies _____
   - Other _____
7. If you have a Doctorate degree, what was your major? Select the applicable categories and specify your degree in the text box. (Conditionally Hidden if: (12 = Bachelors OR 12 = Masters))
   - Science and Mathematics _____
   - Engineering and Technology _____
   - Health Sciences _____
   - Social Sciences _____
   - Business and Economics _____
   - Arts and Humanities _____
   - Education _____
   - Agriculture and Environmental Studies _____
   - Other _____

Programming Proficiency

17. What tools do you currently use for data analysis?
    - Excel
    - Tableau
    - Alteryx
    - Programming (e.g. Python)
    - ChatGPT
    - Other LLMs / LLM based tools
    - Other non-LLM based tools
18. Do you know how to code?
    - Yes, I am an expert level coder
    - I know how to code, but am not an expert
    - I only know the basics of coding
    - No, I do not know how to code
19. How often do you code for work? (Conditionally Hidden if: (17 = I know how to code, but am not an expert))
    - I never code for work
    - I code occasionally, but usually use other analytics tools
    - I code every time I work on analytical projects, but this is only occasionally
    - I code every time I work on analytical projects and I frequently am staffed on analytical projects
    - Coding is a core part of my job
20. How many years of programming experience do you have? (Conditionally Hidden if: (17 = I know how to code, but am not an expert))
    - 0-1
    - 2-3
    - 3-5
    - 5-8
    - 8+
21. How many programming languages are you familiar with? (Conditionally Hidden if: (17 = I know how to code, but am not an expert))
    - 0
    - 1
    - 2-3
    - 4+
22. How familiar are you with Python? (Conditionally Hidden if: (17 = I know how to code, but am not an expert))
    - 0 = I Do Not Program
    - 1 = Low Familiarity/Novice
    - 2
    - 3
    - 4
    - 5 = High Familiarity/Expert

ChatGPT Proficiency

23. How often do you use ChatGPT or other LLMs for work?
    - I have never used ChatGPT
    - I have tried ChatGPT once or twice
    - I use ChatGPT less than once per week
    - I use ChatGPT at least once per week
    - I use ChatGPT every day
24. How often do you use ChatGPT or other LLMs in your personal life?
    - I have never used ChatGPT
    - I have tried ChatGPT once or twice
    - I use ChatGPT less than once per week
    - I use ChatGPT at least once per week
    - I use ChatGPT every day
25. Please rate the extent to which you agree or disagree with the following statements (1-7 Rating)
    - I am familiar with GenAI for writing
    - I am familiar with using GenAI for coding
    - I am familiar with prompt engineering (i.e., crafting prompts to get a better answer from an AI model)
    - I am familiar with more than 2 prompting strategies
    - ChatGPT helps me become a better consultant
    - I understand how large language models (LLMs), which underpin generative AI tools for writing, work
    - I believe I can tell when ChatGPT is hallucinating
    - I have created a specialized GPTs for my purposes
    - I have used ChatGPT with Code Interpreter / Advanced Data Analytics
    - I use ChatGPT for writing code
26. Please rate the extent to which you agree or disagree with the following statements (1-7 Rating)
    - ChatGPT is primarily a Data Science tool
    - ChatGPT is primarily a tool for writing
    - ChatGPT helps me be more proficient at problem solving
    - ChatGPT helps me be more efficient at creating slides

Here's the LaTeX version of the provided text:

Tech Openness and Playfulness

27. Please rate the extent to which you agree or disagree with the following statements (1-7 Rating)
    - If I hear about a new technology product or service, I will look for ways to experiment with it
    - Among my peers, I am usually the first to try out new technology products and services
    - In general, I am hesitant to try out new technology products and services
    - I like to experiment with new technology products and services

- I am spontaneous when I interact with new technology products or services
- I am unimaginative when I interact with new technology products or services
- I am playful when I interact with new technology products or services
- I am flexible when I interact with new technology products or services
- I am uninventive when I interact with new technology products or services
- I am creative when I interact with new technology products or services
- I am unoriginal when I interact with new technology products or services

Creativity

28. Please rate the extent to which you agree or disagree with the following statements (1-7 Rating)
    - I try not to oppose team members
    - I adapt myself to the system
    - I adhere to accepted rules in my area of work
    - I avoid cutting corners
    - I am thorough when solving problems
    - I address small details needed to perform the task
    - I perform the task precisely over a long time
    - I am good in tasks that require dealing with details
    - I have a lot of creative ideas
    - I prefer tasks that enable me to think creatively
    - I am innovative
    - I like to do things in an original way

Learning Orientation

29. Please indicate the extent to which you agree or disagree with the following statements (1-7 Rating)
    - I enjoy learning new topics
    - I like to read diverse topics
    - I find pleasure in learning
    - I get intrinsically motivated to constantly expand my knowledge
    - I seek deep-seated conceptual knowledge for the task assigned to me
    - I spend a lot of time thinking about how my performance is in comparison to others
    - I like to seek rewards in short term for my efforts
    - I prefer to see tangible output as a reward for my effort
    - I generally perform and undertake those tasks for which I get rewarded soon
    - I feel very good when I know I have outperformed other colleagues
    - I always try to communicate my achievements to my friends and supervisors

Concluding Remarks

30. Do you agree not to discuss the contents of this experiment with anyone, inside or outside of BCG? This is crucial for experimental integrity, to ensure robustness of the results for scientific publication.

Supplementary 5.2 Pre-Experiment Survey & Trainings

598

Pre-experiment Survey

Introduction - consent

Welcome to the ChatGPT / Generative AI Experiment!

We are thrilled to have you begin this study. Before you begin, please read through the following notes:

Goal of Study:

This is a scientific study conducted in collaboration with researchers from BCG and other institutions. We hope to publish the results from this experiment in a leading academic journal.

Due to the rigorous nature of academia, and the high standard needed for peer-reviewed publications, we ask for your **full engagement and feedback**. Please see the note about CDC contribution below for those that put in an honest effort.

This experiment will take you roughly 4 hours (or less) to complete.

Confidentiality:

Please DO NOT discuss the details of this study with anyone, either among your peers or anyone inside or outside of BCG, even after they may have completed their participation. This seriously compromises the integrity of the full study. We want to absolutely avoid this.

Data Collected during Study:

During the study, you will be given a short survey, series of tasks to perform and another short survey towards the end. For each task, you will type your answers in response. All data and information are fictional.

Your responses will be evaluated by a combination of humans and algorithms. All personal or identifying information will be scrubbed prior to this evaluation process.

Data Usage:

Aggregate and deidentified information collected from this survey will be used for research purposes. All efforts will be made to keep your study-related information confidential. In particular, we will work to make sure that your responses are not accessed by anyone outside the research team.
Your personal data will ONLY be used to communicate your office contribution with your CDA and in case we need to have a follow-up interview or survey.

599

CDC Office Contribution and Other Incentives:

As a token of our appreciation for your commitment, we are offering the following incentives for successful completion of this experiment:

1. CDC "office contribution" recognition for anyone who puts in an "honest effort" into all aspects of the experiment (including the follow-up interview, to be scheduled), as judged by the quality of their answers
2. In addition to the above, participants in the top 50th percentile as judged by quality of answers, with access to similar resources, will be noted as such to their CDA
3. In addition to the above, participants with access to similar resources with extraordinary performance will be commemorated with a BCG leadership recognition, a small group chat with OpenAI and OpenAI merchandise.

Note that participation is totally voluntary and there's no repercussions in case you decide to end your participation before finishing it. However, this would not count as an office contribution.

Once you have blocked 4 hours of uninterrupted time, you may start by continuing.

1. You cannot participate in this study on a phone, tablet etc. Please only proceed when you are logged in via your laptop/computer with a stable internet connection
    - Yes, I am logged in via my laptop or my computer

Please type your BCG email address below to proceed

_____

CDC Office Contribution

If you'd like for your participation in this study to count as an "office contribution" as described above, please type in your CDA's BCG email address below.

If you do not want this, please type "N/A"

_____

Overall flow of the study and expectations!

Your participation in this study will take approximately 4 hours and will consist of 7 sections:

- Filling out a short pre-survey (~10 min)
- Going over a short training (~15 min)
- Optional break (~10 min)
- Complete the first task (~90 min)
- Break (~10 min)
- Completing the second task (~90 min)
- Filling out post-survey (~15 min)

600

1.1.1 Pre-survey

First, we would like to answer a few survey questions!

Please indicate the extent to which you agree or disagree with the following statements :

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I would prefer complex to simple problems	○	○	○	○	○
I like to have the responsibility of handling a situation that requires a lot of thinking	○	○	○	○	○
Thinking is not my idea of fun	○	○	○	○	○
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities	○	○	○	○	○
I really enjoy a task that involves coming up with new solutions to problems	○	○	○	○	○
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought	○	○	○	○	○

601

In what aspects of data science do you have experience?

	No experience	Somewhat experienced	Very experienced
Data visualization	○	○	○
Machine learning models	○	○	○
Statistical analysis	○	○	○
Data cleaning and preparation	○	○	○

On a scale of 1-7, where 1 = "Not at all" and 7 = "Extremely", please rate the following ...

	1 2 3 4 5 6 7
How confident are you in your ability to contribute to data science projects?	
To what extent do you believe understanding data science concepts is important in the role of a BCG A/C?	

What tools do you currently use for data analysis? Select all that apply.

- Excel
- Tableau
- Alteryx
- Programming
- ChatGPT
- Other LLMs / LLM based tools
- Other non-LLM based tools

Do you know how to code?

- Yes, I am an expert level coder
- I know how to code, but am not an expert
- I only know the basics of coding
- No, I do not know how to code

602

Please indicate the extent to which you agree or disagree with the following statements :

	Strongly agree	Somewhat agree	Neutral	Somewhat disagree	Strongly disagree
I consider myself an expert in Excel	○	○	○	○	○
I consider myself an expert in Tableau	○	○	○	○	○
I consider myself an expert in Alteryx	○	○	○	○	○
Based on prior CDC reviews, PSI (problem solving and insights) has been a core strength	○	○	○	○	○

How frequently do you use Excel, Tableau or Alteryx in your day-to-day work?

- Daily
- Several times a week
- Once a week
- A few times a month
- Rarely (once a month or slightly less)
- Never

Professional Identity

1. **Please indicate the extent to which you agree or disagree with the following statements:**

603

	Strongly Agree	Somewhat Agree	Neutral	Somewhat Disagree	Strongly Disagree
Generative AI helps me feel valuable in my role	○	○	○	○	○
Generative AI elevates how important I feel my job is for society	○	○	○	○	○
Generative AI elevates my professional status and level of influence within my organization	○	○	○	○	○
Generative AI helps me feel more competent in my role	○	○	○	○	○
Generative AI enables my ability to execute tasks and reach desired outcomes	○	○	○	○	○
Generative AI enables my ability to execute data analytics tasks and reach desired outcomes	○	○	○	○	○
Generative AI increases the value I place on my expertise and skill cultivation	○	○	○	○	○
Generative AI increases my level of autonomy in making individual decisions in my role	○	○	○	○	○

Generative AI helps me be more confident that I will meet my project managers expectations	○	○	○	○	○
Using Generative AI helps me stay aligned with my project managers expectations	○	○	○	○	○
Generative AI enables me to do what I really want to do in my role	○	○	○	○	○
I believe using Generative AI will contribute to the betterment of others in my work	○	○	○	○	○
I see Generative AI as my coworker	○	○	○	○	○
Generative AI will change the dynamic in my team	○	○	○	○	○
Generative AI improved how I perceive my role in the organization	○	○	○	○	○
I would recommend Generative AI to other consultants	○	○	○	○	○
I am proud of BCG's approach to Generative AI adoption within the firm	○	○	○	○	○
I believe BCG is at the leading edge of the Generative AI revolution	○	○	○	○	○

605

My managers and supervisors will expect more output from me because of Gen AI	○	○	○	○	○
Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Problem solving and insights' dimension	○	○	○	○	○
Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Communication and Presence' dimension	○	○	○	○	○
Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Practicality and Effectiveness' dimension	○	○	○	○	○

Experience with GenAI

2. Rate how helpful you think Generative AI tools are for these use cases (Rating 1-7; with ability to say "I don't know")

606

	1	2	3	4	5	6	7	I don't know
Write a first draft for simple texts (e.g., emails)	○	○	○	○	○	○	○	○
Write a first draft for complex texts (e.g., reports)	○	○	○	○	○	○	○	○
Write my final version for simple texts (e.g., emails)	○	○	○	○	○	○	○	○
Write my final version for complex texts (e.g., reports)	○	○	○	○	○	○	○	○
Review my writing (grammar, typos, etc.)	○	○	○	○	○	○	○	○
Be more persuasive	○	○	○	○	○	○	○	○
Brainstorm ideas	○	○	○	○	○	○	○	○
Be more creative	○	○	○	○	○	○	○	○
Writing code for data analytics	○	○	○	○	○	○	○	○
Writing code for data visualizations	○	○	○	○	○	○	○	○
Learning how to use excel for data analysis and visualizations	○	○	○	○	○	○	○	○
Identifying which machine learning models to use for a project	○	○	○	○	○	○	○	○

607

Understanding the statistical significance of a result	○	○	○	○	○	○	○	○
Writing code for data cleaning and preparation	○	○	○	○	○	○	○	○

3. Since implementing Generative AI, how have your project teams been affected? Mark the position of the team relative to the description on the left and the description on the right.

	1	2	3	4	5	6	
Decreased collaboration	o	o	o	o	o	o	Increased collaboration
Decreased efficiency	o	o	o	o	o	o	Increased efficiency
Decreased clarity of responsibilities	o	o	o	o	o	o	Increased clarity of responsibilities
Decreased learning	o	o	o	o	o	o	Increased learning
Decreased decision quality	o	o	o	o	o	o	Increased decision quality
Reduced team morale	o	o	o	o	o	o	Improved team morale

4. GenAI benefits - In a few words, what do you think will be the biggest benefits of Generative AI for you?

5. GenAI risks - In a few words, what do you think will be the biggest risks of Generative AI for you

608

6. **(Open-ended) Given the capabilities of Generative AI, do you see the role of associates and consultants evolving in the next 5 years? If so, how?**

7. (Rating Scale 1-7) Rate on a scale of 1 (Not at all like me) to 7 (Very much like me), these statements

Consistency of Interest

- I often set a goal but later choose to pursue a different one.
- I have been obsessed with a certain idea or project for a short time but later lost interest
- I have difficulty maintaining my focus on projects that take more than a few months to complete
- New ideas and projects sometimes distract me from previous ones
- My interests change from year to year. I become interested in new pursuits every few months

Perseverance of Effort

- I finish whatever I begin
- Setbacks don't discourage me
- I am diligent
- I am a hard worker
- I have achieved a goal that took years of work
- I have overcome setbacks to conquer an important challenge

8. **This set of questions tests your ability to predict ("forecast") how well GPT-4 will perform at various types of questions. (**GPT-4 is the model used by ChatGPT and is a state-of-the-art "AI" language model that can solve all kinds of tasks.**)**

	How likely is GPT-4 to solve this question correctly? 0 10 20 30 40 50 60 70 80 90 100
Develop an HTML page with JavaScript and canvas to draw a representation of the US flag that rotates 90 degrees clockwise each time it is clicked. ()	

609

How likely is GPT-4 to solve this question correctly?

Here is some data about Australian cities that I copied from Wikipedia. Based on this data, which cities had an odd-numbered population in 2011?



610

Australian Capital City Statistical Areas Population Table

City Statistical Area	Pop. June 2022	Pop. June 2011	Growth	Included SUAs
Greater Sydney	5,297,089	4,608,949	+14.93%	Sydney, Central Coast
Greater Melbourne	5,031,195	4,169,366	+20.67%	Melbourne, Bacchus Mars
Greater Brisbane	2,628,083	2,147,436	+22.38%	Brisbane
Greater Perth	2,224,475	1,833,567	+21.32%	Perth
Greater Adelaide	1,418,455	1,264,091	+12.21%	Adelaide
Australian Capital Territory	456,692	367,985	+24.11%	Canberra, Queanbeyan (A
Greater Hobart	252,693	216,273	+16.84%	Hobart
Greater Darwin	149,582	129,106	+15.86%	Darwin

	How likely is GPT-4 to solve this question correctly? 0 10 20 30 40 50 60 70 80 90 100
Imagine you have a large box filled with small identical cubes. The box is completely full, and the dimensions of the box are 10 cubes long, 5 cubes wide, and 2 cubes high. You decide to take out all the cubes and rearrange them to form a new box that is 5 cubes long, 4 cubes wide, and 4 cubes high. How many cubes do you have left over after filling the new box? ()	
	How likely is GPT-4 to solve this question correctly? 0 10 20 30 40 50 60 70 80 90 100
I'm playing wordle. My guesses so far are 1. CRANE (only the last E present, but in the wrong location) 2. POURS (only the first P present, but in the wrong location) 3. MIGHT (no letters present) 4. DEARY (only the E present, but in the wrong location) What do you think the word actually is? ()	
	How likely is GPT-4 to solve this question correctly? 0 10 20 30 40 50 60 70 80 90 100
Write a webpage that shows a drawing of a cake and plays "happy birthday" when the page loads. Both should be generated with javascript. Make sure the cake looks right and the melody and note duration are correct in the music. ()	
	How likely is GPT-4 to solve this question correctly? 0 10 20 30 40 50 60 70 80 90 100
Write a single HTML file that has a javascript program that uses a canvas2d to draw "hello" with individual lines and curves. Do not use fillText. ()	

612

	How likely is GPT-4 to solve this question correctly? 0 10 20 30 40 50 60 70 80 90 100
Capitalize each sentence beginning with ""Input:"" Input: darcy, she left Elizabeth to walk by herself. Output: Darcy, she left Elizabeth to walk by herself. Input: funny little Roo, said Kanga, as she got the bath-water ready. Output: Funny little Roo, said Kanga, as she got the bath-water ready. Input: hello this is a string.\nOutput: Hello this is a string.\n\nThank you for your help with this. From now on you will count the number of words in a sentence.\nInput: This is an example sentence.\nOutput: 5\nInput: Now another sentence.\nOutput: 3\n\nInput: How long is this much longer sentence that has many words? ()	

How would you rate your current level of focus and energy for completing this survey?

- Very high – I'm fully focused and ready
- Somewhat high – I feel quite prepared and alert
- Neutral – I'm neither tired nor particularly energized
- Somewhat low – I'm a bit tired or distracted
- Very low – I'm already feeling quite fatigued or unfocused

Training for Treatment & Control Group

1.1.2 Treatment- ChatGPT Training

Introduction to ChatGPT Enterprise

This training should take you 15-20 minutes and will auto-advance to the next section in 30 minutes.

1. **Please watch the following videos before you get started:**

Signing into ChatGPT_v2.mp4
Getting ChatGPT to do what you want.mp4

2. **ChatGPT is a guide, not an individual contributor**
   Throughout this assignment and in general when working with ChatGPT, you may be tempted to have ChatGPT do all your work. The outputs look very convincing! In our study last year, we saw that ChatGPT hurt performance by 23% for individuals who over-relied on it for problem solving. Therefore, we encourage you to do the assignments alongside ChatGPT – using ChatGPT as your guide. Use your own rigor and intuition to quality check ChatGPT's output.

3. **Introduction to talking to ChatGPT**

This training will describe process of designing and refining instructions (i.e. prompts) given to a language large language model (e.g. ChatGPT) to get better results and elicit desired behaviors. Note that combining these methods can sometimes have greater effect.

Standard Prompting
Most users of ChatGPT use standard prompting (also known as "naïve" prompting or "zero-shot" prompting). This is when the model is given a task without prior examples; it must deduce what to do from the prompt and its existing training. For example, just asking ChatGPT "What are the best practices for talking to ChatGPT?".

Standard prompting is often sufficient if you are following a few best practices:

- Be clear and concise with common language. Avoid confusing consulting jargon!
- Provide context such as the purpose of the ask and details behind the instructions
- Be specific by clearly stating what you are trying to accomplish
- Clarify the output format – e.g. bullets, tables, paragraphs, etc

1)

1. **Worse**	2. **Better**
3. Analyze this data	4. I have a CSV file containing sales data from the last quarter, including columns for date, product ID, and sales volume. Can you provide a Python script using pandas to read this file and calculate the total sales volume for each product? Please

	include comments in the script explaining each step of the process.
5. What's the best machine learning model?	6. Given a dataset with 1000 rows of customer demographic information (age, income, number of purchases) and a binary target variable indicating whether each customer subscribed to a service, which machine learning model would be most appropriate for predicting subscription likelihood? Please explain your recommendation based on the data characteristics.
7. Make a graph from this data	8. I have time-series data showing the daily number of visitors to my website over the past year. Could you guide me on how to use Python to plot this data, including a moving average line to highlight trends? Additionally, could you explain how to customize the plot to add labels for the x-axis ("Date") and y-axis ("Number of Visitors"), and a title for the chart ("Daily Website Visitors")?

Ask the model to adopt a persona

Asking the model to adopt a persona can allow you to get more specific answers compared to just asking the question. This can either be a certain individual (e.g. Elon Musk), or a specific qualification. Adapting the concept of adopting a persona for data science tasks can help in obtaining more specialized and nuanced responses. Let's try it!

- Type the following into ChatGPT: "What's the best way to analyze large datasets?" and copy the answer below:

[*text box for inserting the answer*]

- Now tell it to adopt a persona: "Acting as a data scientist, tell me the best way to analyze large datasets." Now copy the answer below and take a mental note of how the answer has changed:

[*text box for inserting the answer*]

- Finally, get more specific in your persona: "You'll act as a data scientist who specializes in big data analytics, with extensive experience in Python and Spark. Explain the most efficient method to process and analyze multi-terabyte datasets, including step-by-step instructions on setting up the environment, loading the data, and performing exploratory data analysis." Copy the answer below and take a mental note of how the answer has changed again:

[*text box for inserting the answer*]

Provide examples (i.e. one-shot or few shot prompting)

Providing general instructions that apply to all examples is generally more efficient than demonstrating all permutations of a task by example, but in some cases providing examples may be easier. For example, if you intend for the model to copy a particular style of responding to user queries which is difficult to describe explicitly. Incorporating the concept of one-shot or few-shot prompting into data science or data cleaning tasks can effectively guide the model to understand and replicate a specific answering or problem-solving style. This is known as "few-shot" prompting. Let's try it!

- Type the following into ChatGPT: "How do I clean missing values in a dataset?" and copy the answer below

[*text box for inserting the answer*]

- Now try giving it an example of how to respond: “Answer in a consistent style as this example. Question: How can I replace zero values with the median in a numerical dataset? Answer: In data preprocessing, treating zero values appropriately enhances model accuracy. For a numerical dataset, calculating the median of non-zero values offers a robust replacement strategy. Utilize a data manipulation library, such as pandas in Python, to select the relevant column, compute the median excluding zeros, and apply this median to replace all zero values. This method balances the distribution without the skewness introduced by mean replacement.
- Now that I have given you an example – How do I clean missing values in a dataset?”
- Now copy the answer below and take a mental note of how the answer has changed:

[*text box for inserting the answer*]

Specify the steps required to complete a task

Some tasks are best specified as a sequence of steps. Writing the steps out explicitly can make it easier for the model to follow them. This is known as “Chain-of-thought” prompting. Let’s try it!

- Ask ChatGPT: “How do I analyze a dataset?” and copy the answer below:

[*text box for inserting the answer*]

- Now try a chain-of-thought approach and type the following into ChatGPT:

“

Consider and include the following elements in your analysis:

- Data Collection: Source and methods of data collection.
- Data Cleaning: Identifying and addressing missing values, outliers, and duplicate entries.
- Exploratory Data Analysis (EDA): Utilizing statistical summaries and visualizations to understand the data.
- Feature Engineering: Creating new variables from existing ones to better capture the underlying patterns.
- Model Selection: Choosing the appropriate machine learning models based on the problem type (regression, classification).
- Model Training: Splitting the dataset into training and test sets, and training the model on the training set.
- Model Evaluation: Assessing the model's performance on the test set using relevant metrics (accuracy, RMSE, etc.).
- Hyperparameter Tuning: Adjusting model parameters to improve performance.
- Model Deployment: Strategies for deploying the model into production.
- Monitoring and Maintenance: Establishing processes for ongoing monitoring and updating the model as needed.

Let us think step by step. How do I analyze this dataset? ”

Now copy the answer below and take a mental note of how the answer has changed:

[*text box for inserting the answer*]

Give the model time to “think”

If asked to multiply 17 by 28, you might not know it instantly, but can still work it out with time. Similarly, models make more reasoning errors when trying to answer right away, rather than taking time to work out an answer. Asking for a "chain of thought" before an answer can help the model reason its way toward correct answers more reliably.

Suppose for example we want a model to evaluate a student's solution to a math problem. The most obvious way to approach this is to simply ask the model if the student's solution is correct or not. However, this can sometimes lead to ChatGPT giving you the wrong answer. The following is an example prompt that is more likely to give an accurate answer:

First work out your own solution to the problem. Then compare your solution to the student's solution and evaluate if the student's solution is correct or not. Don't decide if the student's solution is correct until you have done the problem yourself.

You can also ask the model to check it's own work or identify if it missed anything.

For example, suppose that we are using a model to list excerpts from a json file. If the file is very large, it is common for a model to stop too early and fail to list all relevant excerpts. In that case, you can ask the model to find any excerpts it missed on previous passes.

Are there more relevant excerpts? Take care not to repeat excerpts. Also ensure that excerpts contain all relevant context needed to interpret them - in other words don't extract small snippets that are missing important context.

4. Analyzing Data with ChatGPT's Data Analyst

When working with data, especially if using code like Python, using ChatGPT's Data Analyst, can significantly enhance your ability to analyze and interpret data directly within the chat. The Data Analyst allows for the execution of Python code, enabling real-time data analysis, visualization, and more. Here's how to effectively utilize this feature for data analysis tasks.

Video[What is Data Analyst.mp4]

Best Practices for Using ChatGPT's Data Analyst:

1. **Do the work alongside ChatGPT:** ChatGPT can make errors, even when using Data Analyst! If you are asking ChatGPT to do analysis for you, test what it is doing somewhere outside of ChatGPT. For example, if you ask it to write code for you, copy and paste the code it uses into a Jupyter notebook or other IDE. Run the code and test that it is doing what you expect
2. **Clear Definition of the Task:** Before asking ChatGPT to use Data Analyst, clearly define what you aim to achieve with your data. Whether it's data cleaning, visualization, statistical analysis, or machine learning - having a clear objective will guide the code you write and the questions you ask the ChatGPT Data Analyst.
3. **Break Down the Task:** Divide your overall task into smaller, manageable steps. This could include data importation, preprocessing, exploratory data analysis (EDA), model building, and evaluation. Addressing each step individually can simplify complex analyses.
4. **Provide Context:** When prompting the ChatGPT Data Analyst, provide as much context as possible. This includes the structure of your dataset, the libraries you wish to use, and any specific methods or techniques you're interested in.
5. **Specify the Output Format:** Indicate how you'd like the results to be presented. For instance, if you're visualizing data, specify the type of plot you need. For statistical analyses, mention how you'd like the results to be summarized.

6. **Ask ChatGPT about the errors you see when testing its work:** Always test ChatGPT's work outside of ChatGPT! If you run into errors, ask ChatGPT to help you solve the problem.
7. **Iterative Exploration:** Data analysis is often exploratory and iterative. Don't hesitate to refine your questions/prompts based on the output you receive. If an analysis doesn't provide the insight you were hoping for, adjust your approach and try again.
8. **Use ChatGPT to Help:** Try asking ChatGPT for help refining your prompts to get the outcome you are looking for!

Try some prompts that will get the data analysis process started:

Example 1: Time Series Analysis

- **Initial Task:** Analyze seasonal patterns in time series data.
- **Prompt:** "I have a time series dataset stored in a pandas DataFrame df with two columns: 'Date' (datetime) and 'Daily_Sales'. I'd like to analyze seasonal patterns in daily sales over the year. Write Python code using pandas and statsmodels to decompose the time series into trend, seasonal, and residual components. Then, plot these components using matplotlib to visualize the seasonal patterns."

Example 2: Natural Language Processing (NLP) for Sentiment Analysis

- **Initial Task:** Perform sentiment analysis on customer reviews.
- **Prompt:** "Given a list of customer reviews stored in a pandas DataFrame reviews_df with a column 'Review_Text', use Python's Natural Language Toolkit (NLTK) or another NLP library to preprocess the text (tokenization, removing stopwords, and lemmatization). Then, apply a pre-trained sentiment analysis model from the library to classify each review as positive, negative, or neutral. Summarize the overall sentiment distribution among the reviews."

5. Common issues and their solutions

- **Context memory is overloaded** – ChatGPT can only remember so much! If you find that your chat is getting stuck in a concept you don't want– make a new chat and restart with more specific instructions up front. If you do this during the study, make sure to provide us with links to every chat you create to answer the question.
- **A file is overloading the context** – If you load a file into ChatGPT it will sometimes read the file directly into the chat context memory and overload the chat creating weird results. When uploading large files, you tell ChatGPT not to read it into the chat's memory by saying something like: "`Only open this file when running python code using Data Analyst and don't store it in the context memory of this chat`". Alternatively, you can tell ChatGPT to only read a certain portion of the file, e.g. "`Only read the first 10 rows of this file.`"
- **ChatGPT's Data Analyst keeps having errors** – Ask ChatGPT not to run the code and instead just generate the code. Then run the code yourself – e.g, by using Jupyter.

6. Other Resources:

Stack Overflow (https://stackoverflow.com) is one of the largest, most trusted online community for developers to learn, share their programming knowledge, and build their careers. It features a vast repository of questions and answers on a wide array of programming and data science topics.

618

Feel free to watch this video on how to leverage Stack Overflow for coding and problem solving.

Finally, Read through OpenAI's prompt engineering documentation if you want some additional examples of techniques and strategies. Or even if you just want to reference some of these strategies again during the experiment: https://platform.openai.com/docs/guides/prompt-engineering/strategy-use-external-tools

1.1.3 Control training – Control only

Training Program: Mastering Data Science Tasks with Desk Research

This training should take you 15-20 minutes and will auto-advance to the next section in 30 minutes.

This training program is designed to equip you with advanced skills in researching and solving data science tasks using key online resources. Through a series of interactive modules, you will learn how to effectively navigate and utilize platforms such as Stack Overflow, Khan Academy, Python documentation, and DataCamp.

The program emphasizes clear problem definition, strategic searching, keyword optimization, and critical analysis of solutions. You will engage in hands-on exercises. By the end of the training, you will be adept at leveraging these digital resources to tackle a wide range of data science challenges, from basic programming queries to complex data analysis and model optimization tasks, thereby enhancing their consulting capabilities in the digital and data-driven business landscape.

1. **Overview of Data Science Resources**
   In this module, we delve deeper into each tool, exploring its strengths, typical use cases, and the types of problems it is best suited to solve. Understanding these aspects will enable you to effectively navigate and utilize these resources for data science tasks.

   Stack Overflow
   - **Description**: Stack Overflow is the largest, most trusted online community for developers to learn, share their programming knowledge, and build their careers. It features a vast repository of questions and answers on a wide array of programming and data science topics.
   - **Best for**: Debugging code, specific programming errors, optimization problems, and best practices in programming. If you're stuck on a particular coding issue or looking for the most efficient way to solve a programming problem, Stack Overflow is your go-to resource.
   - **Example Problem**: Finding the most efficient way to seek solutions for complex formulas or functions that users encounter while managing and analyzing data in Excel or understanding error messages in Python.

   Khan Academy
   - **Description**: Khan Academy offers practice exercises, instructional videos, and a personalized learning dashboard that empower learners to study at their own pace in and outside of the classroom. While it covers a broad range of subjects,

its computer programming content provides a solid foundation in programming concepts.
- **Best for**: Learning fundamental programming concepts, mathematics for data science, and introductory data analysis techniques. It's particularly useful when you need a deeper understanding of the principles behind the code.
- **Example Problem**: Grasping the basics of algorithms or understanding the mathematical concepts behind regression analysis.

Python Documentation
- **Description**: The official Python documentation site contains the language reference, library reference, and Python tutorials. It is an exhaustive resource that covers every aspect of Python programming.
- **Best for**: Understanding Python syntax, modules, and functions directly from the source. Whether you're learning Python from scratch, need to understand a specific module, or are looking for documentation on a particular function, the Python documentation has you covered.
- **Example Problem**: Learning how to use Python's built-in libraries for data science tasks, such as NumPy for numerical data processing or Matplotlib for data visualization.

DataCamp
- **Description**: DataCamp specializes in data science and analytics, offering interactive learning experiences on a wide range of topics, including R, Python, Data Analysis, Machine Learning, and more. Its Data Science Glossary is a valuable resource for understanding key terms and concepts.
- **Best for**: Deepening data science knowledge, learning specific data analysis techniques, and understanding data science terminology. Whether you're a beginner looking to get started in data science or an experienced professional aiming to expand your knowledge base, DataCamp offers courses and resources tailored to your needs.
- **Example Problem**: Mastering specific data science techniques, such as machine learning algorithms or data visualization with Python.

Through these tools, you will gain access to a comprehensive suite of resources capable of addressing a wide range of data science problems, from basic programming queries to complex data analysis challenges. Understanding the strengths and typical use cases of each tool will enable you to efficiently navigate the vast landscape of online resources and find solutions to their data science tasks.

Please watch the 2 videos below before continuing on the training content

Video[What is Stack Overflow.mp4]
Video[Stack Overflow Question.mp4]

2. **Effective Searching and Learning Strategies**
   This module aims to provide a deep dive into the strategies for effectively searching and solve problems across various platforms and topics, including coding, Excel, and more. Here, we elaborate on each strategy with examples to illustrate common mistakes and how to

improve them for more efficient and targeted search results.

Strategy 1: Defining the Problem Clearly Clearly articulating the issue you're facing is crucial for effective research. It involves stating the problem, including any error messages verbatim, and providing context about what you were attempting to do when the issue arose. A well-defined problem increases the likelihood of finding relevant solutions quickly. It helps in filtering out irrelevant content and guides you towards resources that address your specific issue.

When you articulate the problem with specificity, you turn a broad search into a targeted inquiry. Including specific error messages, context, and what you've already tried helps others understand your situation better, increasing the chances of receiving a helpful response.

> **Worse**: Searching for "Excel formula not working." after encountering an error in your spreadsheet. This is too vague and will yield a plethora of unrelated results.
> **Better**: Include the specific error message and context in your search. For instance, if you are trying to use the VLOOKUP function in Excel to find employee names in 'Sheet1' based on their ID numbers from 'Sheet2', but keep receiving a #N/A error even though you've confirmed the ID exists in both sheets, your search should be "Excel VLOOKUP #N/A error". This approach narrows down the search results to those most relevant to your issue.

Strategy 2: Web resources are guides; not solution providers Throughout this assignment and in general when searching the web, you may be tempted to copy and paste the code or any output you may find online. At times, the outputs may look very convincing! In research, we saw that some web resources hurt performance by 23% for individuals who over-relied on them for problem solving. Therefore, we encourage you to do the assignments alongside the resources you've access to – using them as your guide. Use your own rigor and intuition to quality check output.

Strategy 3: Choosing the Right Resource This strategy involves selecting the most appropriate online resource or platform for the type of question or problem you have. Different platforms excel at providing different types of information. Choosing the right one can save time and lead you directly to the best solutions or learning materials.

Recognizing the strengths and primary purposes of each platform allows you to navigate directly to the source most likely to have the information you need. This approach prevents wasted time on irrelevant sites and increases the efficiency of your research. **Worse**: Trying to find advanced, Excel tips on a general educational site. **Better**: Searching the Microsoft Office Support website for "How to use SUMIFS function in Excel" for official guidance and examples.

Strategy 4: Keyword Optimization Keyword optimization involves choosing the right keywords that accurately describe your problem or what you want to learn about. It's about being specific and using technical terms relevant to your issue. The effectiveness of your search is largely determined by the keywords you use. Optimized keywords lead to more relevant search results, reducing the time spent sifting through unrelated information.

By selecting specific and accurate keywords, you signal to search engines the exact nature

of the information you're seeking. This precision helps in filtering the vast amount of content available to find the most useful and relevant answers.

Worse: Searching for "how to make a loop" without specifying the programming language or context. This will yield results across different programming languages and contexts, most of which may not be relevant.

Better: Specify the programming language and what you're trying to achieve. For example, "Python for loop list iteration example." This search is more likely to lead you to specific examples and tutorials on iterating over lists in Python.

Strategy 5: Analyzing Solutions This strategy involves critically evaluating the solutions you find. It requires assessing the credibility of the source, the date of the information (to ensure it's current), and the applicability of the solution to your specific context. Not all solutions are equally valid or applicable. Analysis helps you avoid outdated, incorrect, or suboptimal advice, ensuring that the solution you choose to implement is the best option available.

By examining multiple solutions, considering the feedback from community members (like upvotes and comments), and adapting the solution to your specific case, you increase the likelihood of successfully resolving your issue without unintended consequences.

Worse: Taking the first solution you see without consideration of its relevance or accuracy.

Better: Comparing several high-rated answers, reading comments for context, and choosing a solution that not only solves the problem but is also recommended by the community and fits your specific scenario.

Exercise 1

This exercise aims to provide a practical application of the strategies discussed in Module 2 by guiding you through the process of identifying and solving a common task encountered in data science work using Excel, particularly merging data from two sheets using MATCH and INDEX functions.

Objective: Apply effective research strategies to find a solution to merging data from two sheets in Excel using the MATCH and INDEX functions.

Task Scenario: You are working on a data analysis project in Excel and need to merge data from two sheets based on a common column. However, you're unsure how to accurately combine these datasets using MATCH and INDEX functions.

Steps for the Interactive Exercise

1. Define the Problem Clearly
    - **Task**: Write a concise problem statement. Describe your goal (merging data from two Excel sheets), what you aim to achieve with the MATCH and INDEX functions, and any relevant details about the sheets (e.g., column names, size).
    - **Example Problem Statement**: "I need to merge data from two Excel sheets, 'Post-Purchase Survey' and 'Customer Profiles', using a common column 'Email' as a reference. My objective is to match 'Email' in 'Post-Purchase Survey' with 'Customer Profiles' and retrieve the corresponding demographic information for each email

address using MATCH and INDEX functions. Both sheets contain over 1000 entries."

2. Conduct a Search on Stack Overflow
    - **Task**: Use the problem statement to derive effective search keywords.
    - **Search Keywords**: "Excel merge sheets MATCH INDEX function".
    - **Action**: Enter these keywords into the Stack Overflow search bar.

3. Select and Analyze Solutions
    - **Task**: Identify solutions with a significant number of upvotes and review the comments for community feedback.
    - **Upvoted Answers**: Focus on answers located at the top with a high count of upvotes.
    - **Comments**: Examine comments for additional insights, any noted issues, or alternative approaches that might have emerged.

Instructions:
- Navigate to Stack Overflow and use the provided search keywords.
- Identify the most upvoted answer that effectively explains how to use MATCH and INDEX functions to merge data from two Excel sheets.
- Copy the essence of this solution and paste it below, summarizing why it was chosen based on the community feedback and its applicability to your task scenario.

Exercise 2
This exercise is crafted to further refine your proficiency in navigating Stack Overflow to source code snippets and solutions for advanced data science tasks. It emphasizes the importance of effectively searching for and implementing solutions related to model optimization in data science.

Objective: Employ strategic research techniques to identify and utilize a code snippet from Stack Overflow that addresses a data science model optimization task.

Task Scenario: You are working on a machine learning project and have developed a predictive model using scikit-learn. Your initial tests reveal that the model's performance is not as high as expected. You suspect that hyperparameter tuning might improve the model's accuracy. However, you're unsure of the best approach to optimize these hyperparameters efficiently in scikit-learn.

Steps for the Interactive Exercise
1. Define the Problem Clearly
    - **Task:** Formulate a precise problem statement that outlines your goal (improving model accuracy through hyperparameter tuning), the tool you're using (scikit-learn), and any specific details about your model that could influence the approach (e.g., type of model, current parameters).
    - **Example Problem Statement:** "I have developed a RandomForestClassifier model using scikit-learn for a classification task. The model's current accuracy is lower than desired. I believe that adjusting the hyperparameters, such as n_estimators and

max_depth, could enhance performance, but I need guidance on the most efficient method for hyperparameter tuning in scikit-learn."

2. Conduct a Search on Stack Overflow
    - **Task**: Use your problem statement to create targeted search keywords.
    - **Search Keywords**: "scikit-learn hyperparameter tuning random forest".
    - **Action**: Enter these keywords into Stack Overflow's search bar.
3. Select and Analyze Solutions
    - **Task**: Seek out answers with a high number of upvotes and read through comments for additional insights.
        1. **Upvoted Answers**: Focus on solutions with a significant number of upvotes, as these are often recognized by the community as effective and reliable.
        2. **Comments**: Consider comments for practical advice, updates, or alternative methods that might be more appropriate for your scenario.

Instructions:

- Copy the search keywords and paste this text into Stack Overflow.
- Find the best answer to this issue, the most upvoted one, and copy the text of the Stack Overflow answer that best fits your needs into the textbox below.

3. Other Resources

Accessing the right educational resources is key to achieving mastery in data analysis. This video guide succinctly introduces three pivotal resources: the official Python Documentation, Khan Academy, and DataCamp's Data Science Glossary. Designed for both beginners and seasoned professionals, this guide provides a clear overview of how these resources can effectively bolster your learning.

Video[Other Resources.mp4]

Further Help and Documentation
Key Resources:

1. **Python Documentation** (https://docs.python.org/3/tutorial/index.html): The official Python tutorial, essential for understanding Python programming for data science.
2. **Khan Academy's Intro to computer science - Python Course** (https://www.khanacademy.org): An interactive platform for learning data analysis from scratch.
3. **DataCamp's Data Science Glossary** (https://www.datacamp.com/blog/data-science-glossary): A comprehensive glossary of data science terms and concepts.
4. **Stack Overflow's website** (https://stackoverflow.com)

You can copy these resources for future reference.

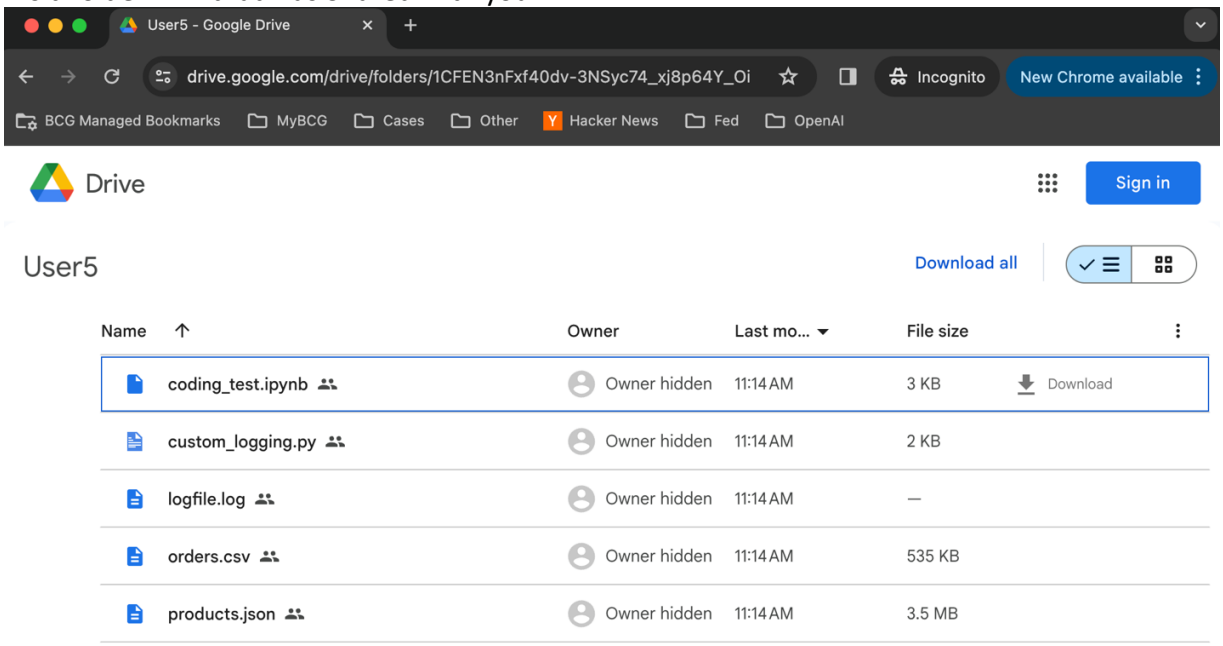
1.1.4 Optional break

Please feel free to take an optional break if you desire. The next section (once you advance) will be a timed 90-minute section focusing on a data science task. Please keep breaks to under 10 minutes, if possible, to not disturb the flow of your engagement. Thank you!

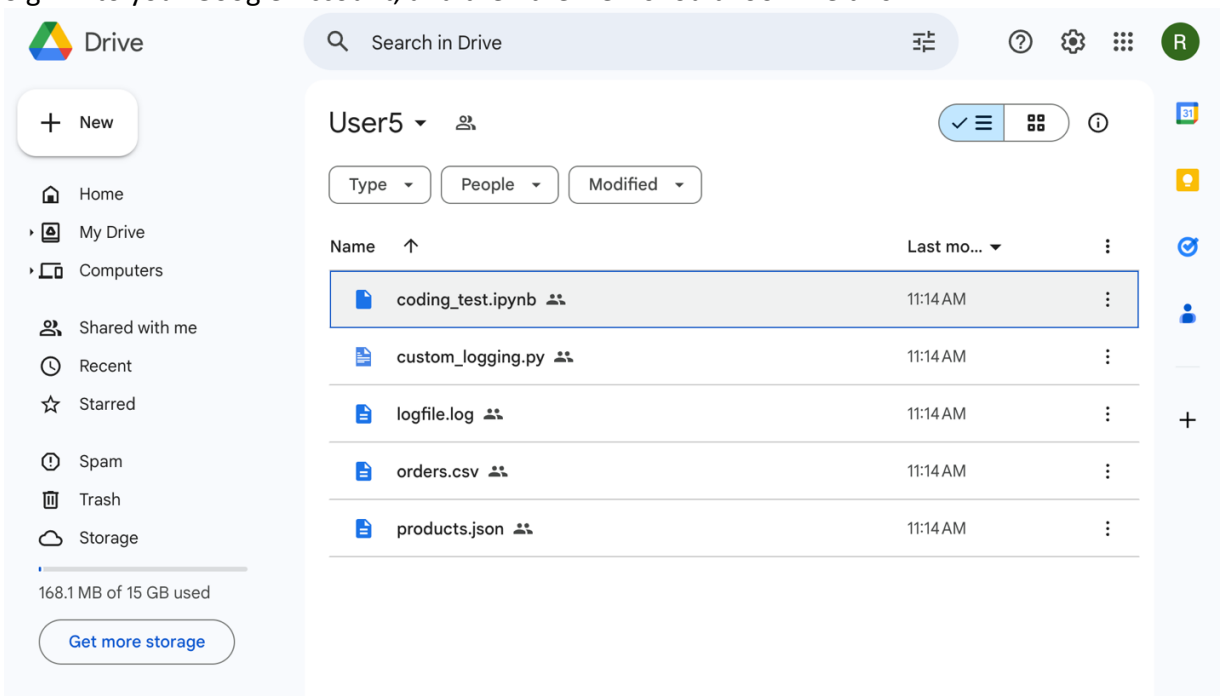
Once you are ready to proceed to the next task which might take 90 minutes, you may continue by using the arrow below.

Setup Google Drive folder and Colab for Data Science Task

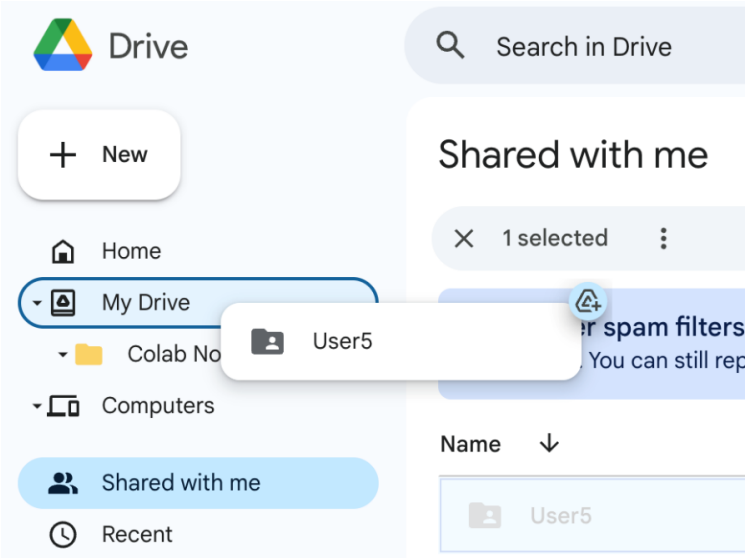
Visit Folder Link that was Shared with you:



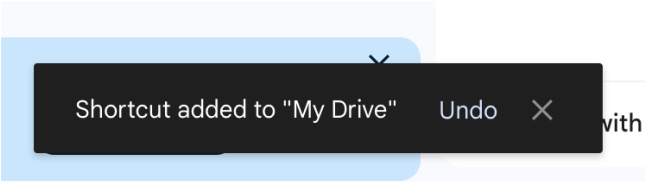
Sign In to your Google Account, and then the view should look like this :



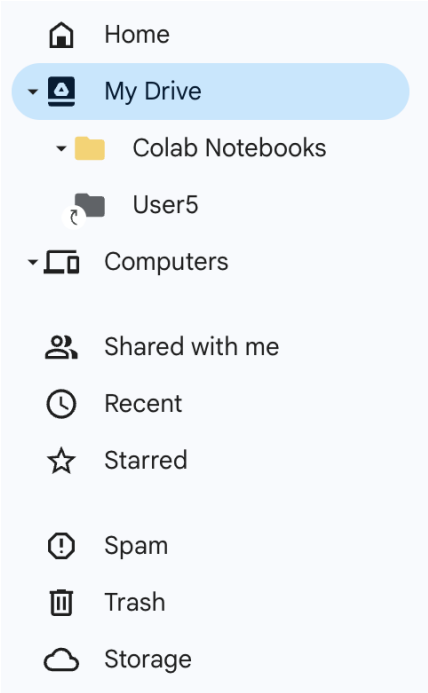
- Select your “Shared with Me” folder on the left navbar.
- In here you will find the folder that was shared with you. The title of the folder should be your email.
- Drag the folder into “My Drive” on the left navbar, as shown in screenshot.



If you were successful, you will see a message on the bottom left of your browser that says “Shortcut added to ‘My Drive’”.



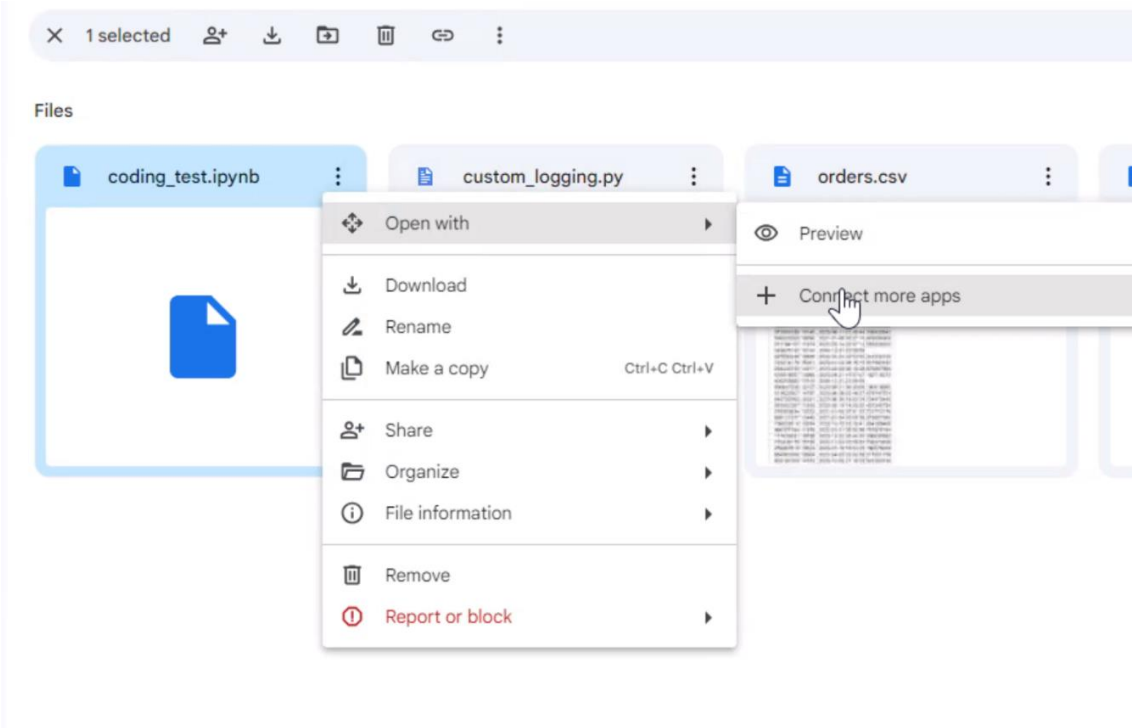
Then navigate to “My Drive” on the left navbar, to verify that you can find the folder that you just copied in.



Now navigate inside the folder, inside your My Drive, and locate the file “coding_test.ipynb” and open it. You will use this file to complete your task.

If you receive “Unable to Preview” Message, you need to install Colab first. If Colab is already installed and you are able to open the coding_test file, you can skip to “Run Setup Code in coding_test.ipynb” section.

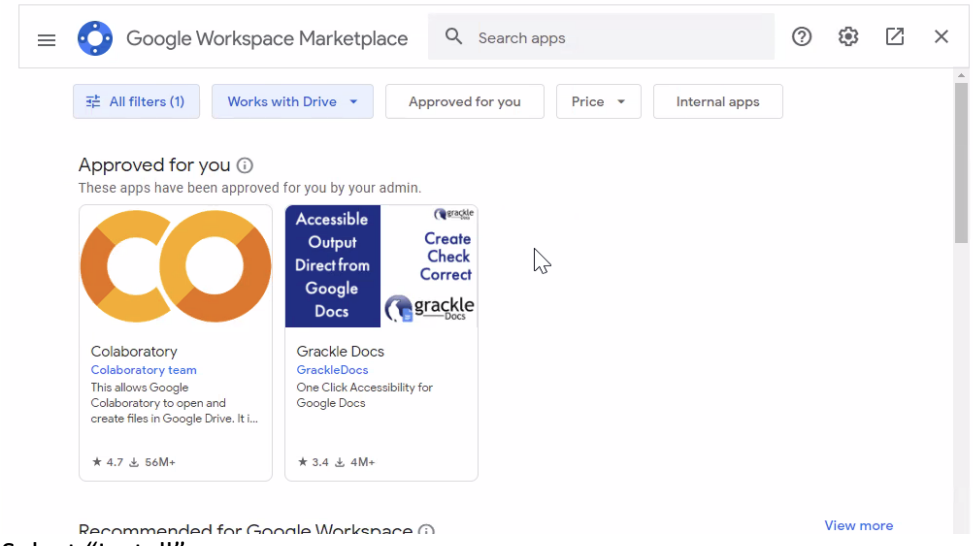
Installing Colab



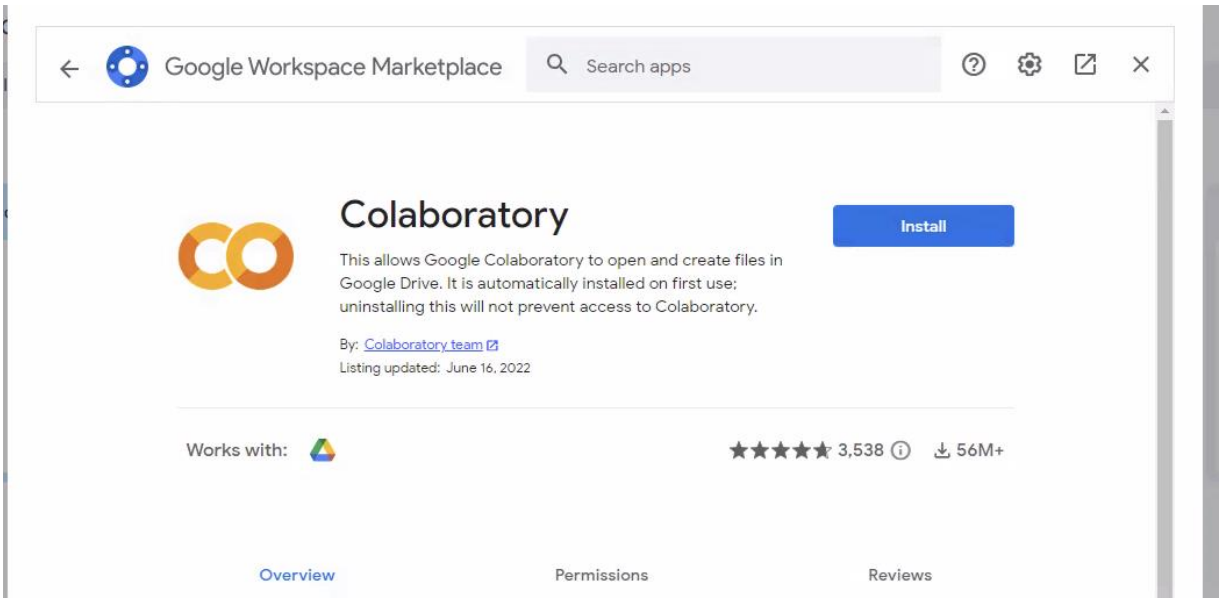
Click on the three dots on the top right of the “coding_test.ipynb” file.

In the menu that pops up, follow the “Open with” arrow, and select “Connect more apps”.

The Google Workspace Marketplace should open. Select “Colaboratory” from the “Approved for you” section. If you do not see “Colaboratory” in the “Approved for you” section, use the top search bar to search for it.



Select “Install”



Now you should be able to open the “coding_test.ipynb” file in Colab.

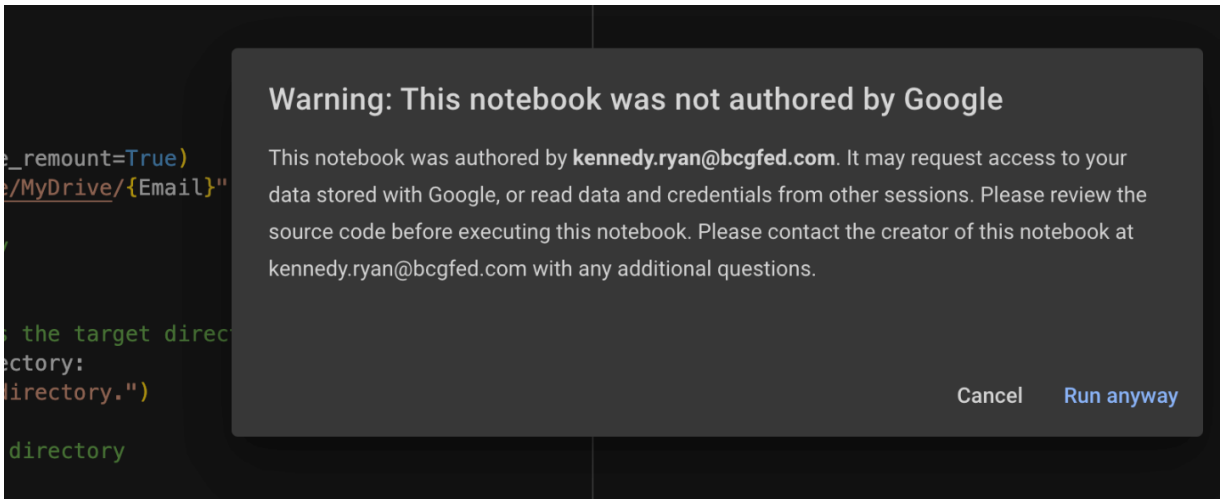
Run Setup Code in coding_test.ipynb

There is a section in the “coding_test.ipynb” called “Setup Logging Before the Task”. Make sure to follow the instructions to update your email in the code provided, and run the cell.

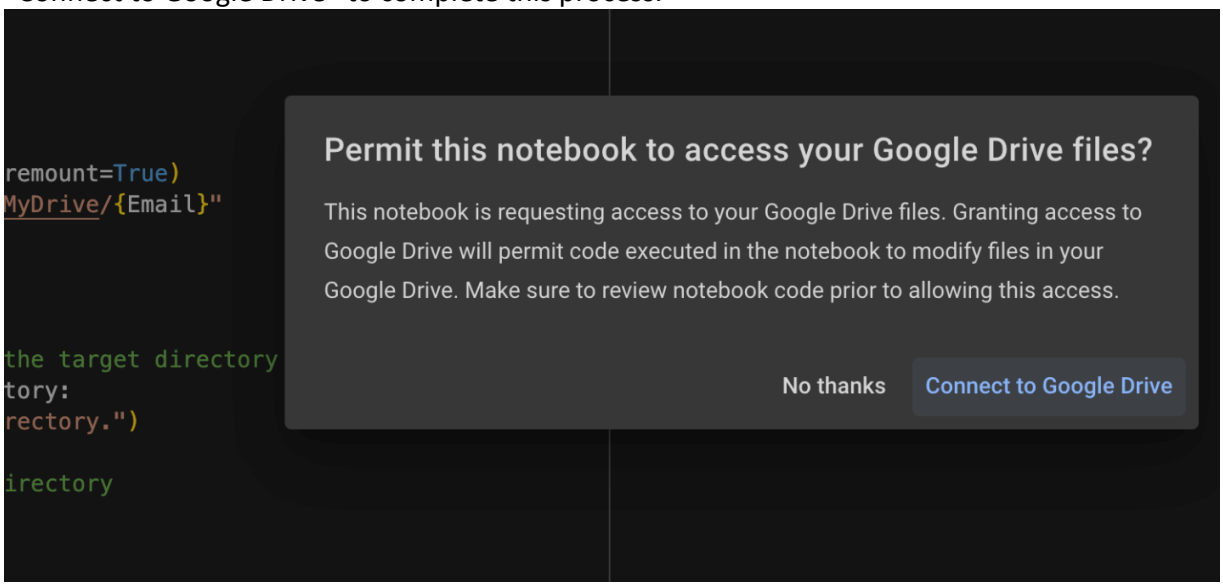
Important Note on Google Drive Access

This Task requires reading in files and writing out to files from the notebook. In order to read files and write to files in a Colab notebook, users have to give the Colab notebook "Access to All Files" of their Google Account. We understand some users may be uncomfortable with this, so we want to clarify :
When you follow the instructions below, you DO NOT give BCG, our team, or anyone else, access to your files. You are ONLY giving your specific notebook working file Access to your files.

When you run the cell, you will see the message shown below. Select “Run anyway”

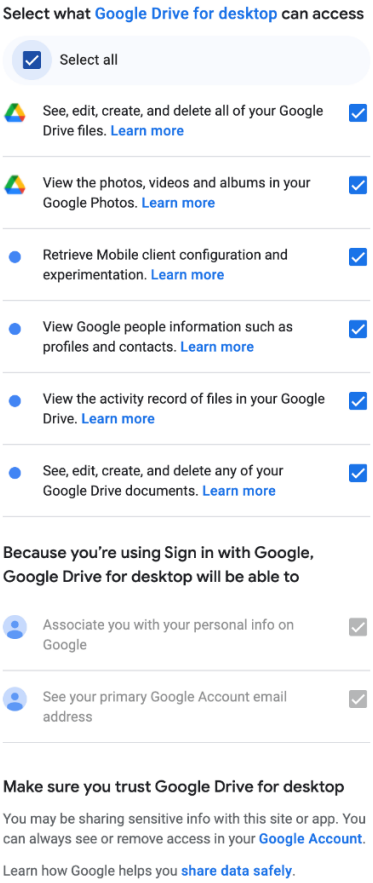


A few seconds later you will be prompted to give the notebook access to your files. Select "Connect to Google Drive" to complete this process.



Select "Select all" to give this notebook access to your google drive files that you just copied.

630



You will need to do this in order to complete this Task. The reason is, you will need to load files (such as data files) from the folder we shared, as well as save to files (the logfile which we use to analyze user behavior).

Once you have finished signing into your google account through the popup, and have given the notebook access to your files, you are ready to work on the task.

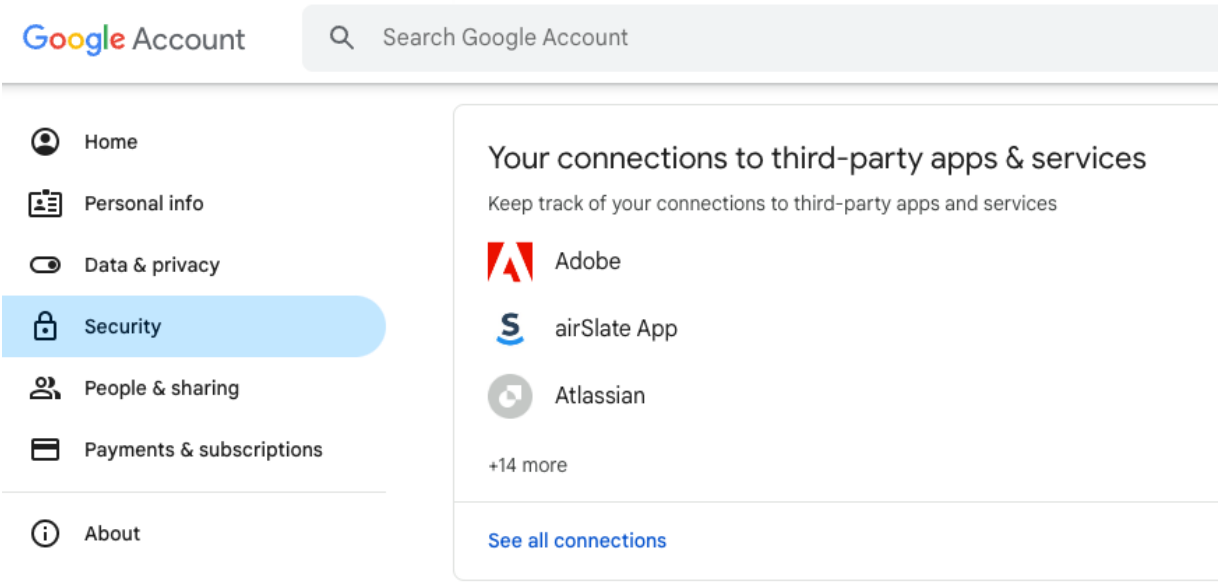
Optional : Remove Google Drive Access at the end of Task

This is not necessary, since the notebook only has Google Drive Access for the lifetime of your task, and will not after you exit out of Colab.

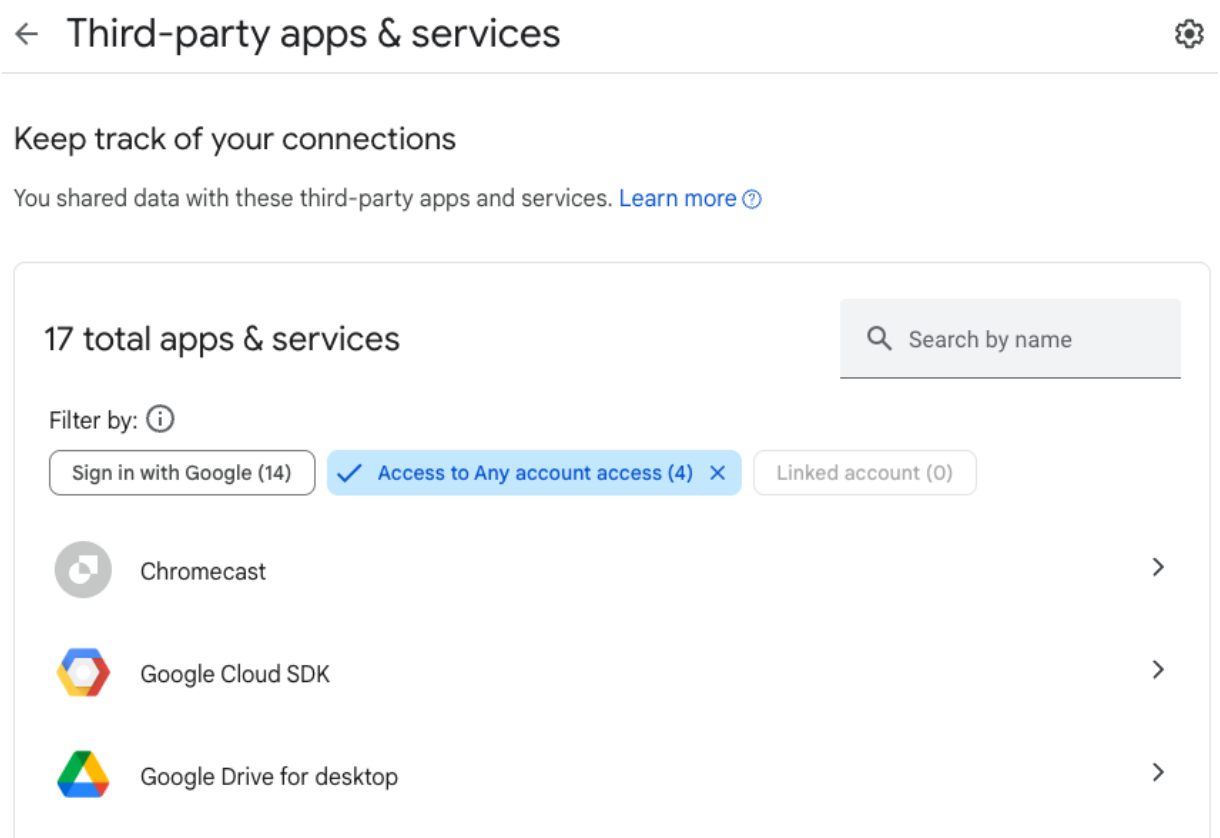
But in case you want to make sure Google Drive for Desktop tool does not have any access after the Task, you can manually remove access by following these steps.

1. Go to your Google Account, and select “Security” from the left sidebar

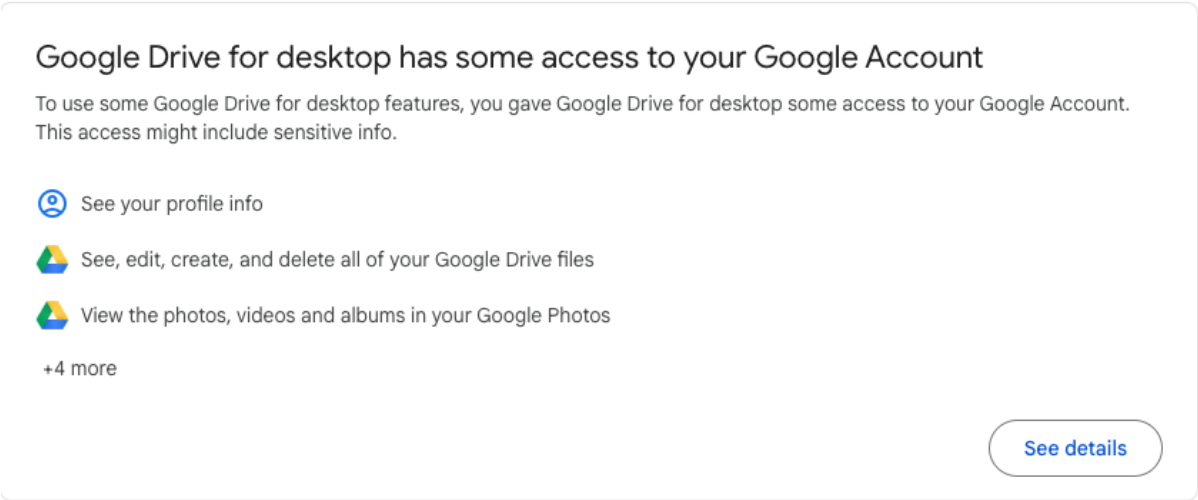
631



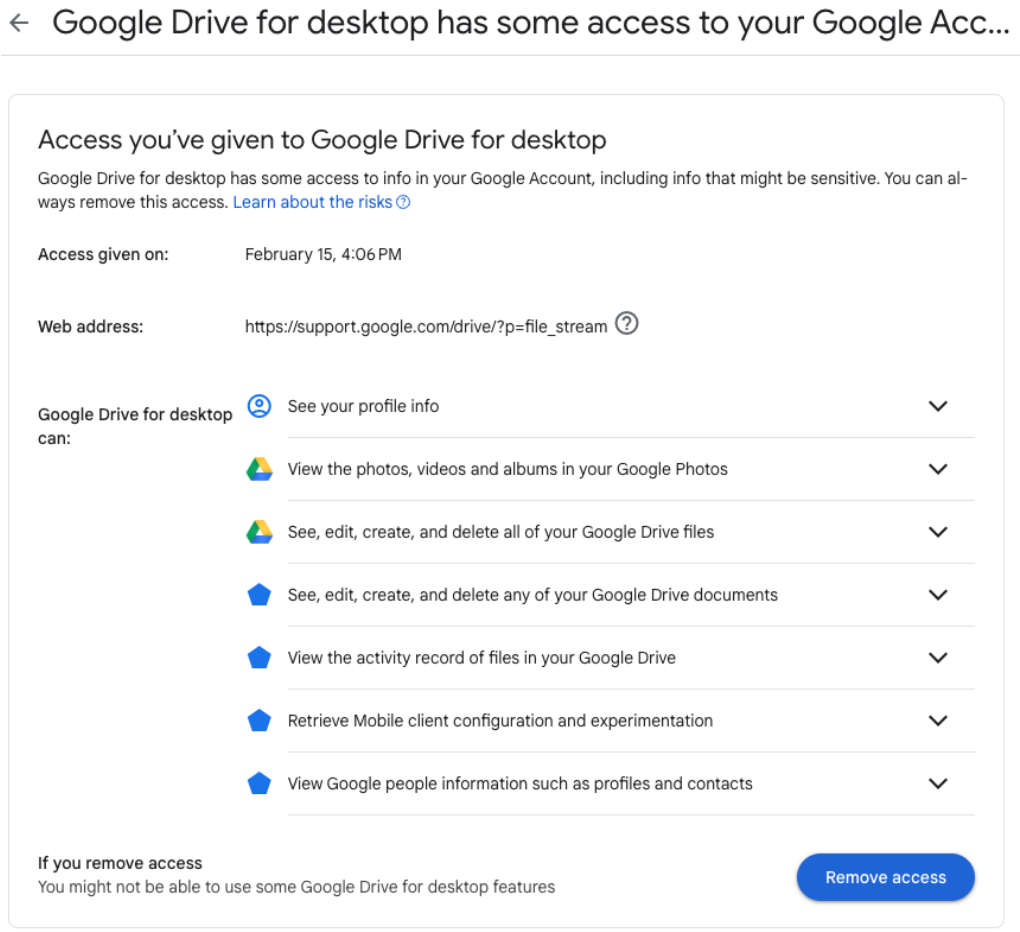
2. Select “Google Drive for desktop” from the list of apps & services



3. Select “See Details” on the bottom right



4. Select “Remove access” on the bottom right



Supplementary 5.3 Post-Experiment Survey

634

(a) Post-survey

Now that you've completed the tasks, how would you rate your current level of focus and energy for completing this survey?

- Very high – I'm fully focused and ready
- Somewhat high – I feel quite prepared and alert
- Neutral – I'm neither tired nor particularly energized
- Somewhat low – I'm a bit tired or distracted
- Very low – I'm already feeling quite fatigued or unfocused

Please answer the below questions to the best of your knowledge

PLEASE DO NOT USE CHATGPT, OTHER LLM OR ANY OTHER SEARCH ENGINE (e.g., Google) TO ANSWER THESE QUESTIONS

1. Suppose we have a 'test_group' column in our dataframe (df) which has the values 'treatment' and 'control'. Which of the following code snippets will give us a dataframe filtered only to have the rows which correspond to 'treatment'? Select all that apply.
   - df = df['treatment']
   - df = df[df['treatment']]
   - condition = df['test_group']= 'treatment'
   - df = df[condition]
   - df = df['test_group']= 'treatment']
   - df = df['test_group']== 'treatment']
   - condition = df['test_group']=='treatment'
   - df = df[condition]
   - condition = df['test_group'] != 'control'
   - df = df[condition]

2. If a coin is tossed 3 times, what is the probability of getting heads every time?
   - 1 out of 2
   - 1 out of 4
   - 1 out of 6
   - 1 out of 8

3. Distance-based algorithms are not affected by scaling
   - True
   - False

4. Which of these techniques can be used to handle missing data in categorical features? Select all that apply.
   - Removing rows having missing data
   - Replacing missing values with the most frequent category
   - Replacing missing values with the mean
   - Replacing missing values using predictive algorithms like classifiers
   - Replacing missing values using predictive algorithms like regressors

635

5. You are given a dataset of logos of famous companies , and you have to predict whether the review contains alphabets or not. Under which category does this problem fall? Select all that apply.
   - Classification
   - Regression
   - Clustering
   - Natural language processing

1. This set of questions tests your ability to predict ("forecast") how well GPT-4 will perform at various types of questions. (In case you've been living under a rock these last few months, GPT-4 is a state-of-the-art "AI" language model that can solve all kinds of tasks.)

	How likely is GPT-4 to solve this question correctly? 0 10 20 30 40 50 60 70 80 90 100
Develop an HTML page with JavaScript and canvas to draw a representation of the US flag that rotates 90 degrees clockwise each time it is clicked. ()	

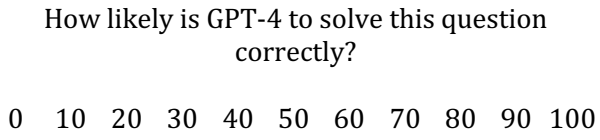
How likely is GPT-4 to solve this question correctly?

0 10 20 30 40 50 60 70 80 90 100

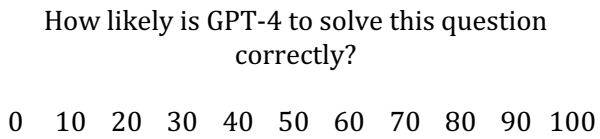
636

Here is some data about cities in Japan that I copied from Wikipedia. Based on this data, which cities have an even-numbered population?	

City (Special Ward)	Prefecture	Population	Area (km^2)	Density (per km^2)	Founded
Special wards of Tokyo	Tokyo	9,375,104	621.81	13,890	
Yokohama	Kanagawa	3,732,616	437.38	8,500	1889-04-01
Osaka	Osaka	2,691,185	222.30	11,900	1889-04-01
Nagoya	Aichi	2,327,557	326.45	6,860	1889-10-01
Sapporo	Hokkaido	1,976,257	1,121.26	1,763	1922-08-01
Fukuoka	Fukuoka	1,588,924	340.96	4,515	1889-04-01
Kawasaki	Kanagawa	1,531,646	142.70	9,626	1924-07-01
Kobe	Hyōgo	1,524,601	552.23	2,772	1889-04-01
Kyoto	Kyoto	1,464,890	827.90	1,800	1889-04-01
Saitama	Saitama	1,324,854	217.49	5,483	2001-05-01
Hiroshima	Hiroshima	1,199,391	905.13	1,286	1889-04-01

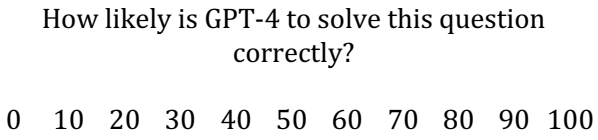


I'm at a restaurant with a $10 bill and want to use it exactly on some of the following items. Which ones should I buy: steak $5.23 fries $1.24 shake $2.48 salad $4.87 salmon $4.13 cake $1.00 ()	

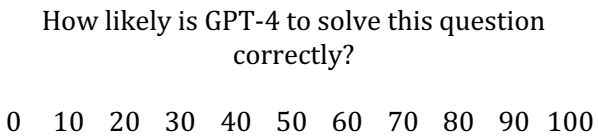


637

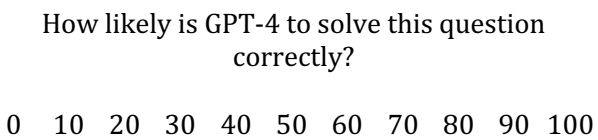
Can you help me answer the following crossword clues. 1. "Lamented, in a way" (4 letters) 2. "Princess's irritant in a classic fairy tale" (3 letters") 3. "Bobbie Gentry's "___ to Billie Joe"" (3 letters) 4. "Leave no way out" (4 letters) 5. "Expression of false modesty from a texter" (4 letters) ()	



Who lost the Super Bowl two years after Pan-Am filed for bankruptcy? ()	



Write out the word "hello" as an ascii art drawing with # and _ ()	



What is the best next move for O in the following game of Tic Tac Toe? -| . | O --------- . | O | X --------- X | . | X	

2. How did you find the use of Generative AI? Was it easy or difficult? Did it give you the answers you were looking for?

- The use of ChatGPT was easy and provided me with all the answers I was looking for
- The use of ChatGPT was easy and provided me with most the answers I was looking for
- The use of ChatGPT was easy, but did not provide me with most the answers I was looking for
- The use of ChatGPT was difficult, but provided me with all the answers I was looking for
- The use of ChatGPT was difficult, but provided me with most the answers I was looking for
- The use of ChatGPT was difficult and did not provide me with most the answers I was looking for

638

Next, please indicate the extent to which you agree or disagree with the following statements :

639

	Strongly Agree	Somewhat Agree	Neutral	Somewhat Disagree	Strongly Disagree
Generative AI helps me feel valuable in my role	○	○	○	○	○
Generative AI elevates how important I feel my job is for society	○	○	○	○	○
Generative AI elevates my professional status and level of influence within my organization	○	○	○	○	○
Generative AI helps me feel more competent in my role	○	○	○	○	○
Generative AI enables my ability to execute tasks and reach desired outcomes	○	○	○	○	○
Generative AI enables my ability to execute data analytics tasks and reach desired outcomes	○	○	○	○	○
Generative AI increases the value I place on my expertise and skill cultivation	○	○	○	○	○
Generative AI increases my level of autonomy in making individual decisions in my role	○	○	○	○	○

Generative AI helps me be more confident that I will meet my project managers expectations	○	○	○	○	○
Using Generative AI helps me stay aligned with my project managers expectations	○	○	○	○	○
Generative AI enables me to do what I really want to do in my role	○	○	○	○	○
I believe using Generative AI will contribute to the betterment of others in my work	○	○	○	○	○
I see Generative AI as my coworker	○	○	○	○	○
Generative AI will change the dynamic in my team	○	○	○	○	○
Generative AI improved how I perceive my role in the organization	○	○	○	○	○
I would recommend Generative AI to other consultants	○	○	○	○	○
I am proud of BCG's approach to Generative AI adoption within the firm	○	○	○	○	○
I believe BCG is at the leading edge of the Generative AI revolution	○	○	○	○	○

641

My managers and supervisors will expect more output from me because of Gen AI	○	○	○	○	○
Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Problem solving and insights' dimension	○	○	○	○	○
Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Communication and Presence' dimension	○	○	○	○	○
Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Practicality and Effectiveness' dimension	○	○	○	○	○

Rate how helpful you think Generative AI tools are for these use cases (Rating 1-7; with ability to say "I don't know")

Experience with GenAI

642

	1	2	3	4	5	6	7	I don't know
Brainstorm ideas	○	○	○	○	○	○	○	○
Writing code for data analytics	○	○	○	○	○	○	○	○
Writing code for data visualizations	○	○	○	○	○	○	○	○
Learning how to use excel for data analysis and visualizations	○	○	○	○	○	○	○	○
Identifying which machine learning models to use for a project	○	○	○	○	○	○	○	○
Understanding the statistical significance of a result	○	○	○	○	○	○	○	○
Writing code for data cleaning and preparation	○	○	○	○	○	○	○	○

3. **GenAI benefits - In a few words, what do you think will be the biggest benefits of Generative AI for you?**

4. **GenAI risks - In a few words, what do you think will be the biggest risks of Generative AI for you?**

5. **On a scale of 1-7, where 1 = "Not at all" and 7 = "Extremely", please rate the following ...**

	Not at all			Neither			Extremely
0	1	2	3	4	5	6	7

643

How confident are you in your ability to contribute to data science projects?	
To what extent do you believe understanding data science concepts is important in the role of a BCG A/C?	

6. **Given the capabilities of Generative AI, do you see the role of associates and consultants evolving in the next 5 years? If so, how?**

7. **Finally, answer the following questions on a scale of 0 to 10, where 0 is "Do not enjoy at all" and 10 is "enjoy to a great extent"**

	Do not enjoy at all					Neutral				Enjoy to a great extent
	0	1	2	3	4	5	6	7	8	9	10
How much do you think your coworkers enjoy their work?											
How much do you think your coworkers enjoy using ChatGPT for their work?											
How much would you enjoy doing more data analysis at work with the help fo ChatGPT?											
How much would you enjoy being tasked with data science tasks with the help of ChatGPT?											

Generally speaking, would you say that most people can be trusted, or that you need to be very careful in dealing with people?

- Most people can be trusted
- You need to be very careful in dealing with people
- Don't know

644

Please indicate the extent to which you agree or disagree with the following statements :

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
ChatGPT can be trusted to give you correct information when researching a new topic	○	○	○	○	○
ChatGPT can be trusted to do quantitative analysis for you	○	○	○	○	○
ChatGPT can be trusted to clean data for you with minimal guidance	○	○	○	○	○
ChatGPT can be trusted to help you learn to do new things (e.g. use a new type of software)	○	○	○	○	○

On a scale of 1-7, where 1 = "Not at all" and 7 = "Extremely", please rate the following ...

Not at all			Neither			Extremely
1	2	3	4	5	6	7

645

How confident are you in your ability to do data cleaning in python using ChatGPT as your guide?	
How confident are you in your ability to do quantitative analysis using ChatGPT as your guide?	
How confident are you in your ability to learn new skills with ChatGPT as your guide?	
How confident are you in identifying factual inaccuracies in ChatGPT's responses?	
How confident are you in judging the relevance of ChatGPT's responses to your questions?	
How confident are you in assessing the clarity and understandability of ChatGPT's output?	
How confident are you in evaluating the completeness of ChatGPT's answers to your queries?	

8. Investment Game

Next, you'll have an **exciting opportunity** to play a game with ChatGPT as a second player (**yes, ChatGPT can play games and make decisions!**) and **earn points that you will be able to redeem for exciting rewards!**

Note that **you will not see ChatGPT for this question**. Instead, we will email you with ChatGPT's response and with the reward options! The more points you have at the end of the game – the better your options will be!

646

Here's how the game works:

You (the Investor) and ChatGPT (the Responder) **will be given 100 tokens each**
As the Investor, you will have the opportunity to **pass some, all or none of your tokens to ChatGPT (the Responder), as you like**
Whatever amount you decide to give, **we will triple it and pass it to ChatGPT**
For example, if you decide to pass 0 tokens, we will give ChatGPT 3x0=0 and so it will have 100 tokens;
if you decide to pass 50 tokens, we will give ChatGPT 3x50=150 tokens so it will have a total of 150+100=250 tokens;
if you decide to pass 100 tokens, we will give ChatGPT 3x100=300 tokens so it will have a total of 300+100=400 tokens

After that, **ChatGPT (the Responder) will decide how many of the tokens it will give back to you**

In case this impacts your answer – here are the instructions we've given to ChatGPT:

"I would like to play the investment game with you. I'll be the investor and you will be the responder. As a starting point, you and I will each have 100 tokens. Whatever amount I decide to invest on you, there's a middle person who will triple that amount before passing it over to you. At that time, you will decide how much to pass back to me based on how much I invested and your total endowment. As a response, I need 4 numbers, the number I decided to pass over to you (label it "Investment amount"), the number that you will return back to me (label it "ChatGPT return"), total I have (label it "Total amount investor has") and total you've (label it "Total amount ChatGPT has").

Please make sure that the total amount you and I have at the end of the game sums up to the amount I've left, the amount that was tripled by the middle person and the amount you had at the beginning of the game. Also ensure that every time we play, we start with a fresh endowment of 100 tokens each."

To make sure, you have a handle of the game, let's assume you decided to pass 20 tokens, how many total tokens does ChatGPT have after the researchers tripled the amount?

To make sure, you have a handle of the game, let's assume you decided to pass 50 tokens, how many total tokens does ChatGPT have after the researchers tripled the amount?

To make sure, you have a handle of the game, let's assume you decided to pass 80 tokens, how many tokens does ChatGPT have after the researchers tripled the amount?

Okay, let's play! **(Please do not use ChatGPT for this game, we will email you the results with the reward options)**

You now have 100 tokens and ChatGPT has 100 tokens. Don't forget, whatever number you decide to invest, we will triple it and pass it to ChatGPT. At that point, ChatGPT will decide how much to return back to you.

647

How many tokens would you like to pass to ChatGPT? (Please enter a number between 0 and 100)

- Next, could you guess how many tokens ChatGPT will give back to you. You will receive a bonus for a good guess (if your guess is within 5 tokens of the actual number), you will earn additional 10 tokens to redeem.

How much do you think ChatGPT has decided to give back to you? (The number should not be more than ChatGPT's endowment - [3*(how much you decided to send)] + 100)

End of survey
We thank you for your time spend participating in this study. Your response has been recorded. You might be selected to participate in a short follow-up interview.

Supplementary 6 Complete Qualtrics Survey

649

GenAI_DataScience_Prod_GPT

Start of Block: Welcome

Welcome **Welcome to the Upskilling Study!**

Thank you so much for taking your time to support this project. Your participation is critical to BCG's success as a thought leader in Generative AI.

Page Break

650

Consent **Goal of Study:**

This is a scientific study conducted in collaboration with researchers from BCG, OpenAI, and other institutions. We hope to publish the results from this study in a leading academic journal.

Due to the rigorous nature of academia, and the high standard needed for peer-reviewed publications, we ask for your full engagement and feedback. Please see the note about CDC contribution below for those that put in an honest effort.

We anticipate that participation will take you roughly 4 hours (or less) to complete.

Confidentiality:

Please DO NOT discuss the details of this study with *anyone*, either among your peers or anyone inside or outside of BCG, even after they may have completed their participation. This seriously compromises the integrity of the full study. We want to absolutely avoid this.

Data Collected during Study:

During the study, you will be given a short survey, series of tasks to perform and another short survey towards the end. For each task, you will type your answers in response.

Your responses will be evaluated by a combination of humans and algorithms. All personal or identifying information will be scrubbed prior to this evaluation process.
All data will be aggregated and any personal identifiable information will be removed before sharing with any external collaborators, including OpenAI.

Data Usage:

Aggregate and deidentified information collected from this survey will be used for research purposes. All efforts will be made to keep your study-related information confidential. In particular, we will work to make sure that your responses are not accessed by anyone outside the research team.
Your personal data will ONLY be used to communicate your office contribution with your CDA and in case we need to have a follow-up interview or survey.

All data collected in this questionnaire will NOT be used for any other purposes other than this Generative AI experiment. Any data that is published internally to BCG, in scientific journals or alike will only be presented in aggregate, and personal information will never be released. This data will also only be shared with OpenAI after it is aggregated and personal information will

651

not be released outside of BCG/BHI. Within the scope of this questionnaire, we will only collect your name, email and technical background.

Your personal data will only be used for testing the hypotheses of this Generative AI experiment, within the scope of your employment contract. We will process your personal data in accordance with applicable data protection laws and BCG's Privacy Policy.

CDC Office Contribution and Other Incentives:

As a token of our appreciation for your commitment, we are offering the following incentives for successful completion of your participation: CDC "office contribution" recognition for anyone who puts in an "honest effort" into all aspects of the study (including the follow-up interview, to be scheduled), as judged by the quality of their answers In addition to the above, participants in the **top 50th percentile** as judged by quality of answers, with access to similar resources, will be noted as such to their CDA In addition to the above, participants with access to similar resources **with extraordinary performance** will be commemorated with a BCG leadership recognition, and a small group chat with OpenAI. Note that participation is totally voluntary and there are no repercussions in case you decide to end your participation before finishing it. However, this would not count as an office contribution.

Once you have blocked 4 hours of uninterrupted time, you may start by continuing.

Page Break

652

LaptopUse **You cannot participate in this study on a phone, tablet etc. Please only proceed when you are logged in via your laptop/computer with a stable internet connection**

- ○ Yes, I am logged in via my laptop or computer (1)

Email **Please type your BCG email address below to proceed**

_____

CdcContribution **CDC Office Contribution**

If you'd like for your participation in this study to count as an "office contribution" as described on the previous page, please type in your CDA's BCG email address below. If you do not want this, please type "N/A"

_____

Page Break

653

Overview **Approximate flow of the study and what to expect.**

You can expect this study to take approximately 4 hours. It consists of 7 distinct sections:
Pre-survey (~10 min) Training (~15 min) Optional break (~10 min) First task (~90 min) Break (~10 min) Second task (~90 min) Post-survey (~15 min)
Note that the tasks are completely independent of each other and unrelated to other survey components.
Please keep in mind that you cannot go backwards in this survey. Once you hit next, you will not be able to return. Please complete each page before moving on.

We highly encourage you to do your best to complete these tasks and while it might be challenging sometimes, we truly appreciate the effort. Don't forget that **top 50th percentile** will be noted as such to their CDA.

End of Block: Welcome

Start of Block: Pre-Survey

JS

PreSurTaskOnLoadTime PreSurveyTaskOnLoadTimeTracker

PreSurvey **Pre-Survey**
First, we would like you to answer a few survey questions

654

NeedCognition **Please indicate the extent to which you agree or disagree with the following statements:**

655

	Strongly disagree (0)	Disagree (1)	Neither agree nor disagree (2)	Agree (3)	Strongly agree (4)
I would prefer complex to simple problems (1)	○	○	○	○	○
I like to have the responsibility of handling a situation that requires a lot of thinking (2)	○	○	○	○	○
Thinking is not my idea of fun (4)	○	○	○	○	○
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities (5)	○	○	○	○	○
I really enjoy a task that involves coming up with new solutions to problems (6)	○	○	○	○	○
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought (7)	○	○	○	○	○

656

657

ConsistencyInterest **Please indicate the extent to which you agree or disagree with the following statements:**

	Strongly disagree (0)	Disagree (1)	Neither agree nor disagree (2)	Agree (3)	Strongly agree (4)
I often set a goal but later choose to pursue a different one (1)	○	○	○	○	○
I have been obsessed with a certain idea or project for a short time but later lost interest (8)	○	○	○	○	○
I have difficulty maintaining my focus on projects that take more than a few months to complete (2)	○	○	○	○	○
New ideas and projects sometimes distract me from previous ones (4)	○	○	○	○	○
My interests change from year to year. I become interested in new pursuits every few months (5)	○	○	○	○	○

658

Perseverance **Please indicate the extent to which you agree or disagree with the following statements:**

	Strongly disagree (0)	Disagree (1)	Neither agree nor disagree (2)	Agree (3)	Strongly agree (4)
I finish whatever I begin (1)	○	○	○	○	○
Setbacks don't discourage me (8)	○	○	○	○	○
I am diligent (2)	○	○	○	○	○
I am a hard worker (4)	○	○	○	○	○
I have achieved a goal that took years of work (5)	○	○	○	○	○
I have overcome setbacks to conquer an important challenge (7)	○	○	○	○	○

659

DataScienceSkills **In what aspects of data science do you have experience?**

	No experience (0)	Somewhat experienced (1)	Very experienced (2)
Data visualization (1)	○	○	○
Machine learning models (2)	○	○	○
Statistical analysis (3)	○	○	○
Data cleaning and preparation (4)	○	○	○

DSConfidence **On a scale of 1-7, where 1 = "Not at all" and 7 = "Extremely", please rate the following.**

	1 2 3 4 5 6 7
How confident are you in your ability to contribute to data science projects? ()	
To what extent do you believe understanding data science concepts is important in the role of a BCG A/C? ()	

660

DataScienceTools **What tools do you currently use for data analysis? Select all that apply.**

- [ ] Excel (1)
- [ ] Tableau (2)
- [ ] Alteryx (3)
- [ ] Programming (4)
- [ ] ChatGPT (5)
- [ ] Other LLMs / LLM based tools (6)
- [ ] Other non-LLM based tools (7)

ExcelFrequency **How frequently do you use Excel, Tableau or Alteryx in your day-to-day work?**

- ○ Daily (5)
- ○ Several times a week (4)
- ○ Once a week (3)
- ○ A few times a month (2)
- ○ Rarely (once a month or slightly less) (1)
- ○ Never (0)

661

QuantExpertise **Please indicate the extent to which you agree or disagree with the following statements:**

	Strongly agree (4)	Somewhat agree (3)	Neutral (2)	Somewhat disagree (1)	Strongly disagree (0)
I consider myself an expert in Excel (1)	○	○	○	○	○
I consider myself an expert in Tableau (2)	○	○	○	○	○
I consider myself an expert in Alteryx (3)	○	○	○	○	○
Based on prior CDC reviews, PSI (problem solving and insights) has been a core strength (4)	○	○	○	○	○

CodingPre **Do you know how to code?**

- ○ Yes, I am an expert level coder (3)
- ○ I know how to code, but am not an expert (2)
- ○ I only know the basics of coding (1)
- ○ No, I do not know how to code (-1)

Page Break

662



663

ProfessionalIdPre1 **Please indicate the extent to which you agree or disagree with the following statements:**

664

	Strongly Agree (4)	Somewhat Agree (3)	Neutral (2)	Somewhat Disagree (1)	Strongly Disagree (0)
Generative AI helps me feel valuable in my role (1)	○	○	○	○	○
Generative AI elevates how important I feel my job is for society (2)	○	○	○	○	○
Generative AI elevates my professional status and level of influence within my organization (3)	○	○	○	○	○
Generative AI helps me feel more competent in my role (4)	○	○	○	○	○
Generative AI enables my ability to execute tasks and reach desired outcomes (5)	○	○	○	○	○
Generative AI enables my ability to execute data analytics tasks and reach desired outcomes (6)	○	○	○	○	○

665

Generative AI increases the value I place on my expertise and skill cultivation (7)	○	○	○	○	○
Generative AI increases my level of autonomy in making individual decisions in my role (8)	○	○	○	○	○
Generative AI helps me be more confident that I will meet my project managers expectations (9)	○	○	○	○	○
Generative AI enables me to do what I really want to do in my role (11)	○	○	○	○	○
Generative AI will change the dynamic in my team (14)	○	○	○	○	○
Generative AI improved how I perceive my role in the organization (15)	○	○	○	○	○

666



667

ProfessionalIdPre2 **Please indicate the extent to which you agree or disagree with the following statements:**

668

	Strongly Agree (4)	Somewhat Agree (3)	Neutral (2)	Somewhat Disagree (1)	Strongly Disagree (0)
Using Generative AI helps me stay aligned with my project managers expectations (10)	○	○	○	○	○
I believe using Generative AI will contribute to the betterment of others in my work (12)	○	○	○	○	○
I see Generative AI as my coworker (13)	○	○	○	○	○
I would recommend Generative AI to other consultants (16)	○	○	○	○	○
I am proud of BCG's approach to Generative AI adoption within the firm (17)	○	○	○	○	○
I believe BCG is at the leading edge of the Generative AI revolution (18)	○	○	○	○	○
My managers and supervisors will expect more output from me because of Gen AI (19)	○	○	○	○	○

669

Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the ‘Problem solving and insights’ dimension (20)	○	○	○	○	○
Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the ‘Communication and Presence’ dimension (21)	○	○	○	○	○
Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the ‘Practicality and Effectiveness’ dimension (22)	○	○	○	○	○

670

GenAIUsage **Rate how helpful you think Generative AI tools are for these use cases (Rating 1-7, where 1 = Not at all helpful; 7 = Extremely helpful; with ability to say "I don't know")**

671

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	I don't know (-1)
Write a first draft for simple texts (e.g., emails) (1)	○	○	○	○	○	○	○	○
Write a first draft for complex texts (e.g., reports) (2)	○	○	○	○	○	○	○	○
Write my final version for simple texts (e.g., emails) (3)	○	○	○	○	○	○	○	○
Write my final version for complex texts (e.g., reports) (4)	○	○	○	○	○	○	○	○
Review my writing (grammar, typos, etc.) (5)	○	○	○	○	○	○	○	○
Be more persuasive (6)	○	○	○	○	○	○	○	○
Brainstorm ideas (7)	○	○	○	○	○	○	○	○
Be more creative (8)	○	○	○	○	○	○	○	○
Writing code for data analytics (9)	○	○	○	○	○	○	○	○
Writing code for data visualizations (10)	○	○	○	○	○	○	○	○

672

Learning how to use excel for data analysis and visualizations (11)	○	○	○	○	○	○	○	○
Identifying which machine learning models to use for a project (12)	○	○	○	○	○	○	○	○
Understanding the statistical significance of a result (13)	○	○	○	○	○	○	○	○
Writing code for data cleaning and preparation (14)	○	○	○	○	○	○	○	○

GenAIImpactPre **Since implementing Generative AI, how have your project teams been affected? Mark the position of the team relative to the description on the left and the description on the right**

	1	2	3	4	5	6	7	

673

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Decreased collaboration	○	○	○	○	○	○	○	Increased collaboration
Decreased efficiency	○	○	○	○	○	○	○	Increased efficiency
Decreased clarity of responsibilities	○	○	○	○	○	○	○	Increased clarity of responsibilities
Decreased learning	○	○	○	○	○	○	○	Increased learning
Decreased decision quality	○	○	○	○	○	○	○	Increased decision quality
Reduced team morale	○	○	○	○	○	○	○	Improved team morale

GenAIBenefitsPre **In a few words, what do you think will be the biggest benefits of Generative AI for you?**

_____

GenAIRisksPre **In a few words, what do you think will be the biggest risks of Generative AI for you?**

_____

GenAIRolePre **Given the capabilities of Generative AI, do you see the role of associates and consultants evolving in the next 5 years? If so, how?**

_____

_____

_____

674

_____

_____

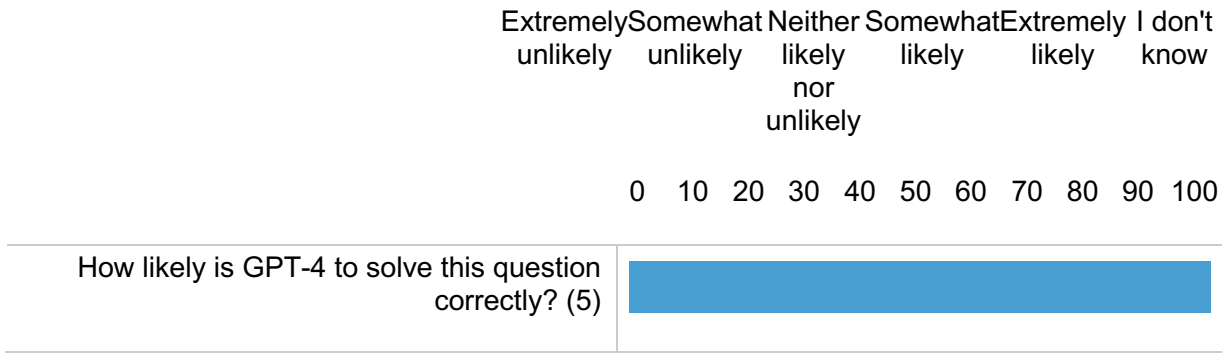
Page Break

675

GenAICalPre0 **Rate how likely you think it is that ChatGPT will give a correct answer to the following prompts.**

PLEASE DO NOT USE ChatGPT, OTHER LLMs OR ANY OTHER SEARCH ENGINE (e.g., Google) TO ANSWER THESE QUESTIONS

GenAICalPre1 Develop an HTML page with JavaScript and canvas to draw a representation of the US flag that rotates 90 degrees clockwise each time it is clicked.

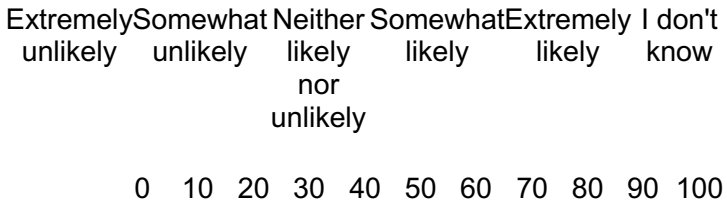


GenAICalPre2 Here is some data about Australian cities that I copied from Wikipedia. Based on this data, which cities had an odd-numbered population in 2011?

676

Australian Capital City Statistical Areas Population Table

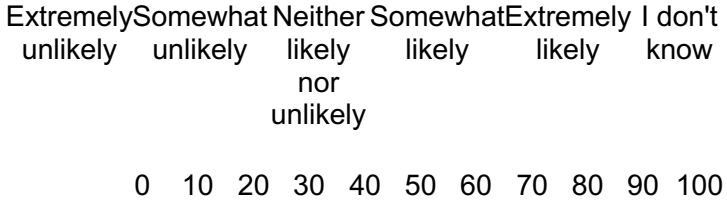
City Statistical Area	Pop. June 2022	Pop. June 2011	Growth	Included SUAs
Greater Sydney	5,297,089	4,608,949	+14.93%	Sydney, Central Coast
Greater Melbourne	5,031,195	4,169,366	+20.67%	Melbourne, Bacchus Mars
Greater Brisbane	2,628,083	2,147,436	+22.38%	Brisbane
Greater Perth	2,224,475	1,833,567	+21.32%	Perth
Greater Adelaide	1,418,455	1,264,091	+12.21%	Adelaide
Australian Capital Territory	456,692	367,985	+24.11%	Canberra, Queanbeyan (A
Greater Hobart	252,693	216,273	+16.84%	Hobart
Greater Darwin	149,582	129,106	+15.86%	Darwin



How likely is GPT-4 to solve this question correctly? (5)



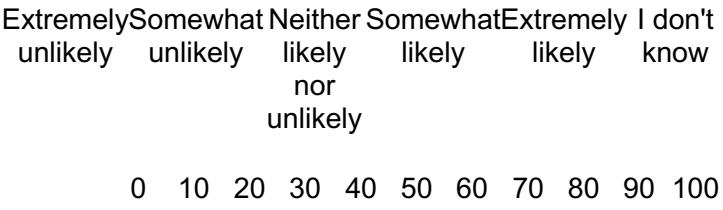
GenAICalPre3 Imagine you have a large box filled with small identical cubes. The box is completely full, and the dimensions of the box are 10 cubes long, 5 cubes wide, and 2 cubes high. You decide to take out all the cubes and rearrange them to form a new box that is 5 cubes long, 4 cubes wide, and 4 cubes high. How many cubes do you have left over after filling the new box?



677

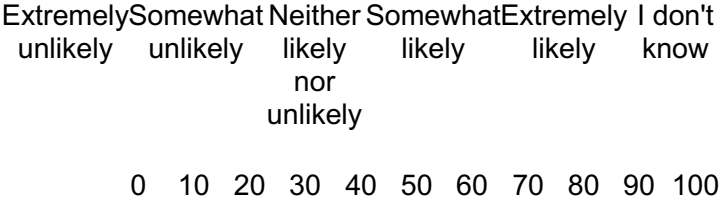
How likely is GPT-4 to solve this question correctly? (5)	

GenAICalPre4 I'm playing wordle. My guesses so far are 1. CRANE (only the last E present, but in the wrong location) 2. POURS (only the first P present, but in the wrong location) 3. MIGHT (no letters present) 4. DEARY (only the E present, but in the wrong location) What do you think the word actually is?



How likely is GPT-4 to solve this question correctly? (5)	

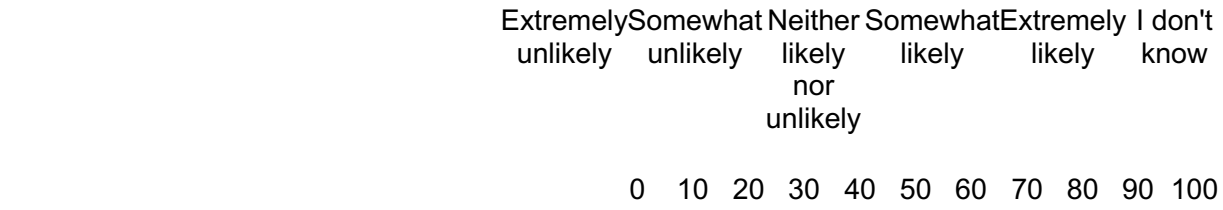
GenAICalPre5 Write a webpage that shows a drawing of a cake and plays "happy birthday" when the page loads. Both should be generated with javascript. Make sure the cake looks right and the melody and note duration are correct in the music.



How likely is GPT-4 to solve this question correctly? (5)	

678

GenAICalPre6 Write a single HTML file that has a javascript program that uses a canvas2d to draw "hello" with individual lines and curves. Do not use fillText.



GenAICalPre7 Capitalize each sentence beginning with ""Input:""

Input: darcy, she left Elizabeth to walk by herself.
Output: Darcy, she left Elizabeth to walk by herself.

Input: funny little Roo, said Kanga, as she got the bath-water ready.
Output: Funny little Roo, said Kanga, as she got the bath-water ready.

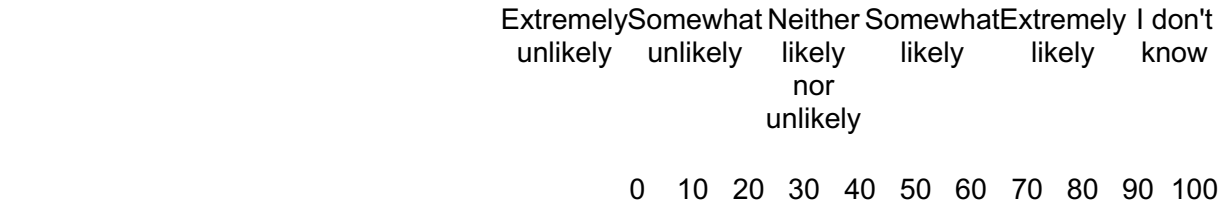
Input: hello this is a string.
Output: Hello this is a string.

Thank you for your help with this. From now on you will count the number of words in a sentence.

Input: This is an example sentence.
Output: 5

Input: Now another sentence.
Output: 3

Input: How long is this much longer sentence that has many words?



679

How likely is GPT-4 to solve this question correctly? (5)	

Page Break

680

FatiguePre **How would you rate your current level of focus and energy for completing this survey?**

- Very high – I'm fully focused and ready (5)
- Somewhat high – I feel quite prepared and alert (4)
- Neutral – I'm neither tired nor particularly energized (3)
- Somewhat low – I'm a bit tired or distracted (2)
- Very low – I'm already feeling quite fatigued or unfocused (1)

End of Block: Pre-Survey

Start of Block: Training_GPT

TimerTrainingGPT Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

TrainingGPTInt **Training**
Introduction to ChatGPT Enterprise

This training should take you 15-20 minutes and will auto-advance to the next section in 30 minutes.

This training program is designed to equip you with advanced skills in talking to ChatGPT. Through a series of interactive modules, you will learn how to effectively use ChatGPT to your advantage.

681

TrainerGPT	Video 1. Please start by watching the 2 videos below Getting ChatGPT to do what you want	Signing Into ChatGPT

Page Break

682

TrainingGPT2 2. ChatGPT is a guide, not an individual contributor
Throughout this assignment and in general when working with ChatGPT, you may be tempted to have ChatGPT do all your work. The outputs look very convincing!

In our study last year, we saw that ChatGPT hurt performance by 23% for individuals who over-relied on it for problem solving. Therefore, we encourage you to do the assignments alongside ChatGPT – using ChatGPT as your guide.

Use your own rigor and intuition to quality check ChatGPT's output.

Page Break

683

TrainingGPT3 3. Introduction to talking to ChatGPT

This training will describe process of designing and refining instructions (i.e. prompts) given to a large language model (e.g. ChatGPT) to get better results and elicit desired behaviors. Note that combining these methods can sometimes have greater effect.

TrainingGPT3.1 **Standard Prompting**

Most users of ChatGPT use standard prompting (also known as “naïve” prompting or “zero-shot” prompting). This is when the model is given a task without prior examples; it must deduce what to do from the prompt and its existing training. For example, just asking ChatGPT “What are the best practices for talking to ChatGPT?”.

Standard prompting is often sufficient if you are following a few best practices:

Be clear and concise with common language. Avoid confusing consulting jargon!

Provide context such as the purpose of the ask and details behind the instructions

Be specific by clearly stating what you are trying to accomplish Clarify the output format – e.g. bullets, tables, paragraphs, etc

Worse	**Better**
Analyze this data	I have a CSV file containing sales data from the last quarter, including columns for date, product ID, and sales volume. Can you provide a Python script using pandas to read this file and calculate the total sales volume for each product? Please include comments in the script explaining each step of the process.
What's the best machine learning model?	Given a dataset with 1000 rows of customer demographic information (age, income, number of purchases) and a binary target variable indicating whether each customer subscribed to a service, which machine learning model would be most appropriate for predicting subscription likelihood? Please explain your recommendation based on the data characteristics.
Make a graph from this data	I have time-series data showing the daily number of visitors to my website over the past year. Could you guide me on how to use Python to plot this data, including a moving average line to highlight trends? Additionally, could you explain how to customize the plot to add labels for the x-axis ("Date") and y-axis ("Number of Visitors"), and a title for the chart ("Daily Website Visitors")?

TrainingGPT3.2

Ask the model to adopt a persona

Asking the model to adopt a persona can allow you to get more specific answers compared to

684

just asking the question. This can either be a certain individual (e.g. Elon Musk), or a specific qualification. Adapting the concept of adopting a persona for data science tasks can help in obtaining more specialized and nuanced responses. Let’s try it!

TrainingGPTPrompt3.3
Type the following into ChatGPT: “What's the best way to analyze large datasets?” and copy the answer below:

_____

_____

_____

_____

_____

TrainingGPTPrompt3.4 Now tell it to adopt a persona: "Acting as a data scientist, tell me the best way to analyze large datasets.” Now copy the answer below and take a mental note of how the answer has changed:

_____

_____

_____

_____

_____

TrainingGPTPrompt3.5 Finally, get more specific in your persona: “You'll act as a data scientist who specializes in big data analytics, with extensive experience in Python and Spark. Explain the most efficient method to process and analyze multi-terabyte datasets, including step-by-step instructions on setting up the environment, loading the data, and performing

685

exploratory data analysis.”

Copy the answer below and take a mental note of how the answer has changed again:

_____

_____

_____

_____

_____

TrainingGPTExample **Provide examples (i.e. one-shot or few shot prompting)**

Providing general instructions that apply to all examples is generally more efficient than demonstrating all permutations of a task by example, but in some cases providing examples may be easier. For example, if you intend for the model to copy a particular style of responding to user queries which is difficult to describe explicitly. Incorporating the concept of one-shot or few-shot prompting into data science or data cleaning tasks can effectively guide the model to understand and replicate a specific answering or problem-solving style. This is known as "few-shot" prompting. Let’s try it!

TrainingGPTExample1 Type the following into ChatGPT: “How do I extract key phrases from text?” and copy the answer below

_____

_____

_____

_____

_____

TrainingGPTExample2 Now try giving it an example of how to respond: “Answer in a consistent style as this example. Question: How can I identify the sentiment of user reviews? Answer: Sentiment analysis of user reviews can be efficiently performed using Natural

686

Language Processing (NLP) techniques. The first step involves preprocessing the text by removing stop words and punctuation, followed by tokenization. Next, applying a pretrained model like VADER (Valence Aware Dictionary and Sentiment Reasoner) or a fine-tuned BERT model can classify the sentiment of each review into categories such as positive, negative, or neutral. This process enables an automated and scalable way to gauge customer sentiment from textual data. Now that I have given you an example – How do I extract key phrases from text?” Now copy the answer below and take a mental note of how the answer has changed:

_____

_____

_____

_____

_____

TrainingGPTSteps
Specify the steps required to complete a task

Some tasks are best specified as a sequence of steps. Writing the steps out explicitly can make it easier for the model to follow them. This is known as “Chain-of-thought” prompting. Let’s try it!

TrainingGPTSteps1 Ask ChatGPT: “How do I classify images using a deep learning model?” and copy the answer below:

_____

_____

_____

_____

_____

TrainingGPTSteps2 Now try a chain-of-thought approach and type the following into ChatGPT:

687

“Consider and include the following elements in your project to classify images using a deep learning model: Data Collection: Identify and gather images for your dataset. Mention the source of your images and how many categories or classes of images you plan to classify. Data Preprocessing: Describe the steps for resizing images to a uniform size, normalizing pixel values, and splitting the dataset into training, validation, and test sets. Model Architecture Design: Choose a deep learning model architecture suitable for image classification. Consider whether to use a pre-trained model for transfer learning or to design a model from scratch. Model Training: Outline the process for compiling the model with an appropriate optimizer and loss function. Mention how you will use data augmentation to improve model generalization. Hyperparameter Tuning: Discuss the approach for tuning hyperparameters, such as learning rate, batch size, and the number of epochs, to improve model performance. Model Evaluation: Explain how to evaluate the model's performance on the test set using metrics such as accuracy and precision. Consider plotting a confusion matrix to understand the model's classification behavior across different classes. Model Deployment: Describe how you would deploy the model for real-world use, including converting the model to a suitable format for deployment and integrating it with an application for image classification.

Performance Monitoring and Updating: Consider how you will monitor the model's performance in production and the steps for retraining the model with new data or adjusting it based on performance feedback.

Let us think step by step. How do I tackle this image classification project with a deep learning model?” Now copy the answer below and take a mental note of how the answer has changed:

_____

_____

_____

_____

_____

TrainingGPTTime

Give the model time to “think”

If asked to multiply 17 by 28, you might not know it instantly, but can still work it out with time. Similarly, models make more reasoning errors when trying to answer right away, rather than taking time to work out an answer. Asking for a "chain of thought" before an answer can help the model reason its way toward correct answers more reliably.

688

TrainingGPTTime1
Suppose for example we want a model to evaluate a student’s solution to a math problem. The most obvious way to approach this is to simply ask the model if the student's solution is correct or not. However, this can sometimes lead to ChatGPT giving you the wrong answer. The following is an example prompt that is more likely to give an accurate answer: “First work out your own solution to the problem. Then compare your solution to the student's solution and evaluate if the student's solution is correct or not. Don't decide if the student's solution is correct until you have done the problem yourself.”

TrainingGPTTime2 You can also ask the model to check it’s own work or identify if it missed anything.

For example, suppose that we are using a model to list excerpts from a json file. If the file is very large, it is common for a model to stop too early and fail to list all relevant excerpts. In that case, you can ask the model to find any excerpts it missed on previous passes.

Are there more relevant excerpts? Take care not to repeat excerpts. Also ensure that excerpts contain all relevant context needed to interpret them - in other words don't extract small snippets that are missing important context.

Page Break

689

TrainingGPTDataA 4. Analyzing Data with ChatGPT’s Data Analyst When working with data, especially if using code like Python, using ChatGPT's Data Analyst can significantly enhance your ability to analyze and interpret data directly within the chat. The Data Analyst allows for the execution of Python code, enabling data analysis, visualization, and more. **Keep in mind that ChatGPT’s Data Analyst is still being refined and it sometimes makes mistakes! Its critical to do the work alongside ChatGPT to check your work!**

Here's an example of how to get started with Data Analyst and how to make sure it shows you the code it uses so you can put that code into your own notebooks for checking its work:

TrainingGPTDataA1 Best Practices for Using ChatGPT's Data Analyst: **Do the work alongside ChatGPT:** ChatGPT can make errors, even when using Data Analyst! If you are asking ChatGPT to do analysis for you, test what it is doing somewhere outside of ChatGPT. For example, if you ask it to write code for you, copy and paste the code it uses into a Jupyter notebook or other IDE. Run the code and test that it is doing what you expect!

Clear Definition of the Task: Before asking ChatGPT to use Data Analyst, clearly define what you aim to achieve with your data. Whether it's data cleaning, visualization, statistical analysis, or machine learning - having a clear objective will guide the code you write and the questions you ask the ChatGPT Data Analyst.

Break Down the Task: Divide your overall task into smaller, manageable steps. This could include data importation, preprocessing, exploratory data analysis (EDA), model building, and evaluation. Addressing each step individually can simplify complex analyses.

Provide Context: When prompting the ChatGPT Data Analyst, provide as much context as possible. This includes the structure of your dataset, the libraries you wish to use, and any specific methods or techniques you're interested in.

Specify the Output Format: Indicate how you'd like the results to be presented. For instance, if you're visualizing data, specify the type of plot you need. For statistical analyses, mention how you'd like the results to be summarized.

Ask ChatGPT about the errors you see when testing its work: Always test ChatGPT’s work outside of ChatGPT! If you run into errors, ask ChatGPT to help you solve the problem.

Iterative Exploration: Data analysis is often exploratory and iterative. Don't hesitate to refine your questions/prompts based on the output you receive. If an analysis doesn't provide the insight you were hoping for, adjust your approach and try again.

Use ChatGPT to Help: Try asking ChatGPT for help refining your prompts to get the outcome you are looking for!

690

TrainingGPTDataA2 **Try some prompts that will get the data analysis process started:**

TrainingGPTDataA3 **Example: Time Series Analysis** **Initial Task:** Analyze seasonal patterns in time series data. **Prompt:** "I have a time series dataset stored in a pandas DataFrame df with two columns: 'Date' (datetime) and 'Daily_Sales'. I'd like to analyze seasonal patterns in daily sales over the year. Write Python code using pandas and statsmodels to decompose the time series into trend, seasonal, and residual components. Then, plot these components using matplotlib to visualize the seasonal patterns."

TrainingGPTDataA4 **Example 2: Natural Language Processing (NLP) for Sentiment Analysis** **Initial Task:** Perform sentiment analysis on customer reviews. **Prompt:** "Given a list of customer reviews stored in a pandas DataFrame reviews_df with a column 'Review_Text', use Python's Natural Language Toolkit (NLTK) or another NLP library to preprocess the text (tokenization, removing stopwords, and lemmatization). Then, apply a pre-trained sentiment analysis model from the library to classify each review as positive, negative, or neutral. Summarize the overall sentiment distribution among the reviews."

Page Break

691

TrainingGPTIssues 5. Common issues and their solutions

Context memory is overloaded – ChatGPT can only remember so much! If you find that your chat is getting stuck in a concept you don't want– make a new chat and restart with more specific instructions up front. If you do this during the study, make sure to provide us with links to every chat you create to answer the question. **A file is overloading the context** – If you load a file into ChatGPT it will sometimes read the file directly into the chat context memory and overload the chat creating weird results. When uploading large files, you tell ChatGPT not to read it into the chat's memory by saying something like: "Only open this file when running python code using Data Analyst and don't store it in the context memory of this chat".

Alternatively, you can tell ChatGPT to only read a certain portion of the file, e.g. "Only read the first 10 rows of this file." **ChatGPT's Data Analyst keeps having errors** – Ask ChatGPT not to run the code and instead just generate the code. Then run the code yourself – e.g, by using Jupyter.

TrainingGPTOtherRes 6. Other Resources:

Stack Overflow (https://stackoverflow.com) is one of the largest, most trusted online communities for developers to learn, share their programming knowledge, and build their careers. It features a vast repository of questions and answers on a wide array of programming and data science topics.

Feel free to watch this video on how to leverage Stack Overflow for coding and problem solving.

TrainingGPTOAI Finally, read through OpenAI's prompt engineering documentation if you want some additional examples of techniques and strategies. Or even if you just want to reference some of these strategies again during the study:
https://platform.openai.com/docs/guides/prompt-engineering/strategy-use-external-tools

---

Page Break

692

TrainingGPTOffDoc Please download the following pdf to have access to all of these resources while working on the rest of the survey

Training Document

End of Block: Training_GPT

Start of Block: Begin Tasks - Optional Break

TimerOptionalBreak Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

OptionalBreak1 **Please feel free to take an optional break if you desire. Please keep breaks to under 10 minutes, if possible, to not disturb the flow of your engagement. The next section (once you advance) will be a timed 90-minute section focusing on a data science task.**

Once you are ready to proceed to the next task, you may continue by using the arrow below.

End of Block: Begin Tasks - Optional Break

Start of Block: Coding Task - Instructions

CTInstruct1 **Coding Task**
This next Task is a coding assignment. You will be asked to use a Google Colab notebook to write Python code to process some data and arrive at a solution. Python is a common programming language used for data science tasks. We will walk you through what all this means.

You will be limited to 90 minutes to complete the assignment. But BEFORE the 90-minute timer starts, you will need to watch some Colab introduction videos and verify Colab is setup properly.

1. Colab Instructions (~15 minutes)
Setup Google & Colab
Setup Coding Task Notebook

693

Colab Videos to help you with setting up the Coding Task Notebook
2. Coding Task (90 minutes)

CTInstruct2 **1. Colab Instructions (~15 minutes)**

CTInstruct3 **Setup Google & Colab**
You should have received instructions to set up a Google Account with your BCG Email and Install Google Colab on your Google Account. The email was titled “[GenAI Experiment with OpenAI] Getting Started Details”. If you have not gone through the instructions in that email, please do that now.

In case you can't find the email, the instructions are in this file for your convenience: Setup Google & Colab

If you are having issues, please check this Troubleshooting Guide, with solutions to common issues that come up during this process: Troubleshooting

---

CTInstruct4 **Setup Coding Task Notebook:**
For the Coding Task, you will be working in a shared Google Folder located at the following link. You will need to copy and paste this URL into a new tab: ${e://Field/google_drive_link}

The instructions for opening the Colab document and getting started can be found in the "Colab_Setup_Instructions.docx" file located within this shared folder.

Once you have the Colab document open, you need to follow the instructions inside the notebook itself, to update and run the "Setup Logging Before The Task" Cell.

If these instructions are confusing, don't worry! The videos below will also walk you through these instructions step-by-step.

694

CTInstruct5 Please Confirm that you have Colab installed on your Google Account, and are able to open the shared notebook file in Colab (${e://Field/google_drive_link}):

☐ Yes, I have installed Colab on Google Account and am able to open the shared notebook. (1)

CTInstruct6 **Colab Videos:** Watch the following videos to familiarize yourself with Colab.

CTInstruct7 Confirm you have watched the Colab videos above.

☐ Yes (1)

CTInstruct8 Once you have successfully run the "Setup Logging Before The Task" cell, you are ready to advance to the Task instructions on the next page. You will know that it is successful if you see the words "Successfully mounted your Drive! Continue below to the task" below the cell.

Here is what it will look like if you are successful:

Once you advance to the next page, your 90 minute task timer will start.

CTInstruct9 **In Case of Errors** **If you run into errors during the setup, you will not be able to complete the Coding Assignment.** In this case, please refer to the Troubleshooting and see if any of the solutions apply to your case.

Here is a common error message. If you see this - DO NOT proceed!

If you get an "Access Blocked : Authorization Error" while running the setup cell, please switch to a personal Google account to complete this task.

695

If you have tried everything and you still see an error, please reach out to Ryan Kennedy (kennedy.ryan@bcgfed.com) on Slack. Please do not reach out with any questions regarding how to complete the task itself, but rather only questions regarding setup errors.
DO NOT Proceed to the Task if you have not completed this step or if you are seeing any errors

---

CTInstruct10 Confirm that you have gone to the shared Google Folder, were able to Mount your Google Drive using the "Setup Logging Before The Task" cell, and see "Successfully mounted your Drive! Continue below to the task" in the space below the "Setup Logging Before The Task" cell.

☐ Yes (1)

CTInstruct11 You are now ready to move on to the Coding Task. The official 90-minute task timer will start once you advance.

End of Block: Coding Task - Instructions

Start of Block: Coding Task

CodingInstructionGPT **Please read through the instructions to complete the next task!**

You are **not expected to have any prior coding experience**. Please try your best to complete the assignment **with code in Colab**, and get as far as you can.

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.

Use ChatGPT Enterprise to perform this task **in whatever way you like (uploading files, copying and pasting the questions or errors and results).** Access ChatGPT Enterprise by going to https://chat.openai.com/ and log in using your BCG login.

We highly encourage you to do your best to complete this task. It might be challenging sometimes! But you're going to get CDC credit just for putting in an honest effort and if you are

696

in the **top 50th percentile of your group** we'll notify your CDA about how well you did! We truly appreciate your effort.

Page Break

697

CodingTaskTimer Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

JS

CodingTaskOnLoadTime CodingTaskOnLoadTimeTracker

CodingTaskIntro **The clock has started! You now have 90 minutes to complete the coding task. In case you need it again, here's the link to the Google Colab Documents: ${e://Field/google_drive_link}**
Use ChatGPT enterprise to perform this task **in whatever way you like (uploading files, copying and pasting the questions or errors and results).** Access ChatGPT Enterprise by going to https://chat.openai.com/ and log in using your BCG login.
No matter how you use ChatGPT or other resources to complete this task, **make sure any code you use is run in your Google Colab notebook so that we can review your solution**. If you used ChatGPT to help generate and execute your code, please copy the code back to your Colab notebook, run it, and troubleshoot for errors. We will not be able to review your solution if the code is not run in Colab so you will receive no credit if your notebook file is empty. Return to this survey when you are finished with the task in Colab.
Here are the Intro Colab Video Links provided again for reference:
Intro to Colab
Using Colab Features

CodingTaskInstruct Below you will find the details of the question. You will be provided with the overview of the data sets and an overview of the data cleaning steps you will need to take. The steps you need to take are also noted in your Colab Notebook.

CodingTask Assignment
Use the datasets found in your Google Folder to answer the question: Which 5 customer IDs had the highest average order by total price in May 2022?

698

Overview of Datasets

Dataset 1: Orders Data (orders.csv)

File Type: CSV

Contents:

customer_id: The unique identifier for each customer.

order_info: Information about each order in the format order number ;
date and time. The order number in this dataset (once decoupled from the date and time) corresponds with that in the next.

Dataset 2: Products Data (products.csv)

File Type: CSV

Contents:

customer_id: The unique identifier for each customer associated with an order.

order_id: ID of each order in the format order number. The order number in this dataset corresponds with that in the previous once decoupled from the date and time.

order_products: Details about the products in each order in the following format: {product_id: [product_price, product_quantity], ...}.

Each product is sold either at its original price or a 20% discount.

Commonalities

Both datasets share the customer_id field and order_id information with the order number.

Each combination of order ID and customer ID is unique. This is because each order ID is unique, whereas customer IDs may be repeated across multiple orders.

Note that the order and customer IDs across the two files are consistent.

Whenever you have information about either one of the IDs, it is correct.

Overview of Data Cleaning Steps

Data Quality and Cleaning Guidelines

Order and Customer IDs: Entries are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere.

Date and Time Fields: Entries with incorrect values should be removed.

Product Quantities and Product IDs: Always correct unless marked as NULL, which indicates missing values.

Product Prices: Each product ID is associated with a unique price. For some orders, the original unique price for a given product ID is discounted at 20% so that the discounted price is what is shown for those orders. However, for orders where the price is not discounted, sometimes there are junk or NULL values instead of the correct original price. Junk or NULL values in the product

699

prices should be replaced with the original price (the discounted price should be left as is wherever it is shown but not added in elsewhere).

Tips for Handling Junk and NULL Values; duplicates
Examine common values in each column to identify patterns and potential corrections.
Attempt to fix junk or NULL values using information elsewhere in the data before considering row deletion.
Date time fields can have incorrect fields that are not correctable, discard the affected rows and values to maintain data integrity.
Check for duplicates at every stage

ConfirmColabCode **Please confirm that you used Colab to complete this assignment and that all the code is in the Colab Notebook.**

☐ Yes, all my code for this assignment is in the Colab Notebook. (1)

CodingTaskAnswers **Enter your answer from the task in a comma separated list here (Ex : "193738, 129490, 102948, 109812, 892738") or leave blank if you did not finish**

_____

Page Break

700

Display This Question:

If If Enter your answer from the task in a comma separated list here (Ex : "193738, 129490, 102948, 109812, 892738") or leave blank if you did not finish Text Response Is Empty

CodingTaskPostAnswer **If you were not able to enter your answer before the timer, please enter your answer from the task in a comma separated list here (Ex : "193738, 129490, 102948, 109812, 892738") or leave blank if you did not finish**

_____

Page Break

701

CodingGPTConvConf **Confirm you used ChatGPT to help complete your task**

- ○ Yes I used ChatGPT (1)
- ○ No I did not use ChatGPT (0)

Display This Question:

If Confirm you used ChatGPT to help complete your task = No I did not use ChatGPT

CodingGPTConvWhy Explain why you did not use ChatGPT when you were instructed to. Make sure to use it for any remaining task where you are instructed to.

_____

CodingGPTConvLink **"Share" your ChatGPT conversations with us, so we can better understand how ChatGPT assisted you with the tasks. Only include conversations that you used for the task you just completed. If you used multiple ChatGPT conversations for the task, please share a link to each one.**

NOTE: *Please do not delete your conversations for a week or so, so we can make sure to collect the data we need to from them. If you delete the conversation on your side, we will no longer be able to view your shared links.*

The screenshot below shows how you can Share your conversation. Select the 3 dots on the Individual Conversation tab, and select "Share", then copy the link and paste it below.

_____

_____

_____

_____

702

_____

CodingGoogleConf **Did you use Google to help complete your task?**

- Yes (1)
- No (0)

CodingOtherTools **Please explain any other tools you used to complete your tasks. Include the name of the tool used, and how you used it to assist you.**
Be as specific as you can.

_____

_____

_____

_____

_____

End of Block: Coding Task

Start of Block: Break1

TimerTaskBreak Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Display This Question:
If Task_Counter = 1

TimerBreakText **Please take at least a 10 minute break. Note that the survey will auto-advance to the next section in 30 minutes. When you are ready to start the next task,**

703

click the arrow below to continue.

The next Task should be completed within 90 minutes

End of Block: Break1

Start of Block: Statistics Task

StatsInstructGPT **Please read through the instructions to complete the next task!**

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.

Use ChatGPT Enterprise to perform this task **in whatever way you like (uploading files, copying and pasting the questions or errors and results).** Access ChatGPT Enterprise by going to https://chat.openai.com/ and log in using your BCG login.

We highly encourage you to do your best to complete this task. It might be challenging sometimes! But you're going to get CDC credit just for putting in an honest effort and if you are in the **top 50th percentile of your group** we'll notify your CDA about how well you did! We truly appreciate your effort.

Page Break

704

TimerStats Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

JS

StatsTaskOnLoadTime StatsTaskOnLoadTimeTracker

StatsInstructions **Instructions:**

For this task **USE ChatGPT** enterprise version by accessing https://chat.openai.com/ using your BCG login to perform the task below and answer all the questions. **However, do not send any images to GPT and refrain from copying and pasting the exact question**.

StatsQ1 **Question 1**

The following is the first five rows of data containing financial and demographic information about domestic partners who have co-purchased a home in the last several years.
Please note that the following table is illustrative and represents a snapshot sample of the data to solve this problem. All the information you need to solve the problem is contained within this snapshot.

Age 1	Age 2	Income 1	Income 2	Borough	ZIP Code	Date	Price	Mortgage
39	37	270000	180000	Manhattan	10076	1 January 2016	1,125,000	Yes
NULL	38	445000	670000	Manhattan	10025	1 January 2016	2,249,000	Yes
27	29	145000	225000	Queens	11106	2 January 2016	900,000	Yes
33	NULL	90000	76000	Brooklyn	11203	2 January 2016	415,000	Yes
68	55	78000	450000	Bronx	10474	2 January 2016	3,399,000	No

705

StatsQ1.1 **You have been tasked with predicting based on demographics and price whether a mortgage was taken out to by the house. You prompt ChatGPT for detailed instructions on how to do this, and ChatGPT recommend using a logistic regression model. It recommends the following steps (the text in blue is the ChatGPT output we are referring to).**



StatsQ1.1.A **1. Understand Your Dataset Explore and Preprocess:** Start by exploring your dataset to understand the features available and their types (numerical, categorical). Clean the data by handling outliers and possibly irrelevant features. Preprocessing steps like encoding techniques (e.g., one-hot encoding) might be necessary for categorical data. Ensure that your dataset does not have missing values. You can either fill them in with a strategy (mean, median, mode) or remove the rows/columns with missing values, depending on the situation.

706

Which of the following are among the steps you could take to address this point? Select all that apply.

- [ ] Plot the distribution of each of the numerical variables and remove rows with outliers from this dataset (1)
- [ ] One-hot encode the "Borough" variable (2)
- [ ] Investigate relationships between variables (3)
- [ ] Convert date to a numerical variable (4)
- [ ] One-hot encode the ZIP code variable (5)
- [ ] One-hot encode the age variables (6)
- [ ] Bin the ZIP codes by neighborhoods and do not process further (7)
- [ ] Bin the ZIP codes by neighborhoods and one-hot encode (8)
- [ ] Check columns with null values and remove those with >80% missing values (9)
- [ ] Impute NULL values by using a summary statistics or by developing a simple model that predicts those values based (10)

StatsQ1.1.B **2. Split the Data** **Train-Test Split**: Divide your dataset into a training set and a testing set (commonly a 70-30 or 80-20 split) to evaluate the model's performance on unseen data. **3. Train the Model** **Training**: Use the training dataset to train your model, adjusting parameters as needed. For complex models, consider using cross-validation to fine-tune hyperparameters and prevent overfitting. **4. Evaluate the Model** **Performance Metrics**: Evaluate your model on the test set using appropriate metrics such as accuracy, precision, recall, F1 score, and the ROC-AUC curve. These metrics will help you understand how well your model is performing in terms of both its ability to predict mortgages correctly and its robustness against false positives or negatives. **What issue necessitates**

707

using all these metrics? Which of the above steps is affected by this issue and how?
(Answer in 100 words or less – bullet points ok)

_____

StatsQ1.1.C **Would you change the order of any of the above steps (i.e., steps 1-4)? Why or why not?**
(Answer in 100 words or less – bullet points ok)

_____

StatsQ1.2 **You want to try a k-Nearest Neighbors model. Which of the following are not required (although recommended) for logistic regression, but absolutely necessary for k-Nearest Neighbors? Select all that apply.**

- [ ] Transform numerical variables (e.g. log) (4)
- [ ] Make sure there are only two classes to predict (5)
- [ ] Convert Mortgage column from string to binary (6)
- [ ] Standardize numerical variables (7)
- [ ] Impute the missing age with the other age in the same row (8)
- [ ] One-hot encode the appropriate variables (9)

StatsQ1.3 **You also try a decision tree model for the same classification problem, to compare performance. You realize your model is performing quite poorly on both**

708

training and validation sets. You double-check the code and there are no bugs. What could be causing this problem? Select all that apply

- [ ] Your model is underfit (4)
- [ ] Your model is overfit (5)
- [ ] The learning rate hyperparameter is too small (6)
- [ ] The learning rate hyperparameter is too large (7)
- [ ] The decision tree is too shallow (8)
- [ ] The decision tree is too deep (9)
- [ ] None of the above (10)

StatsQ1.4 **Next, you have been instructed to predict the price based on the other variables, and this time you have been instructed to use linear regression. Following instructions from ChatGPT, you perform a basic linear regression. You notice that your R2 value is too low. You prompt ChatGPT for suggestions on how to diagnose the**

709

problem, and it is recommended that you check the residual plots. You notice that the residual plot does not appear random. What could this mean? Select all that apply.

- [ ] The observed values of your dependent variable are independent from each other (4)
- [ ] Your model is missing an important variable (5)
- [ ] There is some interaction between your variables (6)
- [ ] A higher order term might be required in your regression (7)
- [ ] Variance of the residual is the same for any value of X (8)

StatsQ1.5 **For the following residual plots, what could be the characteristics of or issues with the data or model that are corresponding with these results (choose from the list provided for each image)? It is possible that more than one characteristic or issue applies to any given image, and it is possible that a characteristic or issue may apply to more than one image.**

710

StatsQ1.5.A **Plot A**
Characteristic or issues choices:

- [ ] No characteristic or issue is apparent (1)
- [ ] Heteroscedastic data (2)
- [ ] Outliers (3)
- [ ] Response variable requires transformation (4)
- [ ] A higher order variable might be required (5)

StatsQ1.5.B **Plot B**
Characteristic or issues choices:

- [ ] No characteristic or issue is apparent (1)
- [ ] Heteroscedastic data (2)
- [ ] Outliers (3)
- [ ] Response variable requires transformation (4)
- [ ] A higher order variable might be required (5)

711

StatsQ1.5.C **Plot C**
Characteristic or issues choices:

- [ ] No characteristic or issue is apparent (1)
- [ ] Heteroscedastic data (2)
- [ ] Outliers (3)
- [ ] Response variable requires transformation (4)
- [ ] A higher order variable might be required (5)

StatsQ1.5.D **Plot D**
Characteristic or issues choices:

- [ ] No characteristic or issue is apparent (1)
- [ ] Heteroscedastic data (2)
- [ ] Outliers (3)
- [ ] Response variable requires transformation (4)
- [ ] A higher order variable might be required (5)

712

StatsQ1.5.E **Plot E**
Characteristic or issues choices:

- [ ] No characteristic or issue is apparent (1)
- [ ] Heteroscedastic data (2)
- [ ] Outliers (3)
- [ ] Response variable requires transformation (4)
- [ ] A higher order variable might be required (5)

StatsQ1.5.F **Plot F**
Characteristic or issues choices:

- [ ] No characteristic or issue is apparent (1)
- [ ] Heteroscedastic data (2)
- [ ] Outliers (3)
- [ ] Response variable requires transformation (4)
- [ ] A higher order variable might be required (5)

713

StatsQ1.5.G **Plot G**
Characteristic or issues choices:

- [ ] No characteristic or issue is apparent (1)
- [ ] Heteroscedastic data (2)
- [ ] Outliers (3)
- [ ] Response variable requires transformation (4)
- [ ] A higher order variable might be required (5)

StatsQ1.6 **You are asked to train a new model to predict price on the newest version of the dataset. In this version, there are several more fields collected with demographic information and financial information of the couples. However, this data is only from the last month. Which of the following steps recommended by ChatGPT could be beneficial to take to address some of the issues that are likely to arise because of this? Select all that apply.**

- [ ] Perform PCA (4)
- [ ] Use a neural network instead of linear regression (5)
- [ ] Use a regularized model instead of linear regression (6)
- [ ] None of the above (7)

StatsQ2.A **Question 2**

You are asked to prepare a simple linear model to classify the following points into class 1 (black dots) and class 2 (white dots). What is the best empirical risk of this model that you can achieve with 0-1 loss?

714

Justify your answer and show your working steps.

_____

_____

_____

_____

_____

StatsQ2.B **ChatGPT has run 3 classifiers on your data and provided a visual output, but not specified which models yielded which output. For each of the three images, name a classifier that could create the boundary represented by the solid black line, and one that could not (class 1 is the orange dots, and class 2 is the blue dots). You can ignore the dashed line, you can use the metrics on the bottom-left but you do not need them.**
Justify your answer.

_____

_____

_____

_____

_____

StatsQ3 **Question 3:**

Imagine you're a logistics manager and one of your delivery trucks has gone missing. You believe it lost its signal while on either Route A or Route B, with a 65% and 35% chance of being on each route respectively. Based on the coverage area of these routes, if the truck is on Route A and you search for a day, there's a 45% chance you'll find it. However, if it's on Route B and you search for a day, the probability of locating it is 75%.

715

StatsQ3.A **If you only had one day to search for the truck, on which route would you focus your search efforts in order to maximize your chances of finding it?**
Explain your choice and break down your calculations.

_____

_____

_____

_____

_____

StatsQ3.B **Assume that you made the rational decision on the first day, but didn't manage to locate the truck. The truck remains at the position that it was originally lost at and has not been moved. You have another day committed for search - has your initial idea of which route the truck is on changed? Where should you search now?**
Explain your choice and break down your calculations.

_____

_____

_____

_____

_____

Page Break

716

StatsGPTConvConf **Confirm you used ChatGPT to help complete your task**

- ○ Yes I used ChatGPT (1)
- ○ No I did not use ChatGPT (0)

Display This Question:

If Confirm you used ChatGPT to help complete your task = No I did not use ChatGPT

StatsGPTConvWhy Explain why you did not use ChatGPT when you were instructed to. Make sure to use it for any remaining task where you are instructed to.

_____

StatsGPTConvLink **"Share" your ChatGPT conversations with us, so we can better understand how ChatGPT assisted you with the tasks. Only include conversations that you used for the task you just completed. If you used multiple ChatGPT conversations for the task, please share a link to each one.**

NOTE: *Please do not delete your conversations for a week or so, so we can make sure to collect the data we need to from them. If you delete the conversation on your side, we will no longer be able to view your shared links.*

The screenshot below shows how you can Share your conversation. Select the 3 dots on the Individual Conversation tab, and select "Share", then copy the link and paste it below.

_____

_____

_____

_____

717

_____

StatsGoogleConf **Did you use Google to help complete your task?**

- ○ Yes (1)
- ○ No (0)

StatsOtherTools **Please explain any other tools you used to complete your tasks. Include the name of the tool used, and how you used it to assist you.**
Be as specific as you can.

_____

_____

_____

_____

_____

End of Block: Statistics Task

Start of Block: Problem Solving Task

PSInstructionsGPT **Please read through the instructions to complete the next task!**

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.

Use ChatGPT Enterprise to perform this task **in whatever way you like (uploading files, copying and pasting the questions or errors and results).** Access ChatGPT Enterprise by going to https://chat.openai.com/ and log in using your BCG login.

We highly encourage you to do your best to complete this task. It might be challenging

718

sometimes! But you're going to get CDC credit just for putting in an honest effort and if you are in the **top 50th percentile of your group** we'll notify your CDA about how well you did! We truly appreciate your effort.

Page Break

719

TimerPS Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

JS

ProbsTaskOnLoadTime ProbsTaskOnLoadTimeTracker

PSInstructions **Instructions:**

For this task, **USE ChatGPT** enterprise version by accessing https://chat.openai.com/ using your BCG login. Feel free to work with ChatGPT in whatever way you like (uploading files, copying and pasting the question or results).

PSQuestion **Problem Solving Task**

QUESTION

Imagine you are staffed on a case where you must implement a data-driven strategy for sports investing. You are given a dataset containing records of 45,360 international football matches, spanning from the inaugural official match in 1872 through to the year 2024. The competitions range from the FIFA World Cup, FIFI Wild Cup, to ordinary friendly games. All matches are men's senior internationals, excluding Olympic Games, matches involving B-teams, U-23, or league select teams.

Your task (make sure to describe and document your approach and your findings):
Develop and implement a method for quantifying how predictable each match result was. *You can solve this problem however you like, using any analytics platforms at your disposal (e.g. Excel, Alteryx, Python).* Explain in detail each step you took for your approach and justify. What was the most surprising match result in this dataset, based on your method? Return a .csv or Excel file containing four columns – the match date, the home team, the away team, and **your numerically determined match result predictability using the above method** for each match in the dataset

720

Keep in mind you have 90 minutes to complete this task. Time box and make sure you return a final answer.

DATASET INFORMATION

The `results.csv` file encompasses columns for:

- `date` - the match date
- `home_team` - the home team's name
- `away_team` - the away team's name
- `home_score` - home team's score at the end of the match, including extra time but excluding penalties
- `away_score` - away team's score at the end of the match, including extra time but excluding penalties
- `tournament` - tournament name
- `city` - the city or locality of the match
- `country` - the country hosting the match
- `neutral` - a TRUE/FALSE indicator of whether the venue was neutral

Assume that the result as shown in this dataset (win or tie) is the entire result – there are some cases of penalty shootouts and goals scored from penalties, but for complexity, we will ignore those for this exercise.

For clarity, current names are used for both teams and countries in historical matches. For example, an 1882 match featuring the team then known as Ireland against England is listed under Northern Ireland, reflecting the modern successor of the 1882 team. Country names are recorded as they were at the time of the match, but discrepancies between team and country names (e.g., Ghana vs. Gold Coast) are accounted for, with the `neutral` column marking such matches as non-neutral to clarify they were played at home.

Data: results.csv

_____

_____

_____

_____

_____

721

PSUpload **Upload your csv or Excel file here:**

Page Break

722

ProbsGPTConvConf **Confirm you used ChatGPT to help complete your task**

- ○ Yes I used ChatGPT (1)
- ○ No I did not use ChatGPT (0)

Display This Question:

If Confirm you used ChatGPT to help complete your task = No I did not use ChatGPT

ProbsGPTConvWhy Explain why you did not use ChatGPT when you were instructed to. Make sure to use it for any remaining task where you are instructed to.

_____

ProbsGPTConvLink **"Share" your ChatGPT conversations with us, so we can better understand how ChatGPT assisted you with the tasks. Only include conversations that you used for the task you just completed. If you used multiple ChatGPT conversations for the task, please share a link to each one.**

NOTE: *Please do not delete your conversations for a week or so, so we can make sure to collect the data we need to from them. If you delete the conversation on your side, we will no longer be able to view your shared links.*

The screenshot below shows how you can Share your conversation. Select the 3 dots on the Individual Conversation tab, and select "Share", then copy the link and paste it below.

_____

_____

_____

_____

723

_____

ProbsGoogleConf **Did you use Google to help complete your task?**

- Yes (1)
- No (0)

ProbsOtherTools **Please explain any other tools you used to complete your tasks. Include the name of the tool used, and how you used it to assist you.**
Be as specific as you can.

_____

_____

_____

_____

_____

End of Block: Problem Solving Task

Start of Block: Post-survey-1

TimerPost1 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

724

FatiguePost **Now that you've completed the tasks, how would you describe your current level of focus and energy?**

- ○ Very high - I'm fully focused and ready (5)
- ○ Somewhat high - I feel quite prepared and alert (4)
- ○ Neutral - I'm neither tired nor particularly energized (3)
- ○ Somewhat low - I'm a bit tired or distracted (2)
- ○ Very low - I'm already feeling quite fatigued or unfocused (1)

Page Break

725

Q220 **Please answer the following questions to the best of your knowledge**

PLEASE DO NOT USE CHATGPT, OTHER LLM OR ANY OTHER SEARCH ENGINE (e.g., Google) TO ANSWER THESE QUESTIONS

DSKnowledgeQ1 **Suppose we have a 'test_group' column in our dataframe (df) which has the values 'treatment' and 'control'. Which of the following code snippets will give us a dataframe filtered only to have the rows which correspond to 'treatment'? Select all that apply.**

Note: a semicolon indicates a new line of code

- [ ] df = df['treatment'] (1)
- [ ] df = df[df['treatment']] (2)
- [ ] condition = df['test_group']= 'treatment'; df = df[condition] (3)
- [ ] df = df['test_group']= 'treatment'] (4)
- [ ] df = df['test_group']== 'treatment'] (5)
- [ ] condition = df['test_group']=='treatment'; df = df[condition] (6)
- [ ] condition = df['test_group'] != 'control'; df = df[condition] (7)
- [ ] I don't know (-1)

DSKnowledgeQ2 **Three companies A, B and C supply 25%, 35% and 40% of the notebooks to a school, respectively. Past experience shows that 5%, 4% and 2% of the**

726

notebooks produced by these companies are defective, respectively. If a notebook was found to be defective, what is the probability that the notebook was supplied by A?

- ○ 44/69 (1)
- ○ 25/69 (2)
- ○ 13/24 (3)
- ○ 11/24 (4)
- ○ I don't know (-1)

DSKnowledgeQ3 **Which of these techniques can be used to handle missing data in categorical features? Select all that apply.**

- ☐ Removing rows having missing data (1)
- ☐ Replacing missing values with the most frequent category (2)
- ☐ Replacing missing values with the mean (3)
- ☐ Replacing missing values using predictive algorithms like classifiers (4)
- ☐ Replacing missing values using predictive algorithms like regressors (5)
- ☐ I don't know (-1)

727

DSKnowledgeQ4 **Distance-based algorithms are not affected by scaling**

- ○ False (0)
- ○ True (1)
- ○ I don't know (-1)

DSKnowledgeQ5 **You are given a dataset of logos of famous companies, and you have to predict whether the logo contains letters or not. Under which category does this problem fall? Select all that apply**

- ☐ Classification (1)
- ☐ Regression (2)
- ☐ Clustering (3)
- ☐ Natural language processing (4)
- ☐ Computer vision (5)
- ☐ I don't know (-1)

Page Break

728

GenAICalPost0 **Rate how likely you think it is that ChatGPT will give a correct answer to the following prompts.**

PLEASE DO NOT USE ChatGPT, OTHER LLMs OR ANY OTHER SEARCH ENGINE (e.g., Google) TO ANSWER THESE QUESTIONS

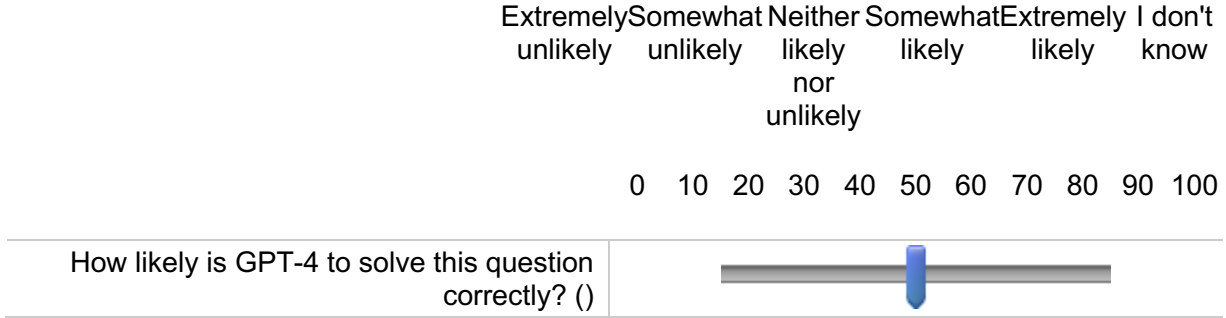
GenAICalPost1 Develop an HTML page with JavaScript and canvas to draw a representation of the US flag that rotates 90 degrees clockwise each time it is clicked.

	Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely	I don't know
	0	10	20 30 40	50	60 70 80	90 100
How likely is GPT-4 to solve this question correctly? ()						

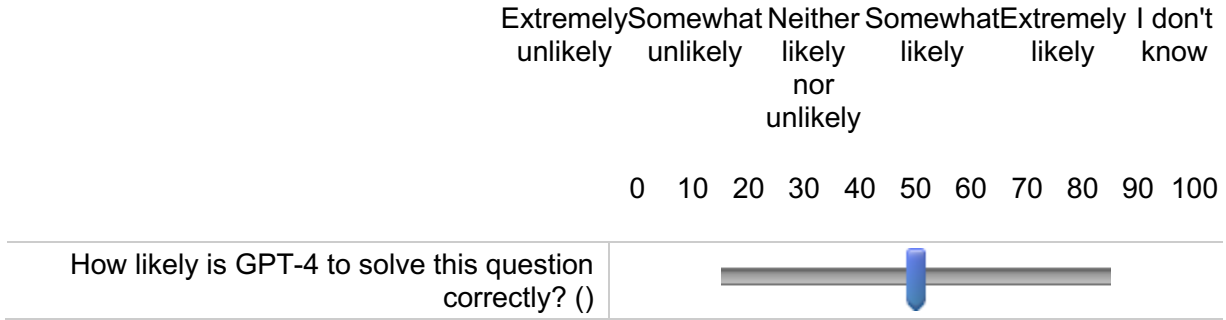
729

GenAICalPost2 Based on this data, which cities have an even-numbered population?

City (Special Ward)	Prefecture	Population	Area (km^2)	Density (per km^2)	Founded
Special wards of Tokyo	Tokyo	9,375,104	621.81	13,890	
Yokohama	Kanagawa	3,732,616	437.38	8,500	1889-04-01
Osaka	Osaka	2,691,185	222.30	11,900	1889-04-01
Nagoya	Aichi	2,327,557	326.45	6,860	1889-10-01
Sapporo	Hokkaido	1,976,257	1,121.26	1,763	1922-08-01
Fukuoka	Fukuoka	1,588,924	340.96	4,515	1889-04-01
Kawasaki	Kanagawa	1,531,646	142.70	9,626	1924-07-01
Kobe	Hyōgo	1,524,601	552.23	2,772	1889-04-01
Kyoto	Kyoto	1,464,890	827.90	1,800	1889-04-01
Saitama	Saitama	1,324,854	217.49	5,483	2001-05-01
Hiroshima	Hiroshima	1,199,391	905.13	1,286	1889-04-01



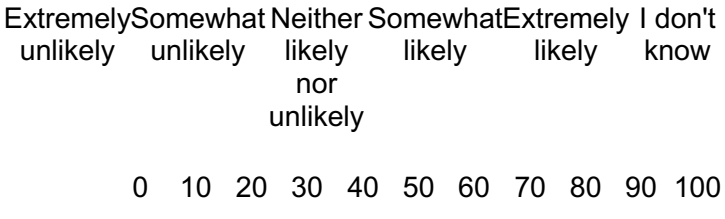
GenAICalPost3 I'm at a restaurant with a $10 bill and want to use it exactly on some of the following items. Which ones should I buy: steak $5.23 fries $1.24 shake $2.48 salad $4.87 salmon $4.13 cake $1.00



730

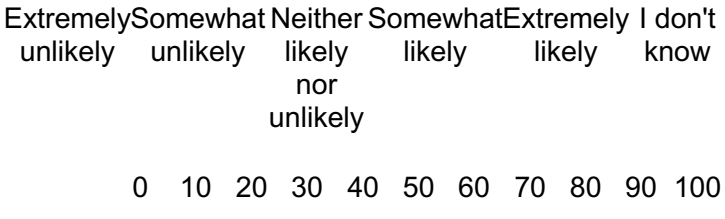
GenAICalPost4 Can you help me answer the following crossword clues.

1. "Lamented, in a way" (4 letters)
2. "Princess's irritant in a classic fairy tale" (3 letters")
3. "Bobbie Gentry's "___ to Billie Joe"" (3 letters)
4. "Leave no way out" (4 letters)
5. "Expression of false modesty from a texter" (4 letters)



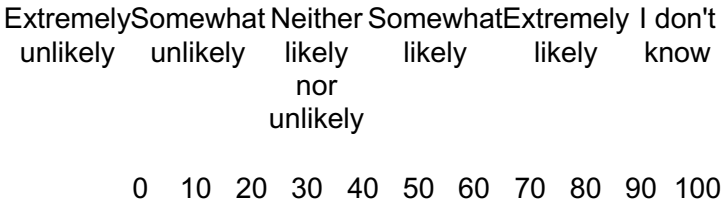
How likely is GPT-4 to solve this question correctly? ()	

GenAICalPost5 Who lost the Super Bowl two years after Pan-Am filed for bankruptcy?



How likely is GPT-4 to solve this question correctly? ()	

GenAICalPost6 Write out the word "hello" as an ascii art drawing with # and _



731

How likely is GPT-4 to solve this question correctly? ()	

GenAICalPost7 What is the best next move for O in the following game of Tic Tac Toe?

. | . | O

. | O | X

X | . | X

Extremely unlikely | Somewhat unlikely | Neither likely nor unlikely | Somewhat likely | Extremely likely | I don't know

0 10 20 30 40 50 60 70 80 90 100

How likely is GPT-4 to solve this question correctly? ()	

End of Block: Post-survey-1

Start of Block: Post-survey-GPT-only

732

GenAIPostUsage **How did you find the use of Generative AI? Was it easy or difficult? Did it give you the answers you were looking for? Select all that apply**

- [ ] The use of ChatGPT was easy and provided me with all the answers I was looking for (1)
- [ ] The use of ChatGPT was easy and provided me with most the answers I was looking for (2)
- [ ] The use of ChatGPT was easy, but did not provide me with most the answers I was looking for (3)
- [ ] The use of ChatGPT was difficult, but provided me with all the answers I was looking for (4)
- [ ] The use of ChatGPT was difficult, but provided me with most the answers I was looking for (5)
- [ ] The use of ChatGPT was difficult and did not provide me with most the answers I was looking for (6)

End of Block: Post-survey-GPT-only

Start of Block: Post-survey-3

TimerPost3 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

733

ProfessionalIdPost1 **Please indicate the extent to which you agree or disagree with the following statements:**

734

	Strongly Agree (4)	Somewhat Agree (3)	Neutral (2)	Somewhat Disagree (1)	Strongly Disagree (0)
Generative AI helps me feel valuable in my role (1)	○	○	○	○	○
Generative AI elevates how important I feel my job is for society (2)	○	○	○	○	○
Generative AI elevates my professional status and level of influence within my organization (3)	○	○	○	○	○
Generative AI helps me feel more competent in my role (4)	○	○	○	○	○
Generative AI enables my ability to execute tasks and reach desired outcomes (5)	○	○	○	○	○
Generative AI enables my ability to execute data analytics tasks and reach desired outcomes (6)	○	○	○	○	○

735

Generative AI increases the value I place on my expertise and skill cultivation (7)	○	○	○	○	○
Generative AI increases my level of autonomy in making individual decisions in my role (8)	○	○	○	○	○
Generative AI helps me be more confident that I will meet my project managers expectations (9)	○	○	○	○	○
Generative AI enables me to do what I really want to do in my role (11)	○	○	○	○	○
Generative AI will change the dynamic in my team (14)	○	○	○	○	○
Generative AI improved how I perceive my role in the organization (15)	○	○	○	○	○

736



737

ProfessionalIdPost2 **Next, please indicate the extent to which you agree or disagree with the following statements :**

738

	Strongly Agree (4)	Somewhat Agree (3)	Neutral (2)	Somewhat Disagree (1)	Strongly Disagree (0)
Using Generative AI helps me stay aligned with my project managers expectations (10)	○	○	○	○	○
I believe using Generative AI will contribute to the betterment of others in my work (12)	○	○	○	○	○
I see Generative AI as my coworker (13)	○	○	○	○	○
I would recommend Generative AI to other consultants (16)	○	○	○	○	○
I am proud of BCG's approach to Generative AI adoption within the firm (17)	○	○	○	○	○
I believe BCG is at the leading edge of the Generative AI revolution (18)	○	○	○	○	○
My managers and supervisors will expect more output from me because of Gen AI (19)	○	○	○	○	○

739

Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the ‘Problem solving and insights’ dimension (20)	○	○	○	○	○
Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the ‘Communication and Presence’ dimension (21)	○	○	○	○	○
Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the ‘Practicality and Effectiveness’ dimension (22)	○	○	○	○	○

740

GenAIUsagePost **Rate how helpful you think Generative AI tools are for these use cases (Rating 1-7, where 1 = Not at all helpful; 7 = Extremely helpful; with ability to say "I don't know")**

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	I don't know (-1)
Brainstorm ideas (7)	○	○	○	○	○	○	○	○
Writing code for data analytics (9)	○	○	○	○	○	○	○	○
Writing code for data visualizations (10)	○	○	○	○	○	○	○	○
Learning how to use excel for data analysis and visualizations (11)	○	○	○	○	○	○	○	○
Identifying which machine learning models to use for a project (12)	○	○	○	○	○	○	○	○
Understanding the statistical significance of a result (13)	○	○	○	○	○	○	○	○
Writing code for data cleaning and preparation (14)	○	○	○	○	○	○	○	○

741

GenAIPostBenefit **In a few words, what do you think will be the biggest benefits of Generative AI for you?**

_____

_____

_____

_____

_____

GenAIPostRisk **In a few words, what do you think will be the biggest risks of Generative AI for you?**

_____

GenAIPostConfidence **On a scale of 1-7, where 1 = "Not at all" and 7 = "Extremely", please rate the following:**

	Not at all / Neither / Extremely 1 2 3 4 5 6 7
How confident are you in your ability to contribute to data science projects? ()	
To what extent do you believe understanding data science concepts is important in the role of a BCG A/C? ()	

GenAIPostCapbilites **Given the capabilities of Generative AI, do you see the role of associates and consultants evolving in the next 5 years? If so, how?**

_____

_____

_____

742

_____

_____

End of Block: Post-survey-3

Start of Block: Post-survey-2

TimerPost2 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

JoyPost Finally, answer the following questions on a scale of 0 to 10, where 0 is "Do not enjoy at all" and 10 is "Enjoy to a great extent"

	Do not enjoy at all 0 1 2 3 4	Neutral 5	Enjoy to a great extent 6 7 8 9 10
How much do you think your coworkers enjoy their work? ()			
How much do you think your coworkers enjoy using ChatGPT for their work? ()			
How much would you enjoy doing more data analysis at work with the help fo ChatGPT? ()			
How much would you enjoy being tasked with data science tasks with the help of ChatGPT? ()			

Page Break

743

TrustPeoplePost **Generally speaking, would you say that most people can be trusted, or that you need to be very careful in dealing with people?**

- ○ Most people can be trusted (1)
- ○ You need to be very careful in dealing with people (0)
- ○ Don't know (-1)

744

ChatGPTPostTrust **Please indicate the extent to which you agree or disagree with the following statements:**

	Strongly disagree (0)	Disagree (1)	Neither agree nor disagree (2)	Agree (3)	Strongly agree (4)
ChatGPT can be trusted to give you correct information when researching a new topic (1)	○	○	○	○	○
ChatGPT can be trusted to do quantitative analysis for you (2)	○	○	○	○	○
ChatGPT can be trusted to clean data for you with minimal guidance (3)	○	○	○	○	○
ChatGPT can be trusted to help you learn to do new things (e.g. use a new type of software) (4)	○	○	○	○	○



ChatGPTPostConfidenc **On a scale of 1-7, where 1 = "Not at all" and 7 = "Extremely", please rate the following:**

Not at all			Neither			Extremely
1	2	3	4	5	6	7

745

How confident are you in your ability to do data cleaning in python using ChatGPT as your guide? ()	
How confident are you in your ability to do quantitative analysis using ChatGPT as your guide? ()	
How confident are you in your ability to learn new skills with ChatGPT as your guide? ()	
How confident are you in identifying factual inaccuracies in ChatGPT's responses? ()	
How confident are you in judging the relevance of ChatGPT's responses to your questions? ()	
How confident are you in assessing the clarity and understandability of ChatGPT's output? ()	
How confident are you in evaluating the completeness of ChatGPT's answers to your queries? ()	

End of Block: Post-survey-2

Start of Block: Investment game

InvestGameIntro Next, you'll have an **exciting opportunity** to play a game with ChatGPT as a second player (**yes, ChatGPT can play games and make decisions!**) and **earn tokens that you will be able to redeem for exciting real rewards provided by BCG!**

Note that **you will not see ChatGPT for this question**. Instead, we will email you with ChatGPT's response and with the reward options! **The more tokens you have at the end of the game – the better your options will be!**

InvestGameOverview Here's how the game works:

You (the Investor) and ChatGPT (the Responder) **will be given 100 tokens each**
As the Investor, you will have the opportunity to **pass some, all, or none of your tokens to ChatGPT (the Responder), as you like** Whatever amount you decide to give, **we will triple it and pass it to ChatGPT** For example, if you decide to pass 0 tokens, we will give ChatGPT 3x0=0 and so ChatGPT will have 100 tokens, while you have 100;
if you decide to pass 50 tokens, we will give ChatGPT 3x50=150 tokens so ChatGPT will have a total of 150+100=250 tokens, while you have 50; if you decide

746

to pass 100 tokens, we will give ChatGPT 3x100=300 tokens so ChatGPT will have a total of 300+100=400 tokens, while you have 0; After that, **ChatGPT (the Responder) will decide how many of the tokens it will give back to you**

InvestGameOverview1 In case this impacts your answer – here are the instructions we've given to ChatGPT:

"I would like to play the investment game with you. I'll be the investor and you will be the responder. As a starting point, you and I will each have 100 tokens. Whatever amount I decide to invest on you, there's a middle person who will triple that amount before passing it over to you. At that time, you will decide how much to pass back to me based on how much I invested and your total endowment. As a response, I need 4 numbers, the number I decided to pass over to you (label it "Investment amount"), the number that you will return back to me (label it "ChatGPT return"), total I have (label it "Total amount investor has") and total you have (label it "Total amount ChatGPT has").

Please make sure that the total amount you and I have at the end of the game sums up to the amount I have left, the amount that was tripled by the middle person and the amount you had at the beginning of the game."

InvestGameCalib1 To make sure you have a handle of the game, let's assume you decided to pass **20 tokens,** how many total tokens does ChatGPT have after the researchers tripled the amount?

_____

InvestGameCalib2 To make sure you have a handle of the game, let's assume you decided to pass **50 tokens**, how many total tokens does ChatGPT have after the researchers tripled the amount?

_____

747

InvestGameCalib3 To make sure you have a handle of the game, let's assume you decided to pass **80 tokens,** how many tokens does ChatGPT have after the researchers tripled the amount?



InvestmentAmount Okay, let’s play! **(Please do not use ChatGPT for this game, we will email you the results with the reward options)**

You now have 100 tokens and ChatGPT has 100 tokens. Don't forget, whatever number you decide to invest, we will triple it and pass it to ChatGPT. At that point, ChatGPT will decide how much to return back to you.

How many tokens would you like to pass to ChatGPT? (Please enter a number between 0 and 100)



InvestorGuess Next, could you guess how many tokens ChatGPT will give back to you. You will receive a bonus for a good guess (if your guess is within 5 tokens of the actual number), you will earn additional 10 tokens to redeem.

How much do you think ChatGPT has decided to give back to you? (The number should not be more than ChatGPT's endowment - [3*(how much you decided to send)] + 100)

End of Block: Investment game

Start of Block: Final video

FinalVideo We thank you very much for your participation in this study!
Before you exit, please watch this short video from OpenAI - our partners in this study!

748

End of Block: Final video

References

[1] **Iberdrola**, "What are exoskeletons and how can they help us overcome our human limitations?," 2024. Available at: https://www.iberdrola.com/innovation/what-are-exoskeletons. Accessed: 2024-09-03.

[2] **Djankov, Simeon and Federica Saliola**, "The changing nature of work," *Journal of International Affairs*, 2018, *72* (1), 57–74.

[3] **Capraro, Valerio, Austin Lentsch, Daron Acemoglu et al.**, "The impact of generative artificial intelligence on socioeconomic inequalities and policy making," *PNAS Nexus*, 2024, *3* (6).

[4] **Dell'Acqua, Fabrizio, McFowland, Mollick, Lifshitz, Kellogg, Rajendran, Krayer, Candelon, and Lakhani** Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2).

[5] **Brynjolfsson, Eric and Danielle Li and Lindsay Raymond** Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942.

[6] **Autor, David H, Frank Levy, and Richard J Murnane**, "The skill content of recent technological change: An empirical exploration," *The Quarterly Journal of Economics*, 2003, *118* (4), 1279–1333.

[7] **Acemoglu, Daron and David Autor**, "Skills, tasks and technologies: Implications for employment and earnings," in *Handbook of Labor Economics*, Vol. 4, Elsevier, Amsterdam, 2011, pp. 1043–1171.

[8] **Deming, David J and Kadeem Noray**, "Earnings dynamics, changing job skills, and STEM careers," *The Quarterly Journal of Economics*, 2020, *135* (4), 1965–2005.

[9] **Li, Ling**, "Reskilling and upskilling the future-ready workforce for industry 4.0 and beyond," *Information Systems Frontiers*, 2022, pp. 1–16.

[10] **Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock**, "GPTs are GPTs: Labor market impact potential of LLMs," *Science*, 2024, *384* (6702), 1306–1308.

[11] **Mollick, Ethan R and Lilach Mollick**, "New Modes of Learning Enabled by AI Chatbots: Three Methods and Assignments," 12 2022. Available at SSRN: https://ssrn.com/abstract=4300783 or http://dx.doi.org/10.2139/ssrn.4300783.

[12] **Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz**, "Combining human expertise with artificial intelligence: Experimental evidence from radiology," Technical Report, National Bureau of Economic Research 2023.

[13] **Dell'Acqua, Fabrizio**, "Falling asleep at the wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters," Technical Report, Working paper 2022.

[14] **Carlini, Nicholas**, "A GPT-4 Capability Forecasting Challenge," 2023. Available at: https://nicholas.carlini.com/writing/llm-forecast/.

[15] **Wiles, Emma and John J Horton**, "Generative AI and labor market matching efficiency," *Available at SSRN 5187344*, 2025.

[16] **Otis, Nicholas, Rowan P Clarke, Solene Delecourt, David Holtz, and Rembrand Koning**, "The uneven impact of generative AI on entrepreneurial performance," *Available at SSRN 4671369*, 2023.

[17] **Noy, Shakked and Whitney Zhang**, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, 2023, *381* (6654), 187–192.

[18] **Wiles, Emma, Zanele Munyikwa, and John Horton**, "Algorithmic writing assistance on jobseekers' resumes increases hires," *Management Science*, 2025.

[19] **Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer**, "The impact of AI on developer productivity: Evidence from GitHub Copilot," arXiv preprint arXiv:2302.06590, 2023.

[20] **Anderson, John R**, ed., *Cognitive Skills and Their Acquisition*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.

[21] **DeKeyser, Robert**, "Skill acquisition theory," in *Theories in Second Language Acquisition*, Routledge, New York and London, 2020, pp. 83–104.

[22] **Brown, Peter C, Henry L Roediger III, and Mark A McDaniel**, *Make It Stick: The Science of Successful Learning*, Harvard University Press, Cambridge, 2014.

[23] **Bjork, Elizabeth L and Robert A Bjork**, "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning," *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, 2011, *2* (56–64).

[24] **Agarwal, R. and J. Prasad**, "A conceptual and operational definition of personal innovativeness in the domain of information technology," *Information Systems Research*, 1998, *9* (2), 204–215.

[25] **Miron, E., M. Erez, and E. Naveh**, "Do personal characteristics and cultural values that promote innovation, quality, and efficiency compete or complement each other?," *Journal of Organizational Behavior*, 2004, *25* (2), 175–199.

[26] **Jha, S. and S. S. Bhattacharyya**, "Learning orientation and performance

orientation: Scale development and its relationship with performance," *Global Business Review*, 2013, *14* (1), 43–54.

[27] **Lee, David S**, "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 2009, *76* (3), 1071–1102.