

543 Appendix A Online Appendix

544 A.1 Task grading rubric and implementation

545 A.1.1 Use of LLMs in grading

546 • LLMs were used in the grading of all three tasks. Specifically, the API endpoint of
547 the model ‘gpt-4-turbo’ was used to grade (which, at the time of grading, pointed
548 to pt-4-turbo-2024-04-09).

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

551 • First, participant answers were preprocessed to minimize GPT-4’s known bias to
552 prefer its own answers and answers of a certain length: For detailed text-based
553 answers (more details on the task-specific descriptions below), the LLM was asked
554 to paraphrase and summarize the student answers, so that there would be no
555 obvious difference between the group assisted by ChatGPT and the group with-
556 out access to ChatGPT. Human validation was done to verify that this step
557 significantly minimized GPT’s bias.

558 • All grading was run using randomized batching

559 – Randomized, overlapping batches of participants’ answers were prepared,
560 with each participant’s answers appearing in a batch at least 5 times. This
561 allows for each participant’s answer to appear in batches with different
562 answers for each of the 5 iterations of being graded for a single answer. This
563 mitigates bias that may be caused by comparison to the answers of the other
564 participants in the batch. For example, the appearance of an average answer
565 in a batch alongside a very poor answer might overinflate the grade by com-
566 parison (due to the LLM perceiving one answer as a lot better than the
567 other). In addition, the appearance of an average answer in a batch with an
568 exceptionally good answer might deflate the grade by comparison. Random-
569 izing batches, having the answer graded 5 times in comparison with different
570 answers, and comparing the scores across batches allows any variability to
571 be caught and studied. Studying variability across batches helped make deci-
572 sions about which prompts or prompting strategies yield the most consistent
573 results.

574 • Prompting strategies were separately defined for each question

575 – For each task, and for each underlying question or question type, several
576 prompting strategies were tested. Some prompting strategies included pro-
577 viding rubrics with how many points should be assigned for the completion
578 of certain logical steps with all steps in one query. Alternatively, each step in
579 a query of its own, asking yes/no questions about whether a student had pro-
580 vided a specific answer or a shown a specific step of work towards calculating
581 a specific answer, yes/no questions about the manner in which the student
582 answered the question, one-shot or few-shot examples of how the student may
583 have solved each question or a whole or sub-step of each question, and even
584 allowing the LLM room to exercise judgment and deviate from the rubric
585 when an answer is not reflected exactly in what the rubric planned for. The

586 accuracy of the prompting strategies were determined by randomly select-
587 ing answers from human grading and comparing the human-assigned grade
588 with the LLM-assigned grade. The final prompting strategies were chosen to
589 minimize variability across batches and maximize accuracy.

- 590 • All steps and final grading were human-validated
 - 591 – For each task, a ground truth set of grades was created by a Data Scientist
592 for each substep of each question consisting of 10% or more of participants.
593 Having a ground truth made it possible to validate each batch of grades
594 output by the LLM with an objective reality of what the grades should be,
595 and batches of LLM outputs of grades of participant responses were able to
596 be judged by common metrics such as mean squared error from the ground
597 truth. This means squared error metric was used in prompt engineering and
598 when deciding what prompting strategy should be used for each question.
 - 599 – The language of the rubrics were adjusted throughout grading to accom-
600 modate the semantic requirements of LLM-based grading After each round
601 of testing of participants’ answers on each task, the rubric was adjusted as
602 needed when the following issues arose
 - 603 * LLM was instructed in most cases to abstain from grading a participants’
604 response if the response was not explicitly reflected in the rubric with
605 instructions on how to score or manage. When answers came up that
606 were not reflected explicitly in the rubric, the answers were flagged for
607 human review and subsequently added to the rubric with the appropriate
608 treatment of points assigned based on human (Data Scientist) judgment.
 - 609 * The original rubric submitted for pre-registration was insufficient for
610 an LLM to provide accurate scores. To ensure accuracy of LLM-based
611 grading, the language and grading-logic was simplified. Specifically, the
612 process of grading each question was broken down into smaller steps.
613 Sub-steps had to be defined with sub-points assigned to them based
614 on human (Data Scientist) judgment. However, the cumulative point
615 distributions and rankings of step-difficulty were kept constant against
616 the pre-registered rubric.

617 A.1.2 Coding task rubric and grading

618 Coding process rubric

619 The coding task is designed such that there is a deterministic set of logical steps
620 that need to be taken to arrive at the correct answer. The correct answer could be
621 identified by executing 10 steps with the provided dataset. Each step was assigned
622 a numerical score based on how difficult that step was to execute. This rubric was
623 validated by a Lead Data Scientist (4+ years of experience at BCG, likely more in
624 the industry). The following is an overarching view of the rubric steps:

625 RUBRIC STEP 1: Load the products and orders data frames (1 point).

627 RUBRIC STEP 2: Breaking up the order_info column into two columns with string
628 manipulation (4 points).

629 RUBRIC STEP 3: Converting the order_date column to datetime WITHOUT coerc-
630 ing errors, with allowance for mixed formatting (2 points).

631 RUBRIC STEP 4: Deleting duplicates in order_id or a set of customer_id and order_id
632 (3 points).

633 RUBRIC STEP 5: Impute NULL values in products dataframe in customer_id column
634 using the references from the orders dataframe where the information is available (5
635 points).

636 RUBRIC STEP 6: Replace values of price for each product in the dataframe with
637 the correct price. This involves studying all the prices and replacing all null and junk
638 ones with the correct price (10 points).

639 RUBRIC STEP 7: Merge the data frames correctly (2 points).

640 RUBRIC STEP 8: Filter the data correctly for the correct date range (1 point).

641 RUBRIC STEP 9: Get the total cost for each order by multiplying the prices and
642 quantities of each product in the order correctly (2 points)

643 RUBRIC STEP 10: Sort the data by total order cost to get the top 5 values (it does
644 not matter if they filter for the top 5 as long as they sort) (1 point).

645 The code was graded with the help of LLMs.

646 LLM Grading for Coding

647 For the coding task, the aforementioned grading architecture was deployed in the
648 following way:

- 649 • Preprocessing
 - 650 – The coding was task was performed by participants on Google Colab, in
651 iPython notebooks. The iPython notebooks were converted to python files,
652 with markdown blocks being converted into quoted comments. The text from
653 the derived python files was the input text for the LLMs to grade.
- 654 • Randomized batching
 - 655 – Batching was not performed for the coding task. Since the size of the input
656 text was quite large, this was done in order to avoid issues with the model's
657 context window that might arise if multiple users' answers were graded at
658 once. Each answer was still graded independently no less than 5 times.
659 The prompting strategies were adjusted appropriately to ensure maximal
660 consistency and accuracy.
- 661 • Prompting
 - 662 – After extensive testing and validation, it was determined that the prompting
663 strategy that should be used is asking, step by step with the past answers
664 in memory, whether each step had been taken, and to identify which lines
665 of code define it. Along with this, a rubric was there with various ways
666 of performing that steps, with examples in code of those ways. With this
667 information, the model was asked based on this to identify how the step was
668 executed if executed, and to assign points based on how as told by the rubric.
669 This means that at first, the model was asked whether step 1 of the rubric
670 was implemented correctly (with examples of correct implementation), to
671 identify the lines of code in the input where this implementation took place,
672 and to assign a score based on where the execution fell in the rubric. Then,
673 with the question and answer about step 1 added to the query context and

674 therefore model memory of the conversation, the model was asked the same
675 question about step 2. Having the information about former steps in memory
676 allowed the model to avoid confusing and conflating steps, especially when
677 they had logical dependency on one another. This method was evaluated on
678 the ground truth set to show 100% accuracy and 0% variability across 5 runs.
679 Here are the prompts for the model up to three steps of execution.

680 **System Prompt (Persona setting)**

681 coding_system_prompt: You will be given code used to solve a problem and an
682 accompanying rubric for each step expected to solve the problem. Assign points for
683 the completion of each step based on the rubric. The code does not have to follow the
684 rubric exactly, but it should follow the same logic.

685 For each step you are asked about, return your answer in this format:

686 SCORE: 1

687 REASONING: reasoning without code

688 LINES EXECUTING THIS STEP: code in the original that is executing that step

689

- 690 • rubric_step_1:

691 RUBRIC STEP 1: Load the products and orders data frames (1 point).

692 Example:

693 orders = pd.read_csv('orders.csv')

694 products = pd.read_csv('products.csv')

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

696

- 697 • rubric_step_2:

698 RUBRIC STEP 2: Breaking up the order_info column into two columns with
699 string manipulation (4 points).

700 EXAMPLE:

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

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

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

704

- 705 • rubric_step_3:

706 RUBRIC STEP 3: Converting the order_date column to datetime WITHOUT
707 coercing errors, with allowance for mixed formatting (2 points).

708 EXAMPLE:

709 orders['order_date'] = pd.to_datetime(orders['order_date'], format='mixed')

710 OR

711 Converting the order_date column to datetime WITH coercing errors (1 point).

712 EXAMPLE:

713 orders['order_date'] = pd.to_datetime(orders['order_date'], errors='coerce')

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

715

- 716 • rubric_step_4:

717 RUBRIC STEP 4: Deleting duplicates in order_id or a set of customer_id and
718 order_id (3 points).

719 EXAMPLE:

```

720 orders=orders[orders[['customer_id','order_id']].apply(frozenset,axis=1).duplicated()]
721 OR
722 orders = orders.drop_duplicates(subset='order_id')
723 HOW MANY POINTS WOULD YOU ASSIGN FOR THIS STEP? WHY?
724
725 • rubric_step_5:
726 RUBRIC STEP 5: Impute NULL values in products dataframe in customer_id
727 column using the references from the orders dataframe where the information is
728 available (5 points).
729 EXAMPLE:
730 customers_dict = dict(zip(orders.order_id, orders.customer_id))
731 products['customer_id']=products['order_id'].apply(lambda
732 x:customers_dict[str(x)])
733 HOW MANY POINTS WOULD YOU ASSIGN FOR THIS STEP? WHY?
734
735 • rubric_step_6:
736 RUBRIC STEP 6: Replace values of price for each product in the dataframe
737 with the correct price. This involves studying all the prices and replacing all null
738 and junk ones with the correct price (10 points).
739 EXAMPLE:
740 import json
741 products['order_products']=products['order_products'].apply(lambda
742 x:json.loads(x.replace("'", "")).replace('NaT', '-100'))
743 from collections import defaultdict
744 from collections import Counter
745 def product_search(x):
746 for key in list(x.keys()):
747 product_dict[key].append(x[key][0])
748 product_dict = defaultdict(list)
749 products['order_products'].apply(lambda x: product_search(x))
750 from statistics import mode
751 for key in product_dict.keys():
752 print(key)
753 print(Counter(product_dict[key]))
754 for key in product_dict.keys():
755 product_dict[key] = mode(product_dict[key])
756 def correct_cost(x):
757 for key in x.keys():
758 if x[key][0] == 100:
759 x[key][0] = product_dict[key]
760 elif x[key][0] == -100:
761 x[key][0] = product_dict[key]
762 elif x[key][0] == 0:
763 x[key][0] = product_dict[key]
764 return x

```

765 products['order_products']=products['order_products'].apply(lambda
766 x:correct_cost(x))

767 OR

768 Replace values of price for each product in the dataframe with the correct price,
769 but this time replacing all values with the correct price (not just null and junk
770 values) (6 points).

771 EXAMPLE:

```
772 import json
773 products['order_products']=products['order_products'].apply(lambda
774 x:json.loads(x.replace("'", "")).replace('NaN', '-100'))
775 from collections import defaultdict
776 from collections import Counter
777 def product_search(x):
778     for key in list(x.keys()):
779         product_dict[key].append(x[key][0])
780     product_dict = defaultdict(list)
781     products['order_products'].apply(lambda x: product_search(x))
782 from statistics import mode
783 for key in product_dict.keys():
784     print(key)
785     print(Counter(product_dict[key]))
786 for key in product_dict.keys():
787     product_dict[key] = mode(product_dict[key])
788 products['order_products']=products['order_products'].apply(lambda
789 x:product_dict(x))
```

790 OR

791 If the above steps are completed correctly but errors='coerce' parameter is
792 passed to any function USED FOR THIS STEP (5 points).

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

794

795 • rubric_step_7:

796 RUBRIC STEP 7: Merge the data frames correctly (2 points).

797 EXAMPLE:

```
798 products.order_id = products.order_id.apply(str)
799 final=orders.merge(products,right_on=['customer_id','order_id'],left_on=['customer_id','order_id'])
```

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

801

802 • rubric_step_8:

803 RUBRIC STEP 8: Filter the data correctly for the correct date range (1 point).

804 EXAMPLE:

```
805 final['order_month']=final['order_date'].apply(lambda x:str(x.year) + ' ' +
806 str(x.month))
```

```
807 may_2022 = final[final['order_month'] == '2022 5']
```

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

809

- 810 • rubric_step_9:
811 RUBRIC STEP 9: Get the total cost for each order by multiplying the prices
812 and quantities of each product in the order correctly (2 points).
813 `final['total_order_cost'] = final['order_products'].apply(lambda x: sum([y[0]*y[1]
814 for y in list(x.values()))))`
815 HOW MANY POINTS WOULD YOU ASSIGN FOR THIS STEP? WHY?
816
- 817 • rubric_step_10: RUBRIC STEP 10: Sort the data by total
818 order cost to get the top 5 values (it does not matter if
819 they filter for the top 5 as long as they sort) (1 point).
820 `pd.DataFrame(may_2022.groupby('customer_id')['total_order_cost'].mean()).sort_values('total_order_cost'
821 ascending=False).head(5)`
822 HOW MANY POINTS WOULD YOU ASSIGN FOR THIS STEP? WHY?
- 823 • Alternative prompting strategies involved asking whether a certain step had been
824 conducted by the participant as a yes/no question, and then asking whether the
825 step had been conducted with a specific coding strategy or with certain paramete-
826 ters as a yes/no question. The yes/no strategy proved unsuccessful due to often
827 conflating and confusing steps. Similarly, asking independently about steps with-
828 out the other steps in context or memory caused confusion, which caused the
829 decision to be made not to ask parallel questions about each step but to ask con-
830 secutive questions about each step. Another strategy employed was providing all
831 steps at once and asking for the output in a certain format to be able to parse
832 results for each step. This performed better than other alternative methods but
833 still showed inconsistency and inaccuracy due to too many logical steps being
834 taken in one query.
- 835 • Validation
 - 836 – For the coding task, ground truth label sets of over 25% of the total corpus
837 of answers was graded by a Data Scientist. Each answer was validated to
838 have maximum accuracy for this 25% with the assumption that the accuracy
839 extended to the rest of the data.
- 840 • Rubric semantic adjustment
 - 841 – The LLM was initially asked to abstain from answering questions not explic-
842 itly mentioned in the rubric. The abstentions were studied in detail, and
843 where an answer or answer type was missing from the rubric, it was added
844 in along with an assignment of points decided by a Data Scientist. This ulti-
845 mately ensured that all answers were scored according to a rubric as intended
846 by the rubric designer, rather than arbitrarily through LLM ‘judgment’.

847 A.1.3 Statistics task rubric and grading

848 Statistics Task

849 Grading

850 An answer key was created for the Statistics task detailing the correct answers to
851 each question and the correct steps it would take to arrive at the correct answer. A
852 grading rubric was assigned to each correct answer in the answer key by assigning
853 working steps with point values based on how difficult they were to execute correctly,

854 and correct answers a point value based on how difficult they were to arrive at (where
855 applicable, correct answers without correct work were not awarded full points). The
856 answer key and the rubric were reviewed and validated for correctness and validity in
857 difficulty assessment and point assignment respectively by a few Lead Data Scientists
858 (or Data Scientists with 4+ years of experience at BCG, likely longer in the industry).

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

861

862 **Statistics task rubric**

863

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.

- Plot the distribution of each of the numerical variables and remove rows with outliers from this dataset **+0.5 points**
- One-hot encode the 'Borough' variable **+0.5 points**
- Investigate relationships between variables **+0.5 points**
- Convert date to a numerical variable **0 points**

- 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

Maximum: 5 points

Minimum: 0 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.

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

Maximum: 4 points

Minimum: 0 points

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 **+2 points**
- b. Your model is overfit **-4 points**
- c. The learning rate hyperparameter is too small **0 points**
- d. The learning rate hyperparameter is too large **0 points**
- e. The decision tree is too shallow **+2 points**
- f. The decision tree is too deep **-4 points**
- g. None of the above **0 points**

Maximum: 4 points

Minimum: 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**

Maximum: 3 points

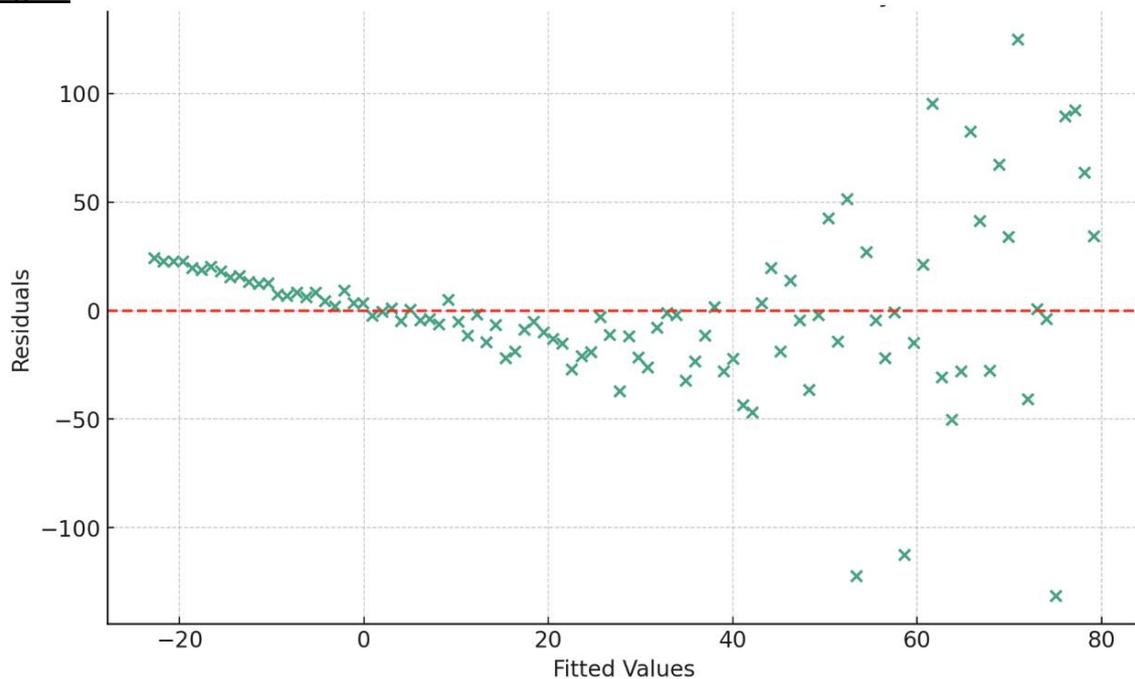
Minimum: 0 points

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 are not an immediate 0 unless both incorrect answers picked or 3 correct answers not picked

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



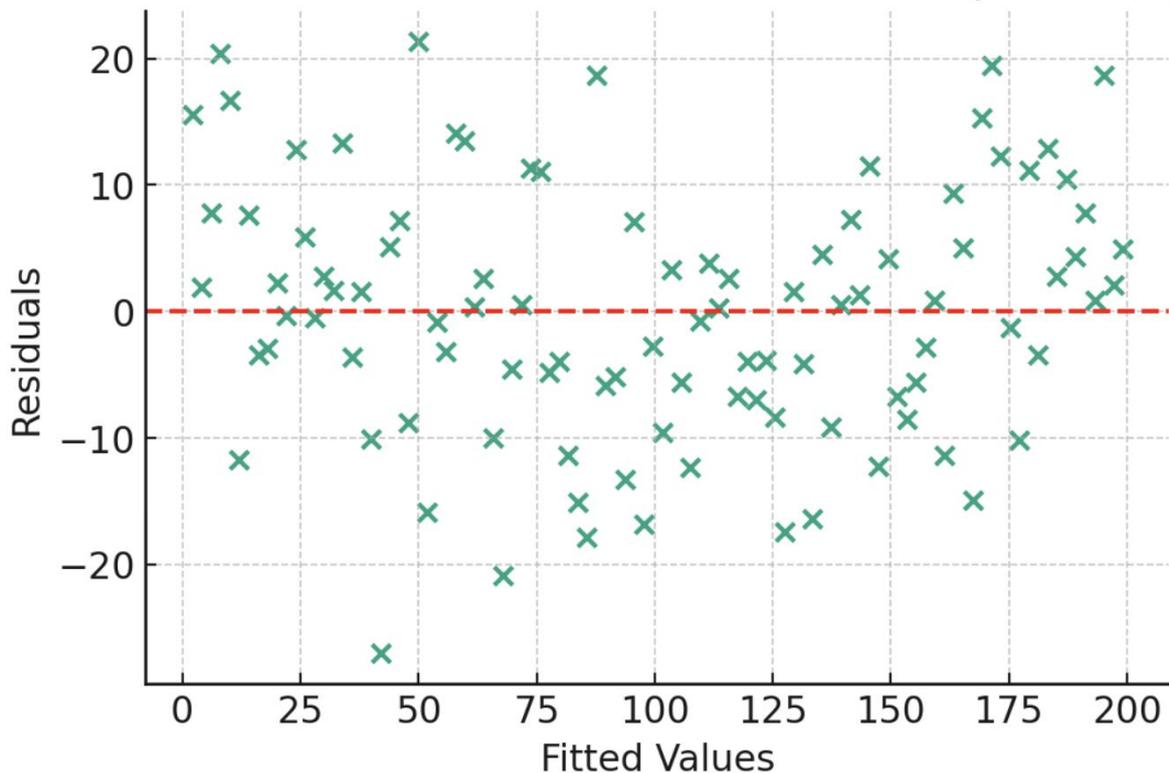
- No characteristic or issue is apparent **-2 points**
- Heteroscedastic data **1 points**
- Outliers **0 points**
- Response variable requires transformation **-1 points**
- A higher order variable might be required **2 points**

Maximum: 3 points

Minimum: 0 points

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



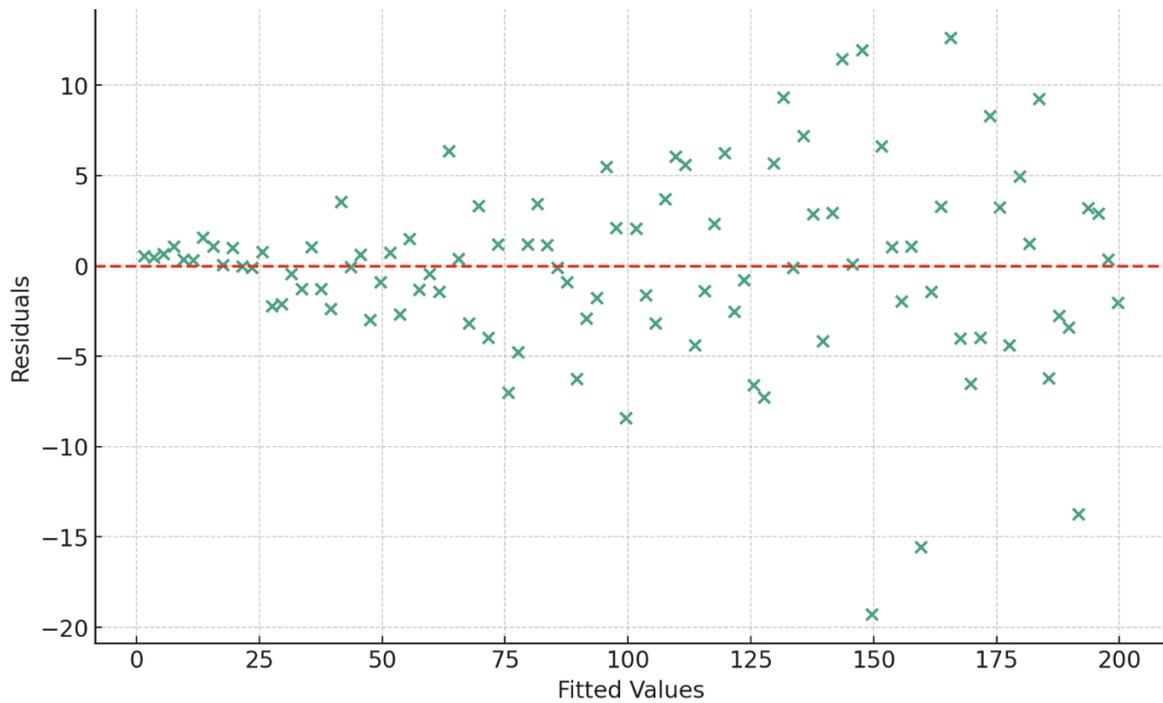
- a. No characteristic or issue is apparent **2 points**
- b. Heteroscedastic data **-2 points**
- d. Outliers **0 points**
- e. Response variable requires transformation **-2 points**
- f. A higher order variable might be required **-2 points**

Maximum: 2 points

Minimum: 0 points

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



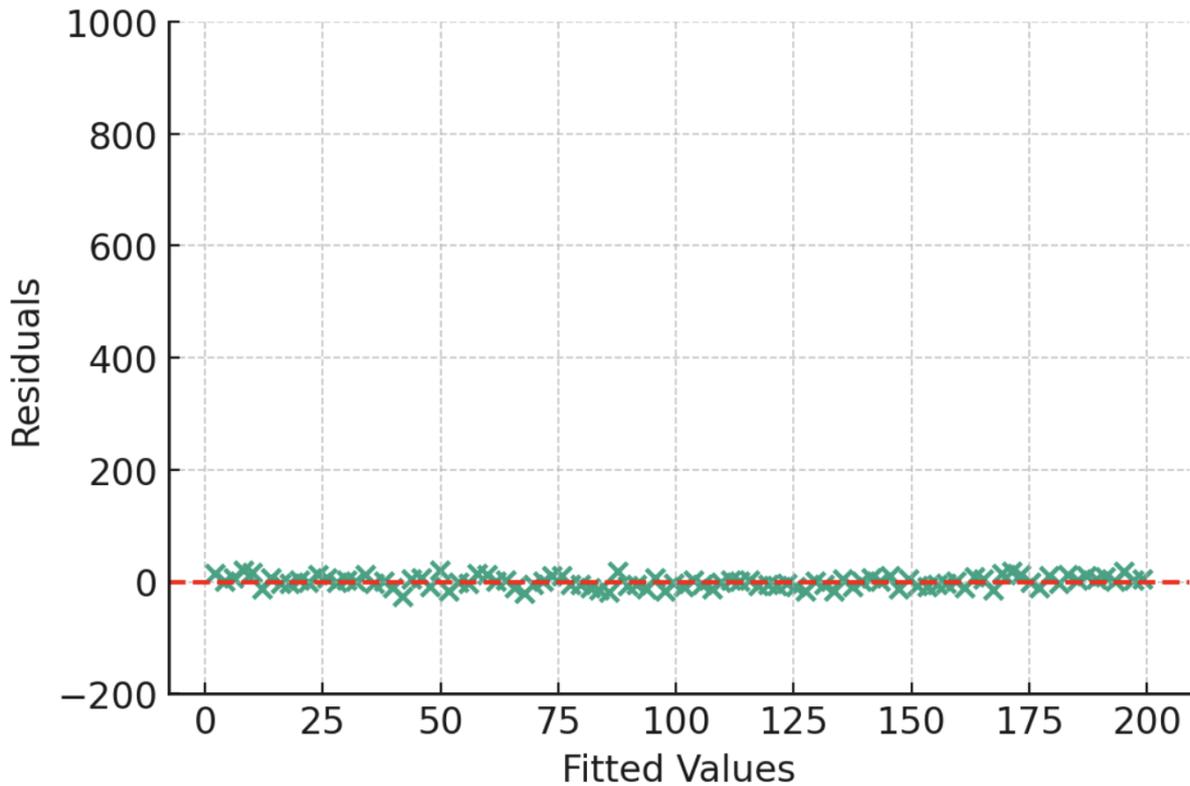
- a. No characteristic or issue is apparent **-2 points**
- b. Heteroscedastic data **2 points**
- e. Outliers **0 points**
- f. Response variable requires transformation **-2 points**
- g. A higher order variable might be required **-2 points**

Maximum: 2 points

Minimum: 0 points

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



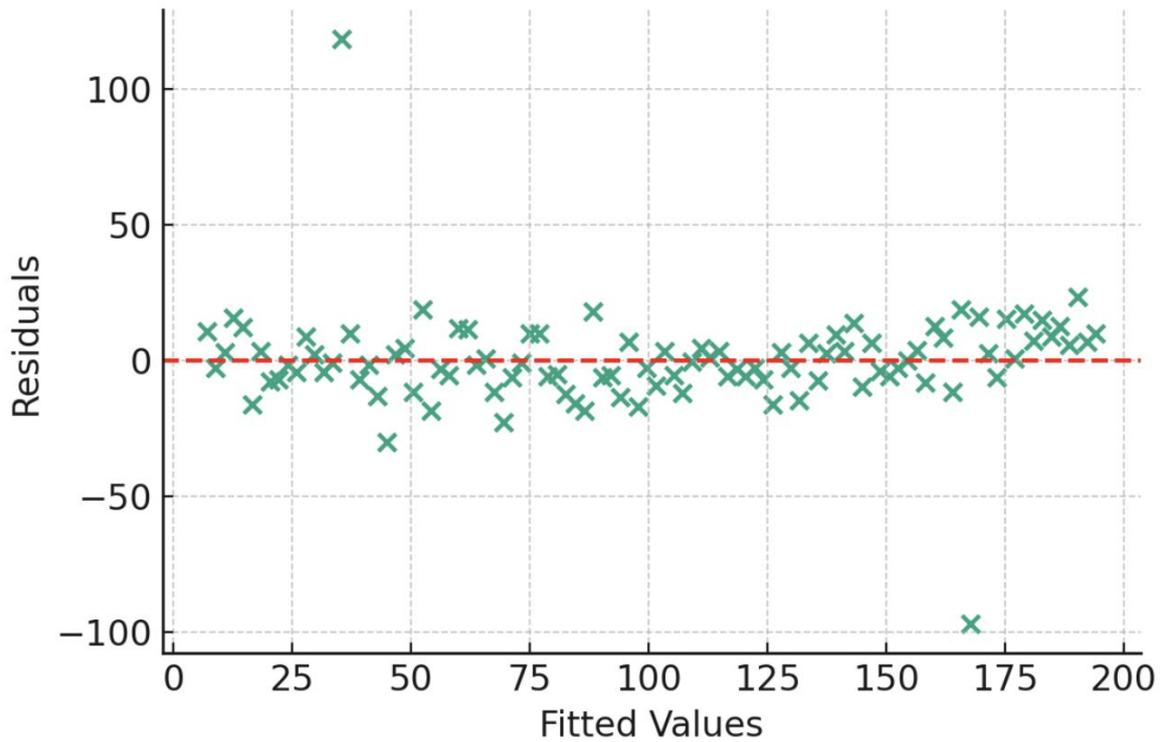
- a. No characteristic or issue is apparent **0 points**
- b. Heteroscedastic data **-2 points**
- c. Outliers **-2 points**
- d. Response variable requires transformation **2 points**
- e. A higher order variable might be required **-2 points**

Maximum: 2 points

Minimum: 0 points

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



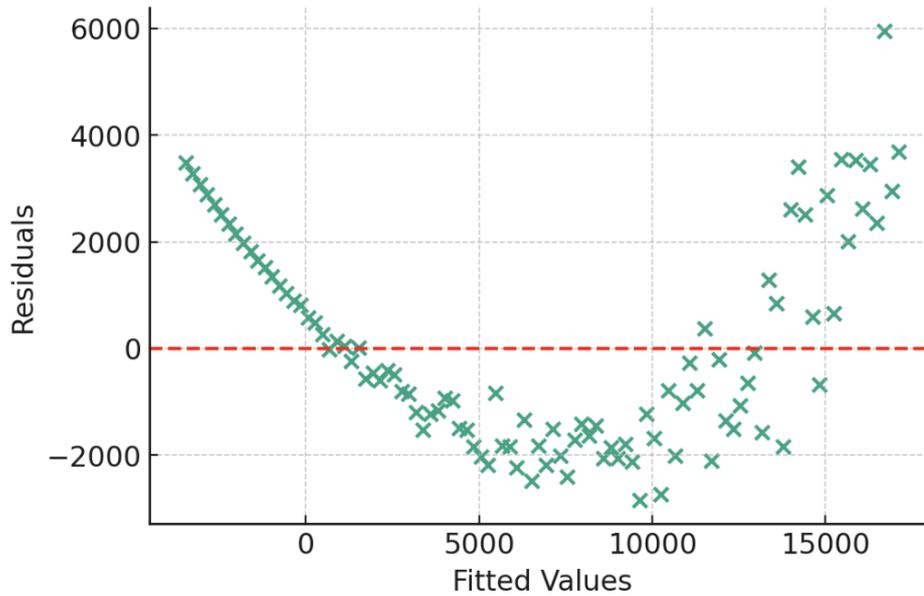
- a. No characteristic or issue is apparent **0 points**
- b. Heteroscedastic data **-2 points**
- c. Outliers **2 points**
- d. Response variable requires transformation **0 points**
- e. A higher order variable might be required **-2 points**

Maximum: 2 points

Minimum: 0 points

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



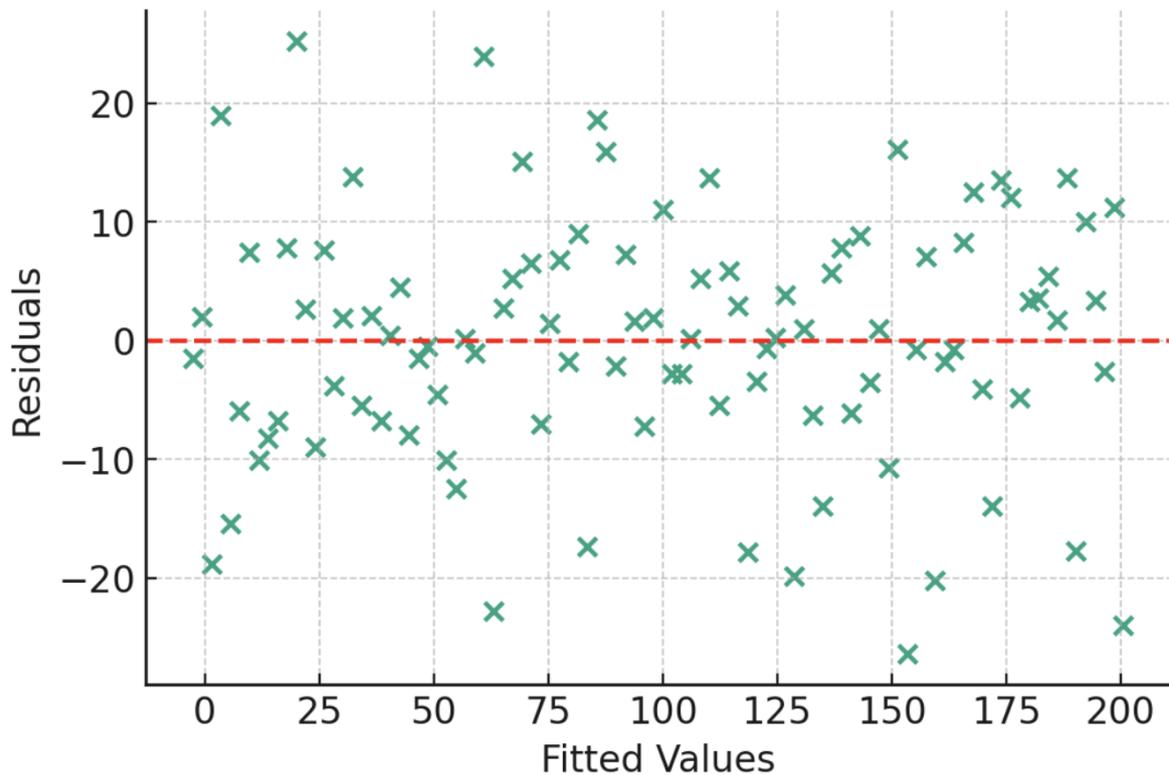
- No characteristic or issue is apparent **-2 points**
- Heteroscedastic data **1 points**
- Outliers **0 points**
- Response variable requires transformation **-2 points**
- A higher order variable might be required **2 points**

Maximum: 3 points

Minimum: 0 points

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



- No characteristic or issue is apparent **2 points**
- Heteroscedastic data **-2 points**
- Outliers **0 points**
- Response variable requires transformation **-2 points**
- A higher order variable might be required **-2 point**

Maximum: 2 points

Minimum: 0 points

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

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 **+2 points**
- Use a neural network instead of linear regression **-4 points**

- c. Use a regularized model instead of linear regression **+2 points**
- d. None of the above **0 points**

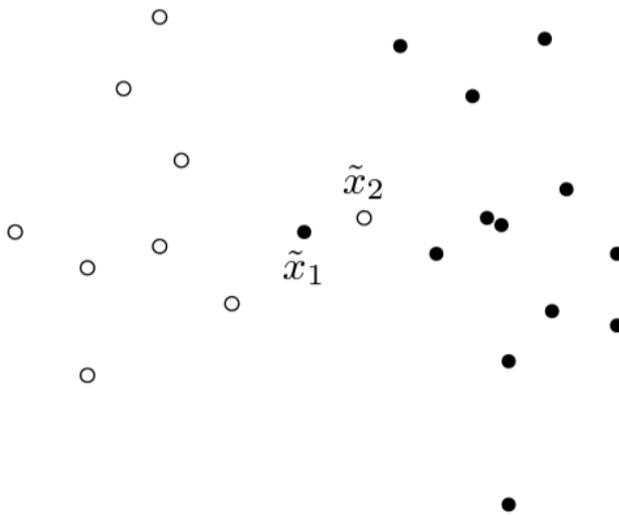
Maximum: 4 points

Minimum: 0 points

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

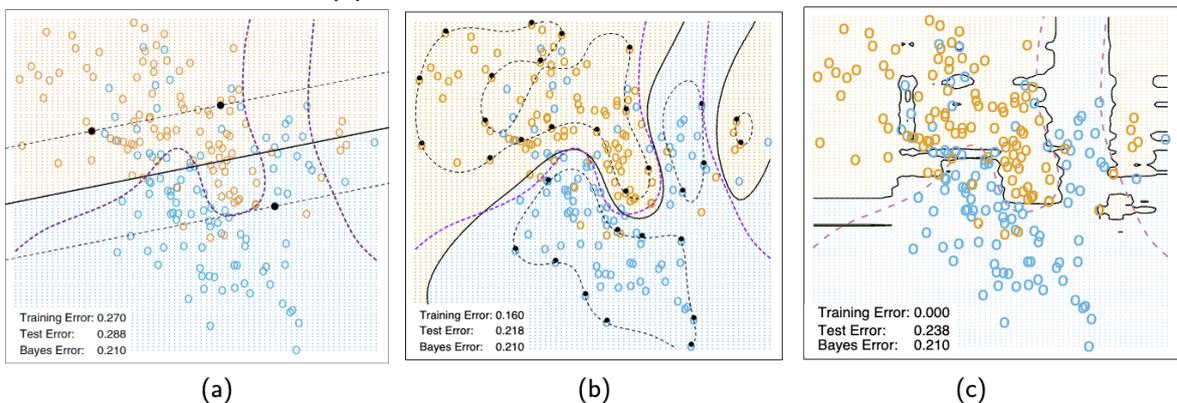
Award 1 point for: Identifying there will be 1 misclassification at best but not knowing what to do further

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)

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

Maximum: 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

Maximum: 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

Maximum: 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 = $P(\text{Find in A} | \text{Truck in Route A}) * P(\text{Truck in Route A}) = 0.65 * 0.45 = 0.2925$ (2 points)

Probability of finding the truck in Route B = $P(\text{Find in B} | \text{Truck in Route B}) * P(\text{Truck in Route B}) = 0.35 * 0.75 = 0.2625$ (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)

$P(\text{Posterior Truck in A} \mid \text{Truck not found on Day 1 in A}) = (P(\text{Truck not found on Day 1 in A} \mid \text{Truck in A}) * P(\text{Prior Truck in A})) / P(\text{Truck not found on Day 1 in A})$ (2 points)

$P(\text{Truck not found on Day 1 in A} \mid \text{Truck in A}) = 0.55$

$P(\text{Prior truck in A}) = 0.65$

$P(\text{Truck not found on Day 1 in A}) = P(\text{Truck not found on Day 1 in A} \mid \text{Route A}) + P(\text{Truck not found in Day 1 in A} \mid \text{Route B}) = 0.55 + 1 = 1.55$

Therefore $P(\text{Posterior Truck in A} \mid \text{Truck not found on Day 1 in A}) = 0.55 * 0.65 / 1.55 = 0.231$ (3 points)

Probability of finding the truck in Route A = $P(\text{Find in A} \mid \text{In Route A}) * P(\text{Posterior Truck in A}) = 0.23 * 0.45 = 0.1035$ (2 points)

Probability of finding the truck in Route B = $P(\text{Find in B} \mid \text{Route B}) * P(\text{Posterior Truck in Route B}) = 0.35 * (1 - 0.23) = 0.269$ (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 (prior probabilities are no longer valid).

Award 10 points for Route B with accompanying process steps.

LLM Grading for Statistics

For the statistics task, the aforementioned LLM grading architecture 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’. 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 extensively). This increased the accuracy

- 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 provide 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:

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

928 EXERCISE QUESTION:

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

933 What issue necessitates using all these metrics?

934 ANSWER RUBRIC:

935 This answer carries a total of 3 points.

936 If the student correctly identifies that the issue with the data is imbalanced or
937 unbalanced, or that the classes of positive and negative are imbalanced or unbalanced,
938 assign 3 points.

939 OR

940 If the student mentions stratified sampling or undersampling, assign 2 points.

941 OR

942 If the student recognizes that overfitting might be an issue, assign 1 point.

943 Return your score assignment in the following format

944 STUDENT 1 SCORE: points

945 STUDENT 1 EXPLANATION: explanation

946 STUDENT 2 SCORE: points

947 STUDENT 2 EXPLANATION: explanation

948 ...

- 949 • The second, performing better on questions expecting a specific set of answers
950 or answers with justification through calculations (less open-ended), yes or no
951 questions were asked about the answers and steps taken to answer the question,
952 with each answer broken up into several substeps asked about separately. Point
953 values were assigned based on a logic defined for each group of yes/no questions.

954 Here is an example of the same for answering one question:

955 StatsQ2.A.1:

956 You are a grader for a batch of students on the following exercise. You must grade
957 the answers of the students to the following exercise question according to the rubric.

958 EXERCISE QUESTION:

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

963 ANSWER RUBRIC:

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

966 Return your answer in the following format

967 STUDENT 1 RESULT: yes or no

968 STUDENT 2 RESULT: yes or no

969 ...

970 StatsQ2.A.2:
971 You are a grader for a batch of students on the following exercise. You must grade
972 the answers of the students to the following exercise question according to the rubric.
973 EXERCISE QUESTION:
974 You are asked to prepare a simple linear model to classify the following points into
975 class 1 (black dots) and class 2 (white dots). What is the best empirical risk of this
976 model that you can achieve with 0-1 loss? Justify your answer and show your working
977 steps.
978 ANSWER RUBRIC:
979 Did the student explicitly return the final answer as 2/22 (0.0909) or 20/22
980 (0.9090)?
981 Return your answer in the following format
982 STUDENT 1 RESULT: yes or no
983 STUDENT 2 RESULT: yes or no
984 ...
985 StatsQ2.A.3:
986 You are a grader for a batch of students on the following exercise. You must grade
987 the answers of the students to the following exercise question according to the rubric.
988 EXERCISE QUESTION:
989 You are asked to prepare a simple linear model to classify the following points into
990 class 1 (black dots) and class 2 (white dots). What is the best empirical risk of this
991 model that you can achieve with 0-1 loss? Justify your answer and show your working
992 steps.
993 ANSWER RUBRIC:
994 Did the student explicitly return the final answer as 21/22 (0.9545)?
995 Return your answer in the following format
996 STUDENT 1 RESULT: yes or no
997 STUDENT 2 RESULT: yes or no
998 ...
999 StatsQ2.A.4:
1000 You are a grader for a batch of students on the following exercise. You must grade
1001 the answers of the students to the following exercise question according to the rubric.
1002 EXERCISE QUESTION:
1003 You are asked to prepare a simple linear model to classify the following points into
1004 class 1 (black dots) and class 2 (white dots). What is the best empirical risk of this
1005 model that you can achieve with 0-1 loss? Justify your answer and show your working
1006 steps.
1007 ANSWER RUBRIC:
1008 Did the student explicitly return the final answer as 21/22 (0.0454)?
1009 Return your answer in the following format
1010 STUDENT 1 RESULT: yes or no
1011 STUDENT 2 RESULT: yes or no
1012 ...
1013 • All prompting strategies were selected for maximal accuracy and minimal
1014 variability across answers in batches of answers (maximal consistency).
1015 • Validation

- 1016 – For each free response answer in the statistics task, ground truth label sets
- 1017 of over 40% of the total corpus of answers was graded by a Data Scientist.
- 1018 Each answer was validated to have maximum accuracy for this 40% with the
- 1019 assumption that the accuracy extended to the rest of the data.
- 1020 • Rubric semantic adjustment
- 1021 – For the questions answered with the scoring rubric strategy, the LLM was
- 1022 initially asked to abstain from answering questions not explicitly mentioned in
- 1023 the rubric. For the yes/no strategy, the LLM initially returned some answers
- 1024 that all contained no responses. The abstentions and no answers were studied
- 1025 in detail, and where an answer or answer type was missing from the rubric, it
- 1026 was added in along with an assignment of points decided by a Data Scientist.
- 1027 This ultimately ensured that all answers were scored according to a rubric
- 1028 as intended by the rubric designer, rather than arbitrarily through LLM
- 1029 ‘judgment’.

1030 **A.1.4 Prediction task rubric and grading**

1031 **Methodology Score**

1032 **Rubric Development**

1033 The methodology score for the problem-solving task is based solely on the par-

1034 ticipants’ description of the methods they each applied to perform the task. The

1035 Data Scientists’ descriptions were studied by a human reviewer (Data Scientist) to

1036 understand the major themes of methods employed and the key decision points differ-

1037 entiating methods among participants. Broad categories were formed based on these to

1038 represent the rubric, and all participants’ data was reviewed manually (by a Data Sci-

1039 entist) and programmatically (with LLM support) to confirm that all methodologies

1040 fell within the categorical spectrum defined by the rubric.

1041 A key limitation of this methodology is that we are limited by the description

1042 provided by the participant. For example, if a participant took several steps to validate

1043 their analysis but failed to mention this in their description of their methodology, this

1044 will be flagged as a participant who did not validate their analysis and they would

1045 not get credit for the same. After careful consideration and several passes through the

1046 data, the categories were finalized so as to be generalizable to the maximum amount

1047 of the data while limiting bias on how detailed the descriptions were.

1048 The categories defined by the rubric are as follows:

1049 1. Predictive methodology

1050 In order to solve the problem posed by the task, each participant has to define

1051 a quantitative, classifying, or gradable outcome related to each soccer match in the

1052 dataset, and perform a prediction of that outcome for each match. The predictive

1053 methodology category refers to the mechanism employed by the participant to arrive

1054 at that prediction. This is the very basis of the methodological evaluation and other

1055 categories are best understood in relation to the predictive methodology. Within this

1056 category are the following methodological subcategories

- 1057 • Machine Learning
- 1058 – Regression

- 1059 * Linear regression on difference of goals scored by each team: The partic-
1060 ipants define the goal difference in each match as the outcome variable
1061 to predict and perform linear regression on a combination of feature
1062 variables to predict this.
- 1063 * Ensemble regression on goal difference: The participants define the goal
1064 difference in each match as the outcome variable to predict and per-
1065 form regression using ensemble methods such as random forest on a
1066 combination of feature variables to predict this.
- 1067 * Linear regression on numerical definitions of classes: The participants
1068 define a categorical outcome in each match as the outcome variable and
1069 define custom classes as a specific numeric (for example, +1 for home
1070 team victory, 0 for draw, -1 for home team loss). Then, the participants
1071 perform linear regression and transform the result to predict the custom
1072 class outcome.
- 1073 – Classification: Perform classification to predict match outcomes as victory or
1074 loss, using one of the following methods
 - 1075 * Logistic regression
 - 1076 * Random forest
 - 1077 * Gradient boosting
- 1078 • Probabilistic Analysis
 - 1079 – Fitting historical outcomes to a Poisson distribution: Participants fit histori-
1080 cal victory or goal statistics per team to a Poisson distribution and use those
1081 calculations to calculate the likelihood of the outcome that did occur or the
1082 most likely outcome.
 - 1083 – Using Elo ratings: Participants use established methodologies in sports or
1084 competition data science such as the calculation of Elo ratings to assess
1085 relative team strengths as part of a probabilistic calculation about outcomes.
- 1086 • Summary statistics: Develop custom summary statistics using combinations and
1087 slices of historical data and perform a probabilistic calculation about outcomes
- 1088 2. Definition of predictability

1089 The problem statement for the problem-solving task explicitly asks the participants
1090 to provide a ‘predictability score’ for each match in the dataset. Depending on the
1091 predictive methodology selected by the participant, the definition of predictability
1092 definition used makes logical sense or does not. The following are the categories for
1093 predictability definition

 - 1094 • Outcome difference: Participants define predictability of a match result as the dif-
1095 ference between their prediction of the outcome and the actual outcome, however
1096 they might define their dependent variable.
 - 1097 • Probability: Participants do not solidly define predictability beyond the raw out-
1098 put of their predictive models. For example, when using classification, participants
1099 return the probability of a home team victory as the predictability of the match
1100 result (a 0% forecasted chance of a home team victory signals a highly predictable
1101 lost but when this raw number is returned as predictability, this is logically and
1102 contextually unsound and signals a lack of understanding of the problem or of
1103 the predictive model functionality). For another example, when using summary

1104 statistics, participants return the raw probability of victory for a certain team
1105 based on historical probability of victory. This does not automatically have any
1106 meaning when it comes to the predictability of a match.

- 1107 • Processed probability: Participants define predictability of a match result as a
1108 function of the predictive power or predictive results of their model. They mean-
1109 ingfully process the probability outcomes or outputs of their models to arrive at
1110 a well-defined quantitative notion of predictability.
- 1111 • Team difference: Participants define their own predictability metric relative to
1112 the data or model they use relating to the difference between predicted outcomes
1113 for each team, where there is relatively subjective logic. For example, those that
1114 choose to regress on goal difference between teams can define predictability such
1115 that a larger predicted goal difference implies a more predictable match. Similarly,
1116 those that choose to create metrics for strengths of teams could define the pre-
1117 dictability of the match as the difference between team strengths (with a higher
1118 difference implying a more predictable result). This method does not make sense
1119 for classifiers due to the fact that you would be predicting the outcome of one
1120 team, and if you predicted the outcomes of both, they are mutually exclusive and
1121 interdependent outcomes.
- 1122 • Z-score: Participants that use historical data to arrive at likelihoods of match
1123 outcomes define predictability of the match as the likelihood of the outcome
1124 occurring, and use z-scores to quantify that likelihood.

1125 3. Categorical feature management

- 1126 • Match-based analysis: Participants use features as they appear on a match basis as
1127 part of their predictions (such as home and away advantage, neutral field or match
1128 type)
- 1129 • Team-based analysis: Participants create features representing team strengths or
1130 statistics based on historical data and incorporate those into predictions for each
1131 match
- 1132 • Both: Participants create features representing team strengths or statistics based on
1133 historical data and incorporate those into predictions for each match, and also use
1134 match-based features (such as home and away advantage, neutral field or match type)

1135 4. Temporal feature management (used date)

- 1136 • Yes / No: Participants did or did not make use of the temporal features. For
1137 example, participants did or did not extract years, months, or decades from
1138 the temporal data to use as predictive features, or participants did or did
1139 not use moving or rolling windows of dates in their summarized statistical
1140 analyses.

1141 5. Method of validation (if any)

- 1142 • Accuracy validation: Participants conducted simple metric measurement of
1143 accuracy to assess the performance of their methodology and made changes
1144 accordingly, to improve performance
- 1145 • Cross validation: Participants conducted cross-validated metric measurement
1146 on rotating training and testing sets to assess the performance of their
1147 methodology and made changes accordingly, to improve performance.

1148 • Model selection: Participants attempted various different methods or models
1149 as their predictive methodology and evaluated each method using one or
1150 more metrics to select the highest performing methodology.

1151 6. Completion of final step (return of what the participant identifies at the most
1152 surprising match):

1153 • Yes / No: Participants did or did not reach the conclusion of their predictive
1154 and predictability analysis to return a final answer, whatever that may be
1155 based on their individual methodology.

1156 The final scoring rubric, determined on the basis of the above categories, is as
1157 follows.

1158 First, a basic score is arrived at based on a combination of predictive methodology
1159 and predictability definition. The reasoning behind the scoring is evident in the method
1160 descriptions above.

1161 Classification Scoring

1162 For those that employed classification, the following points were awarded according
1163 to their classification methodology and predictability definition

- 1164 • Outcome difference for predictability definition
 - 1165 – Logistic regression employed: 7 points
 - 1166 – Random forest employed: 8 points
 - 1167 – XGBoost employed: 8 points
 - 1168 – LightGBM employed: 9 points

1169 These methodologies are ranked in ascending order of robustness and success when
1170 applied to the problem at hand based on Data Scientist testing (this ranking of the
1171 models is specific to this use case and cannot be generalized to all problem types)

- 1172 • Processed probability for predictability definition
 - 1173 – Logistic regression employed: 7 points
 - 1174 – Random forest employed: 8 points
 - 1175 – XGBoost employed: 8 points
 - 1176 – LightGBM employed: 9 points

- 1177 • Team difference for predictability definition:
 - 1178 – Logistic regression employed: 5 points

1179 This method does not make sense for classification.

- 1180 • Probability for predictability definition
 - 1181 – Logistic regression employed: 5 points
 - 1182 – Random forest employed: 5 points
 - 1183 – XGBoost employed: 5 points
 - 1184 – LightGBM employed: 5 points

1185 As described above, the probability for predictability definition is a relatively poor
1186 and illogical definition. No additional points are awarded for model robustness when
1187 the outcome is illogical, but points are awarded for model execution.

1188 Regression Scoring

1189 For those that employed regression, the following points were awarded according
1190 to their regression methodology and predictability definition

- 1191 • Outcome difference for predictability definition
 - 1192 – Linear regression on difference of goals scored by each team: 8 points

1193 – Random forest regression on goal difference: 9 points
1194 – XGBoost regression on goal difference: 9 points
1195 – Linear regression on numerical definitions of classes: 6 points
1196 Regression is a generally more informative and robust way to solve this problem
1197 due to the inherent ability to incorporate more information (e.g. the difference in
1198 goals rather than just categorical feature) into the dependent variable. Therefore, the
1199 regression models generally score higher, even when regression is performed on classes,
1200 which is a poor way to use regression.

- 1201 • Processed probability for predictability definition
 - 1202 – Linear regression on difference of goals scored by each team: 8 points
 - 1203 – Linear regression on numerical definitions of classes: 6 points
- 1204 • Probability for predictability definition
 - 1205 – Linear regression on difference of goals scored by each team: 5 points
 - 1206 – Ensemble regression on goal difference: 5.5 points
 - 1207 – Linear regression on numerical definitions of classes: 4 points

1208 Summary Statistics Scoring

- 1209 • Outcome difference for predictability definition: 6 points
- 1210 • Processed probability for predictability definition: 6 points
- 1211 • Team difference for predictability definition: 5 points
- 1212 • Probability for predictability definition: 4 points
- 1213 • Z-Score scoring: 8 points

1214 Poisson Scoring

- 1215 • Outcome difference for predictability definition: 8 points
- 1216 • Probability for predictability definition: 6 points

1217 Elo Scoring

- 1218 • Outcome difference for predictability definition: 8 points
- 1219 • Processed probability for predictability definition: 8 points
- 1220 • Team difference for predictability definition: 7 points
- 1221 • Probability for predictability definition: 5 points

1222 Categorical Feature Management Scoring

1223 The following points were awarded based on the categorical feature management

- 1224 • If a team-based analysis was used, an additional 3 points were awarded due to
1225 the difficulty of incorporating team-based features and the additional analysis
1226 this would entail
- 1227 • If both team-based and match-based analysis were used, an additional 5 points
1228 were awarded due to the difficulty of incorporating both types of features and
1229 the additional analysis this would entail

1230 Temporal Feature Management Scoring

1231 If temporal features were used in the analysis, an additional 5 points were rewarded
1232 due to the difficulty of incorporating those features and the additional analysis this
1233 would entail.

1234 Method of Validation Scoring

1235 The following points were awarded based on the method of validation employed

1236 • If accuracy was measured to validate the model, an additional 4 points were
1237 awarded due to the effort to validate the model, which is essential in predictive
1238 modeling

1239 • If cross validation was performed or extensive model selection, an additional 7
1240 points were awarded due to the effort to validate, and the robustness of the
1241 methodology chosen to do so

1242 Completion of Final Step Scoring

1243 If the final step was completed, an additional 3 points were rewarded to signal
1244 end-to-end completion of the exercise.

1245 **LLM Grading for Problem-Solving**

1246 For the problem-solving task, the LLM grading did follow the aforementioned
1247 architecture, but was slightly different because each and every LLM output was man-
1248 ually reviewed by a Data Scientist and corrected as needed due to the complexity of
1249 the problem at hand. Therefore, the adoption of the LLM grading architecture for
1250 problem solving is described as follows:

- 1251 • Preprocessing

- 1252 – Each methodology was summarized along different dimensions by the LLM
1253 through the asking of specific questions related to the categories identified
1254 above. This process was akin to a ‘field extraction’ where the methodologies,
1255 initially all formatted differently and largely unstructured, were organized
1256 using the LLM into a more structured format, easier to parse and understand
1257 for grading purposes, making the downstream grading by the LLM more
1258 accurate. Each summary was then manually reviewed and validated. The
1259 prompts used to summarize are as follows.

- 1260 • Randomized batching

- 1261 – Each summarized category answer in the problem-solving task was batched
1262 with four other random answers in the same category and graded in a single
1263 query by the LLM. This was done to minimize the bias arising from com-
1264 parison as described in the overall LLM grading architecture. Each answer
1265 was graded at least 5 times in order to arrive at a consensus similar to the
1266 statistics task

- 1267 • Prompting

- 1268 – The actual grading was performed by prompting involving two tiers of clas-
1269 sification – the first using a yes/no approach, and the next using a simple
1270 classification. At first, the LLM was asked as a yes/no prompt whether each
1271 kind of predictive methodology was employed (whether classification was
1272 employed, regression was employed, Poisson distributions were fitted to, and
1273 so on), and if classification or regression were chosen, it was given choices of
1274 the sub-methodologies to choose from to classify which was used (for exam-
1275 ple, if classification was employed, which kind of model from the list below
1276 was used to classify?). For the categories without sub-categories, a yes/no
1277 answer sufficed.

- 1278 – Many other strategies were tested, but the prompting strategies and results
1279 informed the rubric as well, so for the above rubric, this was the only tested
1280 method that was relevant.

- 1281 • Validation

- 1282 – 100% of the methodology data was manually graded and reviewed by a Data
1283 Scientist. Every LLM output was validated for accuracy, and with under 5%
1284 inaccuracy, the answers were manually changed to the correct answers.
- 1285 • Rubric semantic adjustment
- 1286 – The rubric was informed by how much detail of information could be accu-
1287 rately parsed out by an LLM, and how much information was present
1288 consistently across methodologies. Therefore, the rubric was constantly
1289 adjusted in conjunction with changing prompts till the above described rubric
1290 was arrived at.

1291 **Correctness**

1292 We define correctness on the problem-solving task as closeness to the answers that
1293 the Data Scientists arrive at for the same exercise within the given time. Each Data
1294 Scientists’ predictability result is standardized and then regarded as a baseline. We
1295 collect two measurements of correctness for each participant.

- 1296 1. The standardized predictability result of each participant in the control or treat-
1297 ment group is then compared to each Data Scientist baseline in the data. The
1298 mean absolute error of the participants’ result is calculated in comparison to each
1299 baseline. For each participant, the lowest recorded mean absolute error against
1300 each baseline is recorded as one of their correctness scores.
- 1301 2. For each Data Scientist, and for each participant, we have a breakdown of their
1302 predictive methodology (needed for the predictive methodology score described
1303 above). For the second score, we calculate each participants’ predictability result
1304 only in comparison with the Data Scientists baselines for those Data Scientist who
1305 employed comparative predictive methodology to them. For each participant, the
1306 lowest recorded mean absolute error against the applicable baselines is recorded
1307 as the second correctness score.

1308 Three Data Scientists were excluded from the baselines – of these three, one
1309 returned their match dates as years, having discarded other data information. This
1310 made their dataset impossible to merge with others. Another Data Scientist used
1311 ChatGPT for assistance and was therefore excluded from analysis and baselines. The
1312 last one that was excluded only returned predictability scores for 5 matches, leaving
1313 the rest blank. It is impossible to get a good correctness score from 5 entries given the
1314 initial size of the dataset (40k+ entries).

1315 **A.2 Registration Survey**

1316 Thank you again for taking part in the Generative AI Experiment! The following ques-
1317 tionnaire will take roughly 30 minutes to complete and contains questions about your
1318 background and your experiences. Please take the time to thoroughly and thought-
1319 fully respond to these questions, as it is a crucial part of the overall experiment. We
1320 ask that you please take this questionnaire in one sitting, before February 16, 2024.

1321 Please note that by submitting this questionnaire, you agree to not discuss the
1322 contents of the experiment to anyone, inside or outside of BCG. This is crucial for
1323 experimental integrity, to ensure robustness of the results for scientific publication.

1324 **Data Use and Collection:**

1325 All data collected in this questionnaire will NOT be used for any other purposes other
1326 than this Generative AI experiment. Any data that is published internally to BCG, in
1327 scientific journals or alike will only be done so in aggregate, and personal information
1328 will never be released. This data will also only be shared with OpenAI in aggregate
1329 and personal information will not be released outside of BCG/BHI. Within the scope
1330 of this questionnaire, we will only collect your personal data, listed below.

- 1331 • Name
- 1332 • Email
- 1333 • Location
- 1334 • Gender
- 1335 • Tenure
- 1336 • Title
- 1337 • English proficiency
- 1338 • Education
- 1339 • Proficiency and orientation towards tech

1340 Your personal data will only be used for testing the hypotheses of this Generative
1341 AI experiment, within the scope of your employment contract. We will process your
1342 personal data in accordance with applicable data protection laws and BCG's Privacy
1343 Policy [Link to internal policy]

1344 **CDC Contribution:**

1345 As mentioned in the email, successful completion of participation in the study will
1346 count as an "office contribution" to your CDC to reflect our appreciation for your
1347 efforts. You will have the opportunity to provide your CDA details after completing
1348 the study. However, to avail of this opportunity, you must put in an "honest effort"
1349 throughout, as judged by the quality of your responses.

1350 If there are any questions at all, please contact Lisa Kraye (kraye.lisa@bcg.com)

1351 **Survey**

1352 **Demographics (Role and Location)**

- 1353 1. Please Provide your Name First Name _____
1354 Last Name _____

- 1355 2. Please Provide your BCG Email Address Below
1356 _____
- 1357 3. Please Select Your Home BCG Office Location
- 1358 • Africa
 - 1359 • Asia Pacifia
 - 1360 • Central & South America
 - 1361 • Europe & The Middle East
 - 1362 • North America
- 1363 4. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Africa))
- 1364 • Cairo
 - 1365 • Casablanca
 - 1366 • Johannesburg
 - 1367 • Lagos
 - 1368 • Luanda
 - 1369 • Nairobi
 - 1370 • Other (Please Elaborate) _____
- 1371 5. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Asia Pacific))
- 1372 • Auckland
 - 1373 • Bangkok
 - 1374 • Beijing
 - 1375 • Bengaluru
 - 1376 • Canberra
 - 1377 • Chennai
 - 1378 • Fukuoka
 - 1379 • Ho Chi Minh City
 - 1380 • Hong Kong
 - 1381 • Jakarta
 - 1382 • Kuala Lumpur
 - 1383 • Kyoto
 - 1384 • Manila
 - 1385 • Melbourne
 - 1386 • Mumbai
 - 1387 • Nagoya
 - 1388 • Gurugram
 - 1389 • New Delhi
 - 1390 • Osaka
 - 1391 • Perth
 - 1392 • Seoul
 - 1393 • Shanghai
 - 1394 • Shenzhen
 - 1395 • Singapore
 - 1396 • Sydney
 - 1397 • Taipei
 - 1398 • Tokyo
 - 1399 • Other (Please Elaborate) _____

- 1400 6. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Central &
1401 South America))
- 1402 • Bogota
 - 1403 • Buenos Aires
 - 1404 • Lima
 - 1405 • Panama City
 - 1406 • Rio De Janeiro
 - 1407 • Santiago
 - 1408 • San Jose
 - 1409 • Sao Paulo
 - 1410 • Other (Please Elaborate) _____
- 1411 7. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Europe &
1412 The Middle East))
- 1413 • Abu Dhabi
 - 1414 • Amsterdam
 - 1415 • Athens
 - 1416 • Baku
 - 1417 • Barcelona
 - 1418 • Berlin
 - 1419 • Brussels
 - 1420 • Budapest
 - 1421 • Cologne
 - 1422 • Copenhagen
 - 1423 • Doha
 - 1424 • Dubai
 - 1425 • Dusseldorf
 - 1426 • Frankfurt
 - 1427 • Geneva
 - 1428 • Hamburg
 - 1429 • Helsinki
 - 1430 • Istanbul
 - 1431 • Lisbon
 - 1432 • London
 - 1433 • Madrid
 - 1434 • Milan
 - 1435 • Munich
 - 1436 • Oslo
 - 1437 • Paris
 - 1438 • Prague
 - 1439 • Riyadh
 - 1440 • Rome
 - 1441 • Stockholm
 - 1442 • Stuttgart
 - 1443 • Tel Aviv
 - 1444 • Vienna
 - 1445 • Warsaw

- 1446 • Zurich
- 1447 • Other (Please Elaborate) _____
- 1448 8. Please Select Your BCG Affiliation Below
- 1449 • Traditional BCG Consulting Team
- 1450 • BCG X
- 1451 • BCG Platinion
- 1452 • Other (Please Specify) _____
- 1453 9. Please Select Your Official Title at BCG
- 1454 • Associate
- 1455 • Consultant
- 1456 • BCG X Data Scientist
- 1457 • BCG X Senior Data Scientist
- 1458 • Other (Please Specify) _____
- 1459 10. Please Select Your Total Tenure at BCG (in Years)
- 1460 • 0 to 1 Years
- 1461 • 1 Years to 2 Years
- 1462 • 2 Years to 3 Years
- 1463 • 3 Years to 4 Years
- 1464 • 4 Years to 5 Years
- 1465 • 5+ Years

1466 **Demographics (Education and Language)**

- 1467 1. What is your gender?
- 1468 • Female
- 1469 • Male
- 1470 • Prefer Not to Say
- 1471 • Other
- 1472 2. What is your English proficiency? (Reading, Written, and Spoken Combined)
- 1473 • Beginner
- 1474 • Intermediate
- 1475 • Advanced
- 1476 • Native
- 1477 3. What is your highest education level?
- 1478 • Bachelors
- 1479 • Masters
- 1480 • Professional Degree (e.g., MD, JD etc.)
- 1481 • Doctorate
- 1482 4. If you have a Bachelors degree, what was your major? Select the applicable
- 1483 categories and specify your degree in the text box.
- 1484 • Science and Mathematics _____
- 1485 • Engineering and Technology _____
- 1486 • Health Sciences _____
- 1487 • Social Sciences _____
- 1488 • Business and Economics _____
- 1489 • Arts and Humanities _____

- 1490 • Education _____
 - 1491 • Agriculture and Environmental Studies _____
 - 1492 • Other _____
- 1493 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 =
- 1494 Bachelors))
- 1495 • Science and Mathematics _____
 - 1496 • Engineering and Technology _____
 - 1497 • Health Sciences _____
 - 1498 • Social Sciences _____
 - 1499 • Business and Economics _____
 - 1500 • Arts and Humanities _____
 - 1501 • Education _____
 - 1502 • Agriculture and Environmental Studies _____
 - 1503 • Other _____
- 1504 6. If you have a Professional degree, what was your major? Select the applicable
- 1505 categories and specify your degree in the text box. (Conditionally Hidden if: (12
- 1506 = Bachelors OR 12 = Masters))
- 1507 • Science and Mathematics _____
 - 1508 • Engineering and Technology _____
 - 1509 • Health Sciences _____
 - 1510 • Social Sciences _____
 - 1511 • Business and Economics _____
 - 1512 • Arts and Humanities _____
 - 1513 • Education _____
 - 1514 • Agriculture and Environmental Studies _____
 - 1515 • Other _____
- 1516 7. If you have a Doctorate degree, what was your major? Select the applicable
- 1517 categories and specify your degree in the text box. (Conditionally Hidden if: (12
- 1518 = Bachelors OR 12 = Masters))
- 1519 • Science and Mathematics _____
 - 1520 • Engineering and Technology _____
 - 1521 • Health Sciences _____
 - 1522 • Social Sciences _____
 - 1523 • Business and Economics _____
 - 1524 • Arts and Humanities _____
 - 1525 • Education _____
 - 1526 • Agriculture and Environmental Studies _____
 - 1527 • Other _____
 - 1528

1529 **Programming Proficiency**

- 1530 17. What tools do you currently use for data analysis?
- 1531 • Excel
 - 1532 • Tableau
 - 1533 • Alteryx

- 1534 • Programming (e.g. Python)
- 1535 • ChatGPT
- 1536 • Other LLMs / LLM based tools
- 1537 • Other non-LLM based tools
- 1538 18. Do you know how to code?
- 1539 • Yes, I am an expert level coder
- 1540 • I know how to code, but am not an expert
- 1541 • I only know the basics of coding
- 1542 • No, I do not know how to code
- 1543 19. How often do you code for work? (Conditionally Hidden if: (17 = I know how to
- 1544 code, but am not an expert))
- 1545 • I never code for work
- 1546 • I code occasionally, but usually use other analytics tools
- 1547 • I code every time I work on analytical projects, but this is only occasionally
- 1548 • I code every time I work on analytical projects and I frequently am staffed
- 1549 on analytical projects
- 1550 • Coding is a core part of my job
- 1551 20. How many years of programming experience do you have? (Conditionally Hidden
- 1552 if: (17 = I know how to code, but am not an expert))
- 1553 • 0-1
- 1554 • 2-3
- 1555 • 3-5
- 1556 • 5-8
- 1557 • 8+
- 1558 21. How many programming languages are you familiar with? (Conditionally Hidden
- 1559 if: (17 = I know how to code, but am not an expert))
- 1560 • 0
- 1561 • 1
- 1562 • 2-3
- 1563 • 4+
- 1564 22. How familiar are you with Python? (Conditionally Hidden if: (17 = I know how
- 1565 to code, but am not an expert))
- 1566 • 0 = I Do Not Program
- 1567 • 1 = Low Familiarity/Novice
- 1568 • 2
- 1569 • 3
- 1570 • 4
- 1571 • 5 = High Familiarity/Expert

1572 ChatGPT Proficiency

- 1573 23. How often do you use ChatGPT or other LLMs for work?
- 1574 • I have never used ChatGPT
- 1575 • I have tried ChatGPT once or twice
- 1576 • I use ChatGPT less than once per week
- 1577 • I use ChatGPT at least once per week

- 1578 • I use ChatGPT every day
- 1579 24. How often do you use ChatGPT or other LLMs in your personal life?
- 1580 • I have never used ChatGPT
- 1581 • I have tried ChatGPT once or twice
- 1582 • I use ChatGPT less than once per week
- 1583 • I use ChatGPT at least once per week
- 1584 • I use ChatGPT every day
- 1585 25. Please rate the extent to which you agree or disagree with the following statements
- 1586 (1-7 Rating)
- 1587 • I am familiar with GenAI for writing
- 1588 • I am familiar with using GenAI for coding
- 1589 • I am familiar with prompt engineering (i.e., crafting prompts to get a better
- 1590 answer from an AI model)
- 1591 • I am familiar with more than 2 prompting strategies
- 1592 • ChatGPT helps me become a better consultant
- 1593 • I understand how large language models (LLMs), which underpin generative
- 1594 AI tools for writing, work
- 1595 • I believe I can tell when ChatGPT is hallucinating
- 1596 • I have created a specialized GPTs for my purposes
- 1597 • I have used ChatGPT with Code Interpreter / Advanced Data Analytics
- 1598 • I use ChatGPT for writing code
- 1599 26. Please rate the extent to which you agree or disagree with the following statements
- 1600 (1-7 Rating)
- 1601 • ChatGPT is primarily a Data Science tool
- 1602 • ChatGPT is primarily a tool for writing
- 1603 • ChatGPT helps me be more proficient at problem solving
- 1604 • ChatGPT helps me be more efficient at creating slides
- 1605 Here's the LaTeX version of the provided text:

1606 **Tech Openness and Playfulness**

- 1607 27. Please rate the extent to which you agree or disagree with the following statements
- 1608 (1-7 Rating)
- 1609 • If I hear about a new technology product or service, I will look for ways to
- 1610 experiment with it
- 1611 • Among my peers, I am usually the first to try out new technology products
- 1612 and services
- 1613 • In general, I am hesitant to try out new technology products and services
- 1614 • I like to experiment with new technology products and services
- 1615 • I am spontaneous when I interact with new technology products or services
- 1616 • I am unimaginative when I interact with new technology products or services
- 1617 • I am playful when I interact with new technology products or services
- 1618 • I am flexible when I interact with new technology products or services
- 1619 • I am uninventive when I interact with new technology products or services
- 1620 • I am creative when I interact with new technology products or services
- 1621 • I am unoriginal when I interact with new technology products or services

1622 **Creativity**

1623 28. Please rate the extent to which you agree or disagree with the following statements
1624 (1-7 Rating)

- 1625 • I try not to oppose team members
- 1626 • I adapt myself to the system
- 1627 • I adhere to accepted rules in my area of work
- 1628 • I avoid cutting corners
- 1629 • I am thorough when solving problems
- 1630 • I address small details needed to perform the task
- 1631 • I perform the task precisely over a long time
- 1632 • I am good in tasks that require dealing with details
- 1633 • I have a lot of creative ideas
- 1634 • I prefer tasks that enable me to think creatively
- 1635 • I am innovative
- 1636 • I like to do things in an original way

1637 **Learning Orientation**

1638 29. Please indicate the extent to which you agree or disagree with the following
1639 statements (1-7 Rating)

- 1640 • I enjoy learning new topics
- 1641 • I like to read diverse topics
- 1642 • I find pleasure in learning
- 1643 • I get intrinsically motivated to constantly expand my knowledge
- 1644 • I seek deep-seated conceptual knowledge for the task assigned to me
- 1645 • I spend a lot of time thinking about how my performance is in comparison
1646 to others
- 1647 • I like to seek rewards in short term for my efforts
- 1648 • I prefer to see tangible output as a reward for my effort
- 1649 • I generally perform and undertake those tasks for which I get rewarded soon
- 1650 • I feel very good when I know I have outperformed other colleagues
- 1651 • I always try to communicate my achievements to my friends and supervisors

1652 **Concluding Remarks**

1653 30. Do you agree not to discuss the contents of this experiment with anyone, inside or
1654 outside of BCG? This is crucial for experimental integrity, to ensure robustness
1655 of the results for scientific publication.

1656 **A.3 Pre-experiment Survey & Training**

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

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

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

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

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



NeedCognition **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 would prefer complex to simple problems (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I like to have the responsibility of handling a situation that requires a lot of thinking (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Thinking is not my idea of fun (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I really enjoy a task that involves coming up with new solutions to problems (6)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought (7)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



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)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have been obsessed with a certain idea or project for a short time but later lost interest (8)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have difficulty maintaining my focus on projects that take more than a few months to complete (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
New ideas and projects sometimes distract me from previous ones (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My interests change from year to year. I become interested in new pursuits every few months (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



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)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Setbacks don't discourage me (8)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am diligent (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am a hard worker (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have achieved a goal that took years of work (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have overcome setbacks to conquer an important challenge (7)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



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

	No experience (0)	Somewhat experienced (1)	Very experienced (2)
Data visualization (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Machine learning models (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Statistical analysis (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Data cleaning and preparation (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



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? ()							

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)
-



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)	<input type="radio"/>				
I consider myself an expert in Tableau (2)	<input type="radio"/>				
I consider myself an expert in Alteryx (3)	<input type="radio"/>				
Based on prior CDC reviews, PSI (problem solving and insights) has been a core strength (4)	<input type="radio"/>				



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



ProfessionalIdPre1 **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)
Generative AI helps me feel valuable in my role (1)	<input type="radio"/>				
Generative AI elevates how important I feel my job is for society (2)	<input type="radio"/>				
Generative AI elevates my professional status and level of influence within my organization (3)	<input type="radio"/>				
Generative AI helps me feel more competent in my role (4)	<input type="radio"/>				
Generative AI enables my ability to execute tasks and reach desired outcomes (5)	<input type="radio"/>				
Generative AI enables my ability to execute data analytics tasks and reach desired outcomes (6)	<input type="radio"/>				

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)



ProfessionalIdPre2 **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)
Using Generative AI helps me stay aligned with my project managers expectations (10)	<input type="radio"/>				
I believe using Generative AI will contribute to the betterment of others in my work (12)	<input type="radio"/>				
I see Generative AI as my coworker (13)	<input type="radio"/>				
I would recommend Generative AI to other consultants (16)	<input type="radio"/>				
I am proud of BCG's approach to Generative AI adoption within the firm (17)	<input type="radio"/>				
I believe BCG is at the leading edge of the Generative AI revolution (18)	<input type="radio"/>				
My managers and supervisors will expect more output from me because of Gen AI (19)	<input type="radio"/>				

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)



**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")**

	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)	<input type="radio"/>							
Write a first draft for complex texts (e.g., reports) (2)	<input type="radio"/>							
Write my final version for simple texts (e.g., emails) (3)	<input type="radio"/>							
Write my final version for complex texts (e.g., reports) (4)	<input type="radio"/>							
Review my writing (grammar, typos, etc.) (5)	<input type="radio"/>							
Be more persuasive (6)	<input type="radio"/>							
Brainstorm ideas (7)	<input type="radio"/>							
Be more creative (8)	<input type="radio"/>							
Writing code for data analytics (9)	<input type="radio"/>							
Writing code for data visualizations (10)	<input type="radio"/>							

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)



GenAllImpactPre **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	
--	---	---	---	---	---	---	---	--

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Decreased collaboration	<input type="radio"/>	Increased collaboration						
Decreased efficiency	<input type="radio"/>	Increased efficiency						
Decreased clarity of responsibilities	<input type="radio"/>	Increased clarity of responsibilities						
Decreased learning	<input type="radio"/>	Increased learning						
Decreased decision quality	<input type="radio"/>	Increased decision quality						
Reduced team morale	<input type="radio"/>	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?

Page Break

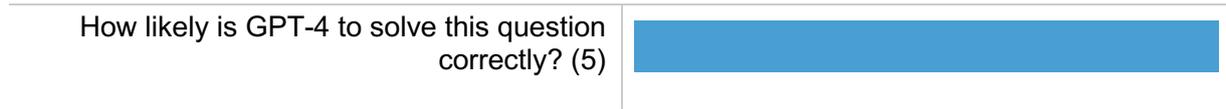
GenAICalPre0 **Rate how likely you think it is that ChatGPT will give a correct answer to the following prompts.**

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.

ExtremelySomewhat Neither SomewhatExtremely I don't
unlikely unlikely likely likely likely know
nor
unlikely

0 10 20 30 40 50 60 70 80 90 100



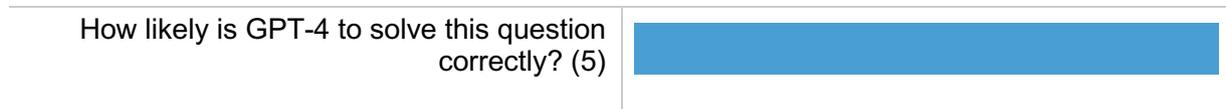
GenAICalPre2 Here is some data about Australian cities that I copied from Wikipedia. Based on this data, which cities had an odd-numbered population in 2011?

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

ExtremelySomewhat Neither SomewhatExtremely I don't
 unlikely unlikely likely likely likely know
 nor
 unlikely

0 10 20 30 40 50 60 70 80 90 100



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?

ExtremelySomewhat Neither SomewhatExtremely I don't
 unlikely unlikely likely likely likely know
 nor
 unlikely

0 10 20 30 40 50 60 70 80 90 100

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?

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? (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.

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? (5)



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.

ExtremelySomewhat Neither SomewhatExtremely I don't
 unlikely unlikely likely likely likely know
 nor
 unlikely

0 10 20 30 40 50 60 70 80 90 100

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



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?

ExtremelySomewhat Neither SomewhatExtremely I don't
 unlikely unlikely likely likely likely know
 nor
 unlikely

0 10 20 30 40 50 60 70 80 90 100

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



Page Break



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.

TrainerGPTVideo 1. Please start by watching the 2 videos below
Getting ChatGPT to do what you want

Signing Into ChatGPT



Page Break

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

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.

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

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

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

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:

“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.

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

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!

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

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

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

*What is Stack Overflow: <https://youtu.be/Eose6jHWC4U>
Stack Overflow Question: <https://youtu.be/na53nc4l5Pw>*

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.

Other Resources: <https://youtu.be/WCOQ16SI6FQ>

Further Help and Documentation

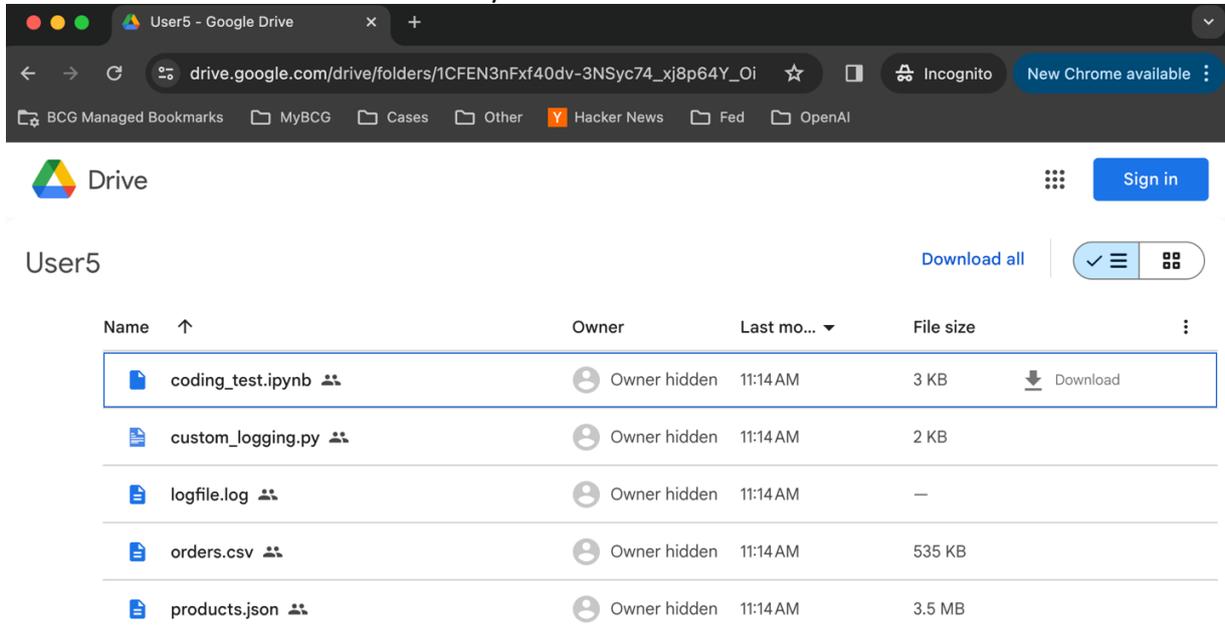
Key Resources:

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** (<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.

Setup Google Drive folder and Colab for Data Science Task

Visit Folder Link that was Shared with you:



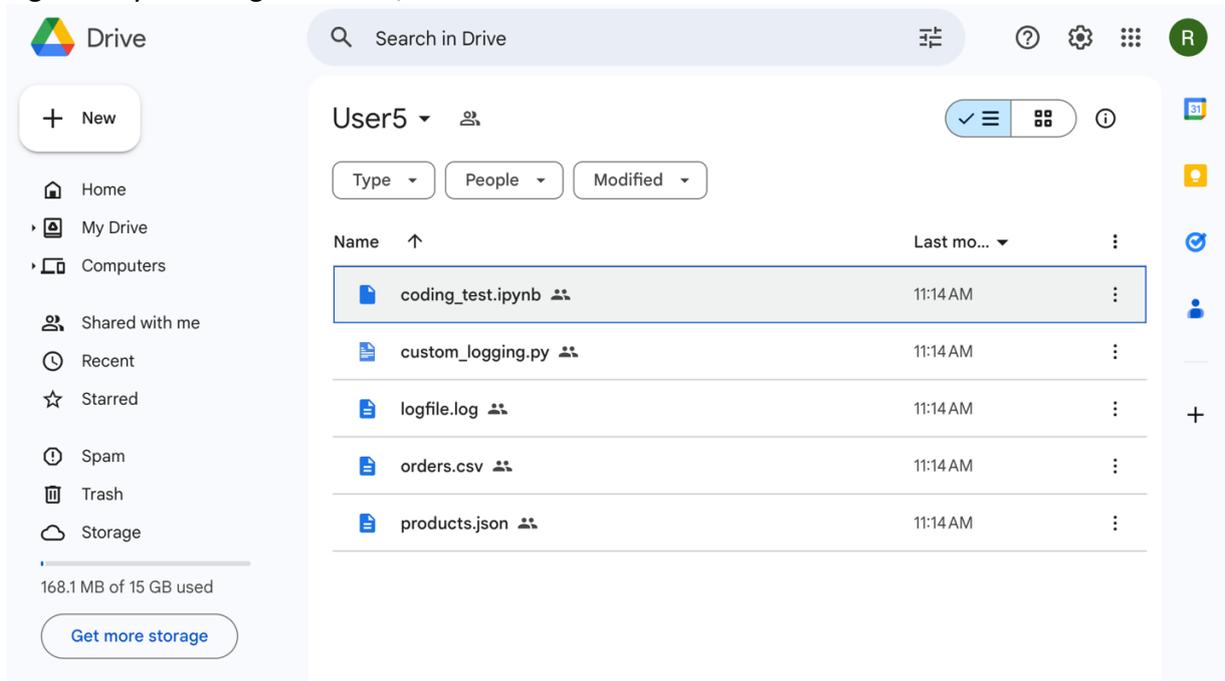
drive.google.com/drive/folders/1CFEN3nFxf40dv-3NSyc74_xj8p64Y_Oi

Drive Sign in

User5 Download all

Name	Owner	Last mo...	File size
coding_test.ipynb	Owner hidden	11:14 AM	3 KB
custom_logging.py	Owner hidden	11:14 AM	2 KB
logfile.log	Owner hidden	11:14 AM	—
orders.csv	Owner hidden	11:14 AM	535 KB
products.json	Owner hidden	11:14 AM	3.5 MB

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



Drive Search in Drive

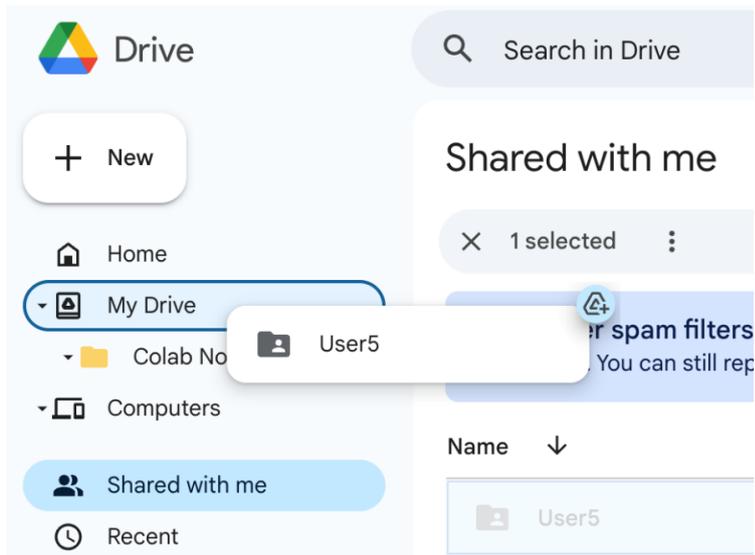
User5

Type People Modified

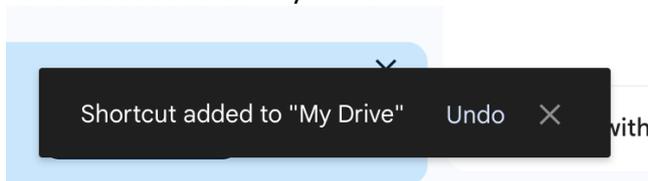
Name	Last mo...
coding_test.ipynb	11:14 AM
custom_logging.py	11:14 AM
logfile.log	11:14 AM
orders.csv	11:14 AM
products.json	11:14 AM

168.1 MB of 15 GB used Get more storage

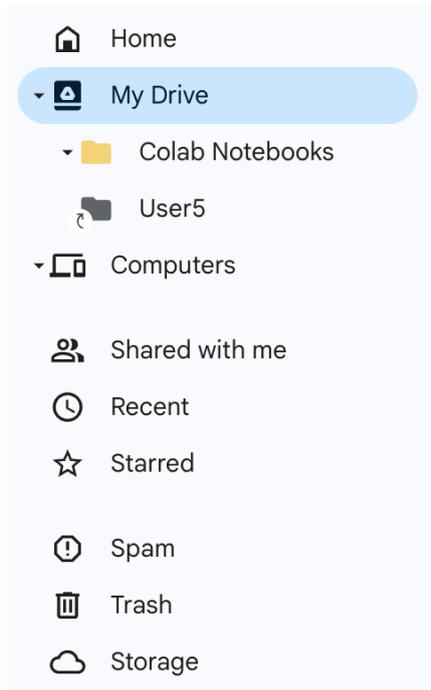
- Select your “Shared with Me” folder on the left navbar.
- In here you will find the folder that was shared with you. The title of the folder should be your email.
- Drag the folder into “My Drive” on the left navbar, as shown in screenshot.



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



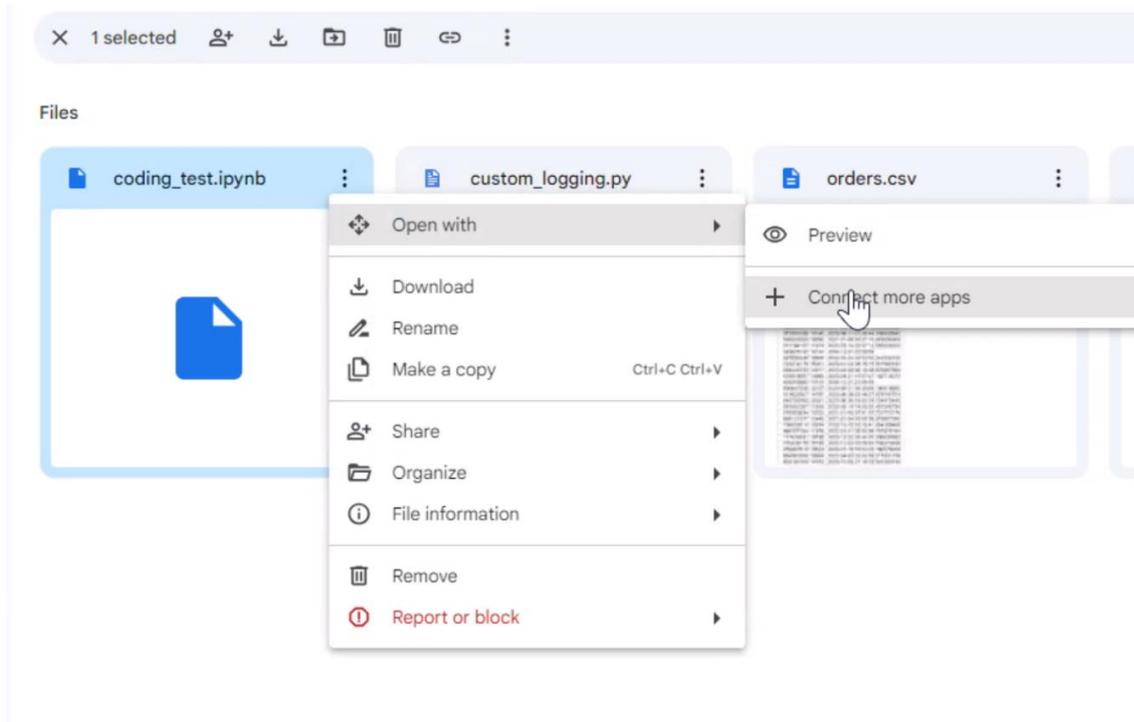
Then navigate to “My Drive” on the left navbar, to verify that you can find the folder that you just copied in.



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.

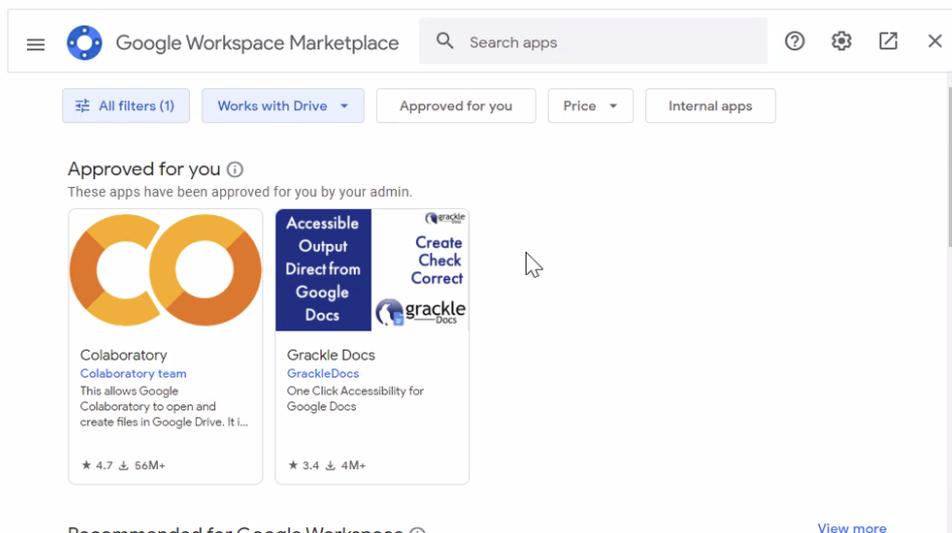
Installing Colab



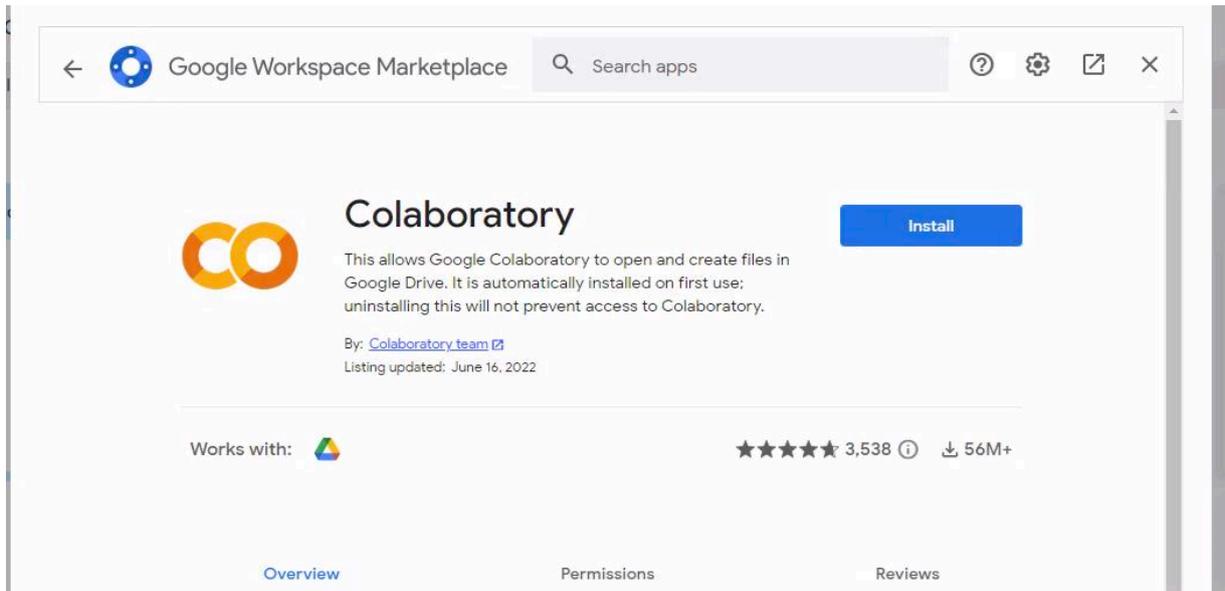
Click on the three dots on the top right of the “coding_test.ipynb” file.

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.



Recommended for Google Workspace
Select “Install”



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

Run Setup Code in coding_test.ipynb

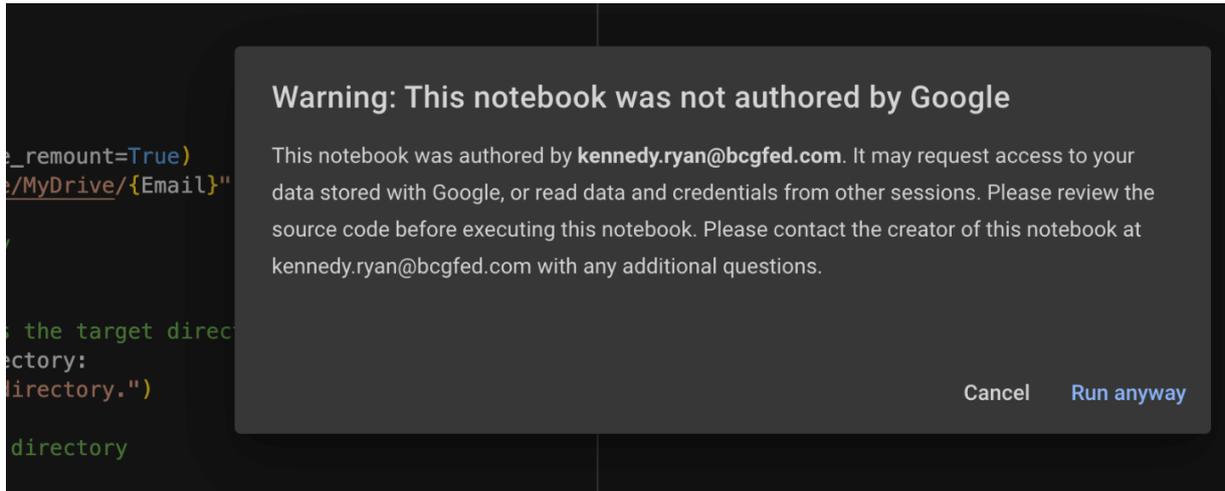
There is a section in the “coding_test.ipynb” called “Setup Logging Before the Task”. Make sure to follow the instructions to update your email in the code provided, and run the cell.

Important Note on Google Drive Access

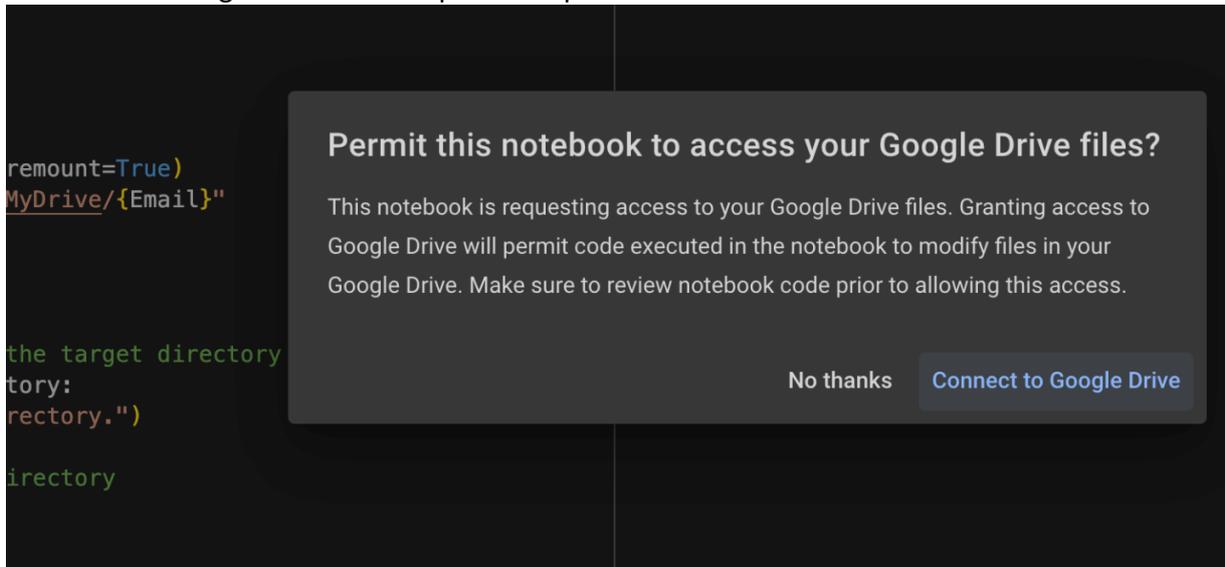
This Task requires reading in files and writing out to files from the notebook. In order to read files and write to files in a Colab notebook, users have to give the Colab notebook "Access to All Files" of their Google Account. We understand some users may be uncomfortable with this, so we want to clarify :

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.

Select what Google Drive for desktop can access

Select all

- See, edit, create, and delete all of your Google Drive files. [Learn more](#)
- View the photos, videos and albums in your Google Photos. [Learn more](#)
- Retrieve Mobile client configuration and experimentation. [Learn more](#)
- View Google people information such as profiles and contacts. [Learn more](#)
- View the activity record of files in your Google Drive. [Learn more](#)
- See, edit, create, and delete any of your Google Drive documents. [Learn more](#)

Because you're using Sign in with Google, Google Drive for desktop will be able to

- Associate you with your personal info on Google
- See your primary Google Account email address

Make sure you trust Google Drive for desktop

You may be sharing sensitive info with this site or app. You can always see or remove access in your [Google Account](#).

Learn how Google helps you [share data safely](#).

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

-  Home
 -  Personal info
 -  Data & privacy
 -  Security
 -  People & sharing
 -  Payments & subscriptions
-
-  About

Your connections to third-party apps & services

Keep track of your connections to third-party apps and services

-  Adobe
-  airSlate App
-  Atlassian
- +14 more

[See all connections](#)

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

← Third-party apps & services

Keep track of your connections

You shared data with these third-party apps and services. [Learn more](#) 

17 total apps & services

Filter by: 

Sign in with Google (14)

Access to Any account access (4) 

Linked account (0)

-  Chromecast >
-  Google Cloud SDK >
-  Google Drive for desktop >

3. Select “See Details” on the bottom right

Google Drive for desktop has some access to your Google Account

To use some Google Drive for desktop features, you gave Google Drive for desktop some access to your Google Account. This access might include sensitive info.

-  See your profile info
-  See, edit, create, and delete all of your Google Drive files
-  View the photos, videos and albums in your Google Photos
- +4 more

[See details](#)

4. Select “Remove access” on the bottom right

← [Google Drive for desktop has some access to your Google Acc...](#)

Access you've given to Google Drive for desktop

Google Drive for desktop has some access to info in your Google Account, including info that might be sensitive. You can always remove this access. [Learn about the risks](#)

Access given on: February 15, 4:06 PM

Web address: https://support.google.com/drive/?p=file_stream

Google Drive for desktop can:	 See your profile info	▼
	 View the photos, videos and albums in your Google Photos	▼
	 See, edit, create, and delete all of your Google Drive files	▼
	 See, edit, create, and delete any of your Google Drive documents	▼
	 View the activity record of files in your Google Drive	▼
	 Retrieve Mobile client configuration and experimentation	▼
	 View Google people information such as profiles and contacts	▼

If you remove access

You might not be able to use some Google Drive for desktop features

[Remove access](#)

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

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.

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.

If you have tried everything and you still see an error, please reach out to Ryan Kennedy (kennedy.ryan@bcgfd.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

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 _____

CodingTaskTimer Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)



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?

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

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

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



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.



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,**

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

TimerStats Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)



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

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.

**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**

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**

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

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.



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)
-

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)
-

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)
-

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?

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%.

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



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.



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

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

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



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

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](#)



PSUpload **Upload your csv or Excel file here:**

Page Break



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.

(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

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

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

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

0 10 20 30 40 50 60 70 80 90 100

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 ()



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

0 10 20 30 40 50 60 70 80 90 100

<p>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) ()</p>	
<p>How likely is GPT-4 to solve this question correctly?</p> <p>0 10 20 30 40 50 60 70 80 90 100</p>	
<p>Who lost the Super Bowl two years after Pan-Am filed for bankruptcy? ()</p>	
<p>How likely is GPT-4 to solve this question correctly?</p> <p>0 10 20 30 40 50 60 70 80 90 100</p>	
<p>Write out the word "hello" as an ascii art drawing with # and _ ()</p>	
<p>How likely is GPT-4 to solve this question correctly?</p> <p>0 10 20 30 40 50 60 70 80 90 100</p>	
<p>What is the best next move for O in the following game of Tic Tac Toe?</p> <pre> - . O ----- . O X ----- X . X </pre>	

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

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

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

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

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

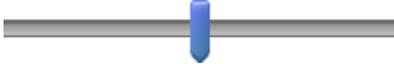
	1	2	3	4	5	6	7	I don't know
Brainstorm ideas	<input type="radio"/>							
Writing code for data analytics	<input type="radio"/>							
Writing code for data visualizations	<input type="radio"/>							
Learning how to use excel for data analysis and visualizations	<input type="radio"/>							
Identifying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

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

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?

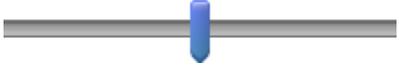
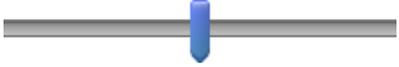
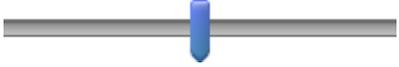
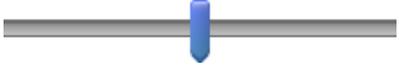
- Most people can be trusted
- You need to be very careful in dealing with people
- Don't know

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	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ChatGPT can be trusted to do quantitative analysis for you	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ChatGPT can be trusted to clean data for you with minimal guidance	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ChatGPT can be trusted to help you learn to do new things (e.g. use a new type of software)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

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 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!

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 $3 \times 0 = 0$ and so it will have 100 tokens;

if you decide to pass 50 tokens, we will give ChatGPT $3 \times 50 = 150$ tokens so it will have a total of $150 + 100 = 250$ tokens;

if you decide to pass 100 tokens, we will give ChatGPT $3 \times 100 = 300$ 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.

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 - $[3 * (\text{how much you decided to send})] + 100$)

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.

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 `gpt-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
 - o 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
 - o 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
 - o 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
 - o 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.
 - o 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.

LLM Grading for Statistics

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

- Preprocessing
 - o 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
 - o 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
 - o 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 stratified 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
 - o The coding 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
 - o 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
 - o 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

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
 - o 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
 - o 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)
 - o Team-based analysis: Participants create features representing team strengths or statistics based on historical data and incorporate those into predictions for each match
 - o 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)
 - o 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)
 - o Accuracy validation: Participants conducted simple metric measurement of accuracy to assess the performance of their methodology and made changes accordingly, to improve performance
 - o 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
 - o Linear regression on difference of goals scored by each team: 8 points
 - o Random forest regression on goal difference: 9 points
 - o XGBoost regression on goal difference: 9 points
 - o 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
 - o Linear regression on difference of goals scored by each team: 8 points
 - o Linear regression on numerical definitions of classes: 6 points
- Probability for predictability definition
 - o Linear regression on difference of goals scored by each team: 5 points
 - o Ensemble regression on goal difference: 5.5 points
 - o 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
 - o 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).