

# Data C100/200, Midterm

Fall 2024

Name: \_\_\_\_\_

Email: \_\_\_\_\_@berkeley.edu

Student ID: \_\_\_\_\_

Name and SID of the person on your right: \_\_\_\_\_

Name and SID of the person on your left: \_\_\_\_\_

## Instructions:

This exam consists of **66 points** spread out over **7 questions** and the Honor Code certification. The exam must be completed in **110 minutes** unless you have accommodations supported by a DSP letter.

Note that you should **select one choice** for questions with **circular bubbles**. There is always at least one correct answer. Please **fully shade** in the circle to mark your answer. For all math questions, **please simplify your answer**. Please also **show your work** if a large box is provided. For all coding questions, you may use commas and/or one or more function calls in each blank.

For all Python questions, you may assume `Pandas` has been imported as `pd`, `NumPy` has been imported as `np`, the Python `RegEx` library has been imported as `re`.

**You MUST write your Student ID number at the top of each page.**

## Honor Code [1 Pt]:

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I am the person whose name is on the exam, and I completed this exam in accordance with the Honor Code.

Signature: \_\_\_\_\_

This page has been intentionally left blank.

# 1 Using Data to Learn Who Asked [5 points]

DATA C100 staff just finished hosting a hackathon that was open to the general public. They collected a dataset, `participant_info`, containing **all** hackathon participants' information.

Each participant gets exactly one row describing them in the dataset, with the following columns:

- `participant_id`: The participant's ID. Each participant has a randomly assigned unique ID. Each row of the dataset corresponds to one unique participant ID. (type = `numpy.int64`)
- `first_name`: The first name of the participant. (type = `str`)
- `last_name`: The last name of the participant. (type = `str`)
- `project_topic`: The project topic that the participant is interested in. Project topics can be one of: "AI", "Education", or "Health". (type = `str`)

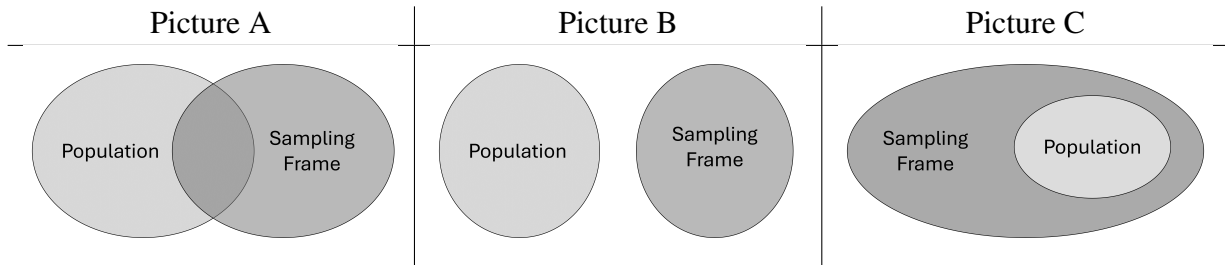
A sample of `participant_info` is shown below:

	<code>participant_id</code>	<code>first_name</code>	<code>last_name</code>	<code>project_topic</code>
0	4	James	Geronimo	AI
1	14	Yewen	Xu	Health
2	17	Julianna	Lee	AI
3	18	Vladyslav	Shevkunov	Health
4	19	Ella	Hammond	Education

**Note:** You can assume that at least one but not all UC Berkeley students and at least one non-UC Berkeley student participated in this hackathon.

- (a) [1 Pt] Select the correct statement regarding the dataset `participant_info` from the following choices:
- The column `participant_id` can serve as a primary key for this dataset.
  - There exist several columns that can serve as unique identifiers of each row, such as `participant_id` or `first_name`.
  - If 30% of the data from `project_topic` column is missing, it would be best to drop rows with missing values rather than imputing them.
  - The granularity of this dataset is such that each row represents one unique project topic.

- (b) [1 Pt] Xiaorui is interested in learning what hackathon participants think about the dinner food quality. He distributed a survey about dinner food quality to participants when he happened to be on campus from 8 pm to 9 pm, right after the dinner event. Which of the following is an accurate description of this sampling method?
- The sampling method guarantees a representative sample of its population.
  - The sampling method guarantees to prevent response bias from its respondents.
  - This method is an instance of convenience sampling.
  - This method is an instance of simple random sampling with replacement.
- (c) [1 Pt] Xiaorui changes the sampling method for the survey described in the previous subpart. This time, he distributed this survey to all UC Berkeley students the day after the dinner event. Which of the following pictures best describes the relationship between the sampling frame and the population of this survey?



- Picture A
  - Picture B
  - Picture C
- (d) [1 Pt] Suppose we want to use a visualization to capture the exact number of participants involved in each project topic. Which of the following is a suitable type of visualization?
- Hexplot
  - Contour Plot
  - Barplot
  - Histogram
- (e) [1 Pt] Evaluate the correctness of each of the following descriptions regarding `participant_info` dataframe.
- True  False The variable `participant_id` is a quantitative variable.
  - True  False The variable `project_topic` is a qualitative ordinal variable.

## 2 Submitting pandas Code at 11 : 59 : 59 PM [16 points]

To assess the traffic that the hackathon website observed during the hackathon, Sarah collected the dataset, `portal_traffic`, with the following columns:

- `visitor_id`: The ID of the website visitor. Hackathon participants have an ID that begins with "100", and judges have an ID that begins with "101". A visitor is either a hackathon participant or a judge, but not both. (type = str)
- `hour_of_visit`: The hour that the user activity occurred at; Values must be an integer within [0, 23]. (type = np.float32)
- `activity`: The activity of the visit. The activity can be one of: "submit", "visit", or "grade". (type = str)

A sample of `portal_traffic` is shown below:

	<code>visitor_id</code>	<code>hour_of_visit</code>	<code>activity</code>
0	101-Sammie-10053	4.0	grade
1	100-James-10258	NaN	submit
2	101-Willy-12930	8.0	visit
3	101-Nehal-12100	2.0	visit
4	100-Brie-11003	19.0	visit

- (a) [2 Pts] Shreya wants to create a DataFrame called `portal_traffic_filtered`, such that there are **only** records from hackathon participants in `portal_traffic`. The resulting `portal_traffic_filtered` should include all columns from `portal_traffic` for such activities. Write one line of pandas code that achieves this task.

You should use at least one of the following variables when constructing your solution:

```
mask_1 = portal_traffic["visitor_id"].str.contains("100")
mask_2 = portal_traffic["visitor_id"].str.startswith("100")
mask_3 = portal_traffic["visitor_id"].str[:3].isin(["100"])
```

- (b) [1.5 Pts] Malavikha wants a DataFrame of only entries with the activity "visit" and the columns `visitor_id` **and** `hour_of_visit`. She decides to do this by running the following code:

```
df = portal_traffic.copy()
is_visit = df["activity"]=="visit"
_____A_____
```

For each option below, determine whether filling it into blank A outputs the desired DataFrame:

- True  False `df.loc[is_visit, [0, 1]]`
- True  False `df[is_visit].iloc[:, [0, 1]]`
- True  False `df[is_visit][["visitor_id", "hour_of_visit"]]`
- (c) Using `portal_traffic`, help Abby create a histogram with 24 bins for the count of visits per hour. Make the **title** of the resulting plot `visits_per_hour`. Assume that all missing entries are already dropped from the `hour_of_visit` column within the following code.

```
import matplotlib.pyplot as plt
plt._____A_____ (
    portal_traffic["hour_of_visit"].dropna(), bins=24
)
plt._____B_____
```

- (i) [0.5 Pts] Fill in blank A:

- (ii) [1 Pts] Fill in blank B:

- (d) Using `portal_traffic`, fill in the blanks to filter out the missing entries from the `hour_of_visit` column, and create a `Series` that contains the count of activities for each hour from `hour_of_visit`. Assume no missing entries are in the `activity` column.

```
(  
    portal_traffic[~_____A_____._____B_____]()  
    ._____C_____  
    ._____D_____["activity"]  
)
```

- (i) [1 Pts] Fill in blank A:

- (ii) [1 Pts] Fill in blank B by selecting the correct choice below:

- `isnull`  
 `dropna`  
 `fillna`

- (iii) [1 Pts] Fill in blank C:

- (iv) [1 Pts] Fill in blank D:

- (e) Fill in the blanks below to create `less_than_three_grade`, a DataFrame containing only rows where its corresponding `hour_of_visit` has less than 3 grade activities.

```
def less_than_three_grading(df):  
    return df[df["activity"]=="grade"]._____A_____
```

```
less_than_three_grade = (  
    portal_traffic.groupby(_____B_____)._____C_____
```

```
)
```

- (i) [1 Pts] Fill in blank A:

- (ii) [1 Pts] Fill in blank B:

- (iii) [1 Pts] Fill in blank C:



- (f) Claire wants to impute the missing entries of `hour_of_visit` with the mode of its same activity category.

For example, if any row with activity "grade" has a missing value for `hour_of_visit`, the missing values should be replaced by the mode of `hour_of_visit` for "grade" activities. Fill in the blanks to accomplish this.

```
modes_of_each_group = (  
    portal_traffic.groupby(_____A_____)  
        .agg(lambda series: series.value_counts().index[0])  
)  
replacement_values = modes_of_each_group[  
    portal_traffic[portal_traffic["hour_of_visit"].isna()][ "activity"  
].values  
portal_traffic.loc[  
    _____B_____._____C_____._____D_____ ) , _____  
] = replacement_values
```

*Hint:* Below is `modes_of_each_group`, a Series where each row corresponds to the mode of `hour_of_visit` for each activity category. Note that the following values are for the entire `portal_traffic` dataframe:

```
activity  
grade    13.0  
submit    2.0  
visit     8.0  
Name: hour_of_visit, dtype: float64
```

- (i) [1 Pts] Fill in blank A:

- (ii) [1 Pts] Fill in blank B:

- (iii) [1 Pts] Fill in blank C:

- (iv) [1 Pts] Fill in blank D:

### 3 How The Tables Have Turned... [14 points]

To learn more about hackathon participants, Sarika collected another dataset called `other_info` with the following columns:

- `participant_id`: The participant's ID. Each participant has a randomly assigned unique ID. Each row of the dataset corresponds to one unique participant ID. (type = `numpy.int64`)
- `age`: The age of the participant. (type = `np.int64`)
- `level_of_profession`: Level of profession of the participant. (type = `str`)
- `email_address`: The email address of the participant. (type = `str`)

A sample of `other_info` is shown below:

	<code>participant_id</code>	<code>age</code>	<code>level_of_profession</code>	<code>email_address</code>
0	4	20	undergrad	james.geronimo@berkeley.edu
1	14	23	undergrad	yewen.xu@berkeley.edu
2	17	30	newgrad	julianna.lee@dataChundred.net
3	18	20	newgrad	vladyslav.shevkunov@dataChundred.net
4	19	25	newgrad	ella.hammond@dataChundred.net

(a) [1 Pt] Fill in the blank to create a `Series` with participants' email suffixes.

Note: The suffix of an email address is any text that comes after the first appearance of the character "@". For example, the suffix of "john.doe@berkeley.edu" is "berkeley.edu".

```
email_suffix = other_info["email_address"]._____A_____.str[1]
```

Fill in blank A:

- (b) Recall the dataset used in question 1, `participant_info`. For reference, a sample of this dataset is shown again below:

	<code>participant_id</code>	<code>first_name</code>	<code>last_name</code>	<code>project_topic</code>
0	4	James	Geronimo	AI
1	14	Yewen	Xu	Health
2	17	Julianna	Lee	AI
3	18	Vladyslav	Shevkunov	Health
4	19	Ella	Hammond	Education

where the `participant_id` column in `participant_info` matches the `participant_id` column in `other_info`.

Fill in the blanks to produce `sorted_merged_info`, a DataFrame where each row represents one participant with information from both `participant_info` and `other_info`. In addition, `sorted_merged_info` should be sorted by the last letter of the participant's first name in descending order and contain a new column `last_letter`.

```
participant_info.loc[:, "last_letter"] = (
    participant_info["first_name"].str.lower()._____A_____
)
sorted_merged_info = participant_info._____B_____ (
    _____C_____ # Feel free to use commas here.
)._____D_____
```

- (i) [1 Pts] Fill in blank A:

- (ii) [1 Pts] Fill in blank B:

- (iii) [1 Pts] Fill in blank C:

- (iv) [1 Pts] Fill in blank D:

(c) [1 Pt] Let `sorted_merged_info` be the correct result of the previous subpart.

Consider this code snippet:

```
sorted_merged_info.pivot_table(
    index="level_of_profession", columns="project_topic",
    values="age", aggfunc="median"
)
```

Which image below is the output of the above code snippet?

project_topic	level_of_profession	age
AI	gradstudent	27.0
	newgrad	24.0
	undergrad	20.0
Education	gradstudent	32.0
	newgrad	22.5
	undergrad	22.0
Health	gradstudent	32.0
	newgrad	27.0
	undergrad	21.0

Picture A

level_of_profession	gradstudent	newgrad	undergrad
AI	27.0	24.0	20.0
Education	32.0	22.5	22.0
Health	32.0	27.0	21.0

Picture B

project_topic	AI	Education	Health
gradstudent	27.0	32.0	32.0
newgrad	24.0	22.5	27.0
undergrad	20.0	22.0	21.0

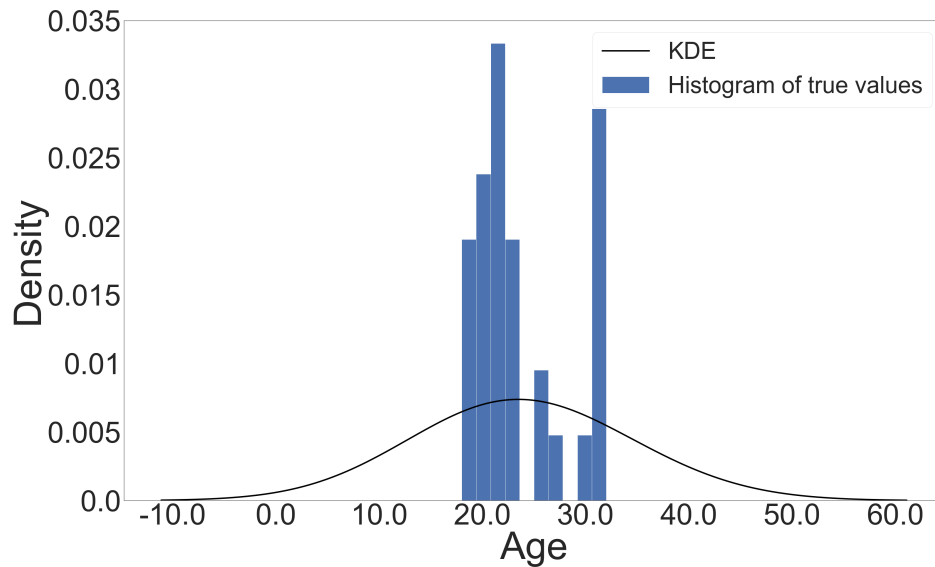
Picture C

- Picture A.  
 Picture B.  
 Picture C.

(d) [1 Pt] Rose is considering a variety of visualizations to use. For each of the following descriptions regarding visualizations, determine whether it is true or false.

- True  False In contour plots, contour lines represent datapoints with the same density.
- True  False The skew of a histogram refers to the ratio between its peak and number of bins.

- (e) [1 Pt] Rose generates the following KDE plot to estimate the distribution of participants' ages:



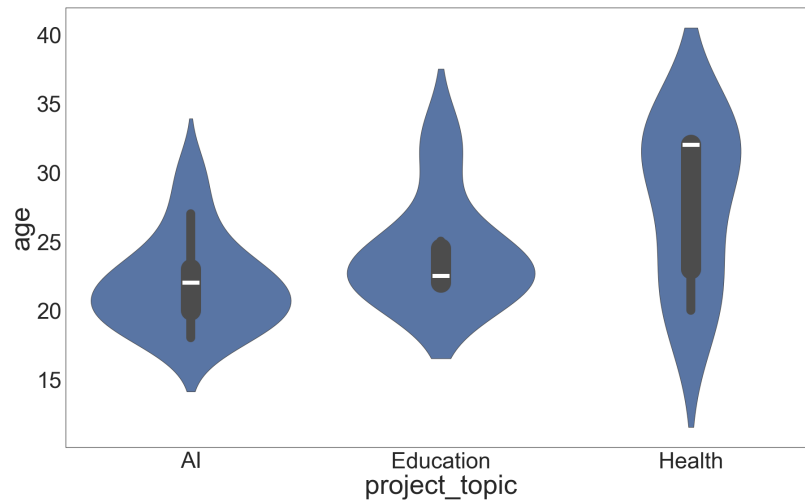
For each of the following descriptions, determine whether it is true or false.

- True  False The  $\alpha$  of this KDE plot is too high, resulting in a high estimated density for numeric values way below 18.
- True  False Based on the histogram, the true distribution of participant age is unimodal.

- (f) Angela noticed that side-by-side violin plots could help effectively visualize the variable age across participants of different project topics.

Recall `sorted_merged_info`, a `DataFrame` where each row represents one participant with information from both `participant_info` and `other_info`.

- (i) [2 Pt] Let `sorted_merged_info` be the correct result of subpart 3(b). Write one line of Python code to recreate the following visualization with the `seaborn` library:



Assume the following line of code has been run:

```
import seaborn as sns
```

- (ii) [1 Pts] For each of the following descriptions regarding the violinplots, determine whether it is true or false.

- True  False Violinplots show the mean of each variable's distribution.
- True  False Based on the violinplot, the 75th percentile of age for participants with project topic "Health" is higher than that for participants in project topic "Education".

- (g) [3 Pts] Provided three participants' age: (20, 22, 23), use a boxcar kernel of width  $\alpha = 2$  to perform KDE. What is the estimated density at age 22.5?

A boxcar kernel is formulated as follows:

$$B_{\alpha}(x, x_i) = \begin{cases} \frac{1}{\alpha}, & \text{if } -\frac{\alpha}{2} \leq x - x_i \leq \frac{\alpha}{2} \\ 0, & \text{otherwise} \end{cases}$$

**Grading will be done based on the work you show in the box below.**

Density at age 22.5 = \_\_\_\_\_

## 4 Pur+egex! [6 points]

To alleviate participants' stress, Aneesh decided to host cat-petting sessions during the hackathon. People provided feedback for this event, and Aneesh is trying to process this text data.

For subparts (a) and (b), you will be provided a table with the following format:

Match all of these below	Match none of these below
✓case 1	✗case 3
✓case 2	✗case 4

You will be asked to provide patterns that match strings in the left column after the ✓ and not match strings in the right column after the ✗. For example, for the above table, you should provide a pattern that matches case 1 and case 2 but not case 3 and case 4.

(a) Aneesh decides to do the following RegEx exercises before processing feedback.

- (i) [2 Pts] In the following blank, write a RegEx pattern that only matches words that contain the substring "cat" and does not contain any uppercase letters or space characters.

Fill in the blank with only the RegEx pattern, **do not make your solution a raw string**. A raw string is in the format of r"\_\_\_\_\_".

Match all of these below	Match none of these below
✓cats	✗Cat
✓concatenate	✗CA Tacoma
✓10cats10	✗CATcatCAT

- (ii) [1 Pts] Assume the correct answer to the previous question is stored in a raw string `pat`, and `cat_responses` is a Series of string objects. Write a line of pandas code that replaces any word that matches `pat` with "dog".



(b) [2 Pts] Here is the provided table:

Match all of these below

✓total cats petted:60,102,305  
✓my cats' names are:amy,baron,carol  
✓the current time is 20:59

For each of the following RegEx patterns, determine if it is true that they fully match all cases listed in the "Match all of these below" column.

- True  False `.*:\w+(,\w+)*`  
 True  False `.*:[A-Za-z\d]+(,[A-Za-z\d]+)*`  
 True  False `.*:\w+(,\w+)+`  
 True  False `\D*:\w+(,\w+)+`

(c) [1 Pt] Given the provided variables:

```
case = "Can you let the mouse go, Jordan?"  
pattern = r"\w?$"
```

What is the output of the following function call?

```
re.search(pattern, case)[0]
```

Select the correct option from below.

- "Jordan"  
 "n"  
 "n?"  
 ""

## 5 Spending a Night in Jupyter [9 points]

It is unhealthy to sleep for less than 8 hours. Regardless, several groups tried working on their projects overnight. Sam collected a dataset to investigate whether the number of hours worked overnight is influential to the project's final grade:

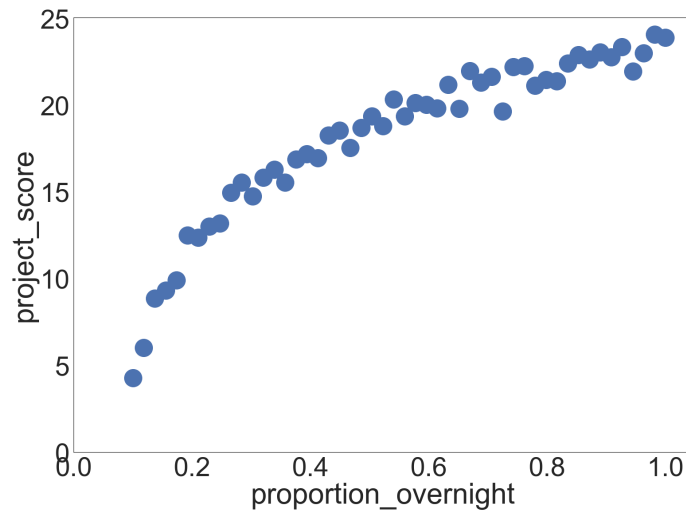
- `team_name`: The team name of the submitted project. (type = `str`)
- `project_score`: The final grade of the submitted project is out of 30 points. It must be a positive value. (type = `np.float32`)
- `hours_spent`: Number of hours the team spent on their project. It must be a positive value. (type = `np.float32`)
- `proportion_overnight`: The proportion of time spent between 12 AM and 8 AM on their project, between 0 and 1 inclusively. (type = `np.float32`)

A sample of the dataset is shown below:

	<code>team_name</code>	<code>project_score</code>	<code>hours_spent</code>	<code>proportion_overnight</code>
0	Pandey	24.0	10.5	0.30
1	Seabirth	18.0	9.0	0.35
2	Numthon	28.0	11.0	0.27

- (a) [1 Pt] Nikhil wants to use a constant model to predict `proportion_overnight`. What should Nikhil pay attention to before fitting the constant model? Select the correct description from below:
- For a constant model, the L1 loss is generally more sensitive to outliers than the L2 loss.
  - For a constant model, the optimal solution to L1 loss is always the mode of the predicted variable's actual values.
  - For a constant model, the optimal solution to L2 loss is always the mean of the predicted variable's actual values.
  - Instead of L1 loss, Nikhil can choose to use a loss function:  $\mathcal{L}(y, \hat{y}) = y - \hat{y}$ .

- (b) [2 Pts] Dan builds a linear regression model to predict `project_score` using the variable `proportion_overnight`, and made a scatterplot as shown below:



For each of the following combination of transformations, determine whether applying that combination alone would linearize the data:

- True  False Applying a cubic root to the `proportion_overnight` variable, and a logarithmic transformation to the `project_score` variable.
- True  False Raising the `proportion_overnight` variable to the third power, and a logarithmic transformation to the `project_score` variable.
- True  False Applying a cubic root to the `proportion_overnight` variable, and raising the `project_score` variable to the second power.
- True  False Raising the `proportion_overnight` variable to the third power, and raising the `project_score` variable to the second power.

(c) [2 Pts] We are provided the following view of our dataset:

	project_score	hours_spent
mean	22.0	8.0
variance	16.0	1.0

The correlation between `project_score` ( $y$ ) and `hours_spent` ( $x$ ) is 0.5. Using Simple Linear Regression (SLR), write the model's equation predicting `project_score` using `hours_spent`. **Grading will be done based on the work you show in the box below.**

The model equation: \_\_\_\_\_

(d) [1 Pt] Minoli proposes an alternative loss function for the linear regression problem between `project_score` ( $y$ ) and `hours_spent` ( $x$ ) as:

$$\mathcal{L}(y_i, \hat{y}_i) = \begin{cases} 3|y_i - \hat{y}_i|, & \text{if } y_i > \hat{y}_i \\ |y_i - \hat{y}_i|, & \text{otherwise} \end{cases}$$

Select the correct description regarding Minoli's proposed loss function.

- Minoli's loss function punishes underprediction and overprediction equivalently.
- Minoli's loss function punishes underprediction more than overprediction.
- Minoli's loss function is concave.
- The optimal solution to Minoli's loss function cannot be found.

- (e) [3 Pts] Minoli wants to fit a constant model with equation  $\hat{y}_i = \theta$ , using the loss function  $\mathcal{L}$  from part (d). What is the derivative of the following empirical risk function  $\mathcal{R}$  with respect to  $\theta$ ?

$$\mathcal{R}(\theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(y_i, \hat{y}_i)$$

Let  $y_+$  be the count of values  $y_i$  where  $y_i > \hat{y}_i$ , and  $y_-$  be the count of  $y_i$  where  $y_i \leq \hat{y}_i$ .

Express and simplify your final answer in terms of  $y_+$  and  $y_-$ . **Grading will be done based on the work you show in the box below.**

$$\frac{d\mathcal{R}}{d\theta} = \underline{\hspace{2cm}}$$

## 6 Ordinarily, or Legendarily Satisfied? [7 points]

Rayna uses the following dataset to train a linear regression model for hackathon participant satisfaction. All variables in the dataset are integers (type = `np.int64`) between 0 and 5 inclusively:

- `overall`: The overall satisfaction of a participant.
- `food`: The participant's satisfaction regarding the provided food.
- `booth`: The participant's satisfaction regarding external booths.

The **full dataset** is shown below:

	<code>overall</code>	<code>food</code>	<code>booth</code>
<b>Datapoint 0</b>	4	2	0
<b>Datapoint 1</b>	2	0	1
<b>Datapoint 2</b>	4	0	0

- (a) [2 Pts] Calculate the coefficients of an Ordinary Least Squares (OLS) model to predict `overall` with all the other features. Do not include a bias column in your design matrix. **Grading will be done based on the work you show in the box below.**

*Hint:*  $\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a} & 0 \\ 0 & \frac{1}{b} \end{bmatrix}$

$\hat{\theta}_{\text{food}} = \underline{\hspace{2cm}}, \hat{\theta}_{\text{booth}} = \underline{\hspace{2cm}}$

**We have now collected more datapoints and will use this larger dataset for the following subparts.**

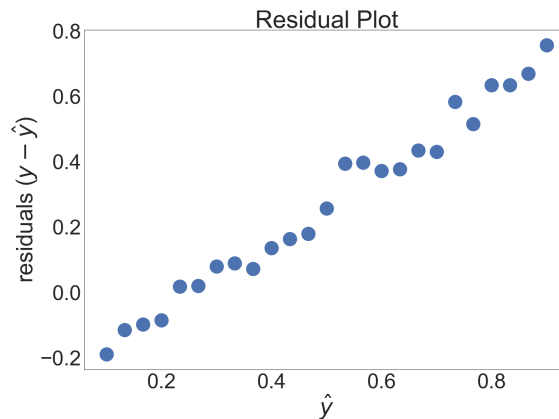
(b) [2 Pts] Rayna processes the dataset before fitting the model. For each suggestion, determine whether it is true or false that Rayna should perform it before applying OLS to the provided dataset.

- True  False If the design matrix is an  $n$ -by- $p$  matrix, then it must be that  $\mathbb{Y} \in \mathbb{R}^n$ .
- True  False Rayna should check if the resulting design matrix has more columns than rows. If so, OLS does not have a unique solution.
- True  False We must have  $\text{rank}(\mathbb{X}) = 1$  for OLS to have a unique solution.
- True  False Rayna should check if the rank of  $\mathbb{X}^T\mathbb{X}$  is equal to the rank of  $\mathbb{X}$ . If so, OLS cannot find a unique solution.

(c) [2 Pts] Rayna produces the OLS model after some work. For each option, determine whether the option holds true **only if** a **unique** OLS solution exists.

- True  False The dot product of residuals  $\vec{e}$  and all features is 0.
- True  False The design matrix  $\mathbb{X}$  has full column rank.
- True  False For the design matrix  $\mathbb{X}$ ,  $\mathbb{X}^T\mathbb{X}$  is invertible.
- True  False For the design matrix  $\mathbb{X}$ , the equation  $\mathbb{X}^T\mathbb{X}\theta = \mathbb{X}^T\mathbb{Y}$  holds.

(d) [1 Pt] Rayna generated a residual plot from her linear regression model, as shown below:



For each of the following descriptions regarding this plot, mark whether it is true or false.

- True  False The above plot shows a trend of overprediction for higher  $\hat{Y}$  values.
- True  False In an ideal residual plot, there should not be a particular pattern with systematic underpredictions or overpredictions.

## 7 What is The Optimal Amount of Free Food? [8 points]

We chose not to offer free food to reduce the cost of hosting a hackathon, but maybe that's not the most optimal choice. Jessica tries to answer this concern via an optimization process.

- (a) Jessica created the following loss function  $\mathcal{L}$ , which describes the overall operational cost of providing free food:

$$\mathcal{L}(\theta_f, \theta_c, \theta_s) = 25\theta_f\theta_c - \ln(\theta_s)$$

Here,  $\theta_f$ ,  $\theta_c$ , and  $\theta_s$  respectively represent the amount of free food, car transportation cost of food, and subsidies for food costs. For this subpart, **grading will be done based on the work you show in the boxes below.**

- (i) [3 Pts] Express the gradient for  $\mathcal{L}$  in terms of  $\theta_f$ ,  $\theta_c$ ,  $\theta_s$ . Please simplify your answer.

$$\nabla \mathcal{L} = [ \quad \quad \quad ]^T$$

- (ii) [3 Pts] Ignoring the previous part, Jessica found the following current values for parameters and gradient at iteration  $t$ :

$$\theta_f^{(t)} = 0.5, \theta_c^{(t)} = -0.5, \theta_s^{(t)} = 1$$
$$\nabla \mathcal{L}(\theta_f, \theta_c, \theta_s) = [0.5 \quad 1.5 \quad -0.5]^T$$

With a learning rate of  $\alpha = 2$ , calculate the value of each variable at iteration  $t + 1$  of gradient descent on  $\mathcal{L}$ .

$$\theta_f^{(t+1)} = \underline{\hspace{2cm}}, \theta_c^{(t+1)} = \underline{\hspace{2cm}}, \theta_s^{(t+1)} = \underline{\hspace{2cm}}$$

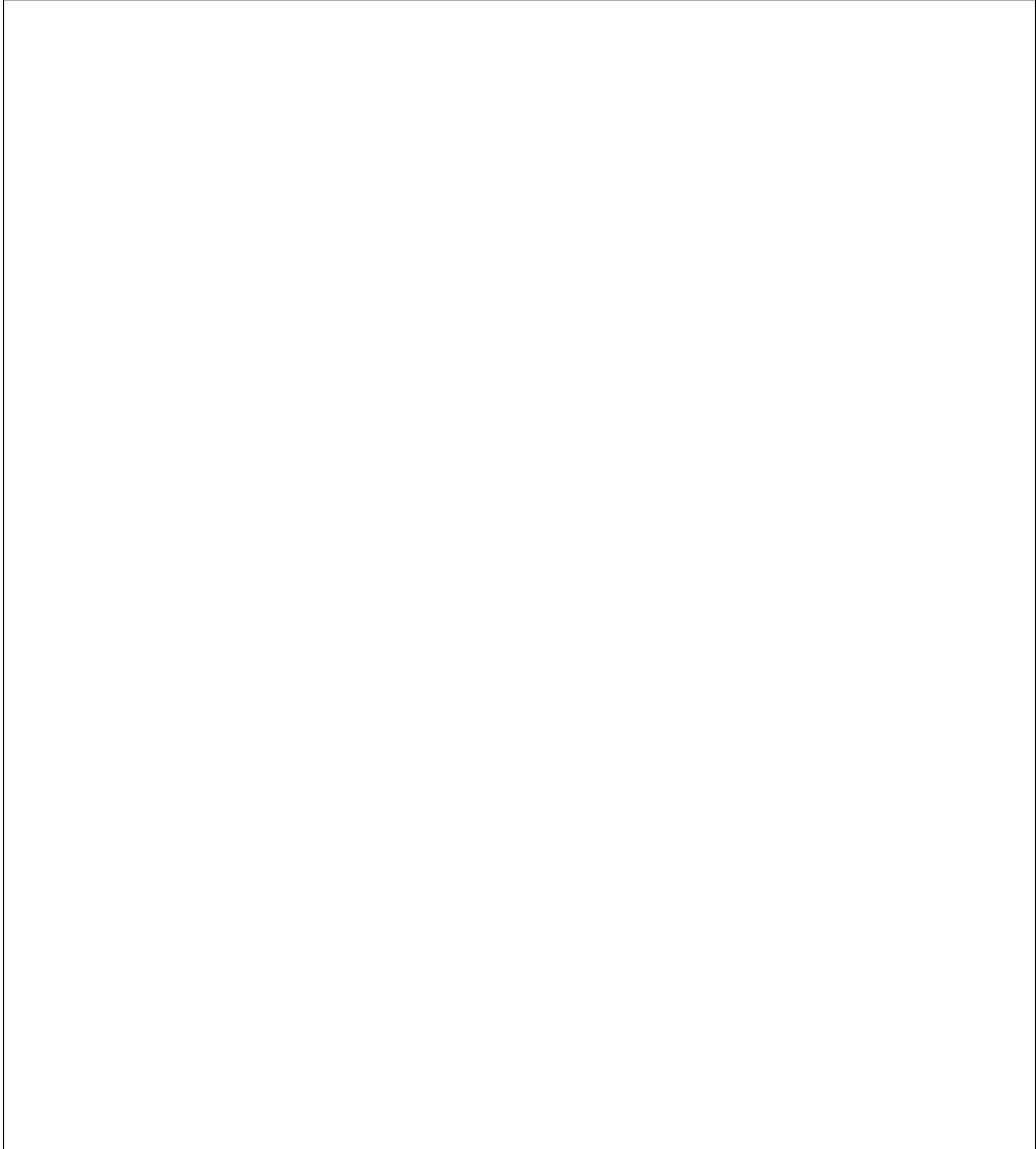


(b) [2 Pts] What possible aspect(s) of gradient descent may lead to a suboptimal solution? For each option, determine whether it is true or false.

- True  False    The starting point of the gradient descent algorithm.
- True  False    The choice of gradient descent (batch, mini-batch, or stochastic gradient descent).
- True  False    Using a fixed number of iterations.
- True  False    The learning rate of gradient descent.

**You are done with the midterm! Congratulations!**

Use this page to draw your favorite Data 100 moment!

A large, empty rectangular box with a thin black border, intended for the student to draw their favorite Data 100 moment.