DATA 8 Spring 2021 — Foundations of Data Science — Midterm

INSTRUCTIONS

This is your exam. Complete it either at exam.cs61a.org, or if that doesn't work, by emailing course staff with your solutions before the exam deadline.

This exam is intended for the student with email address nellepersson@berkeley.edu. If this is not your email address, notify course staff immediately, as each exam is different. Do not distribute this exam PDF even after the exam ends, as some students may be taking the exam in a different time zone.

For questions with **circular bubbles**, you should fill in exactly *one* choice.

○ You must choose either this option

○ Or this one, but not both!

For questions with **square checkboxes**, you may fill in *multiple* choices.

□ You could select this choice.

□ You could select this one too!

You may start your exam now. Your exam is due at [DEADLINE] Pacific Time. Go to the next page to begin.

For fill-in-the-blank coding questions, there will be a template for you to follow. You should copy and paste the provided template, then fill in the ___________ to answer the question. You can put anything inside the blanks, including commas, parentheses, and periods. Note that the length of the blank does not necessarily correspond to the length of the code you should write.

The exam is worth 100 points.

If you encounter any logistical problems during the exam, please contact us at data8berkeley@gmail.com. We will **not** be answering any questions related to the contents of the exam.

(a)

Your name:

(b) Your @berkeley.edu email address:

(c) Honor Code: *All work on this exam is my own.*

By writing your full name below, you are agreeing to this code:

(d) Important: You must **type** the following statement exactly into the box below. Failure to do so, or simply copy+pasting it, may result in points deducted on the exam.

“l certify that all work on this exam is my own. l acknowledge that collaboration of any kind is forbidden, and that l will face severe penalties if l am caught, including at minimum, harsh penalties to my grade and a letter sent to the Center for Student Conduct.”

1. (28.0 points) The Ultimate Ubiquitous Uniqlo

The `uniqlo` table has 3 columns:

- **Store Number**: The store number of a Uniqlo store
- **Item Number**: The item number of a given clothing piece sold at that store
- **Quantity**: The number of a particular clothing piece at the particular store.

Store Number	Item Number	Quantity
446	25	51
391	2	17
446	32	36
172	25	83
172	1	52

... (15042 rows omitted)

The names of the actual Uniqlo stores are contained in the `stores` table, which has one row per unique store and columns for the store number and corresponding store name.

Store Number	Store Name
1	Marronnier Gate Ginza
2	New York City 5th Ave.

... (6298 rows omitted)

The `items` table similarly contains information about each unique item number. It has columns for the item's corresponding clothing name, the type of clothing (either a top or a bottom) and the clothing style (either women's or men's). Assume every unique clothing name is available in both women's and men's styles.

Item Number	Clothing Name	Type	Style
1	Ultra Light Down Jacket	top	women
2	Ultra Light Down jacket	top	men
3	Crew Neck T-Shirt	top	men
4	Classic Straight Leg Jeans	bottom	men

... (5713 rows omitted)

(a) (1.0 pt) Write a line of code that outputs the number of unique (store, item) pairs that are possible.

(b) **(3.0 pt)** Katherine wants to combine the data from the three tables into one table that has five columns in the following order: Store Name, Clothing Name, Type, Quantity, and Style. Write a line of code to create this table and assign it to the name `combination`.

```
combination = __________
```

(c) **(3.0 pt)** Katherine wants to know how much space she is saving by using store numbers in the original `uniqlo` table instead of using the full store names. She writes the following partially completed code to calculate the total number of characters across all store names in all rows of `combination`. (e.g. the string "Mc Donald's" has 11 characters)

```
num_characters_store_names = ______________
```

Copy/paste the code above and fill in the blanks.

(d) **(4.0 pt)** Katherine wants to create a `ratio` function that takes in a store name to return the ratio of the number of tops to the number of bottoms at that store. She writes the following partially completed code:

```
def ratio(store_name):
  num_tops = ______________
  num_bots = ______________
  return ______________
```

Copy/paste the code above and fill in the blanks. You may use the `combination` table in your solution.

(e) **(3.0 pt)** Katherine wants to use `combination` to create a `tops_bots` table that has a column for store names, a column for tops, and a column for bottoms. She wants each row to contain the median quantity of tops and bottoms for each unique store. Fill out the following code to create and assign this table.

```
tops_bots = __________
```

(f) (3.0 pt) Katherine wants to create a `new_ratio` function that takes in a store name to calculate the ratio between the number of median tops and median bottoms at the store. She writes the following partially completed code:

```
def new_ratio(store_name):
    tops_bots_store = ____________________
    return __________________________________
```

Copy/paste the code above and fill in the blanks. You may use the `top_bots` table in your solution.

(g) (3.0 pt) Using the function you just wrote, create a new two-column table called `top_to_bot` that has one row for each store name and a column called "Tops to Bottoms" with the median top to bottom ratio.

```
top_to_bot = ______________________
```

(h) (1.0 pt) Write a line of code to visualize the distribution of median top to bottom ratios.

(i) **(3.0 pt)** Write a line of code to return the unique clothing name of a specific style that has the highest total quantity across all Uniqlo stores.

(j) **(4.0 pt)** Katherine wants to create an array containing the differences between the average quantity of women's and men's styles for each unique clothing piece in the `combination` table. She writes the following partially completed code:

```
all_averages = __________
____________
```

Copy/paste the code above and fill in the blanks.

2. (12.0 points) Programming in Python

For each question below, write Python code to answer the question **using what we have taught you in this class**. If we run your Python code, it should evaluate to the answer to the question.

(a) **(2.0 pt)** Fill in the following code to create a 5 element array such that `np.count_nonzero` and `sum` called on the array will return the same value.

```
make_array(__________)
```

(b) **(2.0 pt)** Fill in the following code to create a 5 element array such that `np.count_nonzero` and `sum` called on the array will return different values.

```
make_array(__________)
```

(c) **(2.0 pt)** Using only one line of Python code, construct an array called `squared`$_{i}ncrementsthatcontainsthesuccessiveinc$ $1^2, 2^2, 3^2, , 10^2$ `For example, the increment from` 1^2 `to` 2^2 `is 3.`

(d) `(2.0 pt) Consider an array, called banking, that contains the values of daily deposits to a bank account. Write a single line of Python code that constructs an array that contains the successive daily balances of the bank account after each deposit. The bank account's starting balance is $29,690.`

(e) `(2.0 pt) To which function is the mystery function below equivalent?`

```
def mystery(arr):
    a = 0
    for item in arr:
        if item != 0:
            a = a + 1
    return a
```

(f) (2.0 pt) The following function foo takes in an array as an argument. What does the following function do?

```
def foo(input_arr):
    arr = make_array()
    length_input = len(input_arr)
    for index in np.arange(length_input):
        arr = np.append(arr, input_arr.item(length_input-index-1))

    return arr
```

3. (11.0 points) WandaVision vs Queens Gambit

Alvin surveyed Data 8 staff members about how much time they spend on various streaming platforms each month. The table, `streaming`, has two columns, "service" and "time", and a few rows are shown below. Each TA was asked for their watch time (in hours) for Disney+ and Netflix, and then the table was reordered to preserve anonymity. There are 35 staff members recorded in the sample, and 2 streaming services, so there are a total of 70 rows.

service	time
Netflix	1
Netflix	11
Disney+	5
Netflix	8
Disney+	3
Disney+	14
Disney+	16
Netflix	16

... (62 rows omitted)

Alvin wants to know if staff members watch Netflix more than they watch Disney+. Alvin asks Sophia to help him to test this claim.

Define a null and alternative hypothesis for this test.

(a) (2.0 pt) Null Hypothesis:

(b) (2.0 pt) Alternative Hypothesis:

(c) **(3.0 pt)** Which of the following could be valid simulated test statistics to test these hypotheses? Select all that apply.

□ The median value of Netflix stream time - the median value of Disney+ stream time in the `streaming` table

□ The number of TAs in the `streaming` table who watched Netflix more than they watched Disney+, minus 10

□ The number of TAs in a copy of `streaming` with shuffled labels who watched Netflix more than they watched Disney+

□ The number of TAs in a copy of `streaming` with shuffled labels who watched Netflix more than they watched Disney+, minus 10

□ The median value of Netflix stream time - the median value of Disney+ stream time in a copy of `streaming` with shuffled labels

□ The number of TAs in the `streaming` table who watched Netflix more than they watched Disney+

(d) **(3.0 pt)** Which of the following are valid ways of shuffling labels for this hypothesis test? Note: `np.random.choice(arr, n, replace=False)` returns n samples from arr without replacement. Select all that apply.

□
```
shuffled_table = streaming.sample()
```

□
```
shuffled_table = streaming.sample(with_replacement=False)
```

□
```
shuffled_labels = streaming.sample().column(time)
shuffled_table = streaming.with_column(time, streaming.column(time))
```

□
```
shuffled_labels = streaming.sample(with_replacement=False).column(time)
shuffled_table = streaming.with_column(time, streaming.column(time))
```

□
```
shuffled_labels = streaming.sample(with_replacement=False).column(service)
shuffled_table = streaming.with_column(service, shuffled_labels)
```

□
```
shuffled_column = streaming.take(np.random.choice(np.arange(streaming.num_rows), 35, replace=Fa
shuffled_table = streaming.with_column(time, shuffled_column)
```

□
```
shuffled_table = streaming.sample(with_replacement=False)
disney_plus_rows = shuffled_table.take(np.arange(34)).select(time)
netflix_rows = shuffled_table.take(np.arange(35, 70)).select(time)
```

(e) **(1.0 pt)** Can Alvin use his hypothesis test to make claims about Data 8 students?

○ Yes, because the experiment is not making any claims about cause and effect, so it doesn't need to be a randomized control trial.

○ Yes, because Data 8 staff are a random sample.

○ No, because the experiment is not a randomized control trial.

○ No, because Data 8 staff are a convenience sample and might not represent Data 8 students.

4. (10.0 points) TikTok Data 8 O'Clock

The Data 8 team has just launched a TikTok account.

You may tackle each of Parts I and II in either order.

(a) Part I

i. Four members of the staff — Brett, Gregory, Kanika, and Sarah — have auditioned to perform an activity in a promotional TikTok video for Data 8. Prof. Sahai will select **exactly one** of the four auditioners at random (with equal probability). The selected auditioner then selects **exactly one** activity at random (with equal probability) — from among *acting*, *dancing*, and *singing* — to perform in the video.

For each of the following problems, express your answer as a Python expression (e.g., 1/8 + 1/9).

If there's not enough information to answer the question, write "N/A"

A. (1.0 pt) Determine the probability that Sarah is selected to perform, **and** they choose to sing in the video.

B. (1.0 pt) Determine the probability that Sarah is selected to perform, but they choose **not** to sing in the video.

C. (1.0 pt) Prof. Sahai did **not** select Kanika to perform in the video. Determine the probability that Sarah was selected.

(b) Part II

The course staff decide to release a TikTok video at the beginning of every week of the semester, for a total of fourteen (14) weeks. Each video has an 15% chance of going viral in **each** of the first 24 hours after its release. Thereafter, it has no chance of going viral.

i. (2.0 pt) Determine the probability that **none** of the fourteen videos goes viral within 24 hours of release.

ii. (2.0 pt) After seeing the Data 8 team transition to a position of social-media influence, the CS 61a course staff turn jealous and decide to pirate Data 8 TikTok videos.

There is a 9% chance that the CS 61a staff copy *three (3) or more* Data 8 TikTok videos, a 21% chance that they copy *exactly two (2)* Data 8 TikTok videos, and a 62% chance that they do **not** pirate (copy) any TikTok video from Data 8.

Determine the probability that the CS61a staff copy **exactly one** Data 8 TikTok video. Express your answer as a percent.

iii. (3.0 pt) To attract future students to Data 8, Babak commits to performing an entire lecture as Gollum, if *each* of the fourteen (14) Data 8 TikTok videos reaches the "For You" page, or if *at least one* of the Data 8 TikTok videos gets over ten million views.

Each video has a 40% chance of reaching the "For You" page. Moreover, and independently, each video has a 7% chance of getting over ten million views.

Determine the probability that Babak must deliver on his promise to conduct an entire lecture as Gollum.

5. (20.0 points) Dreams and Nightmares

Dream is the username of a famous Minecraft speedrunner. Minecraft speedrunners try to win the game as quickly as possible. Dream has been accused of modifying his version of the game to increase the probability that characters called "piglins" return items called "ender pearls" when the player initiates a trade, a "piglin barter", with the piglins.

One of the TAs, King, recorded data from six consecutive live streams of Dream playing Minecraft and found that 42 of the 249 piglin barters dropped ender pearls.

King is a Minecraft expert, so he knows that the true probability a piglin returns an ender pearl in the official version of Minecraft is 5%.

Provide a null and alternate hypothesis King can use to test if Dream's version of Minecraft was modified as described.

(a) (1.0 pt) Null hypothesis:

(b) (1.0 pt) Alternative hypothesis:

(c) (4.0 pt) Which of the following are valid test statistics? Variable names are descriptive. Select all that apply.

□ `42 - 0.05*249`

□ `num_ender_pearls_in_sample - 42`

□ `TVD(sample, observed)`

□ `num_ender_pearls_in_sample`

□ `proportion_ender_pearls_in_sample - 0.05`

□ `proportion_ender_pearls_in_sample`

□ `abs(num_ender_pearls_sample - 42)`

□ `abs(proportion_ender_pearls - 0.05)`

□ `42/249 - 0.05`

(d) **(6.0 pt)** Copy and paste the following code, then fill it in to perform the hypothesis test correctly. As a test statistic, use the number of **non**-ender pearls in a sample.

```
model_proportions = make_array(__________,__________)
test_stats = __________
for i in np.arange(10000):
    prop_non_ender_pearls = __________
    num_non_ender_pearls = __________
    test_stats = __________
observed_stat = __________
```

(e) **(2.0 pt)** Write a line of code to calculate the p-value of the test assuming `test_stats` has been defined.

```
p_value = __________
```

(f) **(4.0 pt)** King calculated a p-value of 0%. Which of the following conclusions can he draw? Select all that apply.

□ With a p-value cutoff of 1%, King fails to reject his null hypothesis (from part 1)

□ Because the study was observational, King cannot make any conclusions

□ With a p-value cutoff of 1%, King can reject his null hypothesis (from part 1)

□ It's statistically unlikely that the version of Minecraft Dream was playing had a 5% likelihood of piglin trades returning ender pearls

□ With a p-value cutoff of 1%, King rejects his alternative hypothesis (from part 2)

□ With a p-value cutoff of 1%, King accepts his alternative hypothesis (from part 2)

□ Dream did not cheat or modify his version of Minecraft

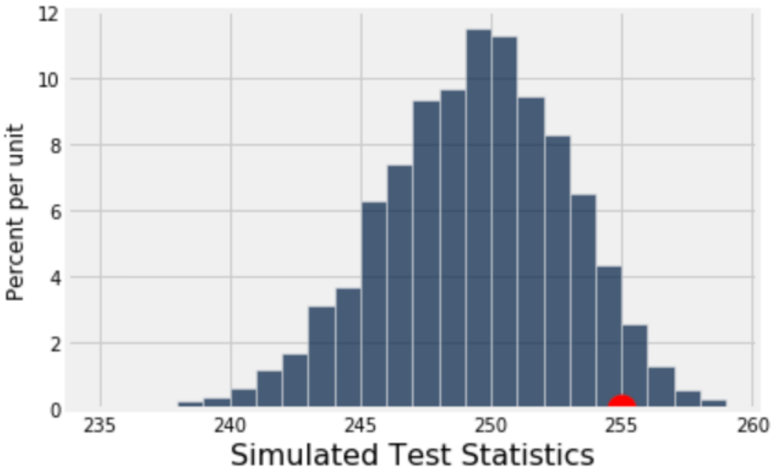
□ Dream cheated by modifying his version of Minecraft to increase the likelihood of piglin trades returning ender pearls.

(g) **(2.0 pt)** King ran another, **separate** hypothesis test on another aspect of Dream's speedrun data and generated the following distribution of test statistics (note: this is a completely separate hypothesis test, meaning the test statistic could be different than the ones before).

Which of the following are potential p values he would calculate for that test?

The observed value is plotted as the red dot on the x-axis and has a value of 255.

The bins are `np.arange(235, 260)`.



□ `2%`

□ `4.75%`

□ `0%`

□ `97.5%`

□ `99%`

□ `65%`

6. (13.0 points) Wall Street Bets

Brett is interested in investing his hard-earned money by buying stocks. His friend Sunny proposes that he does some research and data analysis before he believes all the deep dives that he read about on Reddit.

Right before the stock market closes, Brett collects the daily price data in the table `stocks`, where each row represents one company. Here is a snapshot:

Symbol	Open	High	Low	Close	Volume	Percent Change %
GME	128.17	151.53	127.50	137.74	30,733,670	0.407
BB	9.80	9.85	8.82	9.45	23,723,400	-0.0396
AMC	8.08	8.27	7.63	8.05	59,734,129	0.0025
AMD	79.00	79.48	74.20	78.52	58,548,910	0.0099
TSLA	626.06	627.84	539.49	597.95	89,396,461	-0.0378

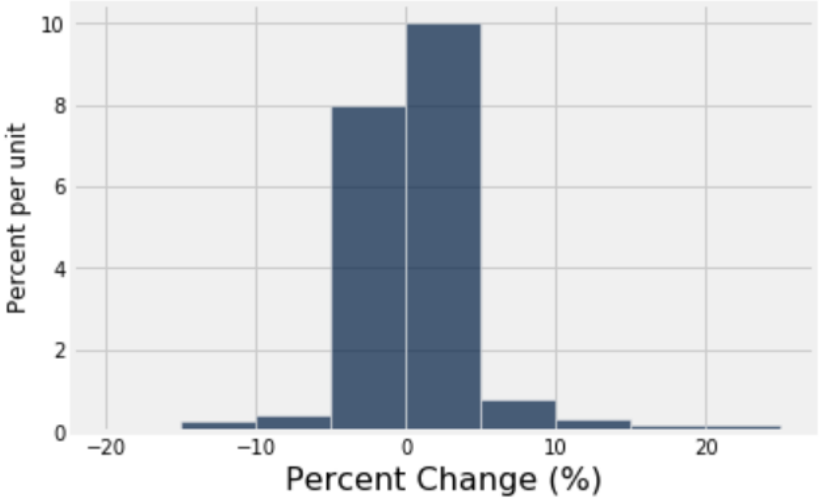
. . . (16767 rows omitted)

The table contains the following columns:

- **Symbol:** The abbreviation that denotes a specific company or index (such as the S&P 500 index, which consists of a basket of specific stocks)
- **Open/Close:** The price of one stock unit at the opening/closing of the market (US$)
- **High/Low:** The highest/lowest price of one stock unit during the day (US$)
- **Volume:** The number of stock units traded during the day
- **Percent Change:** The percentage change, from the the prior-day's close to the current-day's close (%)

Brett decides that the best way to analyze the riskiness of a company's stock is through finding the difference between the high and the low of the day, also known as the day's range.

Brett's other friend Natalie argues that he also needs to take into account the distribution of percentage change. The histogram looks as follows:



(a) **(1.0 pt)** Natalie wants to know the total percentage of companies for which the stock prices did not change by more than 5%. Select the percentage corresponding to this quantity.

○ 30%

○ 90%

○ 10%

○ 50%

(b) **(1.0 pt)** Sunny realizes that the majority of the companies fall within this small range. It is important that we make the bins smaller to get a better understanding. We want to divide the bin [0,5) up into three bins: [0, 1), [1, 3) and [3, 5).

Bin	% Within Original [0,5) Bin
[0,1]	25%
[1,3)	70%
[3, 5)	5%

Given the heights (A) 2.5, (B) 12.5, and (C) 17.5, order the heights from left to right to match the order of the 3 bins above.

○ A, C, B

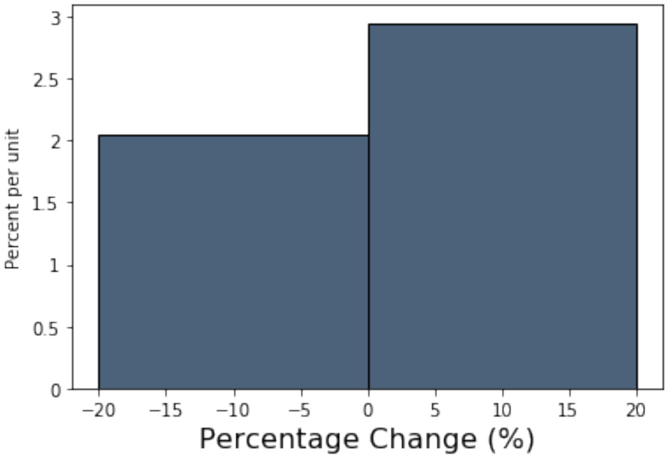
○ B, A, C

○ B, C, A

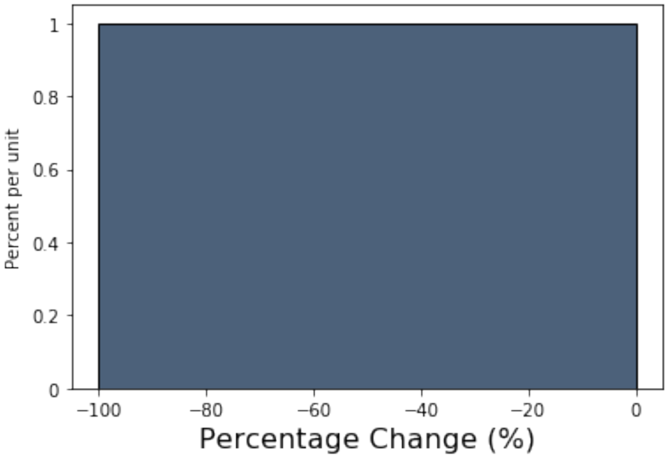
○ A, B, C

(c) **(1.0 pt)** Kevin tries to help them create a new histogram with the smaller bins, but he has never taken

Data 8. He messes up the histogram by incorrectly specifying the bins argument. Which of the following graphs could he have generated, given the data above? Select all that apply.



□



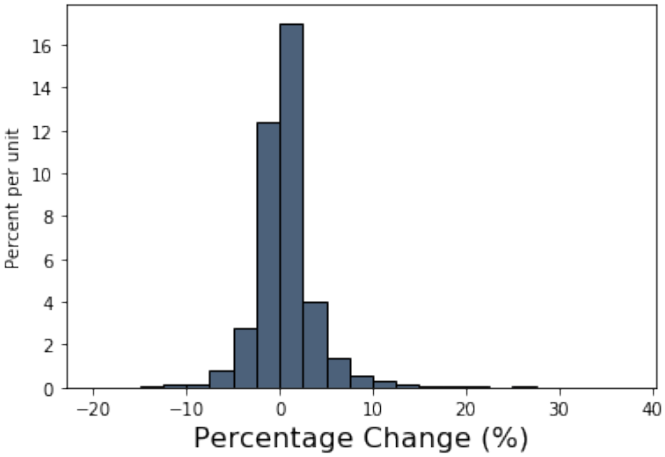
□

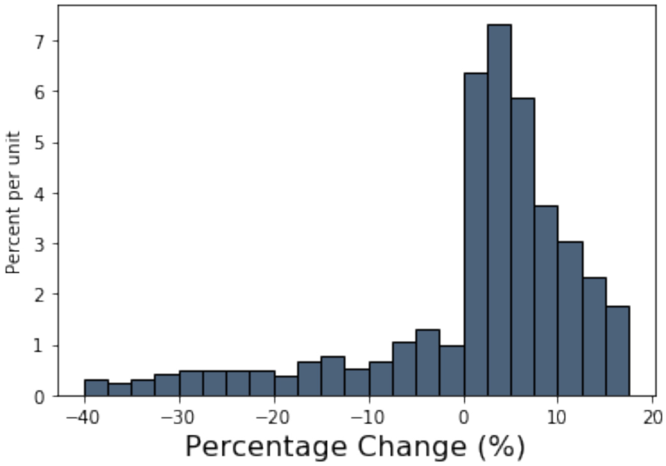
□

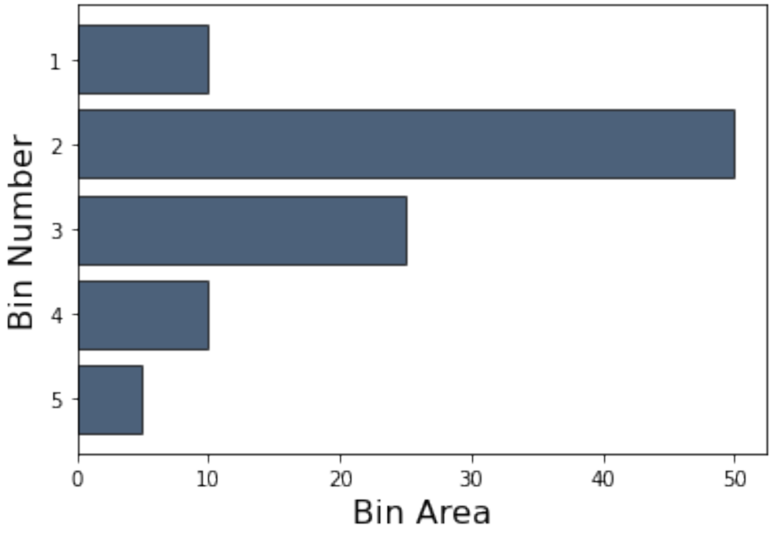
□

(d) (5.0 points)

The four of them together discover an anonymous post on Reddit that claims to have more data about the absolute value of percentage change in stocks that is not publicly known. However, this post requires them to solve a puzzle by piecing together a bar chart and table to see the full histogram. Fill in the table below using the bar chart. (Note: Treat `Bin Number` as a categorical variable.)


Percent per unit
0
2
4
6
8
10
12
14
16
−20
−10
0
10
20
30
40
Percentage Change (%)


Percent per unit
0
1
2
3
4
5
6
7
−40
−30
−20
−10
0
10
20
Percentage Change (%)



Bin Number	Bin Range	Height
1	[0, 2.5)	**a**
2	**b**	10
3	**c**	10
4	[10,15)	**d**
5	**e**	2

i. (1.0 pt) a (your answer should be a single number)

ii. (1.0 pt) b (your answer should be a range in the format [x, y))

iii. (1.0 pt) c (your answer should be a range in the format [x, y))

iv. (1.0 pt) d (your answer should be a single number)

v. (1.0 pt) e (your answer should be a range in the format [x, y))

7. (8.0 points) Visualizations

During shelter-in-place, many Data 8 students explored new hobbies. Angela takes a poll about their hobbies and puts her findings in the `hobbies` table. Angela wonders how to visualize the collected data for presentation at her next discussion.

The first few rows of the `hobbies` table are shown below:

favorite color	favorite hobby	hours of sleep	wake up time	favorite food
blue	baking	7	early	crepe
green	running	10	late	boba
purple	tv shows	8	late	pizza
blue	basketball	7	early	mango
blue	hiking	6	both	pizza

... (282 rows omitted)

(a) (1.0 pt) The best visualization to understand the distribution of the top 5 most popular foods in this discussion is (Choose only one):

○ Scatter Plot

○ Histogram

○ Bar Chart

○ Line Plot

(b) (1.0 pt) Given the `hobbies` table, which methods would you use to help you plot the visualization you chose in part a? (Select all that apply):

□ .sort

□ .join

□ .group

□ .apply

□ .pivot

□ None of the above

(c) (1.0 pt) The best visualization to understand the association between favorite hobby and wake up time in this discussion is (Choose only one.):

○ Bar Chart

○ Overlaid Histogram

○ Scatter Plot

○ Line Plot

○ Overlaid Bar Chart

(d) (1.0 pt) Which visualization would best display the association between hours of sleep and early or late wake up time? (Choose only one.)

○ Scatter Plot

○ Overlaid Bar Chart

○ Bar Chart

○ Line Plot

○ Overlaid Histogram

(e) (1.0 pt) What visualization is impossible to make without modifying the `hobbies` table? (Select all that apply):

□ Line Plot

□ Bar Chart

□ Overlaid Histogram

□ Overlaid Bar Chart

□ Scatter Plot

(f) (1.0 pt) Which of the following variables are categorical? (Select all that apply):

□ Favorite food

□ Hours of sleep

□ Favorite hobby

□ Wake up time

□ Favorite color

(g) (1.0 pt) Select all that are correct:

□ Whether a student attended discussion on a given day is a numerical variable.

□ The size of attendance at a particular discussion section on a given day is a categorical variable.

□ The size of attendance at a particular discussion section on a given day is a numerical variable.

□ Whether a student attended discussion on a given day is a categorical variable.

(h) (1.0 pt) Angela wants to choose a finger food to send to the students in her discussion section. She uses the level of support (i.e., number of rows in the table) for each favorite food (Bagels, Chicken Nuggets, Samosas, Trinidad Cod Fritters, and Wontons) as the basis for her selection. Ultimately, she selects the food option with the most support (favorite).(Select all that apply):

□ Favorite food option is a numerical variable.

□ Level of support for a food option is a categorical variable.

□ Level of support for a food option is a numerical variable.

□ Favorite food option is a categorical variable.

8. (0.0 points) Last Words

(a) If there was any question on the exam that you thought was ambiguous and required clarification to be answerable, please identify the question (including the title of the section, e.g., Visualizations) and state your assumptions. Be warned: We only plan to consider this information if we agree that the question was erroneous or ambiguous and we consider your assumption reasonable.

No more questions.