DATA C8: Summer 2024 Midterm Exam

Full name: ____

Email address: ____@berkeley.edu

Student ID number: ____

Name of the student to your left: _____________________________________

Name of the student to your right: ____________________________________

Instructions:

Do not open the exam until you are instructed to do so.

This exam is **98 points** total, spread out over **6 questions** on **22 pages**, and must be completed in the **110-minute** class period on July 12, 2024, from 11:10 AM to 1:00 PM with **only pencils, erasers, a water bottle, and the midterm reference sheet** unless you have pre-approved accommodations otherwise.

There is a line to write your initials in the upper left-hand corner of each page of the exam. **Please make sure to write your initials on *each page*** to ensure that your exam is graded properly.

Note that some questions require filling in the blank. Please only write your **final answer** in the box.

Note that some questions have circular bubbles to select a choice. This means that you should **select one choice only**. Other questions have square boxes. This means that you should **select all choices that apply**. Please fully shade in the circle or box to mark your answer.

Honor Code [0.5 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: ____

Good luck!

1) Potpourri [18 pts]

a) [2 pt] Charlie creates a table called `crossword` with two columns: `time`, her time (float) for an individual crossword puzzle, and `theme`, whether or not the puzzle had a theme (boolean). Which one of the following would be best for Charlie to visualize the time distributions of both themed puzzles and non-themed puzzles?

○ Line graph
○ Scatter plot
○ Overlaid bar chart
○ Histogram
○ Overlaid histogram

b) [2 pt] An array can contain data that are not similar (e.g., integers and strings).

○ True
○ False

c) [2 pt] If the treatment and control groups are similar in all aspects but the treatment, we may establish that the treatment caused a change in the outcome.

○ True
○ False

d) [2 pt] Pivot tables are strictly for cross-classifying *two* categorical variables.

○ True
○ False

e) [2 pt] What would the following code output?

```
‘20’ + str(round(int(24.8)))
```

○ `2025`
○ `2024`
○ `‘2024.8’`
○ `‘2025’`
○ None of the above

f) [2 pt] Which of the following is correct about Python? **Select all that apply.**

□ Functions need not have a `return` statement.
□ Conditional statements must have an `else` statement.
□ Indenting the body of an `if` statement is optional.
□ A function can take zero arguments.
□ None of the above

g) [2 pt] In a randomized controlled trial, each subject records one observation for the treatment group and one observation for the control group.

○ True
○ False

h) [2 pt] Which one of the following is true?

○ Wesley stands in front of Moffitt Library and surveys nearby students. This is a random sample of UC Berkeley students.
○ Wesley randomly selects students from the full Data 8 Summer 2024 roster by flipping a coin for each student, where “Heads” results in the student being surveyed. This is a random sample of UC Berkeley data science majors.
○ Wesley surveys students who comment on the “Students Materials Megathread” on Ed Discussion. This is a random sample of current Data 8 students.
○ Wesley randomly selects students from his lab section roster by flipping a coin for each student, where “Tails” results in the student being surveyed. This is a random sample of Wesley’s lab section students.
○ None of the above.

i) [2 pt] Mia creates a table `houses` with four columns: `city` (string), `zip code` (int), `price` (float), and `sold` (boolean). Which columns have categorical data, or data which is best treated categorically rather than numerically? **Select all that apply.**

□ `city`
□ `zip code`
□ `price`
□ `sold`
□ None of the above

2) What Would Python Do? [16 pts]

a) [14 pts] For each of the Python expressions below, write the output when the expression is evaluated. If the expression evaluates to an array, you should format your answer like so: `array([..., ..., ...])`. If you think it would produce an error, write `ERROR`. You may assume the standard imports:

```
from datascience import *
import numpy as np
```

And here, we create two arrays, `arr1` and `arr2`.

```
arr1 = np.arange(2, 7, 2)
arr2 = make_array(1, 3, 5)
```

i) `len(np.append(arr1, arr2)) == len(arr1 + arr2)`

ii) `arr1 / make_array(2, 2, 2, 2)`

iii) `sum(make_array(3, 5, 7, 10) % 2 == 0)`

iv) `arr1 * arr2`

v) `"Data" + "Science" * 3`

vi) `str(8) + int('7')`

b) [2 pts] Which of the following functions correctly returns the sum of the squares of the numbers from 0 up to and including n? **Select all that apply.**

□
```
def sum_squares(n):
   total = 0
   for i in np.arange(0, n+1, 1):
      total = total + i**2
   return total
```

□
```
def sum_squares(n):
   return sum(np.arange(1, n+1, 1)**2)
```

□
```
def sum_squares(n):
   total = make_array()
   for i in np.arange(0, n+1, 1):
      total = np.append(total, i**2)
   return sum(total)
```

□
```
def sum_squares(n):
   return sum(np.arange(0, n+1) * np.arange(0, n+1))
```

□ None of the above

3) For Who? [14.5 pts]

Conan and Bing are avid users of TikTok. They are interested in analyzing Conan's For You page (fyp) and uncovering why he likes to spend so much time on the app. Each row in the `tiktok` table below represents one post that Conan has seen in the past month and includes the following columns:

- `creator`: (**string**) the username of the owner of the post.
- `views`: (**int**) the number of views (in thousands) the post received.
- `comments`: (**int**) the number of comments left on the post.
- `tags`: (**string**) all tags associated with the post (e.g., "#fyp #foryou").
- `liked`: (**boolean**) True or False indicating if Conan liked the post or not.

creator	views	comments	tags	liked
crawly_possessed	95800	69856		True
kaybchung	9100	1441	#dentalstudent #dentalschool #allnighter #studytok…	False
thebasementyard	9000	2051		False
allmight_31	40700	34359	#frogs #wrestling #foryou #funny	True
official.mrhungry	32000	59023	#mrhungry #itstimetoeat #easterbunny…	True

(... 4995 rows omitted)

a) [2 pts] Fill in the following blanks to create the most appropriate visualization to display the `comments` and `views` variables on the same plot.

```
tiktok.___(A)___(___(B)___)
```

i) Fill in blank `(A)`:

ii) Fill in blank `(B)`:

b) [3 pts] Fill in the following blanks to return the username of the creator with the most viewed post.

```
tiktok.___(A)___(___(B)___).___(C)___(“creator”).item(0)
```

i) Fill in blank `(A)`:

ii) Fill in blank `(B)`:

iii) Fill in blank `(C)`:

c) [2 pts] Fill in the following blanks to return the number of unique creators in the `tiktok` table.

```
tiktok.___(A)___(“creator”).___(B)___
```

i) Fill in blank `(A)`:

ii) Fill in blank `(B)`:

Conan becomes excited when he stumbles upon a post with very few tags because he feels destined to see that post. He calls such posts “mythical fyp pulls”. To improve Conan’s TikTok experience, Bing suggests that we analyze how Conan’s likelihood of liking a video varies depending on the number of tags a post has. Bing helps us out by defining a function below:

```
def count_tags(string):
     separated_string = string.split(“ ”)
     return len(separated_string)
```

d) [2 pts] Fill in the following blanks to add a new column called **num of tags** to the `tiktok` table. This will be a column of integers containing the number of tags associated with each post.

```
tags_array = tiktok.___(A)___(___(B)___, “tags”)


tiktok = tiktok.___(C)___(“num of tags”, ___(D)___)
```

i) Fill in blank `(A)`:

ii) Fill in blank `(B)`:

iii) Fill in blank `(C)`:

iv) Fill in blank `(D)`:

e) [3.5 pts] Conan is not a fan of posts with numerous tags. First, filter the table to only include posts with fewer than 10 tags. Then, create a table that illustrates how Conan's average number of likes varies by how many tags a post has. Finally, use that table to create a bar chart.

Note: Each bar should represent videos with a certain number of tags, and the lengths of the bars should be determined by Conan's average number of likes for those videos.

```
less_than_ten = tiktok.__(A)__(“num of tags”, __(B)__)


varies_tbl = less_than_ten.select(“num of tags”, “liked”).
             __(C)__(__(D)__)


varies_tbl.__(E)__(__(F)__, “liked average”)
```

i) Fill in blank (A):

ii) Fill in blank (B):

ii) Fill in blank (C):

iv) Fill in blank (D):

v) Fill in blank (E):

vi) Fill in blank `(F)`:

f) [2 pt] Conan and Bing collect more information on three of the creators whose posts they viewed and have stored them in the `creators` table below.

user	followers_mil
kaybchung	1.5
thebasementyard	3.9
official.mrhungry	3

Consider joining *just the five-row excerpt of* `tiktok` and the `creators` tables together.

i) How many rows will the joined table have? Enter a number.

ii) How many columns will the joined table have? Enter a number.

4) Summer, Lovin' Visualizations [17 pts]

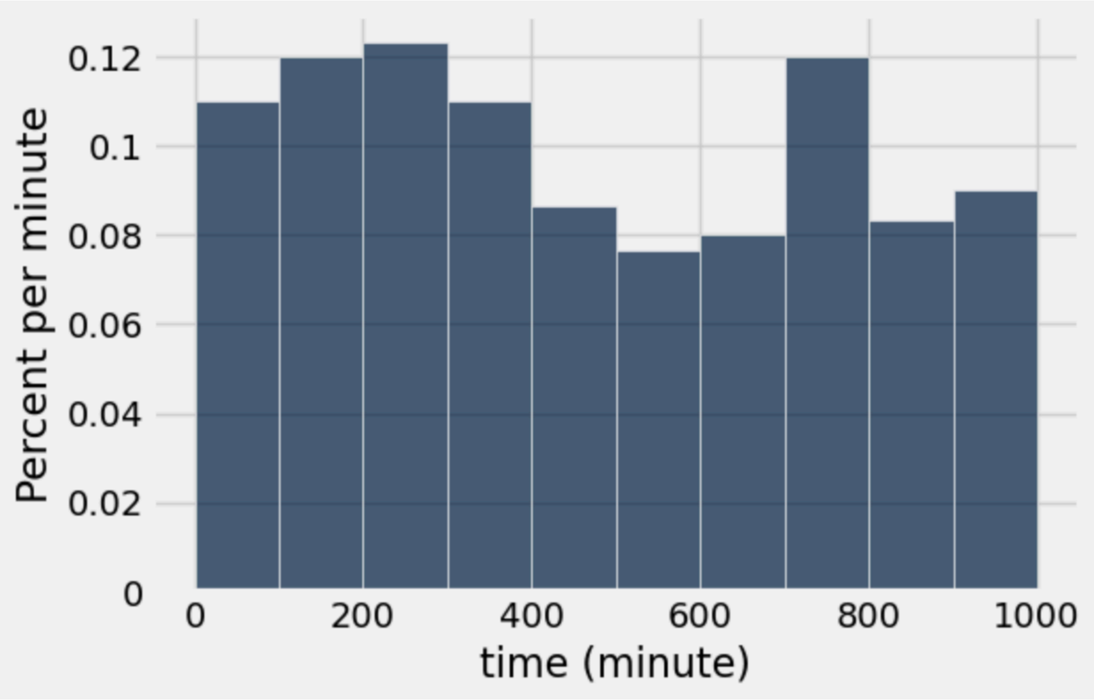
Danny, a Data 8 student this summer, wants to learn more about his classmates' music listening habits. He receives permission from Professor Sanchez to survey the class. All 300 Data 8 summer students respond to his survey, and he organizes the results into a table called `music`, where each row represents a unique student. The first four rows are presented below.

- `studentID`: (**int**) a categorical variable used to identify each student.
- `time`: (**float**) the student's daily music listening time, measured in minutes.
- `platform`: (**string**) the student's favorite music streaming platform.
- `gsi`: (**string**) the student's lab GSI.

studentID	time	platform	gsi
858	56.7	Apple Music	Carisma
402	493.5	Spotify	Andrew
202	806.4	YouTube Music	Kaed
305	273.0	Pandora	Lydia

… (296 rows omitted)

a) [3 pts] Fill in the following blanks to help Danny create the histogram below.
Note: The *x*-axis of the histogram has a range of **[0, 1000]**, and each bin has a width of **100 minutes**.



```
__(A)__.__(B)__(__(C)__, unit = __(D)__, bins = __(E)__)
```

i) Fill in blank `(A)`:

ii) Fill in blank `(B)`:

iii) Fill in blank `(C)`:

iv) Fill in blank `(D)`:

v) Fill in blank `(E)`:

b) [4 pts] GSI Kaed is interested in students' specific daily music listening times. Using only Danny's table and histogram, help Kaed answer the following questions.

i) Which one of the following is the best estimate for the amount of students with a daily music listening time of exactly 100 minutes?

○ 0 students
○ 33 students
○ 24%
○ 28%
○ 32%
○ None of these estimates are appropriate.

ii) Which of the following is the best estimate for the total number of students in our table with a daily music listening time greater than or equal to 500 minutes, but less than 600 minutes?

○ 8 students
○ 16 students
○ 18 students
○ 24 students
○ 30 students
○ 48 students

c) [3 pts] GSI Andrew likes Danny's histogram from part (a), but he challenges Danny to combine the two bins [600, 700) and [700, 800) into one single bin [600, 800).

i) Which of the following is the best estimate of the area of the new **combined** bin?

○ 16%
○ 20%
○ 24%
○ 28%
○ 32%
○ None of the above

ii) Calculate the height of the new **combined** bin, and place your final answer in the box below. You may leave your answer unsimplified or as a fraction. Remember to include the units.

d) [2 pts] GSI Carisma thinks using bins of 60 minutes would make more sense. Which of the following arguments should she use within `music.hist()`? **Select all that apply.**

□ `column = “time” / 60`
□ `unit = “hour”`
□ `bins = np.arange(60)`
□ `bins = np.arange(0, 1000, 60)`
□ `bins = np.arange(0, 1030, 60)`
□ Carisma cannot do this with the `music` table

e) [2 pts] GSI Lydia thinks it would be cool to see the distributions of students' daily music listening times by GSI. She wants to create one plot with a histogram of a different color for each GSI. Which one of the following arguments should she use within `music.hist()`?

○ `column = make_array("time", "gsi")`
○ `group = "gsi"`
○ `group = make_array("gsi")`
○ `group = ["time", "gsi"]`
○ `unit = "gsi"`
○ None of the above

f) [3 pts] Based on the information from Danny's `music` table and histogram from part (a), which of the following statements are true? **Select all that apply.**

□ Being enrolled in Kaed's lab section causes students to listen to more music.
□ Danny's data came from the population of summer 2024 Data 8 students and not a sample of summer 2024 Data 8 students.
□ All UC Berkeley students have daily music listening times ranging from 0 to 1000 minutes.
□ It is appropriate for Danny to create a scatter plot with the `gsi` variable on the *x*-axis and the `platform` variable on the *y*-axis.
□ It is not appropriate for Danny to create a line graph with the `time` variable on the *x*-axis and the `gsi` variable on the *y*-axis.
□ Running this line of code will create a bar chart, where each bar represents a GSI and the bar's length represents the number of students that GSI has: `music.group("gsi").barh("gsi")`

5) Boba Bonanza [13 pts]

Two popular boba tea spots near campus, Taiwanese Professional (TP) Tea and Boba Ninja, are in a friendly competition to see which one is preferred by students. The managers of these two establishments want to determine if there is a significant difference in student ratings between their boba teas.

They collected data from 300 randomly sampled student reviews, with the following columns:

- `name`: (**string**) the name of the boba restaurant.
- `month`: (**string**) the first three letters of the month the review was made.
- `rating`: (**float**) the rating given to the boba tea (on a scale from 1 to 5).
- `topping`: (**string**) the type of topping added ('none', 'regular', or 'extra').

Here are the first five rows of the table, which is named `boba`:

name	month	rating	topping
TP Tea	jan	4.5	regular
Boba Ninja	feb	4.0	none
TP Tea	mar	3.8	extra
Boba Ninja	apr	4.2	regular
TP Tea	may	4.7	none

… (295 rows omitted)

a) [4 pts] Formulate a null hypothesis and an alternative hypothesis that could be used to test whether there is a difference in the average ratings between TP Tea and Boba Ninja.

i) Null Hypothesis:

ii) Alternative Hypothesis:

b) [2 pts] Which one of the following function calls would be helpful for simulating data under the null hypothesis?

○ `boba.sample()`
○ `boba.sample(with_replacement = False)`
○ `boba.sample(300, with_replacement = True)`
○ `sample_proportions(300, make_array(0.5, 0.5))`
○ `np.random.choice(boba.column('rating'))`
○ None of the above

c) [5 pts] Suppose you decide to use a test statistic such that higher values favor the alternative hypothesis. You simulate 1,000 values of the test statistic and assign them to the array `test_stats`. You then write the following code, which assigns the *p*-value to `p_val`.

```
num_extreme_test_stats = __(A)____

p_val = num_extreme_test_stats / __(B)____
```

Assume the observed test statistic has been assigned to `observed_stat`.

i) Fill in blank `(A)`:

ii) Fill in blank `(B)`:

d) [2 pts] Suppose that `num_extreme_test_stats` from part (c) is equal to 45. Assuming you use a 5% cutoff, which of the following can you conclude? **Select all that apply.**

- □ You can reject the null hypothesis.
- □ You can fail to reject the null hypothesis.
- □ You can accept the alternative hypothesis.
- □ There is evidence that the average ratings of TP Tea and Boba Ninja are different in the population.
- □ There is no evidence that the average ratings of TP Tea and Boba Ninja are different in the population.
- □ None of the above

6) Inside Out Lottery [19 pts]

On any given day, one of Andrew's emotions can be found in charge at the control center. The emotion is chosen at random **with replacement** by drawing a ticket out of a box containing 100 tickets total. The tickets in the box have the following breakdown:

Emotion	Tickets
Anger	11
Anxiety	28
Disgust	8
Fear	9
Joy	28
Sadness	16

a) [2 pts] Suppose Fear and Sadness were to pool their tickets together. What is the probability that one of their tickets would be chosen two days in a row?

○ `(0.16 + 0.09)**2`
○ `0.16 + 0.09`
○ `(0.16**2) + (0.09**2)`
○ `0.16 * 0.09`
○ `(0.16 * 0.09)**2`
○ None of the above

b) [2 pts] Out of four consecutive days, what is the probability that Anger is in charge at least once?

○ `0.11 + (0.11)**2 + (0.11)**3`
○ `1 - (0.89)* (0.11)**2`
○ `(3 * 0.11) + (3 * (0.11)**2) + (0.11)**3`
○ `1 - (0.89)**4`
○ `0.11 * (0.89)**3`
○ None of the above

c) [2 pts] Disgust thinks it is unfair that she has fewer tickets than everyone else. She proposes that the tickets should be drawn **without replacement**, and the emotions agree to give it a try. **Select all of the following statements that are true.**

☐ The probability that Disgust will be chosen the first two days in a row is **less than** it was when tickets were being drawn *with replacement*.

☐ The probability that Disgust will be chosen the first two days in a row is **the same as** it was when tickets were being drawn *with replacement*.

☐ The probability of Disgust being chosen on the first day is the same as the probability of her being chosen on the tenth day.

☐ The probability of Disgust being chosen on the second day is higher if she is not chosen on the first day.

☐ In any given six days, it is guaranteed that all emotions will have been chosen once.

☐ None of the statements are true.

The emotions keep losing tickets when drawing **without replacement** and have decided to go back to drawing **with replacement**. Sadness notices that Joy is being drawn many times. She begins to keep track and notices that over the next 50 days, Joy is drawn 25 times. This worries Sadness, who knows the importance of all emotions and suspects Joy might be rigging the system by replacing other tickets with her own tickets to increase her chances of being drawn. Having taken Data 8 in the past, Sadness decides to investigate.

d) [2 pts] Given the information above, provide a clear and complete null hypothesis for Sadness's investigation.

e) [2 pts] Provide an alternative hypothesis.

f) [2 pts] Which one of the following test statistics is the best to assess whether Joy's probability of being chosen is higher than it should be according to the original ticket drawing system?

○ The difference between the sampled proportion of times Joy is drawn and the expected proportion
○ The absolute difference between the sampled proportion of times Joy is drawn and the expected proportion
○ The expected proportion of times Joy is drawn under the alternate hypothesis
○ The expected proportion of times Joy is drawn under the null hypothesis
○ The expected proportion
○ None of the above

Upon closer examination of recent drawing results, Sadness notices that Joy may not be the only one whose probability of being chosen seems to have changed. She is now adjusting her investigation to see if there is a difference between the expected distribution and the one she has observed. Here are the hypotheses she is going to use:

- Null: The probability that each emotion is chosen is the same as described in the original distribution.
- Alternative: The distribution of emotions chosen in these 50 days is different from the original distribution.

g) [2 pts] Sadness wants to use TVD to assess if a difference exists but is rusty on this topic. **Select all of the following statements that are true.**

□ We use `np.abs` when calculating TVD because we only care about the instances when the differences are positive.
□ We use `np.abs` when calculating TVD because otherwise, the differences would sum to 0.
□ TVD will be the same regardless of whether you subtract the expected distribution from the simulated, or vice versa.
□ TVD is a good test statistic to use when assessing whether one emotion is being chosen more than it should be.
□ It is necessary to divide TVD by 2 because TVD takes an average of the two distributions.
□ None of the statements are true.

h) [3 pts] Sadness now writes a function to calculate one test statistic but is struggling. Help her by filling in the blanks.

```
def simulate_test_statistic():
     expected_props = make_array(0.28, 0.16, 0.08, 0.28, 0.11,
                                  0.09)
     simulated_props = __(A)__(__(B)__)
     test_stat = sum(__(C)__(__(D)__)) / 2
     return test_stat
```

i) Fill in blank `(A)`:

ii) Fill in blank `(B)`:

iii) Fill in blank `(C)`:

iv) Fill in blank `(D)`:

i) [2 pts] Suppose after Sadness runs her simulation 10,000 times, she gets a *p*-value of 0.035 and uses a cut-off value of 0.05. Which of the following statements is correct? **Select all that apply.**

- □ There is a 0.035 chance that the alternative hypothesis is true.
- □ There is a 0.035 chance that the null hypothesis is true.
- □ There is a 0.035 chance that the TVD simulated under the null hypothesis is equal to or less than the observed test statistic.
- □ There is a 0.05 chance that Sadness will falsely reject the null hypothesis if the null is true.
- □ Since the *p*-value is less than the cut-off value, Sadness can conclude that the null hypothesis is more reasonable than the alternative hypothesis.
- □ None of the statements are correct.

7) Congratulations [0 pts, *Optional*]

Congratulations! You have completed the Midterm Exam.

- **Please make sure that you have written your initials on *each page* of the exam.** You may lose points on pages where you have not done so.
- Also, make sure that you have **signed the Honor Code** on the cover page of the exam for 0.5 point.

a) [0 pts] 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., "Boba Bonanza"), 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.

b) [0 pts] Draw and caption your favorite Data 8 experience or staff member!