Data C8
Spring 2025

Foundations of Data Science Midterm

7:10-9:00PM, FRIDAY, MARCH 14

Berkeley Honor Code [1 point]

"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."

Initials: _____

Full Name: _____

Student ID Number: _____

Name of person to your left: _____

Name of person to your right: _____

GSI/TAs Name (Write N/A if in self-service lab): _____

INSTRUCTIONS

- You may only have with you: a pencil(s), an eraser(s), your student ID, a water bottle, and your midterm reference sheet, unless you have received pre-approved accommodations otherwise.
- If you need to use the restroom, bring your phone, exam, reference sheet, and student ID to the front of the room.
- Do not open the exam until you are instructed to do so.
- Write your initials at the top of each page.
- There are **5** questions and **16** pages on this exam, including cover page. **Read the instructions and point values carefully** for each question, part and subpart.
- Where relevant, you may assume that all necessary Python modules have been imported. Use of any code which has not been taught in this iteration of the course is prohibited and it will not be graded.
- Where a written (English) answer is expected, you must use complete sentences. Your work will not be graded otherwise.
- **Each coding blank may include multiple arguments/methods/functions.** However, your solution must use every blank available.

MULTIPLE CHOICE QUESTION TYPES

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

◯ You must choose either this option

◯ Or this one, but not both!

For questions with **square boxes**, you may fill in *multiple* choices. **Please fill in completely.**

☐ You could select this choice.

☐ You could select this one too!

1 Multiple Choice [23 points]

For each of the following questions, fill in the bubble/square(s) corresponding to the correct answer.

a. (5 points) The Data 8 staff has a version of the `united` table from lecture that contains all United Airlines flights departing from San Francisco in June, July and August of 2024. They want to answer the following question: were the flights in the table that took place in July delayed for longer, on average, than those in June? Ella decides to report the difference in average flight delay times between the two months. Aileen, however, decides to run a hypothesis test, using the difference Ella calculated as a test statistic. State whose method is more appropriate to answer the question.

◯ Ella's method ◯ Aileen's method

b. (4 points) What will output to the screen once the following Python expression runs?

```
3 * make_array(0,1,2,3) ** 2  - 1
```

◯ An error will be produced. ◯ `array([-3, 0, 9, 24])`
◯ 107 ◯ `array([-1, 2, 11, 26])`
◯ `array([-1, 5, 11, 17])` ◯ `array([-1, 8, 35, 80])`

c. (3 points) On Project 1, you worked with a dataset that included multiple tables, each of them containing different statistics recorded on the same set of countries across many years. Which table method was instrumental in allowing you to combine information from multiple tables in order to answer a particular question?

◯ `group` ◯ `join` ◯ `with_columns` ◯ `select`
◯ `apply` ◯ `pivot` ◯ `where` ◯ `sort`

d. (3 points) When the following Python code is run, how many rows and columns will the table output to the screen have?

```
my_table = Table().with_columns('letters',  make_array('a','b','c','d'),
                                'numbers', np.arange(4))

my_table.select('numbers').where('numbers', are.above(2))
```

◯ 2 rows and 4 columns ◯ 2 rows and 1 column
◯ 4 rows and 2 columns ◯ 4 rows and 1 column
◯ 1 row and 1 column ◯ An error will be produced.

e. (2 points) Each year on Groundhog Day (February 2), the famous groundhog, Punxsutawney Phil, predicts whether or not another six weeks of winter weather will follow. Historically, he is accurate 40% of the time. Assuming that Phil has a 40% chance of predicting correctly each year, independently of other years, write a math expression for the probability that he makes a correct prediction in the years 2025, 2026 and 2027, but then makes incorrect predictions in the years 2028 and 2029.

- ◯ $(\frac{4}{10})^3 * (\frac{6}{10})^2$
- ◯ $(\frac{4}{10}) * 3 + (\frac{6}{10}) * 2$
- ◯ $(\frac{4}{10})^3 + (\frac{6}{10})^2$
- ◯ $(\frac{6}{10})^3 * (\frac{4}{10})^2$
- ◯ $(\frac{4}{10}) * 3 * (\frac{6}{10}) * 2$
- ◯ $(\frac{6}{10}) * 3 + (\frac{4}{10}) * 2$
- ◯ $(\frac{6}{10})^3 + (\frac{4}{10})^2$
- ◯ Not enough information is given to answer.

f. (2 points) Ishani convinced nine other Data 8 staff members to try a secret sushi restaurant in downtown Berkeley with her. At this restaurant, each customer has a 50% chance of getting served tea, a 30% chance of getting served soda and a 20% chance of getting served water, independently of other customers. When no one is served soda, Ishani grows suspicious. She decided to test the hypothesis that the distribution of drinks follows the probabilities listed by the restaurant against the alternative that they follow a different one. What is an appropriate test statistic to use?

- ◯ Difference in means
- ◯ The number of expected sodas given out to a party of ten
- ◯ Total variation distance
- ◯ Absolute value of difference in means

g. (2 points) For which case studies covered across the course materials have we used the `sample_proportions` function? *Select all that apply.*

- ☐ Smoking and birth weights - text and lecture
- ☐ Therapeutic touch - lab
- ☐ Robert Swain jury panel - text and lecture
- ☐ Happiness factors - homework

h. (1 point) After polling a representative, random sample of American adults on their beliefs toward current issues, the Pew Research Center published an article in February 2025 with the headline Americans Continue to View Several Economic Issues as Top National Problems. Which of the three aspects of data science does Pew's study fall under?

◯ Exploration ◯ Inference ◯ Prediction

i. (1 point) The following visualization was created using the `du_bois` table seen in lecture. What kind of visualization is it?

◯ Overlaid histogram

◯ Histogram

◯ Overlaid (side-by-side) bar chart

◯ Bar chart

2 Movie Night [12 points]

Dagny is hosting a movie night with her friends. She is overwhelmed with all the options and wants to use information from the online database IMDb. She has compiled a table containing information on 400 popular movies called `imdb`; a four-row excerpt of the table, as well as a description of its attributes (columns), is below.

- **ID** *(integer)*: A code which uniquely defines a movie's URL on the website.
- **Title** *(string)*: The name of the movie.
- **Year** *(integer)*: The year the movie was released.
- **Genre** *(string)*: The artistic style of the movie.
- **Critics** *(float)*: An aggregated rating, out of 10, given to the movie by critics.
- **Fans** *(integer)*: An aggregated rating, out of 100, given to the movie by fans.
- **Runtime** *(integer)*: The duration of the movie, in minutes.

ID	Title	Year	Genre	Critics	Fans	Runtime
7286456	Joker	2019	Crime	8.4	59	122
1517268	Barbie	2023	Adventure	6.8	81	114
0133093	The Matrix	1999	Action	8.7	83	136
0094721	Beetlejuice	1988	Comedy	7.5	71	92

...(396 rows omitted)

a. (5 points) Which attributes in the table are numerical? *Select all that apply.*

- ☐ **ID**
- ☐ **Title**
- ☐ **Year**
- ☐ **Genre**
- ☐ **Critics**
- ☐ **Fans**
- ☐ **Runtime**

b. (4 points) Dagny wants to see if movies well liked by fans also tended to be well liked by critics.

(i) (2 points) Which of the following visualizations is most appropriate for this purpose?

- ◯ Line plot
- ◯ Overlaid histogram
- ◯ Side-by-side (overlaid) bar chart
- ◯ Scatter plot

(ii) (2 points) The **Critics** and **Fans** ratings are currently on different numerical scales, which might make interpreting a visualization involving the two attributes difficult. Complete the code below to add a new version of the **Fans** column, called **Fans Rescaled**, to the `imdb` table whose values are on the same numerical scale as the values in the **Critics** column.

```
imdb = imdb._________(A)_________(_________(B)_________)
```

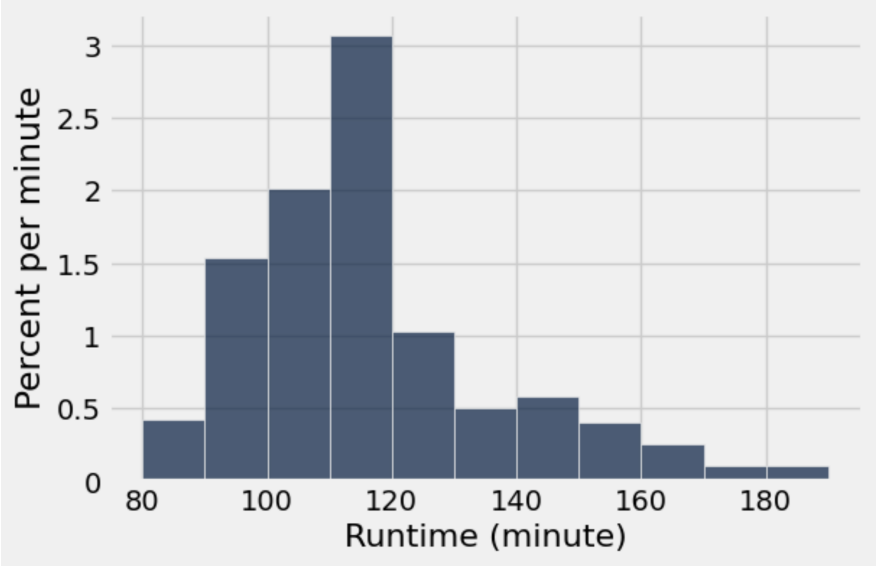
A. Fill in blank (A)

B. Fill in blank (B)

c. (3 points) Dagny wants to take into account the fact that movies vary greatly in length, so she turns her attention to the **Runtime** attribute.

(i) (2 points) Dagny first generates the following histogram of the distribution of runtimes using the `hist` table method. Which of the following questions can be answered based on the histogram? *Select all that apply.*

- ☐ What is the percentage of movies that have a runtime of at least 140 minutes?
- ☐ What is the percentage of movies that have a runtime greater than 160 minutes?
- ☐ What is the percentage of movies that have a runtime of less than 160 minutes?
- ☐ What is the percentage of movies that have a runtime of at least 125 but less than 130 minutes?



(ii) (1 point) Dagny is also interested in labeling the movies as new or old based on their year of release. She creates a function called `year_since_released` to help her. `year_since_released` takes in a movie name as an argument, calculates the years since the release of the movie, labels the movie as *old* if it has been at least 25 years since the release, and labels it *new* otherwise. Once she has finished writing the function and verifies that it works correctly, Dagny wants to look at the distribution of runtimes across both old and new movies. Complete the skeleton code below to create a visualization which accomplishes this task.

```
imdb.with_column('Age',______(A)______).______(B)______(______(C)______)
```

A. Fill in blank (A)

B. Fill in blank (B)

C. Fill in blank (C)

3 Welcome Survey [16 points]

At the beginning of the semester, we asked each of you to fill out a survey about your personal and academic background. Some of the results of this survey have been organized into a table named `survey`. An excerpt and data description of the table is shown below.

- **Major** *(string)*: A student's declared major/set of majors. Students also had the option to select "Undeclared."
- **Experience** *(string)*: Whether or not a student had experience in computer programming prior to taking Data 8.
- **Comfort** *(integer)*: A student's comfort level in computer programming as translated to a scale of 1 to 5, with larger values indicating a higher level of comfort.
- **Languages** *(integer)*: A student's answer to the question: *In how many (speaking) languages do you know how to say the phrase "Hello World"?*

Major	Experience	Comfort	Languages
Public Health	No	2	2
Data Science, Economics	Yes	4	2
Business, Environmental Studies	No	1	1
Applied Math	Yes	5	3

...(1421 rows omitted)

a. (13 points) Fill in the blanks in the Python expressions to compute the described values. You must use only the lines provided. *The last line of the answer should evaluate to the value described.*

(i) (5 points) The most common major/set of majors among Data 8 students this semester.

```
majors = survey._____(A)_____(_____(B)_____)

majors._____(C)_____(_____(D)_____).column(_____(E)_____).item(_____(F)_____)
```

A. Fill in blank (A)

B. Fill in blank (B)

C. Fill in blank (C)

D. Fill in blank (D)

E. Fill in blank (E)

F. Fill in blank (F)

(ii) (4 points) A two-column table with the five majors/set of majors having the lowest average comfort level in computer programming. The first column should contain the five majors/set of majors; the second column should contain the average comfort level for each.

```
majors = survey.select(______(A)______).______(B)______(______(C)______)

majors.______(D)______(______(E)______).______(F)______(______(G)______)
```

A. Fill in blank (A)

B. Fill in blank (B)

C. Fill in blank (C)

D. Fill in blank (D)

E. Fill in blank (E)

F. Fill in blank (F)

G. Fill in blank (G)

(iii) (3 points) The percentage of students in the course that have declared the data science major.

```
part = survey.______(A)______(______(B)______).______(C)______

whole = survey.______(D)______

percentage = (part/whole) * 100
```

A. Fill in blank (A)

B. Fill in blank (B)

C. Fill in blank (C)

D. Fill in blank (D)

(iv) (1 point) A visualization of the distribution of the comfort level when broken down by experience.

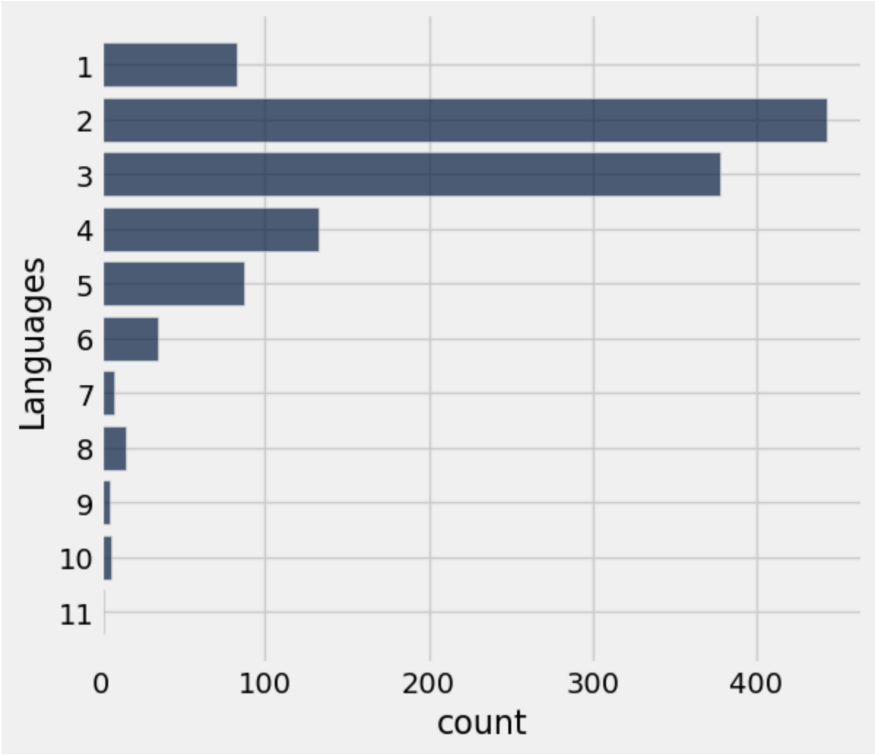
```
survey.______(A)______(______(B)______).______(C)______('Experience')
```

A. Fill in blank (A)

B. Fill in blank (B)

C. Fill in blank (C)

b. (3 points) The following is a visualization made with the **Languages** attribute. In the box below, give the plot a title which summarizes the pattern that the visualization shows. Your title should not be longer than one sentence.



4 Galentine's Day Gone Postal [14 points]

Ramisha wants to build a bouquet of flowers for her friend Cai on Galentine's Day, but falls ill on the morning of February 14. She therefore uses a flower delivery service to send some flowers to Cai. This service sends out a random bouquet of five flowers. When each flower is picked for the bouquet, it has a $\frac{1}{6}$ chance of being a lily, a $\frac{2}{6}$ of being a rose and a $\frac{3}{6}$ chance of being a tulip, independent of other flowers picked.

a. (3 points) The following is a skeleton of a function called `number_of_roses` which simulates the selection of the five flowers for the bouquet (each flower can either be a **Lily**, **Rose**, or **Tulip**) and then computes the number of these flowers that are roses. Complete the function by filling in the blanks.

```
def number_of_roses():

    flower_choices = ______(A)______

    bouquet =  np.random.choice(______(B)______)

    return ______(C)______(______(D)______)
```

A. Fill in blank (A)

B. Fill in blank (B)

C. Fill in blank (C)

D. Fill in blank (D)

b. (3 points) The following is a skeleton of a function called `rose_in_bouquet`. The function computes the number of roses in the bouquet, and returns `True` if there is at least one rose and `False` otherwise. Complete the skeleton code. You may use any functions that were defined previously and assume they work as intended!

```
def rose_in_bouquet():

    rose_sum = ______(A)______

    ______(B)______:
        return True

    ______(C)______:
        return False
```

A. Fill in blank (A)

B. Fill in blank (B)

C. Fill in blank (C)

c. (3 points) The following is a skeleton of a function called `rose_in_bouquet_prop`. The function simulates 1,000 bouquet selections and returns the proportion of these that resulted in a bouquet with at least one rose. Complete the function by filling in the blanks. You may use any functions that were defined previously and assume they work as intended!

```
def rose_in_bouquet_prop():

    rose_in_bouquet_results = ______(A)______

    for i in ______(B)______:

        rose_in_bouquet_results = np.append(______(C)______)

    return ______(D)______
```

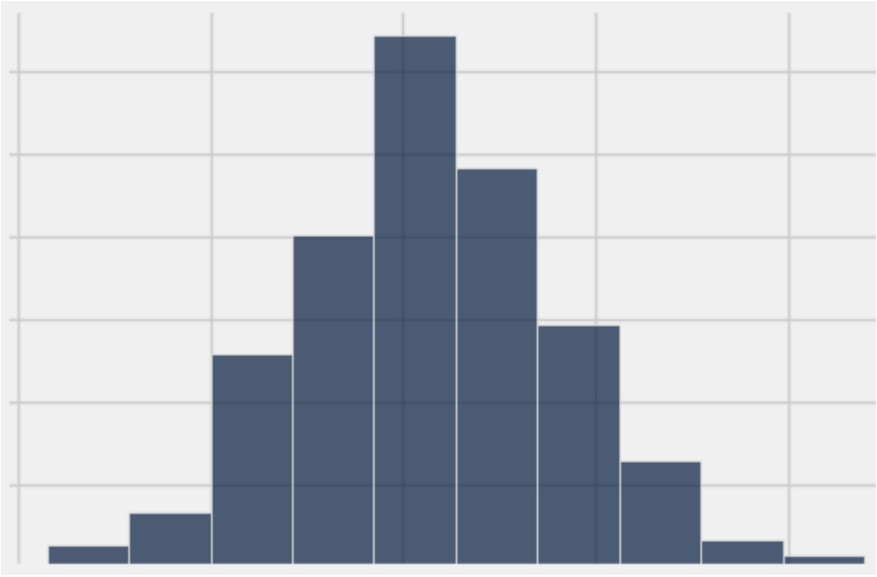
A. Fill in blank (A)

B. Fill in blank (B)

C. Fill in blank (C)

D. Fill in blank (D)

d. (5 points) Ramisha ran the completed `rose_in_bouquet_prop` function 2,000 times. From this, she collected two-thousand proportions and then created the visualization below.



△

(i) (3 points) What kind of distribution is being visualized?

◯ Empirical distribution ◯ Probability distribution

(ii) (2 points) As mentioned earlier, Ramisha has used the code to repeat the the rose-bouquet selection experiment (which consists of 1,000 trials) a large number of times (2,000 times). By the Law of Large Numbers, what should we expect the value of the proportion under the caret (△) to be close to? Show your work in the space below and place a box around your final answer.

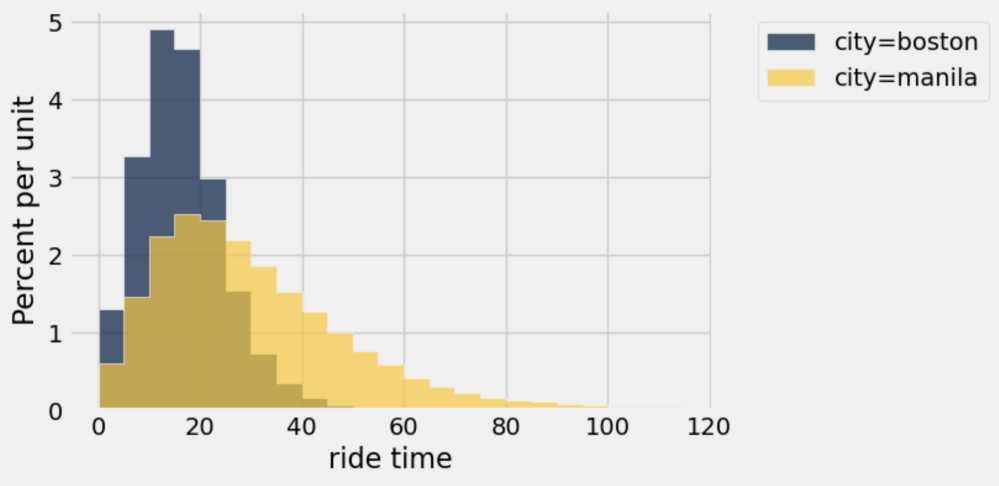
5 *Driving, Just Driving* [19 points]

While listening to SZA's song *Drive*, Sam and Thomas became interested in learning more about the differences in the distributions of Uber passenger ride times for Manila and Boston. They took large, independent, random samples of ride times which had been observed in each city during the year of 2024 and placed them into a new table called `rides`, which contains two columns. A description of these columns and a four-row excerpt of the table are below.

- **City** *(string)*: Indicates whether the Uber ride took place in Manila or Boston.
- **Time** *(integer)*: The length of the Uber ride (in minutes).

City	Time
Boston	36
Manila	45
Boston	12
Manila	64

...(375 rows omitted)



After looking at the above visualization, which displays the distribution of ride times broken down by city, Sam and Thomas are curious if the average ride time for Manila rides is different from the average ride time for Boston rides. They decide to perform an A/B test to assess their claim and settle on a p-value cutoff of 0.05.

a. (4 points) In the following boxes, write the appropriate null and alternative hypotheses for this test:

(i) (2 points) Write the appropriate null hypothesis here.

(ii) (2 points) Write the appropriate alternative hypothesis here.

b. (3 points) Which of the following quantities would be most appropriate as a test statistic?

- ◯ Absolute value of difference in mean ride times
- ◯ Total variation distance between the distributions of ride times
- ◯ Absolute value of difference in the 75th percentile of ride times
- ◯ Difference in mean ride times

c. (3 points) Sam computes one simulated value of the test statistic under the null hypothesis. The simulated value will be obtained from a new table which has been randomly generated using the `rides` table. Which of the following options describes a valid process to generate this table?

- ◯ Split the dataset into Boston and Manila tables, shuffle the **City** column within each of these tables, and then recombine the shuffled tables.
- ◯ Shuffle the **City** column.
- ◯ Select with replacement as many rows as are in the `rides` table.
- ◯ Generate a table where 50 percent of the rows are Manila times and 50 percent of the rows are Boston times.

Thomas computes a large number of simulated test statistics, visualizes the distribution of these statistics along with the observed test statistic, and arrives at a p-value of 0.03.

d. (2 points) Which statements about Thomas's p-value are true? *Select all that apply.*

- ☐ His p-value is the proportion of test statistics that are smaller than their observed test statistic.
- ☐ His p-value suggests that the null hypothesis is not true.
- ☐ His p-value suggests that if the average ride times are the same, it would be unlikely to obtain the observed test statistic that was computed from the data in the `rides` table.
- ☐ None of these statements are true.

e. (2 points) Which statement describes the conclusion of the test, based on the p-value cutoff of 0.05?

- ◯ We would reject the null hypothesis and conclude that the average ride times for the two cities are different.
- ◯ We would retain the null hypothesis and conclude that the average ride for the two cities are the same.
- ◯ The null hypothesis is false. We conclude that the average ride times for the two cities are different.
- ◯ The null hypothesis is true. We conclude that the average ride times for the two cities are different.

f. (5 points) As mentioned before, Sam and Thomas ensured that the ride times in the `rides` table from each city were sampled independently and at random. They also used a random mechanism to obtain each of the simulated test statistics. Should their test results support the alternative hypothesis, would these uses of randomness allow Sam and Thomas to make the conclusion that the difference in city led to the difference in ride times?

- ◯ Yes
- ◯ No

6 Congratulations! [Extra Credit: 1 point]

Extra Credit [1 point]: What is Prof Jeremy allergic to?: *Repeating code*

You have now completed the Midterm Exam. If you have not been told otherwise, you may bring all of your testing materials (reference sheet and this test paper), as well as your student ID, to the front of the room. Once you have been checked off, you may leave quietly.

- 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.**
- Please make sure you have filled in bubbles and squares completely rather than having used a check mark, cross or any other mark.
- Double check that you have not skipped over any questions!

Below, you may draw and caption your favorite Data 8 experience or staff member!