DATA C8
Summer 2026

FOUNDATIONS OF DATA SCIENCE

MIDTERM

11:00AM–1:00PM FRIDAY, JULY 17TH 2026

PRINT Your Name: Data C8

PRINT Your Student ID: _____

PRINT Your Exam Room: _____

PRINT the Name of Person to your Left: _____

PRINT the Name of Person to your Right: _____

PRINT Your GSI's Name : _____

INSTRUCTIONS

You have **110 minutes** to complete the exam. There are **6 questions** and **12 pages** on this exam, including this cover page.

Question	1	2	3	4	5	6	Total
Points	26	33	17	14	10	0	100

- This exam is closed book, closed computer and closed calculator, except the Reference Sheet provided for you.
- You may only have with you: a pencil, an eraser and your student ID, unless you have pre-approved accommodations.
- If you need to use the restroom, bring your phone, exam, reference sheet and student ID to the front of the room.
- For written questions:
  - answers written outside the boxes provided will not be graded;
  - if your answer is ambiguous or you provide multiple answers, the worst interpretation will be graded.
- For coding questions:
  - blank spaces may include multiple arguments or functions per blank, but your solution must use every blank available;
  - you may assume the `datascience` and `numpy` libraries are imported, as seen in class;
  - the use of **any code** which has not been taught in this offering of the course is not allowed and will result in zero credit.
- For multiple choice questions, see question types and instructions below.

Questions with **circular bubbles**: you may select only **1 choice**. Questions with **square boxes**: you may select **1 or more choices**.

○ Unselected option (completely unfilled)

● Single option selected (completely filled)

■ You may select multiple squares

■ as long as they are completely filled

You must fill in the bubbles **completely**. Ticks, crosses, or other check marks will **not** receive credit.

HONOR CODE

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

SIGN Your Name: _____

This page intentionally left blank

The exam begins on the next page.

Initials: [ ]

1. [26.0 points] General

(a) [6.0 pts] *Select all* of the options where performing a hypothesis test is appropriate.

- ☐ There are eight planets in our solar system: four of them rocky and four of them gaseous. Viraj wants to assess whether the rocky planets are smaller in diameter, on average, than the gaseous planets.
- ■ Nao wants to assess whether a new version of her midterm exam is more difficult than the previous one. She randomly assigns half of the students in her class to take the old version and the other half to take the new version.
- ☐ Cynthia wants to assess whether a new location for office hours is more accessible for students than the old location. She sets up a table near the student union and collects data on the first 100 people that pass by.
- ■ Tai wants to assess whether the majority of Berkeley students prefer online classes to in-person classes. He takes a random sample of 100 students from the entire student roster and collects data from each of these students.
- ■ Every discussion, Elizabeth gives out a cookie to one student by spinning a wheel she found online. After the wheel lands on the same student in three straight sections, she wonders whether each student has an equal chance of being selected.
- ☐ Lena wants to predict how many students will enroll in the fall section of Data C8, based on enrollment figures from the summer and the number of incoming majors in relevant fields.

(b) [5.0 pts] What does the following Python expression evaluate to? Give your answer as Python would display it.

```
make_array(-1,-2,-3) + make_array(1,2,3) * 6
```

arr([5, 10, 15])

(c) [5.0 pts] Which methods that can I call on `raghavans_table` may result in a table with less rows? *Select all that apply.*

```
raghavans_table._____
```

☐ `select(...)` ■ `where(...)` ■ `take(...)` ■ `join(...)` ■ `sample(...)`

(d) [3.0 pts] What data type does the following Python expression evaluate to?

```
(4.0 != (True * 8)/2) + 1
```

○ `bool` ○ `float` ○ `numpy array` ○ `str` ● Something else

(e) [3.0 pts] The table below was created directly from the `farmers_markets` table you have seen in your assignments. What can be said about this table? Fill in the blanks in the sentence shown on the next page, from left to right.

Bakedgoods	Vegetables	Seafood	count
Y	Y	N	3379
N	N	N	3202
Y	Y	Y	1248
N	Y	N	564
Y	N	N	80
N	Y	Y	41
Y	N	Y	30
N	N	Y	2

"This is a _____ table. A _____ table _____ be used to display this information."

- ○ grouped; pivot; could
- ● grouped; pivot; could not
- ○ pivot; grouped; could
- ○ pivot; grouped; could not

(f) [2.0 pts] How did John Snow conclude that contaminated water was the cause of the cholera outbreaks in London in the 1850s?

- ○ Snow drew a map of cholera cases during a cholera outbreak in the Soho district in London. He saw that most of the cases centered around a water pump on Broad Street.
- ● Snow analyzed an area of London where residents received water from one of two suppliers. One of these suppliers got water downstream from a contamination site, and the other supplier got water upstream from this site. The downstream supplier was hit much harder by the outbreak.
- ○ Snow performed a randomized controlled trial and saw that the treatment group was hit much harder by the outbreak.
- ○ Snow analyzed an area of London where residents received water from one of two different pumps: a clean pump and a contaminated pump. The contaminated pump was hit much harder by the outbreak.

(g) [2.0 pts] Prof. Jeremy plays chess in his free time against an AI chess robot. Suppose that, one day, Prof. Jeremy plays the robot 3 times, and that during each game he has a 60% chance of winning. What is the probability that *neither Prof. Jeremy nor the AI robot wins all of the games*? Write an arithmetic expression for the probability below and show your work where necessary.

The probability is $1 - (P(JJJ) + P(RRR))$

$= 1 - \left(\left(\frac{6}{10}\right)^3 + \left(\frac{4}{10}\right)^3 \right)$

2. [33.0 points] FIFA World Cup

The FIFA World Cup tournament is the most prestigious international soccer tournament. Running on a four-year cycle, the 2026 edition is currently ongoing across the United States, Canada and Mexico.

The `teams` table contains information on the 48 nations competing in the 2026 tournament. Below is a four-row excerpt of the table, along with short descriptions for each column.

- **Team:** The name of the participating nation.
- **Region:** One of "North America", "South America", "Africa", "Asia," "Europe" or "Oceania".
- **Population:** estimated as of 2026.
- **Group**: A code which indicates what four-team group the nation was drawn into for the group stage of the tournament.
- **Debutant:** "Yes" if the country is competing in its first World Cup tournament, "No" otherwise.
- **Ranking:** FIFA Ranking as of June 2026.
- **Titles**: Number of FIFA World Cups won by the nation.

Team	Region	Population	Group	Debutant	Ranking	Titles
Spain	Europe	49687120	8	No	2	1
Mexico	North America	132997658	1	No	14	0
Tunisia	Africa	12500000	6	No	45	0
Scotland	Europe	5500000	3	No	42	0

`(44 rows omitted...)`

(a) [5.0 pts] How many numerical variables are there in the table?

○ 1 ○ 2 ● 3 ○ 4 ○ 5 ○ 6 ○ 7

(b) [4.0 pts] *Select all of the questions below* that the information in the table provides a sufficient answer to.

☐ In what year did Spain win the FIFA World Cup?

■ Which teams have a good shot at winning the 2026 FIFA World Cup?

■ Which regions represented in the table have won the FIFA World Cup?

☐ What is the total population of Africa?

(c) [5.0 pts] In the box below, write *one line* of code to visualize the distribution of the **Region** variable.

```
teams.group('Region').barh('Region')
```

(d) [4.0 pts] In the box below, write *one line* of code to find the number of debutant countries.

```
teams.where('Debutant', 'Yes').num_rows
```

(e) [3.0 pts] There is one nation in the tournament that has won more FIFA World Cup titles than any other team in the tournament. *Assuming you don't know which nation it is*, write *one line* of code to return the name of this nation as a string.

```
teams.sort('Titles', descending=True).column('Team').item(0)
```

(f) [3.0 pts] Complete the code below, which adds a column named **Population (millions)** to the `teams` table containing the population of each country *in millions*.

```
pop_millions = _____[1]_____
teams = teams._____[2]_____
```

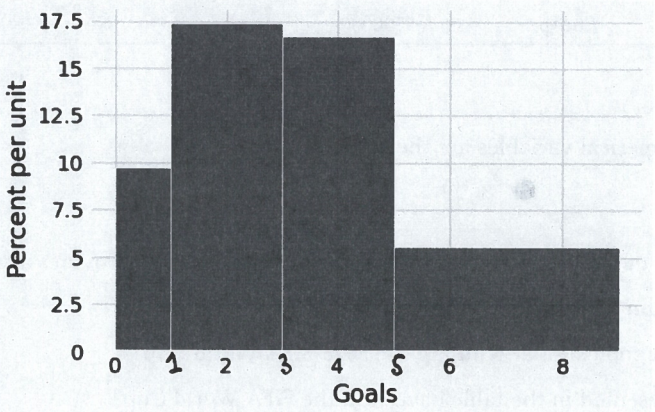
i. [1.5 pts] Fill in Blank [1]:

teams.column("Population") / 1e6

i. [1.5 pts] Fill in Blank [2]:

with_column('Population (millions)', pop_millions)

(g) [2.0 pts] The visualization below displays the distribution of goals scored in the 72 group stage games of the tournament. Assuming the bins include their lower endpoints and do not include their upper endpoints, which of the following questions can be answered using this visualization? *Select all that apply.*



- [x] What percentage of group stage games saw no goals scored?
- [x] What percentage of group stage games saw between 1 and 5 goals scored?
- [ ] What percentage of group stage games saw 3 goals scored?
- [ ] What percentage of group stage games saw less than 7 goals scored?
- [x] What percentage of group stage games saw at least 3 goals scored?

(h) [1.0 pt] In the box below, provide a title for the visualization *in ten words or less.*

The highest percentage of games saw 1 or 2 goals

* The title must feature a claim/summary/takeaway to receive full credit.

(i) [6.0 pts] The `players` table contains the best 100 players in the 2026 FIFA World Cup according to FOX Sports, the main T.V. broadcaster of the tournament in the United States.

Name	Nation	Age	Position
Alexander Isak	Sweden	26	Forward
Mohamed Salah	Egypt	33	Forward
Willian Pacho	Ecuador	24	Defender
Emiliano Martinez	Argentina	33	Goalkeeper

(96 rows omitted...)

i. [3.0 pts] In the box below, write *one line* of code to create a two-column table which contains the name of each position in the `players` table and the average age of the players in that position.

```
players.group('Position', np.average).select('Position', 'Age average')
```

ii. [2.0 pts] Write *one line* of code to create the following table from the `players` table.

Nation	Defender	Forward	Midfielder	Goalkeeper
France	2	8	4	1

```
players.pivot('Position', 'Nation').where('Nation', 'France')
```

iii. [1.0 pt] Fill in the skeleton code below to create a three-column table with the following columns:

- **Nation**
- **Top 100**: the number of players in the top 100 for a given nation.
- **Ranking**

```
player_counts_table = _____[1]_____
teams = teams._____[2]_____
```

[0.5 pts] Blank `[1]`

```
players.group('Nation').relabeled(1, 'Top 100')
```

[0.5 pts] Blank `[2]`

```
.join('Team', player_counts_table, 'Nation').relabled('Team', 'Nation')
.select('Nation', 'Top 100', 'Ranking')
```

3. [17.0 points] Do try this at home

There is a box of four balls: two white and two red. Soumyadeep performs the following experiment.

- Draw one ball from the box at random.
- If the ball is white, keep the ball in the box. If the ball is red, remove the ball from the box.
- Draw a second ball from the box at random.

(a) [3.0 pts] In the box below, write an arithmetic expression for the probability of *drawing one white ball* during the experiment.

P(one W) = P(WR) + P(RW) = $\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{2}{3} = 1/4 + 1/3 = 7/12$

(b) [3.0 pts] Complete the body of the function `game()`, which simulates one round of the game described above and returns the *number of white balls* that the player has drawn in total once the game has finished.

```
def game():

  box = make_array('W','W','R','R')

  first_ball = np._____[1]_____

  if _____[2]_____:
    box = _____[3]_____

  second_ball = _____[4]_____

  hand = make_array(first_ball, second_ball)

  return _____[5]_____
```

i. [0.9 pts] Blank [1]

random.choice(box)

ii. [0.6 pts] Blank [2]

first_ball == 'R'

iii. [0.6 pts] Blank [3]

make_array('W', 'W', 'R')

iv. [0.3 pts] Blank [4]

np.random.choice(box)

v. [0.6 pts] Blank [5]

```
sum ( hand == 'W' )
```

(c) [5.0 pts] Simulate 1,000 rounds of the game in the box below and save the results into an array called `game_results`.

```
game_results = make_array()
for i in np.arange(1000):
    game_results = np.append( game_results , game() )
```

(d) [2.0 pts] Soumyadeep would now like to visualize the data in `game_results`. Namely, he would like to see the different possibilities for the number of white balls he could receive while playing this game (0, 1 or 2) on one axis, and on the other axis, the *number of times* he actually got each of the possibilities on the other. Which plot type(s) are most appropriate for this purpose? *Select one.*

● Bar chart ○ Histogram ○ Line plot ○ Scatter plot

(e) [4.0 pts] How many times out of the experiments that Soumyadeep ran should he expect to have drawn one white ball? In the box below:

- Write an arithmetic expression for the answer.
- Justify your answer with two to three sentences.

Soumyadeep should expect to have drawn a white ball in $1,000 \times 7/12$ of the experiments (so a bit over 500).

By the Law of Large Numbers, the proportion of times that he draws one white ball should approach 7/12, the theoretical probability of drawing one white ball, after a large number of trials.

4. [14.0 points] Jury Duty

Taylor v. Louisiana (1975) was a landmark decision handed down by the U.S. Supreme Court. The decision centered around the case of a man named Billy J. Taylor, who was indicted in St. Tammany Parish, Louisiana for a kidnapping charge. Read the following excerpt, which gives more background on the case, and answer the questions below.

> *"Appellant, Billy J. Taylor... alleged that women were systematically excluded [from the panel of persons who were brought to the courthouse to have a chance to sit on the jury] and that he would therefore be deprived of what he claimed to be his federal constitutional right to "a fair trial by jury of a representative segment of the community. . .*
>
> *. . . 53% of the persons eligible for jury service in these parishes were female . . . During the period from December 8, 1971, to November 3, 1972, 12 females were among the 1,800 persons drawn [to come to the courthouse for a chance to sit on the jury] in St. Tammany Parish."*

(a) [5.0 pts] Consider conducting a hypothesis test to determine whether Taylor received an unfair jury pool or not. What is an appropriate null hypothesis for the test?

- ○ The 1,800 persons selected for juries in St. Tammany parish during the given time period were chosen unfairly; specifically, the jury panels were biased against women.
- ● The 1,800 persons selected for juries in St. Tammany parish during the given time period were chosen fairly.
- ○ None of these hypotheses are appropriate.

(b) [3.0 pts] What is the value of the observed test statistic for this hypothesis test?

○ 0.53 ● 12/1800 (roughly 0.0067) ○ 0.50 ○ Something else

(c) [3.0 pts] Describe how to obtain simulated test statistics under the null hypothesis.

- ○ Draw a sample of 1,800 people at random from a population where ((12/1800)*100)%, or roughly 0.67 percent, of people are women and then calculate the proportion of people in the sample that are women.
- ○ We currently have a sample of 1,800 people where 53% of people are women and 47% of them are men. Shuffle the sex label in our table, and then calculate the difference in the proportions of each sex that are selected to a jury panel.
- ● Draw a sample of 1,800 people at random from a population where 53% of people are women and then calculate the proportion of people in the sample that are women.
- ○ None of these methods to simulate the null hypothesis are appropriate.

(d) [3.0 pts] Suppose we decide to reject the null hypothesis. What does this mean?

- ○ The alternative hypothesis is true; the null hypothesis is false.
- ● If the null hypothesis is true, the set of 1,800 persons that were selected for a chance at jury duty was unlikely.
- ○ Based on the 1,800 persons selected for a chance at jury duty, the probability of the null hypothesis being true is small.
- ○ The trial being in St Tammany parish versus a neighboring parish led to the specific set of persons being selected.

Initials:

5. [10.0 points] Skittles

(a) I walked to 7-Eleven today and bought a bag of Skittles. Before eating them, I took all of the candies out of the bag to see what proportion of each of the five possible candy colors I got. The results I got can be found in the table below.

Color	Distribution in my bag
Red	0.14
Orange	0.22
Yellow	0.18
Green	0.20
Purple	0.26

(b) [2.0 pts] I would like to test the hypothesis that the distribution of colors in my bag is different from the distribution *where the colors appear in equal proportions throughout the bag.* In the box below, compute the value of the observed test statistic for this test.

.14 − .20 = −.06
.22 − .26 = .02
.18 − .20 = −.02
.20 − .20 = .00
.26 − .20 = .06

Adding the positive parts together gives us the value of 0.08 for the observed test statistic.

(c) [7.0pts] Richard sees me running the hypothesis test and decided to collect his own data. He goes to 7-Eleven and buys his own Skittles bag. The value of the observed test statistic that Richard calculates is `0.07`.

i. [4.0 pts] Richard then simulates 10 test statistics under the null hypothesis and reports them below.

T F F T T T T T F F
`0.07, 0.06, 0.05, 0.09, 0.11, 0.08, 0.08, 0.08, 0.04, 0.03` ≥ 0.07

In the box below, write the p-value for Richard's test. You may write it in decimal or fraction form.

~~1/10~~ 6/10

ii. [4.0 pts] Assuming that Richard set a p-value cutoff of 0.10 before running the test, what decision should he make?

- ○ **Reject** the null hypothesis and conclude that the distribution of colors in his bag **is different** than the distribution where the colors appear in equal proportions throughout the bag.
- ○ **Reject** the null hypothesis and conclude that the distribution of colors in his bag **is no different** than the distribution where the colors appear in equal proportions throughout the bag.
- ○ **Retain** the null hypothesis and conclude that the distribution of colors in his bag **is different** than the distribution where the colors appear in equal proportions throughout the bag.
- ● **Retain** the null hypothesis and conclude that the distribution of colors in his bag **is no different** than the distribution where the colors appear in equal proportions throughout the bag.

6. [0.0 points] Congratulations! You have completed the Midterm Exam.

Make sure:

- that you have written your initials on each page of the exam.
- you have filled in bubbles and squares completely rather than having used a check mark, cross or any other mark.
- you double check that you have not skipped over any questions!

Once you've done this, *and you've 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.

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