Data C8
Fall 2025

Foundations of Data Science

Midterm

7:10–9:00pm Friday, October 17th 2025

Print Your Name: ______

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 (Write N/A if in Self-Service): ______

Instructions

You have **110 minutes** to complete the exam. There are **5 questions** and **17 pages** on this exam, including this cover page.

Question	1	2	3	4	5	Total
Points	21	21	20	21	17	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**.

○ Unselected option (completely unfilled)

● Single option selected (completely filled)

Questions with **square boxes**: you may select **1 or more choices**.

■ 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.

ii. [1.0 pt] Imagine that the spreadsheets described in the excerpt were loaded into Python as tables. When coding the "*basic grouping method*" described in the excerpt, which of the aggregation methods taught in this class would be appropriate?

○ `pivot` only　　○ `group` only

○ Both `pivot` and `group`　　○ Neither of these methods are appropriate.

(c) [3.0 pts] What will the following Python expression output to the screen?

```
make_array(8, 24, 8) + np.arange(8, 24, 8)
```

○ `array([16, 40, 32])`

○ `array([16, 48, 16])`

○ `array([8, 24, 8, 8, 16, 24])`

○ `array([8, 24, 8, 8, 24, 8])`

○ This expression produces an error.

(d) [3.0 pts] What will the following Python expression output to the screen?

```
make_array(False, False, True) == np.count_nonzero(make_array(True, False, False))
```

○ `True`　　○ `False`

○ `array([False, False, True])`　　○ `array([True, True, False])`

○ `array([False, False, False])`　　○ This expression produces an error.

(e) For each of the following scenarios, choose the sampling method involved from the items below.

A	Deterministic sample	**B**	Convenience sample
C	Random sample without replacement	**D**	Random sample with replacement

i. [1.0 pt] Rolling a fair, six-sided die 100 times.

○ **A**　○ **B**　○ **C**　○ **D**

ii. [1.0 pt] Simulating 900 pea plant growths under the null hypothesis that each pea plant has a 75 percent chance of blossoming with purple flowers, independent of other plants.

○ **A**　○ **B**　○ **C**　○ **D**

iii. [0.5 pts] A UC Berkeley professor recruits participants for a research study by posting flyers around campus.

○ **A**　○ **B**　○ **C**　○ **D**

iv. [0.5 pts] Cal Athletics hires Qualtrics to conduct a survey on student opinions regarding the direction of the football program. Qualtrics is given a roster of the student ID and randomly selects 5,000 students to participate.

○ **A**　○ **B**　○ **C**　○ **D**

2. [21.0 points] Berkeley Car Crashes

The California Highway Patrol (CHP) compiles data of vehicle accident reports throughout the state. For this problem, you will be working with a table called **`berkeley`**. This table contains information on all 565 crash reports which took place in the city of Berkeley from January 2025 through September 2025. A three-row excerpt of the **`berkeley`** table lies below.

ID	Time	Type	Day of Week	Highway	Latitude	Longitude	Road 1	Road 2
4591937	Afternoon	Side Swipe	Friday	`True`	37.8821	−122.308	I-80 E/B	Buchanan
4649754	Morning	Rear End	Tuesday	`True`	37.8807	−122.296	Gilman	San Pablo
4742576	Late Night	Rear End	Saturday	`False`	37.8645	−122.302	Bolivar	Potter

(a) **[5.0 pts]** *Select all columns* that are numerical variables.

□ ID
□ Time
□ Type
□ Day of Week
□ Highway
□ Latitude
□ Longitude
□ Road 1
□ Road 2

(b) **[5.0 pts]** Based on the excerpt and the information given, which of the following tasks are appropriate to complete using the `berkeley` table when it comes to the first nine months of 2025? *Select all that apply.*

□ Visualizing the distribution of accident types.

□ Conducting a hypothesis test to conclude whether the proportion of highway accidents is equal to 0.5.

□ Finding the name of the road most commonly involved in an accident.

□ Finding the exact time (hours and minutes) of each crash that occurred on a Friday.

(c) Write Python code to make a visualization which displays a rough map of the accidents.

```
berkeley.______[A]______(______[B]______)
```

i. **[2.0 pts]** Fill in blank `[A]`.

ii. **[2.0 pts]** Fill in blank `[B]`.

(d) An intersection is where two roads meet. Write Python code to construct a three-column table that contains the 5 non-highway intersections that had the most crashes, as well as how many crashes occurred at each of these intersections.

```
no_highway_intersections = berkeley.____[A]____(____[B]____).____[C]____(____[D]____)
five_most_crashes = no_highway_intersections.____[E]____(____[F]____).____[G]____(____[H]____)
```

i. **[0.5 pts]** Fill in the blank [A].

ii. **[0.5 pts]** Fill in the blank [B].

iii. **[0.5 pts]** Fill in the blank [C].

iv. **[0.5 pts]** Fill in the blank [D].

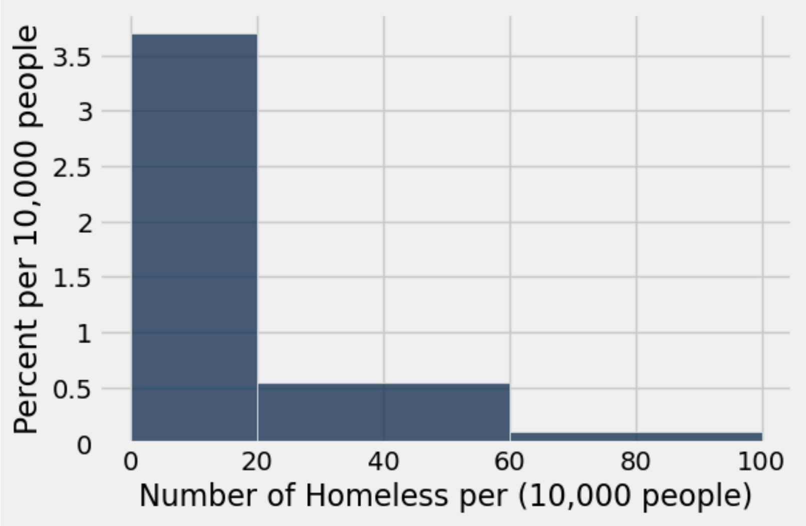
v. **[0.5 pts]** Fill in the blank [E].

vi. **[0.5 pts]** Fill in the blank [F].

vii. **[0.5 pts]** Fill in the blank [G].

viii. **[0.5 pts]** Fill in the blank [H].

(c) The following is the completed visual from **part (b)**, with labels edited for better readability. The data are separated into three bins: [0, 20), [20, 60) and [60, 100). Answer the items below based on this visual.



i. [3.0 pts] Which bin is most dense?

○ [0, 20) ○ [20, 60) ○ [60, 100) ○ An answer cannot be determined.

ii. [3.0 pts] Which bin has the most states in it?

○ [0, 20) ○ [20, 60) ○ [60, 100) ○ An answer cannot be determined.

iii. [2.0 pts] Roughly how many states have between 20 and 60 people experiencing homelessness per 10, 000 people?

○ 5 ○ 10
○ 20 ○ 40
○ An answer cannot be determined.

iv. [2.0 pts] Roughly what percentage of states have between 20 and 40 people experiencing homelessness per 10, 000 people?

○ 0.5 ○ 5
○ 10 ○ 20
○ An answer cannot be determined.

v. [1.0 pt] Consider changing the visual so that the [0, 20) bin is split into [0,10) and [10, 20) bins.
Which of the following statements are true? *Select all that apply.*

□ The combined area of the [0,10) and [10, 20) bins will be equal to the area of the original [0, 20) bin.
□ The height of the [0,10) bin may be 0 percent per 10,000 people.
□ The height of the [10,20) bin may be greater than the height of the original [0,20) bin.

(c) Fill in the blanks of the following statement using the options below. It is possible to use an item more than once.

“By the ________[1]________, ________[2]________ that Jaina wins the game should be ________[3]________ that Jaina wins the game, which is equal to ________[4]________.

A	Inference facet of data science	**B**	Central Limit Theorem	**C**	Law of Large Numbers
D	the proportion of times	**E**	the number of times	**F**	the theoretical probability
G	close to the proportion of times	**H**	equal to the empirical probability	**I**	close to the empirical probability
J	equal to the theoretical probability	**K**	close to the theoretical probability	**L**	$\frac{4}{12}$
M	$\frac{6}{12}$	**N**	$\frac{6}{16}$	**O**	$\frac{10}{16}$

i. [1.0 pt] Fill in blank [1].

○ A ○ B ○ C ○ D ○ E ○ F
○ G ○ H ○ I ○ J ○ K ○ L
○ M ○ N ○ O

ii. [1.0 pt] Fill in blank [2].

○ A ○ B ○ C ○ D ○ E ○ F
○ G ○ H ○ I ○ J ○ K ○ L
○ M ○ N ○ O

iii. [1.0 pt] Fill in blank [3].

○ A ○ B ○ C ○ D ○ E ○ F
○ G ○ H ○ I ○ J ○ K ○ L
○ M ○ N ○ O

iv. [1.0 pt] Fill in blank [4].

○ A ○ B ○ C ○ D ○ E ○ F
○ G ○ H ○ I ○ J ○ K ○ L
○ M ○ N ○ O

Toby wants to try something new. He comes up with another two-player game (Game 2) that has ten index cards, numbered 1 through 10. The rules for one round of this game are below.

Game 2 Rules

1. The ten-card deck is shuffled.
2. Three cards are picked out, one by one.
   - The first card goes to Toby.
   - The second card goes to Jaina.
   - The third card is set aside.
3. Toby privately looks at the number on his card.
   - If the number is 5 or less, he takes Jaina’s card as his new card.
     - Jaina then takes the third card (that was previously set aside) as her new card.
4. Toby and Jaina flip over their cards and compare them. The player with the highest numbered card wins.