Data C8
Spring 2026

Foundations of Data Science

Midterm

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 TA's Name (Write N/A if in Self-Service): _____

Instructions

You have to complete the exam. There are **7 questions** and **10 pages** on this exam, including this cover page.

Question	1	2	3	4	5	6	7	Total
Points	16	10	20	16	10	28	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;
- 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."

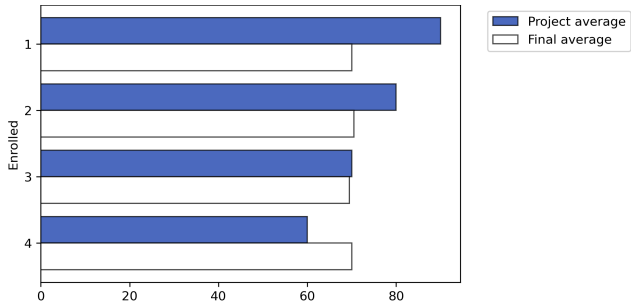
Sign Your Name: _____

This page intentionally left blank

The exam begins on the next page.

1. [16.0 points] Late Enrollment

A course instructor was curious whether student performance in their course was associated with the week in which the student enrolled. Students can enroll on time in week 1 or enroll late in weeks 2, 3, or 4. This chart shows the average scores on a semester-long course `Project` and the `Final` exam for students who `Enrolled` in each of these 4 weeks.



(a) [6.0 pts] What does this chart show? Select all that apply.

- ☐ There is an association between the week they Enrolled and Project score for students in the course.
- ☐ There is *no* association between the week they Enrolled and Project score for students in the course.
- ☐ There is an association between the week they Enrolled and Final score for students in the course.
- ☐ There is *no* association between the week they Enrolled and Final score for students in the course.
- ☐ There is an association between Project score and Final score for students in the course.
- ☐ There is *no* association between Project score and Final score for students in the course.

(b) [4.0 pts] Does this chart show that the students who enrolled in week 4 would have received a higher average project score if they had instead all enrolled in week 1?

- ◯ Yes, because the average project score is lower for students who enrolled in week 4 than week 1.
- ◯ Yes, because enrolling in week 1 gives students more time to complete the project than enrolling in week 4.
- ◯ Yes, because of the clear association between the week they Enrolled and Project score.
- ◯ No, because not all students who enrolled in the same week had the same project score.
- ◯ No, because the week 1 students had more time to complete the project than the week 4 students.
- ◯ No, because that's a causal statement, and this chart doesn't show whether starting in week 4 caused lower scores.

(c) [2.0 pts] The professor wants to design an experiment showing that the difference in average Project scores among those who enroll in week 1 and week 4 is **caused** by when they enroll. Which experiment might show this?

- ◯ Randomly assign students to the week they enroll instead of letting them choose.
- ◯ Randomly assign each student who enrolls during week 1 to one of two different lecture times.
- ◯ Randomly match each student who enrolls in week 4 to one who enrolled in week 1 and compare their Project scores.
- ◯ Students who enroll in week 4 are given an additional 3 weeks to complete the project.

(d) [4.0 pts] A Table called `scores` contains one row per student with columns for the week they **`Enrolled`**, their **`Project`** score, and their **`Final`** score. All values in the table are integers. Write an expression to create the bar chart above.

2. [10.0 points] Percentage Change

Definition. The *percentage change* from an `old` value to a `new` value is the difference (`new` − `old`) as a percentage of the `old` value. For example, the percentage change from 60 to 63 is 5, since 3 is 5% of 60. The percentage change from 60 to 54 is −10, since −6 is −10% of 60.

Define a function `pchange` that takes a table `t` and a column label `c`. It returns an array of the percentage change between values in column `c` for all adjacent pairs of rows in Table `t`. **Assume that all numbers in the table `t` are positive.**

```
def pchange(t, c):
    "Return an array of percentage changes in column c of table t."

    changes = np.diff(t.column(c))

    return _______ / t._______._______
             (a)          (b)     (c)
```

Example: `temps` is a table of high and low temperatures in Berkeley in January 2026.

```
temps = Table.read_table('temps.csv')
temps.show(4)
```

day	high	low
1	60	49
2	63	55
3	62	55
4	55	52

... (27 rows omitted)

```
pchange(temps.take(np.arange(4)), 'high')  # Call pchange on just the first 4 rows
```

```
array([ 5., -1.587, -11.29])
```

(a) [2.0 pts] Fill in blank (a) in the definition of `pchange`.

- ◯ `percentile(changes)`
- ◯ `100 * changes`
- ◯ `100 * make_array(changes)`
- ◯ `make_array(100, 100, 100, 100) * changes`

(b) [6.0 pts] Fill in blank (b)? You **may not** write `[` or `]` or `for` or `:`.

(c) [2.0 pts] Fill in blank **(c)**.

- ◯ `column(c)`
- ◯ `column(c).diff()`
- ◯ `select(c)`
- ◯ `select(c).diff()`

3. [20.0 points] Study Groups

A new Data 8 initiative is attempting to give every study group their own dedicated room. The `students` table has a row for each student enrolled in Data 8 and columns for their **`name`** (`str`) and **`study`** group number (`int`). Students with the same **`study`** group number are in the same study group. The `rooms` table has one row per room and columns for its **`location`** (`str`) and **`capacity`** (`int`).

`students.show(4)`

name	study
Natalie Coughlin	2
Missy Franklin	7
Alex Morgan	2
Collin Morikawa	4

... (1410 rows omitted)

`rooms.show(4)`

location	capacity
Evans 60	3
Evans 71	6
Evans 87	4
Evans B5	5

... (246 rows omitted)

(a) [2.0 pts] Which columns in these two tables correspond to categorical variables? Select all that apply.

☐ `name` ☐ `study` ☐ `location` ☐ `capacity`

(b) [6.0 pts] Write an expression that evaluates to the size of study group 2. You **may not** write `group`.

```
group2_size = _____
```

Write an expression that evaluates to an array of all the locations that have capacity larger than the size of Group 2. Assume `group2_size` is defined correctly.

```
rooms.where(_______).column(_______)
             (c)             (d)
```

(c) [2.0 pts] Fill in blank (c).

◯ `capacity, are.above(group2_size)`

◯ `capacity, are.above('group2_size')`

◯ `'capacity', are.above(group2_size)`

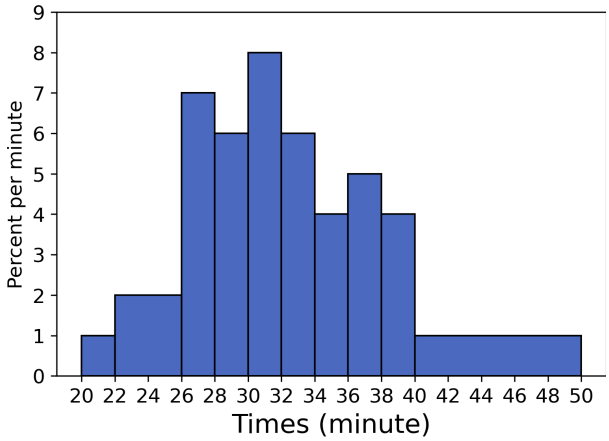
◯ `'capacity', are.above('group2_size')`

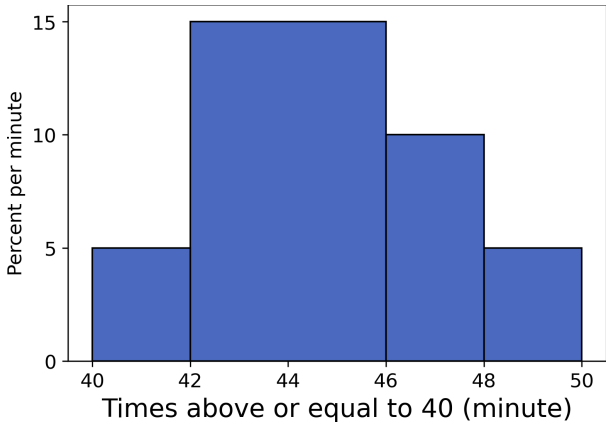
(d) [2.0 pts] Fill in blank (d).

(e) [8.0 pts] Write code to create a table called `fit` with columns **`study`** and **`location`**. It should have one row for every possible pair of a **`study`** group number and a room **`location`**, where the size of the study group is exactly the capacity of the room. It's ok for the `fit` table to have other columns as well. It is possible to solve this in one line, but you may write more than one. You **may not** write `def`, `for`, `if`, `[`, `]`, or `item` or use Python syntax not covered in Data 8.

4. [16.0 points] Berkeley 5k

The Berkeley 5k race had 200 runners who finished in times ranging from 20 to 50 minutes. Their finishing times are stored in a `Table` called `t` with one column labeled `Times`. A histogram of `Times` in `t` appears to the left, and a histogram of `Times` in `t.where('Times', are.above_or_equal_to(40))` appears to the right. Assume that the height of every bar is an integer.





(a) [2.0 pts] What **percentage** of runners finished in at least 30 but less than 34 minutes?

◯ 6% ◯ 7% ◯ 8% ◯ 12% ◯ 14% ◯ 16% ◯ 24% ◯ 28% ◯ 48% ◯ 56% ◯ 72% ◯ 84%

(b) [2.0 pts] If the left histogram had just one bin from 20 to 28, what would this bin's height be?

◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7 ◯ 8 ◯ 10 ◯ 12 ◯ 20 ◯ 30 ◯ 40

(c) [4.0 pts] How many runners finished in less than 46 minutes?

The prize for finishing in less than 26 minutes is a **gold** medallion that costs $10 to make. The prize for finishing in at least 26 but less than 28 minutes is a **blue** medallion that costs $2 to make. Runners finishing in 28 minutes or more get a **certificate** that costs $1 to make. The race ends after 50 minutes, so all runners have a time less than 50 minutes.

(d) [4.0 pts] What is the total cost to make the **gold and blue medallions** given to these 200 runners?

(e) [4.0 pts] **Challenging**: Using the tables `t` and `prizes` below, fill in the blanks in this expression that computes the total cost to make all the prizes given to the runners.

`t.show(3)`

Times
21.38
23.56
27.01

... (197 rows omitted)

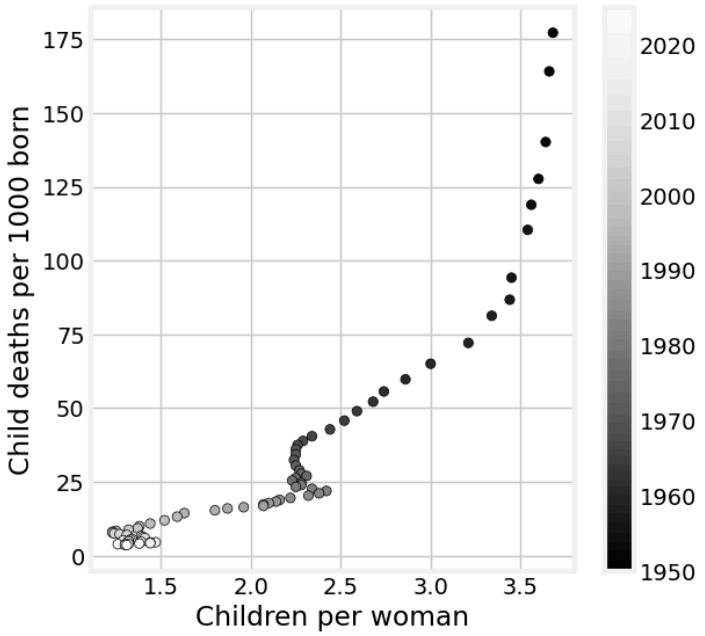
`prizes`

Min Time	Prize	Cost
0	Gold Medallion	10
26	Blue Medallion	2
28	Certificate	1
50	Nothing	0

```
sum(t._____('Times', _____________________________________ ).column(1) * prizes.column('Cost'))
```

5. [10.0 points] Project 1 Review

(a) [4.0 pts] Which of the following can be concluded from the chart below of the data in the table `poland_since_1950` that has one row per year and columns containing fertility and child mortality statistics? Select all that apply.



- ☐ In Poland, having fewer children caused a larger proportion of them to survive childhood.
- ☐ In Poland, increasing childhood survival rates caused women to have fewer children on average.
- ☐ In Poland, as childhood survival rates increased over time, fertility rates decreased.
- ☐ In Poland, as childhood survival rates decreased over time, fertility rates increased.

(b) [6.0 pts] The `poverty` table rows contain estimates of the poverty rate in a country in a year. The result of `poverty.show(3)` appears below.

geo	time	poverty_percent
alb	1996	0.2
alb	2002	0.73
alb	2004	0.53

... (1096 rows omitted)

Complete the code to define `latest_poverty`, a three-column table with one row for each country appearing in the `poverty` table. The first column should contain the 3-letter code for the country. The second column should contain the most recent year for which a poverty rate is available for the country. The third column should contain the poverty rate in that year.

You may use the `first` function, which is defined below.

```
def first(values):
    return values.item(0)


latest_poverty = poverty._____(_____________________________________).group(_____________________)
```

6. [28.0 points] The Goat Whisperer

In Monty Hall's game show, there is a car behind 1 door and a goat behind each of the other 2. The contestant chooses any door (which remains closed for now). The host Monty Hall then opens a different door revealing a goat. The contestant can either choose to open their original door or switch doors and open the remaining one. If the door they open reveals a car, they win.

Assuming a contestant always switches doors and therefore has a 2/3 chance of winning each game...

(a) [2.0 pts] What is the chance of **winning** 3 games out of 3?

- ◯ $\frac{2}{3}$
- ◯ $\frac{2}{3} + \frac{1}{3}$
- ◯ $\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$
- ◯ $\frac{2}{3} \times \frac{2}{3} \times \frac{2}{3}$
- ◯ $3 \times \frac{2}{3}$

(b) [2.0 pts] What is the chance of **winning** 2 games out of 3?

- ◯ $\frac{2}{3} \times \frac{2}{3} \times \frac{1}{3}$
- ◯ $\frac{2}{3} \times \frac{2}{3} \times \frac{2}{3}$
- ◯ $\frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3}$
- ◯ $3 \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3}$
- ◯ $3 \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3}$

(c) [4.0 pts] What is the chance of **losing** at least 1 game out of 5? You can write an expression and leave it unsimplified.

The most recent contestant on Monty Hall's game show (who trains goats professionally) played 12 games, switched doors every time, and won the car 11 out of 12 times! Monty thinks a contestant should win only two thirds of their games, and he decides to conduct a hypothesis test to determine whether this goat training contestant won more often than they should because they were somehow cheating. He chooses a **p-value threshold of 0.01** and the **number of games won** (out of 12) as the test statistic.

(d) [2.0 pts] Which is an appropriate **null hypothesis** for this hypothesis test?

- ◯ The contestant has information about where the goats are, and is using that to her advantage.
- ◯ Monty Hall has information about where the goats are, and is using that to select a door to open.
- ◯ The goats are making noise that the contestant can hear.
- ◯ The game is being played fairly.
- ◯ The car is placed randomly behind one of the doors.

(e) [2.0 pts] Which is an appropriate **alternative hypothesis** for this hypothesis test?

- ◯ The contestant has information about where the goats are, and is using that to her advantage.
- ◯ Monty Hall has information where the goats are, and is using that to select a door to open.
- ◯ The goats are making noise that the contestant can hear.
- ◯ The game is being played fairly.
- ◯ The car is placed randomly behind one of the doors.

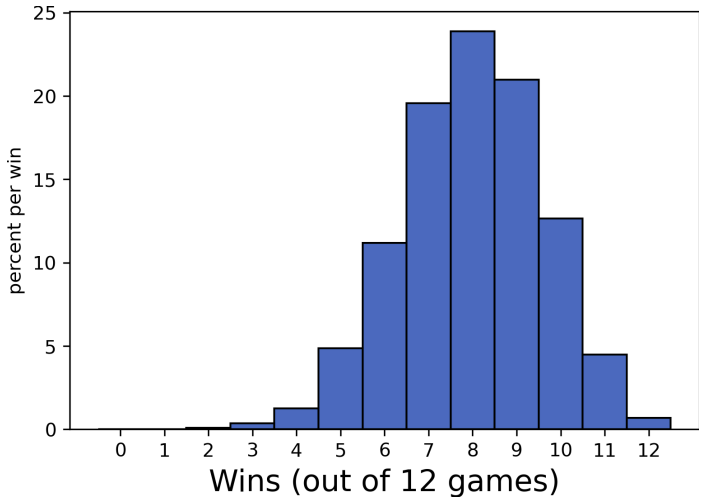
(f) [2.0 pts] What other test statistics would have been equally good choices for this hypothesis test? Mark all that apply.

- ☐ The proportion of games won
- ☐ The number of games played
- ☐ The difference between the proportion of games won and 2/3
- ☐ The absolute value of the difference between the proportion of games won and 2/3

(g) [8.0 pts] Fill in the blank in this simulation of the test statistic under the null hypothesis. Write your answer in the box. If your answer is long, you may write it on multiple lines, but your answer should be a one-line expression.

```
wins_array = make_array()
for i in np.arange(10000):
    wins = _____
    wins_array = np.append(wins_array, wins)
```

The figure below is a histogram created by the following expression:
`Table().with_column('Wins (out of 12 games)', wins_array).hist(bins=np.arange(-0.5, 12.6, 1))`.



(h) [4.0 pts] Write a Python expression using `wins_array` to compute the p-value for this hypothesis test.

(i) [2.0 pts] Can Monty conclude that the contestant was cheating from this histogram?

- ◯ Yes, because 8 out of 12 wins is the most likely outcome.
- ◯ Yes, because 11 or more out of 12 wins is unlikely.
- ◯ No, because 11 out of 12 wins is possible.
- ◯ No, because 11 or more out of 12 wins is not unlikely enough.
- ◯ Maybe; it is very hard to tell from the histogram.

7. Just for Fun (and Completely Optional)

In this 4-door variant of Monty Hall's gameshow, there is a car behind 1 door and goats behind the other 3. The contestant chooses any door (which remains closed for now), then padlocks a different door (which remains closed forever). The host Monty Hall then opens a third door revealing a goat.

(a) What is the chance that the car is behind the fourth door (which was not initially chosen, padlocked, or opened by Monty)?

◯ 1/4 ◯ 1/3 ◯ 1/2 ◯ 2/3 ◯ 3/4

(b) Does using the padlock increase the contestant's chance of winning (assuming they always choose one of the most likely doors) compared to playing without it? If they played without the padlock, then Monty would reveal a goat behind any one of the three doors that the contestant didn't choose initially, and then the contestant would pick a door to open.

◯ Yes, the padlock increases their chance of winning.

◯ No, the padlock does not increase their chance of winning.

(c) Draw a picture of your favorite part of Data 8 so far.