# Fall 2023 Practice Midterm Exam

Foundations of Data Science

Name

## Instructions

- Make sure you have a copy of the Midterm Reference Guide, and rip off the Table Reference page attached at the end of the exam.
- Select the correct response(s) or provide a written response depending on the question type. If a prompt implies you should write code, then you can provide your own code or use the provided template. Try to provide your responses in the spaces provided. If you find that you need additional space, write your extended response(s) on a blank sheet of paper and number them, so we can connect your response to the question.
- You can assume the following code has been run, when you are writing your responses for Section B:

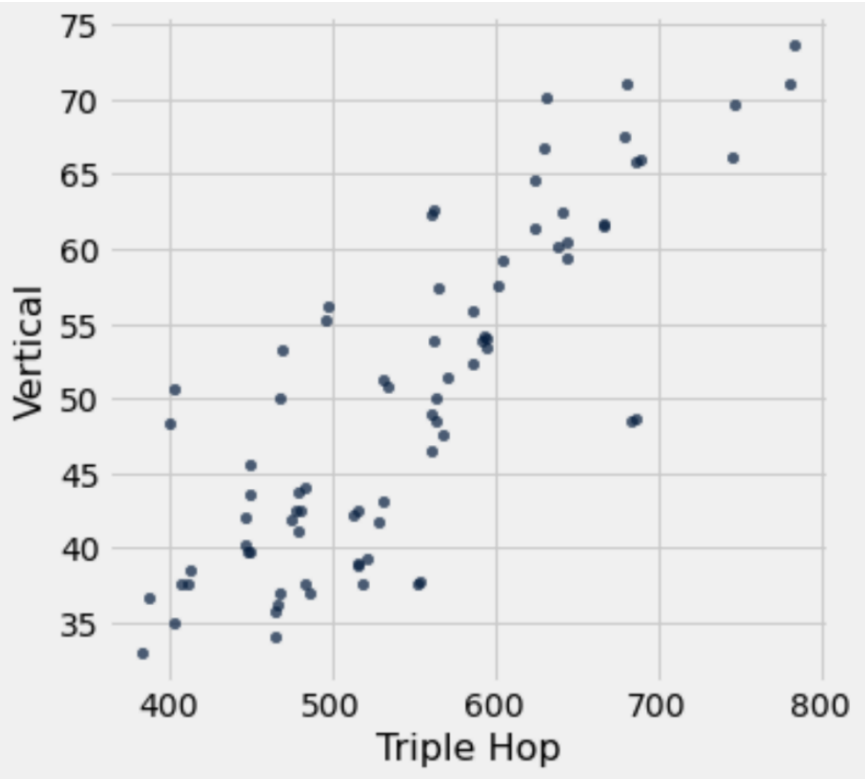
```
from datascience import *
import numpy as np
import matplotlib+
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
```

- Once you have made an attempt on the exam, upload a PDF of your attempt to Canvas for a Complete score. Sample solutions will be released after the study session held on Tuesday evening.

This practice exam contains questions from past DATA 8 midterm exams in addition to a few additional or modified questions.

# Section A

1. A medical institute that specializes in sports medicine has recorded data on athletes with leg injuries. The variables are the distance that the athlete achieved in a test called the triple hop, and how high the athlete could jump vertically. Both distances were measured in centimeters The scatter plot below has a point for each of the athletes.



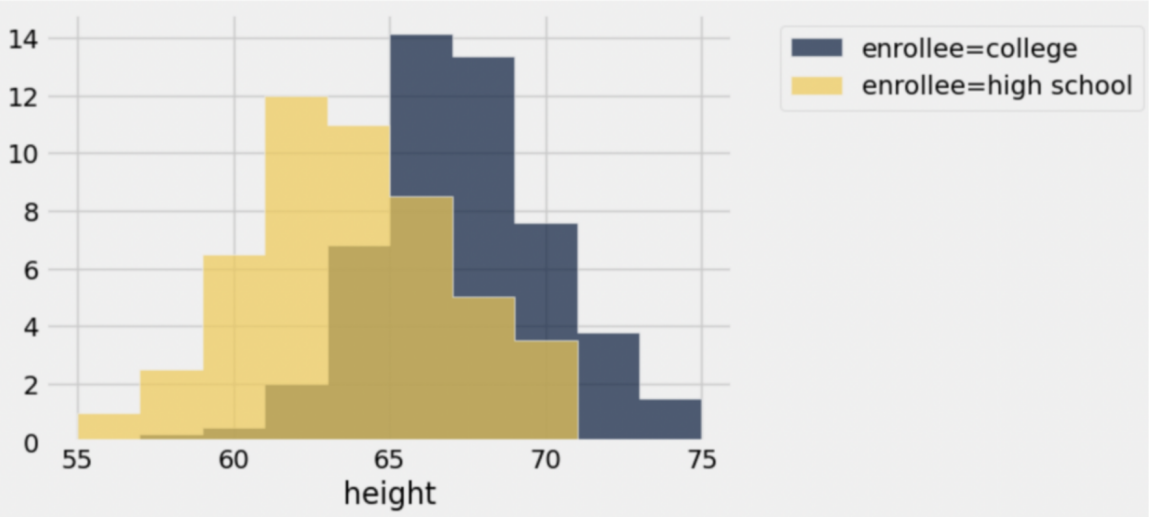

Pick all the conclusions that can be drawn from the scatter plot. Select all that apply.

- ☐ More than half the athletes jumped less than 60 centimeters vertically.
- ☐ Most of the athletes whose triple hop distances were longer than average also jumped higher than average.
- ☐ If athletes were to increase their triple hop distances then they would be able to jump higher.
- ☐ If athletes were to increase the heights of their vertical jumps, they would be able to triple hop longer distances.
- ☐ None of the above conclusions can be drawn from the scatter plot.

2. Dylan, a Data 8 student, wonders if there is a difference in heights between his high school and college classmates. Among Data 8 students, he collects a random sample of 100 high school students and a random sample of 200 college students, and organizes his data in a table called `heights` that is previewed on the Table Reference page.

   He created the following histogram based on the data in `heights`.

   Note: The $x$-axis of the histogram has a range of $[55, 75]$, and each bin is 2 inches in width.




(a) Unfortunately, Dylan forgot to include the y-axis label in his histogram. Which of the following is an appropriate y-axis label? Select one.

- ◯ Percentage of students
- ◯ Percentage of students per height
- ◯ Percentage of students per inch
- ◯ Rate of change in height

(b) Which of the following is closest to the total number of individuals in our table that are greater than or equal to 61 inches, but less than 63 inches in height? Select one.

- ◯ 14
- ◯ 24
- ◯ 28
- ◯ 32

(c) Using only the conclusions he can draw from his histogram, Dylan claims that he is shorter than at least 75% of high school students, but taller than at least one college student. Which of the following bins may Dylan belong in? Select all that apply.

- ☐ [57 - 59)
- ☐ [59 - 61)
- ☐ [61 - 63)
- ☐ [63 - 65)

(d) Which of the following conclusions can you draw from Dylan's overlaid histograms? Select all that apply.

- ☐ Assuming we know the height of each bin, we can calculate the percentage of individuals in any range of heights.
- ☐ College Data 8 students are generally taller than high school Data 8 students.
- ☐ The area of both histograms together sum to 100%, so the area of the blue and yellow histograms are $\frac{100}{300} \times 100$ and $\frac{200}{300} \times 100$, respectively.
- ☐ There is an association between a Berkeley student's height and age.
- ☐ College students enrolled in Data 8 are generally taller than high school students in Data 8 because they are older.

3. Kanu has been hired as a data scientist for Cal Rec Sports! His primary task is to analyze aquatic equipment rental data for the Berkeley Marina. Kanu's team presents him with a `rentals` table that contains the daily number of kayak and windsurfing board rentals, along with information pertaining to the day that the data were collected. In the Table Reference, you can see the first few rows of `rentals`.

   (a) Does a higher wind speed lead to a larger number of windsurfing board rentals? Choose which single visualization is most useful for answering this question. Select one.

   ◯ Histogram
   ◯ Overlaid Histograms
   ◯ Scatter Plot
   ◯ Overlaid Scatter Plots
   ◯ Bar chart
   ◯ Line Plot
   ◯ None of the above

   (b) Do days with more kayak rentals have fewer windsurfing board rentals? Choose which single visualization is most useful for answering this question. Select one.

   ◯ Histogram
   ◯ Overlaid Histograms
   ◯ Scatter Plot
   ◯ Overlaid Scatter Plots
   ◯ Bar chart
   ◯ Line Plot
   ◯ None of the above

   (c) How does the distribution of kayak rentals on weekdays compare to the distribution of kayak rentals on weekends? Choose which single visualization is most useful for answering this question. Select one.

   ◯ Histogram
   ◯ Overlaid Histograms
   ◯ Scatter Plot
   ◯ Overlaid Scatter Plots
   ◯ Bar chart
   ◯ Line Plot
   ◯ None of the above

4. Suppose we have discovered an association between two variables in a dataset. Which of the following would be the best way to test whether it is causal? Choose one.

   ◯ Use hypothesis testing to check whether the association is statistically significant.

   ◯ Run a randomized controlled experiment.

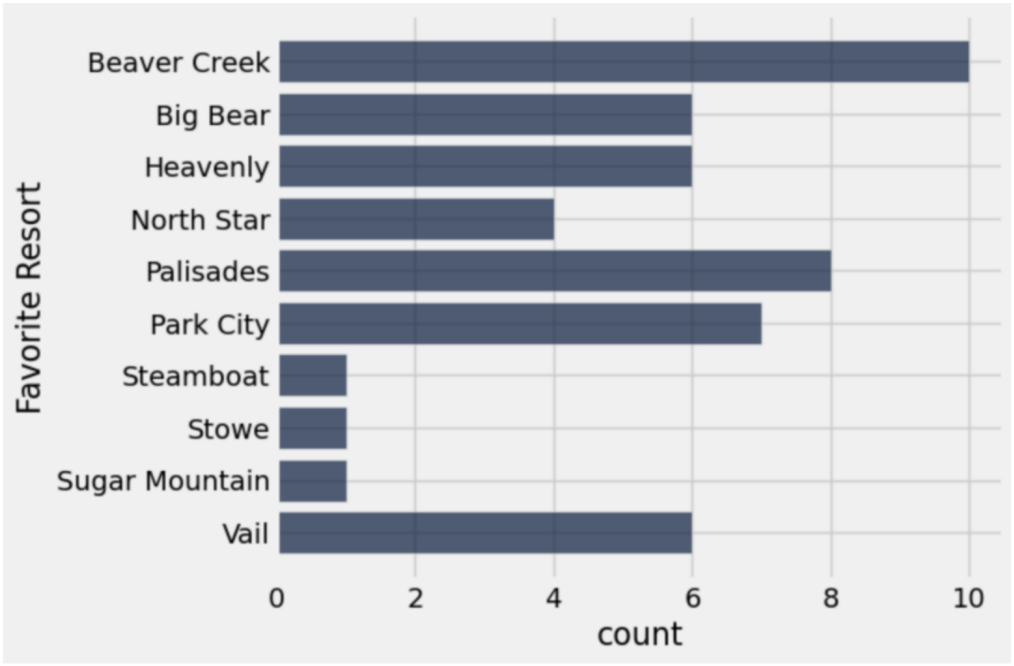
   ◯ Brainstorm some potential confounding factors and test whether any of them has an association with both variables.

   ◯ If both variables are numerical, use a scatter plot to check for a trend.

# Section B

5. The table `skiers` previewed on the Table Reference page contains information about the preferences of several skiers from a convenience sample of Data 8 staff.

   (a) Fill in the blanks to generate the following bar chart showing the popularity of everyone's Favorite Resort given in the skiers table. You may find the axis labels helpful.




```
favorite_counts = skiers.____(A)_____(_______(B)______)
______(C)______._____(D)______(___________(E)___________)
```

(b) The height of the skiers was accidentally inputted into the `skiers` table as strings. Create a function called `str_to_int` that takes an input of a skier height in the format of a string and outputs the height as an integer.

For example `str_to_int('71')` will output `71`.

(c) Update the table `skiers` so that the column `'Height (in)'` has integer values, not string values.

Hint. Consider using the `apply` and `with_column` table methods.

(d) Compute the average skier height for all the skiers in the `skiers` table.

6. The table `orders` previewed on the Table Reference page contains information about food orders that members of Data 8 Course Staff have made this semester.

   (a) Create a scatter plot to visualize the relationship between how much the order costs versus how the user rated the order.

   ```
   orders._____(A)_____(_____(B)_____, _____(C)_____)
   ```

   (b) Assign the variable `usually_friends` to `True` if ordering with friends is more common than not ordering with friends and `False` otherwise.

   ```
   with_friends = orders._____(A)_____(_____(B)_____, True).num_rows
   without_friends = _____(C)_____._____(D)_____ - _____(E)_____
   usually_friends = with_friends _____(F)_____ without_friends
   ```

(c) Assign the variable `frugal_user` to the name of the user who has spent the least over the entire semester:

```
frugal_user = (
    orders.group(______(A)______, ______(B)______)
    .sort(______(C)______, descending=______(D)______)
    .column("User").item(0)
)
```

7. For this problem we are considering restaurants around Berkeley. The `restaurant` table contains information about specific restaurants including their distance from campus in miles. There are no duplicate restaurants in this table. Additionally, the `transport` table contains information about how long it takes to get to each restaurant using various modes of transportation. Each restaurant may appear multiple times in this table with different modes of transportation and time in minutes.

(a) Which code snippet would produce a table containing the fastest time for any type of food (e.g., Pizza, Bagels, Boba, ...)?

◯
```
(transport
    .select("Type", "Time")
    .group("Type", min))
```

◯
```
(restaurant
    .join("Restaurant", transport, "Restaurant")
    .select("Type", "Time")
    .group("Type", min))
```

◯
```
(restaurant
    .join("Restaurant", transport, "Restaurant")
    .pivot("Type", "Time"))
```

(b) Which code snippet would produce a table with columns corresponding to each unique transportation mode (e.g., "Bus", "Drive", ...), rows corresponding to each unique restaurant type (e.g., Pizza, Bagels, Boba, ...) and the cells containing the minimum travel time.

◯
```
(restaurant
    .join("Restaurant", transport, "Restaurant")
    .select("Transportation", "Type", "Time")
    .group("Transportation", "Type", min))
```

◯
```
(restaurant
    .pivot("Transportation", "Type", "Time", min)
    .join("Restaurant", transport, "Restaurant"))
```

◯
```
(restaurant
    .join("Restaurant", transport, "Restaurant")
    .pivot("Transportation", "Type", "Time", min))
```

8. Which of the following will be output by running the following block of code?

```
x = 0
if x == 0:
    x = 1
if x == 1:
    x = 2
elif x < 3:
    x = 3
else:
    x = 0
print("x is", x)
```

◯ x is 0

◯ x is 1

◯ x is 2

◯ x is 3

# Section C

9. Berkeley adds a photo of a staff members pet at the end of each lab assignment. Each pet photo is chosen from a collection of 20 pets with 10 cats, 9 dogs, and 1 bird. For each event below, choose the Python expression that evaluates to the probability of that event. Choose all that apply.

   (a) When one pet is chosen at random, the probability that it is either a cat or a bird.

   ☐ `(9 / 20) ** 2`
   ☐ `(10 / 20) * (1 / 20)`
   ☐ `(10 / 20) + (1 / 20)`
   ☐ `1 - (9 / 20) ** 2`
   ☐ `1 - (10 / 20) * (1 / 20)`
   ☐ `1 - ((10 / 20) + (1 / 20))`

   (b) When two pets are chosen at random with replacement, the probability that they are both dogs.

   ☐ `(9 / 20) ** 2`
   ☐ `(10 / 20) * (1 / 20)`
   ☐ `(10 / 20) + (1 / 20)`
   ☐ `1 - (9 / 20) ** 2`
   ☐ `1 - (10 / 20) * (1 / 20)`
   ☐ `1 - (10 / 20) + (1 / 20)`

   (c) When two pets are chosen at random with replacement, the probability that the first is a cat and the second is not.

   ☐ `10 / 20 + 10 / 20`
   ☐ `(10 / 20) * (10 / 20)`
   ☐ `(10 / 20) * (9 / 20) * (1 / 20)`
   ☐ `1 - (10 / 20) * (10 / 20)`
   ☐ `1 - (10 / 20 + 10 / 20)`
   ☐ `1 - (10 / 20) * (9 / 20) * (1 / 20)`

   (d) When two pets are chosen at random with replacement, the probability that the first chases the second. Assume dogs only chase cats, cats only chase birds, and birds don't chase.

   ☐ `(10 / 20) * (10 / 20)`
   ☐ `(19 / 20) * (10 / 20)`
   ☐ `(10 / 20) * (1 / 20) + (9 / 20) * (10 / 20)`
   ☐ `1 - ((9 / 20) * (1 / 20) + (10 / 20) * (9 / 20))`
   ☐ `1 - ((10 / 20) ** 2 + (9 / 20) ** 2 + (1 / 20) ** 2)`
   ☐ `1 - ((10 / 20) ** 2 + (9 / 20) ** 2 + (1 / 20))`

10. The pygmy hippo is a small, reclusive (and cute) hippopotamid type that is native to the forests and swamps of West Africa. Two teams of zoologists set out to estimate the proportion that are male by sampling at random from the population. The first team samples 100 hippos and finds the proportion of males in their sample to be A. The second team samples 40 hippos and finds the proportion of males in their sample to be B. The full population has all 2,500 wild pygmy hippos; the proportion P of males in the population is 50% (but unknown to the zoologists).

(a) Which of the following are more likely than not? Select all that apply.

- ☐ A is smaller than B.
- ☐ A is larger than B.
- ☐ P is closer to A than B.
- ☐ P is closer to B than A.
- ☐ None of these.

(b) Which of the following is largest?

- ◯ The chance that A is above 55%
- ◯ The chance that B is above 55%
- ◯ The chance that A is above 60%
- ◯ The chance that B is above 60%

11. Complete the code below that uses a simulation repeated $10,000$ times to estimate the chance that the average dice outcome when rolling 5 fair 6-sided dice is within 0.5 of 3.5. (That is, larger than 3 and smaller than 4.) For example, the average dice outcome of rolling $(3, 2, 2, 6, 4)$ from the 5 dice is $(3 + 2 + 2 + 6 + 4)/5 = 3.2$, which is within 0.5 of 3.5.

```
def within(x, y, z):
    "Return whether z is strictly within x of y."
    return ____(a)_____

count = 0
num_reps = 10_000

for i in np.arange(___(b)___):
    if within(0.5, 3.5, np.average(____(c)____(__(d)__, __(e)__))):
        count = count + 1

estimate = count / ___(f)_____
```

# Table Reference

## orders

Here is a preview of the table `orders`:

| User | Restaurant | Total | Receiver | Rating | With Friends |
|---|---|---|---|---|---|
| w3ndyk1m | Sharetea | 5.40 | stephaniekeem | 10 | True |
| s_kw33 | Riceful | 10.24 | haileyyebonjung | 2 | False |
| nikkyp | La Burrita | 14.98 | wfurtaco | 7 | True |
| sonyaki55 | Poke Parlor | 12.86 | oskibear | 8 | False |

*. . . (46 rows omitted)*

The table has 6 columns:

- `User`: (string) username of the user who purchased the order
- `Restaurant`: (string) restaurant name of the order
- `Total`: (float) total amount spent on the order, in dollars
- `Receiver`: (string) username of the user who received the order
- `Rating`: (int) how the user rated their order on a scale of 1-10 (10 being most satisfied)
- `With Friends`: (boolean) whether or not the user placed the order with friends

## heights

Here is a preview of the table `heights`:

- `height` is the **float** height of an individual, measured in inches
- `enrollee` is a **string** containing either "high school" or "college"

| height | enrollee |
|---|---|
| 68.5 | college |
| 62.1 | high school |
| 66.5 | college |

## rentals

Here is a preview of the table `rentals`:

- `date` contains the **string** date on which the rental data were collected
- `weekend` is a **boolean** corresponding to whether the day fell on a weekend
- `num_kayaks` is the **integer** number of kayak rentals on that particular day
- `num_boards` is the **integer** number of windsurfing board rentals on that particular day
- `wind` is the **integer** wind speed on that particular day, in miles per hour

| date | weekend | num_kayaks | num_boards | wind |
|---|---|---|---|---|
| 04/29 | True | 35 | 19 | 4 |
| 05/01 | True | 26 | 24 | 22 |
| 05/02 | False | 17 | 14 | 10 |

## skiers

Here is a preview of the table `skiers`:

| Name | Sport | Height (in) | Downhill Time (s) | Favorite Resort |
|---|---|---|---|---|
| James | Ski | 71 | 90.52 | Vail |
| Eunice | Ski | 66 | 93.64 | Beaver Creek |
| Oscar | Snowboard | 69 | 89.77 | Heavenly |
| Rebecca | Snowboard | 68 | 91.01 | Palisades |
| Ciara | Ski | 70 | 101.34 | Park City |

*. . . (40 rows omitted)*

## restaurant

Here is a preview of the table `restaurant`:

| Restaurant | Type | Distance from Campus |
|---|---|---|
| Round Table | Pizza | 2.2 |
| Panera | Bagels | 2.3 |
| Feng Cha | Boba | 0.13 |
| Boba Guys | Boba | 1.5 |
| Berkeley Thai House | Thai | 0.15 |

*. . . (306 rows omitted)*

## transport

Here is a preview of the table `transport`:

| Restaurant | Transportation | Time |
|---|---|---|
| Panera | Bus | 27 |
| La Burrita | Walk | 5 |
| Panera | Walk | 62 |
| Boba Guys | Drive | 10 |
| Panera | Drive | 12 |

*. . . (1492 rows omitted)*