Fall 2024 Practice Midterm Exam

Foundations of Data Science

Name

Instructions

- Open the Midterm Exam Reference Guide (linked in Canvas). We will provide you with a new copy of this reference to use on the exam.

- Previews of most of the tables you will work with in this exam are at the end of exam in the Table Reference section.

- You can assume the following code has been run, when you are writing your Python code:

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

- 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 template blanks or boxed 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.

- 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 before the Midterm Exam Study Session.

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

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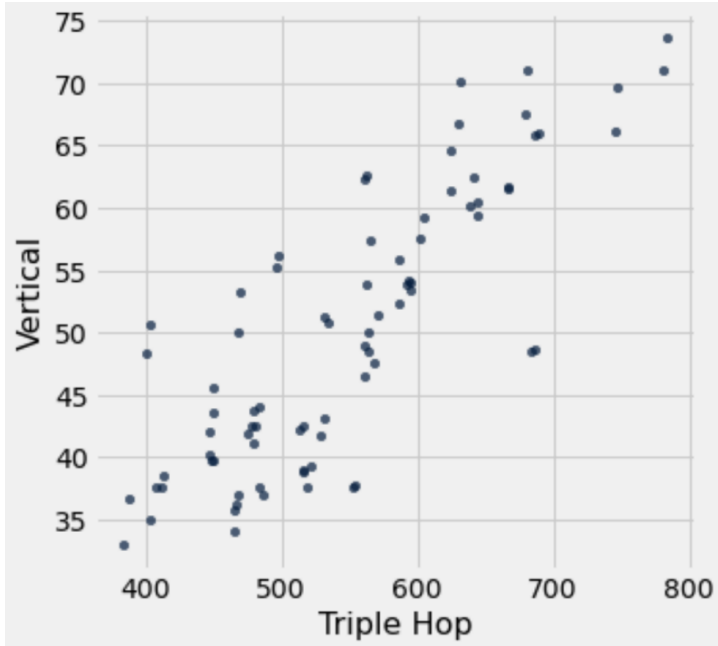
Pick all the conclusions that can be drawn from the scatter plot. Select all that apply.

- ✓ **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 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`. He created the following histogram based on the data in `heights` where the x-axis of the histogram has a range of $[55, 75]$ and each bin is 2 inches in width.
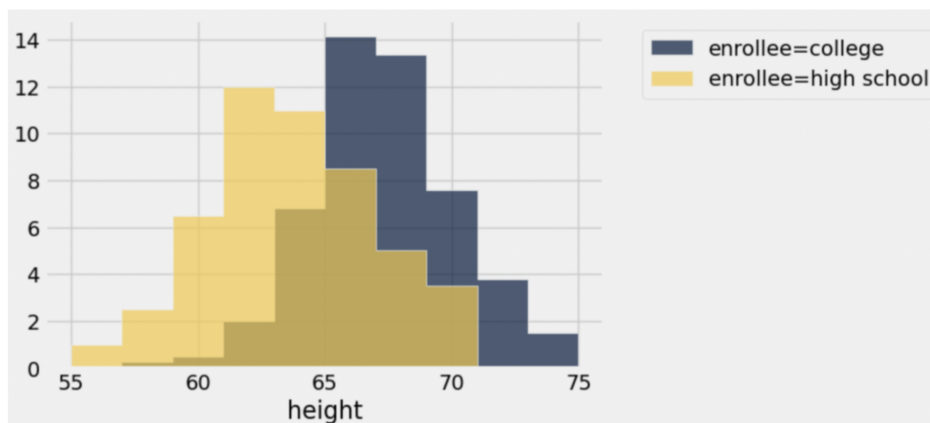
(a) Unfortunately, Dylan forgot to include the y-axis label in his histogram. Which of the following is an appropriate y-axis label? Select one.

○ 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) 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.

(a) Which single visualization is most appropriate to describe the distribution of wind speeds in `rentals`. Select one.

√ **Histogram**

○ Overlaid Histograms

○ Scatter Plot

○ Overlaid Scatter Plots

○ Bar chart

○ Line Plot

(b) 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

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.

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.
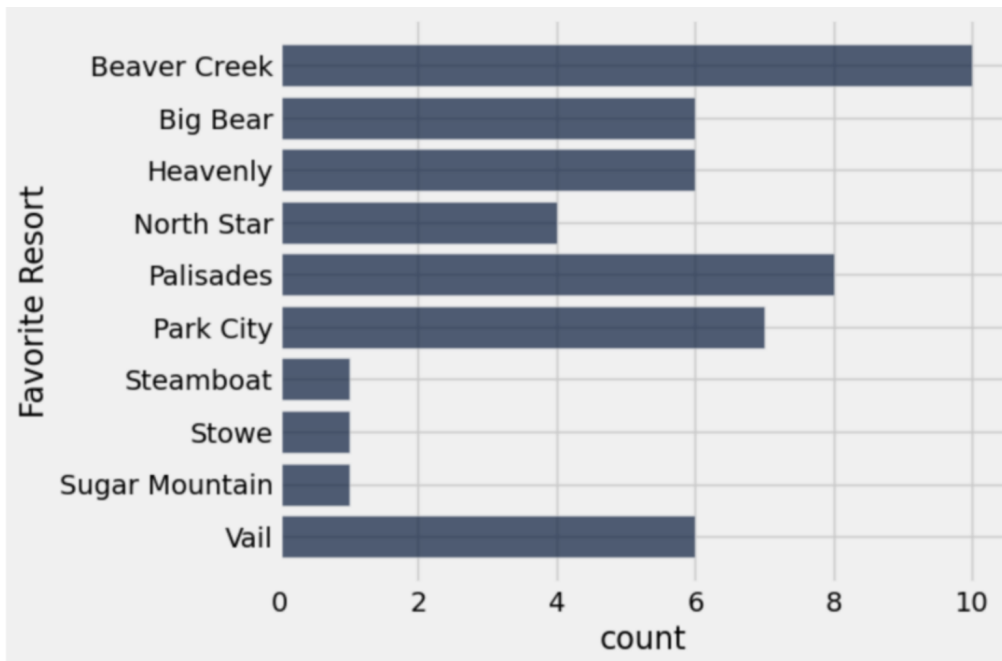

```
favorite_counts = skiers._____(_____)
_____._____(_____)
```

> **Sample Solution:**
>
> ```
> favorite_counts = skiers.group("Favorite Resort")
> favorite_counts.barh("Favorite Resort")
> ```

(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.

Sample Solution:
```
def str_to_int(height):
    return int(height)
```

(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.

Sample Solution:
```
heights = skiers.apply(str_to_int, 'Height (in)')
skiers = skiers.with_column('Height (in)', heights)
```

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

Sample Solution:
```
np.average(skiers.column('Height (in)'))
```

6. If `customers` is a Table with a column called `'Age'`, why does the following code produce an error?

```
customers.where('Age', are.above(24)) / customers.num_rows
```

Sample Solution: This produces an error because `customers.where('Age', are.above(24))` prodcues another Table and `customers.num_rows` produces an integer, yet the division of a Table and an integer is not defined.

7. The empirical distribution of a categorical variable contains the unique values of the variable along with how often each value is observed.

    √ **True**

    ○ False

8. 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._____(_____, _____)
```

> **Sample Solution:**
>
> ```
> orders.scatter("Total", "Rating")
> ```

(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._____(_____, True).num_rows
without_friends = _____._____ - _____
usually_friends = with_friends _____ without_friends
```

> **Sample Solution:**
>
> ```
> with_friends = orders.where("With Friends", True).num_rows
> without_friends = orders.num_rows - with_friends
> usually_friends = with_friends > 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(_____, _____)
    .sort(_____, descending=_____)
    .column("User").item(0)
)
```

9. 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))
   ```

10. 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

11. 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. Select one.

(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.

○ `(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))`

12. 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%

13. Complete the code below to simulate rolling a fair six-sided die 10,000 times and estimate the probability of rolling a 4 or higher.

```
count = 0
num_reps = 10_000
die = np.arange(1, 7)

for i in np.arange(_____):
    random_die_roll = np.random.choice(_____)
    if _____:
        count = count + 1


prob_4_or_higher = _____
```

Sample Solution:

```
count = 0
num_reps = 100
die = np.arange(1, 7)

for i in np.range(num_reps):
    random_die_roll = np.random.choice(die)
    if random_die_roll >= 4:
        count = count + 1

prob_4_or_higher = count / num_reps
```

14. In the game of Wordle, a player guesses up to 6 words until they correctly guess the secret word of the day or run out of guesses. Their guess count is either the guess number that was correct, 1 through 6, or X if all 6 guesses were incorrect.

    The table called `wordle` contains three columns. The first column shows the possible number of guesses. The second column shows the proportion of guess counts for all the players who tried to guess yesterday's word. The third column shows the proportion of all 1,000 Berkeley students with each guess count for yesterday's word.

    The following questions are part of the steps we would use to investigate whether the distribution of guess counts for UC Berkeley students differs from the distribution for all players on yesterday's Wordle.

    (a) Complete the null hypothesis: "The distribution of guess counts for UC Berkeley students is ..."
        ○ uniform with a 1/7 chance for each possible guess count.
        ○ like a random sample from a uniform distribution with a 1/7 chance for each possible guess count.
        ○ different from a uniform distribution with a 1/7 chance for each possible guess count.
        ○ the population of guess counts for all Wordle players.
        √ **like a random sample from the population of guess counts for all Wordle players.**
        ○ different from the population of guess counts for all Wordle players.

    (b) Complete the alternative hypothesis: "The distribution of guess counts for UC Berkeley students is ..."
        ○ the same as the distribution of guess counts for all Wordle players.
        √ **different from the distribution of guess counts for all Wordle players.**
        ○ the same as the uniform distribution.
        ○ different from the uniform distribution.

(c) Which test statistic is best for choosing between the null and alternative hypotheses?
- ○ total guess count
- ○ most common guess count
- ○ guess count
- √ **total variation distance**
- ○ observed average

(d) Which line of code simulates a distribution of proportions for 1000 Berkeley students under the null hypothesis?
- ○ `sample_proportions(1000, wordle.column(2))`
- √ `sample_proportions(1000, wordle.column(1))`
- ○ `sample_proportions(1000, make_array('1', '2', '3', '4', '5', '6', 'X'))`
- ○ `sample_proportions(1000, make_array(1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7))`

(e) Assume the observed test statistic is assigned to `obs`. We simulate under the null hypothesis 10,000 times and append each simulated test statistics to an array `sim`. Complete the following expression that computes the p-value for this hypothesis test.

_____ (_____ >= _____) / len(_____)

> **Sample Solution:**
> `np.count_nonzero(sim >= obs) / len(sim)`

(f) If the null hypothesis is rejected because the p-value of this hypothesis test is very small, what can we conclude? Select all that apply.
- □ Attending Berkeley improves most people's Wordle performance.
- □ Attending Berkeley changes most people's Wordle performance.
- □ Attending Berkeley does not improve most people's Wordle performance.
- □ Attending Berkeley does not change most people's Wordle performance.
- √ **None of these.**

15. Suppose a hypothesis test is proposed and we already know that the null hypothesis is true. If 500 researchers each independently collects a sample of the same size to carry out an experiment and they all use 1% as their p-value cutoff, we should expect around 5 of them to reject the null.
- √ **True**
- ○ False

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)

wordle

Here is a preview of the table `wordle`:

Count	Everyone	Berkeley Students
1	0.0	0.0
2	0.17	0.09
3	0.33	0.25
4	0.27	0.32
5	0.20	0.28
6	0.02	0.03
X	0.01	0.03