Fall 2023 Final Exam

Foundations of Data Science

Name

Total Score: _____ of 100 Points

Instructions

- Make sure to rip off the Table Reference and Final Exam Reference Guide 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 asks you to 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 one of the provided blank sheets 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')
```

- The multiple choice questions (◯) and multiple answer questions (☐) will be scored like they are in Canvas Quizzes.
- The open response questions will be graded as:
  - Full Points: The response is correct and may contain a very very small error.
  - Partial Points: A reasonable response was provided. The partial point value will depend on your response.
  - No Points: No reasonable attempt was provided.
- Once you are finished, turn in your exam and you are welcome to leave.

Section A

1. (2 points) A real estate company has a dataset of all their buildings, with three attributes for each building: its size (in square feet), its type (residential or commercial), and its estimated value (sale price) if sold (in dollars). The standard visualization to understand the distribution of building types is: **Choose one.**

   √ **A bar chart** ◯ A line plot ◯ A scatter plot ◯ A histogram

2. Select True or False for each of the following:

   (a) (2 points) The height of each bar in a histogram represents the proportion of data within the corresponding bin. **Choose one.**

   ◯ True √ **False**

   (b) (2 points) For any distribution, the percentage of data that lies beyond two standard deviations on either side of the mean is at most 25%. **Choose one.**

   √ **True** ◯ False

   (c) (2 points) A classifier is considered to be overfitting if it performs very well on the test set. **Choose one.**

   ◯ True √ **False**

   (d) (2 points) If we use linear regression to predict y-values based on our x-values, the median of our residuals will always be zero. **Choose one.**

   ◯ True √ **False**

   (e) (2 points) For any two events A and B, the probability $P(A \text{ and } B)$ is less than or equal to the probability $P(A \text{ or } B)$ **Choose one.**

   √ **True** ◯ False

3. Cognitive Behavioral Therapy (CBT) is a psycho-social intervention that aims to reduce symptoms of mental health conditions such as depression. As a researcher, you are tasked with designing and implementing a large study of the effect of CBT on reducing such depression symptoms.

   As part of your study, you randomly sample 1,600 individuals seeking treatment for various levels of depression. Currently, you have a table called `patients` containing the following string data:: the first name (`'First Name'`), last name (`'Last Name'`), and email (`'Email'`) for each of the sampled individuals. For a visual reference, use the first 3 columns of the `patients` table in the Table Reference section as an example.

(a) (3 points) First, you need to randomly assign 800 individuals to receive a course of well-studied antidepressant medication and the other 800 individuals to receive a sequence of CBT sessions. Write code that will update the patients table by adding a fourth column called `'CBT'` to the table that will contain `bool` values where 800 of the individuals will randomly be assigned a value of `True` and the rest a value of `False`. Your updated table should resemble the first 4 columns of the `patients` table in the Table Reference section.

Hints: The `np.random.choice` function has a parameter called `replace` that allows you to use sampling with or without replacement. The default value is `True`. Also, the code:

```
np.array([True] * 3 + [False] * 2)
```

will create the following array:

```
array([ True,  True,  True, False, False], dtype=bool).
```

Sample Solution:

```
CBT_values = np.array([True] * 800 + [False] * 800)
patients = patients.with_column(
    'CBT',
    np.random.choice(CBT_values, 1_600, replace=False)
)
```

(b) (2 points) With the treatment assignments, the patients are messaged with their treatment. Those receiving CBT know they are receiving several weeks of therapy sessions, and the others know that they are receiving a course of antidepressant medication that is known to reduce the symptoms of depression at all levels. All 1,600 individuals consent to and complete their assigned treatment. This study design is a randomized controlled experiment. **Choose one.**

√ **True** ◯ False

(c) PHQ-9 is a 9-question patient health questionnaire that, according to the American Psychological Association, offers psychologists concise information about a patient's level of depression. PHQ-9 scores range from 0 to 27, with higher scores indicating a more severe level of depression. PHQ-9 scores are collected for every individual in the study both before and after treatment. The `patients` table is updated with columns labeled 'Pre-PHQ-9'— for the scores before treatment, and `'Post-PHQ-9'` for the scores after treatment. See a preview of the table (just the first 6 columns after this step) in the Table Reference section.

i. (2 points) Which of the following is a correct null hypothesis that could be used to test if CBT reduces depression symptoms more effectively than the antidepressant treatment? **Choose one.**

- √ **There is no difference between the average difference in PHQ-9 scores (Pre-PHQ-9 scores minus Post-PHQ-9 score) for those receiving CBT and the average difference in PHQ-9 scores for those receiving the antidepressants.**
- ◯ There is a positive difference between the average difference in PHQ-9 scores (Pre-PHQ-9 scores minus Post-PHQ-9 score) for those receiving CBT and the average difference in PHQ-9 scores for those receiving the antidepressants.
- ◯ There is a negative difference between the average difference in PHQ-9 scores (Pre-PHQ-9 scores minus Post-PHQ-9 score) for those receiving CBT and the average difference in PHQ-9 scores for those receiving the antidepressants.

ii. (2 points) Please, state a clear and complete alternative hypothesis that could be used to test if CBT reduces depression symptoms more effectively than the antidepressant treatment.

Sample Solution: There is a positive difference between the average difference in PHQ-9 scores (Pre-PHQ-9 scores minus Post-PHQ-9 score) for those receiving CBT and the average difference in PHQ-9 scores for those receiving the antidepressants.

iii. (3 points) Create a python function called `PHQ_diff` that has 2 arguments, `pre` and `post` where:

- `pre` is a Pre-PHQ-9 integer score.
- `post` is a Post-PHQ-9 integer score.

The function should return the Pre-PHQ-9 score minus the Post-PHQ-9 score.

Sample Solution:

```
def PHQ_diff(pre, post):
    return pre - post
```

iv. (3 points) Using the `PHQ_diff` function, update the `patients` table by adding a column called
`'Diff-PHQ-9'` to the `patients` table showing the difference in PHQ-9 scores for all the patients. The table will now resemble all 7 columns in the `patients table` in the Table Reference section.

Sample Solution:

```
patients = patients.with_column(
    'Diff-PHQ-9',
    patients.apply(PHQ_diff, 'Pre-PHQ-9', 'Post-PHQ-9')
)
```

v. (3 points) Using the updated `patients` table, provide code that will generate a histogram showing the distribution of `'Diff-PHQ-9'` values for those receiving CBT overlaid with a histogram showing the distribution of of `'Diff-PHQ-9'` values for those not receiving CBT. Refer to the entire `patients` table in the Table Reference section as an example.

Sample Solution:

```
patients.hist('Diff-PHQ-9', group='CBT')
```

vi. (2 points) If there is an association between the treatment and the difference in PHQ-9 scores, then the histograms will almost perfectly overlay each other. **Choose one.**

◯ True √ **False**

vii. (3 points) Using the `patients` data, complete the following code which will calculate the observed statistic. That is, the difference between the average difference in PHQ-9 score for those patients who received CBT, and the average difference in PHQ-9 score for those who received the antidepressant.

```
averages_tbl = (patients.select('CBT', 'Diff-PHQ-9')
                        .____________(A)___________)
averages = averages_tbl.column('Diff-PHQ-9 average)
observed_diff_aves = ____________(B)___________
```

Sample Solution:

```
averages_tbl = (patients.select('CBT', 'Diff-PHQ-9')
                        .group('CBT', np.average))
averages = averages_tbl.column(1)
observed_diff_aves = averages.item(1) - averages.item(0)
```

viii. (3 points) Ten thousand permutations (reshuffling the `'CBT'` column) of the `patients` table were generated and the difference in average differences was calculated each time and stored in an array called `simulated_diffs_aves`. If the observed difference is named `observed_diff_aves`, write code that will calculate the p-value for this hypothesis test using the simulated data.

Sample Solution:

```
np.count_nonzero(simulated_diffs_aves >= observed_diff_aves) / 10_000
```

ix. (3 points) If the p-value is 4%, which of the following is a valid conclusion for this hypothesis test? **Choose all that apply.**

- ☐ 4% of the test statistics simulated under the null hypothesis were as, or less extreme than the observed test statistic.
- ☐ CBT reduces the symptoms of depression equally as well as the prescribed antidepressant.
- ☐ With a p-value cutoff of 5%, our data are consistent with the null hypothesis.
- √ **There is a statistically significant reduction in depression symptoms for those that follow a sequence of CBT sessions**
- ☐ None of the above.

Section B

4. Meme generation is a high-profit business for the Instagram account @menime. Currently, they have 17 million followers and they post a variety of image-based or video-based memes. In an effort to create a prediction tool that will help determine if a meme will be popular before they release it, they share 200 memes with a randomly generated focus group of 1,000 Instagram users, who rate ('like') the memes. The data is stored in a table called `meme_data`. More details about this table are in the Table Reference section.

   You are interested in building a classification model that will use the numerical features in the `meme_data` table to predict whether or not a meme will be `'popular'`. A meme will be labeled `'popular'` if more than 70% of the focus group liked it.

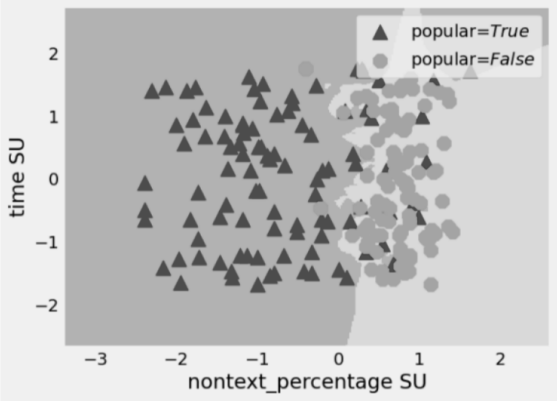
   (a) (3 points) Complete the code below so that `meme_popular` is a copy of `meme_data` with an additional column named `'popular'`, which contain Boolean values that indicate whether a meme is popular (`True`) or not (`False`).

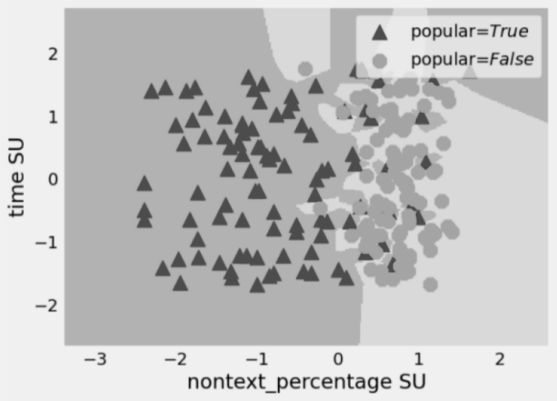
   ```
   pop_arr = ____________________(A)____________________
   meme_popular = ________________________(B)________________________
   ```

   Sample Solution:

   ```
   pop_arr = meme_data.column('rating') > 70
   meme_popular = meme_data.with_column('popular', pop_arr)
   ```

(b) (2 points) After converting the `time` and `nontext` percentage columns to standard units and creating two k-NN classifiers, each with a different value of `k`: `k=3` and `k=1`. Which of the following plots corresponds to the 1-NN classifier?





Choose one.

◯ The left plot

√ **The right plot**

(c) (2 points) After training a 1-NN classifier, you notice that only 10% of the memes in the training data are popular, compared to 50% of memes in the testing data. Additionally, you find that this imbalance is due to an error in your code.

After correcting the error and re-distributing the data to restore the balance of popular memes, you re-train a 1-NN classifier. How would you expect the accuracy to change after re-balancing the data? **Choose one.**

√ **The accuracy increases.**

☐ The accuracy remains the same.

☐ The accuracy decrease.

☐ It is not possible to determine what will happen to the accuracy.

(d) (3 points) It turns out that someone has already developed a similar classification tool called memeNN that has an accuracy of 75%. You notice that when memeNN makes a correct prediction, your 1-NN classifier's correctly predicts the popularity 82% of the time. When memeNN makes an incorrect prediction, your 1-NN classifier correctly predicts the popularity 45% of the time.

If a meme is randomly sampled from the test set and the 1-NN classifier predicts its class correctly, what is the probability that memeNN will also predict the class correctly?

Write your answer as a mathematical/python expression.

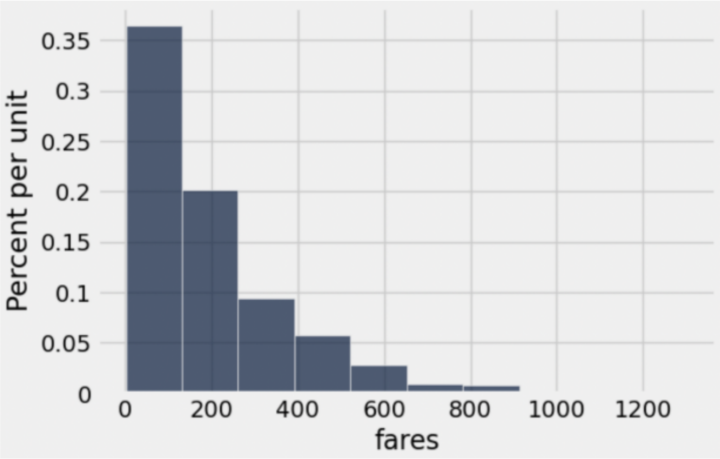
Hint: Making a decision tree may be helpful.

Sample Solution: $\frac{0.75 \times 0.82}{0.75 \times 0.82 + 0.25 \times 0.45}$

Section C

5. Your data analyst team is interested in studying Bay Area public transportation, so you begin analyzing data from the widely-used BART train system and the AC Transit bus services for the year 2022. Unfortunately, due to budget cuts, your available compute power is unable to process all of the data from 2022, so instead, your team is going to work with a large random sample of 1,000 riders. That sample is loaded into a table called `transport`. The first few rows of that table are previewed in the Table Reference section.

   (a) (3 points) Given below is the distribution of the fares column from the `transport` table. Which of the following conclusions can you draw from the plot? **Select all that apply.**



   √ **The distribution of the `'fares'` column in `transport` is right-skewed.**

   ☐ The distribution of the `'fares'` column in `transport` is left-skewed.

   √ **The median of the `'fares'` column in `transport` is less than the mean.**

   ☐ The median of the `'fares'` column in `transport` is greater than the mean.

   (b) (3 points) Which of the following statements must be true? **Select all that apply.**

   ☐ The distribution of fare spending among all riders is approximately normal.

   √ **The distribution of sample means of fare spending is approximately normal for large random samples of data.**

   √ **The distribution of sample sums of fare spending is approximately normal for large random samples of data.**

   ☐ The distribution of sample medians of fare spending is approximately normal for large random samples of data.

   ☐ None of the above.

(c) Your team is interested in estimating the proportion of all riders who had transferred between a BART train and an AC Transit bus at least once. You decide to use your sample of 1,000 riders to estimate this unknown population parameter.

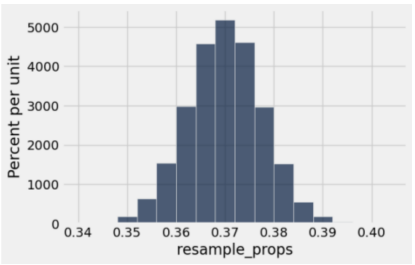
i. (4 points) Provide code that will generate a table of 10,000 bootstrapped proportions of riders who transferred between a BART train and an AC Transit bus at least once.

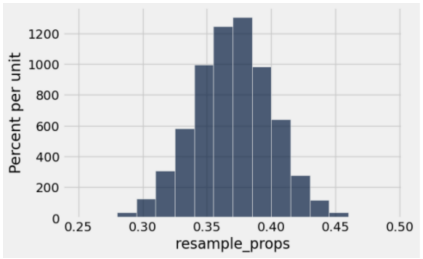
```
resample_props = make_array()

for i in np.arange(10_000):
    resamp = _________________________(A)___________________
    resamp_prop = ____________________(B)___________________
    ___________________________(C)___________________________

resamp_props_tbl =  Table().with_column("resample_props", resample_props)
```

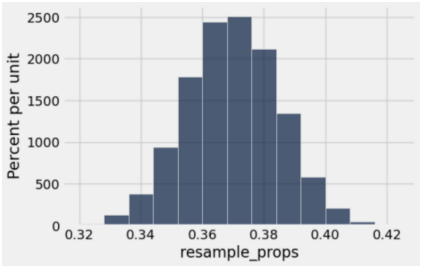
Sample Solution:

```
resample_props = make_array()

for i in np.arange(10_000):
    resamp = transport.sample()
    resamp_prop = np.mean(resamp.column("transfer"))
    resample_props = np.append(resample_props, resamp_prop)

resamp_props_tbl = Table().with_column("resample_props", resample_props)
```

ii. (2 points) You find that the mean and standard deviation of your bootstrapped proportions, resample props is 0.37 and 0.015, respectively. Which of the following most closely resembles the distribution of resample props? **Choose one.**







◯ The left graph ◯ The middle graph √ **The right graph**

iii. (3 points) Write a mathematical or Python expression that evaluates to the probability that the first row in transport is included at least once in a single bootstrap re-sample of size 1,000.

Sample Solution: `1 - (999/1_000) ** 1_000`

iv. (3 points) Provide code that creates the array `interval` which contains the left and right endpoints of a 95% confidence interval estimate for the proportion of riders in the population who transferred at least once.

Note: You may used variable names defined from previous sub-parts in your code.

```
left = ____________________(A)____________________
right = ____________________(B)____________________
interval = make_array(left, right)
```

Sample Solution:

```
left = percentile(2.5, resample_props)
right = percentile(97.5, resample_props)
interval = make_array(left, right)
```

v. (3 points) Which of the following conclusions can you draw using the 95% confidence interval generated in the previous part (iv)? **Select all that apply.**

- ☐ If someone takes the BART train, there is a 95% chance that they transfer to an AC Transit bus.
- √ **If you make confidence intervals from many large random samples from the population, you can expect that roughly 95% of the intervals you create will contain the true population proportion.**
- ☐ There is a 95% chance that the population's true transfer proportion is within the interval generated in part (iv).
- ☐ There is a 95% chance that the sample's true transfer proportion is within the interval generated in part (iv).
- ☐ None of the above.

vi. (3 points) Your team has one last request. They want your 95% confidence interval to be no wider than 5%. What is the smallest sample size that satisfies this requirement? **Express your answer as a whole number or a mathematical/Python expression**.

Sample Solution: `(4 * 0.5 / 0.05) ** 2` or `1600`

Section D

6. Farhana likes to attend a popular class at the local fitness center. She collects data about each class she attends in a table called `workouts`. The table is previewed in the Table Reference section.

   (a) (2 points) Choose the best technique to answer the question "How high will Farhana's exercise heart rate be today given that 30 people attended class?" **Choose one.**

   ◯ Classification

   √ **Linear Regression**

   ◯ Hypothesis Test

   ◯ Randomized Control Experiment

   ◯ Bayes' Rule

   (b) (2 points) Choose the best technique to answer the question "Is there a difference in Farhana's heart rate between sunny and rainy days?" **Choose one.**

   ◯ Classification

   ◯ Linear Regression

   √ **Hypothesis Test**

   ◯ Randomized Control Experiment

   ◯ Bayes' Rule

   (c) (2 points) Choose the best technique to answer the question "What are the most likely weather conditions given that 12 people attended class today?" **Choose one.**

   √ **Classification**

   ◯ Linear Regression

   ◯ Hypothesis Test

   ◯ Randomized Control Experiment

   ◯ Bayes' Rule

   (d) (2 points) Choose the best test statistic for the following alternative hypothesis. **Choose one.**

   Alternative Hypothesis: "The class size is larger on sunny days than it is on rainy days."

   ◯ The total variation distance between the class size distribution of sunny days and class size distribution of rainy days

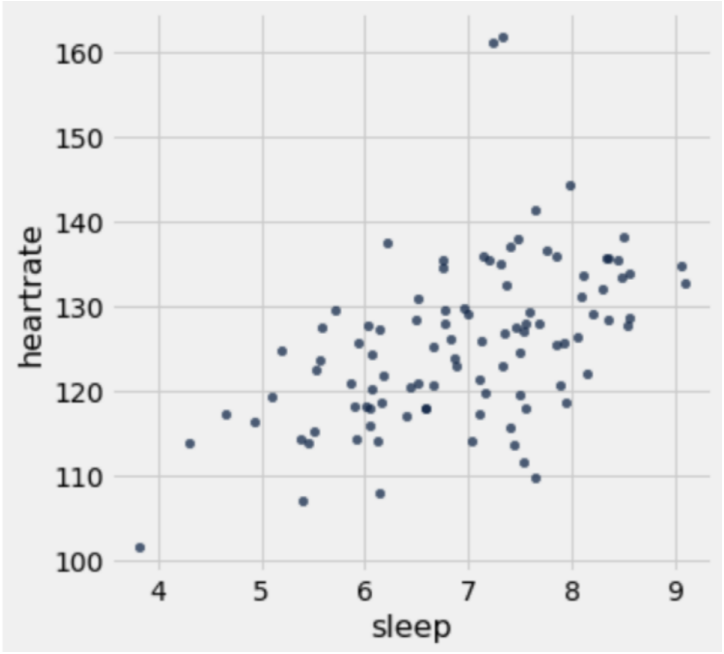
   ◯ The empirical mean of class size on sunny days

   ◯ The empirical mean of class size

   √ **The difference of mean class size between sunny and rainy days**

   ◯ The difference of mean class size between sunny and cloudy days

(e) (3 points) Farhana wants to see if there is a relationship between how much sleep she gets and her heart rate during class, so she creates the following scatter plot.



Write a line of code that would generate the scatter plot above.

Sample Solution:

```
workouts.select(‘sleep‘, ‘heartrate‘).scatter(‘sleep‘)
```

(f) (3 points) Which of the following are valid conclusions from this graph? **Choose all that apply.**

- √ **There is a positive association between her sleep and her heart rate during class**
- ☐ There is a negative association between her sleep and her heart rate during class
- ☐ Getting more sleep causes Farhana to have a higher heart rate during class
- ☐ Getting less sleep causes Farhana to have a higher heart rate during class
- ☐ Fewer people attend class on rainy days

(g) (3 points) Farhana asks her friend to compute the correlation coefficient `r` between these two variables , but her friend's code has at least one mistake in it. In the code below, circle and cross out each mistake and, if applicable, write the correct code immediately above. Alternatively, you can re-write the code in the solution box. You can use the following `standard_units` function from lecture:

```
def standard_units(any_numbers):
    '''Convert any array of numbers to standard units.'''
```

```
    return (any_numbers - np.mean(any_numbers))/np.std(any_numbers)
```

Here is her friend's code:

```
heartrate_in_su = standard_units('heartrate')

sleep_in_su = standard_units('sleep')

r = np.sum(heartrate_in_su * sleep_in_su)
```

Sample Solution:

```
heartrate_in_su = standard_units(workouts.column('heartrate'))

sleep_in_su = standard_units(workouts.column('sleep'))

r = np.mean(heartrate_in_su * sleep_in_su)
```

(h) (3 points) Suppose we know the following:

- Farhana's heart rate has an average of 125 bpm and a standard deviation of 6 bpm
- Farhana's sleep has an average of 7 hours and a standard deviation of 1 hour
- The correlation between Farhana's heart rate and sleep is 0.5

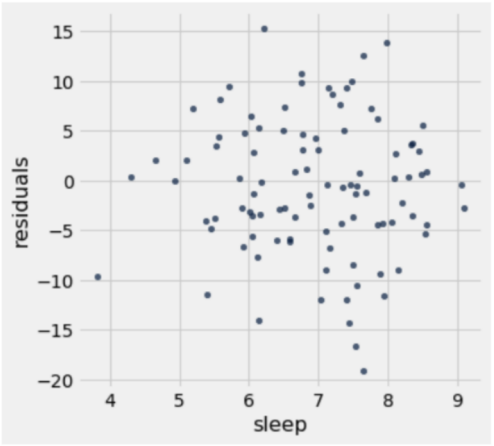
If we were to fit a regression line to the scatterplot in (e), what would the predicted heartrate be when Farhana gets 8 hours of sleep? You may leave your answer as a mathematical/Python expression.

Sample Solution:

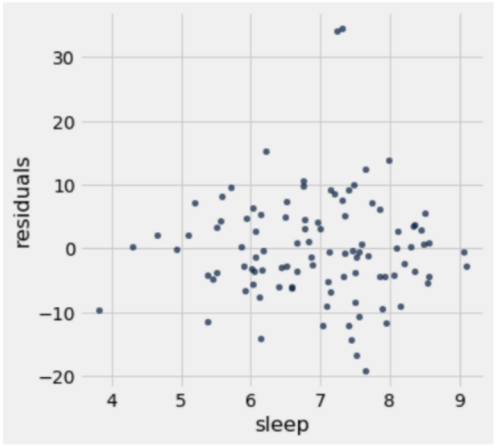
```
(0.5 * 6 / 1) * 8 + (125 - (0.5 * 6 / 1) * 7
```

128 bpm

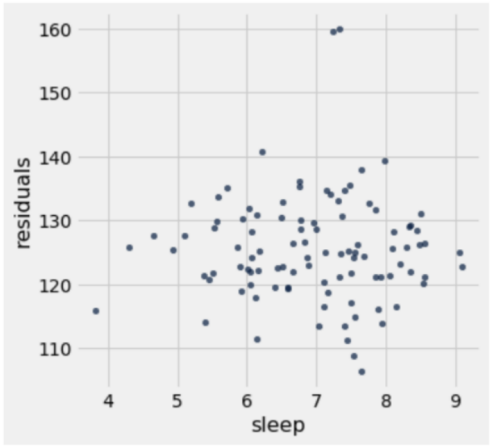
(i) (2 points) Which of the following is the residual plot for the scatter plot shown in part (e)?



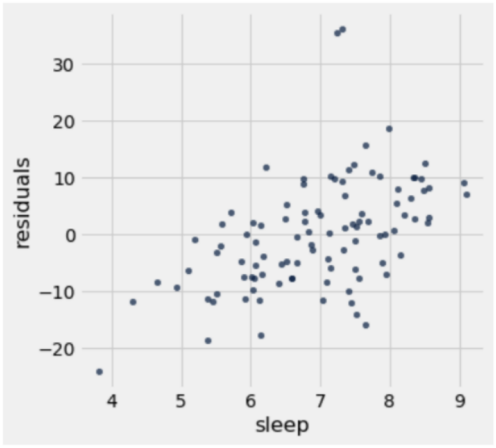
A



B



C



D

Choose one.

◯ Option A ✓ **Option B** ◯ Option C ◯ Option D

(j) (3 points) Farhana begins a research apprenticeship in the School of Public Health, and wants to understand whether the amount of sleep someone gets causes a change in average heart rate during exercise. Her lab starts a study with a random sample of Berkeley undergraduate and graduate students who exercise regularly. Which of the following experiments would be able to answer her causal question? **Choose all that apply.**

☐ Ask the undergraduates to sleep 7 hours per night and the graduate students to sleep 9 hours per night. Then, collect heart rate data during exercise.

☐ Randomly assign the subjects to two groups. Have the first group exercise for 1 hour per day, and have the second (control) group not exercise at all. Then, measure how much sleep they get before and after each exercise session.

✓ **Randomly assign the subjects to two groups. Have the first group sleep for 7 hours or less per night, and the second group sleep 9 hours or more per night. Then, collect heart rate data during exercise.**

☐ It is impossible to determine a causal link between these two variables.

This page was intentionally left blank.

Table Reference

patients

Here is a preview of the `patients` table:

First Name	Last Name	Email	CBT	Pre-PHQ-9	Post-PHQ-9	Diff-PHQ-9
Lily	Smith	lily.smith@email.com	False	23	20	3
Ethan	Garcia	ethan.garcia@email.com	True	18	16	2
Sophia	Lee	sophia.lee@email.com	False	20	19	1
Zoe	Nguyen	zoe.nguyen@email.com	True	15	16	-1
Lily	Smith	lily.smith@email.com	False	23	20	3

... (1595 rows omitted)

- When starting Question 3, the table initially resembles the first 3 columns of this table.
- After successfully completing Question 3 (a), the table resembles the first 4 columns of this table.
- In Question 3 (c), the table resembles the first 6 columns of this table.
- After successfully completing Question 3 (c) iv, the table resembles the first 7 columns of `patients`.

meme_data

Each row of the table `meme_data` represents a meme. The columns are:

- `category` (string): the category of the meme, which is either an "image" or a "video".
- `time` (integer): the duration of the meme, in seconds. Images will have a time value of 0.
- `nontext_percentage` (float): the percentage of the meme that is non-textual content (scale: [0.0 - 100.0]).
- `rating` (float): the percentage of the focus group who liked the meme (scale: [0.0 - 100.0]).

A preview of this table is not provided and should not be needed.

transport

Here is a preview of the `transport` table and some information about the data values:

- `id` (**integer**): identification (id) of the rider.
- `transfer` (**boolean**): whether that particular rider transferred between a BART train and an AC Transit bus at least once during 2022.
- `fares` (**float**): total amount that particular rider spent on fares in 2022, measured in dollars.

id	transfer	fares
32849	True	12.5
29490	False	62
81305	False	131.75
70654	False	43

... (996 rows omitted)

workouts

Here is a preview of the `workouts` table and some information about the data values:

size	heartrate	weather	sleep
33	145.1	sunny	7.3
28	100.7	sunny	6.5
23	124	sunny	5
10	137.8	rainy	9

The table contains four columns:

- **size**: an int, the number of people who attended the class
- **heartrate**: a float, her average heart rate during the class in beats per minute (bpm)
- **weather**: a string, the weather conditions for that day
- **sleep**: a float, the number of hours of sleep she got the night before