Data 8 Foundations of Data Science
Fall 2019

FINAL EXAM SOLUTIONS

INSTRUCTIONS

- The exam is worth 150 points. You have 170 minutes to complete it.
- The exam is closed book, closed notes, closed computer/phone/tablet, closed calculator, except the official final exam reference guide provided with the exam.
- Write/mark your answers on the exam in the blanks/bubbles provided. Answers written anywhere else will not be graded. Unless the question specifically asks you to explain your answer, you do not need to do so, and if you write an explanation it will not be graded.
- If you need scratch paper, you are welcome to use the reference guide and the back of this cover page. Scratch work will not be graded.
- For all Python code, you may assume that the statements `from datascience import *` and `import numpy as np` have been executed. Do not use features of the Python language that have not been described in this course.
- In any part, you are free to use any tables, arrays, or functions that have been defined in previous parts of the same question, and you may assume they have been defined correctly.

Last name	
First name	
Student ID number	
Calcentral email (`_@berkeley.edu`)	
Lab GSI	
Your seat number (e.g. A1) & room	
← Name of the person to your left	
Name of the person to your right →	
All the work on this exam is my own. **(please sign)**	

1. (10 points) Python expressions

For each expression below, say what it evaluates to.

(a) (2 pt) `np.count_nonzero(np.arange(3))`

○ 0 ○ 1 ● 2 ○ Python returns an Error

(b) (2 pt) `make_array(1,2,3) * 2`

○ array([1,2,3,1,2,3]) ● array([2,4,6]) ○ array([1,4,9])

○ Python returns an Error

(c) (2 pt) `np.arange(2,8,4) - np.arange(1,7,5)`

○ array([1,1,-1]) ○ array([1,1]) ● array([1,0])

○ Python returns an Error

(d) (2 pt) `percentile(30, make_array(5,10,30,50))`

○ 5 ● 10 ○ 30 ○ Python returns an Error

(e) (2 pt) `np.std(np.arange(0,200,100))`

○ 5 ● 50 ○ 500 ○ Python returns an Error

2. (5 points) Dorms

Ani is applying for student housing at Berkeley. Since her favorite building on campus is Evans Hall, she hopes to get in to Foothill. Assume that her chance of getting into Foothill each year is 1/6 independent of all other information in her application, and independent of where she lives any other year.

(a) (3 pt) If Ani plans to stay in the dorms for two years, what is the probability that she gets into Foothill in *at least one* of the two years she applies?

○ 1/36 ● 11/36 ○ 2/6 ○ 25/36 ○ 35/36

○ Cannot be calculated with the information given

(b) (2 pt) If Ani plans to stay in the dorms for two years, what is the probability that she gets into Foothill in *both* years?

● 1/36 ○ 11/36 ○ 2/6 ○ 25/36 ○ 35/36

○ Cannot be calculated with the information given

3. (9 points) Scranton Strangler

The Scranton Strangler, a masked criminal, is on the loose in the small town of Scranton, which has a population of 1,000. Volunteer Sheriff Dwight has been assigned the role of catching the Strangler. From witness reports, the police are certain that the Strangler is left handed. Meanwhile, anonymized records from Scranton's primary care providers show that 1 in 20 Scranton residents is left handed.

(a) (2 pt) Suppose Dwight calls a random phone number in Scranton and a person answers the phone. What is the probability that the person is the Scranton Strangler?

○ $\frac{1}{20}$ ● $\frac{1}{1000}$ ○ $(\frac{1}{1000})(\frac{1}{20})$

● $\frac{(\frac{1}{1000})(\frac{1}{20})}{(\frac{1}{1000})(\frac{1}{20})+(\frac{999}{1000})(\frac{1}{20})}$ ○ $\frac{(\frac{1}{1000})(1)}{(\frac{1}{1000})(1)+(\frac{999}{1000})(\frac{1}{20})}$ ○ $\frac{(\frac{1}{1000})(\frac{1}{20})}{(\frac{1}{1000})(\frac{1}{20})+(\frac{999}{1000})(\frac{19}{20})}$

Both of the above filled in options are correct.

(b) (3 pt) Suppose Dwight bumps into a random Scranton resident at the gas station, and notices that they are left handed. What is the probability that they are the Scranton Strangler?

○ $\frac{1}{20}$ ○ $\frac{1}{1000}$ ○ $(\frac{1}{1000})(\frac{1}{20})$

○ $\frac{(\frac{1}{1000})(\frac{1}{20})}{(\frac{1}{1000})(\frac{1}{20})+(\frac{999}{1000})(\frac{1}{20})}$ ● $\frac{(\frac{1}{1000})(1)}{(\frac{1}{1000})(1)+(\frac{999}{1000})(\frac{1}{20})}$ ○ $\frac{(\frac{1}{1000})(\frac{1}{20})}{(\frac{1}{1000})(\frac{1}{20})+(\frac{999}{1000})(\frac{19}{20})}$

Some students interpreted "1 in 20 is left handed" as saying exactly 50/1000 residents that are left handed (instead of each of the 1000 residents independently having a 1/20 chance of being left-handed); given this assumption, then the correct answer is $\frac{1/1000}{1/20}=\frac{1}{50}$. We accepted this alternate answer as long as you provided the correct assumption and/or explanation.

(c) (4 pt) Suppose Dwight has a coworker named Creed, whom he doesn't trust. Before observing Creed's hands, Dwight think there is a 10% chance Creed is the Strangler. Now, suppose Dwight discovers that Creed is actually left handed. Given this new information, what is the probability that Creed is the Strangler?

○ $\frac{1}{20}$ ○ $\frac{1}{1000}$ ○ $(\frac{1}{1000})(\frac{1}{20})$

○ $\frac{(\frac{1}{10})(\frac{1}{1000})(\frac{1}{20})}{(\frac{1}{10})(\frac{1}{1000})(\frac{1}{20})+(\frac{999}{1000})(\frac{1}{20})}$ ● $\frac{(\frac{1}{10})(1)}{(\frac{1}{10})(1)+(\frac{9}{10})(\frac{1}{20})}$ ○ $\frac{(\frac{1}{10})(\frac{1}{1000})(\frac{1}{20})}{(\frac{1}{10})(\frac{1}{1000})(\frac{1}{20})+(\frac{999}{1000})(\frac{19}{20})}$

4. (26 points) Exercise

Farhana likes to attend a popular class at the RSF. She collects data about each class she attends in a table called `workouts`. Here are the first few rows.

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

For parts (a)-(c), choose the best technique to answer the given question. **For each question, choose only one answer.**

(a) (2 pt) *How high will Farhana's exercise heart rate be today given that 30 people attended class?*

- ○ Classification
- ● Linear Regression
- ○ Hypothesis Test
- ○ Randomized Control Experiment
- ○ Bayes' Rule

(b) (2 pt) *Is there a difference in Farhana's heart rate between sunny and rainy days?*

- ○ Classification
- ○ Linear Regression
- ● Hypothesis Test
- ○ Randomized Control Experiment
- ○ Bayes' Rule

(c) (2 pt) *What are the most likely weather conditions given that 12 people attended class today?*

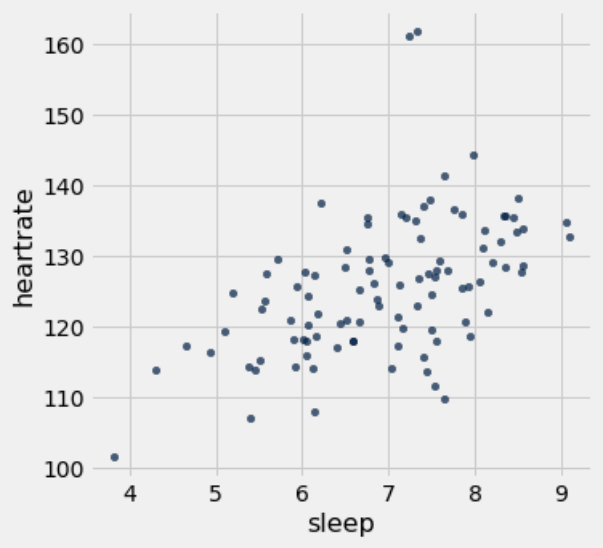
- ● Classification
- ○ Linear Regression
- ○ Hypothesis Test
- ○ Randomized Control Experiment
- ○ Bayes' Rule

(d) (3 pt) Choose the best test statistic for the following alternative hypothesis. **Choose only one answer.**

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) **(2 pt)** 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.

workouts.select('sleep', 'heartrate').scatter('sleep')

(f) **(3 pt)** 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) **(4 pt)** Farhana asks her friend to compute the correlation 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. You should assume the `standard_units` function has been defined correctly as in lecture: it takes in an array of data, and returns the data in standard units.

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

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

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

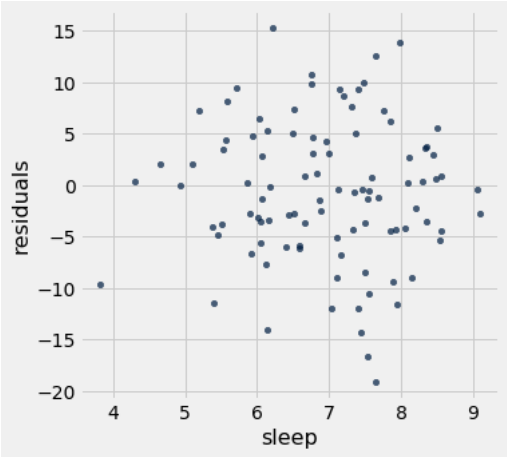
(h) **(2 pt)** Suppose we know the following:

- Farhana's `heartrate` has an average of 125 bpm and a variance of 25 bpm
- Farhana's `sleep` has an average of 7 hours and a variance of 1 hour
- The correlation between Farhana's `heartrate` and `sleep` is 0.1

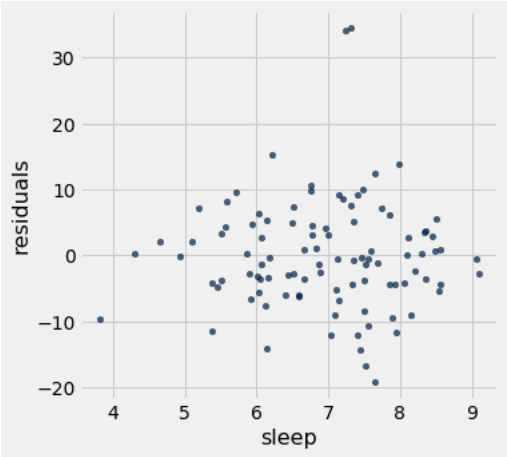
If we were to fit a regression line to the scatterplot in (e), what would the predicted heartrate be when Farhana gets 5 hours of sleep? You may leave your answer as a mathematical expression.

124 bpm

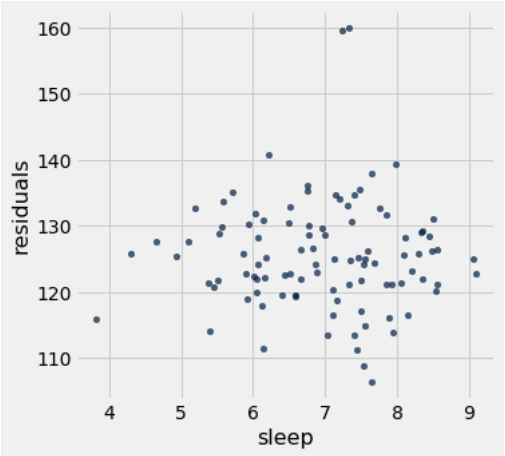
(i) (3 pt) Which of the following is the residual plot for the scatter plot shown in part (e)?



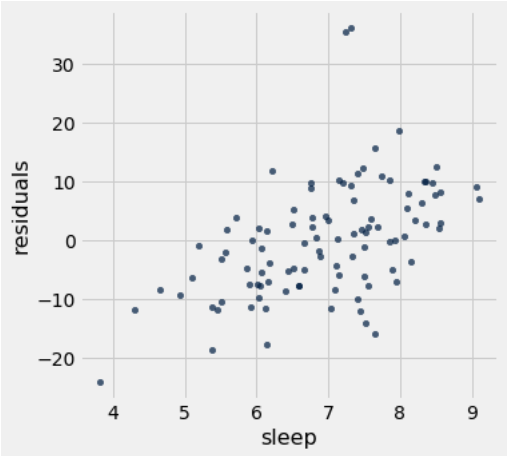
○ A



● B



○ C



○ D

(j) (3 pt) 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 each subject to measure data on their heart rate during exercise and their sleep.
- ○ 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.

5. (24 points) Blood Pressure

A research team conducts a randomized controlled experiment to evaluate a blood pressure treatment. They randomly assign 300 patients to the treatment group (Group A) and 200 to the control group (Group B). At the end of the experiment, they measure each patient's blood pressure. They want to test the claim that the treatment decreases blood pressure. Their null hypothesis states that there is no difference in the distribution of blood pressure between the two groups, and any observed difference is due to chance. As their test statistic, they decide to use the difference between the average blood pressures of Group A and Group B (that is, the average of Group A minus the average of Group B).

The research team asks for your help in simulating under the null hypothesis, so they give you the table `data`, containing 500 rows, one for each patient. It has just one column, labeled `'bp'` for blood pressure, which contains a numerical measurement of blood pressure in mmHg. It doesn't have any labels, so you don't know who was assigned to the treatment group and who was assigned to the control group.

(a) (3 pt) You find that the mean of the patients' measurements is 120, the median is 110, and the SD is 20. Which of the following must be true? **Choose all that apply.**

- ● About 50% of the patients have measurements below 110.
- ○ About 95% of the patients have measurements between 80 and 160.
- ● 75% or more of the patients have measurements between 80 and 160.
- ○ The blood pressure measurements have at least one outlier.
- ○ If we converted the measurements to standard units, the median would be equal to the mean.

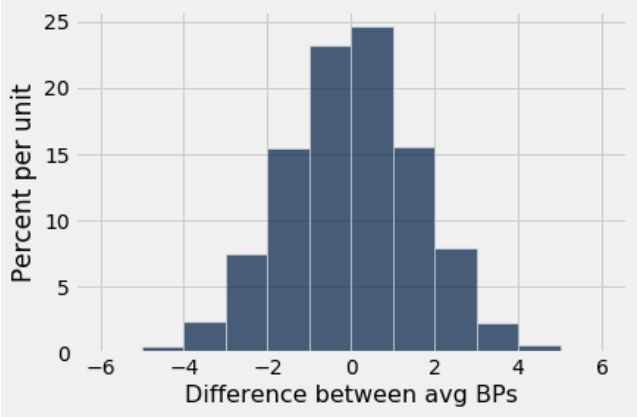
(b) (10 pt) Fill in the blanks in the code below so that the last line evaluates to one value of the test statistic simulated under the null hypothesis.

```
shuffled = data.sample(with_replacement = False)

# Two tables

shuffled_A = shuffled.take(np.arange(300))


shuffled_B = shuffled.take(np.arange(300,500))

# Two averages

mean_A = np.average(shuffled_A.column('bp'))


mean_B = np.average(shuffled_B.column('bp'))

# Test statistic


mean_A - mean_B
```

(c) (4 pt) In part (a) above, why do we shuffle the data? **Choose all that apply.**

- ○ we want to randomize treatment & control to establish causation
- ○ we want to remove all possibilities of non-random selection
- ● under the null hypothesis, the label of being in group A or group B doesn't matter
- ● we want to simulate two groups of people whose expected blood pressure is identical under the null hypothesis

(d) **(3 pt)** Suppose we repeat the procedure in part (a) 10,000 times and we plot a histogram of the simulated test statistics. Assume the histogram shows all of the simulated values.



Which of the following are valid conclusions from this histogram and from the information above? **Choose all that apply.**

- ○ The drug has no effect on blood pressure.
- ○ Because this histogram is bell-shaped, 99.7% of the simulated values must be between -3 and 3.
- ● If the p-value cutoff for the test is 0.05 and the observed test statistic is -3.5, then we should conclude the data are more consistent with the alternative hypothesis.
- ● If the p-value cutoff for the test is 0.01 and the observed test statistic is -2.5, then we should conclude the data are more consistent with the null hypothesis.

(e) **(4 pt)** Suppose the research team now tells you the observed test statistic was -15. Which conclusion(s) are the data consistent with? **Choose all that apply.**

- ● The difference in blood pressures between group A and group B isn't due to chance alone
- ○ The treatment has a positive association with blood pressure
- ● The treatment has an effect on blood pressure
- ● The treatment lowers blood pressure
- ○ There isn't enough information to make a conclusion of any kind

6. (5 points) Final Exam Studying

You ask a random sample of 250 Data 8 students from last semester how long they spent studying for the final exam in minutes. The median of the data in your sample is 9.2 hours. To quantify the uncertainty in your estimate, you create a 90% confidence interval by bootstrapping the 250 sampled students. The interval you obtain is [8.6 hours, 9.9 hours].

(a) **(3 pt)** In the blank below, describe one way to decrease the width of your interval. Your answer must fit in the blank provided.

Decrease confidence level, OR increase sample size.

(b) **(2 pt)** Suppose every single one of the 1300 students in the course this semester repeats the bootstrapping process above, and each one obtains a confidence interval. How many of the confidence intervals would you expect to **not** contain the population's median time spent studying for the final? **Choose the closest answer from the choices below.**

- ○ 650
- ● 130
- ○ 65
- ○ 0

7. (26 points) Bubonic Plague

In the Late Middle Ages (1340-1400), Europe experienced the most deadly disease outbreak in history. The Black Death, the infamous pandemic of bubonic plague, hit in 1347, killing about a third of the European human population. The table `plague`, shown below, stores information about a random sample of 750 citizens in a neighborhood of London in 1349. The first few rows are shown here.

Survived	Street	Sex	Age	Income
0	Oxford	male	22	17.5
1	Holborn	female	23	20.5
1	Bedford	female	19	55.2
1	Holborn	female	35	35.7

The table has five columns:

- **Survived**: an int, 1 if the citizen survived and 0 if the citizen died
- **Street**: a string, the street where the citizen lived
- **Sex**: a string, coded as `'male'` or `'female'`
- **Age**: an int, the age in years of the citizen
- **Income**: a float, the citizen's annual household income (in thousands of pounds)

(a) (2 pt) Suppose we want to test whether there is a difference in the distributions of `'Street'` between those who survived and those who died. Write down the appropriate test statistic to answer this question.

total variation distance between Street distributions of Survived = 0 and Survived = 1 citizens

The plague affected some parts of Europe more than others, and historians disagree over the exact number and the exact proportion of deaths in each location. We will assume that the typical death rate in cities was 33%: that is, 33% of people in cities died due to the Black Death. However, in the sample from this particular neighborhood, 40% of people died.

In parts (b) - (d), we will use a hypothesis test to examine the claim that this neighborhood had an unusually high number of people who died compared to the typical rate of 33%.

(b) (3 pt) Which of the following are correct choices for the null hypothesis? **Choose all that apply.**

- ○ Exactly 33% of the people in this neighborhood died.
- ○ Each person in the sample had a 40% chance of dying, independently of every other person.
- ● Each person in the sample had a 33% chance of dying, independently of every other person.
- ○ Each person in the neighborhood was equally likely to survive or die, independently of every other person.
- ○ The distribution of which street people lived on was the same between people who survived and people who died. Any difference observed in the sample is due to chance.
- ○ Male and female citizens died at equal rates. Any difference observed in the sample is due to chance.

(c) (3 pt) For our test statistic, we choose the percentage of people in the sample who died. Which of the following Python expressions would definitely **not** appear in our code to simulate under the null hypothesis? **Choose all that apply.**

- ○ `sample_proportions`
- ● `plague.sample(with_replacement=False)`
- ● `plague.join`
- ○ `plague.group`
- ○ `np.append`

(d) **(3 pt)** For this part, assume that our simulation under the null hypothesis gives each person a 33% chance of dying, regardless of your answer to part (b).

Suppose we simulate this neighborhood's deaths under the null hypothesis 10,000 times, and record our test statistic for each simulation. We compare our observed test statistic of 40% to these simulated test statistics and find that the p-value is 0.06. If our p-value cutoff is 0.05, which of the following are valid conclusions from the test? **Choose all that apply.**

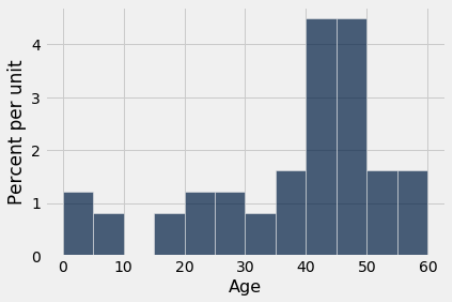
- ○ If the null hypothesis were true and we obtained many random samples, about 95% of the samples would give us a test statistic of exactly 33%.
- ○ If the null hypothesis were true and we obtained many random samples, about 6% of the samples would give us a test statistic of exactly 40%.
- ● If the null hypothesis were true and we obtained many random samples, about 6% of the samples would give us a test statistic of 40% or higher.
- ○ The probability of the null hypothesis being true is 6%.
- ○ The probability of the null hypothesis being true is 5%.
- ● The data are more consistent with the null hypothesis.

(e) **(4 pt)** The `'Sex'` and `'Age'` columns are important predictors of survival, because higher income citizens had access to better nutrition, and men were more likely than women to be merchants on the ships that brought the plague to Europe. Fill in the blank in the code to produce the table below, which gives the average household incomes of male and female citizens who died and who survived. For example, among female citizens who survived, the average household income was 57,000 pounds.

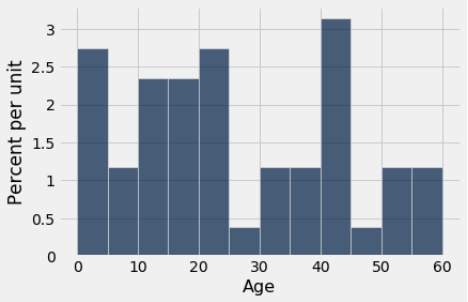
Survived	female	male
0	42.4	45.9
1	57	59.2

```
plague.pivot('Sex', 'Survived', values = 'Income', collect = np.mean)
```

(f) **(5 pt)** The histograms below show the `'Age'` distributions of citizens who survived (left plot) and citizens who died (right plot). Both plots were created using `bins = np.arange(0, 61, 5)`.



Age distribution of survivors



Age distribution of non-survivors

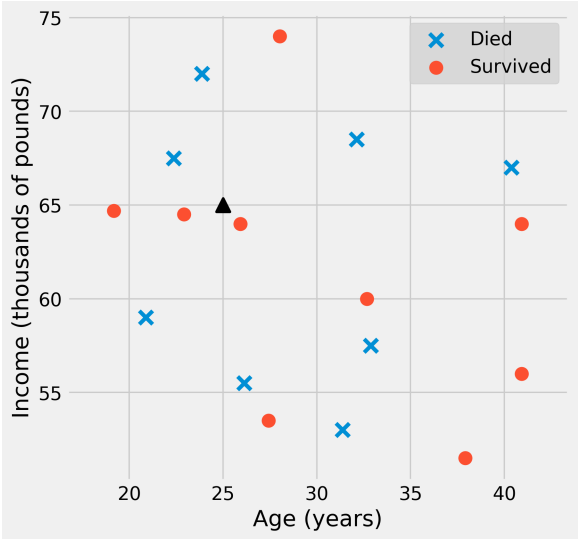
Which of the following statements can we justify based on what we can see in the two above charts? Select **ALL** answers that apply.

Note: all ranges below are inclusive. For example, "citizens between the age of 5 and 9" means their age is greater than or equal to 5 and less than or equal to 9.

- ○ Citizens with a higher income had a higher chance of surviving
- ○ Approximately 1% of the citizens who died were between the age of 5 and 9
- ● Among the citizens between the age of 45 and 49, more survived than died
- ○ Among the citizens who survived, between 10% and 15% were between 20 and 24 years old
- ● All of the citizens between the age of 10 and 14 died
- ○ All of the citizens between the ages of 25 and 29 died

For parts (g)-(h), use the following scatter plot, which contains one point for each female citizen who lived on Holborn Street. The plot shows the `'Age'` and `'Income'` for each citizen, with `'Survived'` represented by the shape of the point. The ×s correspond to citizens who died (0) and the circles correspond to citizens who survived (1).

Suppose we train a k-nearest neighbor classifier to predict whether a given female citizen on Holborn Street survived given her age and household income. We will use our classifier to predict what will happen to a hypothetical 25-year-old woman whose household income was 65,000 pounds (marked by a triangle on the scatter plot).



(g) (3 pt) If we use a 3-nearest-neighbor classifier, what will we predict for the woman marked by a triangle?

● Survive ○ Die ○ The prediction cannot be determined from the plot

(h) (3 pt) If we use a 13-nearest-neighbor classifier, what will we predict for the same woman?

○ Survive ● Die ○ The prediction cannot be determined from the plot

8. (15 points) Nearest Neighbor Regression

Continuing in the same Plague setting as the previous problem, suppose that instead of classifying whether a given citizen will survive, we would like to use nearest neighbors to predict the citizen's probability of survival. We will use the following method:

- Find the k nearest neighbors of the unclassified citizen using the methodology in parts `(g)-(h)` of the previous problem (i.e. find the closest citizens with respect to `'Age'` and `'Income'`).
- For each of the k nearest neighbors, define their *weight* as follows. Let `di` be the difference between the neighbor's household income and the unclassified citizen's household income. Let `da` be the difference between the neighbor's age and the unclassified citizen's age. Calculate the neighbor's weight as

$$\text{weight} = \frac{1}{1 + \sqrt{di^2 + da^2}}$$

  This formula gives a higher weight to closer neighbors.
- Among the k nearest neighbors, find the weighted proportion that survived. For example, if $k = 3$ and the 3 nearest neighbors have `'Survived'` values of 0, 0, & 1 with weights 0.9, 0.3, & 0.2, respectively, then the weighted proportion that survived is

$$\frac{\overbrace{0.9}^{\text{weight}} * \overbrace{0}^{\texttt{Survived}} + \overbrace{0.3}^{\text{weight}} * \overbrace{0}^{\texttt{Survived}} + \overbrace{0.2}^{\text{weight}} * \overbrace{1}^{\texttt{Survived}}}{\underbrace{0.9 + 0.3 + 0.2}_{\text{sum of weights}}}$$

In order to estimate the survival probability of an unclassified citizen using the scheme above, complete the definition of a function `knn_prob` that takes the following arguments:

- `train`: A three-column table in which the first column is labeled `'Income'`, the second column is labeled `'Age'`, and the third column is labeled `'Survived'`. Each row of the table represents a citizen in the training set.
- `citizen`: An array of length two containing the household income and age (in that order) of the citizen to classify.
- `k`: The value of k to use for k-nearest-neighbors.

The function returns the weighted proportion that survived among the k-nearest-neighbors in the scheme proposed above.

(a) (15 pt) Fill in the blanks to implement the function described above.

```
def knn_prob(train, citizen, k):
        income_diffs = train.column('Income') - citizen.item(0)
        age_diffs = train.column('Age') - citizen.item(1)
        distances = np.sqrt(income_diffs ** 2 + age_diffs ** 2)
        train_dist = train.with_column('Distance', distances)
        nn = train_dist.sort('Distance').take(np.arange(k))
        weights = 1/(1 + nn.column('Distance'))
        weighted_sum = np.sum(nn.column('Survived') * weights)
        return weighted_sum / np.sum(weights)
```

9. (30 points) Foreign Aid

Enji, an International Relations major, is writing his Masters thesis on aid given to foreign governments by the World Bank. He finds a sample of donations given to various countries over the last decade and collects these findings into a table called `aid`. Here are the first few rows.

Date	Recipient	Amount	Purpose
May 24, 2016	Zambia	595,321	agriculture
Aug 13, 2012	India	2,571,991	rail
Dec 4, 2018	Bangladesh	1,633,020	agriculture
Dec 16, 2019	Turkey	510,410	manufacturing

The table contains four columns:

- **Date**: a string, the date upon which the donation was made
- **Recipient**: a string, the country receiving the money
- **Amount**: an int, the amount of the donation in USD
- **Purpose**: a string, the reason listed for the aid

For parts (a) and (b), assume that Enji is interested in studying the average `'Amount'` of aid given per donation.

(a) (2 pt) To get a sense of the data, Enji first plots a histogram of the aid `'Amount'` in his sample. He finds that the empirical distribution of `'Amount'` has an average of \$3,532,423 and an SD of \$1,121,240. The distribution of `'Amount'` in his sample is:

- ○ Approximately normal
- ○ Not approximately normal
- ● There isn't enough information to answer this question

(b) (4 pt) Suppose Enji wants to use his sample data to create a 95% confidence interval of the true average amount of aid of all donations. If the distribution of all World Bank donations has an SD of \$1,000,000 and the `aid` table contains 10,000 rows, can Enji create a 95% confidence interval that has a width less than \$25,000?

Note: an interval of [-5,5] has a width of 10.

- ○ He can because the sample size is large enough
- ● He can't because the sample size is too small
- ○ There isn't enough information to answer this question

For the rest of the question, assume that Enji has turned his attention to learning more about aid `'Purpose'`.

(c) (3 pt) As Enji is combing through the data set, he notices that some countries in South Asia appear to have received a disproportionate amount of aid with the purpose of `'rail'` and `'manufacturing'` compared to others in the region. He creates the following table, which displays the aid given to countries in the region with the following proportions. For example, the last column tells us that of the aid that Pakistan received from the World Bank, 20% was for agriculture, 20% was for rail, and 60% was for manufacturing. Note that each country's column adds up to 1.

Purpose	India	Bangladesh	Pakistan
agriculture	0	0.9	0.2
rail	0.4	0.1	0.2
manufacturing	0.6	0	0.6

According to the above distributions, what is the empirical total variation distance of aid `'Purpose'` between India and Bangladesh? You may leave your answer as a mathematical expression (not Python).

0.9

(d) (8 pt) The World Bank claims the total variation distance of aid `'Purpose'` between India and Bangladesh is 0.3. Enji is not sure if his empirical TVD (from part (a)) is different from 0.3 just due to chance, but he thinks he could bootstrap his sample to get a better idea.

Complete the code below to write a function `purpose_tvd` that takes in a table `tbl` with the same column labels as `aid`, two country names, `country_a` and `country_b`, and computes the total variation distance between the two countries' `'Purpose'` distributions.

For example, `purpose_tvd(aid, 'Bangladesh', 'India')` should return your answer from part (c).

```
def purpose_tvd(tbl, country_a, country_b):

    dist_a = tbl.where('Recipient', country_a).group('Purpose')
    counts_a = dist_a.sort('Purpose').column(1)

    dist_b = tbl.where('Recipient', country_b).group('Purpose')

    counts_b = dist_b.sort('Purpose').column(1)

    props_a = counts_a / np.sum(counts_a)

    props_b = counts_b / np.sum(counts_b)

    return 0.5 * np.sum(abs(props_a - props_b))
```

(e) **(7 pt)** Next, complete the code below to simulate 500 bootstrap samples (bootstrap a sample of India and Bangladesh independently), compute the total variation distance between the `'Purpose'` distributions of the aid received by Indian and Bangladesh in each bootstrap sample, and store all of the results in the array `boot_tvds`. You may assume that `purpose_tvd` has been defined correctly.

```
boot_tvds = make_array()

for i in np.arange(500):

    india = aid.where('Recipient', 'India')

    bangladesh = aid.where('Recipient', 'Bangladesh')

    boot_india = india.sample(with_replacement = True)

    boot_bangladesh = bangladesh.sample(with_replacement = True)

    boot_tbl = boot_india.append(boot_bangladesh)

    new_tvd = purpose_tvd(boot_tbl, 'India', 'Bangladesh')

    boot_tvds = np.append(boot_tvds, new_tvd)
```

(f) **(3 pt)** Finally, complete the code below to compute an approximate 95% confidence interval for the population total variation distance between the `'Purpose'` distributions of the aid received by Indian and Bangladesh. After the code is executed, `left` should store the left endpoint of our interval and `right` should store the right endpoint. You may assume that `boot_tvds` has been computed correctly.

```
left = percentile(2.5, boot_tvds)
right = percentile(97.5, boot_tvds)
```

(g) **(3 pt)** Suppose it turns out that the values `left` and `right` are 0.24 and 0.78, respectively, so the confidence interval in part (f) is $[0.24, 0.78]$. Suppose we test whether or not the World Bank actually followed its claims in distributing aid to India and Bangladesh (i.e. the TVD is actually 0.3), using this confidence interval and a 5% cutoff for the P-value. Pick **ALL** the correct ways to complete the sentence:

The test will conclude that the purpose distributions of aid received by India and Bangladesh

- ○ are the same as the World Bank's claims
- ○ are different from the World Bank's claims
- ● could be the same as the World Bank's claims
- ○ probably are different from the World Bank's claims

10. (0 points) Data art (optional) Draw a graph or picture describing your experience in Data 8.

11. (0 points) Write your name in the space provided on one side of every page of the exam. You're done!