Data 8 Foundations of Data Science
Summer 2018 Final

INSTRUCTIONS

- You have 2 hours and 50 minutes to complete the exam.
- The exam is closed book, closed notes, closed computer, closed calculator, except for the official reference guide provided with the exam.
- Mark your answers **on the exam itself**. We will *not* grade answers written on scratch paper.
- You may leave numerical calculations unsimplified throughout the exam.
- You may assume that the statements `import numpy as np` and `from datascience import *` have been executed throughout the exam.
- No questions will be allowed during the time of the final. If something is unclear, you may state an assumption. If the assumption is valid, we will grade based on your assumption.

Question 0 (1 point) Write your name and SID (Student ID Number) in the space provided on one side of every page of the exam.

Last name	
First name	
Student ID number	
CalCentral email (`_@berkeley.edu`)	
GSI name and Lab Time	
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. (13 points) YouTable

You are interested in finding out who the most popular YouTube content producers in the United States are. Your colleague used the YouTube online statistics API to find out the top 250 YouTubers in the United States as of August 2018, based on total number of video views. The first few rows of her dataset `yt` are shown below.

Username	Uploads	Subs	Views	Category
WWE	36170	30296384	24854795294	Sports
Ryan ToysReview	1104	15402660	23566640362	Reviews
PewDiePie	3582	64796939	18465875954	Gaming
JustinBieberVEVO	123	34360581	17646487076	Music
KatyPerryVEVO	125	26779789	16032508300	Music
Movieclips	30147	15412747	15409779576	Movies
TaylorSwiftVEVO	86	28223429	15293096543	Music
shakiraVEVO	142	19720506	14187025758	Music
FunToys Collector Disney Toys...	2395	10563819	14111999766	Reviews
BuzzFeedVideo	5434	17031579	13741195398	News

... (240 rows omitted)

In the table above, the entries in the columns **Username** and **Category** are of type `string` while the entries in all other columns are of type `int`. **Uploads** corresponds to number of video uploads per artist, **Subs** is number of active subscribers, and **Views** is total amount of views in all the videos a YouTube artist has released.

Fill in the blanks of the Python expressions to compute the described values. You must use only the lines provided to get full credit.The last line of each answer should evaluate to the value described. Assume that the statements `import numpy as np` and `from datascience import *` have been executed. You may enter anything you would like in the blanks below, but you may not add code outside of the blanks.

(a) (1 pt) The total number of YouTubers in this table who have more than 1000 video uploads.

```
yt._____________________________________________________________________________
```

(b) (2 pt) You realize that the music video streaming platform VEVO contributes a lot of the top US YouTube artist accounts like "TaylorSwiftVEVO" and "shakiraVEVO". Create a table that only contains channels from VEVO, with the artists with the most subscribers at the top.

```
yt._____________________________________________________________________________
```

(c) (3 pt) You are interested in finding the category (ex. Sports, Music, News) with the largest average number of subscribers. Fill in the lines below so that the last line evaluates to said category.

```
only_sub_and_cat = yt._____________________________________________________________

only_sub_and_cat._____________________________________________________________
```

(d) (3 pt) Your friend proposes that a more effective popularity metric for YouTube content creators might be view count per video upload. Create a table with one column that contains only the usernames of the 5 top US YouTubers, sorted by highest view count per video upload.

```
views_per_upload = ____________________________________________________________________

yt_new_metric = yt.with_column(_____________________________________________________)

yt_new_metric._____________________________________________________________________
```

In addition to the top 250 US YouTube accounts table from before, your friend finds a second table for the top 250 UK (United Kingdom) YouTube accounts, sorted by subscriber count. The first few rows of the `yt_uk` table are shown below.

Username	Uploads	Subs	Views
Little Baby Bum - Nursery Rhymes	651	16037925	17249720731
Ed Sheeran	122	33076017	14132593402
DanTDM	2889	19827701	13316910259
OneDirectionVEVO	171	23533270	8204851709
Coldplay	202	12357179	7704244805
AdeleVEVO	31	16305109	7304218600

... (244 rows omitted)

(e) (4 pt) Find the number of YouTubers that are **only present in one** of the top 250 subscriber lists between the two countries.

```
num_youtubers_in_both =____________________________________________________________

_____________________________________________________________________________________
```

2. (10 points) Spotify Shuffle

Fahad creates a Spotify music playlist by randomly sampling 100 songs from a large collection of tunes from the Data 8 staff music library. It is known that staff library contains 20% Hip-Hop songs, 35% Pop songs, 15% Rock songs, 25% Dance songs, and 5% Country songs.

Vinitra notices that in Fahad's playlist, there are 30% Hip-Hop songs, 30% Pop songs, 20% Rock songs, 18% Dance songs, and 2% Country songs.

Initially, Vinitra believes that Fahad's claim that he randomly sampled 100 songs from the large staff library is false. Specifically, she believes he purposefully sampled more Hip-Hop songs.

(a) (3 pt) Describe a test statistic that we could compute given a sample that can help us pick between the two viewpoints: Fahad sampled the songs randomly and any difference is due to chance, or he was biased in his sampling technique and sampled more Hip-Hop songs.

(b) (2 pt) Fill in the blank with the best number possible. No need to show your work.
If the null hypothesis was true, we expect our test statistic to be roughly equal to _____,
and any difference in our sample is just due to random chance.

Now, Vinitra decides that she does not care about Fahad's bias towards Hip-Hop. Instead, she simply does not believe that Fahad truly created his playlist by randomly sampling the 100 songs from the large staff library. She does not know how he sampled, but she believes it is not completely random.

(c) (3 pt) Describe a test statistic that can help pick between the two viewpoints: Fahad sampled the songs randomly and any difference is due to chance, or he did not sample the songs randomly from the staff playlist.

(d) (2 pt) Fill in the blank with the best number possible. No need to show your work.
If the null hypothesis was true, we expect our test statistic to be roughly equal to _____,
and any difference in our sample is just due to random chance.

3. (8 points) You're Samply Too Mean

The average weights of all people in Berkeley is 146 lb. Through some experimentation, we note that the probability of seeing a sample of 100 people (with replacement) in Berkeley whose average weight is higher than 151 lb is 2.5%.

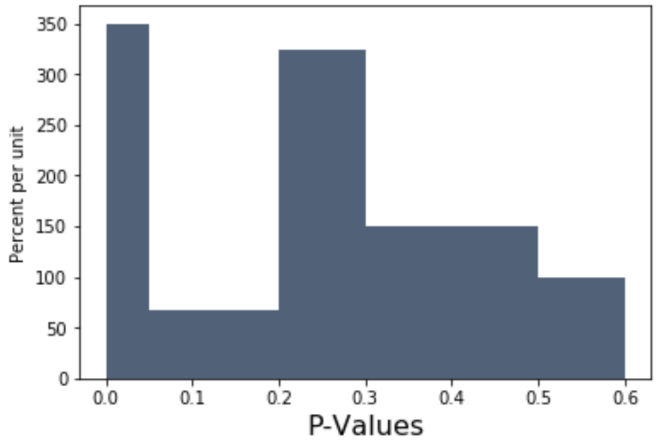
(a) (4 pt) Solve for the standard deviation of all weights in Berkeley. Show your work and box your final answer, which may be left unsimplified.

(b) (4 pt) We would like to reduce the chance of seeing an average weight of a sample (drawn with replacement) that is higher than 151 lb to roughly .15%. Solve for the smallest sample size which would achieve this goal. Show your work and box your final answer, which may be left unsimplified. Assume your answer to Part A is assigned to the variable *PopSD* and use this variable in your final expression, if needed.

4. (6 points) Error Probabilities Probably

A researcher conducts 1000 of the same hypothesis tests (same null, alternative, and test statistic) with different samples of observed data from a population. Throughout her hypothesis tests, she uses a p-value cutoff of .25.

Suppose the following is a histogram of the p-values recorded during the conclusion for each of the 1000 hypothesis tests. The bin sizes are all a multiple of .05.



(a) (3 pt) Assume the null hypothesis in these tests happened to be true. If you can tell, did we reject the null hypothesis more than expected, less than expected, or exactly as many times as we would expect? **Explain** your answer. If we can not determine with this information, **explain** why not.

- ◯ More than expected
- ◯ Less than expected
- ◯ Exactly equal to expected
- ◯ Cannot determine

Explanation:

(b) (3 pt) Assume the alternative hypothesis in these tests happened to be true. If you can tell, did we fail to reject the null hypothesis more than expected, less than expected, or exactly as many times as we would expect? **Explain** your answer. If we can not determine with this information, **explain** why not.

- ◯ More than expected
- ◯ Less than expected
- ◯ Exactly equal to expected
- ◯ Cannot determine

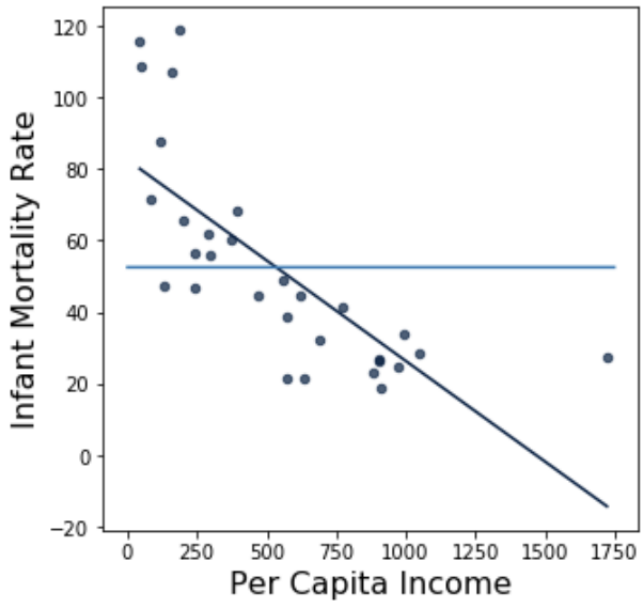
Explanation:

5. (19 points) Linear Regret-ssion

We are interested in seeing how the Per Capita Income was related with Infant Mortality rate for various countries in the early 1950s. The data is encapsulated in a table named `countries`, whose first few rows are shown below:

Country	Per Capita Income	Infant Mortality Rate
Venezuela	392	68.5
Mexico	118	87.8
Ecuador	44	115.8
Colombia	158	106.8
Ceylon	81	71.6

We are interested in looking at the relationship between Per Capita Income and Infant Mortality Rate. We plot two lines that we will use to predict infant mortality rate in a country (on average), given a per capita income. One is the regression line, and one is the constant line at the average of infant mortality rate.



Here are some relevant lines of code and their outputs. Assume all of the functions below (descriptions of which appear on your study guide) are defined already.

Expression	Output
`correlation(countries, 1, 2)`	-.74
`np.mean(countries.column(1))`	534.1
`np.mean(countries.column(2))`	52.5
`np.std(countries.column(1))`	384.8
`np.std(fitted_values(countries, 1, 2))`	21.6

(a) **(4 pt)** If possible, calculate the slope of the regression line in the plot above. Show your work with arithmetic. Please **do not** write any Python expressions. Box your final answer, which you may leave unsimplified. If it is not possible, explain why not.

(b) **(2 pt)** At what value of per capita income would we predict the infant mortality rate to be 0 given the regression line on the previous page?

- ◯ 0
- ◯ 52.5
- ◯ 80
- ◯ 534.1
- ◯ 1500
- ◯ None of the above

(c) **(2 pt)** At what value of per capita income would we predict the infant mortality rate to be 0 given the average line on the previous page?

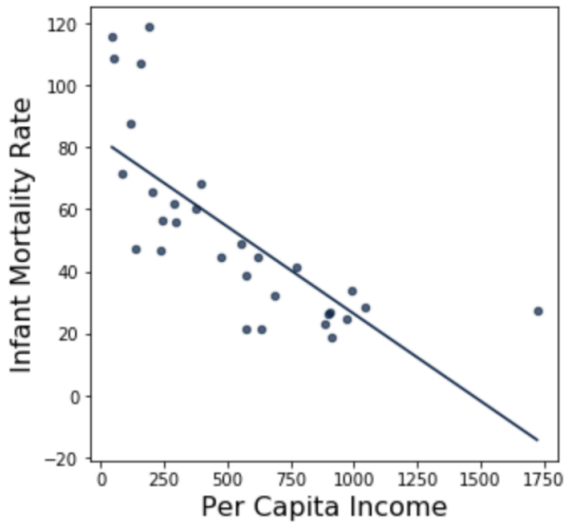
- ◯ 0
- ◯ 52.5
- ◯ 80
- ◯ 534.1
- ◯ 1500
- ◯ None of the above

(d) **(3 pt)** Assume the RMSE (root mean squared error) of the regression line is 19.3. If possible, calculate the RMSE of the constant line representing the average of the infant mortality rate on this data. Show your work with arithmetic. Please **do not** write any Python expressions. Box your final answer, which you may leave unsimplified. If it is not possible, explain why not.

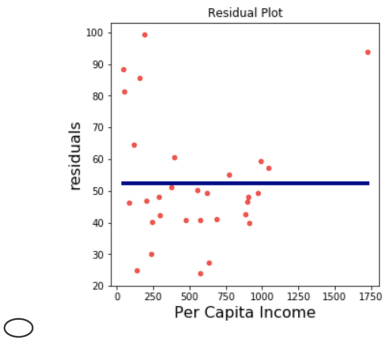
(e) **(3 pt)** If possible, fill in the blank with a mathematical expression. Box your final answer, which you may leave unsimplified. If it is not possible, explain why not.

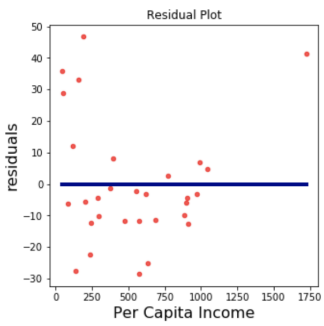
We use our regression line to estimate infant mortality rate using per capita income. At-least 88.9% of our predictions of infant mortality rate will be correct within plus or minus _____ units.

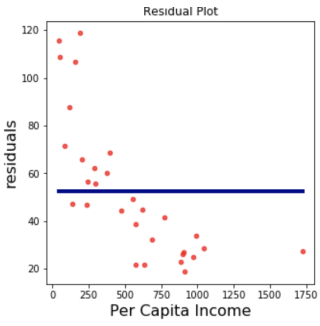
We move our attention to only the regression line and we would like to determine whether or not linear regression was a good idea to begin with. We show the regression line plotted on the scatter plot below again for convenience.



(f) (2 pt) Which of the following plots is the best approximation to the residual plot, along with the line y=0?







(g) (1 pt) What is the approximate average value of the residuals?

- ◯ -10
- ◯ 0
- ◯ 52.5
- ◯ 534.1
- ◯ Need more information

(h) (2 pt) Based on the information above, select the most appropriate statement below.

- ◯ Our residual plot shows no obvious pattern, so linear regression is a good fit for this data.
- ◯ Our residual plot shows a definitive pattern, but based on our original plot, linear regression is still a good fit for this data.
- ◯ Our residual plot shows a linearly decreasing trend, so linear regression is not a good fit for this data.
- ◯ Our residual plot shows a definitive pattern, so linear regression is not a good fit for this data.
- ◯ None of the above

6. (18 points) HA/Biness Among Countries

We are interesting in studying the distribution of human happiness across the world. To do so, we have a table called `happiness`. The table has 140 rows, and the first few rows are shown below:

Region	Happiness Score
Southern Asia	5.196
Latin America and Caribbean	5.538
Sub-Saharan Africa	3.916
Middle East and Northern Africa	5.303
Sub-Saharan Africa	4.121

Each row contains the region and happiness score for a sampled country. The countries are not shown in the table above. Throughout this question, our goal will be to measure if the distribution of happiness scores across regions in the whole world are roughly equivalent.

(a) (2 pt) Define a function `region_happiness` which takes in a table like `happiness` and returns a two column table. The output table should have one column with the names of the unique regions, and a second column with the average happiness score in that region. The first few rows of an output table are shown below, with one row for each unique region. Note that the input table may not have the same labels as `happiness`, but will have the same ordering of columns.

Australia and New Zealand	7.334
Central and Eastern Europe	5.34732
Eastern Asia	5.414
Latin America and Caribbean	5.98987
Middle East and Northern Africa	5.31077

```
def region_happiness(tbl):

    return ____________________________________________________________________________
```

We are in an abnormal situation, in which we have 10 different regions and we are interested in figuring out if the numerical distributions of all of the regions happiness scores are roughly equivalent. To do this, we will use a variant of A/B testing called multivariate testing. The only difference between the two methods is we have more than two groups (in this case, 10), and we are comparing the numerical distributions of all of these groups at the same time.

The important thing to notice is that our null hypothesis and alternative hypotheses retain the same structure as they have for A/B testing, as well as our method for simulating a sample under the null. The only difference will be the test statistic we decide to use to differentiate between our two viewpoints.

We are interested in testing if the distributions of happiness scores of all the regions come from the same underlying population distribution or not using our sample we have above.

(b) (4 pt) Describe a specific null and alternative hypothesis that will help us pick between the two viewpoints presented above.

Null:

Alternative:

(c) (2 pt) To help us differentiate between our two hypotheses, we need to choose a test statistic. Choose the best test statistic and corresponding explanation.

- ⬭ We should choose the TVD between the average happiness of each region and the expected average happiness of each region as our test statistic because only small values of our test statistic will point to the alternative hypothesis.
- ⬭ We should choose the TVD between the average happiness of each region and the expected average happiness of each region as our test statistic because only large values of our test statistic will point to the null hypothesis.
- ⬭ We should choose the standard deviation of the average happiness per region as our test statistic because only large values of our test statistic will point to the alternative hypothesis.
- ⬭ We should choose the standard deviation of the average happiness per region as our test statistic because only small values of our test statistic will point to the alternative hypothesis.
- ⬭ We should choose the median of the average happiness per region as our test statistic because only large values of our test statistic will point to the alternative hypothesis.

(d) (2 pt) Define a function `test_statistic` which takes in a table like `happiness` and returns the test statistic you chose from the last question. Note that the input table will not have the same labels, but will have the same ordering of columns.
You may use your `region_happiness` function from above and assume it is fully correct.

```
def test_statistic(tbl):

    regions = _____________________________________________________________________

    return _____________________________________________________________________
```

(e) (3 pt) Complete the definition of `sample_under_null`, which takes in no arguments and returns one value of the test statistic applied to a random sample simulated under the null hypothesis.
You may use your `test_statistic` function from above and assume it is fully correct.

```
def sample_under_null():

    shuffled_hap = happiness._____________________________________________________

    r_and_h = Table().with_column(_____________________________________________)

    return _____________________________________________________________________
```

(f) (3 pt) We would now like to simulate 1000 test statistics under the null hypothesis. Complete the code below to assign `simulated_stats` to 1000 values of test statistics applied to different samples simulated under the null hypothesis.
You may use your `sample_under_null` function from above and assume it is fully correct.

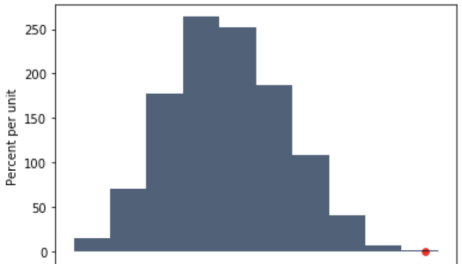
```
simulated_stats = ____________________________________________________

for i in np.arange(1000):

    stat = ____________________________________________________________________________

    simulated_stats = ____________________________________________________________________________
```

The following is a histogram of the simulated test statistics, and the dot (on the right) is the value of your calculated test statistic on the original observed data from the sample.



(g) (2 pt) Select the best conclusion from the options below:

- ◯ Our data is more consistent with the null hypothesis.
- ◯ Our data is more consistent with the alternative hypothesis.
- ◯ It is impossible to decide given just the histogram above.

7. (8 points) You Studied for these Studies

(a) (2 pt) A medical observational study finds a high positive association between drinking coffee and having lung cancer in a group of individuals. Which of the following statements, if true, could present a confounding factor to the study? Choose the best answer.

- ◯ There is a strong positive association between having lung cancer and drinking coffee in this group.
- ◯ There is a strong positive association between drinking coffee and smoking in this group.
- ◯ There is a no association between smoking and having lung cancer in this group.
- ◯ All of the above
- ◯ None of the above

(b) (2 pt) Researchers decide to test their new disease treatment equipment against their older equipment. However, their new disease equipment requires patients remain hydrated. Researchers select a group of 300 patients for this study; some of the participants will use the new equipment, while the others will use the old equipment.
Mark all of the following options which would help control for confounding factors in this study.

- ◯ Requiring patients who are treated with the older equipment remain hydrated.
- ◯ Requiring patients who are treated with the newer equipment remain dehydrated.
- ◯ Assigning patients to groups at random.
- ◯ Assigning different amounts of water randomly to individuals in both groups.
- ◯ None of the above

(c) (4 pt) A local store is the only umbrella store in a large radius. They decide to test whether or not the quality of their umbrellas impacts their yearly sales. In 2016, their umbrellas were of a very high quality and they sold 1,500 umbrellas. In 2017, their umbrellas were of average quality and they sold 1,000 umbrellas.

The store is tempted to explain this phenomenon by claiming that a higher quality of umbrellas causes more sales. As data scientists, however, we are more hesitant to come to this conclusion. Give **two different specific** alternative explanations that could explain the results of the experiment above. You may **not** simply claim that this is an observational study.

1.

2.

8. (8 points) P-Value Puzzle

Define the function `p_value_calculation`, which takes in the following 3 arguments:

- `sim_vals` is an array of simulated test statistics under a specific null hypothesis.
- `observed_ts` is the value of a test statistic from a specific observed sample.
- `larger_alt` is a boolean which is either `True` if only large values of the test statistic above point towards the alternative, and `False` otherwise (small values of the test statistic point towards the alternative).

The function should calculate the P-Value of the `observed_ts` (observed statistic), with respect to the null hypothesis under which `sim_vals` was simulated. You may not need all of the lines.

```
def p_value_calculation(sim_vals, observed_ts, larger_alt):

    ------------------------------------------------------------------------

    ------------------------------------------------------------------------

    ------------------------------------------------------------------------

    ------------------------------------------------------------------------
```

9. (6 points) Minimize That!

(a) (2 pt) Assume we have a table, `tbl`, which has two columns. We also have a function which takes in two numbers that represent column indices, and returns the **negative of the correlation** between the two columns specified by the indices in `tbl`. The first number passed in to the function is the column index for x, and the second number passed in is the column index for y.
Assume the correlation between the two columns in `tbl` is .7. If multiple arguments minimize a function, `minimize` randomly chooses one. Which of the following could be the result of calling `minimize` on our function above?

- ◯ `array([0,1])`
- ◯ `array([1,0])`
- ◯ Both of the above.
- ◯ None of the above.
- ◯ Cannot tell with this information.

(b) (4 pt) We are interested in using `minimize` to find the column with the smallest element in a table `tbl2`. Assume `tbl2` only contains numbers. Define a function `minimized_fn` which, when passed in to minimize, gives the column index (number) of the column which contains the smallest element in the table. You may not need all of the lines provided below. Afterwards, use your new function to assign `smallest` to the smallest entry in `tbl2`.

```
def minimized_fn(_____________________________________):

        ____________________________________________________________________________

        ____________________________________________________________________________


smallest = ____________________________________________________________________________
```

10. (16 points) Confidence in Crime

In 1968, the United States Census Bureau took a large random sample of metropolitan areas and measured their crime rates, measured in proportion per 100,000. The information is encapsulated in a one column table called `crimes`.

(a) (2 pt) Which of the following pieces of information can we determine given the sample? **Mark all that apply**.

- ◯ The average crime rate in all metropolitan areas in the US in 1968.
- ◯ The approximate distribution of crime rates in metropolitan areas in the US in 1968.
- ◯ The range of crime rates in all metropolitan areas in the US in 1968.
- ◯ None of the above

Assume we bootstrap our sample many times to get an approximate 90% confidence interval for the average crime rate of the population. The resulting interval is (0.0256, 0.0287).

(b) (2 pt) What is the probability that the true average crime rate in the population lies in this interval?

- ◯ 100%
- ◯ 0%
- ◯ 90%
- ◯ 95%
- ◯ None of the above

(c) (2 pt) True or False: Approximately 90% of the population crime rates lie between .0256 and .0286.

- ◯ True
- ◯ False

(d) (2 pt) True or False: Approximately 90% of the sample crime rates lie between .0256 and .0286.

- ◯ True
- ◯ False

(e) (3 pt) Suppose the Census Bureau went out and actually sampled two times as many metropolitan areas, so our sample became larger. Would our 90% confidence interval we calculate using the new data be larger, smaller, around the same size as our original interval, or can we not tell?

- ◯ Larger
- ◯ Smaller
- ◯ Same size
- ◯ Cannot determine

(f) (3 pt) Assume we want to test the hypothesis that the average population crime rate is .03, with our alternative being that it's not. Give an interval of p-value cutoffs such that we can reject the null hypothesis that the average population crime rate is .03 given our confidence interval above. Explain your answer. Your range should be contained within the interval [0,1].

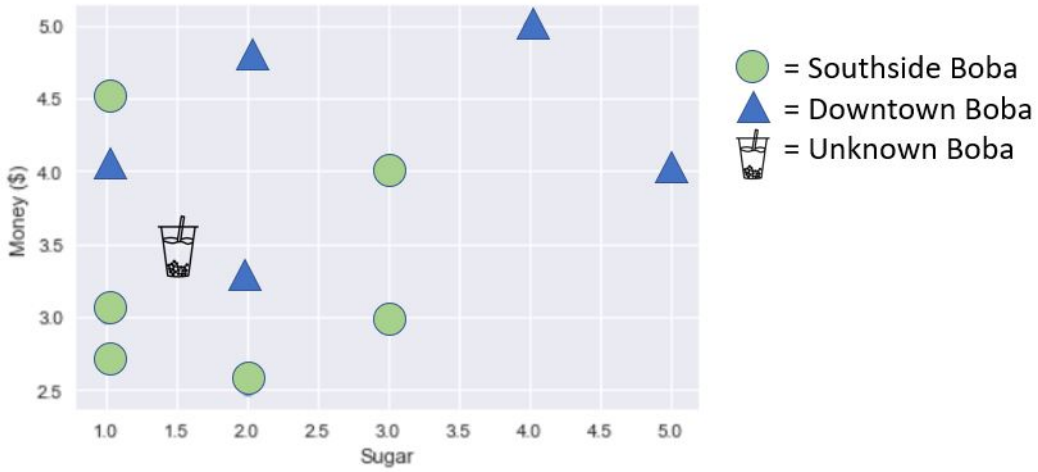
(g) (2 pt) Suppose we repeat the process of creating 90% confidence intervals many times using different samples from the population each time, with the hope of approximately 2,700 intervals containing the true population average crime rate. How many confidence intervals should we create?

- ◯ 900
- ◯ 2,430
- ◯ 2,700
- ◯ 2,850
- ◯ 3,000
- ◯ 3,300
- ◯ None of the above

11. (7 points) K-Nearest Bobas

Tawainese Pearl Milk Tea (also known as boba) is a very common drink on the Berkeley campus. The two main concentrations of boba cafes are on Southside and Downtown Berkeley. Your friend is a Berkeley boba expert. He says that two useful features for classifying boba are the amount of sugar (y-axis) and the amount of money (x-axis) for a standard black milk tea with pearls. He identifies several boba places for you, but leaves one unknown boba cafe for you to classify, because he hasn't tried it yet. No two boba places have exactly the same features.

You decide to construct a classifier for the unknown boba cafe using all of the known boba cafes as the training set. Note that the x and y axes have different scales.



(a) (1 pt) The prediction scheme outlined above is a unsupervised machine learning algorithm.

- ◯ True
- ◯ False
- ◯ Cannot determine

(b) (1 pt) Using a 5-nearest neighbor classifier, what would the unknown boba cafe be classified as?

- ◯ Southside
- ◯ Downtown
- ◯ Cannot determine

(c) (3 pt) What is the problem with using a 4-nearest neighbor classifier for our unknown boba cafe? Provide one possible solution to this problem.

(d) (2 pt) Every possible pair of feature values for a new boba cafe would be classified as Southside for a 11-nearest neighbor classifier using this training set.

- ◯ True
- ◯ False
- ◯ Cannot determine

12. (0 points) (Optional) Data Visualization!

Draw a data visualization about Data 8.