Data 8 Summer 2019

Foundations of Data Science

FINAL EXAM SOLUTIONS

INSTRUCTIONS

- The exam is worth 140 points. You have 170 minutes to complete it.
- The exam is closed book, closed notes, closed computer/phone/tablet, closed calculator, except the official midterm exam reference guide provided with the exam.
- Write/mark your answers on the exam in the blanks/bubbles provided. Throughout this exam, we will give you checkboxes (☐) for "select all that apply" questions and bubbles (⬭) for questions with a single answer. 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`)	
Name of 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)**	

This page is intentionally left blank. You can use it for scratch work but it will not be graded.

1. (6 points) What Would Python Do?

For parts **a** and **b**, choose the bubble corresponding to what Python would output after the expression is evaluated. If the expression would cause an error, select `Error`.

(a) (1 pt) `make_array(9, 8, 7) + 3`

- ○ `Error`
- ○ `np.array([9,8,7,3])`
- ● `np.array([12,11,10])`
- ○ `np.array([9,9,9,8,8,8,7,7,7])`

(b) (1 pt) `make_array(1, 3, 4) + np.arange(0, 4, 2)`

- ● `Error`
- ○ `np.array([1,7,6])`
- ○ `np.array([1,5,8])`
- ○ `np.array([1,3,0,3,4,2])`

(c) (2 pt)

What is the value of `z` after the four following lines of code have been run?

```
x = make_array(0, 3, 4, 2, 1)
y = x > 2
z = np.count_nonzero(y)
z
```

- ○ The code causes an `Error`
- ● 2
- ○ 3
- ○ `False`

(d) (2 pt) You are keeping track of every new TV show you binge watch in an array. You have already seen *Friends* and *How I Met Your Mother*, so in your jupyter notebook you have:

```
tv_shows = make_array("Friends","How I Met Your Mother")
```

Suppose you just finished watching *The Office* and want to add it to your `tv_shows` array, so you type:

```
np.append(tv_shows, "The Office")
```

After typing and running the code exactly as above, what would the array `tv_shows` contain? Write the contents of the array on the following blank:

```
np.array(["Friends", "How I Met Your Mother"])
```

2. (10 points) Shoot Your Shot

The `shots` table has information for every shot taken (including free throws) in the 2014 - 2015 NBA season. Assume that this table contains information regarding every single point scored in a game this season. `GAME_ID` is the unique identifier for each game, and each team has a unique three letter code.

GAME_ID	QUARTER	shot_value	SHOT_RESULT	player_name	team	MINUTES	SECONDS
21400533	3	2	missed	klay thompson	GSW	12	0
21400431	3	2	missed	klay thompson	GSW	12	0
21400406	2	2	made	gerald green	PHX	12	0
21400908	4	2	made	chris kaman	POR	11	47
21400908	2	2	made	chris kaman	POR	11	35
21400908	3	2	missed	lamarcus aldridge	POR	11	1
21400908	3	2	made	lamarcus aldridge	POR	11	29
21400908	3	2	missed	lamarcus aldridge	POR	11	35
21400908	1	2	made	lamarcus aldridge	POR	11	44
21400908	2	2	missed	spencer hawes	LAC	11	10

... (128059 rows omitted)

(a) (2 pt) Write code to produce a table called `points` that only contains shots that were "made" (points are only earned if a shot is made in the basket). The resulting table should have three columns: `GAME_ID`, `shot_value`, and `team`.

```
points = shots.where("SHOT_RESULT", are.equal_to("made")).select("GAME_ID", "shot_value", "team")
```

(b) (2 pt) Assuming the `points` table was implemented correctly, write code to produce a table `scores` that contains the total number of points scored *per team per game*.

```
scores = points.group(make_array("GAME_ID", "team"), sum)
```

(c) (2 pt) Assuming the `scores` table was implemented correctly, write code to produce a table `sorted_scores` where within each game (two rows with the same `GAME_ID` but different teams), the row representing the team that scored the most points comes before the row representing the team that scored the least points.

```
sorted_scores = scores.sort("shot_value sum", descending = True).sort("GAME_ID")
```

(d) (4 pt) Assuming the `sorted_scores` table was implemented correctly, write code to produce a table called `win_count`, which should contain two columns:

- `team`: The team name
- `count`: the number of games the team won

The team that won the most games should be the first row of the table.

```
winners_of_games = sorted_scores.take(np.arange(0,sorted_scores.num_rows,2))
unsorted_winners = winners_of_games.group("team")
win_count = unsorted_winners.sort("count", descending=True)
```

3. (18 points) What's The Chance I Like Boba?

UGSI Katherine loves getting boba in Berkeley, and she goes to three different boba shops. Since her favorite boba place is Gong Cha, half of the time that she gets boba she goes to Gong Cha, and the other half of the time she randomly chooses between Yi Fang and U-Cha, with an equal chance of choosing either shop. Katherine has very specific boba tastes. At U-Cha, she gets milk tea 2/3 of the time and fruit tea 1/3 of the time. At Yi Fang, she only gets fruit tea. At Gong Cha she chooses between fruit tea and milk tea with equal probability of choosing each one.

(a) (3 pt) If Katherine gets boba twice, what is the probability that she goes to Yi Fang both times?

- ○ $(\frac{1}{4}) \times 2$
- ● $(\frac{1}{4})^2$
- ○ $1 - ((\frac{1}{4})^2)$
- ○ $1 - ((\frac{1}{4}) \times 2)$
- ○ $(\frac{1}{4}) \times (\frac{1}{2})$
- ○ $(\frac{1}{4}) + (\frac{1}{2})$

(b) (3 pt) If Katherine gets boba once, what is the probability that she goes to U-Cha and gets fruit tea?

- ● $\frac{1}{4} \times \frac{1}{3}$
- ○ $\frac{1}{4} + \frac{1}{3}$
- ○ $\frac{1}{3}$
- ○ $1 - ((\frac{1}{4}) \times (\frac{1}{3}))$

(c) (3 pt) What is the probability that Katherine gets fruit tea if she got boba once?

- ○ $(\frac{1}{2})^2 \times (\frac{1}{4})^2 \times (\frac{1}{3})$
- ○ $\frac{1}{2} + 1 + \frac{1}{3}$
- ○ $\frac{1}{2} + \frac{1}{3}$
- ● $(\frac{1}{2})^2 + (\frac{1}{4}) + (\frac{1}{4}) \times (\frac{1}{3})$

(d) (3 pt) Given that Katherine bought milk tea, what is the probability that she went to Gong Cha?

- ○ $\frac{1}{2}$
- ○ $(\frac{1}{4} \times \frac{2}{3})/[(\frac{1}{2} \times \frac{1}{2}) + (\frac{1}{4} \times \frac{2}{3})]$
- ● $(\frac{1}{2} \times \frac{1}{2})/[(\frac{1}{2} \times \frac{1}{2}) + (\frac{1}{4} \times \frac{2}{3})]$
- ○ $\frac{1}{2} \times \frac{1}{2}$

(e) (3 pt) If Katherine got boba 3 times, what is the probability that she went to UCha at least twice?

- ○ $1 - (\frac{3}{4})^3$
- ○ $\frac{1}{4}^3 + \frac{1}{2} \times \frac{1}{3} \times \frac{1}{4}$
- ● $(\frac{1}{4})^3 + 3 \times (\frac{1}{4}^2 \times \frac{3}{4}))$
- ○ $\frac{1}{4}$

(f) (3 pt) Read the following code and determine what value `final_value` best approximates.

```
counter = 0
for i in np.arange(1000):
    if np.random.choice(np.arange(12)) < 3:
        if np.random.choice(np.arange(6)) > 3:
            counter = counter + 1
final_value = counter / 1000
```

- ○ The number of times that Katherine gets milk tea in 1000 visits.
- ○ The probability that Katherine gets fruit tea on any particular visit.
- ● The probability that Katherine goes to U-Cha and gets fruit tea.
- ○ The number of times that Katherine gets fruit tea when she goes to Yi Fang.
- ○ The probability that Katherine goes to Gong Cha and gets fruit tea or she goes to U-Cha and gets milk tea.
- ○ The probability that Katherine gets milk tea on any particular visit.

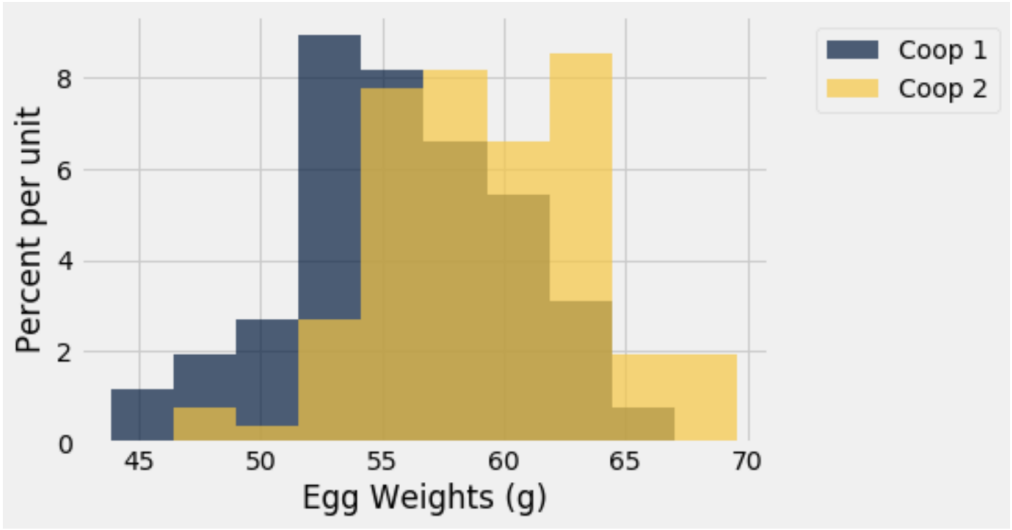
4. (22 points) Don't Count Your Chickens

You're an avid chicken farmer. You built two coops for two groups of chickens (your Bantams, which lay small eggs, and your Jersey Giants, which lay big eggs) since you think they would prefer different sized nesting boxes. Unfortunately, when you visit your coops during the day, you are dismayed to see that it seems like chickens choose coops at random to spend time in, and your hard work was for nothing.

However, your data scientist friend helps you collect eggs one morning and notices that the eggs in one coop seem to tend to be larger than the eggs in the other! Perhaps the chickens do show preference in where they sleep at night and lay their eggs. Overjoyed, you decide to investigate: one morning, after all the chickens have exited their coops, you measure the weight of each egg, and record in which coop it was found. You create the following table called `eggs` to store the data, and then plot this histogram to help you visualize the difference:

Weight (g)	Coop
52.1923	1
61.6388	1
61.3706	1
53.0071	1
55.213	2
58.1972	1
61.4176	1
62.2451	2
60.8448	2
57.4096	1

... (190 rows omitted)



You want to conduct a hypothesis test to see if the distributions of egg weights come from one underlying distribution and thus support the idea that your chickens have no coop preference, or if instead the distributions of egg weights in the two coops are truly different.

(a) (2 pt) If we took a single bootstrapped resample of the Coop 1 eggs, what would the histogram of egg weights in our resampled table look like?

- ○ The histogram would look like the Coop 1 histogram above, but narrower.
- ○ The histogram would look like the Coop 1 histogram above, but more normal.
- ○ The histogram would be like the Coop 1 histogram above, but wider.
- ● The histogram would not change very much from the Coop 1 histogram above.

(b) (2 pt) How would you perform a single simulation under the null hypothesis that the distribution of weights in the two coops is the same?

- ○ Create two normal distributions centered at the mean of Coop 1 and Coop 2 and see if they overlap.
- ● Shuffle the coop labels of your collected eggs and compute your test statistic on the shuffled table.
- ○ Bootstrap the egg weights and calculate a new average egg weight.

(c) (2 pt) What test statistic would **best** help differentiate between the null and the alternative hypothesis?

- ○ The Total Variation Distance (TVD) between the distribution of Coop 1 egg weights and Coop 2 egg weights.
- ● The absolute value of the difference between the mean egg weight in Coop 1 and the mean egg weight in Coop 2.
- ○ The mean egg weight of Coop 2.
- ○ The difference between the mean egg weight in Coop 2 and Coop 1.

(d) (4 pt) Write a function `compute_one_test_stat` that will take in the table with the same columns as `eggs` above and returns one simulated value of your test statistic.

```
def compute_one_test_stat(tbl):

    new_labels = tbl.sample(with_replacement=False).column("Coop")

    new_table = tbl.with_columns("Shuffled labels", new_labels)

    means_col = new_table.group("Shuffled labels", np.mean).column('Weight (g) mean')

    test_stat = abs(means_col.item(0) - means_col.item(1))

    return test_stat
```

(e) (2 pt) You use the `compute_one_test_stat` function to compute the observed test statistic, which is stored in the variable `obsv_test_stat`, as well as 10,000 simulated test statistics, which are stored in an array called `stats`. Which of the following lines of code correctly computes the empirical p-value of this test?

- ○ `np.count_nonzero(stats <= obsv_test_stat)/10000`
- ● `np.count_nonzero(stats >= obsv_test_stat)/10000`
- ○ `np.count_nonzero(stats == obsv_test_stat)/10000`
- ○ `np.count_nonzero(stats >= 0.05)/100000`

(f) (4 pt) Select one of the options from parts i-ii to fill in the corresponding blanks in the sentence below.

A larger observed difference in mean egg sizes would result in a ____(i)____ p-value. We would see a larger observed difference in mean egg sizes if Jersey Giant chickens exhibited a ____(ii)____ preference for one coop, e.g. Coop 2.

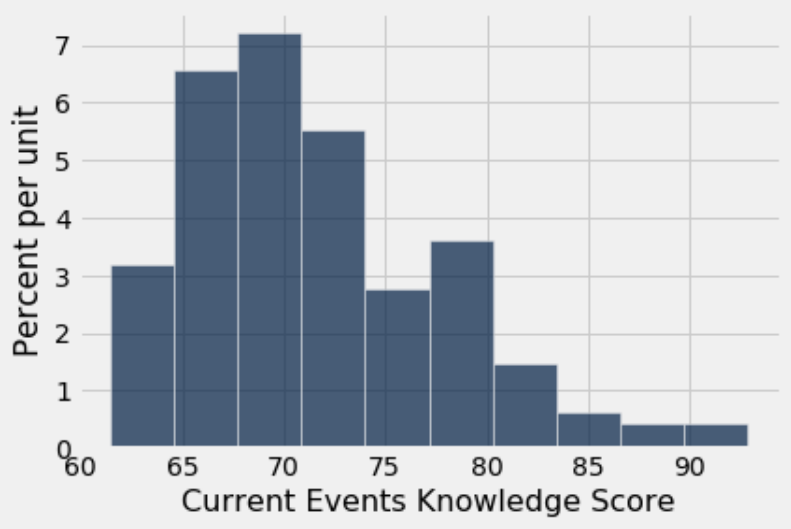
i. ● smaller ○ larger
ii. ○ weaker ● stronger

(g) (6 pt) You calculate a p-value of 0.04 for this test. Select **all** true statements that we can justifiably conclude from this p-value. If you do not have enough information to evaluate whether the statement is true or false, DO NOT select it.

- ☐ There is a 4% chance that the chickens have no preference for which coop they lay eggs in.
- ☐ There is a 96% chance that the chickens have no preference for which coop they lay eggs in.
- ☐ If chickens actually have no coop preference, there is a 4% chance we incorrectly reject the null hypothesis.
- ■ We would reject the hypothesis that the chickens have no coop preference if we used a 5% p-value cutoff.
- ■ We would fail to reject the hypothesis that the chickens have no coop preference if we used a 1% p-value cutoff.
- ■ 4% of the simulated test statistics were greater than or equal to the observed test statistic.
- ☐ We cannot conclude any of the above options.

5. (14 points) I'll Take Civic Participation for 100 Points Please

A Berkeley political science professor is concerned about their students' knowledge of current U.S political events. To assess their civic participation, the professor sent a survey to 150 randomly-selected students quizzing them on their knowledge of recent news out of a total score of 100. In the 150-student sample, the mean score was 71.7 and standard deviation was 6.2. Here is a plot of the distribution of the 150 scores:



For parts a - c, choose the option that best completes the blank.

(a) (2 pt) This distribution is _____.

- ○ Left skewed
- ● Right skewed
- ○ Normally distributed

(b) (2 pt) The mean of the above histogram is _____ the median.

- ● higher than
- ○ lower than
- ○ equal to

(c) (2 pt) The professor can expect at least _____ of the scores to be between 59.3 and 84.1.

- ○ 95%
- ○ 89%
- ● 75%
- ○ We don't have enough information to tell.

Although the professor only ended up with 150 entries for the survey, the professor decided to assume that this distribution of survey scores is representative of the entire UC Berkeley student population. To get a better estimate of the entire student populations knowledge, the professor conducts the following analysis:

- The professor resamples 150 scores with replacement from the original survey distribution.
- The professor calculate the average of those 150 scores and store it in an array called `scores`.
- The professor repeat the above process 10,000 times.

Thus, `scores` is an array of 10,000 average survey scores.

(d) **(2 pt)** If the professor graphed the `scores` distribution as a histogram, what would the shape of the histogram look like?

- ○ Left skewed
- ○ Right skewed
- ● Normally distributed
- ○ We don't have enough information to tell.

(e) **(2 pt)** If the professor graphed the `scores` distribution as a histogram, which of the following would be the best approximation of the mean/center of the distribution?

- ○ 7.17
- ● 71.7
- ○ 50
- ○ We don't have enough information to tell.

(f) **(2 pt)** If the professor graphed the `scores` distribution as a histogram, which of the following is the best approximation for the standard deviation of the distribution?

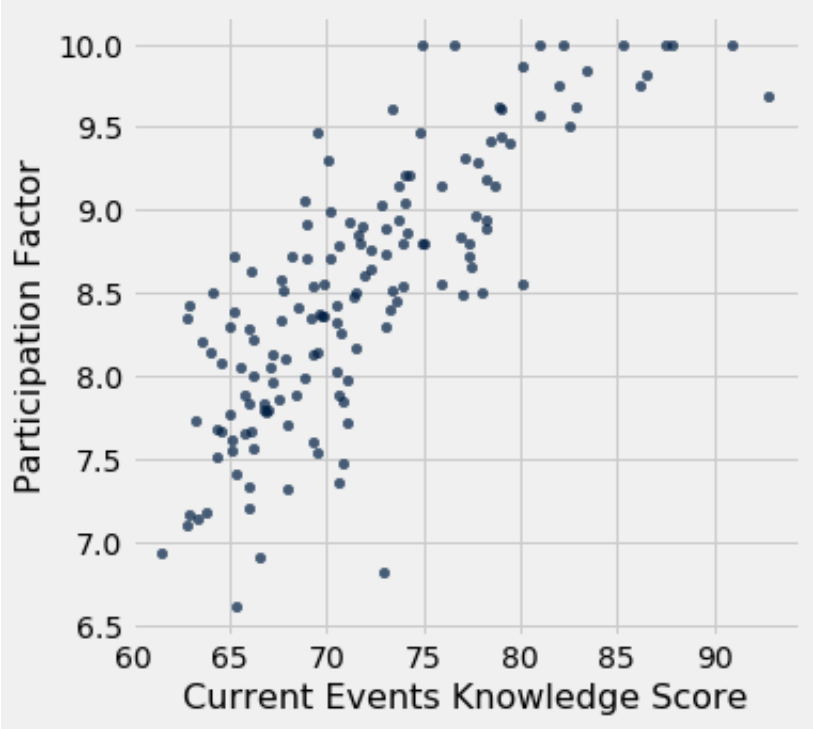
- ○ 6.2
- ● $\frac{6.2}{\sqrt{150}}$
- ○ $\frac{\sqrt{150}}{6.2}$
- ○ We don't have enough information to tell.

(g) **(2 pt)** Still looking at the `scores` distribution (i.e. the distribution of average survey scores), if we take only the average scores that are 1 SD of sample means away from the true average, roughly how much of the data (average survey scores) would we be looking at?

- ○ 10%
- ● 68%
- ○ 75%
- ○ 95%

6. (16 points) Knowledge versus Participation

Berkeley is an interdisciplinary place! A rhetoric professor has been collecting data on students' use of social media to discuss current events, and decides to combine her data with the political scientist's data in the previous question. The data are not anonymous, so the professors are able to link student's current events knowledge scores to a "participation factor" (a continuous value from 0 to 10) that assesses how often students engage with political causes online. The following is a scatter plot of the current events knowledge scores (from question 5) versus the participation factor.



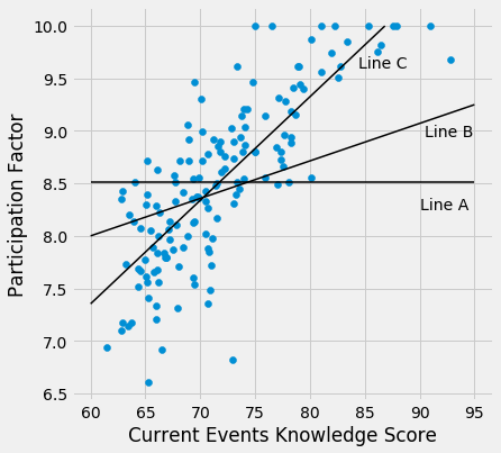
Here is some summary information about the data:

- The average current events knowledge score is 71.7, and the standard deviation of the current events knowledge score is 6.2.
- The average participation factor was 8.5, and the standard deviation of the participation factor is 0.78.
- The correlation coefficient between current events knowledge and the participation factor across the 150 students is 0.8.

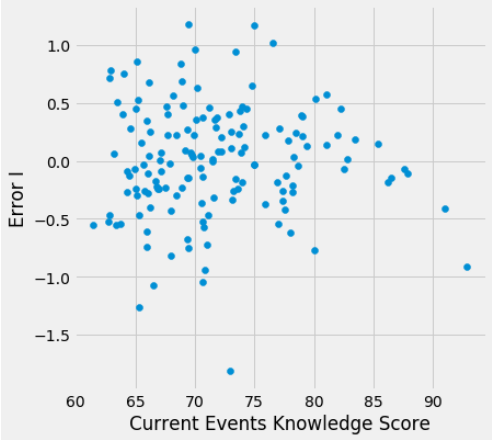
(a) (2 pt) What is the slope of the least squares regression line predicting participation factor from current events knowledge score? Do not simplify your answer (show your work!).
$0.8 \times \frac{0.78}{6.2}$

(b) (2 pt) What is the intercept of the least squares regression line predicting participation factor from current events knowledge score? Do not simplify your answer (show your work!).
$8.5 - (0.8 \times \frac{0.78}{6.2}) \times 71.7$

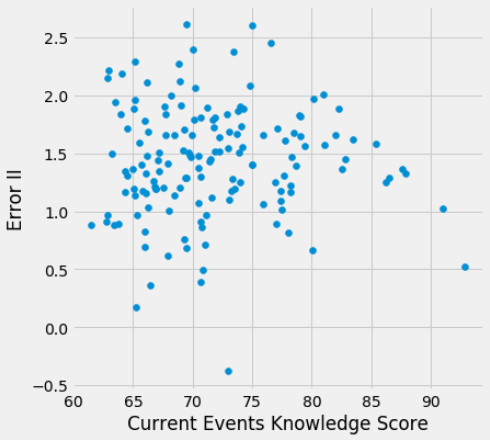
(c) (6 pt) The rhetoric professor plotted three lines on the scatter plot. They then plotted the errors between these lines and the data. **Match the three lines (A, B, C) to their corresponding error plots.** One error plot does not have a matching line.



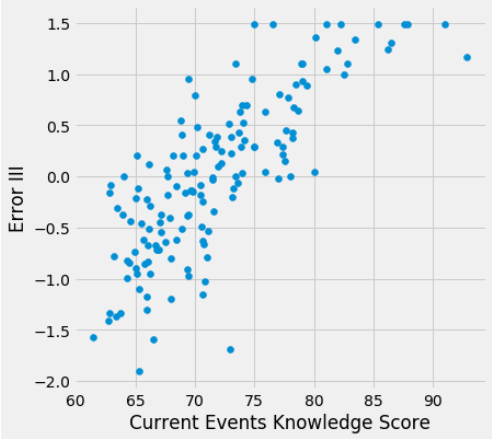
Pick one of the bubbles for each of the following error plots.



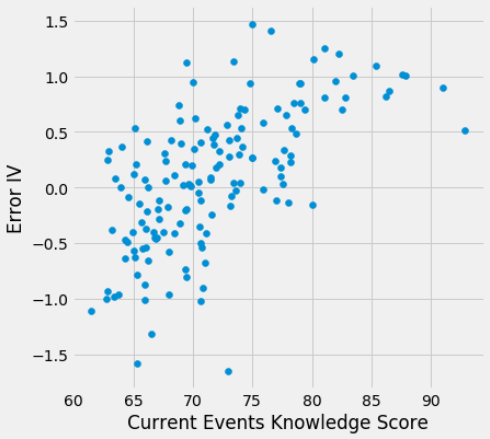
◯Line A ◯Line B
●Line C ◯No match



◯Line A ◯Line B
◯Line C ●No match



●Line A ◯Line B
◯Line C ◯No match



◯Line A ●Line B
◯Line C ◯No match

(d) (2 pt) One of the three lines in the plot in part **c** is the least squares regression line, and one line is the line *predicted y = average(y)*. Which one is which?

Least Squares Regression Line

○ Line A ○ Line B ● Line C

predicted y = average(y)

● Line A ○ Line B ○ Line C

(e) (1 pt) Rank the mean squared error of the three lines from smallest to largest:

○ A, B, C ○ B, C, A ○ A, C, B

○ B, A, C ○ C, A, B ● C, B, A

(f) (3 pt) Select **all** true statements we can justifiably conclude. If you do not have enough information to evaluate whether the statement is true or false, DO NOT select it.

■ The root mean squared error (RMSE) of line A is equal to the standard deviation of the participation score.

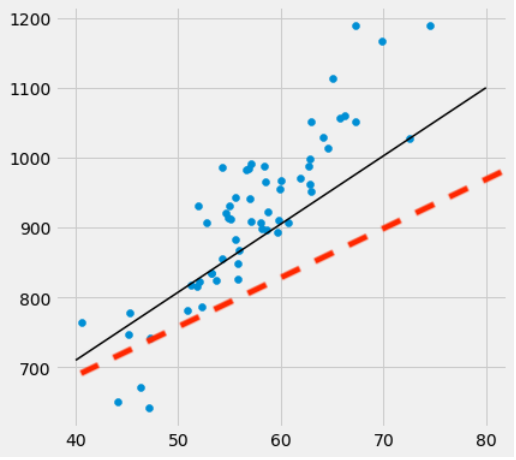
☐ The root mean squared error of line B is equal to the standard deviation of the fitted values from line B.

■ The root mean squared error of line C is equal to the standard deviation of the residuals.

■ It is impossible to create a linear model that will give you a smaller root mean squared error than the root mean squared error from line C.

7. (12 points) Some More Regression

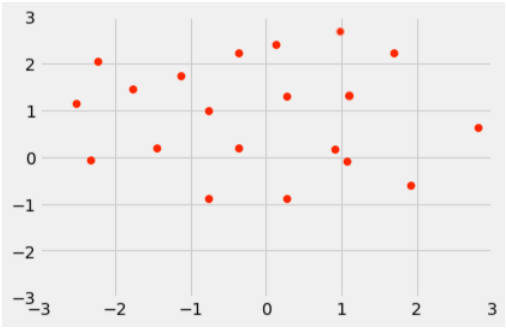
(a) (3 pt) The following depicts a scatter plot between two numerical variables (labeled x and y). A line (labeled "Line A") has been drawn through the scatter plot, but note that Line A is not the least squares regression line between x and y. **On the plot itself (do not make a separate plot), draw a line with a *larger* root mean square error than the existing line.**



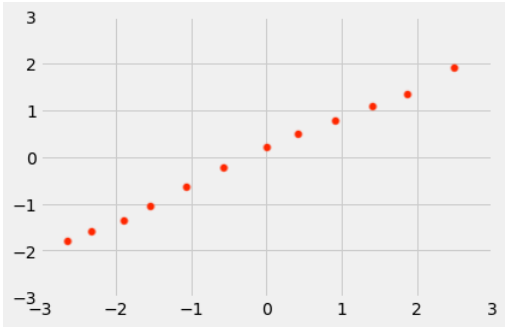
The line at left is one example of a valid answer.

(b) (6 pt) We have provided three empty plots, with axes in standard units. Draw a dataset of ten to twenty points with the specified value for r. Make sure your points are clearly visible and large enough to see when scanned.

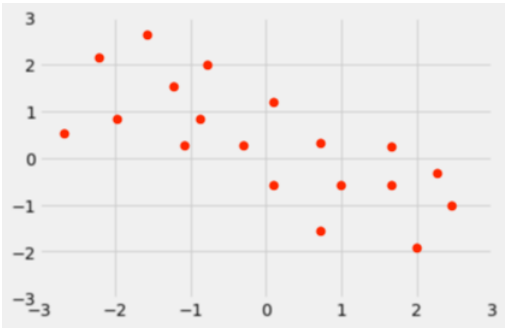
Plot A: Draw data points in a scatter plot such that the correlation of the points could be close to **0.**
Plot B: Draw data points in a scatter plot such that the correlation of the points could be close to **0.99.**
Plot C: Draw data points in a scatter plot such that the correlation of the points could be close to **-0.5.**



Plot A



Plot B



Plot C

The points above are examples of valid solutions.

(c) (3 pt) A researcher is interested in the relationship between the number of bumblebees and the number of ground squirrels observed in an alpine meadow over the course of one day. They calculate the regression line between the number of bees and the number of squirrels, and finds that the slope of the regression is 0.0314. Curious about the small value of the slope, the researcher wants to test the following hypotheses:

- Null: The slope of the true line describing the relationship between number of bees and number of squirrels in alpine meadows is zero.
- Alternative: The slope of the true line describing the relationship between between number of bees and number of squirrels in alpine meadows is not zero.

The researcher finds that an approximate 90% confidence interval for the true slope is [0.013, 0.045]. Select **all** the p-value cutoffs for which the researcher will reject the null hypothesis:

- [x] 20%
- [x] 10%
- [ ] 5%
- [ ] 1%

8. (24 points) Rock Solid Confidence

A petrologist (someone who studies rocks) has collected a random sample of 500 sandstone rocks from a desert region. They are interested in knowing the mean density of sandstone rocks in that region. They find that the mean density of the sample is 2.4 grams/cm^3. Instead of giving just one estimate, they wish to provide a range of values.

(a) (6 pt) Select one of the options from each parts i-vi to fill in the corresponding blanks in the sentence below.

The actual mean density of the sandstone rocks in the region is the population ___(i)___ . Rather than going back to the original population and taking a new sample, the scientist will use the sample they already have. This technique is known as bootstrapping. To use the bootstrap method, the scientist will take many samples ___(ii)___ from the ___(iii)___ to create one bootstrapped sample. We find the ___(iv)___ of each bootstrap sample. After we've completed this process, we can compute a ___(v)___. We are more confident that our procedure will generate an interval that captures the actual mean density when we use a ___(vi)___ interval.

i. ● parameter ○ statistic
ii. ○ without replacement ● with replacement
iii. ○ population ● original sample
iv. ○ middle 95% ● mean
v. ● confidence interval ○ p-value
vi. ● wider ○ narrower

(b) (4 pt) Select all of the following conditions under which bootstrapping would not be an effective estimation technique.

☐ The original sample is very big.
■ You are trying to estimate the minimum value of a population.
■ The original sample is very small.
☐ The original sample is a random sample from the population.
☐ You are trying to estimate the median value of a population.
☐ The distribution of your population is not roughly bell shaped.

(c) (6 pt) The table called `rocks` has one column, `density`, containing the density of the 500 sampled rocks in the region. Write code such that `left_end` and `right_end` evaluate to the endpoints of a ninety percent **(90%)** confidence interval for the mean density of rocks in the region using 10,000 bootstrapped resamples.

```
means = make_array()

for i in np.arange(10000):

        resample = rocks.sample()

        resample_mean = np.mean(resample.column("density"))

s

        means = np.append(means, resample_mean)

left_end = percentile(5, means)

right_end = percentile(95, means)
```

(d) (4 pt) Select **all** answers that we can justifiably conclude. If you do not have enough information to evaluate whether the statement is true or false, DO NOT select it.

☐ If the petrologist convinces 100 of their colleagues to independently take new random samples of 500 rocks, and each colleague generates one approximate 80% confidence interval for the true mean rock density in the region, about 95 of the 100 intervals will contain the true mean rock density of the region.

■ If the petrologist convinces 100 of their colleagues to independently take new random samples of 500 rocks, and each colleague generates one approximate 90% confidence interval for the true mean rock density in the region, about 90 of the 100 intervals will contain the true mean rock density of the region.

■ If the petrologist's colleague runs the code from part (c), but changes the endpoints of the interval (in the last two lines) to the 0.5th and 99.5th percentile, the resulting interval will be wider than the petrologist's original interval.

☐ If the interval calculated in part c) is [1.9g/cm^3, 2.8g/cm^3], there is a 90% chance that the true mean rock density of the region is in that interval.

(e) (4 pt) If the width of a 90% confidence interval you calculated was 1 gram/cm^3, **what is an estimate of the standard deviation of the rock densities in the population** of rocks from which we drew our sample of size 500?

The table below shows percentages of values in a certain range under the normal curve, in addition to those already in the exam reference guide.

Percent in Range	**Normal Distribution**
average ± 1.3 SDs	about 80%
average ± 1.65 SDs	about 90%

Show your calculations below. You should NOT simplify arithmetic expressions. Please draw a box around your final answer.

$1 = 2 \times 1.65 \times$ SD of sample means

SD of sample means $= \frac{1}{3.3}$

$\frac{1}{3.3} = \frac{\text{population SD}}{\sqrt{500}}$

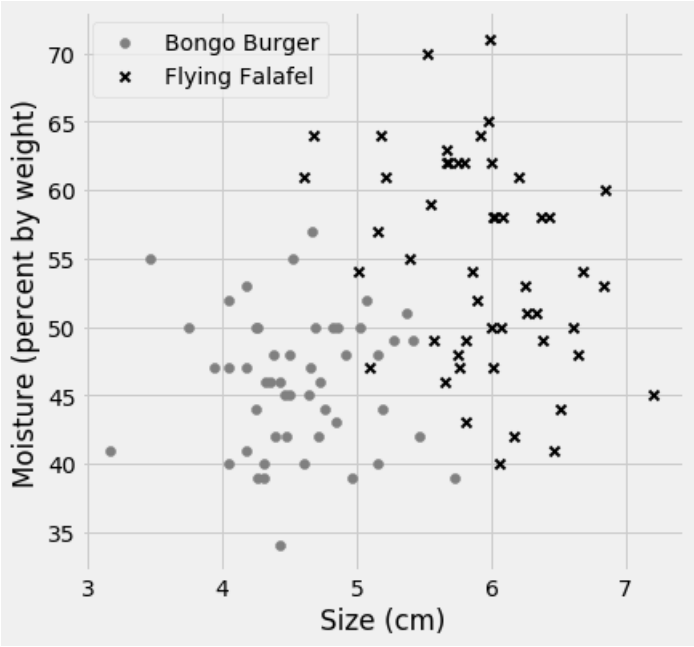
population SD $= \frac{\sqrt{500}}{3.3}$

9. (18 points) Find My Falafel

Francie and Natalia both love falafel (a delicious fried chickpea snack), but they often argue about the best falafel in Berkeley. Natalia likes the falafel from Bongo Burger (BB) and Flying Falafel (FF) equally, but Francie says the BB falafel are too small and dry compared to the good stuff at FF. The Data 8 summer staff are fed up with Francie and Natalia arguing about falafel during staff meetings, so they decide to collect falafel from each location and train a classifier to determine where future falafel come from.
The staff purchase 100 falafels, 50 from each of BB and FF, and measure

- `size` : diameter of falafel, in millimeters.
- `moisture` : the moisture content of each piece, reported as a percentage (by weight)

The following is a plot of the 100 falafel and their two characteristics. Circles are Bongo Burger falafel and crosses are Flying Falafel falafel.



(a) (3 pt) The two classes of falafel cannot be perfectly separated by a line on the plot above. Is it still appropriate to use k-nearest neighbors classification to classify the falafel?

● Yes, because k-nearest neighbors still works even if the data cannot be perfectly separated.

○ Yes, because having individuals from different classes that overlap in their features allows you to use a higher value for k.

○ Yes, because k-nearest neighbors transforms the data so that it can be separated by a decision boundary.

○ No, because k-nearest neighbors does not work if the data cannot be perfectly separated.

○ No, because the accuracy on the test set is guaranteed to be lower if the data cannot separated by a decision boundary.

Parts B and C refer to the information below.
Natalia found another falafel on the ground. She wants to use the dataset we currently have to classify which restaurant the new falafel came from. The following is an **unsorted** table of the seven pieces of falafel in our dataset closest (in terms of Euclidean distance between size and moisture content) to the new piece of falafel:

Restaurant	Size (cm)	Moisture Content	Distance to New Piece
FF	5.57242	48.7916	0.810133
FF	5.80844	48.6107	0.734703
FF	5.75313	48.2827	0.452357
BB	4.37966	47.7014	1.06312
BB	4.50055	47.8064	0.920049
BB	4.91239	48.2195	0.53473
BB	5.14879	48.3998	0.472151

(b) **(2 pt)** If we used a 3 nearest neighbor classifier, would we classify the new piece of falafel as coming from Bongo Burger or Flying Falafel? ●Bongo Burger ○Flying Falafel

(c) **(2 pt)** If we used a 5 nearest neighbor classifier, would we classify the new piece of falafel as coming from Bongo Burger or Flying Falafel? ○Bongo Burger ●Flying Falafel

(d) **(3 pt)** Shoumik is a visual learner and wrote out each step the classifier needs to perform on a different slip of paper and placed them on a desk. Before implementing the classifier Shoumik takes a break to teach his lab section. While he was gone, Shoumik's roommate rearranges the slips of paper while cleaning the desk, and now all the steps are out of order. Label the following steps as 1 ("do this first") to 6 ("do this last") in order to correctly implement a k nearest neighbors classifier.

4 Take the top k rows of the sorted table

6 Calculate the classifier accuracy on all points in your test set

5 Classify the new point as the majority class in top k rows

1 Split the original data set into training and testing set

3 Sort the distances from smallest to largest

2 Calculate the distance between the new point and all points in the training set.

(e) **(2 pt)** Based on the scatter plot, which of the two features is more useful for differentiating falafel between the restaurants (e.g. if you could only use one feature to classify, which should you choose)?

○ moisture ● size

(f) **(3 pt)** The staff notice that falafel made in the morning are a different size than falafel made in the evening, as if restaurants are trying to conserve falafel batter at the end of the day. If you used a classifier trained on morning falafel to predict the label of an evening falafel, how might the accuracy of your classifier be affected?

○ The test accuracy of your classifier would not change - you would do just as well using a classifier trained on morning falafel as a classifier trained on evening falafel.

○ The test accuracy would increase because you are including more data.

○ The test accuracy would increase because your accuracy would not depend on random variation in the morning falafel.

● The test accuracy would decrease because the typical values of features for each restaurant might be different in the evening relative to the morning.

(g) (3 pt) The staff decided to investigate evening falafel and created a new classifier using falafel collected in the evening. However, on the night of data collection Bongo Burger had a machine malfunction so the staff were only able to collect 10 BB falafel, and 90 FF falafel. A GSI tried many different values for k (how many nearest neighbors to look at) when training their classifier, and noticed that above some value of k, the classifier always classified any new falafel as coming from FF, regardless of its features. What is this value of k?

- ○ 11
- ○ 20
- ● 21
- ○ 51

10. (0 points) Data art (optional) Draw a visualization or graph 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!