Data 8 Foundations of Data Science
Summer 2022

FINAL EXAM

INSTRUCTIONS

- You must **write your name in the space provided on one side of every page of the paper exam.**
- The exam is worth 100 points. You have 170 minutes to complete it.
- An official final exam reference is provided. You may not use any other paper, reference, source, or computational device or system apart from those permitted for the online exam.
- 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.
- In starter code we provide, _____ can mean any code, including commas and periods. You do not need to simplify math expressions.
- **Important**: Please completely **fill in** circles and squares to indicate answers and cross out or erase mistakes.
  - **Valid** : ■ or ●
  - **Invalid** : ⃠, ☑, ☒ or ⊗

Last name	
First name	
Student ID number	
Calcentral email (`_@berkeley.edu`)	
Lab GSI	
Your room number and building (e.g. 155 Dwinelle)	
Your seat number (e.g. A1)	
← 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. (6 points) Python Potpourri

In a notebook, Jeffrey has run the following code:

```
first_array = make_array(1, 2, 3, 4)

second_array = make_array(5, 6, 7)

third_array = make_array('My', 'name', 'is', 'Jeffrey')
```

For each of the below lines of code, explain what it would evaluate to, or explain why it would error. You are free to represent Python data types in any reasonable way (e.g. both `[]` and `array([...])` are valid representations of an array), as long as it is clear what you mean.

(a) (1 pt) `first_array * 2`

(b) (1 pt) `2 ** first_array`

(c) (1 pt) `first_array + second_array`

(d) (1 pt) `'hi' * 3`

(e) (1 pt) `make_array(1, 2, 3/4)`

(f) (1 pt) `int('2 + 3')`

2. (12 points) Snow Days

Ellen is visiting Sweden this winter, and wants to know how much snow to expect. She surveys 1,000 Swedes from the city Stockholm, selected at random, and asks them for the snow depth they saw on December 1st, in millimeters (mm).

(a) (2 pt) Ellen decides to create a confidence interval for the mean snow depth on December 1st using this data. Which of the following are reasons she would want to do this? **SELECT ALL THAT APPLY.**

- ☐ To quantify the uncertainty in her estimate
- ☐ To understand how the estimate might have been different, had the sample been different
- ☐ To understand how the estimate varies over time
- ☐ To estimate the true mean snow depth on December 1st, in Stockholm

(b) (2 pt) Using her sample, Ellen computed a 95% confidence interval of (-5.4mm, 21.2mm). This interval is too wide, so she wants to redo the entire process to create a narrower interval.

Which of the following modifications to the process would result in a narrower interval? **SELECT ALL THAT APPLY.**

- ☐ Smaller sample size, same level of confidence
- ☐ Larger sample size, same level of confidence
- ☐ Same sample size, lower confidence level
- ☐ Same sample size, higher confidence level

(c) (3 pt) Instead of making a confidence interval for the mean snow depth, Oscar suggests that Ellen instead makes a confidence interval for the proportion of days in December that have any snow at all. Ellen agrees, but doesn't want her confidence interval to have a width of more than 0.02.

What is the smallest sample size such that a 95% confidence interval for the **true proportion of snow days in December** has width of at most 0.02? Show your calculations; but you don't need to simplify your answers.

(d) (3 pt) Ellen decides that's too large of a sample size, and wants to make a 68% confidence interval instead.

What is the smallest sample size such that a 68% confidence interval for the **true proportion of snow days in December** has width of at most 0.02? Show your calculations; but you don't need to simplify your answers.

(e) (2 pt) For which of the following parameters would it be reasonable to use the bootstrap method to estimate the true value? **SELECT ALL THAT APPLY.**

- ☐ Population Mean
- ☐ Population Maximum
- ☐ Population Median
- ☐ 1st percentile of the population

3. (15 points) Wishing for Rain

Tara travels to work by car or bus, and their decision depends on the weather. If it's sunny outside, they're more likely to travel by bus. If it's raining, they're more likely to travel by car. For parts (a)-(d), you may assume that on any day, there is a 40% chance of rain, and a 60% chance of sun, independently of all other days.

Furthermore, Tara chooses their transportation randomly, according to the weather and the following probabilities:

- If it is rainy, there is a 70% chance that Tara takes their car, and a 30% chance that they take the bus, independently of other days.
- If it is sunny, there is a 20% chance that Tara takes their car, and a 80% chance that they take the bus, independently of other days.

For each question, leave your answer as an unsimplified expression, or write "Need more information" if the probability can't be calculated.

(a) (1 pt) What is the probability that it will rain on a randomly selected day?

(b) (1 pt) On another randomly selected day, it was rainy. Given this information, what is the probability that Tara took the bus on that day?

(c) (2 pt) On a randomly chosen day, what is the probability that Tara took the bus?

(d) (2 pt) We pick a day of the year at random and find that Tara took their car to work on that day. Given this information, what is the probability that it was raining on that day?

(e) (3 pt) In the real world, the probability of it raining on a given day is not fixed, but it depends on the season (e.g. it is more likely to rain in the spring). However, over the whole year, we still know that approximately 40% of days are rainy and approximately 60% are sunny.

If Tara took their car to work on a randomly chosen day in the spring, what is the probability that it was raining on that day?

(f) (2 pt) Suppose Tara moves to London where it rains every day. Their commuting preferences remain the same. What is the probability that Tara takes the bus at least once over their 5-day workweek?

(g) (4 pt) Tara wants to write a function to help them decide which type of transportation they should use, given the weather. The function `choose_transport` should return either the string `'car'`, or the string `'bus'`. The returned vehicle should be chosen at random, according to Tara's preferences from the table on the previous page. You may assume that the argument `weather` is either the string `'sunny'` or the string `'rainy'`.

Fill in the blanks in the below code so that the function works as described.

Hint: `np.random.choice` takes in an optional argument `p`, an array that specifies the probability of selecting the corresponding item. If no argument is supplied to `p`, then each item is equally likely.

```
def choose_transport(weather):

    choices = make_array('car', 'bus')

    if weather == 'sunny':

        probs = _____

        return _____

    else:

        probs = _____

        return _____
```

4. (15 points) Chess

Several Data 8 staff members have recently started playing chess. Being data scientists, they want to study previous games to figure out what strategies will help them win more. They download the Table `chess`, containing 11,250 games played online in 2022, shown below:

Turns	Outcome	Winner	White Rating	Black Rating	Opening	First Move
13	outoftime	white	1500	1191	Scandinavian Defense	e4
16	resign	black	1322	1261	Queen's Pawn Game	d4
61	mate	black	1496	1500	Sicilian Defense	e4
61	mate	white	1439	1454	Scotch Game	e4
95	mate	black	1469	1523	Van't Kruijs Opening	e3

`...(11, 245 rows omitted)`

(a) (2 pt) Write a line of code that returns a two-column table, displaying the average number of turns for each potential outcome. The first column should contain the outcome, and the second column its corresponding average number of turns. The names of the columns don't matter.

(b) (1 pt) Which of the following visualizations would be most appropriate for studying the *counts* of each type of outcome? **SELECT ONE.**

- ◯ Histogram
- ◯ Bar Chart
- ◯ Pie Chart
- ◯ Line Plot

(c) (2 pt) In chess, the white pieces always move first, and many people wonder if there's an advantage to going first. In our dataset, what proportion of games were won by the white player? Write a line of code that returns a float corresponding to this value.

(d) (2 pt) Jonathan's two favorite openings are the "Sicilian Defense" and the "Scotch Game", and they want to know how many times those two were used by other players. Write a line of code to count how many times the opening used was the "Sicilian Defense" or the "Scotch Game" in our dataset. You may continue your code onto the second line if you run out of room, but do not define any new variables.

(e) **(6 pt)** Define an "upset" as a game where the lower-rated player won. Fill in the blanks in the code below to so that the last line evaluates to the number of games in our Table that were upsets.

Hint: The function `lower_rated` should take in a Row object from the `chess` table, and output a String corresponding to the lower rated player in that row, or `'tie'` if both players have the same rating.

```
def lower_rated(game_row):
    white_rating = _____
    black_rating = _____
    if _____:
        return 'white'
    elif _____:
        return 'black'
    else:
        return 'tie'
ratings_array = _____
sum(_____)
```

(f) **(2 pt)** How many unique openings were used by chess players in our dataset? Write a line of code that returns an array of all the unique openings, in alphabetical order.

_____

5. (14 points) Stranger Songs

Max, Lucas, and Dustin are very interested in music, and they've collected a random sample of 250 Pop and Rock songs from their library's music collection. The data is stored in a table called `songs`, of which the first few rows are displayed below:

Song Name	Genre	Length
Running Up That Hill	Pop	298
California Dreamin'	Rock	162
American Pie	Rock	251

`...(247 rows omitted)`

Lucas thinks that both genres have long and short songs, and the distribution of length is the same between the two genres. Dustin disagrees, and thinks that the song lengths are different, on average. They decide to use a hypothesis test to figure out which of them is correct.

(a) (4 pt) In the _____, the distribution of song lengths is _____ for Rock songs and Pop songs..

Fill in the blanks above to complete the null hypothesis by selecting from the following options.

(i) Blank 1 (make **exactly one** choice):

◯ population ◯ sample ◯ underlying distribution ◯ true distribution

(ii) Blank 2 (make **exactly one** choice):

◯ different ◯ the same ◯ on average longer ◯ on average shorter

(b) (2.5 pt) Which of the following could be an alternative hypothesis to decide between Dustin and Lucas's opinions? **SELECT ALL THAT APPLY.**

☐ The song genres are chosen at random, with a 50% chance of being Rock and a 50% chance of being Pop, independently of song length.

☐ Song lengths are not evenly distributed within each genre.

☐ The distribution of song lengths is different between the two genres.

☐ Rock songs are longer than Pop songs, on average.

☐ The data the kids collected does not support the claim that the genres have the same length, on average.

(c) (2.5 pt) The three friends agree to use absolute difference in mean song length as their test statistic. As the next step to testing his friends' claims, Max wants to simulate 10,000 values of this test statistic under the null hypothesis.

Which of the following expressions will simulate a new sample of size 250 to use for the test? **SELECT ALL THAT APPLY.**

☐ `np.random.choice(['Rock', 'Pop'], 250)`

☐ `np.random.choice(songs.column('Length'), 250)`

☐ `sample_proportions(250, [0.5, 0.5])`

☐ `songs.with_column('Genre', songs.sample(with_replacement=False).column('Genre'))`

☐ `songs.with_column('Length', songs.sample(with_replacement=False).column('Length'))`

(d) (2.5 pt) Max then runs a simulation and calculates the p-value for the experiment described above. Which of the following is true about the result of his experiment? **SELECT ALL THAT APPLY.**

☐ The p-value cutoff is the proportion of simulated test statistics that are greater than or equal to the observed statistic.

☐ If he redid the experiment with an increased sample size, and the null hypothesis was actually true, the chance of rejecting the null would increase.

☐ A p-value close to 1 means that the the data are consistent with the null hypothesis.

☐ A p-value close to 0.5 means that the data are consistent with the null hypothesis.

☐ A p-value close to 0 means that the data are consistent with the null hypothesis.

(e) (2.5 pt) At a p-value cutoff of 5%, Max's results show that the data are more consistent with the alternative, and concludes that Dustin was right all along. Lucas objects to this approach, and asks him to redo the test using a 95% confidence interval for the difference in mean song length in the entire music collection instead. Is this a valid complaint? **SELECT ONE.**

○ Yes, the choice of testing method might influence the result of the test.

○ Yes, hypothesis tests are more likely to find the data consistent with the null.

○ No, since confidence intervals can only be used when the underlying distribution is normal.

○ No, since the two methods will result in the same conclusion.

6. (10 points) Experiments

(a) (2 pt) Rebecca wants to know the average commute time of Data C8 students, but she only has time to study 100 people. Which of the following would be valid ways to obtain a random sample? **SELECT ALL THAT APPLY.**

- ☐ Sample the first 100 people who attend the last day of lecture
- ☐ Shuffle the full student roster, and sample the first 100 students in the shuffled roster.
- ☐ Ask all students for their commute time, and only use the first 100 responses.
- ☐ Line up the students by height, and ask every other student until you get 100 responses.

(b) (2 pt) Atticus is trying to figure out if the height of campus buildings can be used to predict the number of students who enter it daily, and notices a linear relationship between the two. Which of the following values would never change if Atticus changed the units that he measured the height in (e.g. from feet to meters)? **SELECT ALL THAT APPLY.**

- ☐ Slope of the best fit line for predicting number of students from height
- ☐ Slope of the best fit line for the variables in standard units
- ☐ Correlation Coefficient between number of students and height.
- ☐ RMSE of the best fit line for predicting number of students from height

(c) (2 pt) Raymond takes a random sample of 250 UC Berkeley students and records whether they prefer Sliver pizza (represented as 0) or Cheeseboard pizza (represented as 1), storing the data in an array called `pizza`. Which of the following lines of code will evaluate to the proportion of students who prefer Cheeseboard? **SELECT ALL THAT APPLY.**

- ☐ `np.mean(pizza)`
- ☐ `np.sum(pizza) / len(pizza)`
- ☐ `np.count_nonzero(pizza == 1) / len(pizza)`
- ☐ `1 - np.count_nonzero(pizza == 0) / len(pizza)`

(d) (2 pt) UC Berkeley has 32 different libraries, and Padma wants to pick one at random using her array of library names, `librs`. Which of the following lines of code will output a String corresponding to the name of one randomly selected library? **SELECT ALL THAT APPLY.**

- ☐ `np.random.choice(librs, 5).item(0)`
- ☐ `Table().with_column('Name', librs).sample(1, with_replacement=False).column(0).item(0)`
- ☐ `Table().with_column('Name', librs).sample().column(0).item(0)`
- ☐ `sample_proportions(1, librs).item(0)`

(e) (2 pt) Sara wants to use a k-NN classifier to find the dining hall that will get her food as fast as possible. She is choosing between Crossroads or Foothill, with the hour of day and her current location on campus (latitude, longitude) as her three attributes. The k-NN classifier should return one of `'Crossroads'` or `'Foothill'`. Is this a valid use of this type of classifier? **SELECT ONE.**

- ○ Yes, this is a valid use of a k-NN classifier.
- ○ No, since hour of day and location are measured in different units.
- ○ No, since the time to get food is numerical.
- ○ No, because she has more than 2 features.

7. (17 points) Fishy Business

A local fisher has recorded the price (in US$) for each fish he sold over the past year, as well as its length (in inches). The data is stored in the table `fish`, shown below:

Fish Name	Length	Price
Albacore Tuna	39	15
Sea Bass	12.5	7.35
Rainbow Trout	18	9.10

`...(1497 rows omitted)`

You may assume that these fish were drawn randomly from all the possible fish in his area. Furthermore, we have the following information about the entire fish population:

(a) In the sample, fish length is normally distributed with a mean of 25 and a standard deviation of 5.

(b) In the sample, fish price has a mean of 40 and a standard deviation of 10.

(c) The correlation between length and price, in the sample, is 0.75.

(a) (2 pt) Which of the following statements are correct? **SELECT ALL THAT APPLY.**

- ☐ The distribution of Price in the sample is normal because of the Central Limit Theorem.
- ☐ 95% of the fish lengths in the sample are contained within 2 SDs of the mean fish length.
- ☐ At most 25% of the lengths in the sample are lower than 15 or higher than 35.
- ☐ At most 25% of the prices in the sample are lower than 20 or higher than 60.

(b) (1 pt) You are interested in predicting the Price based on the Length of the fish. Calculate the slope for the regression line in standard units. You may leave your answer as a mathematical expression.

(c) (1 pt) What is the slope of the regression line for predicting Price from Length, both in original units? You may leave your answer as a mathematical expression.

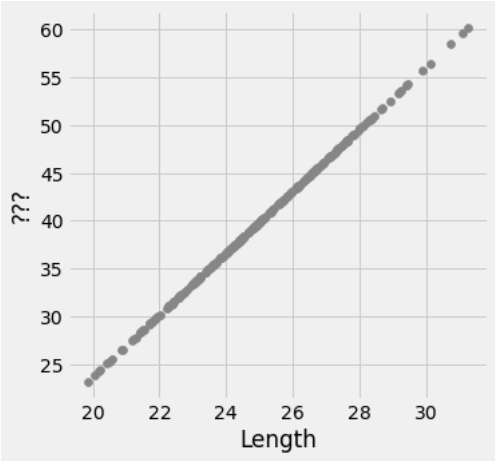
(d) (1 pt) What are the units of the slope of the regression line calculated in part (c)? **SELECT ONE.**

- ○ dollars per inch
- ○ inches per dollar
- ○ inches
- ○ dollars
- ○ unitless

(e) (4 pt) Suppose you perform the regression described in the previous parts, and create a few scatter plots to investigate its fit.

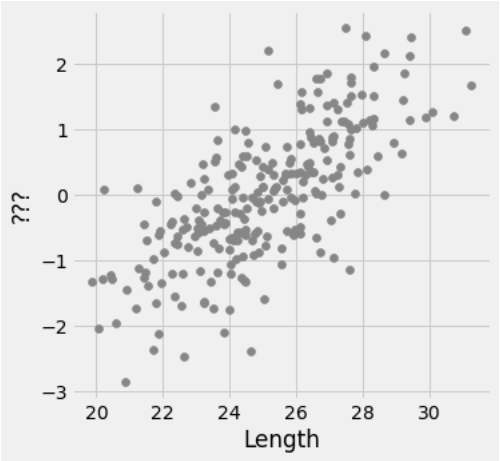
Each of the below plots represents a scatter plot one of Price (in standard units), Price (US $), Predicted Price (US $), and Residual (US $) versus Length. For each one, indicate which of the variables below is the label of the y-axis. Each variable should be used exactly once.

(a) (1 pt)



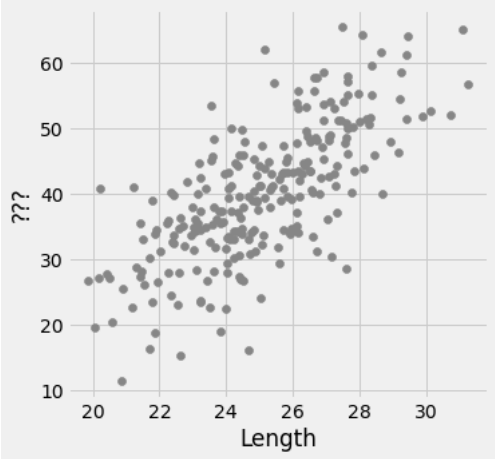
- ◯ Price (standard units)
- ◯ Price (US $)
- ◯ Predicted Price (US $)
- ◯ Residual (US $)

(b) (1 pt)



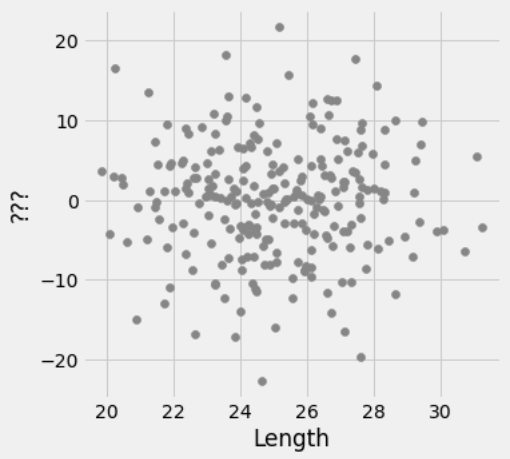
- ◯ Price (standard units)
- ◯ Price (US $)
- ◯ Predicted Price (US $)
- ◯ Residual (US $)

(c) (1 pt)



- ◯ Price (standard units)
- ◯ Price (US $)
- ◯ Predicted Price (US $)
- ◯ Residual (US $)

(d) (1 pt)



- ◯ Price (standard units)
- ◯ Price (US $)
- ◯ Predicted Price (US $)
- ◯ Residual (US $)

(f) (2 pt) Which of the following statements are always true for correctly performed linear regression? **SELECT ALL THAT APPLY.**

☐ The average of the residuals will be negative if our regression line overestimates the y-values.

☐ There is zero correlation between the residuals and the fitted values.

☐ There is zero correlation between the residuals and the actual y-values.

☐ If the observed correlation between the x-values and the y-values is zero, then the intercept of the line of best fit is zero.

☐ None of the above.

(g) (2 pt) Which of the following would be an indication that linear regression is **not** a good fit for a dataset? **SELECT ALL THAT APPLY.**

☐ The x-variable is not normally distributed.

☐ The true correlation is not significantly different from 0, as determined by using a hypothesis test.

☐ There is a clear pattern in the plot of the residuals.

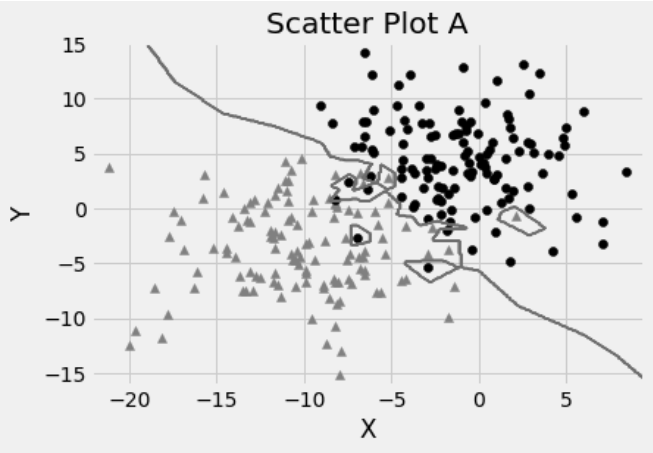
☐ The two variables are measured in units with very different scales.

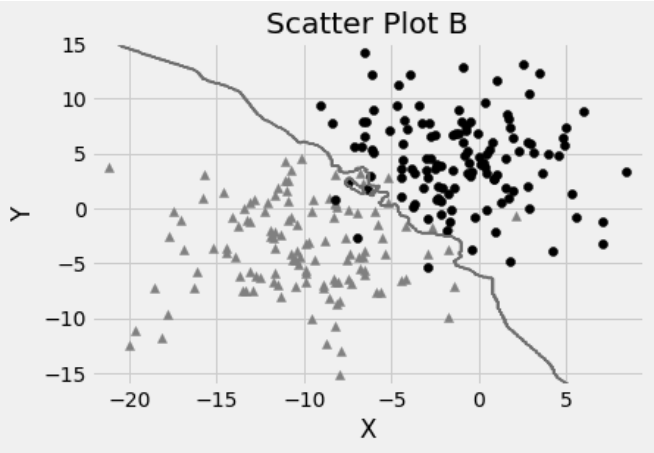
(h) (4 pt) Finally, let's write the code to perform least squares. As a reminder, we are predicting Price from Length, both in their original units. Using the `fish` table, fill in each of the blanks so that `parameters` evaluates to an array of the least-squares estimate of the slope and intercept for the best fit line.

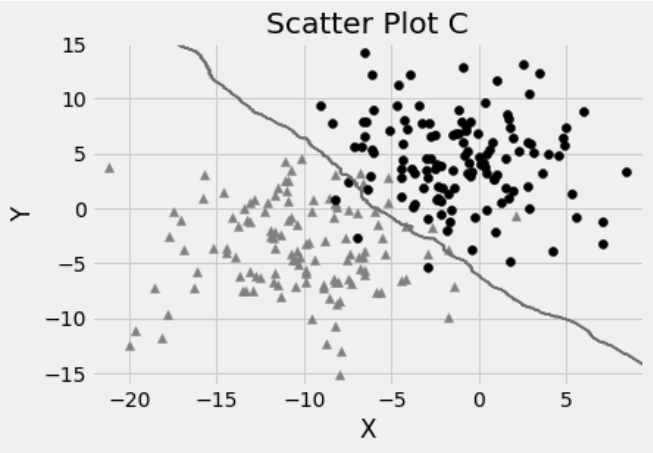
```
def rmse(_____):

    y_fitted = _____

    return _____

parameters = _____( _____)
```

8. (11 points) Boundary Disputes

(a) (3 pt) Kinsey creates three different k-NN classifiers to predict a label (0 or 1) based on two features X and Y. The scatter plots shown below show the training points and their labels, as well as decision boundaries for the models. One of her models used k=1, one of her models used k=7, and one of her models used k=50. Select which boundary corresponds to which model.

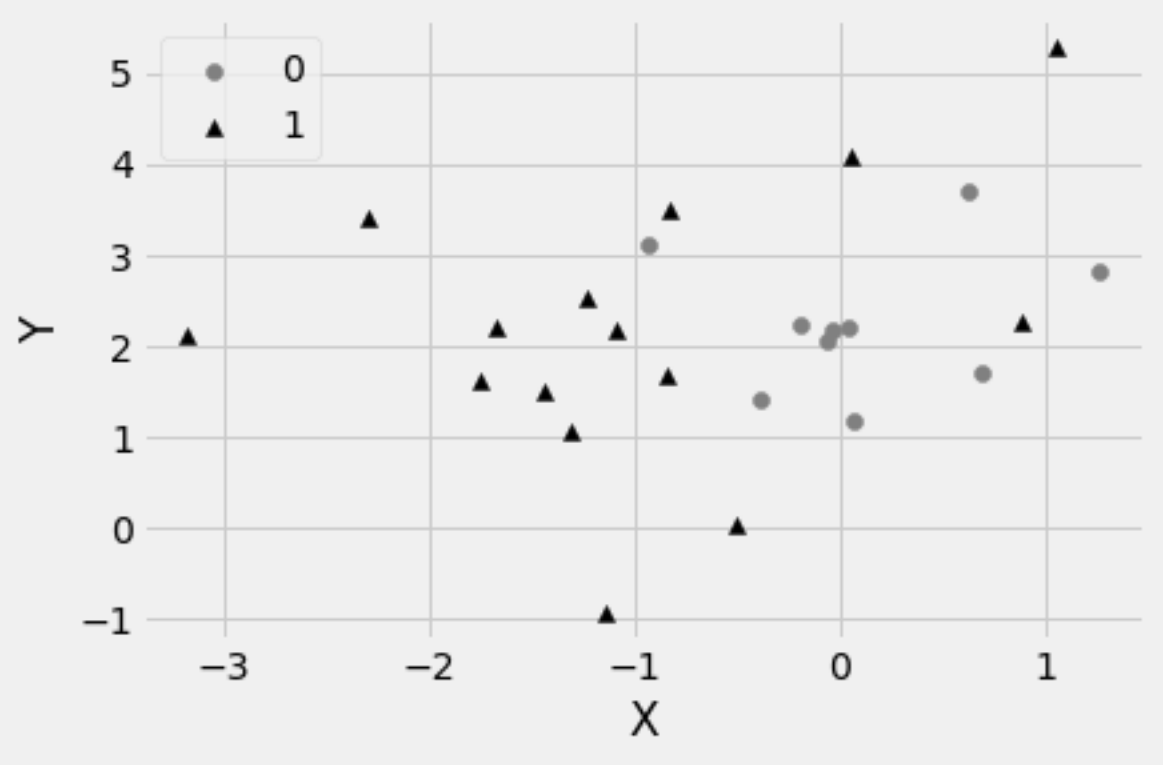






	Scatter Plot A	Scatter Plot B	Scatter Plot C
1-NN Classifier	◯	◯	◯
7-NN Classifier	◯	◯	◯
50-NN Classifier	◯	◯	◯

(b) (3 pt) In the below training set, there are 10 points with a label of 0, and 15 points with a label of 1. Draw the approximate decision boundary for a 21-neighbor classifier (that is, a k-NN classifier with k=21), or write "Impossible" below if you cannot draw a decision boundary for this classifier. **Explain your reasoning in the space below.**



(c) (3 pt) The last step in building a classifier is to compute the most common label in the nearest k points. Suppose we have a Table in the below format, with two columns "Distance" and "Label".

Distance	Label
0.455	0
2.224	0
1.431	1

Three students each wrote code to return the most frequent Label, but some of them made mistakes in their code. Select which functions **correctly** return the most common label in this table. You may assume that the input `tbl` always has an odd number of rows. **SELECT ALL THAT APPLY.**

☐
```
def most_common_1(tbl):
    label_counts = tbl.group('Label').sort('count')
    return label_counts.column('Label').item(0)
```

☐
```
def most_common_2(tbl):
    count0 = 0
    for row in tbl.rows:
        if row.item('Label') == 0:
            count0 = count0 + 1
    if count0 > tbl.num_rows / 2:
        return 0
    else:
        return 1
```

☐
```
def most_common_3(tbl):
    count1 = np.count_nonzero(tbl.column('Label'))
    if count1 > tbl.num_rows / 2:
        return 1
    else:
        return 0
```

(d) (2 pt) When building a k-NN classifier, increasing k will **always** result in which of the below? **SELECT ALL THAT APPLY.**

☐ A higher training accuracy
☐ A higher test accuracy
☐ A lower training accuracy
☐ A lower test accuracy
☐ None of the above

9. (0 points) Data Art (optional) Draw a picture (or graph) describing your experience in Data 8.

10. (0 points) Write your name in the space provided on one side of every page of the paper exam. You're done!