Data 100, Final

Summer 2021

Name: ______________________________

Email: ______________________________@berkeley.edu

Student ID: ______________________________

Exam Time: ______________________________

All work on this exam is my own ***(please sign)***: ______________________________

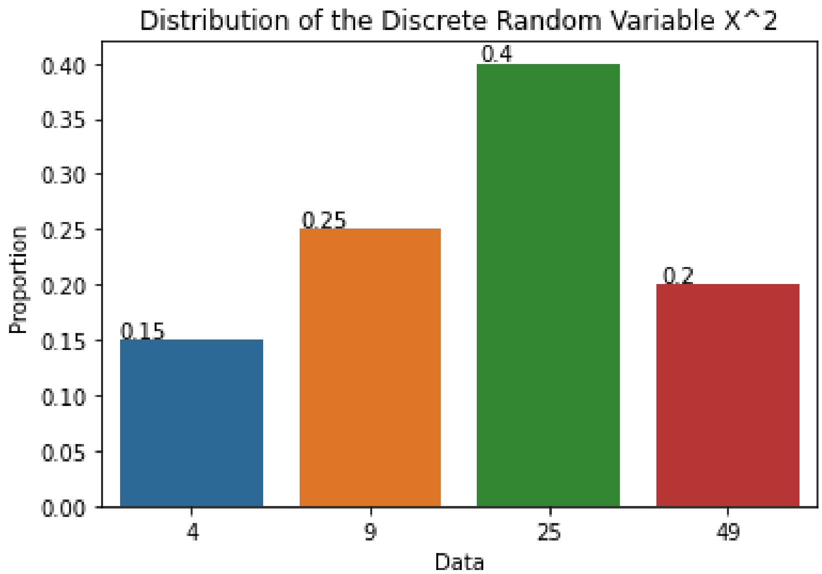
Honor Code [1 Pt]

1. *As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I will not communicate with any other individual during the exam, current student or otherwise. All work on this exam is my own.*

   (a) Please confirm your agreement with the above statement by writing your name in the space below.

Probability Potpourri [8 Pts]

2. Suppose X is a discrete, positively valued random variable. The following graph describes the probability distribution of X^2.



   (a) [2 Pts] What is the expected value of X? Round your answer to two decimal places.

   (b) [2 Pts] Following your answer to the previous question, what is the variance of X? Round your answer to two decimal places.

3. Oh no! Our friend Kanu has decided to take the Data 100 final without studying at all. He believes he can pass the course by simply guessing uniformly at random on every question. Assume Kanu needs a **10%** on the final to pass. The test consists of **20 MCQ** questions and **4 FRQ** questions. The grading scheme is as follows:

   - MCQ
     - 5 points are awarded for each correct answer.
     - $-\frac{1}{3}$ points are awarded for each incorrect answer.
     - 0 points are awarded for each blank answer.
   - FRQ
     - 10 points are awarded for each correct answer.
     - $-\frac{1}{3}$ points are awarded for each incorrect answer.
     - 0 points are awarded for each blank answer.

   There are 140 points available, so Kanu needs at least a 14 to pass.

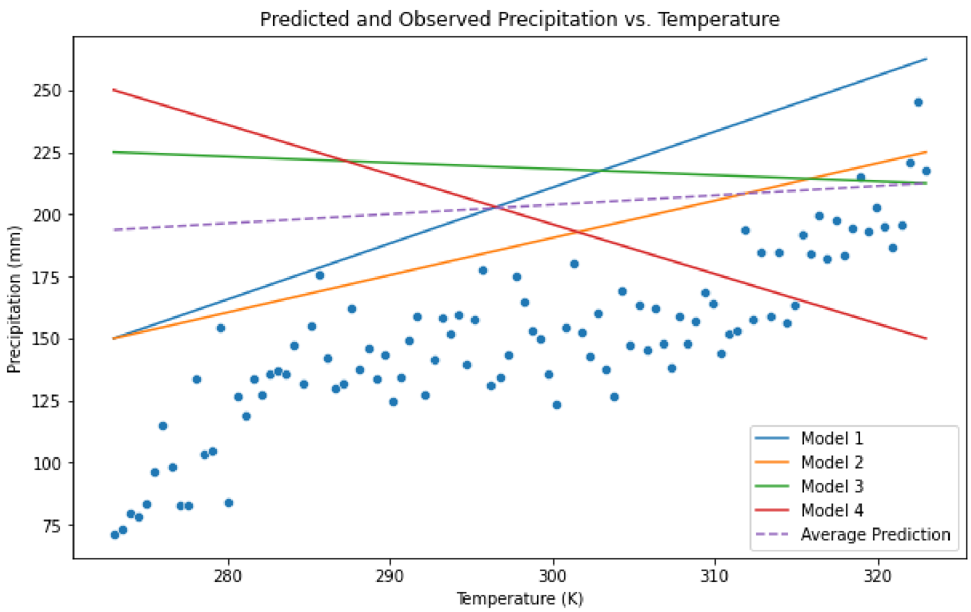
   (a) [4 Pts] Each MCQ question has **4** possible answers, one of which are correct. Each FRQ question has **10** possible answers, one of which is correct. On average, which of the following test taking strategies will help Kanu pass the class? Select all that apply.

- ☐ Guess randomly on all MCQ and FRQ.
- ☐ Guess randomly on all MCQ and leave the FRQ blank.
- ☐ Guess randomly on all FRQ and leave the MCQ blank.
- ☐ Guess randomly on all MCQ and $\frac{1}{2}$ of the FRQ. Leave the other $\frac{1}{2}$ of the FRQ blank.
- ☐ Guess randomly on $\frac{3}{4}$ of the MCQ and all the FRQ. Leave the other $\frac{1}{4}$ of the MCQ blank.

Extreme Tradeoffs [12 Pts]

4. Mr. Bean wants to model extreme precipitation events, which are historically difficult to predict accurately. To attempt to create the world's best model, he tries training multiple models using bootstrap sampling and regularization.

   (a) [2 Pts] Mr. Bean designs 4 different models by bootstrap sampling 30% of the total training data to train each model on. On the test set, he creates the following plot displaying the temperature feature against the model's predictions and true observed values. The dotted line shows the average prediction across all 4 models. Which of the following does the figure indicate? Select all that apply.



   ☐ High model variance
   ☐ Low model variance
   ☐ High model bias
   ☐ Low model bias

   (b) [4 Pts] Mr. Bean decides to diagnose the issue further. He increases the number of trained models to 100 and evaluates the models on the point (x_i, y_i). Using historical data, he assumes that measurement errors follow a normal distribution with mean 0 and standard deviation $\sigma = 4$ mm. Given the below statistics calculated using `pd.describe` on the predictions and loss for these models, estimate the magnitude of the empirical bias. Round to 3 decimal places.

   Hint: Think of how bias is calculated in our bias-variance decomposition and relate the quantities below to the terms in the decomposition.

This question is difficult, so if you are not sure how to start then skip it for now and come back to the question later.

	preds (mm)	MSE
count	100.000000	100.000000
mean	104.130417	101.930101
std	8.255889	112.418423
min	90.819579	0.023824
50%	103.941588	51.771039
max	119.474415	367.888568

In the box below, show how you obtained the value above. Specifically, write down the bias-variance decomposition, substituting in the relevant quantities. No $\LaTeX$ is required, you can use plain English.

(c) [2 Pts] He decides to change his models to add L2 regularization. What behavior is expected in the training set compared to the unregularized models?

- ☐ The model bias will decrease.
- ☐ The model bias will increase.
- ☐ The model variance will decrease.
- ☐ The model variance will increase.
- ☐ The observational variance will decrease.
- ☐ The observational variance will increase.

(d) [2 Pts] Regardless of your answer to the previous question, assume that after implementing regularization, the model bias is too high. Which of these solutions helps reduce the model bias?

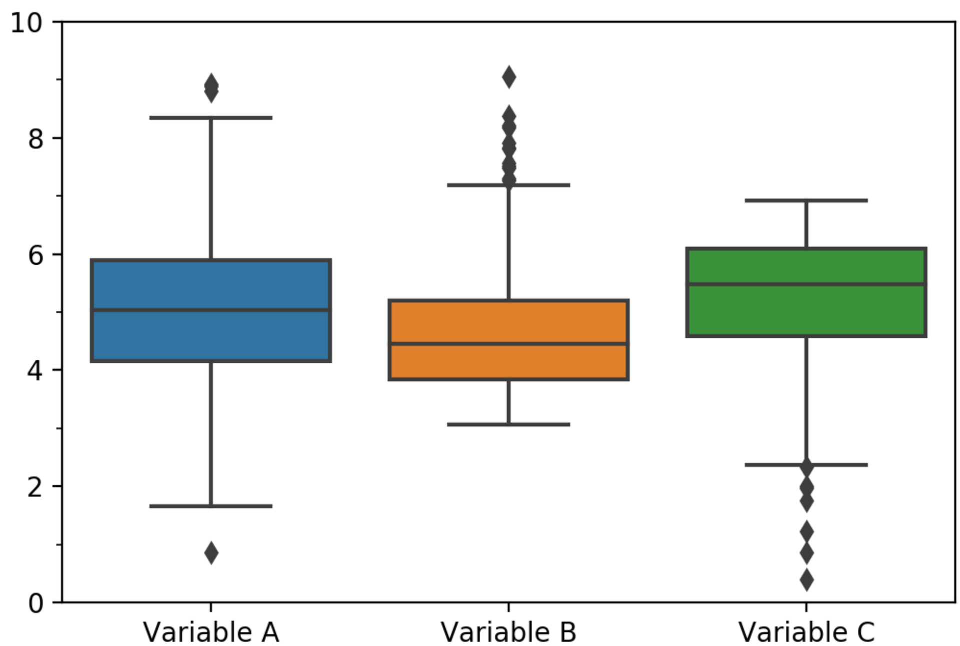
- ☐ Add an intercept term.
- ☐ Use a decision tree with the same features.
- ☐ Increase the regularization hyperparameter.
- ☐ Decrease the regularization hyperparameter.

(e) [2 Pts] Assume that we fixed the previous issue by changing to a *different* unspecified regression model, and the model bias decreased. Which of the following could have happened as a result?

- ☐ The model variance increased.
- ☐ The model variance decreased.
- ☐ The model variance stayed the same.
- ☐ The observational variance decreased.

Thinking Inside the Box [6 Pts]

5. Below are boxplots showing the distributions for three different quantitative variables. We will name these variables Variable A, Variable B, and Variable C.



Some of these distributions may be skewed—if a distribution is skewed, we want to apply a transformation to symmetrize it.

The following three parts will ask which transformations may be suitable for symmetrizing each distribution. If no transformation is necessary, select "No transformation necessary."

(a) [1 Pt] Which of the following transformations may symmetrize Variable A?

☐ $\log(x)$ ☐ x^2 ☐ $\sqrt{x}$ ☐ x^3 ☐ No transformation necessary.

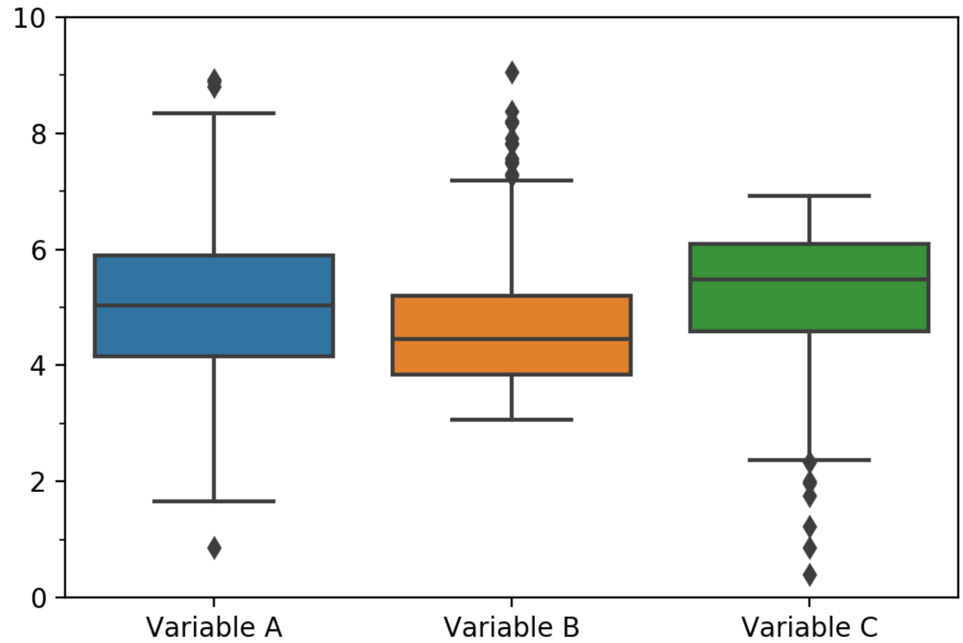
(b) [1 Pt] Which of the following transformations may symmetrize Variable B?

☐ $\log(x)$ ☐ x^2 ☐ $\sqrt{x}$ ☐ x^3 ☐ No transformation necessary.

(c) [1 Pt] Which of the following transformations may symmetrize Variable C?

☐ $\log(x)$ ☐ x^2 ☐ $\sqrt{x}$ ☐ x^3 ☐ No transformation necessary.

The boxplots are repeated here for your convenience.



In each of the following parts, you will see a statement about the boxplots above. Determine if each statement is True, False, or Impossible to tell.

(d) [1 Pt] Variable B has the lowest first quartile among all three variables.

◯ True ◯ False ◯ Impossible to tell

(e) [1 Pt] Variable A is unimodal.

◯ True ◯ False ◯ Impossible to tell

(f) [1 Pt] Variable C contains zero points greater than $1.5 * \text{IQR}$ above its median.

◯ True ◯ False ◯ Impossible to tell

Night Owl or Early Bird? [13 Pts]

6. Suriya and Meghna are Data 100 students, and they have a prediction task where we wish to predict whether people are night owls or early birds using their favorite color. They're given a shortened training set with 5 data points where $X = [\text{'blue'}, \text{'green'}, \text{'pink'}, \text{'purple'}, \text{'red'}]$ and they wish to predict $y = [0, 1, 1, 1, 0]$.

   (a) [1 Pt] What type of variables does X contain?

   ◯ Quantitative continuous
   ◯ Quantitative discrete
   ◯ Qualitative discrete
   ◯ Qualitative nominal
   ◯ Qualitative ordinal

   (b) [1 Pt] They decide to one-hot encode the data in X into a design matrix X', with the categories being ordered alphabetically from left to right. How many values in X' are zero?

   Suriya and Meghna decide to use logistic regression with no intercept term, where the predicted probabilities are rounded to the nearest whole number. Suriya decides to try L0 regularization for their logistic regression model. Unlike L1 and L2 regularization, L0 regularization does not add a term to the loss function. Instead, it specifies a constraint that only some k elements in our parameter θ for the model can be non-zero.

   Hint: $\sigma(0) = .5$

   (c) [2 Pts] Suppose he applies L0 regularization where $k = 5$, and finds the optimal θ for X' and y using logistic regression. How many points does he misclassify?

   (d) [2 Pts] Suppose he applies L0 regularization where $k = 1$ and find the optimal θ for X' and y using logistic regression. How many points does he misclassify?

   Since linear regression using mean square error is easier to solve than logistic regression, Meghna tries to use that instead to create a quick model.

   (e) [1 Pt] Is $X'^T X'$ invertible? ◯ Yes ◯ No

   (f) [2 Pts] What is the optimal value of θ if we use mean square error as the loss function? Your answer should be a sequence of 5 elements, e.g. $[1, 2, 3, 4, 5]$.

(g) [2 Pts] For the optimal value of θ, what is the mean squared error on X'?

Since they can't possibly train a great model with 5 data points, they seek out the full training set and discover that it has 1,000,000 training points with the same colors from before. They decide to use logistic regression, without regularization and without an intercept term, for the remaining parts of the question.

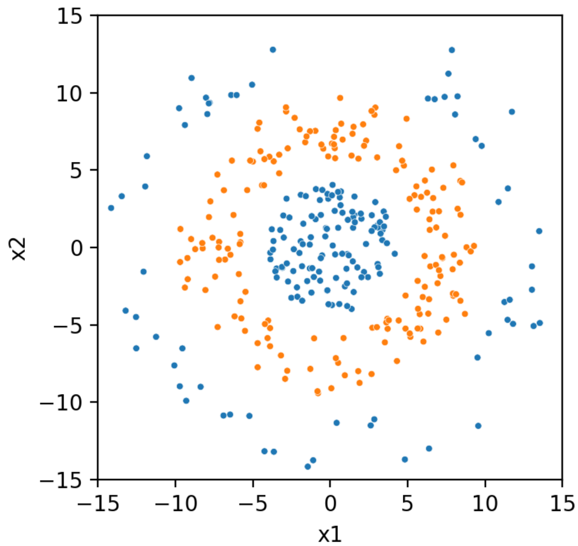
(h) [1 Pt] What is the column rank of the new one-hot encoded dataset?

(i) [1 Pt] Meghna wants to use a gradient method to discover the optimal θ. Which of the following options is the best suited to this training set and problem?

- ◯ Stochastic gradient descent (batch size = 1)
- ◯ Gradient descent, on the complete dataset
- ◯ Stochastic gradient descent (batch size = 32)

Donut Decisions [9 Pts]

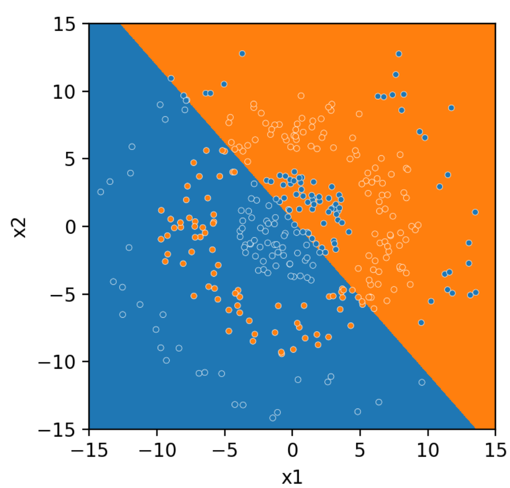
7. Below is a dataset from which we want to create a classifier. We have two features, x_1 and x_2, and two classes. Assume the orange points are in class 1 and the blue points are in class 0. The points displayed here are training data.



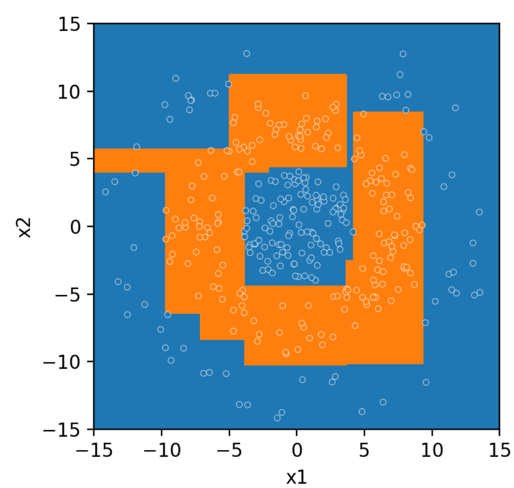
(a) [1 Pt] Is this dataset linearly separable?

◯ Yes ◯ No

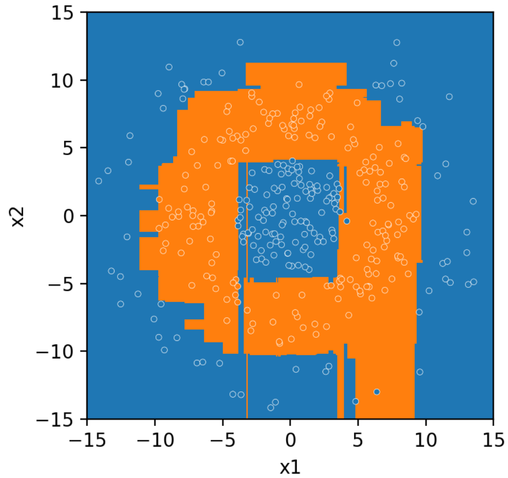
Below are 4 different possible decision boundaries (a.k.a. classifiers) we can generate. The orange regions are areas where new points would be classified as class 1, and the blue regions are areas where new points would be classified as class 0.



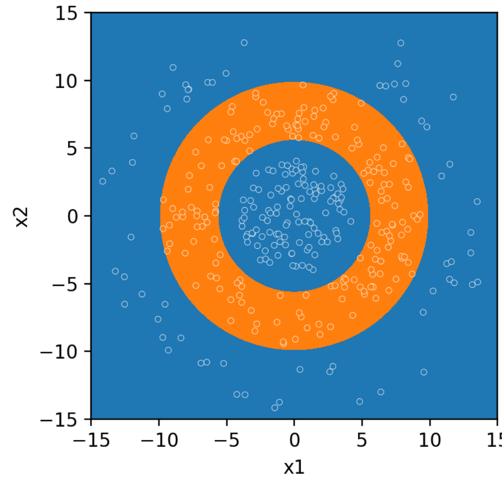
A.



B.



C.



D.

(b) [2 Pts] Which of the above classifiers (A, B, C, D) has perfect accuracy on the training set? Select all that apply.

Note: Do not try to distinguish borderline points—you may assume points right on the boundary are classified correctly.

☐ A ☐ B ☐ C ☐ D

The following parts will ask you about which model(s) could generate each of the boundaries. For each part, *assume* x_1 *and* x_2 *are the only features in our model.*

(c) [1 Pt] Which of the following models could have generated boundary A?

☐ Logistic Regression ☐ Decision Tree ☐ Random Forest ☐ None of the above

(d) [1 Pt] Which of the following models could have generated boundary B?

☐ Logistic Regression ☐ Decision Tree ☐ Random Forest ☐ None of the above

(e) [1 Pt] Which of the following models could have generated boundary C?

☐ Logistic Regression ☐ Decision Tree ☐ Random Forest ☐ None of the above

(f) [1 Pt] Which of the following models could have generated boundary D?

☐ Logistic Regression ☐ Decision Tree ☐ Random Forest ☐ None of the above

(g) [2 Pts] Suppose we add a new feature x_3, which is some function of x_1 and x_2. **Now, assume** x_3 **is a feature in our model.** Which of the following models could have generated boundary D?

☐ Logistic Regression ☐ Decision Tree ☐ Random Forest ☐ None of the above

Feature Engineering [8 Pts]

8. The following dataset contains information about passengers on the Titanic. There are 20 rows in this dataset, and you may assume there are no missing or null values in the dataset. The first 5 rows are shown below.

	sex	age	fare	class	embark_town
0	male	22.0	7.2500	Third	Southampton
1	female	38.0	71.2833	First	Cherbourg
2	female	26.0	7.9250	Third	Southampton
3	female	35.0	53.1000	First	Southampton
4	male	35.0	8.0500	Third	Southampton

A brief description of the columns:

- `age` and `fare` are strictly positive
- `sex` takes on values $\in$ {male, female}
- `class` takes on values $\in$ {First, Second, Third}
- `embark_town` takes on values $\in$ {Southampton, Cherbourg, Queenstown, London, Oxford}

(a) [2 Pts] Suppose we one-hot encode the `sex` column to get a design matrix Φ_1 with 2 columns, `sex_male` and `sex_female`, where values can be 0 or 1 within each column. Note that Φ_1 does NOT contain an intercept term.

Select all of the following statements that are true about Φ_1.

☐ Φ_1 has 20 rows
☐ Φ_1 is full column rank
☐ $\Phi_1^T \Phi_1$ is invertible
☐ None of the above

(b) [2 Pts] Suppose we one-hot encode the `sex` and `embark_town` column and include an intercept term in the model. This results in a design matrix Φ_2 with 8 columns.

Select all of the following statements that are true about Φ_2.

☐ Φ_2 has 20 rows
☐ Φ_2 is full column rank
☐ $\Phi_2^T \Phi_2$ is invertible
☐ None of the above

(c) [2 Pts] **(Hard)** Suppose we one-hot encode the `sex` and `embark_town` column and do **NOT** include an intercept term in the model. This results in a design matrix Φ_3 with 7 columns.

Select all of the following statements that are true about Φ_3.

- ☐ Φ_3 has 20 rows
- ☐ Φ_3 is full column rank
- ☐ $\Phi_3^T \Phi_3$ is invertible
- ☐ None of the above

(d) [2 Pts] Suppose we one-hot encode all the categorical columns (`sex`, `class`, and `embark_town`) and compute the following nonlinear transformations for the quantitative columns `age` and `fare`:

- x^2
- x^3
- $\log(x)$
- $\sin(x)$
- $\cos(x)$

Additionally, we include an intercept term in the model. This results in a design matrix Φ_4.

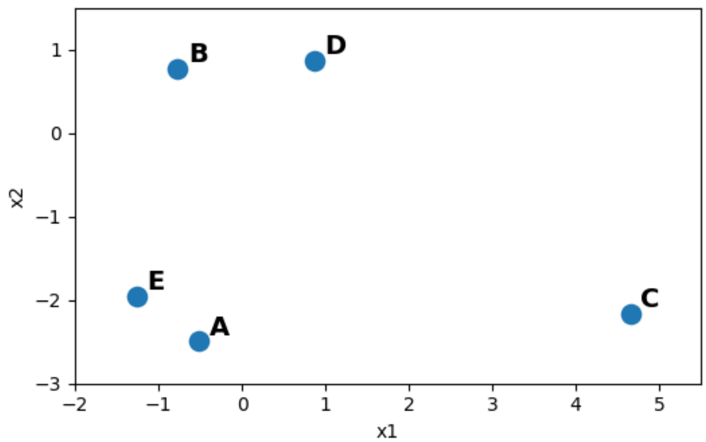
We cannot use the normal equations to minimize the MSE loss and solve for $\hat{\theta}$ because Φ_4 is not full rank. Identify one reason why Φ_4 is not full rank.

Agg(ravating) Clustering [7 Pts]

9. Below is a dataset of 5 points, which we want to group into $k = 3$ clusters.

We will use agglomerative clustering. Our criterion for the distance between clusters is the maximum distance between points from each cluster, as used in lecture.

Below is a scatterplot of the data. Note that both axes have the same scale. Although relative distances should be discernible from the plot, we have provided the Euclidean distance matrix as well. For example, the distance between points **A** and **B** is 3.3.



	A	B	C	D	E
A	0.0	3.3	5.2	3.6	0.9
B	3.3	0.0	6.2	1.6	2.8
C	5.2	6.2	0.0	4.9	5.9
D	3.6	1.6	4.9	0.0	3.5
E	0.9	2.8	5.9	3.5	0.0

In this question, we will walk through each step of the agglomerative clustering algorithm. When two clusters are merged together, the name of the new cluster should be the lowest letter in the cluster. For example, if we merge clusters **Y** and **Z** together, the new cluster should be called **Y**.

(a) [2 Pts] Which two clusters will be merged together first? Your answer should be two letters in alphabetical order. For example, a possible answer would read **YZ**.

____

(b) [2 Pts] Which two clusters will be merged together next? Your answer should be two letters in alphabetical order.

____

(c) [1 Pt] Have we completed the algorithm?

○ Yes ○ No

(d) [2 Pts] Now, we want to cluster this dataset with spectral clustering instead. Remember that when presented with Euclidean point data, we need to construct a graph. In this graph, each vertex corresponds to a point, and the weight of the edge between any two vertices is some function of the Euclidean distance between those points.

In this new graph, which edge will have the **smallest** weight? Your answer should be two letters in alphabetical order.

____

Clustering Facts [7 Pts]

10. Fill in the blank for the following statements about clustering algorithms.

    (a) [1 Pt] When clustering, the algorithm ________ sees data labels.

    ◯ always ◯ sometimes ◯ never

    (b) [1 Pt] A ________ is what we use to represent data when an individual is not represented by its numerical features, but by its relationship to other individuals.

    ◯ graph ◯ design matrix ◯ relational database

    (c) [1 Pt] K-means clustering will ________ find the optimal clustering in terms of inertia.

    ◯ always ◯ sometimes ◯ never

    (d) [1 Pt] When choosing k for k-means clustering, we want to pick a k with a ________ average silhouette score.

    ◯ smaller ◯ larger

    (e) [1 Pt] In spectral clustering, the number of clusters k is ________.

    ◯ predetermined ◯ returned by the algorithm

    (f) [1 Pt] As part of the spectral clustering algorithm, we find the eigenvalues and eigenvectors of the ________ matrix.

    ◯ distance ◯ adjacency ◯ Laplacian

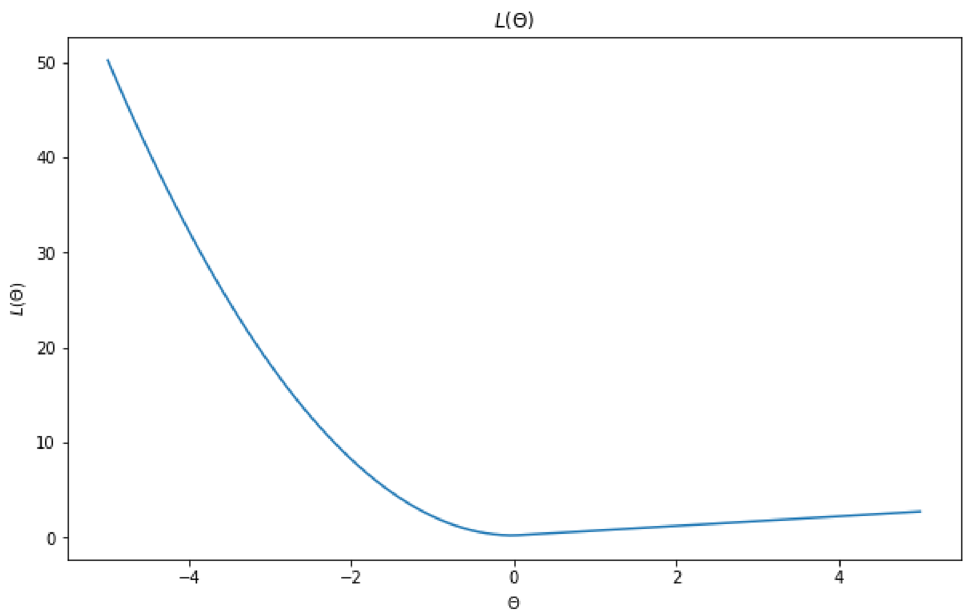
    (g) [1 Pt] After calculating the spectral coordinates of each vertex, we use ________ to determine the clusters.

    ◯ agglomerative clustering ◯ k-means clustering

Infinite Descent [10 Pts]

11. Curious George decides to use gradient descent to minimize his loss function $L(\theta)$ with respect to θ a few times, each time using a different set of hyperparameters. Unfortunately, he refreshed his Jupyter Notebook too early and forgot to save his learning rates and initializations. However, he wrote down his loss function $L(\theta)$, and plotted it with respect to θ (displayed below). In the following parts, we will investigate the behavior of his chosen hyperparameters. Assume the derivative at $\theta = 0$ with respect to θ is 0.

$$L(\theta) = \begin{cases} 2\theta^2 + \frac{3}{16} & \theta \leq 0 \\ \frac{\theta}{2} + \frac{3}{16} & \theta > 0 \end{cases}$$



(a) [1 Pt] George remembers his first set of hyperparameters! He used a learning rate of 0.1 and initialized $\theta^{(0)}$ to -1. After the first iteration of gradient descent, what is the new value of θ? In other words, what is $\theta^{(1)}$?

(b) [1 Pt] After the second iteration of gradient descent, what is our value of θ? In other words, what is $\theta^{(2)}$?

(c) [2 Pts] For his second set of hyperparameters, George remembers he used a learning rate of $\alpha = 0.5$, but he cannot remember the initialization of θ (denoted as $\theta^{(0)}$), where $\theta^{(0)} > 0$. However, he notices after many updates, θ keeps oscillating between the two unique values $\theta^{(0)}$, the initial value, and $\theta^{(1)}$, the value after one gradient update.

Which of the following could cause gradient descent to oscillate between two unique values in this example?

☐ The sign of the gradient oscillates.

☐ The sign of the gradient does not oscillate.

☐ The magnitude of the gradient oscillates.

☐ The magnitude of the gradient does not oscillate.

☐ The learning rate is inappropriate.

(d) [2 Pts] For his third set of hyperparameters, suppose George used a fixed learning rate with gradient descent, and he initialized θ at 4. Assume convergence to the global minimum took 12 steps—that is, $\theta^{(12)} = 0$. Given this information, what is a fixed learning rate that he used? Round to 3 decimal places, or write your answer as a fraction - e.g. "$10/9$".

After these experiments, George decides to upgrade his gradient descent techniques to the next level.

(e) [2 Pts] George doesn't want to use a fixed learning rate for gradient descent anymore. Assuming t represents the timestep corresponding to each gradient step, with $t = 1$ corresponding to the first gradient step and $t = 2$ corresponding to the second, which of the following function(s) are reasonable to describe the learning rate $R(t)$ such that gradient descent will typically converge?

Hint: Think about how the learning rate should behave as we keep iterating.

☐ $R(t) = 2t$

☐ $R(t) = -2t$

☐ $R(t) = \frac{1}{2t}$

☐ $R(t) = e^{2t}$

☐ $R(t) = e^{-2t}$

(f) [2 Pts] George modifies his loss function and learning rate, while using stochastic gradient descent. He notices that gradient descent converges to a real number every time, but depending on his random initialization, the resulting θ changes, with *different corresponding values of the loss function*. Which of the following could cause this behavior?

Hint: Recall that gradient descent "converges" when the model weights do not change from one iteration to the next.

☐ The loss function is convex.

☐ The batch size is too small.

☐ The batch size is too large.

☐ The loss function is not convex.

Decisions, Decisions [13 Pts]

12. Unfortunately, Clippy and some of his classmates did not reach the required accuracy on their spam-ham email model from Homework 11. To get partial credit, they are required to collect their own data and build their own classifiers. Clippy decides to surf his email inbox for data and comes up with a new training dataframe, called `train`, that consists of 7 rows and 2 features. Similarly, he has a testing dataframe, called `test`, with 3 rows and 2 features.

    - The feature `A` represents the polarity of the email **body**, as determined by VADER
    - The feature `B` represents the polarity of the email **subject**, as determined by VADER

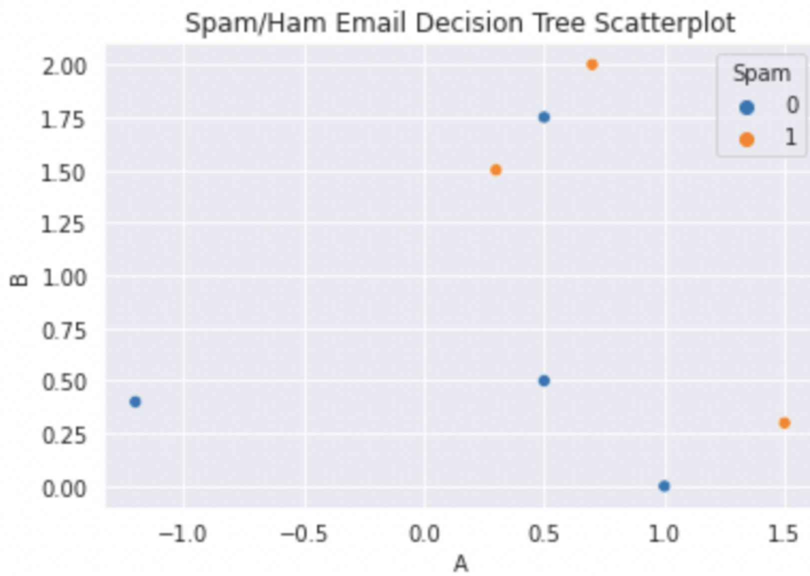
`train`

	A	B	Spam
0	1.5	0.30	1
1	0.5	0.50	0
2	0.3	1.50	1
3	0.7	2.00	1
4	-1.2	0.40	0
5	1.0	0.00	0
6	0.5	1.75	0

`test`

	A	B
0	-1.0	2.00
1	-0.5	0.25
2	1.2	-0.20

(a) [2 Pts] Say Clippy wanted to visualize his training data using a 2D scatter plot, with feature A on the x-axis and feature B on the y-axis. He wants to color his data points differently to distinguish between emails labeled as spam or ham. Write a single line of code using Seaborn to return the following visualization.



(b) [2 Pts] What is the minimum number of linear splits that completely separate the two classes?

(c) [2 Pts] What is the minimum depth of the decision tree that completely separates the two classes? Assume a tree with exactly one split has depth 1.

(d) [2 Pts] Among all possible decision trees that Clippy can train from his training set, what is the maximum testing accuracy Clippy can achieve from any of these models? **Assume the first two emails are truly non-spam (ham) emails, and the third email is spam.** Write your answer as a percentage, including the % symbol.

(e) [2 Pts] What is the weighted entropy of the child nodes after splitting on the feature $\texttt{B} < 1.00$? Round your answers to 3 decimal places, and omit the leading 0.

(f) [1 Pt] Fill in the blank. Using entropy as an indicator of split effectiveness, the proposed split on $\texttt{A} < 1.00$ is __________ to that of the proposed split on $\texttt{A} > 1.25$.

◯ Better than ◯ Worse than ◯ Equal ◯ Cannot be determined

(g) [2 Pts] Clippy realizes the data he's working with is not very representative of spam and ham emails. He is thinking of possible solutions to build a more accurate decision tree which is generalizable to the larger dataset provided in Homework 11. **When implemented alone**, select all of the following possible solutions.

- ☐ Increase the number of features in his training dataset.
- ☐ Increase the size of his test set to get a more accurate testing error.
- ☐ Collect more emails to use for his training and testing dataset.
- ☐ Use LASSO regularization.

(R)ussian (O)lympic (C)ommittee [10 Pts]

13. It's almost time for the gold-medal match of the Data Science Olympiad, with the United States set to take on Russian Olympic Committee.

We have collected lots of data from other matches at this tournament, which we will use to create a logistic regression classifier to predict the outcome of the gold-medal match. Consider the training data to be all previous matches in the tournament, and the gold-medal match as a single test point.

For the gold-medal match, if our classifier outputs 1, we will predict Russian Olympic Committee will win the match, and if our classifier outputs 0, we will predict United States wins the match.

However, we do not know what threshold, T, to use for our model.

(a) [1 Pt] True or False? One way to select T is to try different candidate values of T, create a ROC curve for each one, and pick the T with the largest area under the curve.

◯ True ◯ False

(b) [1 Pt] Suppose we decide that predicting Russian Olympic Committee to win, and being wrong, is a better scenario than predicting the United States to win, and being wrong. Of the following three possible thresholds, which makes the most sense to use for our model?

◯ $T = .3$ ◯ $T = .5$ ◯ $T = .7$

Below is the output of our model on 10 randomly selected training points. The first column contains the true Y, and the second column contains our model's estimate of $\mathbb{P}(Y = 1)$.

y	P(Y = 1)
0	0.133205
0	0.151632
0	0.209790
0	0.361822
1	0.488413
1	0.521694
0	0.721061
1	0.765493
1	0.868745
1	0.925884

(c) [1 Pt] If we set $T = .5$, what is our model's accuracy on these ten training points?

(d) [2 Pts] If we set $T = .5$, what is our model's precision on these ten training points?

(e) [2 Pts] If we set $T = .5$, what is our model's recall on these ten training points?

(f) [3 Pts] Give a value of T that maximizes the accuracy on these 10 training points.

Rare Red Rabbits [20 Pts]

14. Kermit the Frog and Miss Piggy join a job as data science consultants at a national reserve where there live rare red and blue rabbits. They are tasked with exploring yearly data so that they can develop models to help predict future red and blue rabbit population. They are provided with the following DataFrame, `rabbit`. Specifically, `rabbit` includes the location, year, unique ID, name and color of each rabbit.

	Location	Year	ID	Name	Color
0	Site A	2018	A192391839	Alvin	B
1	Site B	2018	B1827333	Kelly	R
2	Site A	2019	M318774443	Peter	R
3	Site C	2017	P8773323	Lapine-Rouge	R
4	Site B	2020	Q382819311	Daisy	B
5	Site A	2019	A192391839	Alvin	B

Additionally, they are provided with another DataFrame, `metadata`. `metadata` contains the year in consideration, carrots eaten in that year, and the number of animal tracks found.

	Year	Carrots Eaten	Tracks Found
0	2018	3817	81273
1	2019	10283	150281
2	2020	30381	300372

Fill out the Pandas expression below to generate a DataFrame where each row corresponds to a **unique** year. Each row should also contain the number of animal tracks, carrots eaten, total red rabbit population, and total blue rabbit population for that year. Fill all missing values with 0. You may not need to use all the provided blanks.

```
rabbit.__A__(_________________B_________________) \
       .merge(_________________C_________________) \
       .__D__(0)
```

Here is what your output should look like:

	B	R	Year	Carrots Eaten	Tracks Found
NaN	0.0	1.0	2017	0	0
0.0	1.0	1.0	2018	3817	81273
1.0	1.0	1.0	2019	10283	150281
2.0	1.0	0.0	2020	30381	300372

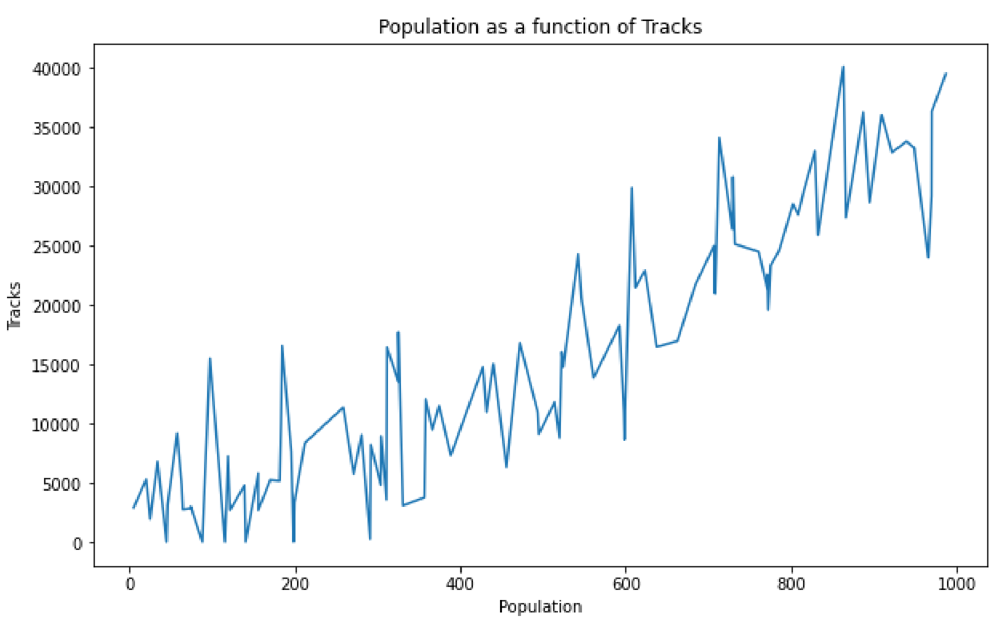
(a) [2 Pts] What goes in the blank indicated by the letter A?

(b) [2 Pts] What goes in the blank indicated by the letter B?

(c) [2 Pts] What goes in the blank indicated by the letter C?

(d) [1 Pt] What goes in the blank indicated by the letter D?

Kermit decides to perform some EDA before diving into some modeling. He decides to focus on the relationship between the population and the number of tracks.



(e) [2 Pts] Describe an issue with the visualization above.

(f) [1 Pt] Which of the following is most likely the correlation coefficient r between population and number of tracks?

◯ -1 ◯ $-.9$ ◯ $-.3$ ◯ 0 ◯ $.3$ ◯ $.9$ ◯ 1

Using his findings from above, Kermit wishes to predict the total rabbit population, p, using the number of animal tracks, t, found in the year. He decides to use Ridge regression **without an intercept term**, and with regularization hyperparameter λ.

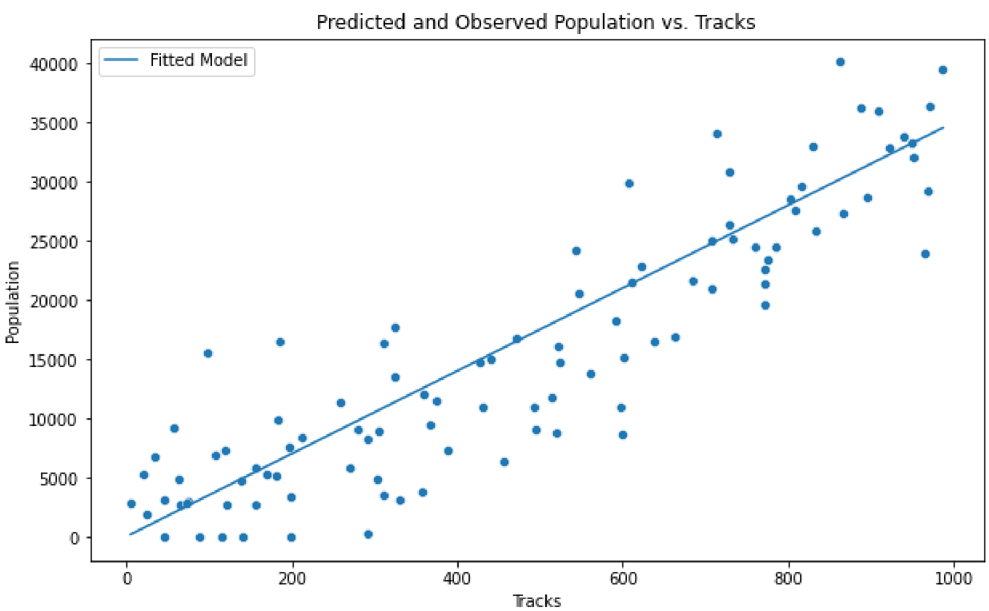
(g) [2 Pts] Which of the following is the optimal $\hat{\theta}$?

◯ $\frac{t^T p}{t^T t + \lambda n}$

◯ $\frac{t^T t}{t^T p + \lambda n}$

◯ $\frac{1}{t^T p + \lambda n}$

(h) [2 Pts] After fitting his model, he plots his predictions against his observations as shown below. Which of the following are true? Select all that apply.



☐ The fit is inaccurate due to the large amounts of scatter around the line, which suggests that linear regression is inappropriate.

☐ There is a curvature in the relationship, which suggests that linear regression is inappropriate.

☐ The training loss would decrease if the regularization hyperparameter λ was 0.

☐ The plot displays high variance, which suggests that λ should be increased to reduce model complexity.

15. Using the same data we constructed in part a of the previous question, Kermit wishes to create a second model to predict the *proportion* of red rabbits given the previous year's proportion of red rabbits. Using the proportion of red rabbits from this new model and the total population from the model trained in part (b), he can calculate the red and blue rabbit population for any year!

   (a) [2 Pts] Which machine learning model(s) would **not** be suitable for predicting the proportion of red rabbits?

   ☐ `KMeans`
   ☐ `LASSO Regression`
   ☐ `Logistic Regression`
   ☐ `Ridge Regression`
   ☐ `PCA`

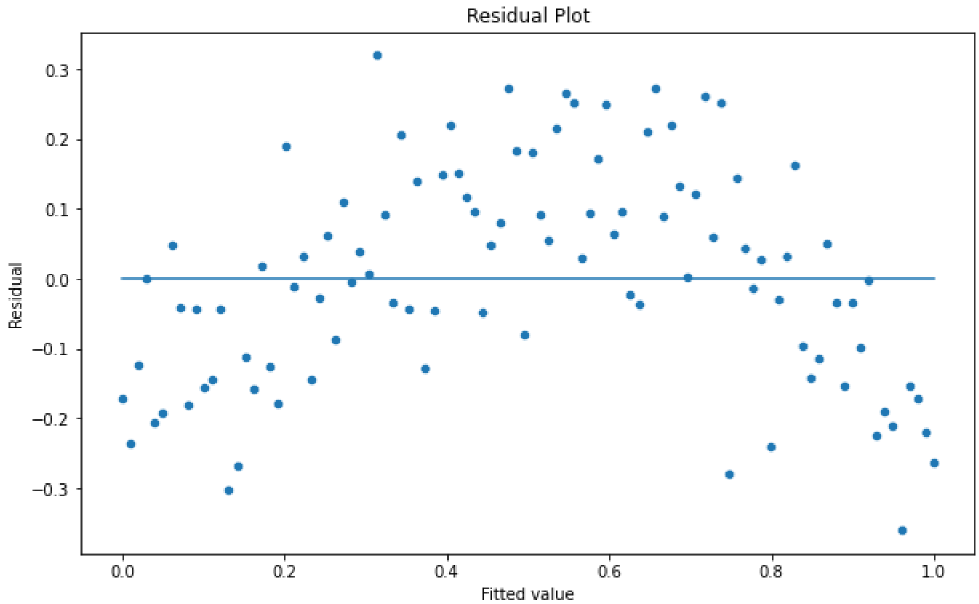
   (b) [2 Pts] Regardless of your answer to the previous question, Miss Piggy invents her own (bad!) loss function to train a linear model **with an intercept term**. Recall that her only feature x represents the proportion of red rabbits from the previous year. The linear model is written as $f_{\vec{\theta}}(x) = \theta_0 + \theta_1 x$ and the loss function is written as:

$$L(y, f_{\vec{\theta}}(x)) = \frac{1}{n}\sum_{i=1}^{n}(f_\theta(x_i) - \theta_1 y_i) + \sum_{j=0}^{1}\bar{y}e^{\theta_j}$$

   Calculate the optimal value for $\hat{\theta}_1$.

   ○ $\ln(1 + \frac{\bar{x}}{\bar{y}})$
   ○ $\ln(1 - \frac{\bar{x}}{\bar{y}})$
   ○ $\ln(1 + \frac{\bar{y}}{\bar{x}})$
   ○ $\ln(1 - \frac{\bar{y}}{\bar{x}})$

   (c) [2 Pts] Unfortunately, Miss Piggy's loss function didn't work, so she resorts to minimizing the MAE on her linear model instead. She generates predictions on the test set and create a residual plot as shown below. What does this visualization indicate?



- ◯ The residual plot displays a roughly normal distribution, which suggests that linear regression is appropriate.
- ◯ The residual plot displays roughly uniformly random scatter, which suggests that linear regression is appropriate.
- ◯ The residual plot has many outliers, which suggests that linear regression is inappropriate.
- ◯ The residual plot has a weak relationship with the fitted value, which suggests that linear regression is inappropriate.
- ◯ The residual plot has curvature, which suggests that linear regression is inappropriate.

Rare Red Rabbits Return [10 Pts]

16. Kermit wants to apply PCA to the rare rabbit dataset from the previous question to understand patterns in rabbit population per location as a function of the year. Provided is a Pandas DataFrame, `rabbit_pop` (shown below), which contains the rabbit population for every particular year and location. Note that not every year and location is shown here.

	2017	2018	2019	2020
Site A	8789	29372	49271	101822
Site B	18573	38317	102847	192742
Site C	402	3928	20212	80272
Site D	4392	28172	93172	203082

Kermit needs to preprocess his current dataset in order to use PCA.

(a) [1 Pt] Select all appropriate preprocessing steps used for PCA.

- ☐ Transform each row to have a magnitude of 1 (Normalization)
- ☐ Transform each column to have a mean of 0 (Centering)
- ☐ Transform each column to have a mean of 0 and a standard deviation of 1 (Standardization)
- ☐ None of the above

(b) [1 Pt] Kermit wishes to apply a transformation to the rabbit population at each site for each year so that he can apply PCA more effectively. What transformation function $f(p)$ would be most effective to apply to the rabbit population p for using PCA?

Hint: Notice the population at each site appears to grow exponentially every year.

- ◯ `f = np.log`
- ◯ `f = np.exp`
- ◯ `f = np.square`
- ◯ `f = np.sqrt`
- ◯ `f = lambda x:  x`

(c) [2 Pts] Assume we decide to standardize the data, regardless of your answer to the previous subparts. Select the line of Pandas code that preprocesses the population for PCA into the DataFrame `rabbit_PCA`.

- ◯ `rabbit_PCA = f(rabbit_pop - rabbit_pop.mean()) / rabbit_pop.std()`
- ◯ `rabbit_PCA = f((rabbit_pop - rabbit_pop.mean()) / rabbit_pop.std())`
- ◯ `rabbit_PCA = f(rabbit_pop - rabbit_pop.mean()) / f(rabbit_pop.std())`

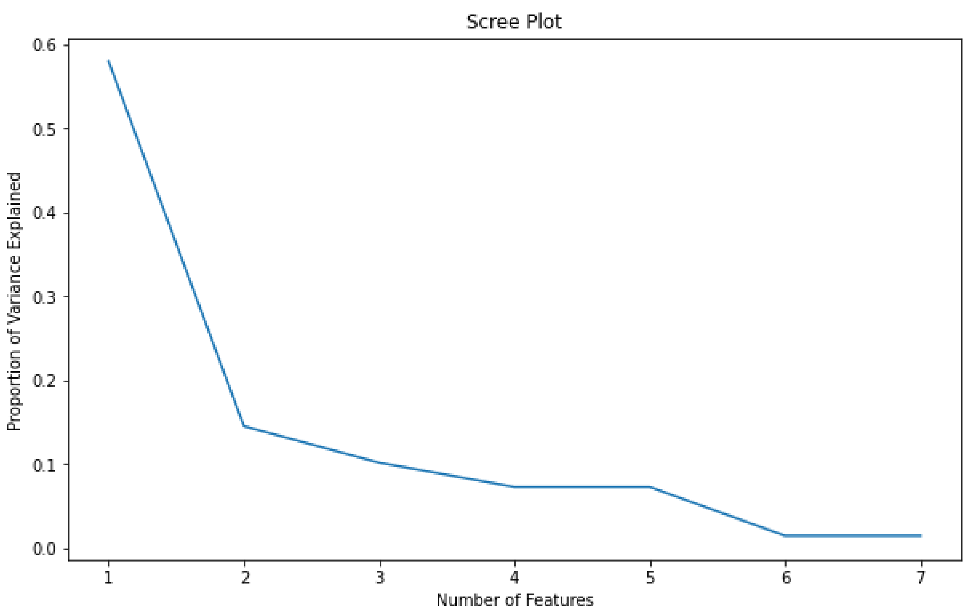
◯ `rabbit_PCA = (f(rabbit_pop) - f(rabbit_pop).mean()) / f(rabbit_pop).std()`

◯ `rabbit_PCA = f(rabbit_pop - rabbit_pop.mean()) - f(rabbit_pop.std())`

(d) [2 Pts] Write a line of code that returns the first 3 principal components assuming you have `rabbit_PCA` and the following variables returned by SVD.

```
u, s, vt = np.linalg.svd(rabbit_PCA, full_matrices = False)
first_3_pcs = ________________________________
```

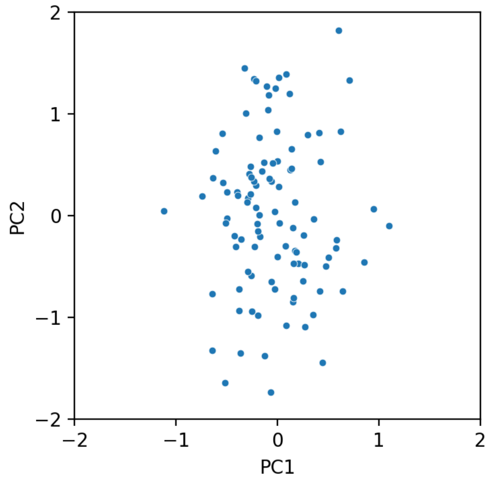
(e) [2 Pts] Kermit successfully applies PCA and makes a scree plot that is displayed below. How many principal components should Kermit use to capture at least 80% of the variance in the rabbit population data?



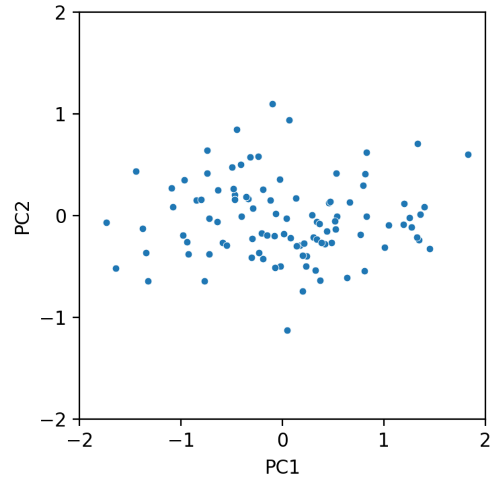
◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7

(f) [2 Pts] We now wish to display the first two principal components in a scatterplot. Which of the following plots could potentially display the first two principal components?

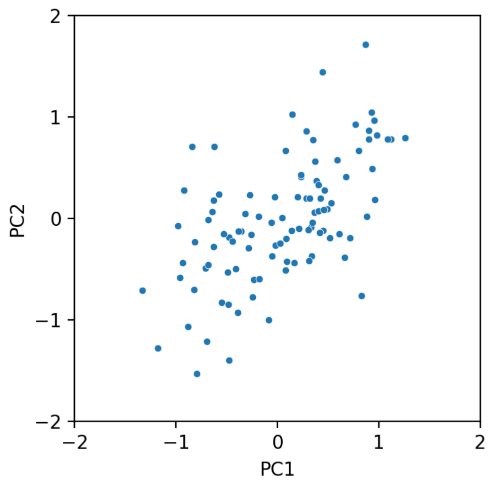
Hint: The above scree plot may be helpful.


PC2
PC1
2
1
0
−1
−2
−2
−1
0
1
2

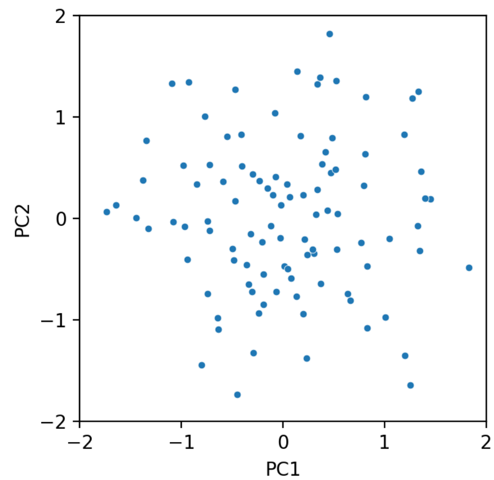
A.


PC2
PC1
2
1
0
−1
−2
−2
−1
0
1
2

B.


PC2
PC1
2
1
0
−1
−2
−2
−1
0
1
2

C.


PC2
PC1
2
1
0
−1
−2
−2
−1
0
1
2

D.

◯ A ◯ B ◯ C ◯ D