Data 100, Spring 2022

Discussion #1

Welcome to Data 100!

Linear Algebra Fundamentals

1. Linear algebra is what powers linear regression, logistic regression, and PCA (concepts we will be studying in this course). Moving forward, you will need to understand how matrix-vector operations work. That is the aim of this problem.

   Josh, Lisa, and Kobe are shopping for fruit at Berkeley Bowl. Berkeley Bowl, true to its name, only sells fruit bowls. A fruit bowl contains some fruit and the price of a fruit bowl is the total price of all of its individual fruit.

   Berkeley Bowl has apples for \$2, bananas for \$1, and cantaloupes for \$4. (expensive!). The price of each of these can be written in a vector:

   $$\vec{v} = \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}$$

   Berkeley Bowl sells the following fruit bowls:

   1. 2 of each fruit
   2. 5 apples and 8 bananas
   3. 2 bananas and 3 cantaloupes
   4. 10 cantaloupes

   (a) Define a matrix B such that $B\vec{v}$ evaluates to a length 4 column vector containing the price of each fruit bowl. The first entry of the result should be the cost of fruit bowl 1, the second entry the cost of fruit bowl 2, etc.

(b) Josh, Lisa, and Kobe make the following purchases:

- Josh buys 2 fruit bowl 1s and 1 fruit bowl 2.
- Lisa buys 1 of each fruit bowl.
- Kobe buys 10 fruit bowl 4s (he really like cantaloupes).

Define a matrix A such that the matrix expression $AB\vec{v}$ evaluates to a length 3 column vector containing how much each of them spent. The first entry of the result should be the total amount spent by Josh, the second entry the amount sent by Lisa, etc.

(c) Let's suppose Berkeley Bowl changes their fruit prices, but you don't know what they changed their prices to. Josh, Lisa, and Kobe buy the same quantity of fruit baskets and the number of fruit in each basket is the same, but now they each spent these amounts:

$$\vec{x} = \begin{bmatrix} 80 \\ 80 \\ 100 \end{bmatrix}$$

In terms of A, B, and $\vec{x}$, determine $\vec{v_2}$ (the new prices of each fruit).

2. As a warm up for the homework, we will introduce matrix inverses and matrix rank.

- The inverse of a square invertible matrix M, M^{-1} is defined as a matrix such that $MM^{-1} = I$ and $M^{-1}M = I$. The matrix I is a special matrix denoted as the identity matrix where the diagonal elements are 1 and the non-diagonal elements are 0.
- Linear dependence among a set of vectors $\{v_1, v_2, v_3, ..., v_n\}$ is defined as follows. If any (non-trivial) linear combination of the vectors can produce the zero vector, then the set of vectors is linearly dependent.

  In other words, if we can multiply the vectors v_i with some scalar α_i and sum the quantity to obtain the zero vector (given at least one $\alpha_j \neq 0$, then the set is linearly dependent.

$$\sum_{i=1}^{n} \alpha_i v_i = 0 \text{ such that some } \alpha_j \neq 0 \implies \text{linear dependence}$$

Any set of vectors such that we cannot obtain the zero vector as described above is linearly independent.

- The (column) rank of a matrix M is the maximal number of linearly independent column vectors in M. A full rank matrix has a column rank equal to the number of column vectors.

We will go over all of these definitions applied to relevant practical examples in the following subparts.

(a) Consider the matrix $M = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$ containing two column vectors $v_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $v_2 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$. Is it possible to construct the zero vector using a linear combination of the column vectors? What can be concluded about the rank of the matrix M?

(b) Consider the inverse matrix $M^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ of M. Carry out the matrix multiplication MM^{-1}, and determine what M^{-1} must be.

(c) Consider a different matrix $Q = \begin{bmatrix} 1 & 0 & 5 \\ 0 & 1 & 5 \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}$. What is the column rank of the matrix? Is the matrix invertible?

(d) Consider a matrix R, which is equal to the transpose of the matrix Q: $R = Q^T$. What is the column rank of the matrix R? Is the matrix R invertible?

3. *(Bonus)* We will explore a few properties of a special symmetric matrix that we will use quite a bit when we cover linear regression and regularization. Consider a matrix X, of shape m by n for some $m \geq n$. We will work with a matrix that is created by matrix multiplying X^T with X: X^TX. We will prove an interesting property:

   If X is of full column rank, then X^TX is invertible.

   (a) Explain why the matrix X^TX is symmetric. Recall that the transpose operation turns the first column into the first row, second column into the second row, and so on.

   (b) Recall that invertibility for square matrices depends on the matrix's rank. Simply put, redundant (or linearly dependent) vectors with respect to a matrix's span reduce its rank. We will determine whether X^TX is invertible by using an equivalent condition without proof (for those curious, this leverages the fact that a square matrix is positive definite implies invertibility).

   Lemma 0.1. *If for all non-zero vectors v, $v^TMv > 0$, then M is invertible.*
   What condition is required to invert the matrix X^TX?

   (c) The nullspace of a matrix M is defined as the set of all vectors that when multiplied by the matrix M yield 0. In other words, it is the set $\{v : Mv = 0\}$. Prove that if X has a trivial nullspace (i.e. a nullspace with only the zero vector), then X^TX is invertible.
   In case you're curious, this is the exact condition that we will need for least squares regression to work later on!
   Hint: Using the fact that $(Av)^T = v^TA^T$, simplify the condition in the previous part by creating a new vector w and leverage the fact that $w^Tw = \|w\|^2$.

   (d) Do the same analysis with $X^TX + \lambda I$ for $\lambda > 0$. You should notice that the conditions required for invertibility are much more lax; this is the setup for ridge regression, which we'll study later too!

Calculus

In this class, we will have to determine which inputs to a functions minimize the output (for instance, when we choose a model and need to fit it to our data). This process involves taking derivatives.

In cases where we have multiple inputs, the derivative of our function with respect to one of our inputs is called a *partial derivative*. For example, given a function $f(x, y)$, the partial derivative with respect to x (denoted by $\frac{\partial f}{\partial x}$) is the derivative of f with respect to x, taken while treating all other variables as if they're constants.

4. Suppose we have the following scalar-valued function on x and y:

$$f(x, y) = x^2 + 4xy + 2y^3 + e^{-3y} + \ln(2y)$$

(a) Compute the partial derivative of $f(x, y)$ with respect to x.

(b) Compute the partial derivative of $f(x, y)$ with respect to y.

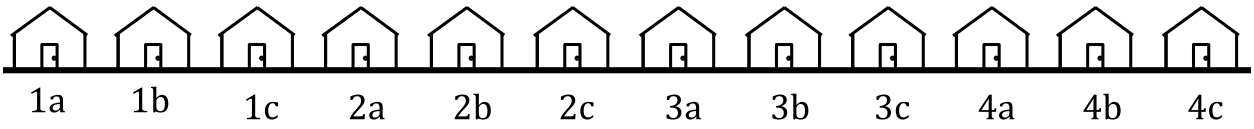
(c) The gradient of a function $f(x, y)$ is a vector of its partial derivatives. That is,

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T$$

$\nabla f(x, y)$ tells us the magnitude and direction in which f is moving, at point (x, y). This is analogous to the single variable case, where $f'(x)$ is the rate of change of f, at the point x.

Using your answers to the above two parts, compute $\nabla f(x, y)$ and evaluate the gradient at the point $(x = 2, y = -1)$.

Probability & Sampling



5. Kalie wants to measure interest for a party on her street. She assigns numbers and letters to each house on her street as illustrated above. She picks a letter "a", "b", or "c" at random and then surveys every household on the street ending in that letter.

   (a) What is the chance that two houses next door to each other are both in the sample?

   (b) Now, suppose that Kalie decides to collect an SRS of one house instead. What is the probability that house 1a is **not** selected in Kalie's SRS of one house?

   (c) Kalie decides to collect a SRS of four houses instead of a SRS of one house. What is the probability that house 1a is **not** in Kalie's simple random sample of four houses?

   (d) Instead of surveying every member of each house from the SRS of four houses, Kalie decides to only survey two members in each house. Four people live in house 1a, one of whom is Bob. What is the probability that Bob is **not** chosen in Kalie's new sample?

Proportions (Bonus)

In Data 100 you will typically work with multiple variables and large data sets. But before we get carried away by complexity, let's make sure we have our feet on the ground when it comes to interpreting simple quantities like proportions.

6. Investigators at the scene of a crime find a footprint that shows a distinctive pattern on the sole. They identify the type of shoe, and then they find a person owns that kind of shoe and could have committed the crime. They put this person on trial for the crime.

   After looking at sales patterns and so on, the investigators find that of the 10,000 other people who could have committed the crime, 1 in 1,000 own that kind of shoe.

   The prosecution says that given these findings, the chance that the defendant is not the guilty person is 1 in 1,000.

   The prosecution has made an error called the "prosecutor's fallacy." Unfortunately it's rather common. Let's see what the error is and what conclusions we can draw from the evidence.

   (a) There are 10,001 people who could have committed the crime. Define a person to be "Matching the Footprint" if the person owns the kind of shoe identified by the investigators. Fill in the table below with the counts of people in the four categories. The four counts should add up to 10,001, and you should assume, as the prosecution did, that only one person is guilty.

	Guilty	Not Guilty
Matching the Footprint		
Not Matching the Footprint		

   (b) The prosecution has reported a proportion as a chance. Whether they know it or not, this implies they are assuming that the defendant is like a person drawn at random from the group who could have committed the crime. So let's assume that too. That is, we assume the defendant is drawn at random from 10,001 people of whom 1 is guilty.

   Use the table in Part **a** to fill in the blanks with choices from among "Guilty", "Not Guilty", "Matching the Footprint", and "Not Matching the Footprint". The vertical bar is the usual notation for "given".

   Under this assumption, $\frac{1}{1000} = P(\_\_\_\_\_ \mid \_\_\_\_\_)$.

(c) What the investigators know is that the defendant has the fateful type of shoe. Fill in the blanks:

Given the findings of the investigators, the chance that the defendant is not guilty is $P(\_\_\_\_\_ \mid \_\_\_\_\_) = \_\_\_\_\_$.

The last blank should be filled with a fraction, and the first two should be filled choices from among "Guilty", "Not Guilty", "Matching the Footprint", and "Not Matching the Footprint".

Note: The prosecution's error is to confuse the probabilities in Parts **b** and **c**.

Election Forecasts (Bonus)

7. People have a hard time understanding polls. In September 2016, the New York Times tried to explain aspects of polls that tend to get overlooked. To illustrate the issues, they gave all the data in one of their own polls to four well-known forecasters and asked them to make predictions.

   The data were from a poll of 867 Florida voters and the exercise was to predict Trump/-Clinton result in Florida. In the election, Trump beat Clinton in Florida, 49% to 47.8%.

   Here are the forecasts. The Times' own results, derived jointly with researchers at Siena College, are in the last line of the table.

Pollster	Clinton	Trump	Margin
Charles Franklin Marquette Law	42%	39%	Clinton +3%
Patrick Ruffini Echelon Insights	39%	38%	Clinton +1%
Omero, Green, Rosenblatt Penn Schoen Berland Research	42%	38%	Clinton +4%
Corbett-Davies, Gelman, Rothschild Stanford University/Columbia University/Microsoft Research	40%	41%	Trump +1%
NYT Upshot/Siena College	41%	40%	Clinton +1%

(a) Pick one of the options (i) and (ii); if you pick (ii), provide the reason.

The predictions were different from each other because

(i) samples come out differently due to randomness so the forecasters all had different data.

(ii) _____

(b) Point out one other interesting aspect of the data in the table. This question doesn't have just one right answer; just describe something you noticed.

(c) If you were going to forecast an election result, which of the following groups would you most want to focus on, and why? Pick at most two groups.

(i) adults in the Census

(ii) eligible voters

(iii) registered voters

(iv) likely voters

(v) undecided voters

(d) Of the two main methods for identifying likely voters, described below, one does a better job at predicting whether the person will show up and vote. Which do you think it is, and why? Could it systematically exclude some likely voters?

- Self-reported voting intention
- Voting history (in which past elections did the person vote; data available in the voter registration database)

Data 100, Spring 2022 **January 28, 2021**

Discussion #2

Finding Chances

Golden rules for finding the probability of an event:

- Addition Rule: list all the distinct ways the event can happen, and add the chances of all the ways. Note that all events must be mutually exclusive for this rule to apply!
- Complement Rule: if the list above looks long and complicated, make the list of ways in which the event *doesn't* happen and calculate its probability q; it might be simpler. The probability of the original event happening p is the complement of q: $p = 1 - q$.
- Multiplication Rule: If an event involves multiple independent trials, like a number of random draws, imagine yourself conducting the experiment one trial at a time. The probability of the event is the product of the probabilities of each trial.

1. Consider a sample of size n where n is a positive integer drawn at random with replacement from a population in which a proportion p of the individuals are called successes.
   (a) For an integer k such that $0 \leq k \leq n$, which of the following are equal to the chance of getting exactly k successes in the sample?
   (i) $p^k(1-p)^{n-k}$
   (ii) $\binom{n}{k}p^k(1-p)^{n-k}$
   (iii) $\binom{n}{n-k}p^k(1-p)^{n-k}$
   (iv) $\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$
   (b) Which of the following are equal to the chance of getting at least one success in the sample?
   (i) $np(1-p)^{n-1}$
   (ii) $\sum_{k=2}^{n}\binom{n}{k}p^k(1-p)^{n-k}$
   (iii) $\sum_{k=1}^{n}\binom{n}{k}p^k(1-p)^{n-k}$
   (iv) $1-p^n$
   (v) $1-(1-p)^n$

Sampling and Bias

2. It's time for the Data 100 midterm and the professor wants to estimate the difficulty of the exam. They decide to survey students on the exam's difficulty with a 10-point scale and then use the mean of the students' responses as the estimate.

   (a) What is the population the professor is interested in trying to understand?

   ◯ A. Students in Data 100
   ◯ B. Students enrolled in the Data 100 Piazza
   ◯ C. Students who attend the Data 100 lecture
   ◯ D. Students who took the Data 100 midterm

   (b) The professor considers a few different methods for collecting the survey data. Which of the following methods is best? (think through which considerations go into "best")

   ◯ A. The professor sends a Zoom poll to all students in an optional midterm debrief lecture after going over exam solutions.
   ◯ B. The professor adds a question to the homework assignments of a simple random sample of students within every discussion section.
   ◯ C. The professor makes a post on Piazza asking students to submit an anonymous Google Form containing the survey question.
   ◯ D. The professor chooses a simple random sample of discussion sections, goes to each selected section, and asks each student in the group as part of the final discussion question.

3. A campus organization wants to take a sample of Berkeley students who are registered for classes this semester. To do this, the organization takes a simple random sample of 20 classes from among all classes offered this semester, and then takes all students in those classes. You can assume that the organization has access to complete enrollment information all classes.

   (a) Is this a simple random sample of students? Explain.

   (b) Is this a probability sample of students? Explain.

4. The Current Population Survey is a national survey run by the Census Bureau. It is thorough and reliable, and thus is sometimes used as a benchmark to assess the accuracy of other surveys.

   As part of an assessment of its own phone surveys, the Pew Research Center found that the response rates have been dropping over the years. Still, on most measures, its estimates were comparable to those of the Current Population Survey. For example, 55% of respondents in the most recent Pew Survey said they did some type of volunteer

work for or through an organization in the past year in a phone survey, compared to 27% in the Current Population Survey.

How do you think this difference might have arisen?

Pandas Practice

Below are the first few rows of the `elections` DataFrame from lecture.

	Year	Candidate	Party	Popular vote	Result	%
0	1824	Andrew Jackson	Democratic-Republican	151271	loss	57.210122
1	1824	John Quincy Adams	Democratic-Republican	113142	win	42.789878
2	1828	Andrew Jackson	Democratic	642806	win	56.203927
3	1828	John Quincy Adams	National Republican	500897	loss	43.796073
4	1832	Andrew Jackson	Democratic	702735	win	54.574789

5. We want to select the "Popular vote" column as a `pd.Series`. Which of the following lines of code will error?

   A) `elections['Popular vote']`

   B) `elections.iloc['Popular vote']`

   C) `elections.loc['Popular vote']`

   D) `elections.loc[:, 'Popular vote']`

   E) `elections.iloc[:, 'Popular vote']`

6. Write one line of Pandas code that returns a `pd.DataFrame` that only contains election results from the 1900s.

7. Write one line of Pandas code that returns a `pd.Series`, where the index is the Party, and the values are how many times that party won an election.

   `Hint:` use `value_counts()`.

Grading Assistance (Bonus)

8. Fernando is writing a grading script to compute grades for students in Data 101. Recall that many factors go into computing a student's final grade, including homework, discussion, exams, and labs. In this question, we will help Fernando compute the homework grades for all students using a DataFrame, `hw_grades`, provided by Gradescope.

   The Pandas DataFrame `hw_grades` contains homework grades for all students for all homework assignments, with one row for each combination of student and homework assignment. **Any assignments that are incomplete are denoted by NaN (missing) values, and any late assignments are denoted by a True boolean value in the Late column.** You may assume that the names of students are unique. Below is a sample of `hw_grades`.

	Name	Assignment	Grade	Late
16	Ash	Homework 7	97.734029	False
14	Ash	Homework 5	68.715955	True
9	Meg	Homework 10	88.405920	False
3	Meg	Homework 4	74.420033	True
13	Ash	Homework 4	64.538548	False

   (a) Assuming there is a late penalty that causes a 10% grade reduction to the student's current score (i.e. a 65% score would become a 65% - 6.5% = 58.5%), write a line of Pandas code to calculate all the homework grades, including the late penalty if applicable, and store it in a column named `'LPGrade'`.

   (b) Which of the following expressions outputs the students' names and number of late assignments, from least to greatest number of late assignments?

   ◯ A. `hw_grades.groupby(['Name']).sum().sort_values()`

   ◯ B. `hw_grades.groupby(['Name', 'Late']).sum().sort_values()`

◯ C. `hw_grades.groupby(['Name']).sum()['Late'].sort_values()`

◯ D. `hw_grades.groupby(['Name']).sum().sort_values()['Late']`

(c) If each assignment is weighted equally, fill in the blanks below to calculate each student's overall homework grade, including late penalties for any applicable assignments.

Hint: Recall that incomplete assignments have NaN values. How can we use `fillna` to replace these null values?

```
hw_grades._____(_____) \
          .groupby(_____)[_____] \
          .agg(_____)
```

(d) Of all the homework assignments, which are the most difficult in terms of the median grade? Order by the median grade, from lowest to greatest. Do not consider incomplete assignments or late penalties in this calculation.

Fill in the blanks below to answer this question.

Hint: Recall that incomplete assignments have NaN values. How can we use `dropna` to remove these null values?

```
hw_grades._____() \
          .groupby(_____)[_____] \
          .agg(_____) \
          .sort_values()
```

Data 100, Spring 2022 **January 4, 2022**

Discussion #3

Pandas Bootcamp

Throughout this section you'll be working with the babynames (left) and elections (right) datasets as shown below (only the first five rows are shown):

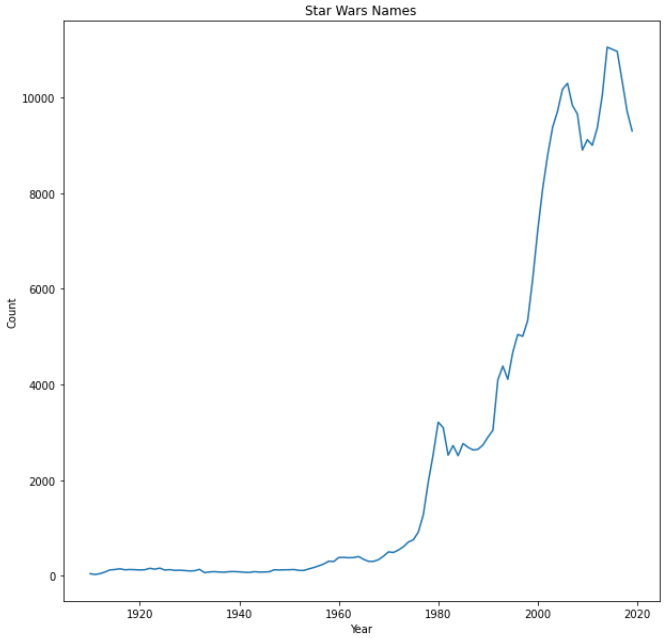
	State	Sex	Year	Name	Count
0	CA	F	1910	Mary	295
1	CA	F	1910	Helen	239
2	CA	F	1910	Dorothy	220
3	CA	F	1910	Margaret	163
4	CA	F	1910	Frances	134

	Year	Candidate	Party	Popular vote	Result	%
0	1824	Andrew Jackson	Democratic-Republican	151271	loss	57.210122
1	1824	John Quincy Adams	Democratic-Republican	113142	win	42.789878
2	1828	Andrew Jackson	Democratic	642806	win	56.203927
3	1828	John Quincy Adams	National Republican	500897	loss	43.796073
4	1832	Andrew Jackson	Democratic	702735	win	54.574789

1. (a) We perform some basic EDA on this data, and we decide to visualize the popularity of the names Luke, Leia, and Han from Star Wars to see if there is a relationship with the release of the major films with the popularity of these names.

   Fill in the blanks to output a Series that contains the year as the index and the number of total Star Wars names as the value, so we can make the plot below!

   Hint: `babynames['Name'].isin(['Helen', 'Jon'])` returns `[False, True, ...]`.



```
sw_names = ['Luke', 'Leia', 'Han']
babynames[_____________________________________] \
         .groupby(_______)_____________________ \
         .plot(ylabel = 'Count', title = 'Star Wars Names',
               figsize = (8, 8))
```

(b) Define the fluctuation of a baby name as the mathematical range of its count per year throughout its history (i.e. maximum count subtracted by minimum count). Write a line of Pandas code to determine **per-state** fluctuations for all baby names, sorted from greatest to least.

(c) Define an upset as an election result for a party that is an outlier vote share attained in that party's history. Fill in the blanks below to find all the rows in `elections` corresponding to election upsets in American history per this definition.

Hint: the `quantile` function can return the quartiles of the data; for example, `elections['%'].quantile(0.25)` returns the first quartile (Q_1). Recall that a point is an outlier if it is outside the interval $[Q_1 - 1.5\text{IQR}, Q_3 + 1.5\text{IQR}]$.

```
def outlier(subdf):
    q1, q3 = _________________________, _________________________
    iqr = ____________________________________________________
    return subdf[_____________________________________________
                 _____________________________________________]
elections.groupby(_________).apply(_________)
```

(d) Write a line of code to output a DataFrame showing the average winning and losing vote share for every party that has won an election (a sample of 5 rows are shown below).

Hint: The arguments to `pivot_table` are `index`, `columns`, `values`, and `aggfunc`.

Result	loss	win
Party		
Democratic	43.697060	51.441864
Democratic-Republican	57.210122	42.789878
National Union	NaN	54.951512
Republican	42.047791	52.366967
Whig	35.258650	50.180255

(e) Fill in the blanks below to create a new column `Middle Name` containing every candidate's middle name (or middle initial). If a candidate has no middle name, that entry should be NaN.

Hint: The default entry of any element in a DataFrame if unspecified is NaN!

```
mid = ________________________________________
elections.loc[_____________________, 'Middle Name'] = _________________
```

(f) Define election twins as two candidates that share the same middle name (or middle initial). Fill in the code below to determine the number of election twins.

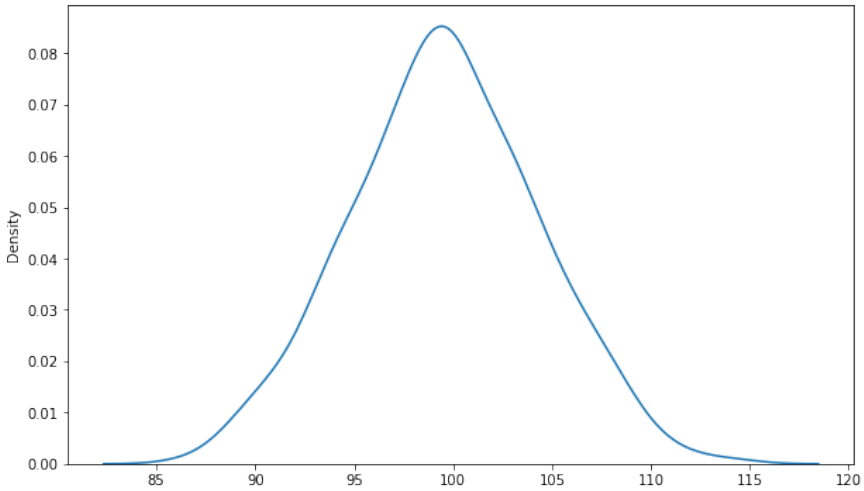
Hint: Try to use a merge, and recall that for merges with non-unique column names, Pandas will rename the non-unique column name with an `_x` suffix for the left table and `_y` suffix for the right table (i.e. for a column `col`, the resulting names would be `col_x` and `col_y`).

```
def election_twins(elections):
    elections = ____________________________________________
    twins = ________________________________________________
    twins = twins[_________________________________________]
    return len(twins)
```

Dealing with Missing Data

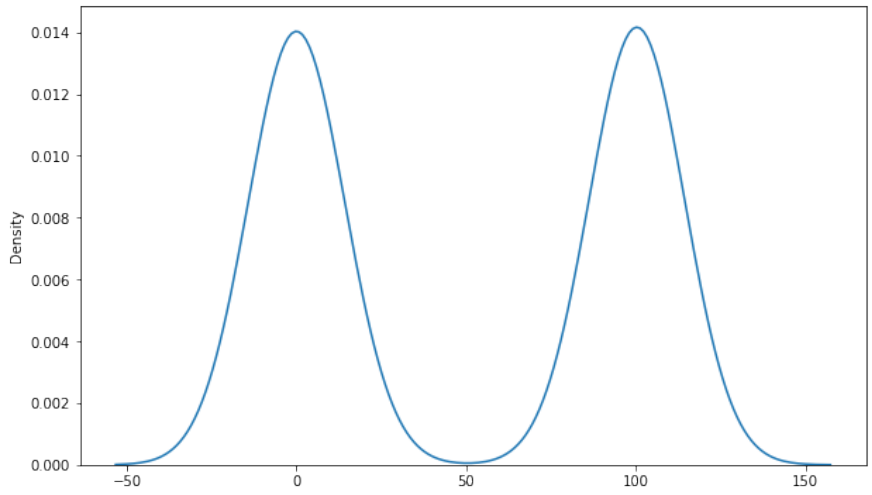
2. While working with a movie dataset from IMDB, Anirudhan realizes that nearly 20% of the votes field is missing with NaN values (however, none of the other fields have null values)! He wants to use the dataset for modeling, so he must impute or drop the missing values. Help him make the correct decision to solve the missing data problem in these subparts given the distribution of the variable.

   (a) Suppose that the distribution of this variable (i.e. `df['votes'].plot.kde()`) is given by the following figure. Which of the following techniques would be **most** effective in solving this issue of missing data?



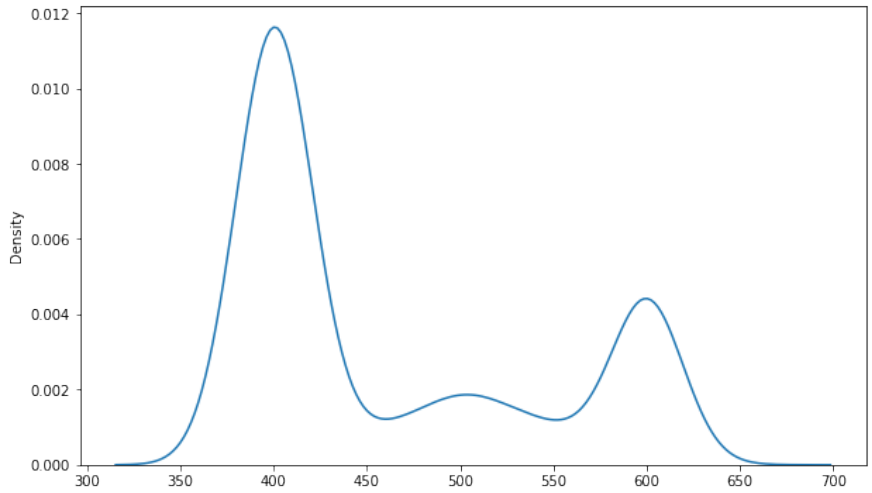
   ☐ A. Using the mean to impute the missing values
   ☐ B. Using the mode to impute the missing values
   ☐ C. Using the median to impute the missing values
   ☐ D. Dropping any rows with missing values
   ☐ E. Imputing missing values with zero

   (b) Suppose that the distribution of this variable is given by the following figure. Which of the following techniques would be **most** effective in solving this issue of missing data?



☐ A. Using the mean to impute the missing values
☐ B. Using the mode to impute the missing values
☐ C. Using the median to impute the missing values
☐ D. Dropping any rows with missing values
☐ E. Imputing missing values with zero

(c) Suppose that the distribution of this variable is given by the following figure. Which of the following techniques would be **reasonably** effective in solving this issue of missing data?



☐ A. Using the mean to impute the missing values
☐ B. Using the mode to impute the missing values
☐ C. Using the median to impute the missing values
☐ D. Dropping any rows with missing values
☐ E. Imputing missing values with zero

Pandas: Olympics (Bonus)

3. We will work with an Olympics dataset containing the names of all athletes who participated in the Olympic Games, including all the Games from Athens 1896 to Tokyo 2020. The first 5 lines of the table are shown below. You may assume that the ID column is the primary key of the table and that the only column with null values are height, weight, and medal.

	ID	Name	Sex	Age	Height	Weight	Team	Year	Season	City	Sport	Event	Medal
0	1	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
1	2	Minna Maarit Aalto	F	30.0	159	55.5	Finland	1996	Summer	Atlanta	Sailing	Sailing Women's Windsurfer	NaN
2	3	Minna Maarit Aalto	F	34.0	159	55.5	Finland	2000	Summer	Sydney	Sailing	Sailing Women's Windsurfer	NaN
3	4	Kjetil Andr Aamodt	M	20.0	176	85	Norway	1992	Winter	Albertville	Alpine Skiing	Alpine Skiing Men's Super G	Gold
4	5	Ragnhild Margrethe Aamodt	F	27.0	163	NaN	Norway	2008	Summer	Beijing	Handball	Handball Women's Handball	Gold

(a) Write a line of Pandas code to determine the 10 most common middle names among gold medal winners. You may assume that all athletes in the table have a middle name.

(b) What are the oldest athletes to participate in each sport along with the corresponding year in which they participated? Fill in the blanks below to answer the question.

```
ath.groupby(_____) \
   .apply(_____) \
   [['Name', 'Year']]
```

(c) Fill in the blanks below to return the names of all the athletes who won a medal after a gap of 10 years or more of not winning any Olympics medals. You may assume that each individual's name is unique to them.

```
def filter_func(subframe):
    return ____________________________________________________

ath.sort_values(_____________________) \
   [_____________________] \
   .groupby(_________) \
   .filter(filter_func)['Name'] \
   .unique()
```

(d) For all athletes that performed in more than 2 Olympics, which athletes had a **better** Olympics debut year than any of their subsequent performances in terms of total medal count? Sort by least difference in medal count from their debut year from any of their performances from subsequent years, from greatest to least.

Hint: The instructor solution uses `groupby` three times (you may not need to though)!

```
def f1(s):
    return ____________________________________________
def f2(s):
    ____________________________________________
    return ____________________________________________
____________________________________________
____________________________________________
____________________________________________
____________________________________________
____________________________________________
names = ____________________________________________
```

Data 100, Spring 2022 **February 11, 2022**

Discussion #4

Regular Expressions

Here's a complete list of metacharacters:

`.  ^  $  *  +  ?  { }  [ ]  \  |  ( )`

Some reminders on what each can do (this is not exhaustive):

- **`"^"`** matches the position at the beginning of string (unless used for negation `"[^]"`)
- **`"$"`** matches the position at the end of string character.
- **`"?"`** match preceding literal or sub-expression 0 or 1 times.
- **`"+"`** match preceding literal or sub-expression *one* or more times.
- **`"*"`** match preceding literal or sub-expression *zero* or more times
- **`"."`** match any character except new line.
- **`"[ ]"`** match any one of the characters inside, accepts a range, e.g., `"[a-c]"`.
- **`"( )"`** used to create a sub-expression
- **`"\d"`** match any *digit* character. `"\D"` is the complement.
- **`"\w"`** match any *word* character (letters, digits, underscore). `"\W"` is the complement.
- **`"\s"`** match any *whitespace* character including tabs and newlines. `\S` is the complement.
- **`"*?"`** Non-greedy version of *. Not fully discussed in class.
- **`"\b"`** match boundary between words. Not discussed in class.
- **`"+?"`** Non-greedy version of +. Not discussed in class.
- **`"{m,n}"`** The preceding element or subexpression must occur between m and n times, inclusive.

Some useful `re` package functions:

- **`re.split(pattern, string)`** split the `string` at substrings that match the `pattern`. Returns a list.
- **`re.sub(pattern, replace, string)`** apply the `pattern` to `string` replacing matching substrings with `replace`. Returns a string.
- **`re.findall(pattern, string)`** Returns a list of all matches for the given `pattern` in the `string`.

Regular Expressions

1. Which strings contain a match for the following regular expression, `"1+1$"`? The character `"␣"` represents a single space.

   ◯ A. `What␣is␣1+1` ◯ B. `Make␣a␣wish␣at␣11:11` ◯ C. `111␣Ways␣to␣Succeed`

2. Write a regular expression that matches strings (including the empty string) that only contain lowercase letters and numbers.

3. Given `sometext = "I've␣got␣10␣eggs,␣20␣gooses,␣and␣30␣giants."`, use `re.findall` to extract all the items and quantities from the string. The result should look like **`['10 eggs', '20 gooses', '30 giants']`**. You may assume that a space separates quantity and type, and that each item ends in s.

4. For each pattern specify the starting and ending position of the first match in the string. The index starts at zero and we are using closed intervals (both endpoints are included).

	`abcdefg`	`abcs!`	`ab␣abc`	`abc,␣123`
`abc*`	$[0, 2]$	_____	_____	_____
`[^\s]+`	_____	_____	_____	_____
`ab.*c`	_____	_____	_____	_____
`[a-z1,9]+`	_____	_____	_____	_____

5. *(Bonus)* Given the following text in a variable `log`:

```
169.237.46.168 - - [26/Jan/2014:10:47:58 -0800]
"GET␣/stat141/Winter04/␣HTTP/1.1" 200 2585
"http://anson.ucdavis.edu/courses/"
```

   Fill in the regular expression in the variable `pattern` below so that after it executes, day is 26, month is Jan, and year is 2014.

```
pattern = ...
matches = re.findall(pattern, log)
day, month, year = matches[0]
```

6. *(Bonus)* Given that `sometext` is a string, use `re.sub` to replace all clusters of non-vowel characters with a single period. For example `"a␣big␣moon,␣between␣us..."` would be changed to `"a.i.oo.e.ee.u."`.

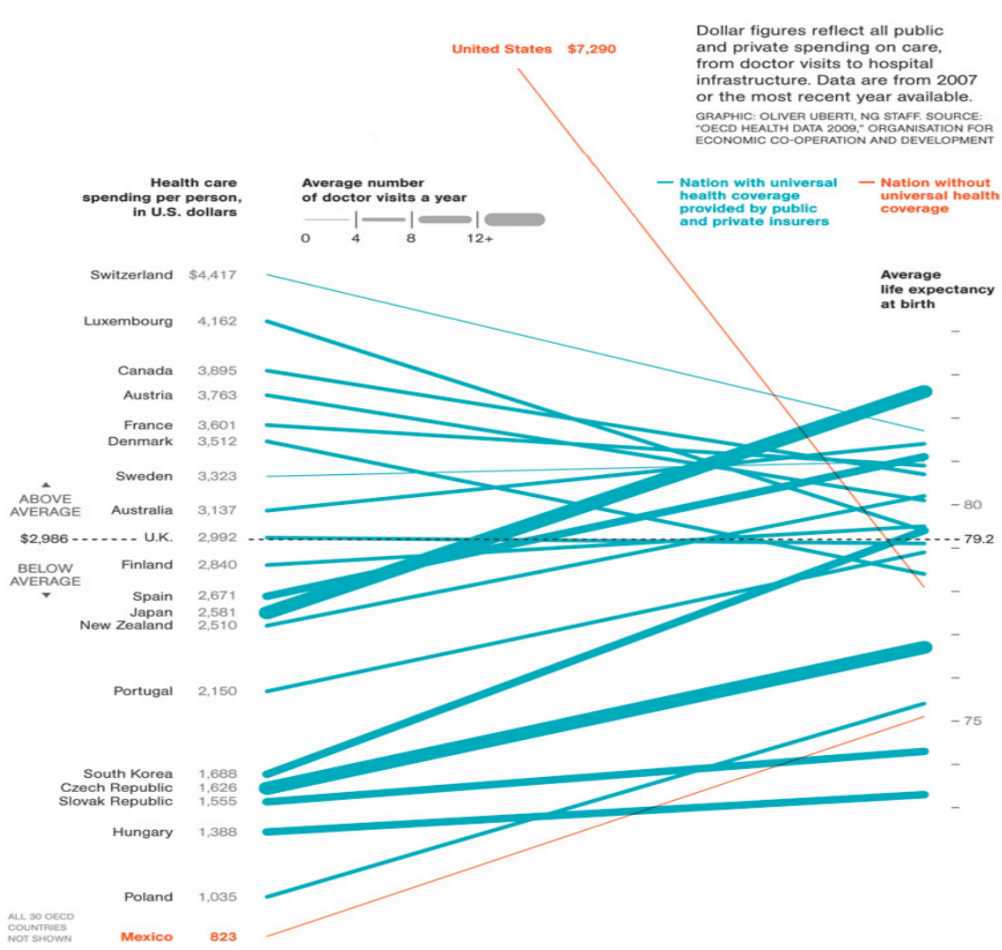
7. *(Bonus)* Given the text:

   ```
   "<record>␣Josh␣Hug␣<hug@cs.berkeley.edu>␣Faculty␣</record>"
   "<record>␣Lisa␣Yan␣<lisa.yan@berkeley.edu>␣Instructor␣</record>"
   ```

   Which of the following matches exactly to the email addresses (including angle brackets)?

   ◯ A. `<.*@.*>` ◯ B. `<[^>]*@[^>]*>` ◯ C. `<.*@\w+\..*>`

Data Visualization



8. The first part of the discussion will be centered on the above visualization.

   (a) Five variables are being represented visually in this graphic. What are they and what are their feature types (ie qualitative, quantitative, nominal, ordinal)?

   (b) How are the variables represented in the graphic, e.g., the variable `XXX` is mapped to the x-axis, the variable `WWW` is mapped to the y-axis, the variable `ZZZ` is conveyed through color, etc.?

(c) How can we figure out how to interpret the visual qualities of the plot, e.g., how do we know what a color represents?

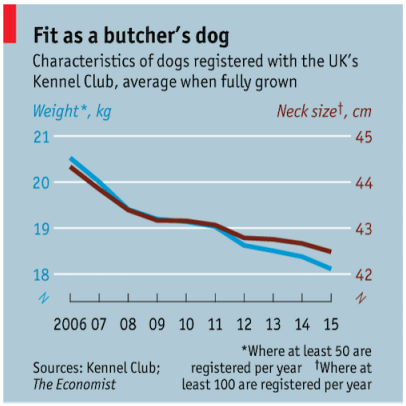
(d) What purpose does the comment at the top right of the plot serve?

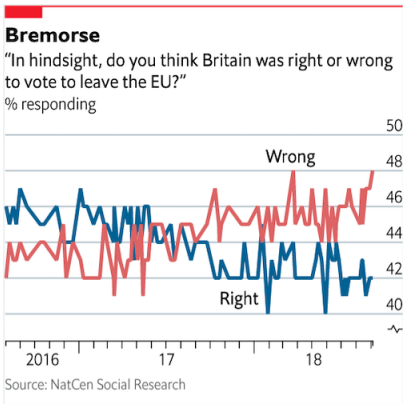
(e) Make 3 observations about the figure. Describe the feature that you are basing your observation on.

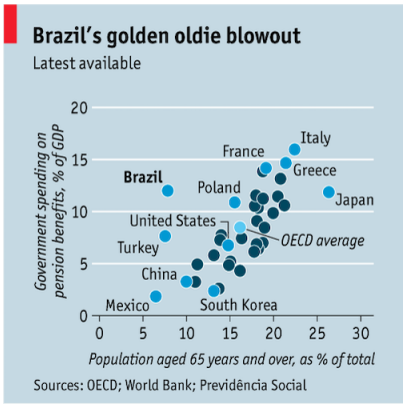
For example, South Korea's expenditure on health care is comparable to Eastern European countries (and among the lowest of all countries plotted), but the life expectancy is much higher than the Eastern European countries. In the plot we see that the left endpoint of South Korea's line segment is near the Eastern European countries, but the slope of the line segment is much steeper.

(f) Consider the steep negative slope and narrowness of the line segment that represents the data for the United States. What systemic, social, or societal issues might explain this?

9. Creating visualizations that represent data accurately and that support the narrative we wish to create is no easy task. Even the journalists and editors at *The Economist*, a newspaper known for it's compelling, data-driven articles, have been known to make blunders. Three of their ill-thought-out plots are presented below. Consider what aspects of the visualizations are misleading, and think of ways in which you can remedy them.







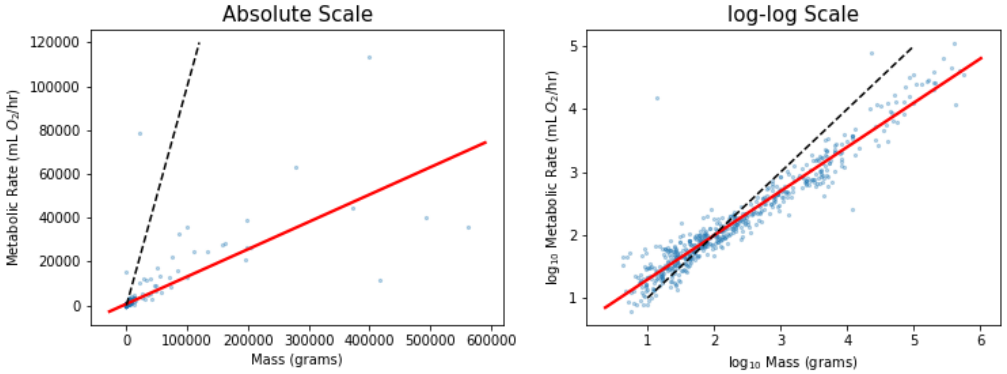
Hint: The datapoints in the rightmost plot are shaded based on whether or not they are labeled.

Data 100, Spring 2022 **February 18, 2022**

Discussion #5

Logarithmic Transformations

1. One of your friends at a biology lab asks you to help them analyze `panTHERIA`, a database of mammals. They are interested in the relationship between mass, measured in grams, and metabolic rate ("energy expenditure"), measured by oxygen use per hour. Originally, they show you the data on a linear (absolute) scale, shown on the left. You notice that the values on both axes vary over a large range with many data points clustered around the smaller values, so you suggest that they instead plot the data on a log-log scale, shown on the right. The solid red line is a "line of best fit" (we'll formalize this later in the course) while the black dashed line represents the identity line $y = x$.



(a) Let C and k be some constants and x and y represent mass and metabolic rate, respectively. Based on the plots, which of the following best describe the pattern seen in the data? Reminder: $log(ab) = log(a) + log(b)$.

◯ A. $y = C + kx$ ◯ B. $y = C \times 10^{kx}$ ◯ C. $y = C + k \log_{10}(x)$ ◯ D. $y = Cx^k$

(b) What parts of the plots could you use to make initial guesses on C and k?

(c) Your friend points to the solid line on the log-log plot and says "since this line is going up and to the right, we can say that, in general, the bigger a mammal is, the greater its metabolic rate". Is this a reasonable interpretation of the plot?

Kernel Density Estimation

2. We wish to compare the results of kernel density estimation using a Gaussian kernel and a boxcar kernel. For $\alpha > 0$, which of the following statements are true? Choose all that apply.

   Gaussian Kernel:

$$K_\alpha(x, z) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{(x-z)^2}{2\alpha^2}\right)$$

   Boxcar Kernel:

$$B_\alpha(x, z) = \begin{cases} \frac{1}{\alpha} & \text{if } -\frac{\alpha}{2} \leq x - z \leq \frac{\alpha}{2} \\ 0 & \text{else} \end{cases}$$

   A. Decreasing α for a Gaussian kernel decreases the smoothness of the KDE.

   B. The Gaussian kernel is always better than the boxcar kernel for KDEs.

   C. Because the Gaussian kernel is smooth, we can safely use large α values for kernel density estimation without worrying about the actual distribution of data.

   D. The area under the boxcar kernel is 1, regardless of the value of α.

   E. None of the above.

Driving with a Constant Model

3. Adam is trying to use modeling to drive his car autonomously. To do this, he collects a lot of data where he drives around his neighborhood, and he wants your help to design a model that can drive on his behalf in the future using the outputs of the models you design. We will tackle two aspects of this autonomous car modeling framework: going forward and turning.

   We show some statistics from the collected dataset below using `pd.describe`, which returns the mean, standard deviation, quartiles, minimum and maximum for the two columns in the dataset: `target_speed` and `degree_turn`.

	target_speed	degree_turn
count	500.000000	500.000000
mean	32.923408	143.721153
std	46.678744	153.641504
min	0.231601	0.000000
25%	12.350025	6.916210
50%	25.820689	45.490086
75%	39.788716	323.197168
max	379.919965	359.430309

   (a) Suppose the first part of the model predicts the target speed of the car. Using constant models trained on the speeds on the collected data shown above with L_1 and L_2 loss functions, which of the following is true?

   ◯ A. The model trained with the L_1 loss will always drive slower than the model trained with L_2 loss.

   ◯ B. The model trained with the L_2 loss will always drive slower than the model trained with L_1 loss.

   ◯ C. The model trained with the L_1 loss will sometimes drive slower than the model trained with L_2 loss.

   ◯ D. The model trained with the L_2 loss will sometimes drive slower than the model trained with L_1 loss.

   (b) Finding that the model trained with the L_2 loss drives too slowly, Adam changes the loss function for the constant model where the loss is penalized **more** if the speed is higher.

That way, the model wants to optimize more for the case where we wish to drive faster since the loss is higher, accomplishing his goal.

Find the optimal $\hat{\theta}$ for the constant model using the new loss function below:

$$L(\theta) = \frac{1}{n} \sum_i y_i (y_i - \theta)^2$$

(c) Suppose he is working on a model that predicts the degree of turning at a particular time between 0 and 359 degrees using the data in the `degree_turn` column. Explain why a constant model is likely inappropriate in this use case.

Extra: If you've studied some physics, you may recognize the behaviour of our constant model!

(d) Suppose we finally expand our modeling framework to use simple linear regression (i.e. $f_\theta(x) = \theta_{w,0} + \theta_{w,1} x$). For our first simple linear regression model, we predict the turn angle (y) using target speed (x). Our optimal parameters are: $\hat{\theta}_{w,1} = 0.019$ and $\hat{\theta}_{w,0} = 143.1$.

However, we realize that we actually want a model that predicts target speed (our new y) using turn angle, our new x (instead of the other way around)! What are our new optimal parameters for this new model?

Standardized SLR (Bonus)

4. We will experiment with standardizing the x (explanatory) and y (response) variables for a simple linear model with and without an intercept term (i.e. $f_\theta(x) = \theta x$). Recall that standardizing a variable V involves subtracting its mean $\bar{V}$ and dividing by its standard deviation σ_V as follows: $\tilde{V} = \frac{V - \bar{V}}{\sigma_V}$.

   (a) What is the optimal solution for θ with $f_\theta(x) = \theta x$ with a standard MSE loss for standardized x and y?

   (b) Calculate the optimal θ_0 and θ_1 for an SLR model with an intercept using if x and y were standardized.

   (c) Show that the optimal θ for a linear model without an intercept (i.e. $f_\theta(x) = \theta x$) is the same as the optimal θ for standardized SLR from the previous subpart.

   Hint: Given that the variance of the values in a vector $\vec{v}$ is $\sigma_v^2 = \frac{1}{n}\sum_i (v_i - \bar{v})^2$, simplify the denominator of the optimal θ from part (a).

Data 100, Spring 2022 **February 25, 2022**

Discussion #6

Linear Models

1. Which of the following models are linear? Select all that apply.

   ☐ A. $\hat{y} = \theta_1 x + \theta_2 \sin(x)$

   ☐ B. $\hat{y} = \theta_1 x + \theta_2 \sin(x^2)$

   ☐ C. $\hat{y} = \theta_1$

   ☐ D. $\hat{y} = (\theta_1 x + \theta_2)\, x$

   ☐ E. $\hat{y} = \ln(\theta_1 x + \theta_2) + \theta_3$

2. Which of the following are true about the optimal solution $\hat{\theta}$ to ordinary least squares (OLS)? Recall that the least squares estimate $\hat{\theta}$ solves the normal equation $(\mathbb{X}^T\mathbb{X})\theta = \mathbb{X}^T\mathbb{Y}$.

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

   ☐ A. Using the normal equation, we can derive an optimal solution for simple linear regression with an L_2 loss.

   ☐ B. Using the normal equation, we can derive an optimal solution for simple linear regression with an L_1 loss.

   ☐ C. Using the normal equation, we can derive an optimal solution for a constant model with an L_2 loss.

   ☐ D. Using the normal equation, we can derive an optimal solution for a constant model with an L_1 loss.

   ☐ E. Using the normal equation, we can derive an optimal solution for the model specified option B in question 1 ($\hat{y} = \theta_1 x + \theta_2 \sin(x^2)$).

3. Which of the following conditions are required for the least squares estimate in Question 2?

   ☐ A. $\mathbb{X}$ must be full column rank.

   ☐ B. $\mathbb{Y}$ must be full column rank.

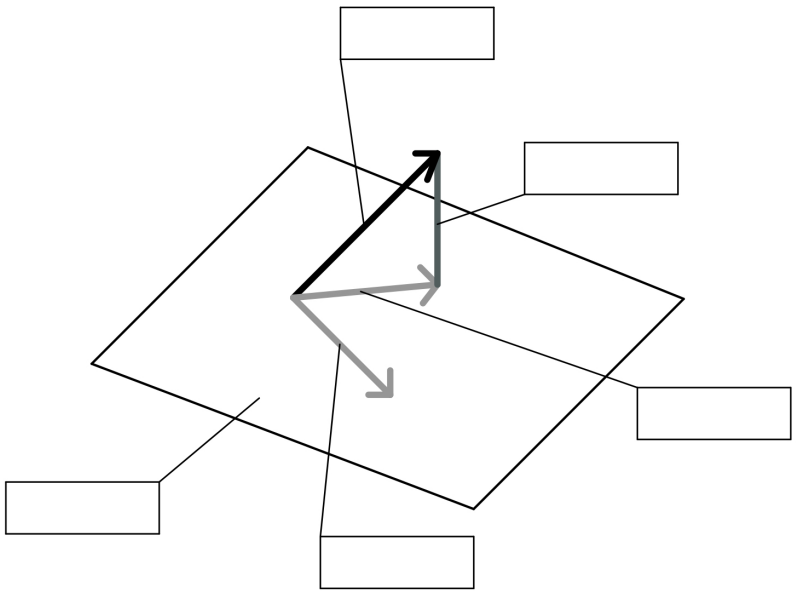
   ☐ C. $\mathbb{X}$ must be invertible.

   ☐ D. $\mathbb{X}^{\mathbb{T}}$ must be invertible.

Geometry of Least Squares

4. Suppose we have a dataset represented with the design matrix span($\mathbb{X}$) and response vector $\mathbb{Y}$. We use linear regression to solve for this and obtain optimal weights as $\hat{\theta}$. Label the following terms on the geometric interpretation of ordinary least squares:

   - $\mathbb{X}$ (i.e., span($\mathbb{X}$))
   - The response vector $\mathbb{Y}$
   - The residual vector $\mathbb{Y} - \mathbb{X}\hat{\theta}$
   - The prediction vector $\mathbb{X}\hat{\theta}$ (using optimal parameters)
   - A prediction vector $\mathbb{X}\alpha$ (using an arbitrary vector α).



(a) What is always true about the residuals in least squares regression? Select all that apply.

☐ A. They are orthogonal to the column space of the design matrix.
☐ B. They represent the errors of the predictions.
☐ C. Their sum is equal to the mean squared error.
☐ D. Their sum is equal to zero.
☐ E. None of the above.

(b) Which are true about the predictions made by OLS? Select all that apply.

☐ A. They are projections of the observations onto the column space of the design matrix.
☐ B. They are linear combinations of the features.
☐ C. They are orthogonal to the residuals.
☐ D. They are orthogonal to the column space of the features.

☐ E. None of the above.

(c) We fit a simple linear regression to our data $(x_i, y_i), i = 1, 2, 3$, where x_i is the independent variable and y_i is the dependent variable. Our regression line is of the form $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$. Suppose we plot the relationship between the residuals of the model and the $\hat{y}s$, and find that there is a curve. What does this tell us about our model?

☐ A. The relationship between our dependent and independent variables is well represented by a line.

☐ B. The accuracy of the regression line varies with the size of the dependent variable.

☐ C. The variables need to be transformed, or additional independent variables are needed.

(d) Which are the following is true of the mystery quantity $\vec{v} = (I - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T)\mathbb{Y}$?

☐ A. The vector $\vec{v}$ represents the residuals for any linear model.

☐ B. If the $\mathbb{X}$ matrix contains the $\vec{1}$ vector, then the sum of the elements in vector $\vec{v}$ is 0 (i.e. $\sum_i v_i = 0$).

☐ C. All the column vectors x_i of $\mathbb{X}$ are orthogonal to $\vec{v}$.

☐ D. If $\mathbb{X}$ is of shape n by p, there are p elements in vector $\vec{v}$.

☐ E. For any α, $\mathbb{X}\alpha$ is orthogonal to $\vec{v}$.

Linear Regression Fundamentals (Extra)

5. In this problem, we will review some of the core concepts in linear regression.

   Suppose we create a linear model with parameters $\hat{\theta} = [\hat{\theta_0}, \ldots, \hat{\theta_p}]$. As we saw in lecture, given an observation $\vec{x}$, such a model makes predictions $\hat{y} = \hat{\theta} \cdot \vec{x} = \hat{\theta_0} + \hat{\theta_1} x_1 + \hat{\theta_2} x_2 + \ldots + \hat{\theta_p} x_p$.

   (a) Suppose $\hat{\theta} = [2, 0, 1]$ and we receive an observation $\vec{x_1} = [1, 2, 3]$. What $\hat{y_1}$ value will this model predict for the given observation?

   (b) Suppose the true y_1 was 3.5. What will be the L_2 loss for our prediction $\hat{y_1}$ from the previous part?

   (c) Suppose we receive another observation $\vec{x_2} = [1, 5, 1]$. What $\hat{y_2}$ value will this model predict for the given observation?

   (d) Suppose the true y_2 was 4. What will be the mean squared error of our model, given the two observations?

Data 100, Spring 2022 **March 4, 2022**

Discussion #7

Ordinary Debugging

1. Anirudhan is fitting a multiple linear regression model with Scikit-learn, but he is having a few bugs and issues along the way. Help him debug his code and his logic!

   (a) Suppose he runs the code below to fit on design matrix X of shape 250 by 3 with corresponding response variable y of shape 250. We wish to use our model to predict on a new dataset X_t with 50 data points, storing the predictions in a variable `final_predictions`. What are 2 potential issues with this code?

```
model = LinearRegression(fit_intercept = False)
final_predictions = model.predict(X_t)
model.fit(X_t, y)
```

   (b) Suppose he forgets about the dataset X_t and wishes to focus only on dataset X. Realizing he did not use an intercept term in part (a), he decides to add one using the `add_intercept` function from the lab. What are 2 potential issues with this new code?

   Note: one of these may not break Scikit-learn, but it's an issue nevertheless!

```
def add_intercept(X):
    # Concatenates "ones" vector to design matrix X
    return np.concatenate([X, np.ones(shape = (n, 1))],
                          axis = 1)
model = LinearRegression()
n, p = X.shape
model.fit(add_intercept(X), y)
final_predictions = model.predict(X)
```

Dive into Gradient Descent

2. Given the following loss function and $\vec{x} = [x_i]_{i=1}^n$, $\vec{y} = [y_i]_{i=1}^n$, and $\theta^{(t)}$, explicitly write out the update equation for $\theta^{(t+1)}$ in terms of x_i, y_i, $\theta^{(t)}$, and α, where $\alpha = 0.5$ is the constant learning rate.

$$L(\theta, \vec{x}, \vec{y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\theta^2 x_i - \log(y_i) \right)$$

Bonus: As $t \to \infty$, what are the required conditions for $\theta^{(t)}$ to converge? To what can it converge?

3. We want to minimize the loss function $L(\theta) = (\theta_1 - 1)^2 + |\theta_2 - 3|$. While you may notice that this function is not differentiable anywhere, we can still use gradient descent wherever the function *is* differentiable!

   Recall that for a function $f(x) = k|x|$, $\frac{df}{dx} = k$ for all $x > 0$ and $\frac{df}{dx} = -k$ for all $x < 0$.

   (a) What are the optimal values $\hat{\theta}_1$ and $\hat{\theta}_2$ to minimize $L(\theta)$? At that point $\hat{\theta}$, what is the gradient ∇L?

   (b) Suppose we initialize our gradient descent algorithm randomly at $\theta_1 = 2$ and $\theta_2 = 5$. Calculate the gradient $\nabla L = \left[\frac{\partial L}{\partial \theta_1} \quad \frac{\partial L}{\partial \theta_2} \right]^T \Big|_{\theta_1 = 2, \theta_2 = 5}$ at the specified θ_1 and θ_2 values.

   (c) Apply the first gradient update with a learning rate $\alpha = 0.5$. In other words, calculate $\theta_1^{(1)}$ and $\theta_2^{(1)}$ using the initializations $\theta_1^{(0)} = 2$ and $\theta_2^{(0)} = 5$.

(d) How many gradient steps does it take for θ_1 and θ_2 to converge to their optimal values obtained in part (a) assuming we keep a constant learning rate of $\alpha = 0.5$?

Hint: After part (c), what is the derivative $\frac{\partial L}{\partial \theta_1}$ evaluated at $\theta_1^{(1)}$?

The Cook County Housing Dataset

4. In Project 1 we will work with real world housing data from Cook County, Illinois. Analyze the dataframe on the next page and address the following questions:

   (a) Based on the columns presented in this dataset and the values that they take, what do you think each row represents? That is, what is the granularity of this dataset?

   (b) Why do you think this dataset was collected? For what purposes? By whom? This question calls for your speculation and is looking for thoughtfulness, not correctness.

   (c) Certain variables in this dataset contain information that either directly contains demographic information (data on people) or could when linked to other datasets. Identify at least one and explain the nature of the demographic data it embeds.

   (d) Craft at least two questions about housing in Cook County that can be answered with this dataset and provide the type of analytical tool you would use to answer it (e.g. "*I would create a plot of . . . and . . .*" or "*I would calculate the* [summary statistic] *for . . . and . . .*"). Be sure to reference the columns that you would use and any additional datasets you would need to answer that question.

Cook County Assessor's Office Dataset (204792 rows x 62 columns): 5 randomly sampled rows

	PIN	Property Class	Neighborhood Code	Land Square Feet	Town Code	Apartments	Wall Material	Roof Material	Basement	Basement Finish	Central Heating	Other Heating
59751	20302160680000	203	121	3750.0	72	0.0	2.0	1.0	1.0	3.0	2.0	5.0
198610	6222130070000	207	13	5640.0	18	0.0	1.0	1.0	2.0	3.0	1.0	5.0
143557	13263020230000	203	70	3125.0	71	0.0	2.0	1.0	1.0	1.0	1.0	5.0
107891	30201100080000	202	101	6600.0	37	0.0	1.0	1.0	1.0	3.0	1.0	5.0
109977	5321220050000	203	150	9864.0	23	0.0	2.0	1.0	1.0	3.0	1.0	5.0

	Central Air	Fireplaces	Attic Type	Attic Finish	Design Plan	Cathedral Ceiling	Construction Quality	Site Desirability	Garage 1 Size	Garage 1 Material	Garage 1 Attachment	Garage 1 Area
	0.0	0.0	3.0	0	0.0	0.0	2.0	2.0	4.0	1.0	2.0	2.0
	1.0	0.0	3.0	0	2.0	0.0	2.0	2.0	3.0	1.0	1.0	1.0
...	1.0	0.0	3.0	0	2.0	0.0	2.0	2.0	3.0	1.0	2.0	2.0
	0.0	0.0	3.0	0	2.0	0.0	2.0	2.0	3.0	1.0	2.0	2.0
	1.0	1.0	2.0	3	2.0	2.0	2.0	2.0	2.0	2.0	1.0	2.0

	Garage 2 Size	Garage 2 Material	Garage 2 Attachment	Garage 2 Area	Porch	Other Improvements	Building Square Feet	Repair Condition	Multi Code	Number of Commercial Units	Estimate (Land)	Estimate (Building)
	7.0	0.0	0.0	0.0	3	0.0	1077.0	2.0	2.0	0.0	22500	70770
	7.0	0.0	0.0	0.0	3	0.0	1902.0	2.0	2.0	0.0	29610	212670
...	7.0	0.0	0.0	0.0	3	0.0	1260.0	2.0	2.0	0.0	43750	254010
	7.0	0.0	0.0	0.0	3	0.0	952.0	2.0	2.0	0.0	21450	61680
	7.0	0.0	0.0	0.0	3	0.0	1307.0	2.0	2.0	0.0	103570	305320

	Deed No.	Sale Price	Longitude	Latitude	Census Tract	Multi Property Indicator	Modeling Group	Age	Use	O'Hare Noise	Floodplain	Road Proximity
	1710934064	1	-87.673073	41.760296	672000.0	0	SF	58	1	0.0	0.0	0.0
	1318416011	173500	-88.188488	42.024901	804310.0	0	SF	24	1	0.0	0.0	0.0
...	1624234087	372500	-87.722925	41.930439	220702.0	0	SF	24	1	0.0	0.0	0.0
	1321946032	56000	-87.535489	41.596277	826202.0	0	SF	56	1	0.0	0.0	0.0
	1801946026	394875	-87.743717	42.072883	800900.0	0	SF	60	1	0.0	0.0	0.0

	Sale Year	Sale Quarter	Sale Half-Year	Sale Quarter of Year	Sale Month of Year	Sale Half of Year	Most Recent Sale	Age Decade	Pure Market Filter	Garage Indicator	Neigborhood Code (mapping)	Town and Neighborhood	Description	Lot Size
	2017	82	41	2	4	1	1.0	5.8	0	1.0	121	72121	This property, sold on 04/19/2017, is a one-st...	3750.0
	2013	67	34	3	7	2	1.0	2.4	1	1.0	13	1813	This property, sold on 07/03/2013, is a two-st...	5640.0
...	2016	79	40	3	8	2	1.0	2.4	1	1.0	70	7170	This property, sold on 08/29/2016, is a one-st...	3125.0
	2013	67	34	3	8	2	1.0	5.6	1	1.0	101	37101	This property, sold on 08/07/2013, is a one-st...	6600.0
	2018	85	43	1	1	1	0.0	6.0	1	1.0	150	23150	This property, sold on 01/19/2018, is a one-st...	9864.0

Dummy Variables/One-hot Encoding (Bonus)

In order to include a qualitative variable in a model, we convert it into a collection of dummy variables. These dummy variables take on only the values 0 and 1. For example, suppose we have a qualitative variable with 3 possible values, call them A, B, and C, respectively. For concreteness, we use a specific example with 10 observations:

$$[A, A, A, A, B, B, B, C, C, C]$$

We can represent this qualitative variable with 3 dummy variables that take on values 1 or 0 depending on the value of this qualitative variable. Specifically, the values of these 3 dummy variables for this dataset are x_A, x_B, and x_C, arranged from left to right in the following design matrix, where we use the following indicator variable:

$$x_{k,i} = \begin{cases} 1 & \text{if } i\text{-th observation has value } k \\ 0 & \text{otherwise.} \end{cases}$$

This representation is also called one-hot encoding. It should be noted here that $\vec{x_A}$, $\vec{x_B}$, and $\vec{x_C}$ are all vectors.

$$\mathbb{X} = \begin{bmatrix} | & | & | \\ \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

We will show that the fitted coefficients for $\vec{x_A}$, $\vec{x_B}$, and $\vec{x_C}$ are $\bar{y}_A$, $\bar{y}_B$, and $\bar{y}_C$, the average of the y_i values for each of the groups, respectively.

5. Show that the columns of $\mathbb{X}$ are orthogonal, (i.e., the dot product between any pair of column vectors is 0).

6. Show that

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

Here, n_A, n_B, n_C are the number of observations in each of the three groups defined by the levels of the qualitative variable.

7. Show that

$$\mathbb{X}^T\mathbb{Y} = \begin{bmatrix} \sum_{i \in A} y_i \\ \sum_{i \in B} y_i \\ \sum_{i \in C} y_i \end{bmatrix}$$

where i is an element in group A, B, or C.

8. Use the results from the previous questions to solve the normal equations for $\hat{\theta}$, i.e.,

$$\begin{aligned} \hat{\theta} &= [\mathbb{X}^T\mathbb{X}]^{-1}\mathbb{X}^T\mathbb{Y} \\ &= \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} \end{aligned}$$

9. *(Extra)* Show that if you augment your $\mathbb{X}$ matrix with an additional $\vec{1}$ bias vector as shown below, $\mathbb{X}^T\mathbb{X}$ is not full rank. Conclude that the new $\mathbb{X}^T\mathbb{X}$ is not invertible, and we cannot use the least squares estimate in this situation.

Hint: Use the original computation of this matrix from question 6 to help you!

$$\mathbb{X} = \begin{bmatrix} | & | & | & | \\ \vec{1} & \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | & | \end{bmatrix}$$

Data 100, Spring 2022 **March 11, 2022**

Discussion #8

Fairness in Housing Appraisal

Recall that Project 1's dataset comes from the Cook County Assessor's Office (CCAO) in Illinois, a government institution that determines property taxes across most of Chicago's metropolitan area and its nearby suburbs. In the United States, all property owners are required to pay property taxes, which are then used to fund public services including education, road maintenance, and sanitation.

1. "How much is a house worth?" If you were a homeowner, why would you want your property to be valued high? Why would you want your property to be valued low?

2. Which of the following scenarios strike you as unfair and why? You can choose more than one. There is no single right answer but you must explain your reasoning.

    A. A homeowner whose home is assessed at a higher price than it would sell for.
    B. A homeowner whose home is assessed at a lower price than it would sell for.
    C. An assessment process that systematically overvalues inexpensive properties and undervalues expensive properties.
    D. An assessment process that systematically undervalues inexpensive properties and overvalues expensive properties.

3. Imagine your home is assessed at a higher value than you believe it would sell for on the market. What might that concretely mean to you, as an individual homeowner?

Dummy Variables/One-hot Encoding

In order to include a qualitative variable in a model, we convert it into a collection of dummy variables. These dummy variables take on only the values 0 and 1. For example, suppose we have a qualitative variable with 3 possible values, call them A, B, and C, respectively. For concreteness, we use a specific example with 10 observations:

$$[A, A, A, A, B, B, B, C, C, C]$$

We can represent this qualitative variable with 3 dummy variables that take on values 1 or 0 depending on the value of this qualitative variable. Specifically, the values of these 3 dummy variables for this dataset are x_A, x_B, and x_C, arranged from left to right in the following design matrix, where we use the following indicator variable:

$$x_{k,i} = \begin{cases} 1 & \text{if } i\text{-th observation has value } k \\ 0 & \text{otherwise.} \end{cases}$$

This representation is also called one-hot encoding. It should be noted here that $\vec{x_A}$, $\vec{x_B}$, and $\vec{x_C}$ are all vectors.

$$\mathbb{X} = \begin{bmatrix} | & | & | \\ \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

We will show that the fitted coefficients for $\vec{x_A}$, $\vec{x_B}$, and $\vec{x_C}$ are $\bar{y}_A$, $\bar{y}_B$, and $\bar{y}_C$, the average of the y_i values for each of the groups, respectively.

4. Show that the columns of $\mathbb{X}$ are orthogonal, (i.e., the dot product between any pair of column vectors is 0).

5. Show that

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

   Here, n_A, n_B, n_C are the number of observations in each of the three groups defined by the levels of the qualitative variable.

6. Show that

$$\mathbb{X}^T\mathbb{Y} = \begin{bmatrix} \sum_{i \in A} y_i \\ \sum_{i \in B} y_i \\ \sum_{i \in C} y_i \end{bmatrix}$$

   where i is an element in group A, B, or C.

7. Use the results from the previous questions to solve the normal equations for $\hat{\theta}$, i.e.,

$$\begin{aligned} \hat{\theta} &= [\mathbb{X}^T\mathbb{X}]^{-1}\mathbb{X}^T\mathbb{Y} \\ &= \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} \end{aligned}$$

8. *(Extra)* Show that if you augment your $\mathbb{X}$ matrix with an additional $\vec{1}$ bias vector as shown below, $\mathbb{X}^T\mathbb{X}$ is not full rank. Conclude that the new $\mathbb{X}^T\mathbb{X}$ is not invertible, and we cannot use the least squares estimate in this situation.

   Hint: Use the original computation of this matrix from question 6 to help you!

$$\mathbb{X} = \begin{bmatrix} | & | & | & | \\ \vec{1} & \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | & | \end{bmatrix}$$

Ridge and LASSO Regression

9. Earlier, we posed the linear regression problem as follows: Find the θ value that minimizes the average squared loss. In other words, our goal is to find $\hat{\theta}$ that satisfies the equation below:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L(\theta) = \underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

Here, $\mathbb{X}$ is a $n \times (p+1)$ matrix, θ is a $(p+1) \times 1$ vector and $\mathbb{Y}$ is a $n \times 1$ vector. Recall that the extra 1 in $(p+1)$ comes from the intercept term. As we saw in lecture, the optimal $\hat{\theta}$ is given by the closed form expression $\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$.

To prevent overfitting, we saw that we can instead minimize the sum of the average squared loss plus a regularization term $\lambda\mathcal{R}(\theta)$.

- If use the function $\mathcal{R}(\theta) = \sum_{j=0}^{p} \theta_j^2 = ||\theta||_2^2$, we have "ridge regression". Recall that $\mathcal{R}$ is the ℓ_2 norm of θ, so this is also referred to as "ℓ_2 regularization".
- If we use the function $\mathcal{R}(\theta) = \sum_{j=0}^{p} |\theta_j| = ||\theta||_1$, we have "LASSO regression". Recall that $\mathcal{R}$ is the ℓ_1 norm of θ, so this is also referred to as "ℓ_1 regularization".

Note that in both of the above formulations, we are regularizing the intercept term to simplify the mathematical formulation of ridge and LASSO regression. In practice, we would not actually want to regularize the intercept term.

For example, if we choose $\mathcal{R}(\theta) = ||\theta||_2^2$, our goal is to find $\hat{\theta}$ that satisfies the equation below:

$$\begin{aligned}\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L(\theta) &= \underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2 + \lambda||\theta||_2^2 \\ &= \underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{X}_{i,\cdot}^T\theta)^2 + \lambda\sum_{j=0}^{p}\theta_j^2\end{aligned}$$

Recall that λ is a hyperparameter that determines the impact of the regularization term. Like ordinary least squares, we can also find a closed form solution to ridge regression: $\hat{\theta} = (\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I})^{-1}\mathbb{X}^T\mathbb{Y}$. It turns out that $\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I}$ is guaranteed to be invertible (unlike $\mathbb{X}^T\mathbb{X}$ which might not be invertible).

(a) As model complexity increases, what happens to the bias and variance of the model?

(b) In ridge regression, what happens if we set $\lambda = 0$? What happens as λ approaches ∞?

(c) If we have a large number of features (10,000+) and we suspect that only a handful of features are useful, which type of regression (Lasso vs Ridge) would be more helpful in interpreting useful features?

(d) What are the benefits of using ridge regression over OLS?

Weighted Least Squares for Housing (Bonus)

10. Anirudhan wants to extend his multiple linear regression modeling framework to incorporate more explicit outlier desensitization for predicting housing prices. One of the ways to do this is to remove outliers, but instead of removing them entirely, perhaps we can choose to "care" less about them through our loss function.

    In other words, we can change our loss function slightly to assign *less* of a weighting to the loss of these outliers. To do this, he decides to weight each sample by a particular amount α_i in the calculation of the loss function. In other words, we augment the loss function as follows:

    $$L(\theta) = \sum_i \alpha_i (y_i - x_i^T \theta)^2$$

    (a) Show that the augmented loss function can be written as follows in matrix/vector notation (i.e. without any summations) for some matrix A that you will find. Assume that α is a vector such that the ith element contains α_i.

    $$L(\theta) = \|A(y - X\theta)\|_2^2$$

    (b) Using the loss vector specified in matrix/vector notation, derive the optimal solution for θ in terms of the appropriate variables (i.e. X, y, α).
    Hint: You should not be doing any optimization (i.e. calculus) in this part!

    (c) True/False: The weighting function $\alpha_i = f(y_i)$ must be linear in terms of X and y for the optimal solution derived to hold. Why or why not?

    (d) Suggest a usage for the following weighting functions $\alpha_i = f(y_i)$:

    $$f(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

    $$f(y_i) = \frac{e^{-y_i}}{\sum_j e^{-y_j}}$$

    (e) The weighting function $\alpha_i = f(...)$ can be a function of the following variables while being a linear model:

☐ A. X
☐ B. y
☐ C. θ

Data 100, Spring 2022 **March 18, 2022**

Discussion #9

Cross Validation

1. After running 5-fold cross validation, we get the following mean squared errors for each fold and value of λ when using Ridge regularization:

Fold Num	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Avg
1	80.2	70.2	91.2	91.8	83.4
2	76.8	66.8	88.8	98.8	82.8
3	81.5	71.5	86.5	88.5	82.0
4	79.4	68.4	92.3	92.4	83.1
5	77.3	67.3	93.4	94.3	83.0
Col Avg	79.0	68.8	90.4	93.2	

   How do we use the information above to choose our model? Do we pick a specific fold? a specific λ? or a specific fold-λ pair? Explain.

2. In the typical setup of k-fold cross validation, we use a different parameter value on each fold, compute the mean squared error of each fold and choose the parameter whose fold has the lowest loss.

   ◯ A. True

   ◯ B. False

Guessing at Random

3. A multiple choice test has 100 questions, each with five possible answers of which one is right. The grading scheme is as follows:

   - 4 points are awarded for each right answer.
   - For each other answer (wrong, missing, etc), one point is taken off; that is, -1 points are awarded.

   A student hasn't studied at all and therefore guesses each answer uniformly at random, independently of all the other answers.

   Define the following random variables:

   - R: the number of answers the student gets right
   - W: the number of answers the student does not get right
   - S: the student's score on the test

   We analyze the random variable R, which denotes the number of answers the student got right.

   (a) What is the distribution of R? Provide the name and parameters of the appropriate distribution. Explain your answer.

   (b) Find $\mathbb{E}(R)$.

4. True or False: $\mathbb{SD}(R) = \mathbb{SD}(W)$.

5. Find $\mathbb{E}(S)$, the student's expected score on the test.

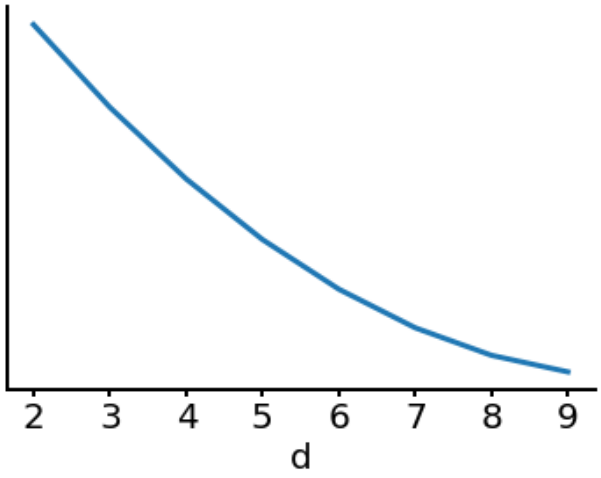
6. Find $\mathbb{SD}(S)$.

Bias-Variance Trade-Off

7. Your team would like to train a machine learning model in order to predict the next YouTube video that a user will click on based on the videos the user has watched in the past. We extract m attributes (such as length of video, view count etc) from each video and our model will be based on the previous d videos watched by that user.

   Hence the number of features for each data point for the model is $m \cdot d$. Currently, you're not sure how many videos to consider.

   (a) Your colleague generates the following plot, where the value d is on the x-axis. However, they forgot to label the y-axis.



   Which of the following could the y-axis represent? Select all that apply.

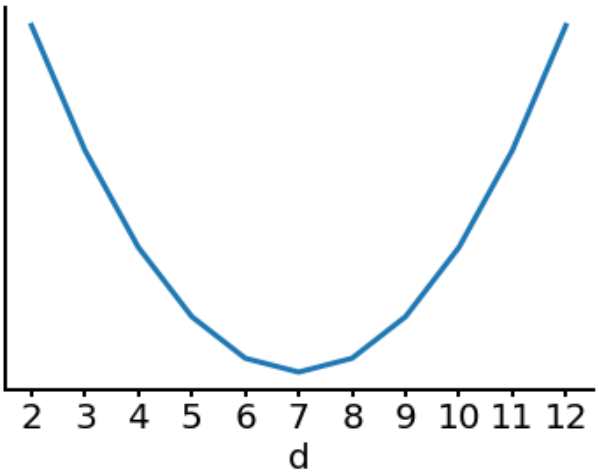
   ☐ A. Training Error

   ☐ B. Validation Error

   ☐ C. Bias

   ☐ D. Variance

   (b) Your colleague generates the following plot, where the value d is on the x-axis. However, they forgot to label the y-axis again.



Which of the following could the y axis represent? Select all that apply.

- ☐ A. Training Error
- ☐ B. Validation Error
- ☐ C. Bias
- ☐ D. Variance

8. We randomly sample some data $(x_i, y_i)_{i=1}^n$ and use it to fit a model $f_{\hat{\theta}}(x)$ according to some procedure (e.g. OLS, Ridge, LASSO). We then sample a new point that is independent from our existing points, but sampled from the same underlying truth as our data. Furthermore, assume that we have a function $g(x)$ and some noise generation process that produces ϵ such that $\mathbb{E}[\epsilon] = 0$ and $\text{var}(\epsilon) = \sigma^2$. Every time we query mother nature for Y at a given a x, she gives us $Y = g(x) + \epsilon$. (The true function for our data is $Y = g(x) + \epsilon$.) A new ϵ is generated each time, independent of the last. In class, we showed that

$$\underbrace{\mathbb{E}\left[(Y - f_{\hat{\theta}}(x))^2\right]}_{} = \underbrace{\sigma^2}_{} + \underbrace{(g(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2}_{} + \underbrace{\mathbb{E}\left[(f_{\hat{\theta}}(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2\right]}_{}$$

(a) Label each of the terms above.

Word Bank: observation variance, model variance, observation bias^2, model bias^2, model risk, empirical mean square error.

(b) What is random in the equation above? Where does the randomness come from?

(c) Calculate the value of $\mathbb{E}[\epsilon f_{\hat{\theta}}(x)]$.

Data 100, Spring 2022 **April 1, 2022**

Discussion #10

Regularization and Bias-Variance Tradeoff

1. We will use a simple constant model $f_\theta(x) = \theta$ to show the effects of regularization on bias and variance. For the sake of simplicity, we will assume that there is no noise or observational variance, so the ground truth output is equal to the observed outputs: $g(x) = Y$.

   (a) Recall that the optimal solution for the constant model with an MSE loss and a dataset $\mathcal{D}$ with $y_1, y_2, ..., y_n$ is the mean $\bar{y}$.

   Suppose that we use L-2 regularization with coefficient $\lambda > 0$ for training another constant model. Derive the optimal solution to this new constant model **with L-2 regularization** to minimize the objective function below.

$$\arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2 + \lambda\theta^2$$

   (b) Using the bias-variance decomposition, show that the optimal constant model's expected loss on a sample test point (x, Y) in terms of the training data y is equal to the following.

$$\mathbb{E}_\mathcal{D}[(Y - f_{\hat{\theta}}(x))^2] = (Y - \mathbb{E}_\mathcal{D}[\bar{y}])^2 + \text{Var}_\mathcal{D}(\bar{y})$$

   Note: the subscript next to the expectation and variance simply lets you know what is random inside the expectation (i.e. what is the expectation taken over?). In this case, we calculate the expectation and variance of $\bar{y}$ across datasets D.

(c) Redo part (b) with the constant model with the optimal constant model with L-2 regularization to derive the expected loss on a sample point. Show that the obtained bias-variance decomposition is the following.

$$\mathbb{E}_{\mathcal{D}}[(Y - f_{\hat{\theta}}(x))^2] = (Y - \frac{1}{1+\lambda}\mathbb{E}_{\mathcal{D}}[\bar{y}])^2 + \frac{1}{(1+\lambda)^2}\text{Var}_{\mathcal{D}}(\bar{y})$$

Hint: Use your result from part (a), along with some properties of expectations and variances!

(d) Remark on how regularization has affected the model bias and model variance as λ increases. Consider what would happen to these quantities as $\lambda \to \infty$.

SQL Syntax

All SQL queries should follow this basic framework. Note that the order of the clauses matter.

```
SELECT [DISTINCT] ___<columns>___
FROM ___<tables>___
[WHERE ___<predicate>___]
[GROUP BY ___<columns>___]
[HAVING ___<predicate>___]
[ORDER BY ___<columns>___]
[LIMIT ___<number of rows>___]
```

2. For this question, we will be working with the UC Berkeley Undergraduate Career Survey dataset, named `survey`. Each year, the UC Berkeley career center surveys graduating seniors for their plans after graduating. Below is a sample of the full dataset. The full dataset contains many thousands of rows.

j_name	**c_name**	**c_location**	**m_name**
Llama Technician	Google	MOUNTAIN VIEW	EECS
Software Engineer	Salesforce	SF	EECS
Open Source Maintainer	Github	SF	Computer Science
Big Data Engineer	Microsoft	REDMOND	Data Science
Data Analyst	Startup	BERKELEY	Data Science
Analyst Intern	Google	SF	Philosophy

Each record of the `survey` table is an entry corresponding to a student. We have the job title, company information, and the student's major.

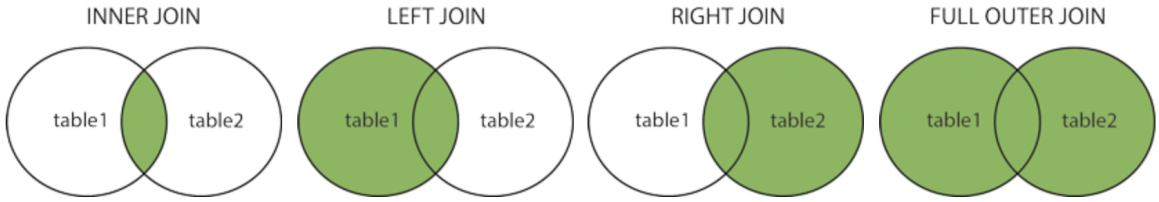
(a) Write a SQL query that selects all data science major graduates that got jobs in Berkeley. The result generated by your query should include all 4 columns.

```
_________________ FROM survey
_____________________________
_____________________________
```

(b) Write a SQL query to find the top 5 popular companies that data science graduates will work at, from most popular to 5th most popular.

```
SELECT c_name, _____________ AS count
FROM survey
WHERE _____________ = "Data Science"
GROUP BY _____________
ORDER BY _____________
LIMIT 5
```

SQL Joins



Note: You do not need the JOIN keyword to join SQL tables. The following are equivalent:

```
SELECT column1, column2
FROM table1, table2
WHERE table1.id = table2.id;
```

```
SELECT column1, column2
FROM table1 JOIN table2
ON table1.id = table2.id;
```

3. In the figure above, assume table1 has m records, while table2 has n records. Describe which records are returned from each type of join. What is the maximum possible number of records returned in each join? Consider the cases where on the joined field, (1) both tables have unique values; and (2) both tables have duplicated values.

4. Consider the following real estate schema:

```
Homes(home_id int, city text, bedrooms int, bathrooms int,
area int)
Transactions(home_id int, buyer_id int, seller_id int,
transaction_date date, sale_price int)
Buyers(buyer_id int, name text)
Sellers(seller_id int, name text)
```

Fill in the blanks in the SQL query to find the id and selling price for each home in Berkeley. If the home has not been sold yet, **the price should be NULL**.

```
SELECT _____
FROM _____
_____ JOIN _____
ON _____
WHERE _____;
```

SQL Queries (Extra)

5. Examine this schema for these two tables:

```
CREATE TABLE owners (
    id integer,
    name text,
    age integer,
    PRIMARY KEY (id)
);
```

```
CREATE TABLE cats (
    id integer,
    owner_id integer,
    name text,
    breed text,
    age integer,
    PRIMARY KEY (id),
    FOREIGN KEY (owner_id) REFERENCES owners
);
```

(a) Write a SQL query to figure out the number of cats, over the age of 10, of each breed of cat.

(b) Write a SQL query to figure out the number of cats each owner owns for owners whose id is greater than 10.

(c) Write a SQL query to figure out the ownerid/owner of the one cat owner who owns the most cats.

(d) Write a SQL query to figure out the names of all of the cat owners who have a cat named Apricot.

(e) It is possible to have a cat with an owner not in the owners table.

◯ A. True ◯ B. False

(f) Write a SQL query to get a random sample of 5 random Siamese Cat (a cat breed) with a name that starts with the letter A.

(g) (Challenge) Write a SQL query to create an almost identical table as cats, except with an additional column 'Nickname' that has the value 'Kitten' for cats less than or equal to the age of 1, 'Catto' for cats between 1 and 15, and 'Wise One' for cats older than or equal to 15.

(h) (Challenge) Write a SQL query to select all rows from the `cats` table that have cats of the top 5 most popular cat breeds.

SQL (Bonus)

6. We wish to convert each Pandas expression to SQL assuming `data` represents a Pandas DataFrame containing 3 columns: `name`, `rank`, and `year`. Both the rank and year are stored as integers.

   (a) `(data['rank'].T.dot(data['year'])) / (data['rank'] ** 2).sum()`

   (b)
   ```
   data.loc[data[‘rank’] < 10, ‘name’] \
       .value_counts() \
       .reset_index()
   ```
   Hint: Remember that `value_counts` returns a sorted output!

   (c)
   ```
   data.merge(data, on = ‘name’) \
       .sort_values(by = ‘name_x’, ascending = False)
   ```

   (d)
   ```
   data.groupby([‘name’, ‘rank’]) \
       .apply(lambda sdf: sdf['year'].max()) \
       .reset_index().head(5)
   ```

   (e)
   ```
   data.groupby(['name', 'year']) \
       .filter(lambda sdf: len(sdf) > 5) \
       .groupby(['name', 'year'])['rank'] \
       .min() \
       .reset_index().head(5)
   ```

Data 100, Spring 2022 **April 8, 2022**

Discussion #11

PCA

1. Consider the following dataset X:

Observations	Variable 1	Variable 2	Variable 3
1	-3.59	7.39	-0.78
2	-8.37	-5.32	0.90
3	1.75	-0.61	-0.62
4	10.21	-1.46	0.50
Mean	0	0	0
Variance	63.42	28.47	0.68

After performing the SVD on this data, we obtain $X = U\Sigma V^T$, where:

$$U = \begin{bmatrix} -0.25 & 0.81 & 0.20 \\ -0.61 & -0.56 & 0.24 \\ 0.13 & -0.06 & -0.85 \\ 0.74 & -0.18 & 0.41 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 13.79 & 0 & 0 \\ 0 & 9.32 & 0 \\ 0 & 0 & 0.81 \end{bmatrix}$$

$$V^T = \begin{bmatrix} 1.00 & 0.02 & 0.00 \\ -0.02 & 0.99 & -0.13 \\ 0.00 & 0.13 & 0.99 \end{bmatrix}$$

Note: Values were rounded to 2 decimals, U and V^T are not perfectly orthonormal due to approximation error.

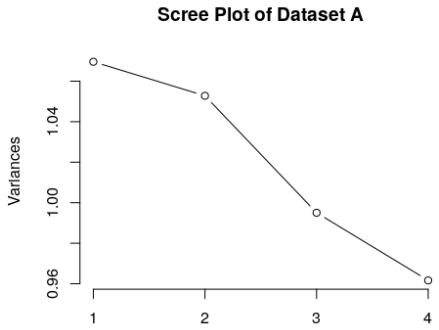
(a) Recall that XV contains the principal components of dataset X. and that we can alternatively calculate it given that $XV = U\Sigma$. Show that $XV = U\Sigma$.

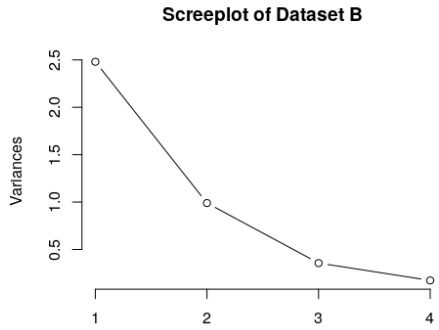
(b) Compute the first principal component (round to 2 decimals).

(c) Given the results of (a), how can we interpret the rows of V^T? What do the values in these rows represent?

(d) Show that the principal component vectors are orthogonal. In other words, for all vectors v_i, v_j that are principal components of dataset X for $i \neq j$, it is true that $v_i^T v_j = 0$.
To show this easily, we can demonstrate that given the principal component matrix P, $P^T P$ is a diagonal matrix. Show that $P^T P$ is diagonal and justify why this proves that the principal component vectors are orthogonal.
Hint: The fact that $(AB)^T = B^T A^T$ might help!

2. Compare the scree plots produced by performing PCA on dataset A and on dataset B. For which dataset would PCA provide the most informative scatter-plot (i.e. plotting PC1 and PC2)? Note that the columns of both datasets were centered to have means of 0 and scaled to have a variance of 1. This means you can interpret the vertical axis as proportional to the fraction of variance captured by each principal component.





Application of PCA

3. Anirudhan wants to apply PCA to a dataset of rare rabbits 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

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

(a) Select all appropriate preprocessing steps used for PCA.

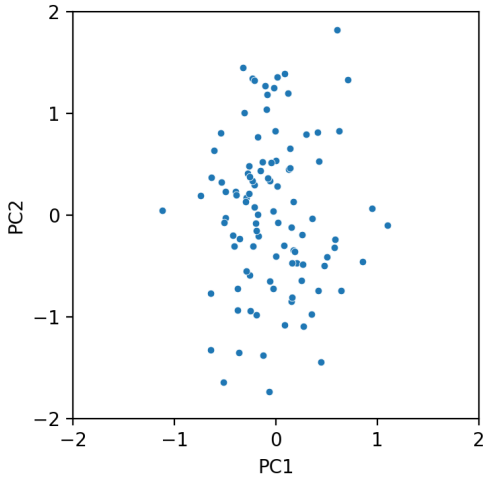
- ☐ A. Transform each row to have a magnitude of 1 (Normalization)
- ☐ B. Transform each column to have a mean of 0 (Centering)
- ☐ C. Transform each column to have a mean of 0 and a standard deviation of 1 (Standardization)
- ☐ D. None of the above

(b) Assume you have correctly preprocessed your data using the correct response in part (a). Write a line of code that returns the first 3 principal components assuming you have the correctly preprocessed DataFrame `rabbit_PCA` and the following variables returned by SVD.

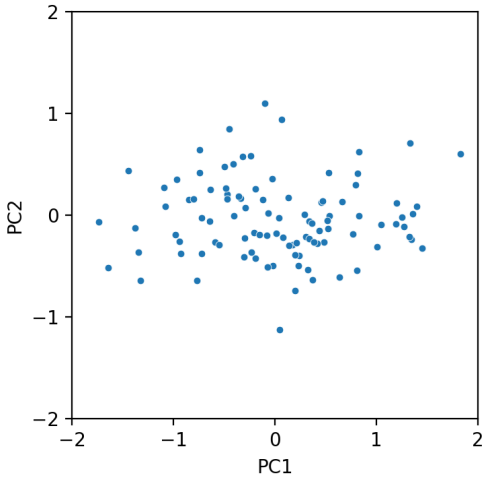
```
u, s, vt = np.linalg.svd(rabbit_PCA, full_matrices = False)
first_3_pcs = _____________________________________
```

(c) 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 given that the first principal component captures 60% of the variance and the second principal component captures 15% of the variance?

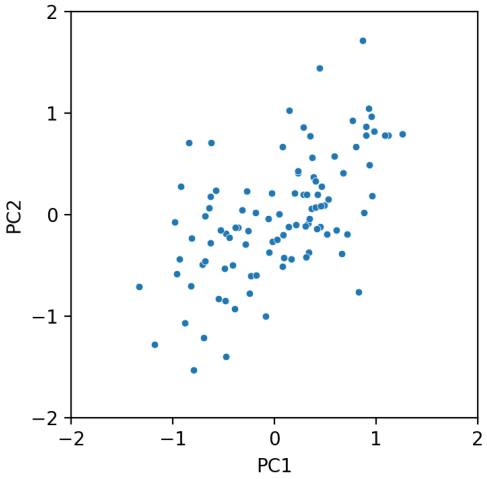
A.


PC2
PC1
2
1
0
−1
−2
−2
−1
0
1
2

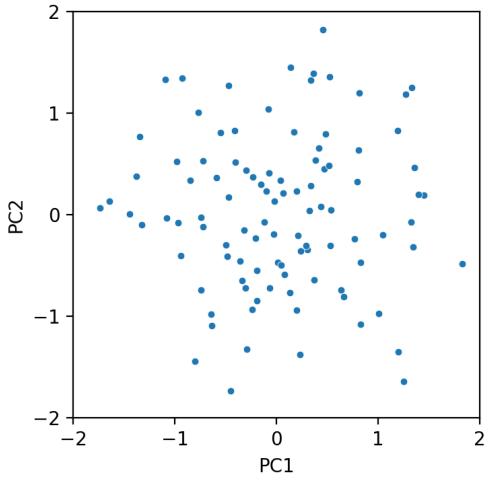
B.


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.


PC2
PC1
2
1
0
−1
−2
−2
−1
0
1
2

Data 100, Spring 2022 **April 15, 2022**

Discussion #12

Logistic Regression

1. Suppose we are given the following dataset, with two features ($\mathbb{X}_{:,0}$ and $\mathbb{X}_{:,1}$) and one response variable (y).

$\mathbb{X}_{:,0}$	$\mathbb{X}_{:,1}$	y
2	2	0
1	-1	1

Here, $\vec{x}$ corresponds to a single row of our data matrix, not including the y column. We write all vectors for this class as column vectors. Thus, we can write, $\vec{x}_1$, as $\vec{x}_1 = [2 \quad 2]^T$. Note that there is no intercept term!

You run an algorithm to fit a model for the probability of $Y = 1$ given $\vec{x}$:

$$\mathbb{P}(Y = 1 \mid \vec{x}) = \sigma(\vec{x}^T\theta)$$

where

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

Your algorithm returns $\hat{\theta} = \left[-\frac{1}{2} \quad -\frac{1}{2}\right]^T$

(a) Calculate $\hat{\mathbb{P}}\left(Y = 1 \mid \vec{x} = [1 \quad 0]^T\right)$.

(b) The empirical risk using log loss (a.k.a., cross-entropy loss) is given by the following expression. Remember, whenever you see $\log$ in this course, you must assume the natural logarithm (base-e), unless explicitly told otherwise.

$$R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log \hat{\mathbb{P}}\left(Y = 1 \mid \vec{x_i}\right) + (1 - y_i) \log \hat{\mathbb{P}}\left(Y = 0 \mid \vec{x_i}\right)\right)$$

Suppose we run a different algorithm and obtain $\hat{\theta} = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$. Calculate the empirical risk for this new $\hat{\theta}$ on our dataset.

(c) Is our dataset linearly separable? If so, write the equation of a hyperplane that separates the two classes.

(d) Does either of our fitted models $\hat{\theta} = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$ or $\hat{\theta} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$ minimize cross-entropy loss?

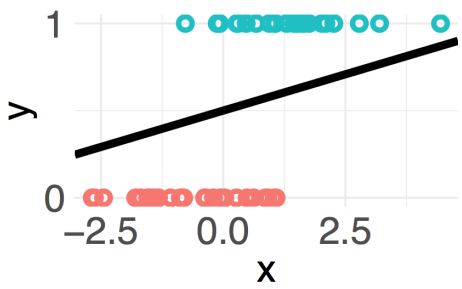
(e) Consider using a linear regression model such that $\mathbb{X}\hat{\theta} = \hat{\mathbb{Y}}$. What is the MSE of the optimal linear regression model given the dataset we used above for logistic regression?

(f) Assume we add the data point $[3, -2]$ to the design matrix such that our resulting design matrix is as follows.

$\mathbb{X}_{:,0}$	$\mathbb{X}_{:,1}$	y
2	2	0
1	-1	1
3	-2	0

Comment on whether it is possible to achieve perfect accuracy using a logistic regression model and whether it is possible to achieve zero MSE using a linear regression model trained on the new data. Assume that we don't use an intercept term.

2. Suppose your friend obtains a dataset with a single feature x. Your friend argues that the data are linearly separable by drawing the line on the following plot of the data.



(a) Argue whether or not your friend is correct. Note: this question refers to a binary classification problem with a single feature.

(b) Suppose you use gradient descent for a **fixed number of iterations** to train a logistic regression model on two design matrices $\mathbb{X}_a$ and $\mathbb{X}_b$. After training, you find that the training accuracy for $\mathbb{X}_a$ is 100% and the training accuracy for $\mathbb{X}_b$ is 98%. What can you say about whether the data is linearly separable for the two design matrices?

3. Suppose we train a binary classifier on this dataset. Suppose y is the set of true labels, and $\hat{y}$ is the set of predicted labels.

y	0	0	0	0	0	1	1	1	1	1
$\hat{y}$	0	1	1	1	1	1	1	0	0	0

Determine each of the following quantities.

(a) Confusion matrix.

Hint: The first row contains the true negatives and false positives, and the second row contains false negatives and true positives (in that order).

(b) The precision of our classifier. Write your answer as a simplified fraction.

(c) The recall of our classifier. Write your answer as a simplified fraction.

Data 100, Spring 2022 **April 22, 2022**

Discussion #13

Decision Trees and Random Forests

1. (a) When creating a decision tree for classification, give two reasons why we might end up having a terminal node that has more than one class.

   (b) Suppose we have a decision tree for classifying the iris data set. Suppose that one terminal decision tree node contains 22 setosas and 13 versicolors. If we're trying to make a prediction and our sequence of yes/no questions leads us to this node, what should we do?

   ◯ A. predict that the class is setosa
   ◯ B. give a probability of setosa = $\sigma(22/35)$
   ◯ C. refuse to make a prediction
   ◯ D. other (describe)

   (c) What techniques can we use to avoid overfitting decision trees?

   (d) Suppose we limit the complexity of our decision tree model by setting a maximum possible node depth d, i.e. no new nodes may be created with depth greater than d. What technique should we use to pick d?

   (e) What is the advantage of a random forest over a decision tree?

   ☐ A. lower bias
   ☐ B. lower variability
   ☐ C. lower bias and variability
   ☐ D. none of these

   (f) Let's say we have a dataset with 3 classes, labeled A, B, and C, from which we want to build a classifier using a decision tree. The following table displays how many points are in each class:

Class	Number of Points
A	4
B	4
C	2

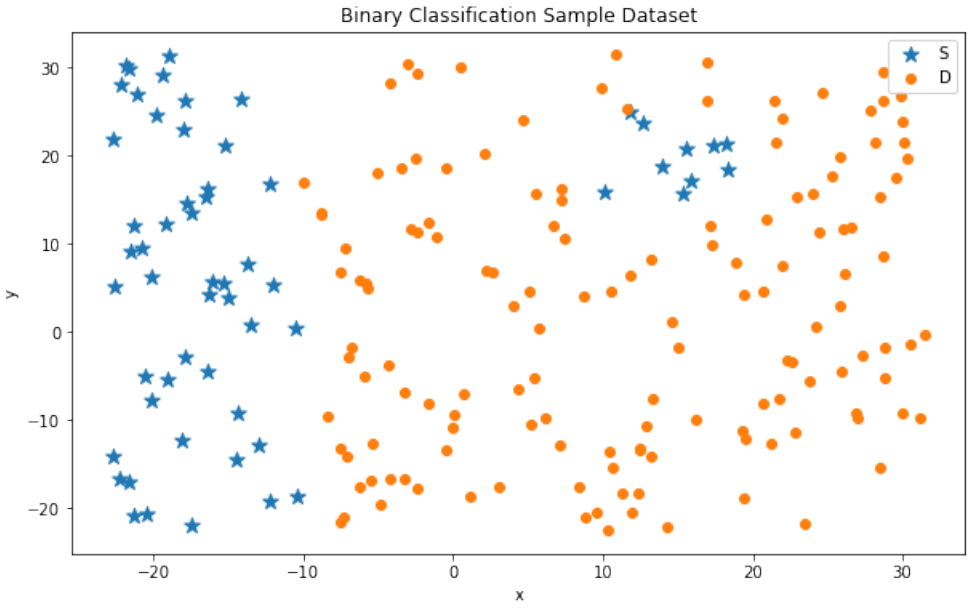
Recall from lecture that we want to minimize the weighted entropy of our splits. What is the weighted entropy of the following split?

Node 1: 4 in class A, 0 in class B, 2 in class C; **Node 2:** 0 in class A, 4 in class B, 0 in class C.

- ◯ A. 0
- ◯ B. 10
- ◯ C. $-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$
- ◯ D. $-\frac{2}{5}(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3})$
- ◯ E. $-\frac{3}{5}(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3})$
- ◯ F. None of the Above

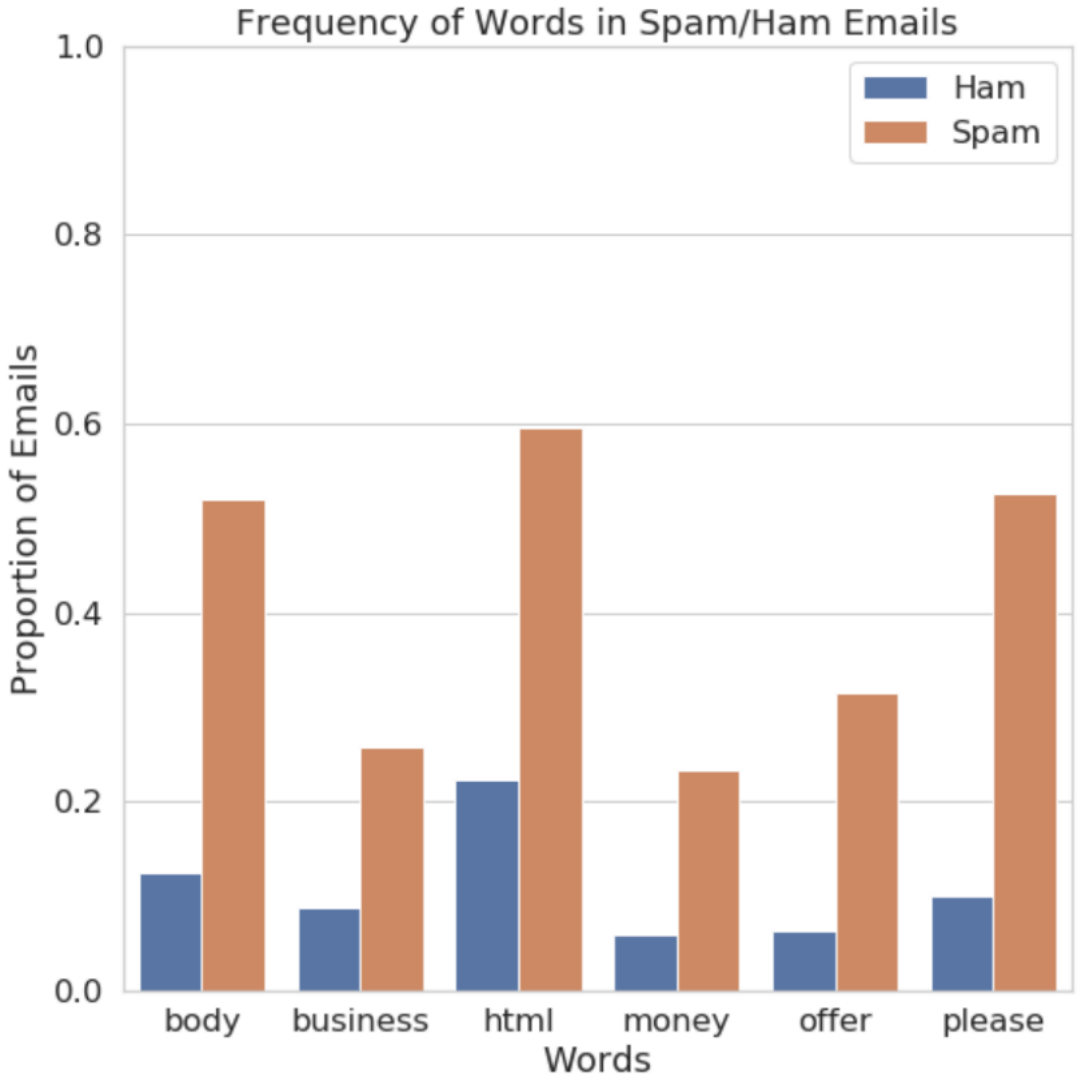
(g) Suppose we wish to train a classifier on a dataset containing classes S and D. The dataset contains two features that are plotted below (x and y), along with their respective classes denoted by stars (S) and dots (D).

Draw the approximate optimal decision boundaries chosen by a decision tree and a logistic regression model trained on the data shown by the figure below.



Spam vs. Ham (Easy/Moderate)

2. We will borrow some techniques from decision trees to build the best possible spam/ham detection classifier possible. Consider the visualization of the words that occur frequently in spam emails but infrequently in ham emails (or perhaps vice versa). These are relevant since it provides the model with word features that differentiate between the classes.



We will study this in-depth in the following parts using some of the concepts that we have learned from our study of decision trees! Assume that our spam/ham dataset contains 20,000 emails, with 10,000 spam emails and 10,000 ham emails (this isn't true - but we will pretend it is to make calculations easier).

(a) Suppose that we are building a decision tree of whether an email is spam or ham, where

the decision tree can read the text in emails. Estimate the weighted node entropy of a split in a decision tree, where the left split corresponds to emails containing the word ”html” and the right split corresponds to emails not containing the word ”html”.

(b) What split word among those shown in the figure is the most effective using the same calculations as we performed in the previous subpart?

(c) In general, what kinds of words in the text would it find most useful to differentiate or decide between the two classes? Describe a procedure to select the best words for the spam/ham logistic classifier.

Data 100, Spring 2022 **Name:**

Discussion #14

Clustering

1. (a) Describe the difference between clustering and classification.

   (b) Given a set of points and their labels (or cluster assignments) from a K-Means clustering, how can we compute the centroids of each of the clusters?

   (c) The process of fitting a K-means model outputs a set of k centers. We can compute the quality of the output by computing the distortion on the dataset. A Data 100 student suggests that distortion is not well-defined when evaluating the output of any agglomerative clustering algorithm because the algorithm doesn't return centers, but simply labels each point individually. Is the student correct?

   (d) Describe qualitatively what it means for a data point to have a negative silhouette score.

   (e) Suppose that no two points have the same distance from each other. Are the cluster labels computed by K-means always the same for a given dataset? What about for max agglomerative clustering?