Data C100, Final Exam

Summer 2024

Name: ______________________________ Seat Number: ________

Email: ______________________________@berkeley.edu

SID: ______________ Check □ if incomplete student (circle term): fa23/sp24

Name of the student to your left: ______________________________

Name of the student to your right: ______________________________

Instructions:

Do not open the exam until instructed to do so.

This exam consists of **68 points** spread out over **7 questions** on **30 pages** and must be completed in the **110 minute** time period on August 8, 2024, from 9:10 AM to 11:00 AM unless you have pre-approved accommodations otherwise.

Note that some questions have circular bubbles to select a choice. This means that you should only **select one choice**. Other questions have boxes. This means you should **select all that apply**. Please shade in the box/circle to mark your answer.

There is space to write your student ID number (SID) in the upper right-hand corner of each page of the exam. **Make sure to write your SID on each page** to ensure that your exam is graded.

Honor Code [0pt or $-\infty$]:

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I am the person whose name is on the exam, and I completed this exam in accordance with the Honor Code.

Signature: ______________________________

This page has been intentionally left blank.

1 Taco Chewsday [13 Points]

Data 100 staff likes to go to Taco Tuesdays for the yummy and cheap tacos ($2.50 tacos!). Every week Angela sends a little blurb on Slack inviting staff to join her. Kevin wants to join his lovely course staff, but he would need to drive all the way to Berkeley after work. Kevin, unsure about whether or not he should go, decides to use a logistic regression model to reach a decision. He randomly selects 4 course staff members and collects data from them. Each person i is assigned a label y where $y_i = 1$ denotes that they're going and $y_i = 0$ that they're not going.

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

Throughout the question, you can assume that the optimal θ value is $\hat{\theta} = [1 \quad 2 \quad 1]^T$.

(a) [2 Pts] Calculate $P_{\hat{\theta}}(Y = 0|\vec{x}^T = [1 \quad 0 \quad 1])$. *Write your answer as a mathematical expression.*

Solution:

$$\begin{aligned} P_{\hat{\theta}}(Y = 1|\vec{x}^T = [1 \quad 0 \quad 1]) &= \sigma\left(\begin{bmatrix}1 & 0 & 1\end{bmatrix}\begin{bmatrix}1\\2\\1\end{bmatrix}\right) \\ &= \sigma\left(1 \cdot 1 + 0 \cdot 2 + 1 \cdot 1\right) \\ &= \sigma\left(2\right) \\ &= \frac{1}{1+\exp(-2)} \end{aligned}$$

$$\begin{aligned} P_{\hat{\theta}}(Y = 0|\vec{x}^T = [1 \quad 0 \quad 1]) &= 1 - P_{\hat{\theta}}(Y = 1|\vec{x}^T = [1 \quad 0 \quad 1]) \\ &= 1 - \frac{1}{1+\exp(-2)} \end{aligned}$$

(b) [2 Pts] Which of the following statements below are true about logistic regression? **Select all that apply**.

□ The logistic regression model estimates the odds ratio of the outcomes as a linear combination of the input features $\mathbb{X}$ and coeffients θ.

□ The MSE Loss is a suitable alternative for logistic regression because of its convex loss surface.

□ Linearly separable training data always guarantees faster convergence than non-linearly separable data.

□ The intercept term θ_0 controls the translation on the y-axis of the sigmoid function.

□ **None of the above.**

Solution:

1. Statement is false. The logistic regression model estimates the **log odds ratio** of outcomes as a linear combination of $\mathbb{X}$ and θ.
2. Statement is false. The MSE's nonconvexity using the logistic regression is precisely our motivation for introducing the Cross Entropy Loss.
3. Statement is false. Due to the stochastic nature of SGD, linearly seperable training data might not converge faster necessarily.
4. The θ_0 term controls the shift along the x-axis in logistic regression. See Logistic Regression 1 Lecture demo.

(c) [3 Pts] In lecture, we saw how minimizing cross-entropy loss is equivalent to maximizing the likelihood of the training data. Let's say that our likelihood is given by

$$L(\theta) = (\frac{2}{3}\theta)^a \cdot (\frac{1}{3}\theta)^b \cdot (\frac{2}{3}(1-\theta))^c \cdot (\frac{1}{3}(1-\theta))^d$$

Calculate the MLE estimate of θ. **Simplify your final answer.**

Solution:

$$\begin{aligned}
\ell(\theta) &= \left(\frac{2}{3}\theta\right)^a \left(\frac{1}{3}\theta\right)^b \left(\frac{2}{3}(1-\theta)\right)^c \left(\frac{1}{3}(1-\theta)\right)^d \\
L(\theta) &= \log(\ell(\theta)) \\
&= \log\left(\left(\frac{2}{3}\theta\right)^a \left(\frac{1}{3}\theta\right)^b \left(\frac{2}{3}(1-\theta)\right)^c \left(\frac{1}{3}(1-\theta)\right)^d\right) \\
&= a\left[\log\left(\frac{2}{3}\right) + \log(\theta)\right] + b\left[\log\left(\frac{1}{3}\right) + \log(\theta)\right] \\
&\quad + c\left[\log\left(\frac{2}{3}\right) + \log(1-\theta)\right] + d\left[\log\left(\frac{1}{3}\right) + \log(1-\theta)\right] \\
L(\theta) &= (a+c)\log\left(\frac{2}{3}\right) + (b+d)\log\left(\frac{1}{3}\right) \\
&\quad + (a+b)\log(\theta) + (c+d)\log(1-\theta) \\
\frac{d}{d\theta}L(\theta) &= \frac{d}{d\theta}((a+c)\log\left(\frac{2}{3}\right) + (b+d)\log\left(\frac{1}{3}\right) + (a+b)\log(\theta) \\
&+ (c+d)\log(1-\theta)) \\
&= \frac{d}{d\theta}\left((a+c)\log\left(\frac{2}{3}\right)\right) + \frac{d}{d\theta}\left((b+d)\log\left(\frac{1}{3}\right)\right) \\
&\quad + \frac{d}{d\theta}\left((a+b)\log(\theta)\right) + \frac{d}{d\theta}\left((c+d)\log(1-\theta)\right) \\
0 &= \frac{a+b}{\theta} - \frac{c+d}{1-\theta} \\
&= \frac{a+b}{\theta} - \frac{c+d}{1-\theta} \\
\theta(c+d) &= (1-\theta)(a+b) \\
\theta(c+d) &= a+b-\theta(a+b) \\
\theta(c+d+a+b) &= a+b \\
\theta &= \frac{a+b}{a+b+c+d}
\end{aligned}$$

Alternative solution to (c): instead of calculating the derivative of $\log(L(\theta))$, another approach involved calculating derivative of $L(\theta)$ directly using the product rule.

Solution:

$$L(\theta) = \left(\frac{2}{3}\theta\right)^a \cdot \left(\frac{2}{3}\theta\right)^b \cdot \left(\frac{2}{3}(1-\theta)\right)^c \cdot \left(\frac{1}{3}(1-\theta)\right)^d$$

$$L(\theta) = \left(\frac{2}{3}\right)^{a+b} \cdot \left(\frac{1}{3}\right)^{c+d} \cdot \theta^{a+b} \cdot (1-\theta)^{c+d}$$

$$\frac{dL(\theta)}{d\theta} = \frac{d}{d\theta}\left[\left(\frac{2}{3}\right)^{a+b} \cdot \left(\frac{1}{3}\right)^{c+d} \cdot \theta^{a+b} \cdot (1-\theta)^{c+d}\right]$$

$$\frac{dL(\theta)}{d\theta} = \left(\frac{2}{3}\right)^{a+b} \cdot \left(\frac{1}{3}\right)^{c+d} \cdot \left[\frac{d}{d\theta}\left(\theta^{a+b} \cdot (1-\theta)^{c+d}\right)\right]$$

$$\frac{dL(\theta)}{d\theta} = \left(\frac{2}{3}\right)^{a+b} \cdot \left(\frac{1}{3}\right)^{c+d} \cdot \left[(a+b)\theta^{a+b-1} \cdot (1-\theta)^{c+d} - (c+d)\theta^{a+b} \cdot (1-\theta)^{c+d-1}\right]$$

$$\frac{dL(\theta)}{d\theta} = \left(\frac{2}{3}\right)^{a+b} \cdot \left(\frac{1}{3}\right)^{c+d} \cdot \theta^{a+b-1} \cdot (1-\theta)^{c+d-1} \cdot [(a+b)(1-\theta) - (c+d)\theta]$$

$$\frac{dL(\theta)}{d\theta} = L(\theta) \cdot \frac{(a+b)(1-\theta) - (c+d)\theta}{\theta \cdot (1-\theta)}$$

To find θ, we set $\frac{dL(\theta)}{d\theta} = 0$:

$$(a+b)(1-\theta) = (c+d)\theta$$

Simplifying:

$$a + b - (a+b)\theta = (c+d)\theta$$

$$a + b = \theta(a+b+c+d)$$

$$\theta = \frac{a+b}{a+b+c+d}$$

(d) [2 Pts] Which of the following is true about performing logistic regression on linearly separable data? **Select all that apply**.

□ **We are guaranteed to achieve a 100% training accuracy.**

□ We can find the θ that will give us a cross-entropy loss of 0.

□ **Data is linearly separable if we are able to draw a decision boundary that perfectly separate the classes.**

□ **The data Kevin collected (at the beginning of this question) is linearly separable.**

□ None of the above.

Solution:

- This statement is true. If we have linearly separable training data, we can find a decision boundary that perfectly separates the two classes, yielding a 100% training accuracy.
- This statement is false. Due to the $\log$ terms in the cross entropy loss, it is impossible to get a 0 loss.
- This statement is true. It states the definition of the decision boundary verbatim.
- This statement is true. We can clearly demarcate a boundary that divides the Y classes of the 4 data points.

(e) [2 Pts] Kevin initially started out with a threshold of 0.5, but after realizing that he really didn't want to be the only one at Taco Tuesday, what threshold(s) should he choose so that there are fewer false positives? **Select all that apply.**

□ 0.2

□ 0.4

□ 0.5

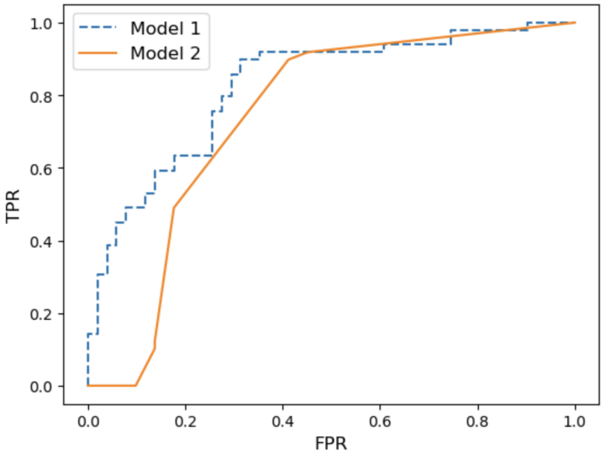
□ **0.6**

□ **0.8**

□ None of the above

Solution: Increasing the threshold means that only those predictions that have a higher confidence will be classified as positive now. In other words, we get fewer positives. As a result, we also yield a lower number of false positives. In this question, in order to get fewer false positives than $T = 0.5$, we select $T > 0.5$.

(f) [2 Pts] Maya trains two mystery models and provides Kevin with the following ROC curve. Which of the following statements are true? **Select all that apply.**



□ Based on the ROC curve, model 2 performs better than model 1.

□ **Model 1's high-confidence predictions (those with higher probabilities) are generally more accurate than those of model 2.**

□ **At lower thresholds, the performances of model 1 and model 2 are similar.**

□ A binary classifier with an AUC close to 0 behaves like a random predictor, i.e. it behaves similarly to a model outputting probabilities uniformly between $[0, 1]$.

□ None of the above.

Solution:

- Statement 1 is false. Model 2 has a lower AUC than 1.
- Statement 2 is true. The false positive rate increases as we decrease our classification threshold. As we can see, at the lower FPR rates (higher thresholds and hence, higher confidence scores) on the left side of the plot, we can infer that model 1 has higher accompanying TPR scores, suggesting a higher accuracy.
- Statement 3 is true. Lower thresholds are mapped out on the right side of the plot, where both plots yield similar FPR and TPR performances.
- Statement 4 is false. A random predictor gets an AUC around 0.5.

2 Pikachu, Charizard, and Arceus [10 Points]

James is a Pokémon trainer who recently traveled to the Safari Zone. Instead of catching wild Pokémon, he decided to study Pokémon he encounters.

He consolidates his data into matrix X, where each row represents some Pokémon he's observed in the wild, and each column represents an attribute of that Pokémon.

(a) [2 Pts] James wants to validate all of the matrices that will be calculated using SVD. However, his Pikachu used Thunder Wave on his notebook, causing the matrix orders to get shuffled! For this question only, you can assume that X is square. Which of the following are always true? **Select all that apply.**

- □ $U^{-1}U = UU^T$
- □ $VX^T = USX^T$
- □ $XVV^T = USV^{-1}VV^T$
- □ $U^{-1}XX^T = SV^TX^T$
- □ None of the above

(b) [2 Pts] James now wants to extract the second principal component. Team Rocket sabotaged his notebook, so he only has access to the matrix U shown below. Given $s_2^2 = 0.25$, calculate the second principal component.

$$U = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \end{bmatrix}$$

Solution:

$$= \sqrt{\frac{1}{4}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2\sqrt{2}} \\ \frac{-1}{2\sqrt{2}} \\ 0 \end{bmatrix}$$

(c) [2 Pts] Which of the following statements below are true about PCA and SVD, in general? Assume that $X \in \mathbb{R}^{n \times d}$. **Select all that apply.**

□ Failure to center the design matrix $\mathbb{X}$ might lead to U and $V^\top$ matrices that are not orthonormal.

□ **If rank$(\mathbb{X}) < d$, we can conclude that the diagonal of S will contain one or more zero values.**

□ V can always be interpreted as a rotation and scaling of $\mathbb{X}$ such that the axis with the most variation is aligned with our basis.

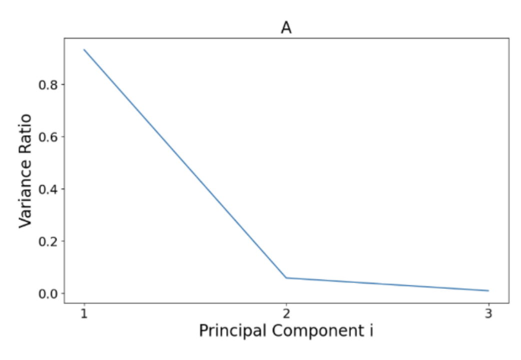
□ **The i-th row in V indicate how each feature contributes to the i-th specific principal component.**

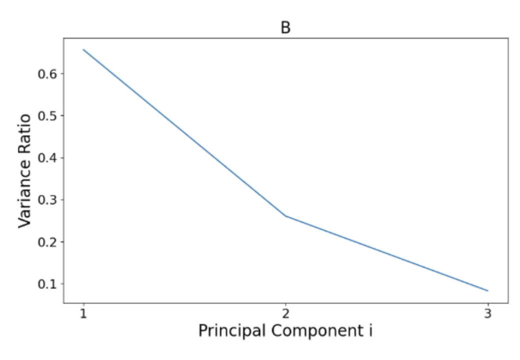
□ None of the above.

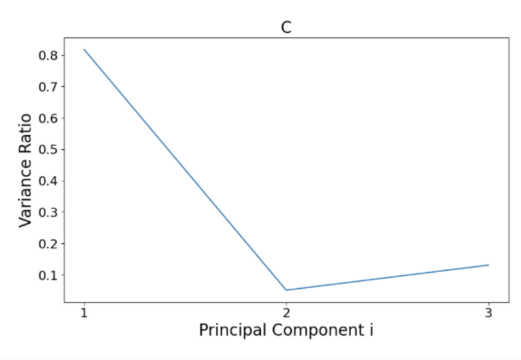
(d) [1 Pt] Pikachu was able to help James recover the following S matrix.

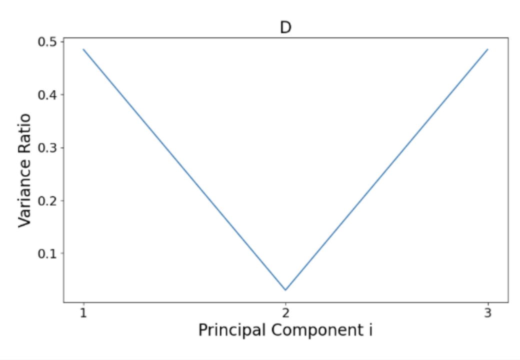
$$S = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.2 \end{bmatrix}$$

Based on the S matrix above, which of the following plots is the corresponding scree plot? **Select one.**



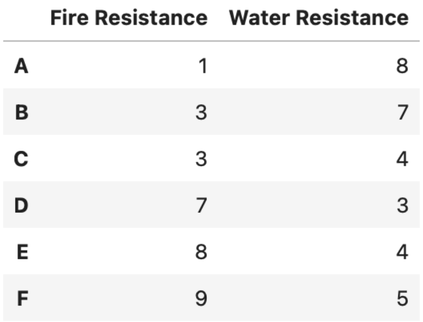




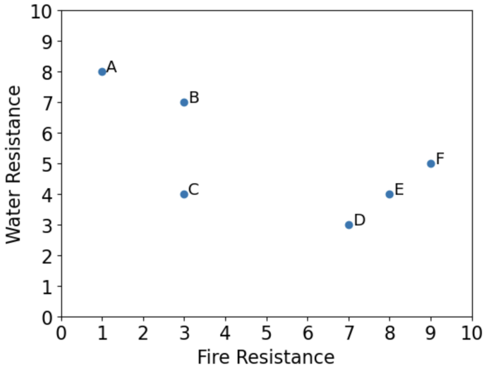


○ **A** ○ B ○ C ○ D

Joining James' adventure in the Safari Zone, Willy wants to use clustering methods to see whether he can group individual Pokémon together. After collecting more data points, he logs two features: a Pokémon's resistance against water and fire, measured on a continuous scale from 1 to 10. The dataframe and a scatter plot have been added for your convenience. For the following questions, assume that distances are calculated using Euclidean distance.



	Fire Resistance	Water Resistance
A	1	8
B	3	7
C	3	4
D	7	3
E	8	4
F	9	5



(e) [1 Pt] Willy decides to perform hierarchical agglomerative clustering using complete linkage.

(i) Which two points are the first to merge into a new cluster? **Write your answer using the provided labels.**

Solution: `(E, F)` or `(E, D)`

(ii) Willy merges two clusters again. Which points are included in this most recent merged cluster? **Write your answer using the provided labels.**

Solution: `A, B`

(f) [2 Pts] Willy decides to start over using K-Means clustering to separate data into two clusters, with the left cluster being centered at $(1, 4)$ and the right cluster being centered at $(6, 7)$.

(i) What points are the left clusters composed of? **Write your answer using the provided labels.**

Solution: `A, C`

(ii) What is the new centroid of the left cluster?

Solution: $(2, 6)$

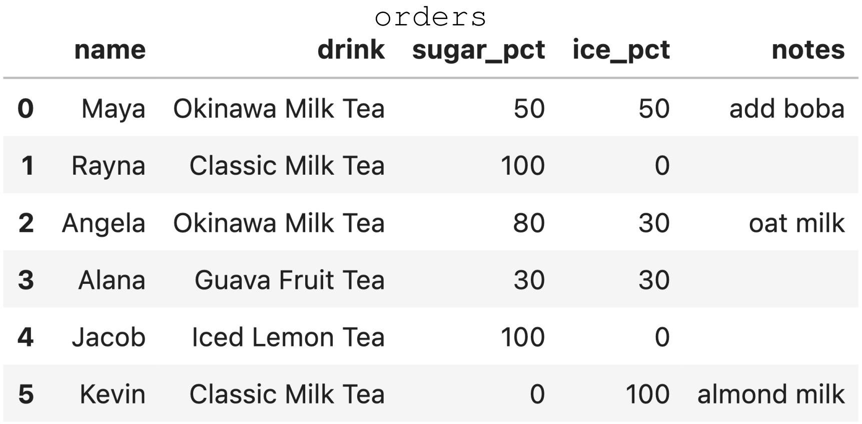
(iii) Calculate the inertia of the left cluster, using the centroid from (ii).

Solution:

$$(1-2)^2 + (3-2)^2 + (8-6)^2 + (4-6)^2 = 1 + 1 + 4 + 4 = 10$$

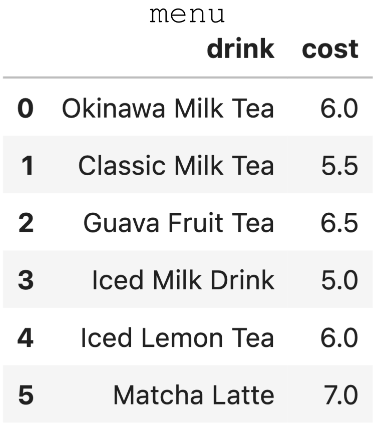
3 Ain't No Free Boba, the SQL [8 Points]

Vicky is ordering boba for the Data 100 staff from her favorite boba shop, DaTea. She organizes data about the orders into two tables in a SQL database: `orders`, containing information about each person's order, and `menu`, containing information about the cost of each drink on DaTea's menu. The full tables are shown below:



orders

	name	drink	sugar_pct	ice_pct	notes
0	Maya	Okinawa Milk Tea	50	50	add boba
1	Rayna	Classic Milk Tea	100	0	
2	Angela	Okinawa Milk Tea	80	30	oat milk
3	Alana	Guava Fruit Tea	30	30	
4	Jacob	Iced Lemon Tea	100	0	
5	Kevin	Classic Milk Tea	0	100	almond milk



menu

	drink	cost
0	Okinawa Milk Tea	6.0
1	Classic Milk Tea	5.5
2	Guava Fruit Tea	6.5
3	Iced Milk Drink	5.0
4	Iced Lemon Tea	6.0
5	Matcha Latte	7.0

(a) [1 Pt] Xiaorui forgot to place an order, but he still wants a drink. He decides to steal somebody else's order, and uses Vicky's data to decide who to steal from. Xiaorui is picky about his boba: he has certain preferences about his drink, and he's also allergic to guava. In the end, Xiaorui executes the following SQL query:

```
SELECT name FROM orders
WHERE ice_pct < 50 AND orders.drink != 'Guava Fruit Tea'
ORDER BY sugar_pct
LIMIT 1;
```

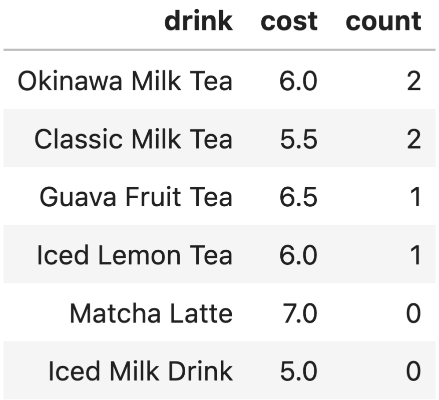
This query outputs a table with one row and one column: that is, a single value. What value is outputted by the query?

Solution: Angela

(b) [2 Pts] Vicky wants to compute the quantity and cost of each drink she needs to buy. Write a SQL query to return a table that, for each item on DaTea's menu, will tell Vicky the cost for that item and how many to order. Your table should have three columns:

- `drink`: the name of the menu item
- `cost`: the cost for one order of that menu item
- `count`: the quantity of that item ordered by the Data 100 staff

The rows should be ordered first by count then cost, both in descending order. Note that not every item on DaTea's menu is ordered by a staff member; items that are not ordered should have a count of 0.
Your table should look similar to the one below:



drink	cost	count
Okinawa Milk Tea	6.0	2
Classic Milk Tea	5.5	2
Guava Fruit Tea	6.5	1
Iced Lemon Tea	6.0	1
Matcha Latte	7.0	0
Iced Milk Drink	5.0	0

Fill in the blanks below (you may not need all the blanks):

Solution:

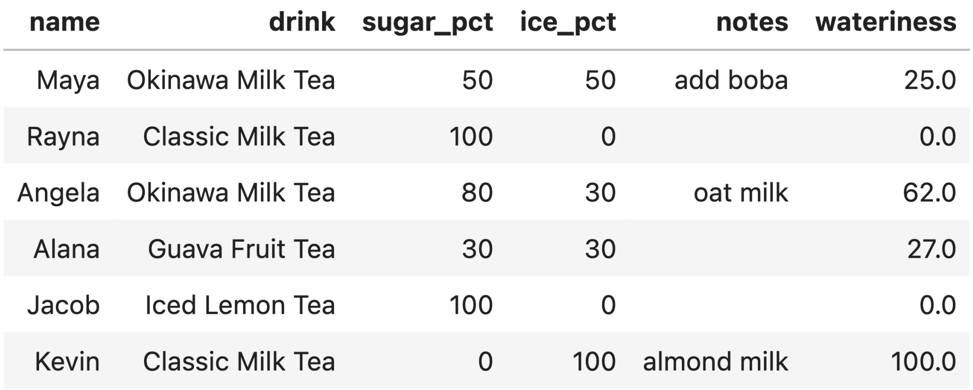
```
SELECT M.drink, FIRST(M.cost) AS cost, COUNT(O.drink)
AS count
FROM menu AS M
LEFT JOIN orders AS O
ON M.drink = O.drink
GROUP BY M.drink
ORDER BY count DESC, cost DESC;
```

(c) [3 Pts] Using his vast knowledge of math and boba, Kevin has defined a new formula for determining how watery a boba drink is. The formula for the wateriness score W of a boba drink is as follows:

$$W = \begin{cases} 80 + 20p_{\text{ice}} - 30p_{\text{sugar}}, & \text{if contains oat or almond milk} \\ 30 + 20p_{\text{ice}} - 30p_{\text{sugar}}, & \text{otherwise} \end{cases}$$

where p_{ice} is the percentage of ice in the drink and p_{sugar} is the percentage of sugar (both in decimal points). Note that the formula for wateriness changes if the drink contains oat or almond milk.

Write a SQL query to return a table that is the same as orders, but with a new column, `wateriness`, containing Kevin's wateriness score W for each drink. Assume that in `orders`, `sugar_pct` and `ice_pct` contain the sugar level and ice level of the drinks in percentage points as integers. Your table should look similar to the one below:



name	drink	sugar_pct	ice_pct	notes	wateriness
Maya	Okinawa Milk Tea	50	50	add boba	25.0
Rayna	Classic Milk Tea	100	0		0.0
Angela	Okinawa Milk Tea	80	30	oat milk	62.0
Alana	Guava Fruit Tea	30	30		27.0
Jacob	Iced Lemon Tea	100	0		0.0
Kevin	Classic Milk Tea	0	100	almond milk	100.0

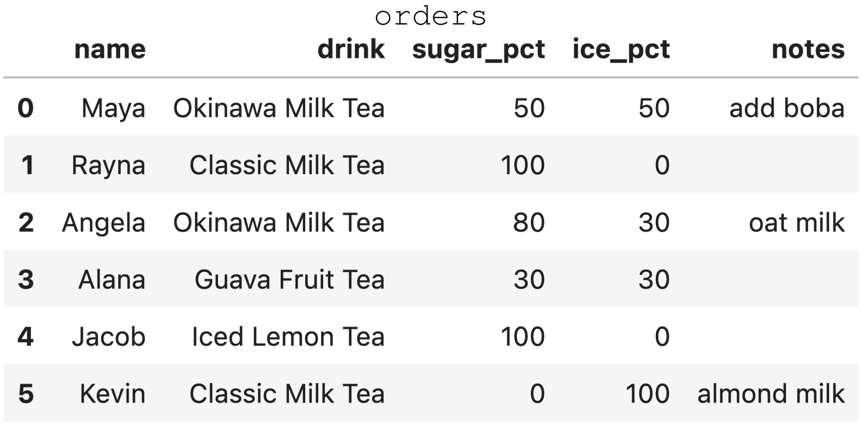
Fill in the blanks below (you may not need all the blanks):

Solution:

```
SELECT *,
CASE WHEN notes = 'oat milk' OR notes = 'almond milk'
THEN 80 + ((ice_pct / 100) * 20) - ((sugar_pct / 100) * 30)
ELSE 30 + ((ice_pct / 100) * 20) - ((sugar_pct / 100) * 30)
END AS wateriness
FROM orders;
```

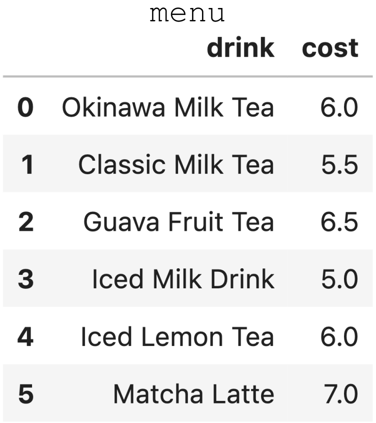
(d) [2 Pts] When joining the tables `orders` and `menu` in SQL, how many rows are outputted by each type of join? Assume that in each case, `orders` is the left table and `menu` is the right table, and that the tables are joined on the condition `orders.drink = menu.drink`.

For your convenience, the tables are reproduced below:



`orders`

	name	drink	sugar_pct	ice_pct	notes
0	Maya	Okinawa Milk Tea	50	50	add boba
1	Rayna	Classic Milk Tea	100	0	
2	Angela	Okinawa Milk Tea	80	30	oat milk
3	Alana	Guava Fruit Tea	30	30	
4	Jacob	Iced Lemon Tea	100	0	
5	Kevin	Classic Milk Tea	0	100	almond milk



`menu`

	drink	cost
0	Okinawa Milk Tea	6.0
1	Classic Milk Tea	5.5
2	Guava Fruit Tea	6.5
3	Iced Milk Drink	5.0
4	Iced Lemon Tea	6.0
5	Matcha Latte	7.0

(i) `INNER JOIN`

Solution: 6

(ii) `LEFT JOIN`

Solution: 6

(iii) `RIGHT JOIN`

Solution: 8

(iv) `FULL (OUTER) JOIN`

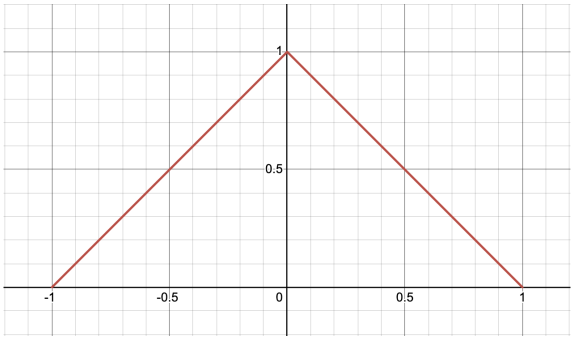
Solution: 8

4 Kermit Density Estimation [4 Points]

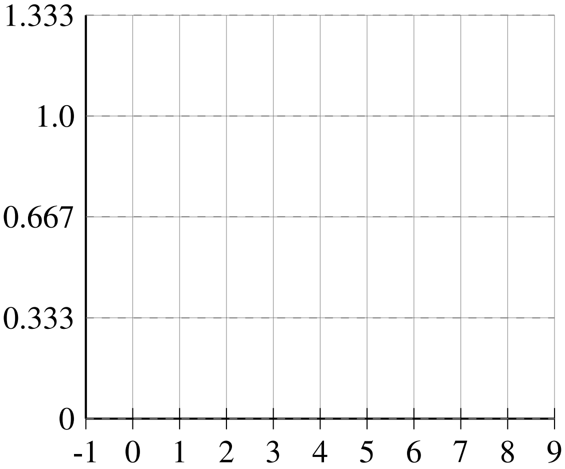
Kermit, an avid pet enjoyer, wants to understand the distribution of pets amongst students taking Data 100 in Summer 2024. He is interested in plotting his data with various KDEs. Below, we introduce a new kernel function called the **Triangular Kernel** which follows the kernel functions below:

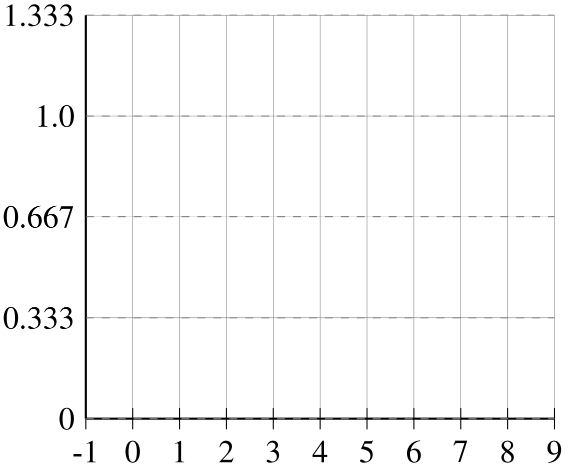
Triangular Kernel : $T_\alpha(x, x_i) = \begin{cases} \frac{1}{\alpha}(1 - \frac{|x-x_i|}{\alpha}) & \text{if } x_i - \alpha \leq x \leq x_i + \alpha \\ 0 & \text{else} \end{cases}$

The Triangular Kernel centered at $x_i = 0$ and $\alpha = 1$ is shown below.

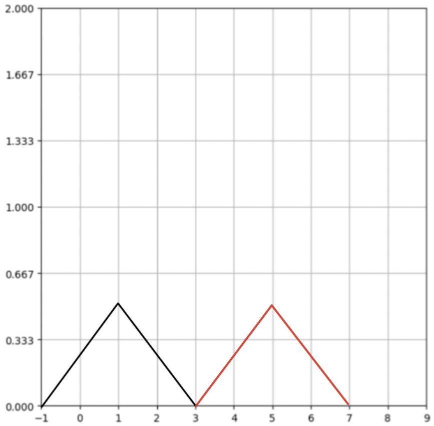


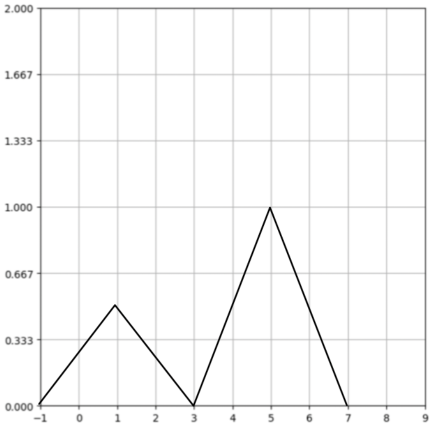
(a) [4 Pts] Kermit wants a KDE plot for data points of `[1, 5, 5]` using a **Triangular Kernel** and $\alpha = 2$. Draw the resulting KDE plot below in the empty plots provided below. Two empty plots are provided for your convenience. **If you choose to use both, please indicate which one graders will grade. Otherwise graders will default to the right plot.** Take note of the pre-ticked axes!



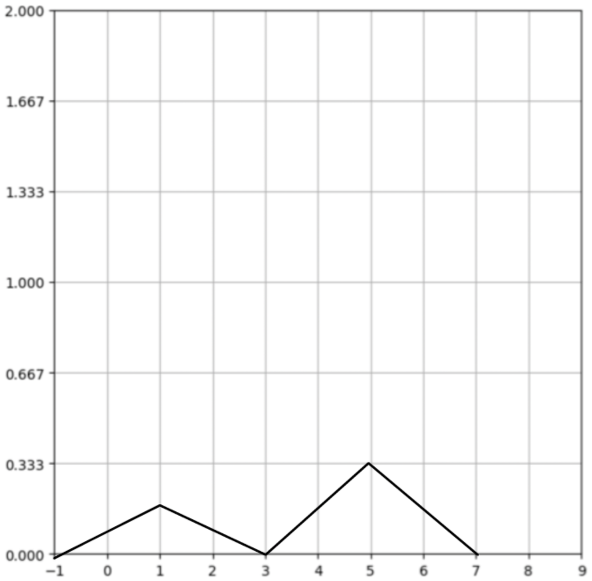


Solution: Notice that a singular Triangular Kernel is centered around x_i and α determines its base. In the provided example of $\alpha = 1$, its base is 2α. This can be alternatively confirmed through finding the bounds through its definition, where $x_i - \alpha \leq x \leq x_i + \alpha$. Thus, since $\alpha = 2$, each individual kernel has a base of 4. Plotting the kernels individually then adding **before normalizing** (each individual kernel's area is still 1), results in the graphs below. Graph 1 shows the individual kernels before adding. Graph 2 shows the kernels added together. Notice that the area under the KDE is 3 ($2 * 0.5 + 2 * 1 = 3$). These intermediate steps need not be shown.





Then, normalize by multiplying the density by $\frac{1}{3}$. This changes the plot by scaling down the kernel heights by a factor of $\frac{1}{3}$. This results in the final plot below. Notice how the area under the KDE properly sums to 1 ($\frac{1}{6} * 2 + \frac{1}{3} * 2 = 1$).



5 We did it, Trader Joe [12 Points]

Jake is obsessed with shopping at Trader Joe's. He puts together a `Pandas DataFrame` called `joe` that contains selling data from various Trader Joe's from 2010-2024. You can find the descriptions of its columns below. We have also provided the first couple of rows.

- `Name`: The name of the item sold (type = `str`).
- `TransactionID`: The transaction ID that the purchase was part of. Each transaction is assigned an unique ID (type = `numpy.int64`).
- `CustomerID`: ID of the customer that made the purchase (type = `numpy.int64`).
- `Price`: The price of the item purchased (type = `numpy.float64`).
- `Location`: The Trader Joe's location where the item was purchased. (type = `str`).

	Name	Date	TransactionID	CustomerID	Price	Location
0	Cacio e Pepe	7-1-2022	94112	612341	5.50	University
1	Chili Lime Tortilla Chips	7-1-2022	94112	612341	2.50	University
2	Apple Cider	4-25-2019	25324	841924	2.75	Rockridge
3	Hashbrowns	2-18-2024	41234	415123	3.25	University
4	Cacio e Pepe	1-1-1970	34123	821832	5.50	Emeryville

(a) [2 Pts] Assuming each customer only buys one of each item at max per transaction, select the minimum number of columns that would form the primary key for the `joe` table.

□ **`Name`**

□ `Date`

□ **`TransactionID`**

□ `CustomerID`

□ `Price`

□ `Location`

(b) [1 Pt] Jake wants to conduct some analysis utilizing the `Date` column, so he decides to add another column that contains the year of the transaction. Which of the following would complete the code below and add the year as its own new column? **Select one.**

`joe[‘Year’] =` ______________________________

○ `joe['Date'].str.split('-').item(2).astype(int)`

○ `joe['Date'].str.extract(r'-\d+$').astype(int)`

○ **`joe['Date'].str.split('-').str[2].astype(int)`**

○ `joe['Date'].str.extract(r'(-\d+)$').astype(int)`

Solution:
Option 1 is incorrect. `item` used incorrectly.

Option 2 is incorrect. There are no capture groups.

Option 3 is correct.

Option 4 is incorrect. Here there is a capture group, but the dash is included.

(c) [1 Pt] After looking through the data, Jake notices that around 4% of the rows have the date `1-1-1970`. Select the best option below on what Jake should do regarding these data points. **Select one.**

○ Replace the `1-1-1970` values with `NaN`.

○ Replace the `1-1-1970` values with the mean date across all rows.

○ **Drop the rows with `1-1-1970` values.**

○ Nothing.

Solution:
Credit was given to all options as it is ambiguous what the best course of action is here. It depends on what the analysis / goals are.

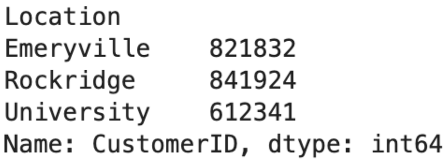
(d) [2 Pts] Jake wants to see who the Trader Joe's Superfans are for each location. A Superfan is a customer who has purchased the most items from a particular location. Assuming there are no ties between any two customers, fill in the code below so that `superfan_custIDs` is assigned to a `Series` where the index is each location and the value is that location's Superfan's `CustomerID`.

```
superfans = joe.groupby(['Location', 'CustomerID']).size() \
                .reset_index(name='Count')

superfans = superfans.sort_values([’Location’, ’Count’], ascending=False)

superfan_custIDs = ___________________(A)___________________
```

The desired output (`superfan_custIDs`):



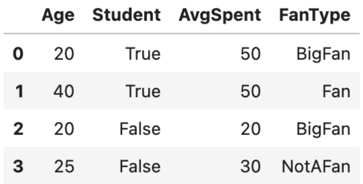
```
Location
Emeryville    821832
Rockridge     841924
University    612341
Name: CustomerID, dtype: int64
```

Fill in the blank `(A)`

Solution:
```
superfans.groupby('Location')['CustomerID'].first()
```

(e) [3 Pts] Jake goes to the University TJ's to collect some data. He asks people for their age (`Age`), if they're a student or not (`Student`), how much they spend on average each TJ's trip (`AvgSpent`), and how big a TJ's fan they are (`FanType`).

He's curious if he can use age (`Age`) and whether or not someone's a student (`Student`) to predict what kind of fan they are (`FanType`). Jake's decides to fit a decision tree using **weighted entropy** to determine splits. Partway through fitting, his decision tree code breaks and Jake must construct the final split himself. The table below shows the samples in the node Jake must split.



	Age	Student	AvgSpent	FanType
0	20	True	50	BigFan
1	40	True	50	Fan
2	20	False	20	BigFan
3	25	False	30	NotAFan

What feature and split (β) should Jake choose? Circle one feature and write the split value β and weighted entropy of your choice on the provided lines (be sure to include units!).

Show all work in the space below in order to receive credit. Note: $\log_2(\frac{1}{3}) = -1.58$.

Solution: Recall the following two definitions:

$$\text{Entropy} = -\sum_C p_C \log_2(p_C) \text{ and Weighted Entropy } = \frac{N_1 S(X) + N_2 S(Y)}{N_1 + N_2}$$

Additionally,

- $\log_2(\frac{1}{3}) = -1.58$ (Given)
- $\log_2(\frac{2}{3}) = \log_2(\frac{1}{3} \cdot 2) = \log_2(\frac{1}{3} \cdot 2) = \log_2(\frac{1}{3}) + \log_2(2) = -1.58 + 1 = -0.58$ (By log rules: $\log(AB) = \log(A) + \log(B)$)

A trick with this one was to remember that pure nodes have an entropy of 0 and notice that a split of `Age` somewhere in [20, 25), e.g. 22.5, resulted in a pure node of size two (and a node with two samples with different labels).

The only other split that resulted in a pure node is `Age` somewhere in [25, 40), e.g. 32.5, but that is a pure node of size one and then a node of made up of three samples (two of one label and one of another). A node with three samples and two labels does have lower entropy than a node with two samples and two labels ($\frac{2.74}{3}$ vs 1) but because we look at *weighted* entropy across the two nodes, the split that results in a pure node with two samples still gives a lower weighted entropy. This was slightly trickier to see without doing the math.

The possible splits (nodes are WLOG):

- `Age` at somewhere in [20, 25), e.g. 22.5:
  resulting in nodes $X = \{\texttt{BigFan}, \texttt{BigFan}\}$ and $Y = \{\texttt{Fan}, \texttt{NotAFan}\}$
  $S(X) = -1 \cdot \log_2(1) = -1 \cdot 0 = 0$ bits
  $S(Y) = -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = -(\frac{1}{2} \cdot -1) - (\frac{1}{2} \cdot -1) = \frac{1}{2} + \frac{1}{2} = 1$ bit
  $L = \frac{(2 \cdot 0) + (2 \cdot 1)}{2+2} = \frac{1}{2}$ bits

- `Age` at somewhere in [25, 40), e.g. 32.5:
  resulting in nodes $X = \{\texttt{BigFan}, \texttt{BigFan}, \texttt{NotAFan}\}$ and $Y = \{\texttt{Fan}\}$
  $S(X) = -\frac{2}{3} \cdot \log_2(\frac{2}{3}) - \frac{1}{3} \cdot \log_2(\frac{1}{3}) = -\frac{2}{3}(-0.58) - \frac{1}{3}(-1.58) = \frac{1.16}{3} + \frac{1.58}{3} = \frac{2.74}{3}$ bits
  $S(Y) = -1 \cdot \log_2(1) = -1 \cdot 0 = 0$ bits
  $L = \frac{(3 \cdot \frac{2.74}{3}) + (1 \cdot 0)}{3+1} = \frac{2.74}{4} > \frac{1}{2}$ bits

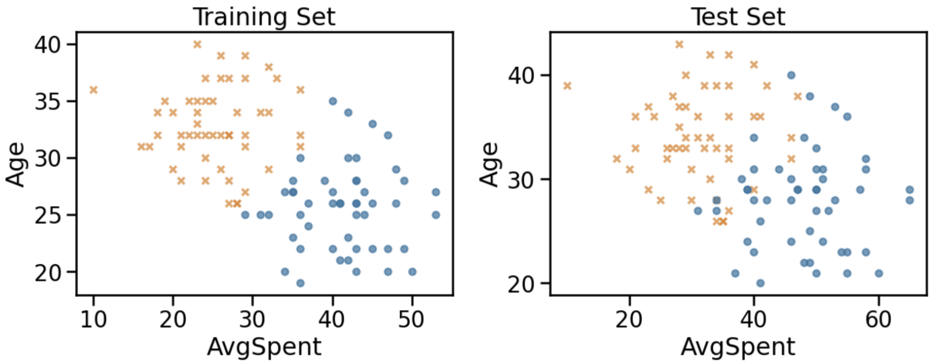
- `Student` at `True` or `False`:
  resulting in nodes $X = \{\texttt{BigFan}, \texttt{Fan}\}$ and $Y = \{\texttt{BigFan}, \texttt{NotAFan}\}$
  $S(X) = -\frac{1}{2} \log_2\left(\frac{1}{2}\right) - \frac{1}{2} \log_2\left(\frac{1}{2}\right) = -\frac{1}{2}(-1) - \frac{1}{2}(-1) = 1$ bits
  $S(Y) = S(X) = 1$ bit
  $L = \frac{(2 \cdot 1) + (2 \cdot 1)}{2+2} = \frac{4}{4} = 1$ bit

After actually working the math out, we can see that the split that minimizes the weighted entropy is

Feature = `Age`
β = somewhere in [20, 25), e.g. 22.5
Weighted Entropy = $\frac{1}{2}$ bits.

Note: We awarded to credit for people who picked Feature = `Age` and $\beta = 25$ as long as it was clear that they meant to split the node between 20 and 25 (e.g. shown work, correct weighted entropy).

(f) [1 Pt] Boyu goes to collect more data but he only asks students if they are a fan or not. In order to combine their datasets, Jake converts his dataset to be binary as well (BigFan and Fan = Fan vs NotAFan). Boyu and Jake split their combined dataset into a training set and test set. The training and test sets are plotted below for `Age` and `AvgSpent`.



Which of the following model(s) are guaranteed to achieve perfect accuracy on the training set? **Select all that apply.**

□ Logistic regression

□ **Decision tree**

□ Random forest

□ None of the above

Solution: Option 1 is incorrect. The data is not linearly separable so logistic regression would not be able to achieve perfect accuracy.

Option 2 is correct. Since there are no samples with the exact same features and different labels, we are guaranteed to achieve perfect accuracy on the training set with a decision tree.

Option 3 is incorrect. Due to the randomness in the random forest model (bootstrapping for each tree, random sample of features at each node), we are not guaranteed to achieve perfect accuracy on the training set.

(g) [1 Pt] Which of the following model(s) are guaranteed to achieve perfect accuracy on the test set? **Select all that apply.**

□ Logistic regression

□ Decision tree

□ Random forest

□ **None of the above**

Solution: We are never *guaranteed* to achieve perfect accuracy on the test set with any model (not without additional knowledge of/assumptions on the test set).

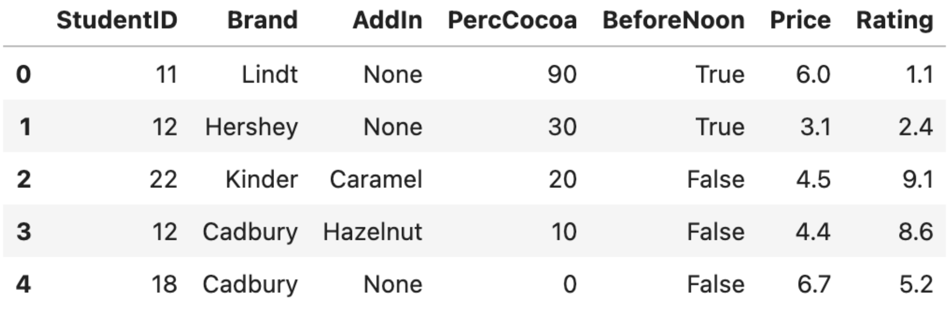
(h) [1 Pt] List two ways in which using random forests decreases variance (as compared to decision trees).

Solution:

- Multiple trees
- Bootstrapping sample for each tree
- Random sample of features at each node.

6 One-Hot-chocolatE [10 Points]

Gisella is curious about the chocolate preferences of U.S. university students. After sourcing chocolate samples from 10 different brands, she goes around Cal campus asking students to rate the chocolate. No student tries the same chocolate sample twice. At the end of the semester she compiles her data into a `Pandas DataFrame` with 88 rows. You can find the descriptions of its columns in part (f). The first couple rows are shown below:



	StudentID	Brand	AddIn	PercCocoa	BeforeNoon	Price	Rating
0	11	Lindt	None	90	True	6.0	1.1
1	12	Hershey	None	30	True	3.1	2.4
2	22	Kinder	Caramel	20	False	4.5	9.1
3	12	Cadbury	Hazelnut	10	False	4.4	8.6
4	18	Cadbury	None	0	False	6.7	5.2

(a) [2 Pts] Gisella starts out by trying to predict `Rating` based on `PercCocoa` using a simple linear regression model:

$$\texttt{Rating} = \theta_0 + \theta_1 \texttt{PercCocoa}$$

Based on the table below and $r = 0.5$, compute the SLR model parameters, $\hat{\theta}_0$ and $\hat{\theta}_1$.

	PercCocoa	Rating
mean	27.7	7.5
std	1.5	3

Solution:
$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x} = 0.5 \cdot \frac{3}{1.5} = 0.5 \cdot 2 = 1$
$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} = 7.5 - 1 \cdot 27.7 = -20.2$
$\hat{\theta}_0 = -20.2, \hat{\theta}_0 = 1$

(b) [1 Pt] Let's assume $\theta_1 = C$. Would it be correct to say: "A unit change in percent cocoa causes an increase in rating of C."? **Select one.**

○ Yes

○ [✓ **No**]

Solution: No, this is a causal statement which we want to avoid stating unless we have further information. This is something we discussed in lecture.

(c) [1 Pt] Gisella is curious if the true parameter θ_1 is 0 or not. She defines the null hypothesis as $\text{H}_0 : \theta_1 = 0$ and the alternative hypothesis as $\text{H}_\text{A} : \theta_1 \neq 0$. Gisella goes ahead with using the bootstrap and gets the following 95% confidence interval: [-0.02, 8]. Should she reject or fail to reject the null hypothesis at cutoff of 5%? **Select one.**

○ Reject

○ **Fail to reject**

Explain your answer to the previous part. No credit will be awarded to the previous part if no explanation is provided. *Do not use more than two sentences.*

Solution: Fail to reject, 0 is in CI

(d) [1 Pt] Gisella decides to use all the other features to predict `Rating`. She has a feeling she should one-hot-encode some of the features but it turns out it's *really* expensive to perform the one-hot-encoding preprocessing step (for each original feature).
Here is a description of the columns:

- `Brand`: Brand of the chocolate sample (type = `str`).
- `AddIn`: What additional ingredients the chocolate sample has added in (type = `str`).
- `PercCocoa`: Percent of cocoa in chocolate sample (%) (type = `numpy.int64`).
- `BeforeNoon`: Whether the student tried the sample before or after noon (type = `numpy.bool_`).
- `Price`: Price of chocolate sample per 10oz (type = `numpy.float`).
- `Rating`: Student's rating of chocolate sample out of 10 (type = `numpy.float`).

Which of the following features should she one-hot-encode? **Select all that apply.**

□ `Brand`

□ `AddIn`

□ `PercCocoa`

□ `BeforeNoon`

□ `Rating`

Solution: `Brand` and `AddIn` are categorical features with more than two values so we should one hot encode them. `PercCocoa` and `Rating` are numerical so we do not need to one hot encode them. Because it's *really* expensive to perform the one hot encoding and it's redundant to do so (boolean feature), we should not one hot encode `BeforeNoon`.

(e) [1 Pt] Gisella change her mind, she now wants to use a single feature, `Brand`, to predict `Rating`. She does some preprocessing including one-hot-encoding `Brand`. She then goes to run the following line of code:

```
choc_reg = LinearRegression().fit(X, y)
```

How many columns should `X` have? Why? *State the number of columns and explain in no more than two sentences.*

Solution: 9 columns. After one-hot-encoding she will have 10 columns, one for each brand, however she needs to remove a column otherwise the features will be linearly dependent.
Note: `LinearRegression()` includes an intercept term by default (`fit_intercept=True`).

(f) [2 Pts] Gisella finally decides on a feature matrix and determines that the columns are not linearly dependent. She decides to include an intercept term in her design matrix and then fits a linear regression using sklearn. Which of the following are true? **Select all that apply.**

- □ **If we multiply the column of the feature `Price` by our residuals, we will get the 0 vector.**
- □ **If we add up the residuals, the sum will be 0.**
- □ **If we multiply our design matrix by non-optimal parameters, the resulting vector is not orthogonal to our residual vector.**
- □ The dot product of our true y values and our residual vector is the 0 vector.
- □ None of the above.

Solution:
Option 1 - We gave credit to both choices as students may have assumed it was a dot product and thought it was the 0 scalar.

Option 2 is correct.
Option 3 - We gave credit to both choices as it was unclear which residual vector we were referring to.
Option 4 is incorrect.

(g) [2 Pts] Gisella is now considering centering (subtracting the respective means from) her features and response variable. Which of the following are true? **Select all that apply.**

- □ If Gisella centers her features, the resulting intercept will be 0.
- □ If Gisella centers her response variable, the resulting intercept will be 0.
- □ **If Gisella centers her features and her response variable, the resulting intercept will be 0.**

□ **If Gisella centers her features, the resulting slope will be the same as before (without centering).**

□ **If Gisella centers her response variable, the resulting slope will be the same as before (without centering).**

□ **If Gisella centers her features and her response variable, the resulting slope will be the same as before (without centering).**

□ None of the above.

7 Silly 🦭 🦭 🦭 [11 Points]

Alana wants to predict flipper length of seals in the Monterey Bay Aquarium using their weight. She plans on using $\hat{Y}(x) = f_\theta(x) = x * \theta$ as her model, a simple linear regression model with no intercept term. Throughout the question, assume that we see noisy observations $Y = g(x) + \epsilon$, where $g(x)$ denotes the fixed, true underlying relationship between x and y, and ϵ is random noise with $\mathbb{E}[\epsilon] = 0$ and $Var(\epsilon) = 1$.

(a) [2 Pts] Which of the following expressions below are random variables? **Select all that apply.**

- □ $f_{\hat{\theta}}(x) - E[f_{\hat{\theta}}(x)]$
- □ $E[f_{\hat{\theta}}(x)] - g(x)$
- □ $E[Y]$
- □ $E[f_{\hat{\theta}}(x)]$
- □ $f_{\hat{\theta}}(x)$
- □ Y
- □ None of the above

(b) [2 Pts] Which of the following statements about model risk, bias and variance are true? **Select all that apply.**

- □ **The model variance is caused by the randomness present in our predictions $f_{\hat{\theta}}(x)$.**
- □ **Observation variance comes from the noise (ϵ) present in Y. This noise is irreducible.**
- □ **Since the model risk is a fixed number, increasing the model variance leads to a lower model bias and vice versa.**
- □ **If we collect more data and fix the model complexity, we can reach a lower model variance and bias.**
- □ None of the above.

(c) [2 Pts] Alana is interested in adding a new feature S, which indicates whether the seal is silly ($S = 1$) or not ($S = 0$), to her model. Since Sillymaxxing is all the rage, Alana wants to know the expected value of the square of silliness, $E(S^2)$. Assume that 80% of the seals are silly and 20% are not. Calculate $E(S^2)$. **Simplify your answer.**

Solution:

$$\begin{aligned}
\text{Var}(S) &= \mathbb{E}(S^2) - (\mathbb{E}(S))^2 \\
\text{Var}(S) &= p(1-p) = 0.8 \cdot 0.2 = 0.16 \\
\mathbb{E}(S) &= p = 0.8 \\
\mathbb{E}(S^2) &= \text{Var}(S) + (\mathbb{E}(S))^2 = 0.16 + (0.8)^2 = 0.16 + 0.64 = 0.8
\end{aligned}$$

For the following questions, you can assume that we have collected additional features and switched to modeling with OLS.

(d) [2 Pts] Instead of using L1 or L2 regularization, Zekai proposes to use the L0 norm to regularize instead. The L0 norm, denoted by $\|\theta\|_0$, is defined as the number of nonzero entries in θ. What happens if we run linear regression with L0 regularization? Can we use gradient descent (with a sufficiently small learning rate)? Explain your answer.

Solution: L_0 regularization induces sparsity, if we find the true global optimum. However, it might not always run with Gradient Descent, as the gradient will either be 0 or `nan` (nonexistent). Points were also awarded if students mentioned nonconvexity, as Gradient Descent might not converge with L_0 regularization.

(e) [3 Pts] In lecture, we have seen that we can model our observations $Y = X\theta_g + \epsilon$ where θ_g corresponds to the parameters of the true underlying relationship and the random variable ϵ denotes the observational noise where $\mathbb{E}[\epsilon] = 0$. Show that the OLS solution $\hat{\theta}_{OLS}$ is an unbiased estimator for θ_g. In other words, show that $\mathbb{E}[\hat{\theta}_{OLS}] = \theta_g$.

Solution:

$$
\begin{aligned}
Y &= X\theta_g + \epsilon \\
\hat{\theta}_{OLS} &= (X^TX)^{-1}X^TY \\
\hat{\theta}_{OLS} &= (X^TX)^{-1}X^T(X\theta_g + \epsilon) \\
\hat{\theta}_{OLS} &= (X^TX)^{-1}(X^TX\theta_g + X^T\epsilon) \\
\hat{\theta}_{OLS} &= \theta_g + (X^TX)^{-1}X^T\epsilon \\
\mathbb{E}[\hat{\theta}_{OLS}] &= \mathbb{E}[\theta_g + (X^TX)^{-1}X^T\epsilon] \\
\mathbb{E}[\hat{\theta}_{OLS}] &= \mathbb{E}[\theta_g] + \mathbb{E}[(X^TX)^{-1}X^T\epsilon] \\
\mathbb{E}[\hat{\theta}_{OLS}] &= \theta_g + (X^TX)^{-1}X^T\mathbb{E}[\epsilon] \\
\mathbb{E}[\hat{\theta}_{OLS}] &= \theta_g + (X^TX)^{-1}X^T \cdot 0 \\
\mathbb{E}[\hat{\theta}_{OLS}] &= \theta_g
\end{aligned}
$$

End of Graded Questions

8 Congratulations [0 Pts]

Congratulations! You have completed the Final Exam.

- **Make sure that you have written your student ID number on *each page* of the exam.** You may lose points on pages where you have not done so.
- Make sure to **sign the Honor Code** on the cover page of the exam. Failure to comply may result in the exam being invalidated.

[Optional, 0 pts] Draw a picture (or graph) describing your experience in Data 100.