Data C100/200, Final

Fall 2024

Name: ____

Email: ____________________________________@berkeley.edu

Student ID: ____

Exam Room:______________________, Seat Number: ______________

Instructions:

This exam consists of **89 points** spread out over **8 questions** and the **Honor Code certification**. The exam must be completed in **170 minutes** unless you have accommodations supported by a DSP letter. Note that you should:

- **Select one choice** for questions with **circular bubbles**. There is always at least one correct answer. Please **fully** shade in the circle to mark your answer.
- For all math questions, **please simplify your answer**. Please also **show your work** if a large box is provided.
- For all coding questions, you may use commas and/or one or more function calls in each blank.
- **You MUST write your Student ID number at the top of each page.**
- You should not use a calculator, scratch paper, or notes you own other than the reference sheets distributed at the beginning of the exam.

For all Python questions, you may assume `Pandas` has been imported as `pd`, `NumPy` as `np`, the Python RegEx library as `re`, `matplotlib.pyplot` as `plt`, and `seaborn` as `sns`.

Honor Code [1 Pt]:

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 Welcome to the Olympics [10 points]

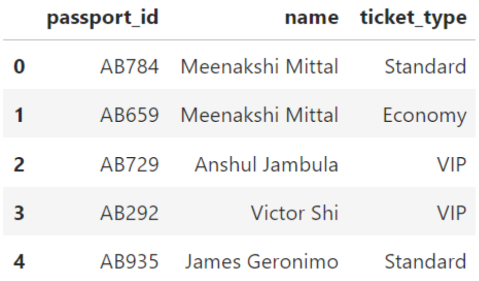
DATA C100 staff are hosting a Winter Olympics event. All attendees and their tickets are described in the `DataFrame` named `ticket`. Each attendee has one ticket. If attendees bought jackets in the event shop, the number of jackets they bought is recorded in the `DataFrame` named `jacket`. Each row of `jacket` describes one unique attendee.

The columns of `ticket` are as described below:

- `passport_id`: The attendee's passport ID. (type = `str`)
- `name`: The attendee's name. Attendees with different passport IDs may have the same name. (type = `str`)
- `ticket_type`: The attendee's ticket type. Values can be either `"Economy"`, `"Standard"`, or `"VIP"`, listed in order of increasing price. (type = `str`)

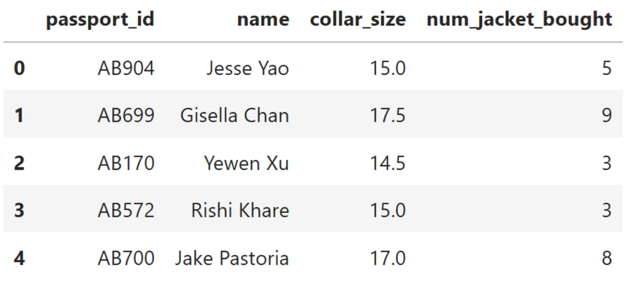
The columns of `jacket` are as described below:

- `passport_id`: The attendee's passport ID. (type = `str`)
- `name`: The attendee's name. Attendees with different passport IDs may have the same name. (type = `str`)
- `collar_size`: The attendee's jacket collar size. (type = `np.float32`)
- `num_jacket_bought`: The number of jackets the attendee bought. (type = `np.int64`)



	passport_id	name	ticket_type
0	AB784	Meenakshi Mittal	Standard
1	AB659	Meenakshi Mittal	Economy
2	AB729	Anshul Jambula	VIP
3	AB292	Victor Shi	VIP
4	AB935	James Geronimo	Standard

Above is a sample of `ticket`



	passport_id	name	collar_size	num_jacket_bought
0	AB904	Jesse Yao	15.0	5
1	AB699	Gisella Chan	17.5	9
2	AB170	Yewen Xu	14.5	3
3	AB572	Rishi Khare	15.0	3
4	AB700	Jake Pastoria	17.0	8

Above is a sample of `jacket`

(a) [0.5 Pts] What type of variable is `ticket_type` in `ticket`?

○ Qualitative Nominal [✓ ○ **Qualitative Ordinal**] ○ Quantitative

Solution: `ticket_type` is a qualitative ordinal variable. This is because the values of `ticket_type` can be inherently ordered by its expense.

(b) Xiaorui wants to know about attendee's experiences buying jackets. He sent a survey to all attendees in `ticket`.

(i) [0.5 Pts] What is the sampling frame of this survey?

○ **All attendees who own a ticket to the event.**

○ All attendees who visited the event shop.

○ All attendees who have purchased jackets

Solution: Because the survey was distributed to all attendees of the Olympics event in `ticket`, its sampling frame is attendees of the Olympics event who own a ticket to the event.

(ii) [0.5 Pts] What is the population of this survey?

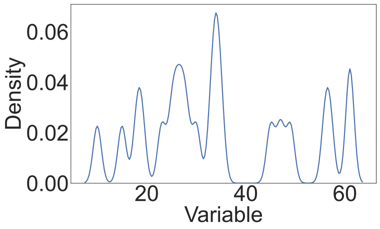
○ All attendees who own a ticket to the event.

○ All attendees who visited the event shop.

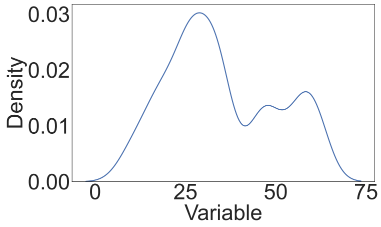
○ **Attendees who have bought jackets in the event shop.**

Solution: The survey's population of interest is the jacket purchasers, hence "Attendees who have bought jackets in the event shop".

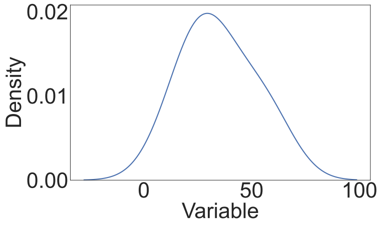
(c) Answer the following questions about Kernel Density Estimation (KDE).



This KDE plot uses bandwidth α = _____ `(x)` _____



This KDE plot uses bandwidth α = _____ `(y)` _____



This KDE plot uses bandwidth α = _____ `(z)` _____

(i) [0.5 Pts] Select the correct statement about kernel functions in a KDE process.

○ A kernel function has to be differentiable across its domain.

○ A kernel function has to be Gaussian.

○ **A kernel function has an area under the curve of 1.**

(ii) [0.5 Pts] Which of the following sequences correctly fills blanks `(x)`, `(y)`, `(z)` in the above plot's labels?

○ **`(x)` is 0.1, `(y)` is 0.5, `(z)` is 1.5**

○ `(x)` is 1.5, `(y)` is 0.5, `(z)` is 0.1

○ `(x)` is 1.5, `(y)` is 0.1, `(z)` is 0.5

(iii) [0.5 Pts] For kernel bandwidth α, let the kernel centered at observation x_i be $K_\alpha(x, x_i)$, and the KDE estimated distribution be $f_\alpha(x)$. When there are n datapoints in the KDE process, what is the correct expression of $f_\alpha(x)$ in terms of $K_\alpha(x, x_i)$ and n?

○ $f_\alpha(x) = \frac{1}{n} K_\alpha(x, x_1)$

○ $f_\alpha(x) = \sum_{i=1}^{n} K_\alpha(x, x_i)$

[✓ ○ $f_\alpha(x) = \frac{1}{n} \sum_{i=1}^{n} K_\alpha(x, x_i)$]

Solution: For (c)(i), the correct choice is "an area under the curve of 1". The kernel function need not be differentiable across its domain and is not confined to being Gaussian.

For (c)(ii), larger bandwidths (α) lead to smoother KDE results. Therefore, the correct choice is the sequence where `(x)` is 0.1, `(y)` is 0.5, and `(z)` is 1.5.

For (c)(iii), we need to normalize and sum over all placed kernels. The third choice corresponds to this operation.

(d) An inner join between `jacket` and `ticket` on `passport_id` contains 52 rows.

Meanwhile, an inner join between `jacket` and `ticket` on `name` contains 20 rows.

Given `jacket` has 92 rows, and `ticket` has 53 rows, answer the following questions about performing **full outer joins** on these two tables. If there is not enough information, write `N/A`.

(i) [1 Pts] What is the number of rows from a full outer join on `passport_id`?

Answer: ______________________

(ii) [1 Pts] What is the number of rows from a full outer join on `name`?

Answer: ______________________

Solution: The solution to these subquestions highly depends on the fact that there is a unique `passport_id` per row. Therefore, we know how many rows from each table are excluded from an inner join on column `passport_id`.

Neither of the tables has a unique `name` per row. Furthermore, the attendee recorded in `jacket` is an unknown subset of the attendee recorded in `ticket`. Therefore, we **do not** know how many rows from each table are excluded from an inner join on column `name`.

Concretely, we know from the inner join on `name` that 20 rows match, but that could be 1 row from `jacket` that matches 20 rows from tickets; or, it could be 20 rows from each table that matched. Due to this ambiguity, we don't know how many remaining rows from each table were not matched in that inner join.

Left blank: 93. There are 52 non-missing equal `name`, 40 other from `jacket`, and 1 other from `ticket`.

Right blank: `N/A`. There is not enough information to deduce the answer.

Aneesh has a broken copy of the `jacket`, called `jacket_miss`. This broken copy has some missing values in `num_jacket_bought`. No other columns have missing values.

(e) Aneesh wants to produce a contour plot that visualizes the joint distribution of `collar_size` and `num_jacket_bought` from `jacket_miss`. Fill all missing values with 0. Do not include any datapoint where the value of `num_jacket_bought` is higher than 10. Fill in the blanks to achieve this:

```
df_to_plot = jacket_miss.______________(i)______________
______________(ii)______________(
    data = df_to_plot______________(iii)________________,
    x = "collar_size", y="num_jacket_bought"
)
```

(i) [1 Pts] Fill in blank `(i)`:

Solution: `fillna(0)`

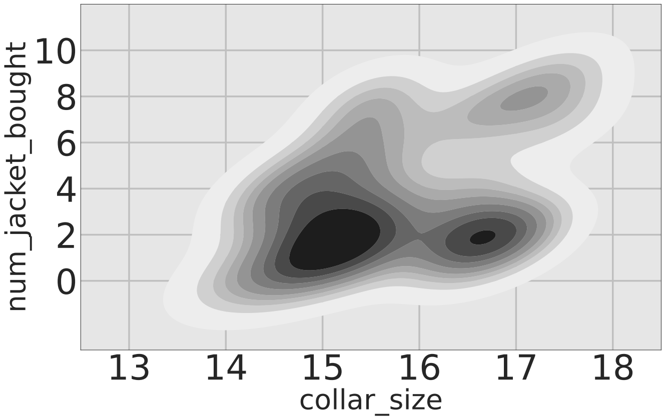
(ii) [0.5 Pts] Fill in blank `(ii)`:

Solution: `sns.kdeplot` or `sns.jointplot`. The latter is out of our reference sheet, but we will accept it with full credit.

(iii) [1 Pts] Fill in blank `(iii)`:

Solution: `[df_to_plot["num_jacket_bought"] <= 10]`
If use `jointplot` above, must attach `, kind="kde"` after the indexing.

(f) [1 Pt] Without using the above code, Aneesh produces his own contour plot of the joint distribution, as shown below:



Evaluate the following statements about the above plot.

○ True [✓ ○ **False**] When `num_jacket_bought` is 2, `collar_size` has a unimodal distribution.

○ True [✓ ○ **False**] To impute the missing values of `num_jacket_bought`, we may interpolate those missing values using its average.

Solution: The first statement is false. The distribution of `collar_size` at the assigned value of `num_jacket_bought` is bimodal with peaks at a region between collar size 15, and at approximately a collar size 16.7.

The second statement is false. The distribution of `num_jacket_bought` is skewed, so the average is not a reliable or representative estimate for the missing values.

(g) Fill in the blanks to create a `DataFrame` with the following properties:

1. Contains all rows and columns from `ticket`.
2. Includes a column `num_jacket_bought` that contains the number of jackets bought by attendees.
3. If an attendee did not buy any jacket, the corresponding value in `num_jacket_bought` for that attendee should be `NaN`.

Note: Some attendees did not buy any jackets.

```
ticket.__________________(i)__________________(
     jacket,
     __________________(ii)__________________
)
```

(i) [0.5 Pts] Fill in blank `(i)`:

Solution: `merge`
`join` is not accepted as it cannot lead to correct results.

(ii) [1 Pts] Fill in blank `(ii)`:

Solution: `how="left", on="passport_id"`

2 I Will Have Order [11 points]

All code for this question must be written as SQL queries.

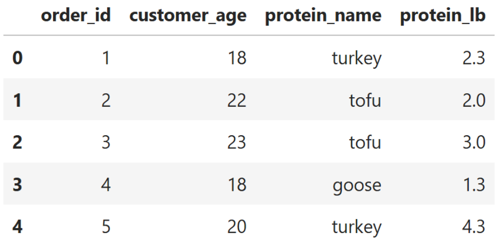
Before Thanksgiving, Claire invited her friends to a group dinner order. Each order contains one protein type. Related records are organized into two tables: `orders` and `protein_menu`.

`orders` contains orders. Its columns are:

- `order_id`: The ID of the order. (type=`INT`)
- `customer_age`: The age of the order's buyer. (type=`INT`)
- `protein_name`: The type of protein ordered. (type=`VARCHAR`)
- `protein_lb`: Pounds of protein bought in the order. (type=`FLOAT`)

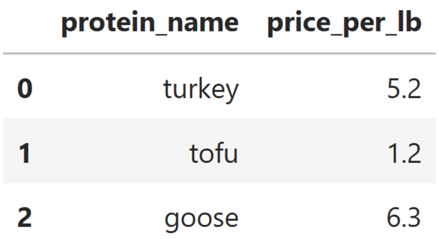
`protein_menu` describes protein options and their prices. Its columns are:

- `protein_name`: The type of protein. (type=`VARCHAR`)
- `price_per_lb`: The price of the protein per pound. (type=`FLOAT`)



	order_id	customer_age	protein_name	protein_lb
0	1	18	turkey	2.3
1	2	22	tofu	2.0
2	3	23	tofu	3.0
3	4	18	goose	1.3
4	5	20	turkey	4.3

A sample of `orders`



	protein_name	price_per_lb
0	turkey	5.2
1	tofu	1.2
2	goose	6.3

The entire `protein_menu`

(a) [0.5 Pts] With `protein_menu` as the foreign table, list the foreign key(s) of `orders` in the following box:

Answer: ____

Solution: `protein_name` is a foreign key in `orders` because it is a primary key in `protein_menu`.

The following definition is used for the next page in subpart 2(b).

The **customer type** of an order is defined as “upper” if the buyer is at least 21 years old, “lower” if the buyer is between 18 and 20 years old inclusive, and “child” otherwise.

(b) Claire wants to create a SQL table called `orders_with_menu`. For each order, this table contains the `order_id`, `customer_type` (containing customer type of each order), `protein_name`, `protein_lb`, and `price_per_lb`.

The columns in the output table must be ordered according to the order described above. Fill in the blanks to achieve this.

Note: All positions denoting blank `(i)` should be filled in with the same content.

```
SELECT od.order_id,
CASE
    ___________(i)___________ 18 THEN 'child'
    ___________(i)___________ 21 THEN 'lower'
    ___________(ii)___________ END AS customer_type,
  ___________(iii)___________
FROM ___________(iv)___________ protein_menu AS m
  ___________(v)___________;
```

(i) [0.5 Pts] Fill in blank `(i)`:

Solution: `WHEN od.customer_age <`

This space is kept empty for scanning purposes

(ii) [0.5 Pts] Fill in blank `(ii)`:

Solution: `ELSE 'upper'`

(iii) [1 Pts] Fill in blank `(iii)`:

Solution: `od.protein_name, od.protein_lb, m.price_per_lb`

(iv) [1 Pts] Fill in blank `(iv)`:

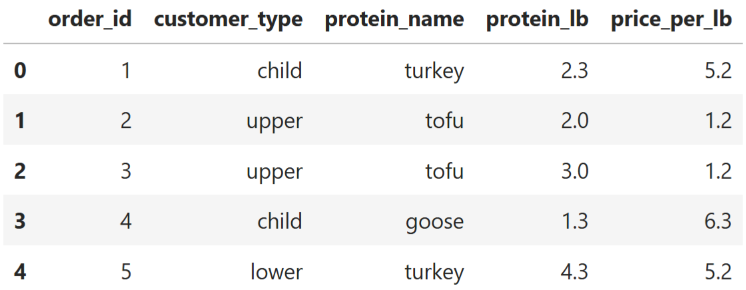
Solution: Set (A): `orders AS od LEFT JOIN`
Set (B): `orders AS od,`

Solution: Alternate solution ruling for set (A): `RIGHT JOIN`, `INNER JOIN`, and `JOIN` are all accepted.

(v) [1 Pts] Fill in blank `(v)`:

Solution: Set (A): `ON od.protein_name = m.protein_name`
Set (B): `WHERE od.protein_name = m.protein_name`

For subparts 2(c) and 2(d), you may assume the table `orders_with_menu` from subpart 2(b) has been correctly created. A sample of this table follows:



	order_id	customer_type	protein_name	protein_lb	price_per_lb
0	1	child	turkey	2.3	5.2
1	2	upper	tofu	2.0	1.2
2	3	upper	tofu	3.0	1.2
3	4	child	goose	1.3	6.3
4	5	lower	turkey	4.3	5.2

(c) Complete the SQL query below to create a table with the following conditions:

1. Contains all columns from `orders_with_menu`.
2. Adds a new column called `expense` that contains the total price of the order.
3. Includes only rows that have a value of `expense` larger than 10.

```
SELECT _______________(i)_______________ FROM orders_with_menu
_______________(ii)_______________;
```

This space is kept empty for scanning purposes

(i) [1 Pts] Fill in blank `(i)`:

Solution: `*, protein_lb * price_per_lb AS expense`

(ii) [1 Pts] Fill in blank `(ii)`:

Solution: `WHERE expense > 10`

(d) Complete the SQL query below to find the customer type with the highest total turkey purchases in pounds, excluding customer types with fewer than 5 orders of turkey. Assume each customer type has a unique total turkey purchase in pounds.

Hint: Your output should only contain one row containing the customer type you found.

```
SELECT customer_type FROM orders_with_menu
_______________(i)_______________
_______________(ii)_______________
_______________(iii)_______________
ORDER _______________(iv)_______________
_______________(v)_______________;
```

Solution: Please note that several alternative solutions exist for each item. This can be found in your grading rubric.

(i) [1 Pts] Fill in blank `(i)`:

Solution: `WHERE protein_name == 'turkey'`. For this context, == is interchangeable with `LIKE`.

This space is kept empty for scanning purposes

(ii) [1 Pts] Fill in blank `(ii)`:

Solution: `GROUP BY customer_type` The GROUP BY-HAVING clause must come after WHERE.

(iii) [1 Pts] Fill in blank `(iii)`:

Solution: `HAVING COUNT(*) >= 5`

(iv) [1 Pts] Fill in blank `(iv)`:

Solution: `BY SUM(protein_lb) DESC`

(v) [0.5 Pts] Fill in blank `(v)`:

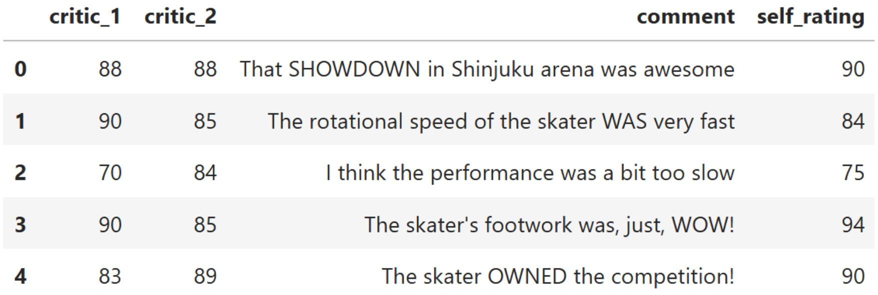
Solution: `LIMIT 1`

3 Skater and Critic [18 points]

We have `skate_record`, a `DataFrame` where each row describes a skating performance with the following columns:

- `critic_1`: Critic 1's rating on the performance. (type = `numpy.int64`)
- `critic_2`: Critic 2's rating on the performance. (type = `numpy.int64`)
- `comment`: An audience comment on that performance. (type = `str`)
- `self_rating`: The skater's self-rating on the performance. (type = `numpy.int64`)

All ratings are between 0 to 100 inclusive. A sample of `skate_record` is shown below:



	critic_1	critic_2	comment	self_rating
0	88	88	That SHOWDOWN in Shinjuku arena was awesome	90
1	90	85	The rotational speed of the skater WAS very fast	84
2	70	84	I think the performance was a bit too slow	75
3	90	85	The skater's footwork was, just, WOW!	94
4	83	89	The skater OWNED the competition!	90

(a) Given the following statistics of `critic_1`:

Statistic	Mean	Standard Deviation	Median
Value	85	13	82

For a constant model that predicts `critic_1`, fill in the correct values:

(i) [0.5 Pts] What is the model's optimal predicted value with L1 loss?

$\hat{\theta} =$ ______________________

(ii) [0.5 Pts] What is the model's optimal predicted value with L2 loss?

$\hat{\theta} =$ ______________________

Solution: The optimal predicted value under an L1 loss is the median of the response variable's known values: 82.

The optimal predicted value under an L2 loss is the mean of the response variable's known values: 85.

(b) [1 Pt] Training the above model with L2 loss, Rose found a model bias of 3, a model variance of 5, and a model risk of 15. Calculate the observational variance of this model.

Observational Variance = ____________

Solution: The observational variance is equivalent to the model risk subtracted by the model bias **squared** and the model variance. The answer is then $15 - 3^2 - 5 = 1$.

(c) Fill in the blanks to add the two following columns to `skate_record`:

- `number_of_W`: The number of words where an uppercase `W` is surrounded by 1 or more adjacent uppercase letters on each side (e.g., `UWU`, `OWNed`, `SHOWDOWN`) in comments.
- `mean_Q`: The average of two critics' ratings for each row.

You are not allowed to use functions from the library `re` in this subpart (3(c)).

```
skate_record["number_of_W"] = (
    ____________(i)____________.findall(____________(ii)____________)
        .____________(iii)____________()
)
____________(iv)____________ = (____________(v)____________) / 2
```

(i) [1 Pts] Fill in blank `(i)`:

Solution: `skate_record["comment"].str`

This space is kept empty for scanning purposes

(ii) [2 Pts] Fill in blank `(ii)`:

Solution: Model solution: `r"\S*[A-Z]+W[A-Z]+\S*"`
For detailed credit ruling, read the following box.

Solution: Partial credit rulings:

- `r"[A-Z]?W[A-Z]?"` will receive 0.5/2 points.
- `r"[A-Z]*W[A-Z]*"` will receive 1/2 points.
- `r"[A-Z]+W[A-Z]+"` will receive 1.5/2 points.
- Replacing + with `{1,}` is permissible.
- Missing the raw string along any backslash usages is a 0.5-point deduction.
- `r"[A-Z]+W[A-Z]+\S*"` and `r"\S*[A-Z]+W[A-Z]+"` are full credit solutions.
- For the above solution, replacing `\S` with `\w` or `[a-z]` is acceptable too.

(iii) [0.5 Pts] Fill in blank `(iii)`:

Solution: `str.len`

(iv) [0.5 Pts] Fill in blank `(iv)`:

Solution: `skate_record["mean_Q"]`

(v) [0.5 Pts] Fill in blank `(v)`:

Solution: `skate_record["critic_1"] + skate_record["critic_2"]`

(d) [2 Pts] We now use the following model equation to predict `self_rating`:

$$\widehat{\texttt{self\_rating}} = \hat{\theta}_0 + \hat{\theta}_1 \times \texttt{critic\_1}$$

For each action, determine whether it would change the $\hat{\theta}_1$ value.

○ True ○ **False** Subtracting all values of `critic_1` by its median.

○ **True** ○ False Dividing all values of `critic_1` by 2.

○ **True** ○ False Using the square of `critic_1` as the predictor variable.

○ True ○ **False** Subtracting all values of `self_rating` by its mean.

Solution: Remember that the SLR formula for $\hat{\theta}_1$ in a model equation $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$ would be:

$$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$$

The first statement is false. Subtracting all values of a predictor variable by a constant changes neither its standard deviation nor its correlation with the response variable.

The second statement is true. The proposed operation changes the standard deviation of the predictor variable without changing its correlation with the response variable. Consequently, it changes the regression line's slope.

The third statement is true. The proposed operation changes the predictor variable, and may therefore change the predictor variable's standard deviation and correlation with the response variable.

The fourth statement is false. The reason is as portrayed in the first option.

(e) We now use the following model to predict `self_rating`:

$$\widehat{\texttt{self\_rating}} = \hat{\theta}_0 + \sum_{i=1}^{d} \hat{\theta}_i \times (\texttt{critic\_1})^i$$

and trains it with L2 regularization and hyper-parameter λ. We perform 4-fold cross-validation and find the following table of cross-validation errors for each (λ, d) pair:

	$\lambda = 0.01$	$\lambda = 0.05$	$\lambda = 0.1$
$d = 1$	5.5	1.6	2.5
$d = 2$	5	0.6	3.3
$d = 3$	6.5	3.5	6.3

(i) [1 Pts] How many times do we need to evaluate a model on a validation set to construct the above table?

Answer = ____________________

(ii) [1 Pts] From the above table, which pair of hyperparameters (λ, d) is expected to lead to the highest model bias squared?

Answer: (λ = __________, d = __________)

(iii) [1 Pts] From the above table, which pair of hyperparameters (λ, d) is expected to lead to the highest model variance?

Answer: (λ = __________, d = __________)

Solution: For subpart (e)(i), provided there is a total of 3 options per hyperparameter and the cross-validation is four-fold, the answer is $3 \times 3 \times 4 = 36$.

For subpart (e)(ii) and (e)(iii), note that a higher λ will present a higher model bias, and a higher d provides more model features and will therefore give higher model variance.

Therefore, subpart (e)(ii) has answer $(\lambda = 0.1, d = 1)$.

Then, subpart (e)(iii) has the answer $(\lambda = 0.01, d = 3)$.

(f) [1 Pt] We now train the following model to predict `self_rating`:

$$\widehat{\texttt{self\_rating}} = \hat{\theta}_1 \times (\texttt{critic\_1}) + \hat{\theta}_2 \times (\texttt{critic\_1})^2$$

Evaluate the following statements about training this model with L1 regularization.

○ True ○ **False** The optimal values of $\hat{\theta}_1$ and $\hat{\theta}_2$ have closed-form solutions.

○ True ○ **False** With regularization hyper-parameter λ, the objective function of this model is augmented by adding $\lambda(\hat{\theta}_1 + \hat{\theta}_2)$.

Solution: The first statement is false. L1 regularization has no closed-form solutions.

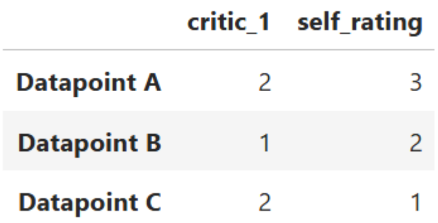
The second statement is false. The augmented term is $\lambda(|\hat{\theta}_1| + |\hat{\theta}_2|)$.

(g) [1.5 Pts] We now train the following model to predict `self_rating`:

$$\widehat{\texttt{self\_rating}} = \hat{\theta}_0 + \hat{\theta}_1 \times (\texttt{critic\_1})^2$$

We find the following optimal parameters for the model: $\hat{\theta}_0 = 2,\ \hat{\theta}_1 = 1$.

We evaluate our model on the following dataset:



	critic_1	self_rating
Datapoint A	2	3
Datapoint B	1	2
Datapoint C	2	1

Using ridge regression with hyper-parameter $\lambda = \frac{1}{3}$, what is the empirical risk of our model on the above dataset?

Answer = ______________________

Solution: Note that we don't regularize the intercept of the model.

The empirical risk of the model can be computed as:

$$\begin{aligned}
R(\vec{\hat{\theta}}) &= \left(\frac{1}{3}\sum_{i=1}^{3}(\hat{y}_i - y_i)^2)\right) + \lambda(\hat{\theta}_1)^2 \\
&= \frac{1}{3}\left[\big((2 + 1 \times 2^2) - 3\big)^2 + \big((2 + 1 \times 1^2) - 2\big)^2 + \big((2 + 1 \times 2^2) - 1\big)^2\right] + \lambda(\hat{\theta}_1)^2 \\
&= \frac{1}{3}\left[3^2 + 1^2 + 5^2\right] + \lambda \times 1^2 \\
&= \frac{35}{3} + \frac{1}{3} = 12
\end{aligned}$$

(h) [1 Pt] We would like to investigate the following model to predict `self_rating`. We train it with L2 loss:

$$\widehat{\texttt{self\_rating}} = \hat{\theta}_0 + \hat{\theta}_1 \times \texttt{critic\_1}$$

To test the null hypothesis that $\theta_0 = 0$, we use bootstrap samples drawn from `skate_record`. Evaluate the following statements about this inference process.

○ **True** ○ False Bootstrap samples are sampled with replacement from the dataset.

○ True ○ **False** If the 95% confidence interval of $\hat{\theta}_0$ from the bootstrapped samples includes 0, then the true value of θ_0 is 0.

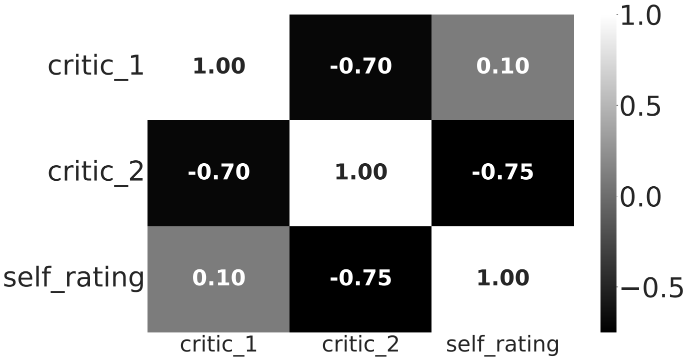
Solution: The first statement is true. Bootstrap samples are defined as samples that are drawn from an equivalently sized or larger sample, sampled with replacement.

The second statement is false. Although we cannot reject the null hypothesis, there is still a possibility that the true parameter value is not 0.

(j) [3 Pts] We now use a subset of `skate_record` to train the following model:

$$\widehat{\texttt{self\_rating}} = \hat{\theta}_0 + \hat{\theta}_1 \times (\texttt{critic\_1}) + \hat{\theta}_2 \times (\texttt{critic\_2})$$

The correlation matrix for the predictor and response variables in our subset of `skate_record` is shown below:



	critic_1	critic_2	self_rating
critic_1	1.00	-0.70	0.10
critic_2	-0.70	1.00	-0.75
self_rating	0.10	-0.75	1.00

Based on the above correlation matrix, evaluate the following statements about our model.

○ **True** ○ False The selected features make our model prone to collinearity.

○ **True** ○ False In our subset, `critic_2` is a more helpful predictor variable than `critic_1`.

○ True ○ **False** If the model's solution is non-unique, `critic_1` must be a scalar multiple of `critic_2` for all rows.

○ True ○ **False** There needs to be at least 2 distinct data points for having a unique model solution.

○ **True** ○ False If the model's solution is unique, standardizing `critic_1` and `critic_2` results in different values for $\vec{\hat{\theta}}$.

○ True ○ **False** If the model's solution is unique, standardizing `critic_1` and `critic_2` results in different values for $\widehat{\texttt{self\_rating}}$

.

Solution: The first statement is true, as the predictor features are highly linearly related ($r = -0.7$).

The second statement is true, as `critic_2` is strongly inversely correlated to the response variable `self_rating`, while `critic_1` is weakly correlated.

The third statement is false. It could be that the weighted sum of critic rating columns is a scalar multiple of the bias column.

The fourth statement is false. The minimum amount of distinct data points for the design matrix to have a chance of being linearly independent is three, as there are three columns in the design matrix.

The fifth statement is true. Standardization would change the standard deviation and scale of the variable, which then changes the linear regression coefficients.

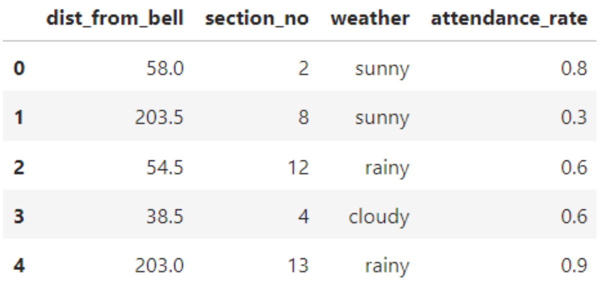
The sixth statement is false. Standardization doesn't change the columnspace of the design matrix, and will therefore result in a model that predicts exactly as before standardization.

4 So... How Many of You Came to the Discussion? [6 points]

We have `discussions`, a `DataFrame` where each row describes one discussion section with the following columns:

- `dist_from_bell`: Distance between discussion section location and the Campanile in meters. (type=`np.float32`)
- `section_no`: The week of the discussion section, between 1 and 13 inclusive. (type=`np.int64`)
- `weather`: Weather during discussion section. Can be `"sunny"`, `"cloudy"`, or `"rainy"`. (type=`str`)
- `attendance_rate`: The attendance rate of the discussion section, between 0 and 1 inclusive. (type=`np.float32`)

A sample of `discussions` is shown below:



	dist_from_bell	section_no	weather	attendance_rate
0	58.0	2	sunny	0.8
1	203.5	8	sunny	0.3
2	54.5	12	rainy	0.6
3	38.5	4	cloudy	0.6
4	203.0	13	rainy	0.9

(a) [1.5 Pts] Evaluate the following statements about training a ridge regression model with regularization hyper-parameter λ.

[✓ ○ **True**] ○ False As λ increases, model variance tends to 0.

○ True [✓ ○ **False**] As λ decreases, observational variance tends to 0.

○ True [✓ ○ **False**] When $\lambda = 0$, there always exists a unique model solution.

Solution: The first statement is true, as very large regularization parameters encourage the model parameters to be near-zero, thus outputting model predictions with very small values.

The second statement is false, as the observational variance of a model is the property of its dataset and is unaffected by regularization.

The third statement is false, as ridge regression with parameter $\lambda = 0$ stands equivalent to ordinary least squares and may not have a unique model solution.

(b) [1.5 Pts] Sam uses the following model to predict `attendance_rate`:

$$\widehat{\texttt{attendance\_rate}} = \hat{\theta}_0 + \hat{\theta}_1 \times (\texttt{dist\_from\_bell}) + \hat{\theta}_2 \times (\texttt{section\_no})$$

Sam one-hot encoded `weather` into three new features. Evaluate the following statements.

○ **True** ○ False If Sam adds all three one-hot encoded features to his current model, a unique OLS solution will not exist.

○ True ○ **False** If Sam adds any one of the three one-hot encoded features to his current model, the model bias squared will increase.

○ **True** ○ False With the same training set, adding more features to Sam's model will not result in higher training loss.

Solution: The first statement is true. The three one-hot encoded columns sum up to a bias column, which would cause the OLS model to have a linearly dependent design matrix and not find a unique solution.

The second statement is false. Adding features to a model would make the model complexity increase, and model bias squared would in turn decrease or stay the same.

The third statement is true. Adding features to an ordinary least squares model grants a residual vector with equal or shorter magnitude, thus a lower or equal training loss.

(c) Sam uses the following model to predict `attendance_rate`:

$$\widehat{\texttt{attendance\_rate}} = \hat{\theta}_0 + \hat{\theta}_1 \times (\texttt{dist\_from\_bell}) + \hat{\theta}_2 \times (\texttt{section\_no})$$

He trains three versions of this model: one without regularization, one with L1 regularization, and one with L2 regularization. He finds different optimal parameters from each version of training: $\hat{\vec{\theta}}^{(a)}$, $\hat{\vec{\theta}}^{(b)}$, and $\hat{\vec{\theta}}^{(c)}$.

$$\hat{\vec{\theta}}^{(a)} = \begin{bmatrix} 0 \\ 0.1 \\ 0 \end{bmatrix}, \quad \hat{\vec{\theta}}^{(b)} = \begin{bmatrix} 1 \\ 0.5 \\ 2 \end{bmatrix}, \quad \hat{\vec{\theta}}^{(c)} = \begin{bmatrix} 0.05 \\ 0.08 \\ 0.03 \end{bmatrix}$$

(i) [0.5 Pts] The optimal parameter from training without regularization is most likely:

○ $\hat{\vec{\theta}}^{(a)}$ [✓ ○ $\hat{\vec{\theta}}^{(b)}$] ○ $\hat{\vec{\theta}}^{(c)}$

(ii) [0.5 Pts] The optimal parameter from training with L1 regularization is most likely:

[✓ ○ $\hat{\vec{\theta}}^{(a)}$] ○ $\hat{\vec{\theta}}^{(b)}$ ○ $\hat{\vec{\theta}}^{(c)}$

This space is kept empty for scanning purposes

(iii) [0.5 Pts] The optimal parameter from training with L2 regularization is most likely:

○ $\hat{\vec{\theta}}^{(a)}$ ○ $\hat{\vec{\theta}}^{(b)}$ [✓ ○ $\hat{\vec{\theta}}^{(c)}$]

Solution: (a)(i) uses $\hat{\theta}^{(b)}$, (a)(ii) uses $\hat{\theta}^{(a)}$, and (a)(iii) uses $\hat{\theta}^{(c)}$.

The larger set of parameters comes from the unregularized model. Then, the set of parameters with a smaller norm and mostly 0 values should come from LASSO (L1) Regularization due to its sparsifying property. The one left with a smaller norm and small but non-zero values would be there result of L2 regularization.

(d) [1.5 Pts] Malavikha trains the following model without regularization:

$$\widehat{\texttt{attendance\_rate}} = \hat{\theta}_0 + \hat{\theta}_1 \times (\texttt{dist\_from\_bell})^2 + \hat{\theta}_2 \times \sqrt{\texttt{section\_no}}$$

After applying bootstrapping to this model, Malavikha observed a high model bias squared. Evaluate whether the following suggestions can reduce model bias squared.

○ True [✓ ○ **False**] Remove the intercept from this model.

○ True [✓ ○ **False**] Train the model with regularization instead.

○ **True** ○ False Use `dist_from_bell` as an additional predictor variable.

Solution: In this question, the model bias squared is high, which means the expected difference between the model's prediction and the true value is high. The question states that we want to reduce the model bias squared.

The first statement is false, as removing the intercept is equivalent to removing a feature, which increases model bias squared.

The second statement is false, as introducing regularization would increase the model bias squared of this model.

The third statement is true, as it does the opposite of the first statement.

5 I'm Not Throwin' Away My Shot [10 points]

Answer the following questions to help Shreya prepare for archery competitions.

(a) [1 Pt] Evaluate the following statements about the advantages of using stochastic gradient descent on a convex loss function.

○ **True** ○ False With the same initialization and number of epochs, stochastic gradient descent can find lower loss than batch gradient descent.

○ True ○ **False** Stochastic gradient descent is computationally more expensive per update than batch gradient descent.

Solution: The first statement is correct. The inherent stochasticity of mini-batch gradient descent can sometimes lead the algorithm to arrive at a smaller loss value before convergence.

The second statement is false. Stochastic gradient descent only computes the gradient of one datapoint, so it is less expensive per iteration.

(b) [3 Pts] To adjust her bowstring, Shreya uses the following model with parameters θ_1 and θ_2:

$$\hat{y} = \theta_2 \theta_1 x$$

Shreya came up with the following loss function to optimize:

$$\mathcal{L}(\hat{y}, y) = \hat{y} - \theta_2 y$$

Perform one iteration of stochastic gradient descent on this loss function with learning rate $\alpha = 0.5$ to obtain the parameters at iteration $t + 1$. The parameter values at iteration t and the datapoint to compute gradients with are:

$$\theta_1^{(t)} = 3, \theta_2^{(t)} = \pi, \quad (x, y) = (2, 1)$$

Grading will be done based on your work in the box below.

$\theta_1^{(t+1)}$ = ______________________, $\theta_2^{(t+1)}$ = ______________________

Solution: The loss function's expression can also be evaluated as follows:

$$\theta_2\theta_1 x - \theta_2 y$$

The gradient of this function can be derived as follows:

$$\frac{\partial \mathcal{L}}{\partial \theta_1} = \theta_2 x$$
$$\frac{\partial \mathcal{L}}{\partial \theta_2} = \theta_1 x - y$$

Consequently, the gradient of this function at parameter values of timestep t are:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \theta_1} &= \theta_2 x \\ &= \pi \times x = 2\pi \\ \frac{\partial \mathcal{L}}{\partial \theta_2} &= \theta_1 x - y \\ &= 3 \times x - y = 6 - 1 = 5\end{aligned}$$

Thus,

$$\theta_1^{(t+1)} = \theta_1^{(t)} - \alpha(2\pi) = 3 - \pi$$
$$\theta_2^{(t+1)} = \theta_2^{(t)} - \alpha(5) = \pi - \frac{5}{2}$$

(c) Shreya collected the following data about her points from the competition:

Probability of scoring this value	Obtained points
0.25	0
0.5	10
0.25	12

(i) [1 Pts] What is the expected number of points Shreya scores from the competition? Grading will be done based on your work in the box below.

Answer = ____________________

This space is kept empty for scanning purposes

(ii) [2 Pts] What is the standard deviation in the number of points Shreya scores from the competition? Grading will be done based on your work in the box below.

Answer = ____________________

Solution: ***Subpart (i).*** Let X be the random variable representing the number of points Shreya obtains from a competition.

The expected value of such a variable can be computed as follows:

$$\begin{aligned}\mathbb{E}[X] &= 0.25 \times 0 + 0.5 \times 10 + 0.25 \times 12 \\ &= 0 + 5 + 3 = 8\end{aligned}$$

Subpart (ii). Let X be the random variable representing the number of points Shreya obtains from a competition.

The variance of such a variable can be computed by first computing the expectation of X^2:

$$\begin{aligned}\mathbb{E}[X^2] &= 0.25 \times 0 + 0.5 \times 100 + 0.25 \times 144 \\ &= 0 + 50 + 36 = 86\end{aligned}$$

Next,

$$\begin{aligned}Var(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= 86 - 8^2 = 22\end{aligned}$$

The standard deviation, which is the square root of variance, is therefore $\sqrt{22}$.

(d) Shreya wants to model her rate of perfect shots using the estimator:

$$\hat{p} = \frac{X}{n}$$

X **is a random variable** that represents the number of times Shreya makes a perfect shot out of n trials. Each trial is independent.

Let the actual probability of Shreya making a perfect shot be an unknown constant p. For each subpart below, grading will be done based on the work you show in the boxes below.

(i) [2 Pts] What is the variance of $\hat{p}$? **Express and simplify your answer in terms of** n, $\hat{p}$, p, and any real number.

Answer = ______________________

(ii) [1 Pts] What is the bias of this estimator: $\mathbb{E}[\hat{p}] - p$? **Express and simplify your answer in terms of** n, $\hat{p}$, p, and any real number.

Answer = ______________________

Solution: ***Subpart (i).*** The variance of this estimator can be computed as follows:

$$\begin{aligned} Var(\hat{p}) = Var(\frac{X}{n}) &= \frac{1}{n^2} Var(X) \\ &= \frac{1}{n^2} \times np(1-p) = \frac{p(1-p)}{n} \end{aligned}$$

Subpart (ii). The expected value of a binomial variable with n trials and p probability of success is np. Therefore,

$$\begin{aligned}\mathbb{E}[\hat{p}] - p &= \mathbb{E}\left[\frac{X}{n}\right] - p \\ &= \frac{1}{n} \times np - p = 0\end{aligned}$$

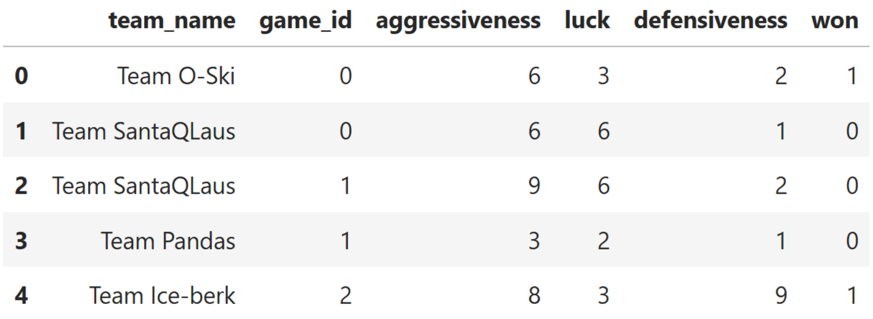
The estimator $\hat{p}$ is unbiased.

6 But Would O-Ski Lose? [17 points]

The dataset `hockey_record` describes the performance data of ice hockey teams across many games. The columns of `hockey_record` are as described below:

- `team_name`: The team name. (type = `str`)
- `game_id`: The ID of the game. (type = `np.int64`)
- `aggressiveness`: The team's aggressiveness in this game. All values are between 0 and 10 inclusive, with 0 being the least aggressive. (type = `np.float32`)
- `luck`: The luckiness of the team in this game. All values are between 0 and 10 inclusive, with 0 being the least lucky. (type = `np.float32`)
- `defensiveness`: The defensiveness of the team in this game. All values are between 0 and 10 inclusive, with 0 being the least defensive. (type = `np.float32`)
- `won`: Indicates if the team won the game: `0` for a loss and `1` for a win. (type = `int`)

A sample of `hockey_record` is shown below:



	team_name	game_id	aggressiveness	luck	defensiveness	won
0	Team O-Ski	0	6	3	2	1
1	Team SantaQLaus	0	6	6	1	0
2	Team SantaQLaus	1	9	6	2	0
3	Team Pandas	1	3	2	1	0
4	Team Ice-berk	2	8	3	9	1

(a) [1 Pt] Evaluate the following statements about training a logistic regression model on a linearly separable dataset.

[✓ ○ **True**] ○ False It is possible to find a classifier that achieves 100% accuracy.

○ True [✓ ○ **False**] For all linearly separable datasets, there exists a finite number of optimal solutions for model parameters.

Solution:

The first statement is true. A linearly separable dataset allows us to find a threshold $z = \vec{x}^T\theta$ such that $z = \sigma(T)$ achieves 100% accuracy.

The second statement is incorrect. There can exist infinite optimal solutions for model parameter values in linearly separable datasets. For example, consider the following toy example of datapoints (x, y):

$$\mathcal{D} = \{(2, 1), (1, 0)\}$$

Any line that lies between $(2, 1)$ and $(1, 0)$ can serve as a decision boundary. Therefore, there can be infinitely many optimal solutions.

(b) [1 Pt] Evaluate the following statements about components of a logistic regression model.

○ **True** ○ False For any $t \geq 0$, the sigmoid function σ satisfies $\sigma(t) \in [0.5, 1)$.

○ **True** ○ False In binary classification, a logistic regression model can predict the probability of a datapoint belonging to class 0.

Solution: The first statement is true. A sigmoid function's expression is:

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

Therefore, for all $z \geq 0$, $\sigma(z) \geq 0.5$.

The second statement is true. This is because the logistic regression model conventionally predicts the probability of a datapoint belonging to class 1, which is the complement of a datapoint belonging to class 0.

(c) [1 Pt] Jessica designed the following loss function to optimize her logistic regression model.

$$\mathcal{L}(\theta) = -\left[y \log(p) + (1 - y) \log\left((1 - p)^3\right)\right]$$

Where p is the probability that the corresponding datapoint belongs to class 1.

Evaluate the following statements about $\mathcal{L}$ and cross-entropy loss.

○ **True** ○ False The loss function $\mathcal{L}$ incurs more cost on making false positive predictions than cross-entropy loss.

○ True ○ **False** Whenever a logistic regression model classifies a datapoint into class 0, the loss component $y \log(p)$ is equal to 0.

Solution: The first statement is true. The term $(1 - y) \log\left((1 - p)^3\right)$ can be rewritten as $3 \times (1 - y) \log(1 - p)$, which is 3 times higher than its corresponding term in cross-entropy loss.

This term is active when the true y value is 0 (negative), yet the probability at which this point is positive is nonzero, such that the classifier thinks this data point could be positive. Therefore, this term punishes false positive terms.

The second statement is false. Even when a logistic regression model classifies a point to class 0, the true class of that point could be class 1, which makes $y \log(p)$ non-zero for any $p < 1$.

(d) Fill in the blanks to create a `DataFrame` that contains all rows from `hockey_record` where `aggressiveness` is greater than 7 and the team's `luck` average across all games is greater than 5.

```
(
    hockey_record[_______________(i)_______________]
        .groupby("team_name")
        ._____________(ii)_____________(_____________(iii)_____________ 5)
)
```

(i) [1 Pts] Fill in blank `(i)`:

Solution: `hockey_record["aggressiveness"] > 7`

(ii) [0.5 Pts] Fill in blank `(ii)`:

Solution: `filter`

(iii) [1 Pts] Fill in blank `(iii)`:

Solution: `lambda sf: sf["luck"].mean() >`

(e) [2 Pts] What will happen to a logistic regression classifier's metrics if the classifier's threshold is lowered? Evaluate the following statements.

○ True ○ **False** The precision will never increase.

○ **True** ○ False The recall will never decrease.

○ True ○ **False** The confidence our model needs to predict $\hat{y} = 1$ will increase.

○ True ○ **False** The classifier becomes closer to a perfect predictor.

Solution: Remember that:

$$\text{Precision} = \frac{TP}{TP + FP}, \text{ Recall} = \frac{TP}{TP + FN}$$

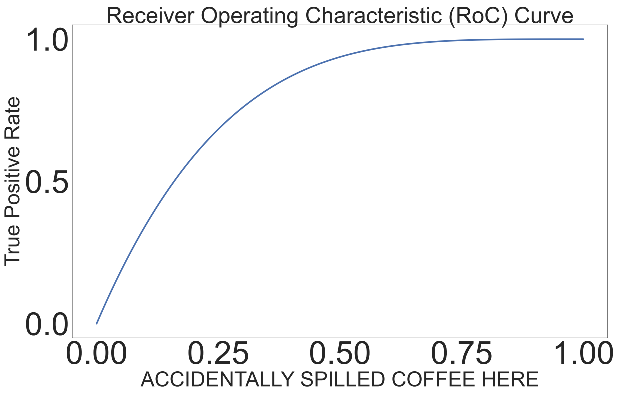
The first statement is false. Lowering the classifier threshold can lead to increases in both True Positives and False Positives. As both the numerator and the denominator are shifting, the precision can either increase or decrease as we tune the classifier threshold.

The second statement is true. In recall metric, the denominator never changes as it is the number of data points belonging to class 1. By lowering the classifier threshold, the number of true positives will either stay the same or increase.

The third statement is false. The confidence our model needs for predicting $\hat{y} = 1$ is decreased.

The fourth statement is false. A perfect predictor is defined as a model with an area under the ROC curve of 1. Changing the threshold of a classifier doesn't change its area under the RoC curve.

(f) [1.5 Pts] The RoC curve of Jessica's logistic regression model is shown below:



The RoC curve is plotted correctly. However, the x-axis label of the plot is incorrect.

Evaluate the following statements regarding the above RoC curve and the model it describes.

○ **True** ○ False The x-axis measures the proportion of datapoints belonging to Class 0 that were classified into Class 1.

[✓ ○ **True**] ○ False A classifier that randomly predicts $p(y = 1)$ to be uniformly between 0 and 1 for all data points would have an area under the ROC curve that's lower than the area under the above ROC curve.

[✓ ○ **True**] ○ False Only a perfect predictor's ROC curve can contain the point $(0, 1)$ on the ROC curve.

Solution: The first statement is true. The x-axis label is supposed to be "False Positive Rate", which can be computed as 1 - specificity.

The second statement is true. The entirety of our RoC curve is above the RoC curve of a random predictor, which is a diagonal line running from $(0, 0)$ to $(1, 1)$. Consequently, the area under the RoC curve for a random predictor must be lower.

The third statement is true. Note that the ROC curve is non-decreasing. Therefore, once it contains the point $(0, 1)$, it will only contain points whose y-coordinate is 1, making the corresponding classifier a perfect predictor.

(g) Jessica's logistic regression model makes predictions on a test set and outputs the following results. What are the values of the following metrics on these datapoints?

True Label	1	1	1	0	0	1	0	1	1
Prediction	1	0	1	0	0	0	1	1	1

(i) [1 Pts] The model's recall:

Recall: ________________________

(ii) [1 Pts] The model's False Positive Rate:

FPR: ________________________

Solution: There are 4 true positives, 1 false positive, 2 true negatives, and 2 false negatives resulting from the classifier.

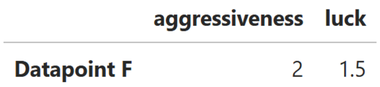
The recall is equivalent to $\frac{TP}{TP+FN} = \frac{4}{4+2} = \frac{2}{3}$.

The false positive rate is equivalent to $\frac{FP}{FP+TN} = \frac{1}{1+2} = \frac{1}{3}$.

For subparts 6(h) and 6(j), we use a logistic regression model to predict if a team wins or not. This model is called **Model D**, and uses the following model to predict the probability that a team wins:

$$P(y=1|\vec{x}) = \sigma(\hat{\theta}_0 + \hat{\theta}_1 \times \texttt{aggressiveness} + \hat{\theta}_2 \times \texttt{luck}); \;\; \hat{\theta}_0 = 1, \hat{\theta}_1 = \frac{1}{2}, \hat{\theta}_2 = \frac{1}{3}$$

(h) [2 Pts] Using **Model D**, what's the predicted probability that the team in **Datapoint F doesn't** win? Grading will be based on your work below. Your answer **must not** include σ.



	aggressiveness	luck
Datapoint F	2	1.5

Answer = ______________________

Solution: The probability is:

$$1 - \sigma(\vec{x}^T\theta) = 1 - \sigma\left(1 \cdot 1 + \frac{1}{2} \cdot 2 + \frac{1}{3} \cdot 1.5\right)$$
$$= 1 - \sigma(2.5) = \frac{e^{-2.5}}{1 + e^{-2.5}}$$

(j) The following classifier has a 100% validation accuracy:

$$\text{classify } \hat{y} = 1 \text{ when } \frac{1}{4} \times (\texttt{aggressiveness}) + \frac{1}{6} \times (\texttt{luck}) \geq 1$$

(i) [2 Pts] Setting the threshold of **Model D** to a value T_D makes **Model D** have the same decision boundary as the above classifier. What is T_D? You may use σ in your answers.

T_D = ______________

Solution: For any decision boundary in the form:

$$\frac{1}{2} \times (\texttt{aggressiveness}) + \frac{1}{3} \times (\texttt{luck}) + 1 = k$$

Using a threshold in our logistic regression classifier $T = \sigma(k)$ would result in the same classifier.

The classifier above has a decision boundary corresponds to the line:

$$\frac{1}{2} \times (\texttt{aggressiveness}) + \frac{1}{3} \times (\texttt{luck}) = 2$$

or alternatively,

$$\frac{1}{2} \times (\texttt{aggressiveness}) + \frac{1}{3} \times (\texttt{luck}) + 1 = 3$$

Therefore, using a threshold of $T_D = \sigma(3)$ results in an equivalent classifier.

(ii) [2 Pts] If the validation set contains datapoints from Class 1, what is the largest **known** range of logistic regression thresholds T for **Model D** to achieve 100% validation accuracy? Express your answer as an interval. You may use σ in your answers.

Note: An interval can appear in one of these formats: (x, y), $(x, y]$, $[x, y)$, $[x, y]$

Answer = ________________, ________________

Solution: Since the validation set only involves datapoints from Class 1, the true positive rate (TPR) is equivalent to the accuracy. TPR is synonymous with recall.

As quoted from 6(d):

> In the recall metric, the denominator never changes as it's the amount of data points belonging to class 1. By lowering the classifier threshold, the number of true positives will either stay the same or increase. Therefore, the recall will increase.

Therefore, for any threshold T' lower than the maximum known threshold T_{max} at which the validation accuracy is 100%, using T' grants 100% validation accuracy as well. Consequently, the answer is $[0, \sigma(3)]$.

7 PCA in Cooking [7.5 points]

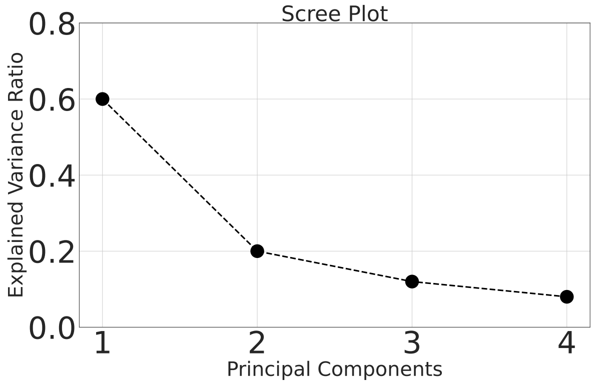
Minoli is using PCA to find two-dimensional representations of cooking show contestants. She collected the following dataset, `cooks`, where each row describes a contestant. `cooks` has the following columns:

- `license_grade`: The contestant's culinary license grade, between 1 and 3 inclusive. (type=`np.int64`)
- `num_stores`: Number of restaurants the contestant owns. (type=`np.int64`)
- `avg_ratings`: Contestant's online rating, between -5 and 5 inclusive. (type=`np.float32`)
- `num_points`: Total number of points the contestant won in the show, between 0 and 5 inclusive. (type=`np.float32`)

Below is a sample of `cooks`:

	license_grade	num_stores	avg_ratings	num_points
0	1	5	4.2	0.0
1	3	0	2.1	2.0
2	2	2	-1.2	1.2
3	2	1	-3.6	3.0
4	2	0	0.3	1.2

(a) [1.5 Pts] Evaluate the statements below based on the following scree plot that Minoli obtained from applying PCA on `cooks`:



○ **True** ○ False We should use 2 principal components for PCA on this dataset.

○ **True** ○ False The rank of this dataset is at least 4.

○ True ○ **False** There can be 5 principal components for `cooks`.

Solution:

The first statement is true. This is because $x = 2$ is at the elbow of the scree plot.

The second statement is true. A dataset's rank is at least as much as the number of nonzero singular values. As long as the explained variance ratio is nonzero, the principal component's corresponding singular value would also be nonzero.

The third statement is false. There can be at most 4 principal components for `cooks` because it has 4 columns. The number of principal components a dataset can have is upper-bounded by its number of columns.

(b) [1.5 Pts] Given matrix X has a singular value decomposition of $X = USV^T$, evaluate the following statements regarding the properties of matrices U, S, and V.

○ True [✓ ○ **False**] The columns of matrix U are always the eigenvectors of $X^T X$.

[✓ ○ **True**] ○ False S is a diagonal matrix.

○ True [✓ ○ **False**] The first row of V is the direction of maximum variance.

Solution: The first statement is false. The columns of U are eigenvectors of XX^T. Here is a quick view of this fact (which was mentioned in lecture slides):

$$\begin{aligned} XX^T U &= USV^T V S^T U^T U \\ &= USS^T \\ &= \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_d \end{bmatrix} \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_d \end{bmatrix} = \begin{bmatrix} s_1 \vec{u}_1 & \cdots & s_d \vec{u}_d \end{bmatrix} \end{aligned}$$

Therefore,

$$\forall i \in \{1, \dots, d\} : XX^T \vec{u}_i = s_i \vec{u}_i$$

The second statement is true. By definition of PCA, S is a diagonal matrix, where the values on the diagonal are singular values.

The third statement is false. The first column of V would be the vector that captures the maximum variance in the dataset, otherwise known as the first principal component.

Your answers in subparts 7(c), 7(d) should be based on the information below.

Angela collected a dataset of 3 data points. Upon applying PCA to this dataset, Angela obtains matrices U, S, and V. For your convenience, matrices U and V have been rounded:

$$U \approx \begin{bmatrix} 0.8 & 0.6 & 0 \\ -0.4 & 0.6 & -0.7 \\ -0.4 & 0.6 & 0.7 \end{bmatrix}, S = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, V \approx \begin{bmatrix} 0.8 & -0.3 & -0.3 \\ 0.4 & 0.6 & 0.6 \\ 0 & 0.7 & -0.7 \end{bmatrix}$$

This space is kept empty for scanning purposes

(c) Answer the following questions about using principal components based on the rounded results of PCA above. Grading will be done based on your work in the box below.

(i) [0.5 Pts] What is the component score of the first principal component?

Answer = ______________________

Solution: The component score of the first principal component is $\frac{s_1^2}{n} = \frac{16}{3}$

(ii) [0.5 Pts] What is the variance captured by the second principal component?

Answer = ______________________

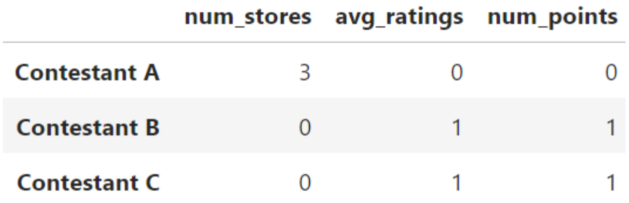
Solution: The total variance captured by the second principal component is equivalent to the component score of the second principal component is $\frac{s_2^2}{n} = \frac{1^2}{3} = \frac{1}{3}$

(iii) [0.5 Pts] What is the reconstruction loss from using 2 principal components?

Answer = ______________________

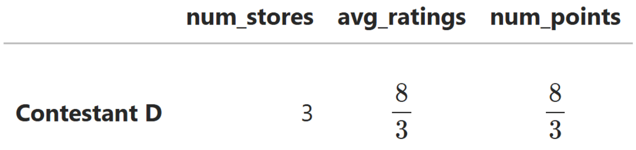
Solution: Based on the results above, the rank of our dataset is 2. Therefore, using 2 principal components would reconstruct the entire dataset perfectly. The reconstruction loss is 0.

(d) [3 Pts] The dataset Angela used to compute the above PCA is as follows:



	num_stores	avg_ratings	num_points
Contestant A	3	0	0
Contestant B	0	1	1
Contestant C	0	1	1

Angela receives a new contestant's performance as shown below:



	num_stores	avg_ratings	num_points
Contestant D	3	$\frac{8}{3}$	$\frac{8}{3}$

What is this contestant's coordinate (x, y) on a PCA plot based on Angela's PCA results from the previous page? Grading will be done based on the work you show in the box below.

This space is kept empty for scanning purposes.

$(x, y) =$ ______________________________

Solution: First, let us decentered our data point using the feature-wise mean we find from the dataset above:

$$\begin{bmatrix}3 & \frac{8}{3} & \frac{8}{3}\end{bmatrix} - \begin{bmatrix}1 & \frac{2}{3} & \frac{2}{3}\end{bmatrix} = \begin{bmatrix}2 & 2 & 2\end{bmatrix}$$

We are asked for two coordinates on the PCA plot, so we need to obtain the projection score of our data point onto the first two principal components. To get the decentered datapoint's projection score onto the first two principal components, simply multiply the datapoint with the first two columns of V:

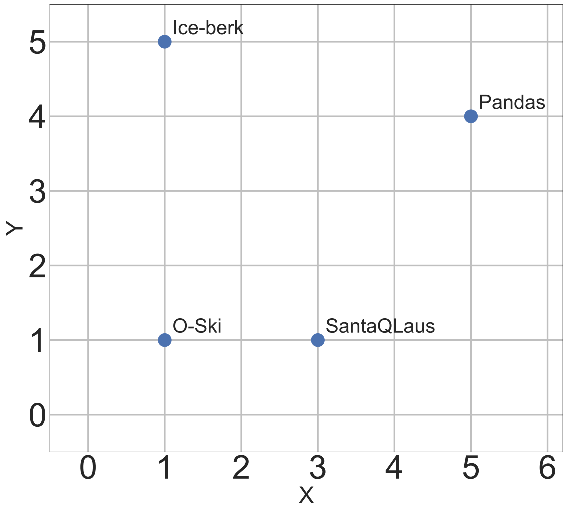
$$\begin{bmatrix}2 & 2 & 2\end{bmatrix}\begin{bmatrix}0.8 & -0.3 \\ 0.4 & 0.6 \\ 0 & 0.7\end{bmatrix} = \begin{bmatrix}2.4 & 2\end{bmatrix}$$

Therefore, the answer is $(x, y) = (2.4, 2)$

8 Soccer Player Cluster [8.5 points]

Using unsupervised learning algorithms, Abby obtained two-dimensional representations for four soccer teams:

Team Name	Two-dimensional Coordinates
Team O-Ski	$(1, 1)$
Team SantaQlaus	$(3, 1)$
Team Pandas	$(5, 4)$
Team Ice-berk	$(1, 5)$



Your work in the above coordinate plane will not be graded.

Abby wants to cluster the above two-dimensional representations.

(a) Abby wants to use k-Means clustering to group teams into two clusters, A and B. Initially, cluster A has a center at $(0, 5)$, and cluster B has a center at $(6, 0)$.

(i) [1 Pts] Which cluster is Team SantaQLaus assigned to in the first iteration of k-Means clustering?

○ Cluster A

○ **Cluster B**

(ii) [1 Pts] Which cluster is Team Pandas assigned to in the second iteration of k-Means clustering?

○ Cluster A

○ **Cluster B**

(iii) In the following blanks, write the centers of each cluster when the algorithm converges.

(i) [1 Pts] Center of cluster A at convergence:

(__________ , __________)

(ii) [1 Pts] Center of cluster B at convergence:

(__________ , __________)

Solution: The overall updates of running k-Means clustering are as follows.

Iteration 1. The distances between each datapoint and centers are as established in the following table:

Team Name	Distance to A	Distance to B	Cluster Assigned
O-Ski	$\sqrt{17}$	$\sqrt{26}$	A
SantaQlaus	$\sqrt{25}$	$\sqrt{10}$	B
Pandas	$\sqrt{26}$	$\sqrt{17}$	B
Ice-berk	1	$\sqrt{50}$	A

Then, cluster centers are updated as follows:

Cluster Name	Original Center	New Center
A	$(0, 5)$	$(1, 3)$
B	$(6, 0)$	$(4, 2.5)$

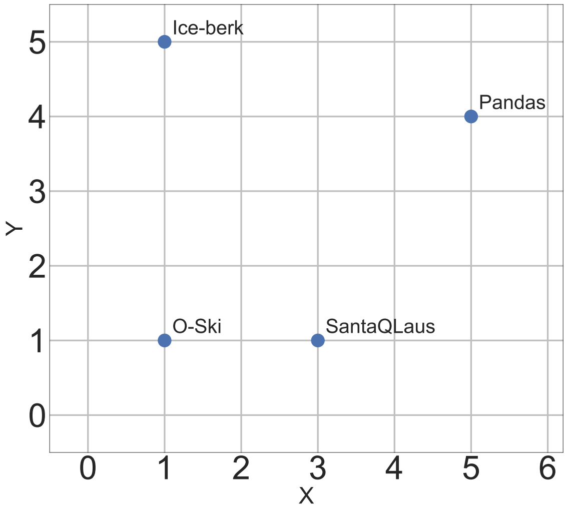
Iteration 2. The distances between centers and each datapoint are as established in the following table:

Team Name	Distance to A	Distance to B	Cluster Assigned
O-Ski	2	larger than 2	A
SantaQlaus	larger than 2	$\sqrt{3.25}$, smaller than 2	B
Pandas	larger than 2	$\sqrt{3.25}$, smaller than 2	B
Ice-berk	2	larger than 2	A

With the new cluster centers, no clustering assignments are updated. The clustering algorithm converges.

For your convenience, the data points of each team are repeated below:

Team Name	Two-dimensional Coordinates
Team O-Ski	$(1, 1)$
Team SantaQlaus	$(3, 1)$
Team Pandas	$(5, 4)$
Team Ice-berk	$(1, 5)$



Your work in the above coordinate plane will not be graded.

This space is kept empty for scanning purposes

(b) Next, Abby uses hierarchical agglomerative clustering to cluster teams together.

(i) [1.5 Pts] If Abby uses **single** linkage, up to the iteration where there are two clusters, which teams are in the same cluster as Team Pandas?

○ True ○ **False** Team Ice-berk
○ **True** ○ False Team O-Ski
○ **True** ○ False Team SantaQlaus

(ii) [1.5 Pts] If Abby uses **complete** linkage, up to the iteration where there are two clusters, which teams are in the same cluster as Team O-Ski?

○ True ○ **False** Team Ice-berk
○ True ○ **False** Team Pandas
○ **True** ○ False Team SantaQlaus

Solution: In the first iteration, regardless of whether single or complete linkage was used, the points for Team O-Ski and Team SantaQlaus will be merged into one cluster. Let us call this cluster "Cluster A", then the table of distances between the current three clusters is established as the following table:

Cluster 1	Cluster 2	Distance (Single Linkage)	Distance (Complete Linkage)
Cluster A	Pandas	$\sqrt{13}$	$\sqrt{25}$
Cluster A	Ice-berk	$\sqrt{16}$	$\sqrt{20}$
Pandas	Ice-berk	$\sqrt{17}$	$\sqrt{17}$

From the above table, we see that under a single linkage, Cluster A (O-Ski and SantaQlaus) will be merged with Pandas.

Under a complete linkage, Pandas and Ice-berk would be merged instead.

(c) [1.5 Pts] Evaluate the following statements regarding assessments of clustering outcomes.

○ **True** ○ False The elbow method can be used to choose a value for k in k-Means clustering.

○ True ○ **False** A data point with a high silhouette score is far from other points in its cluster.

○ **True** ○ False Dendrograms present the process of hierarchical agglomerative clustering as trees.

Solution:

The first statement is true. This is as presented in the lecture slides and used along with other metrics like inertia and silhouette score.

The second statement is false. The silhouette score is high in data points close to its cluster-mates.

The third statement is true. The dendrogram's tree-like structures can resemble the merging process of hierarchical agglomerative clustering.

You are done with the final! Congratulations!

Draw next semester's DATA 100/200 Logo OR your favorite DATA 100/200 memories!