

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`)

	<code>passport_id</code>	<code>name</code>	<code>ticket_type</code>
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`

	<code>passport_id</code>	<code>name</code>	<code>collar_size</code>	<code>num_jacket_bought</code>
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`

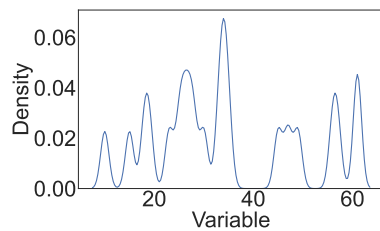
(a) [0.5 Pts] What type of variable is `ticket_type` in `ticket`?

- Qualitative Nominal
 Qualitative Ordinal
 Quantitative

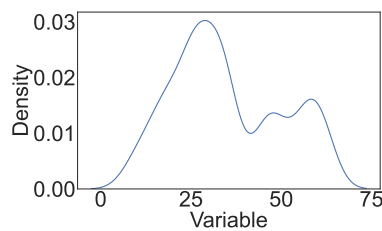
(b) Xiaorui wants to know about attendee's experiences buying jackets. He sent a survey to all attendees in ticket.

- (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
- (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.

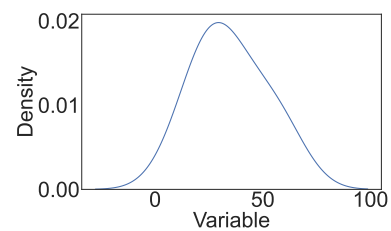
(c) Answer the following questions about Kernel Density Estimation (KDE).



This KDE plot uses bandwidth $\alpha =$ _____ (x) _____



This KDE plot uses bandwidth $\alpha =$ _____ (y) _____



This KDE plot uses bandwidth $\alpha =$ _____ (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 ?
- $f_\alpha(x) = \frac{1}{n} K_\alpha(x, x_1)$
 - $f_\alpha(x) = \sum_{i=1}^n K_\alpha(x, x_i)$
 - $f_\alpha(x) = \frac{1}{n} \sum_{i=1}^n K_\alpha(x, x_i)$

(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`?

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

Answer: _____

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:

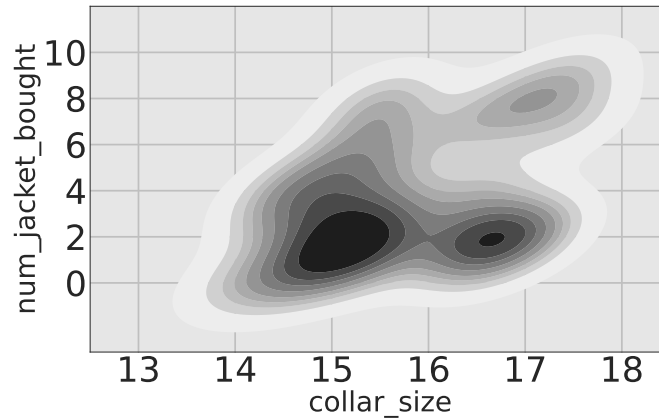
```
df_to_plot = jacket_miss._____ (i) _____
_____ (ii) _____ (
    data = df_to_plot_____ (iii) _____,
    x = "collar_size", y="num_jacket_bought"
)
```

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

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

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

- (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.

- (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):

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

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)

	<code>order_id</code>	<code>customer_age</code>	<code>protein_name</code>	<code>protein_lb</code>
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

	<code>protein_name</code>	<code>price_per_lb</code>
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: _____

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):

This space is kept empty for scanning purposes

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

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

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

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

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:

	<code>order_id</code>	<code>customer_type</code>	<code>protein_name</code>	<code>protein_lb</code>	<code>price_per_lb</code>
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):

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

- (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) _____;
```

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

This space is kept empty for scanning purposes

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

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

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

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

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:

	<code>critic_1</code>	<code>critic_2</code>	<code>comment</code>	<code>self_rating</code>
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?

$\hat{\theta} =$ _____

(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 = _____

(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):

This space is kept empty for scanning purposes

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

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

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

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

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

$$\widehat{\text{self_rating}} = \hat{\theta}_0 + \hat{\theta}_1 \times \text{critic_1}$$

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.

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

$$\widehat{\text{self_rating}} = \hat{\theta}_0 + \sum_{i=1}^d \hat{\theta}_i \times (\text{critic_1})^i$$

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: ($\lambda =$ _____, $d =$ _____)

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

Answer: ($\lambda =$ _____, $d =$ _____)

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

$$\widehat{\text{self_rating}} = \hat{\theta}_1 \times (\text{critic_1}) + \hat{\theta}_2 \times (\text{critic_1})^2$$

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 $\lambda(\hat{\theta}_1 + \hat{\theta}_2)$.

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

$$\widehat{\text{self_rating}} = \hat{\theta}_0 + \hat{\theta}_1 \times (\text{critic_1})^2$$

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 = _____

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

$$\widehat{\text{self_rating}} = \hat{\theta}_0 + \hat{\theta}_1 \times \text{critic_1}$$

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.

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

$$\widehat{\text{self_rating}} = \hat{\theta}_0 + \hat{\theta}_1 \times (\text{critic_1}) + \hat{\theta}_2 \times (\text{critic_2})$$

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



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

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:

	<code>dist_from_bell</code>	<code>section_no</code>	<code>weather</code>	<code>attendance_rate</code>
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.

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

$$\widehat{\text{attendance_rate}} = \hat{\theta}_0 + \hat{\theta}_1 \times (\text{dist_from_bell}) + \hat{\theta}_2 \times (\text{section_no})$$

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.

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

$$\widehat{\text{attendance_rate}} = \hat{\theta}_0 + \hat{\theta}_1 \times (\text{dist_from_bell}) + \hat{\theta}_2 \times (\text{section_no})$$

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: $\vec{\hat{\theta}}^{(a)}$, $\vec{\hat{\theta}}^{(b)}$, and $\vec{\hat{\theta}}^{(c)}$.

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

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

- $\vec{\hat{\theta}}^{(a)}$ $\vec{\hat{\theta}}^{(b)}$ $\vec{\hat{\theta}}^{(c)}$

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

- $\vec{\hat{\theta}}^{(a)}$ $\vec{\hat{\theta}}^{(b)}$ $\vec{\hat{\theta}}^{(c)}$

This space is kept empty for scanning purposes

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

- $\vec{\hat{\theta}}^{(a)}$ $\vec{\hat{\theta}}^{(b)}$ $\vec{\hat{\theta}}^{(c)}$

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

$$\widehat{\text{attendance_rate}} = \hat{\theta}_0 + \hat{\theta}_1 \times (\text{dist_from_bell})^2 + \hat{\theta}_2 \times \sqrt{\text{section_no}}$$

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.

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.

- (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:

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

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

$$\theta_1^{(t+1)} = \underline{\hspace{2cm}}, \theta_2^{(t+1)} = \underline{\hspace{2cm}}$$

(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 = _____

(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 = _____

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:

	<code>team_name</code>	<code>game_id</code>	<code>aggressiveness</code>	<code>luck</code>	<code>defensiveness</code>	<code>won</code>
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.

(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.

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

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

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.

(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):

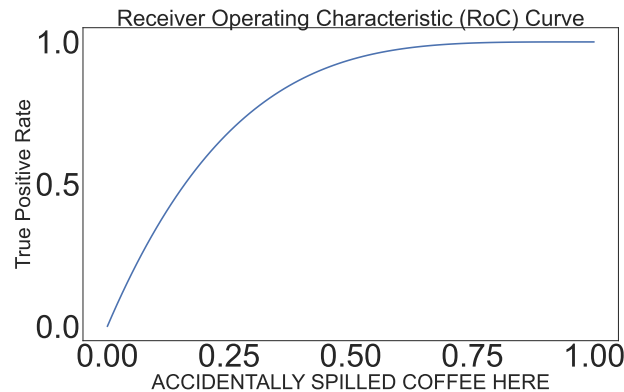
(ii) [0.5 Pts] Fill in blank (ii):

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

(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.

(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.

(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: _____

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:

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

- (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 = _____

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

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

- (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.

$T_D =$ _____

- (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 = _____, _____

7 PCA in Cooking [7.5 points]

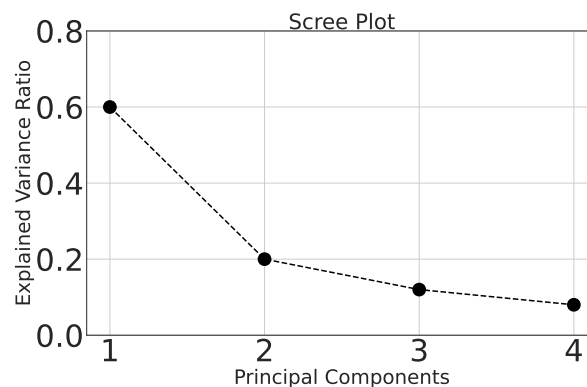
Minoli is using PCA to find two-dimensional representations of cooking show contestants. She collected the following dataset, `cooks`, where each row describes a contestant. `cooks` has the following columns:

- `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`:

	<code>license_grade</code>	<code>num_stores</code>	<code>avg_ratings</code>	<code>num_points</code>
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`.

(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.

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 = _____

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

Answer = _____

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

Answer = _____

(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.

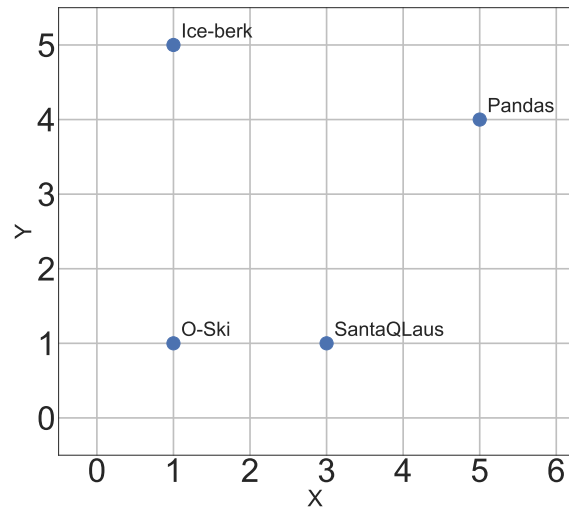
This space is kept empty for scanning purposes.

$(x, y) =$ _____

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:

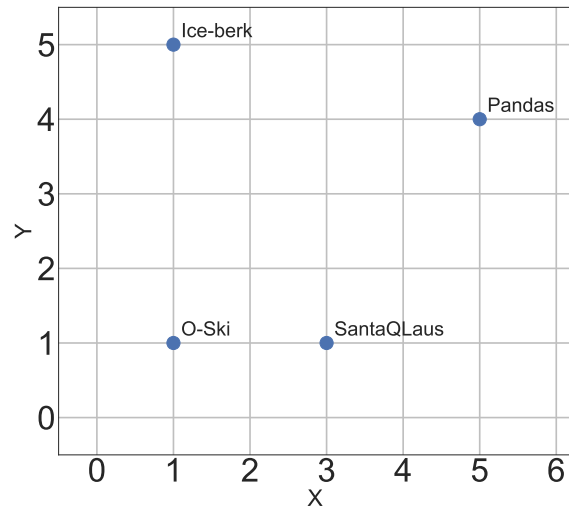
(_____, _____)

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

(_____, _____)

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?

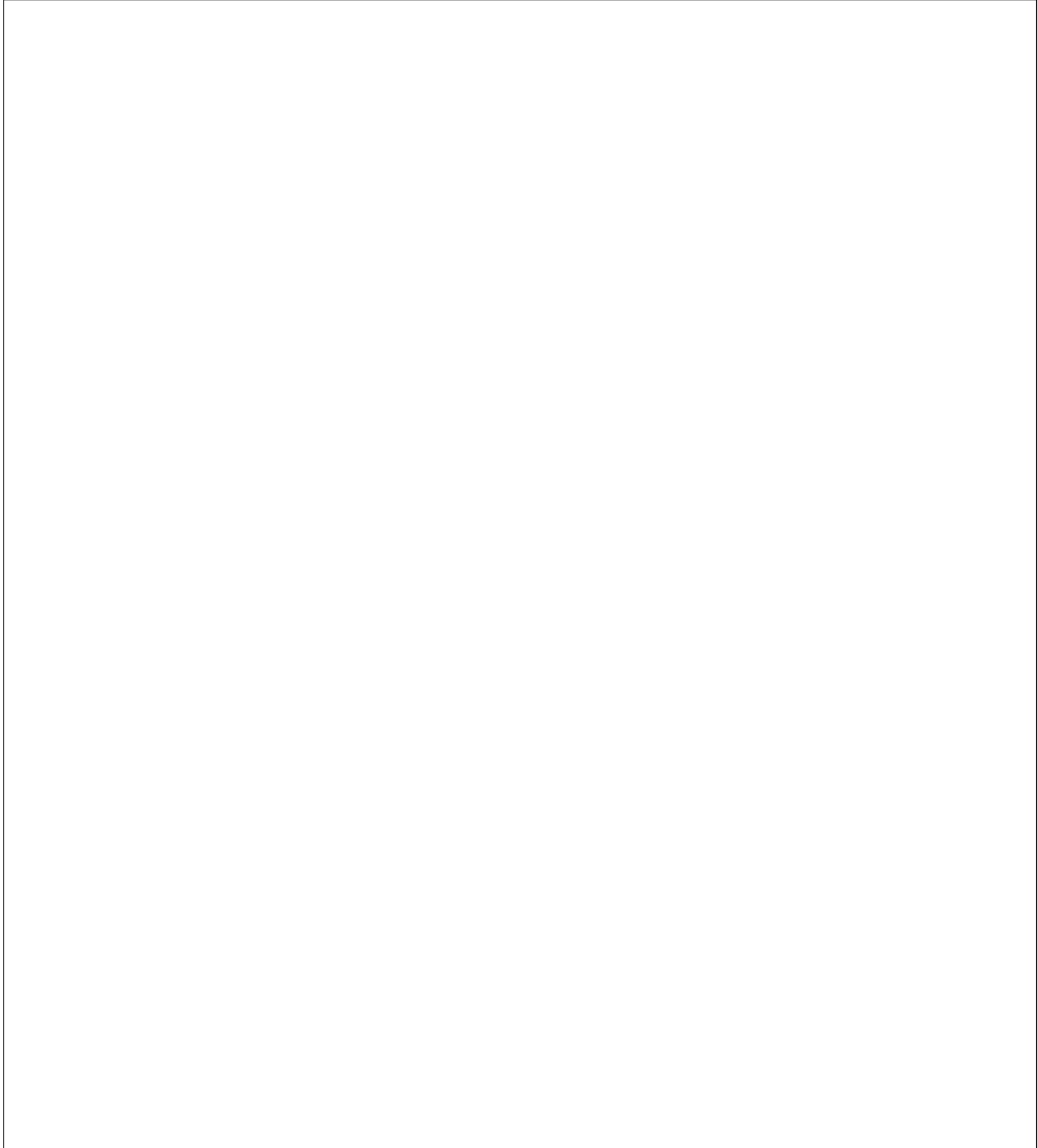
- True False Team Ice-berk
 True False Team Pandas
 True False Team SantaQlaus

(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.

You are done with the final! Congratulations!

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

A large, empty rectangular box with a thin black border, intended for the student to draw their next semester's DATA 100/200 logo or their favorite memories from the course.