DATA 8 Foundations of Data Science

Spring 2022

Final Exam

INSTRUCTIONS

You have 2 hours and 50 minutes to complete the exam.

- The exam is closed book, closed notes, closed computer/calculator, except the provided final reference sheet.
- Mark your answers on the exam itself in the spaces provided. We will not grade answers written on scratch paper or outside the designated answer spaces.
- If you need to use the restroom, bring your phone and exam to the front of the room.

For questions with **circular bubbles**, you should fill in exactly *one* choice.

○ You must choose either this option

○ Or this one, but not both!

For questions with **square checkboxes**, you may fill in *multiple* choices.

□ You could select this choice.

□ You could select this one too!

****Important****: Please **fill in** circles and squares to indicate answers and cross out or erase mistakes.

Preliminaries

You can complete these questions before the exam starts.

(a) What is your full name?

(b) What is your student ID number?

(c) Who is sitting to your left? (Write *no one* if no one is next to you.)

(d) Who is sitting to your right? (Write *no one* if no one is next to you.)

1. **(28.0 points) True or False**

(a) **(2.0 pt)** The slope and intercept that result from minimizing the mean squared error are always the same as the slope and intercept that result from minimizing the root mean squared error.

○ True

○ False

(b) **(2.0 pt)** In order to build a k nearest neighbors classifier, you do not need to know the class value of any of the training examples.

○ True

○ False

(c) **(2.0 pt)** A classifier is considered to be overfitting if it performs very well on the training set, but not very well on the test set.

○ True

○ False

(d) **(2.0 pt)** If I am using Home Price and Household Income (both in US dollars) as two features for my classifier, I do not need to standardize them since they have the same units.

○ True

○ False

(e) **(2.0 pt)** Entering your name and birthday when signing up for the Chipotle rewards program is an example of inference (in terms of data privacy).

○ True

○ False

(f) **(2.0 pt)** Suppose you have a set of points (x, y) with the mean of x being 10 and the mean of y being 8. If you perform linear regression on this dataset, the confidence interval for the height of the regression line at $x = 9$ will be narrower than the corresponding confidence interval at $x = 18$.

○ True

○ False

(g) **(2.0 pt)** In linear regression, slope is always measured in the same units as intercept.

○ True

○ False

(h) **(2.0 pt)** According to the Central Limit Theorem, if a sample is large, and drawn at random from the population with replacement, then the probability distribution of the sample average is roughly normal.

○ True

○ False

(i) **(2.0 pt)** For any distribution, the percent of data that lies within 2 SDs of the average is at least 75% and at most 95%.

○ True

○ False

(j) **(2.0 pt)** The median of a set of 25 integers will always be an integer.

○ True

○ False

(k) **(2.0 pt)** Suppose a hypothesis test is proposed and we already know that the null hypothesis is true. If 100 researchers each independently collect a large random sample of the same size to carry out an experiment and they all use 5% as their p-value cutoff, we should expect around 5% of them to reject the null.

○ True

○ False

(l) **(2.0 pt)** The chance of an event happening is always the number of outcomes that make the event happen, divided by the total number of possible outcomes.

○ True

○ False

(m) **(2.0 pt)** If we use linear regression to predict y-values based on our x-values, the average of our residuals will always be zero.

○ True

○ False

(n) **(2.0 pt)** Given a function `error(a, b, c, d)` which computes some error based on its input arguments, a valid output from the call `minimize(error)` could be an array containing two elements: 25 and 4.

○ True

○ False

2. (22.0 points) Shorts

UBA, a media company based in New York, wants to start integrating Youtube videos into its morning show.

To evaluate some propsects, Cory, the CEO, puts together a table called `videos` that contains a random sample of Youtube videos published in the last year. The first few rows are shown here:

Name	ID	Views	#Shorts	Date
INSANE strawberry trick! #Shorts	UC6D1L2vxEAg	329461822	True	05-08
Adele - Easy On Me (Official Video)	UComP_epzeKz	272980726	False	10-14
BTS () ‘Permission to Dance’	UC3IZKseVp	480881920	False	07-09
BTS () ‘Butter’	UC3IZKseVp	746499500	False	05-20

The table has the following columns:

- *Name*: (string) the **video**’s name
- *ID*: (string) the **channel**’s ID in Youtube’s database
- *Views*: (int) the number of times the video has been watched
- *#Shorts*: (bool) whether the video is a short
- *Date*: (string) the month and day the video was published

(a) **(3.0 pt)** Complete this Python expression, which evaluates to the name of the most watched video.

```
videos._________.item(0)
```

(b) **(4.0 points)**

Complete this Python expression, which evaluates to an array with two items (in any order): the average number of views for #Shorts and the average number of views for non-#Shorts.

```
videos._________(_________)._________('Views average')
          (a)        (b)         (c)
```

i. **(1.0 pt)** Which of these could fill in blank (a)?

○ `sort`

○ `where`

○ `group`

ii. **(2.0 pt)** Fill in blank (b).

iii. **(1.0 pt)** Which of these could fill in blank (c)?

○ `select`

○ `take`

○ `column`

(c) **(3.0 pt)** Complete this Python expression, which visualizes the distribution of the number of views for videos with names that include a `#` character (i.e. a hashtag).

```
videos._________
```

Recall: The `videos` table has columns *Name*, *ID*, *Views*, *#Shorts*, and *Date*.

(d) **(4.0 points)**

Complete this Python expression, which visualizes the distribution of the counts of #Shorts vs. non-#Shorts, including only videos that have more than 10,000 views.

```
videos.where(_________)._________._________
                (a)         (b)        (c)
```

Recall: The `videos` table has columns *Name*, *ID*, *Views*, *#Shorts*, and *Date*.

i. (2.0 pt) Fill in blank (a).

ii. (1.0 pt) Which of these could fill in blank (b)?

○ `column('#Shorts')`

○ `column('Views')`

○ `pivot('#Shorts', 'Views')`

○ `pivot('Views', '#Shorts')`

○ `group('Views')`

○ `group('#Shorts')`

iii. (1.0 pt) Which of these could fill in blank (c)?

○ `hist('count')`

○ `hist('#Shorts')`

○ `hist('Views')`

○ `barh('#Shorts', 'count')`

○ `barh('Views', 'count')`

(e) (3.0 pt) Complete this Python expression, which evaluates to a table with one row for each unique date and three columns: the date, as well as the total number of views of #Shorts and non-#Shorts videos on that date. Any column labels are acceptable.

```
videos._______________
```

Recall: The `videos` table has columns *Name*, *ID*, *Views*, *#Shorts*, and *Date*.

(f) (5.0 points)

Cory's colleague Bradley notices that the `videos` table doesn't contain the names of the channels.

Bradley creates an additional table called `channels` that contains all Youtube channels and has two columns:

- *Identifier*: (string) the channel's ID in Youtube's database
- *Channel*: (string) the channel's name

Bradley suspects that channels with exactly 4 characters in their name such as Vevo could be good channels to partner with since even their lowest performing videos have many views.

Complete her code, which assigns `min_views` to the minimum number of views received by a video posted to a channel with a 4-character name.

Hint: Calling `len` on a string returns the number of characters in the string.

```
channels = channels.with_column('Name Length', _________)
                                                   (a)

min_views = np.min(videos._________.where('Name Length', 4)._________)
                              (b)                                (c)
```

Recall: The `videos` table has columns *Name*, *ID*, *Views*, *#Shorts*, and *Date*.

i. **(2.0 pt)** Fill in blank (a).

ii. **(2.0 pt)** Fill in blank (b).

iii. **(1.0 pt)** Fill in blank (c).

3. (26.0 points) Supes

Hughie and Kimiko are investigating the Supes, a group of paid superheroes. On a regular basis, the Supes are alerted by emergencies in their area and try to arrive on the scene as fast as possible to save the day.

To get a sense of the Supes' speed, Hughie plans to randomly sample response times from public records.

(a) **(2.0 pt)** Suppose that Hughie wants to randomly sample 100 Supe response times to create a confidence interval for the **population mean** of Supe response times.

Which of the following could be used to help him create such a confidence interval? *Select all that apply.*

□ A/B Testing

□ Bootstrapping

□ Classification

□ Central Limit Theorem

□ None of the above

(b) **(2.0 pt)** Suppose Hughie wants to randomly sample Supe response times to create a **95%** confidence interval for the **population median** of response times, and he knows that the population SD is 40 seconds.

What is the minimum sample size he needs to create a confidence interval that has a width of 8 seconds?

○ 3200

○ 1600

○ 800

○ 400

○ 200

○ 100

○ 40

○ 8

○ There is not enough information to answer

(c) **(2.0 pt)** Suppose Hughie wants to randomly sample Supe response times to create a **95%** confidence interval for the **population mean** of response times, and he knows that the population SD is 20 seconds.

What is the minimum sample size he needs to create a confidence interval that has a width of 4 seconds?

○ 3200

○ 1600

○ 800

○ 400

○ 200

○ 100

○ 20

○ 4

○ There is not enough information to answer

(d) (4.0 pt) Suppose that Hughie randomly samples 100 Supe response times and uses the Central Limit Theorem to create a **95%** confidence interval for the **population mean** Supe response time, which he finds is (450, 550).

Which of the following can be concluded from the confidence interval above?

Select all that apply.

□ If Hughie's friend Kimiko repeats this process 500 times, she can expect that roughly 95% of the intervals she creates will contain the true population mean.

□ If Hughie randomly samples 1,000 response times without replacement, he can expect roughly 95% of the Supe response times to be between 450 and 550 seconds.

□ 95% of Supe response times in the population are between 450 and 550 seconds.

□ The average Supe response time in Hughie's sample was exactly 500 seconds.

□ None of the above.

(e) (4.0 pt) Hughie suspects that the Supes' average response time is slower than the 7 minute (420 seconds) average response time for local law enforcement.

Based on a **95%** confidence interval for the **population mean** Supe response time of (450, 550) seconds, if his p-value cutoff is **1%**, what should he conclude?

Select all that apply.

□ The data are consistent with the hypothesis that Supes respond to emergencies more slowly than local enforcement do.

□ The data are consistent with the hypothesis that the distribution of the emergency response times is the same for both the Supes and local law enforcement.

□ The data are consistent with the hypothesis that the Supes respond to emergencies more quickly than local enforcement do.

□ There is not enough information to make any of these conclusions.

(f) (3.0 pt) Suppose that Kimiko creates her own random sample of 100 Supe response times. She observes a sample average of 450 seconds for response time and she also knows that the population SD is 20 seconds.

What is her **95%** confidence interval for the true **population mean** of Supe response time (in seconds)?

○ (446, 454)

○ (448, 452)

○ (449.6, 450.4)

○ (449.8, 450.2)

○ (460, 540)

○ (480, 520)

○ None of the above.

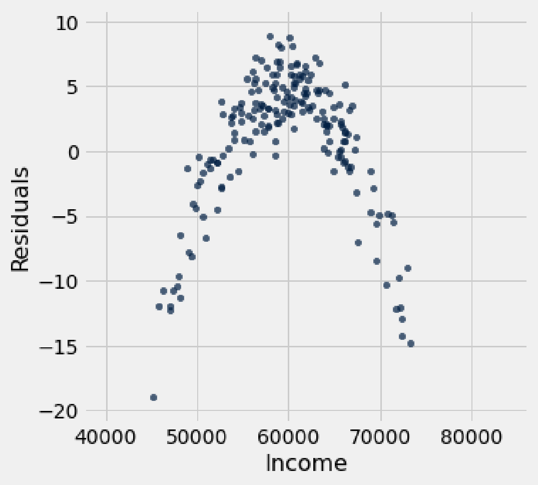
(g) **(3.0 pt)** Suppose that Kimiko creates two different random samples of 100 Supe response times and constructs a **95%** confidence interval for the true **population mean** of Supe response time using each one. What is the chance that at least one of them contains the population mean?

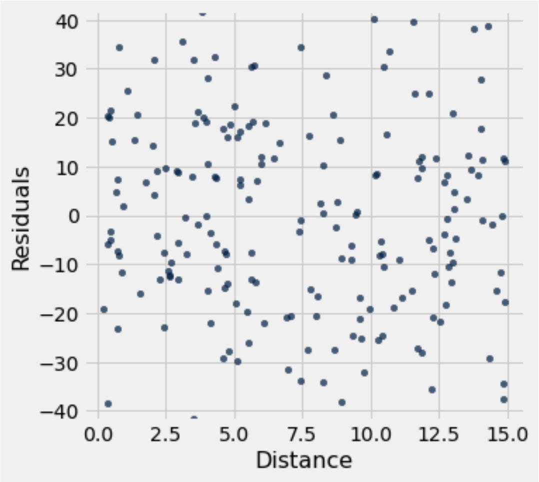
(h) **(6.0 points)**

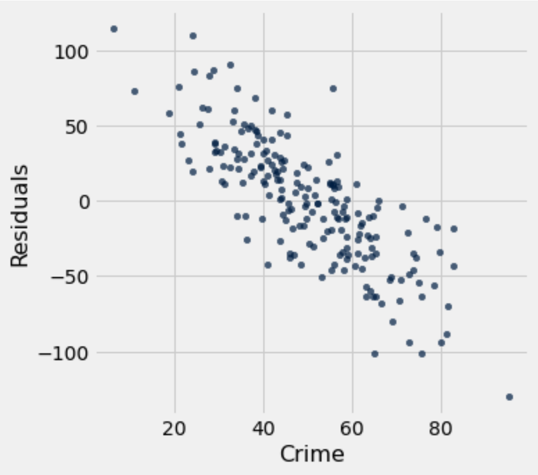
Suppose Kimiko tries to predict the emergency response time of the Supes from each of the following four different variables:

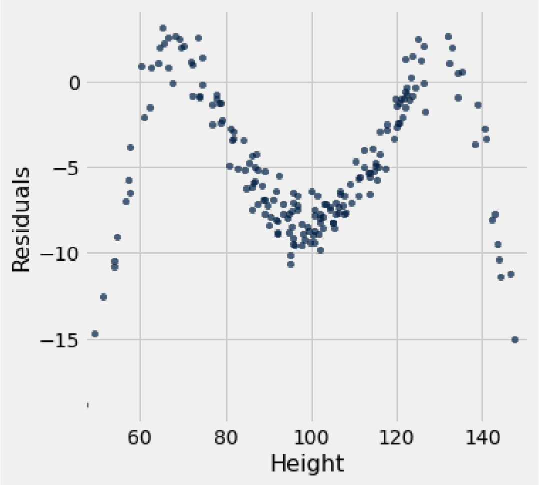
- *Income*: (float) the average income in the emergency location's neighborhood
- *Distance*: (float) the distance in miles from the emergency location to their headquarters
- *Crimes*: (int) the average weekly number of crimes reported in the emergency location's neighborhood last year
- *Height*: (float) the average building height (in feet) in the emergency location's neighborhood

To assess her predictions from each variable, she creates the following residual plots:









i. (2.0 pt) Which of the plots above indicate that the variable is linearly associated with response time?

Select all that apply.

□ Income

□ Distance

□ Crimes

□ Height

□ None of the above.

ii. (2.0 pt) Kimiko suspects she may have made a mistake in her code when plotting the residuals. Which of the plots above are impossible residual plots?

Select all that apply.

□ Income

□ Distance

□ Crimes

□ Height

□ None of the above.

iii. (2.0 pt) For which of the above plots should Kimiko try a quadratic equation for regression?

Recall: a quadratic equation is one of the form $ax^2 + bx + c$.

Select all that apply.

□ Income

□ Distance

□ Crimes

□ Height

□ None of the above.

4. (16.0 points) Spiderverse

Dr. Strange, a former neurosurgeon, is building a device to help him travel to parallel universes, each of which contains an alternate variant of our planet Earth (for example, in Earth-833, the CEO of Twitter is female entrepreneur Elona Musk).

There is a small percentage of these parallel Earths (exactly *1%*) that contains a superhero named Spiderman.

Dr. Strange wants to find a Spiderman, but he doesn't have time to explore every Earth variant.

Suppose Dr. Strange knows that if an Earth variant has a Spiderman, 95% of the time it has a man named Peter Parker. If an Earth variant doesn't have a Spiderman, it has a Peter Parker only 7% of the time.

(a) (3.0 pt) If Dr. Strange randomly selects an Earth variant to visit, what is the probability that it does not have a Peter Parker?

○ 0.95×0.93

○ 0.01

○ 0.05

○ 0.93

○ $0.01 \times 0.05 + 0.99 \times 0.93$

○ 0.01×0.93

○ 0.01×0.05

○ There is not enough information to answer.

(b) (3.0 pt) If Dr. Strange randomly selects an Earth variant to visit, what is the probability that it has a Spiderman **and** does not have a Peter Parker?

○ 0.95×0.93

○ 0.01

○ 0.05

○ 0.93

○ $0.01 \times 0.05 + 0.99 \times 0.93$

○ 0.01×0.93

○ 0.01×0.05

○ There is not enough information to answer.

(c) **(3.0 pt)** Suppose Dr. Strange's random selection leads him to visit the Zombie Earth variant, which is inhabited by zombies instead of humans.

Dr. Strange discovers that Zombie Earth does **not** have a Peter Parker in it.

Given this information, what is the probability that Zombie Earth has a Spiderman?

○ 0.05

○ $\dfrac{0.01 \times 0.95}{0.01 \times 0.95 + 0.99 \times 0.07}$

○ $\dfrac{0.99 \times 0.95}{0.99 \times 0.95 + 0.01 \times 0.07}$

○ $\dfrac{0.01 \times 0.05}{0.01 \times 0.05 + 0.99 \times 0.93}$

○ 0.93

(d) **(3.0 pt)** Suppose Dr. Strange's random selection leads him to visit Earth-42, which he discovers has a girl named MJ.

Prior to inspecting all of Earth-42's inhabitants, the information above makes Dr. Strange believe there is a 60% probability that it has a Spiderman.

Suppose that upon inspection, Dr. Strange learns that Earth-42 does **not** have a man named Peter Parker.

Given this new information, and assuming the conditional probabilities in the problem statement are still valid, what is Dr. Strange's subjective probability that Earth-42 does **not** have a Spiderman?

○ $0.4 \times 0.93 + 0.6 \times 0.05$

○ $\dfrac{0.4 \times 0.93}{0.4 \times 0.93 + 0.6 \times 0.05}$

○ $\dfrac{0.6 \times 0.05}{0.6 \times 0.05 + 0.4 \times 0.93}$

○ $\dfrac{0.4 \times 0.95}{0.4 \times 0.95 + 0.6 \times 0.07}$

○ $\dfrac{0.6 \times 0.95}{0.4 \times 0.95 + 0.6 \times 0.07}$

(e) **(4.0 pt)** Bayes' Rule can be used to do which of the following?

Select all that apply.

□ Quantify subjective beliefs about uncertainty.

□ Update probabilities based on new information.

□ Calculate conditional probabilities.

□ Determine, in multi-stage random experiments, the probability of an earlier stage outcome given the outcome of a later stage.

□ None of the above.

5. **(12.0 points) Data 8 Merch**

Berkeley plans to randomly select two people **without replacement** on Data 8 course staff to each receive a Data 8 sweater. There are 2 professors, 30 tutors, and 50 TAs on staff.

Important: Assume that each individual staff member is equally likely to be selected.

(a) **(3.0 pt)** Suppose we know that one professor received a sweater and the other didn't. Given this information, what is the probability that no tutor gets a sweater?

○ $\frac{30}{82}$

○ $\frac{52}{82}$

○ $\frac{50}{82}$

○ $\frac{50}{80}$

○ $\frac{52}{82} \times \frac{51}{81} \times \frac{50}{80}$

○ There is not enough information to answer.

(b) **(3.0 pt)** What is the probability that only TAs receive sweaters?

○ $\frac{50}{82} \times \frac{50}{81}$

○ $\frac{50}{82} \times \frac{50}{82}$

○ $\frac{50}{82} \times \frac{49}{81}$

○ $2 \times \frac{50}{82}$

○ $1 - \frac{50}{82} \times \frac{50}{81}$

○ $1 - \frac{50}{82} \times \frac{50}{82}$

○ $1 - \frac{50}{82} \times \frac{49}{81}$

○ There is not enough information to answer.

(c) **(3.0 pt)** What is the probability that at least one professor receives a sweater?

○ $1 - \frac{80}{82} \times \frac{79}{81}$

○ $\frac{80}{82} \times \frac{79}{81}$

○ $\frac{2}{82}$

○ $1 - \frac{2}{82} \times \frac{2}{81}$

○ $\frac{2}{82} \times \frac{2}{81}$

○ There is not enough information to answer.

(d) **(3.0 pt)** What is the probability that either both professors receive a sweater or neither do?

○ $\frac{1}{82} + \frac{80}{82} \times \frac{79}{81}$

○ $\frac{2}{82} + \frac{80}{82} \times \frac{80}{81}$

○ $\frac{1}{82} \times \frac{1}{81} + \frac{80}{82} \times \frac{79}{81}$

○ $\frac{1}{82} \times \frac{1}{81} + \frac{80}{82} \times \frac{80}{81}$

○ $\frac{2}{82} \times \frac{1}{81} + \frac{80}{82} \times \frac{79}{81}$

○ $\frac{2}{82} \times \frac{1}{81} + \frac{80}{82} \times \frac{80}{81}$

6. (34.0 points) Waystar

Waystar is an organization that owns businesses and trades stocks in a variety of sectors including media, entertainment, tech, etc. Its executive team wants to understand the performance of its businesses.

Waystar executives Siobhan and Roman put together a table called `performance`, which contains randomly sampled public information about the various businesses' performance over the last 40 years. Here are the first few rows:

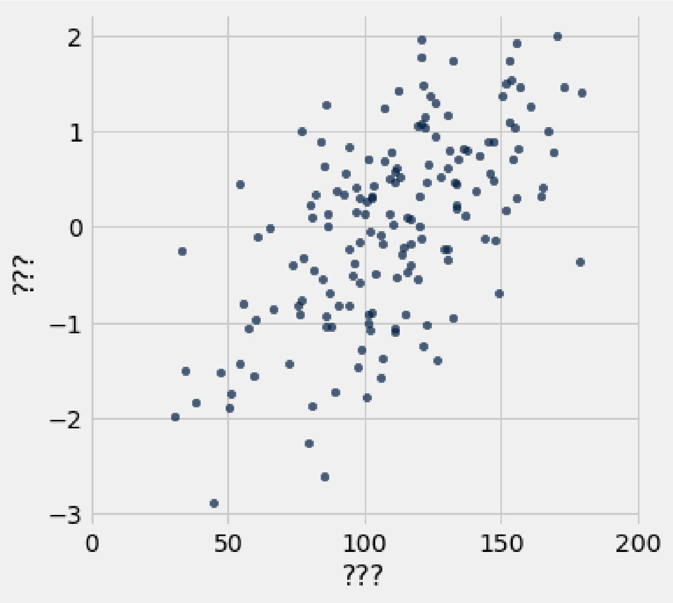
Name	Year	Revenue	Profit	Sector	Advice
Brightstar Cruises	2019	52.3	18.9	Entertainment	Hold
Adventure Parks	2018	42.8	16.3	Entertainment	Sell
Vaulter	2021	150.9	80.2	Tech	Buy
ATN	2020	61.7	48.7	Media	Hold
Adventure Parks	2019	49.2	15.8	Entertainment	Sell

. . . (158 rows omitted)

The table contains the following columns:

- *Name*: (string) the name of the business
- *Year*: (int) the year of the financial performance
- *Revenue*: (float) the business's revenue (in millions of USD)
- *Profit*: (float) the business's profit margin (a **percentage** between `0` and `100`)
- *Sector*: (string) the business's sector in the industry
- *Advice*: (string) Wall Street's stock purchase advice (`'Buy'`, `'Hold'` or `'Sell'`)

Roman uses the table to make the following scatterplot. Unfortunately, he forgets to label the axes of the plot!



(a) **(3.0 pt)** Assuming that all 163 values are shown on the plot, which of the following are possible variables that could be shown along the horizontal x-axis and vertical y-axis?

Select all that apply.

□ $x =$ *Year*, $y =$ *Profit* in standard units

□ $x =$ *Profit*, $y =$ *Profit* in standard units

□ $x =$ *Profit*, $y =$ *Revenue* in standard units

□ $x =$ *Revenue*, $y =$ *Profit* in standard units

□ None of the above

(b) **(6.0 points)**

For the next three questions, assume you know the following:

- the *Profit* column has a mean of 50 and a standard deviation of 10
- the *Year* column has a mean of 2000 and a standard deviation of 5
- the correlation between the *Profit* and *Year* columns is 0.5

i. (2.0 pt) Suppose Siobhan wants to predict *Profit* from *Year* and decides to fit a regression line.

What is the **intercept** of this regression line?

○ 2050

○ 1950

○ 1996.25

○ 0

○ -1950

○ -2050

○ None of the above

ii. (2.0 pt) For Waystar businesses in 2008, what would this regression line predict as the profit?

○ 50

○ 52

○ 54

○ 58

○ 66

○ None of the above

iii. (2.0 pt) What are the units for the residuals of this regression line?

○ Years

○ Dollars

○ Millions of Dollars

○ Shrute Bucks

○ Dogecoin

○ None of the above

(c) (7.0 points)

To forecast Waystar's performance, Siobhan wants to understand what the businesses' profits might be in future years.

She wants to construct a 95% confidence interval for the true height of the regression line between *Year* and *Profit* by bootstrapping the regression line 10,000 times.

She first creates a `correlation` function, which returns the correlation between two numerical arrays.

Complete her `profit_interval` function, which takes in a *year* (int) and returns the confidence interval for the true profit (as an array) for that year:

```
def profit_interval(year):
    profits = make_array()
    for i in np.arange(10000):
        boot_data = performance._________
                                   (a)
        boot_x = boot_data.column('Year')
        boot_y = boot_data.column('Profit')
        slope = correlation(boot_x, boot_y) * _________
                                                 (b)
        intercept = _________
                       (c)
        profit = _________
                    (d)
        profits = np.append(profits, profit)
    left = percentile(2.5, profits)
    right = percentile(97.5, profits)
    return make_array(left, right)
```

Recall: The `performance` table has columns *Name*, *Year*, *Revenue*, *Profit*, *Sector* and *Advice*.

i. (1.0 pt) Fill in blank (a).

ii. (2.0 pt) Fill in blank (b).

iii. (2.0 pt) Fill in blank (c).

iv. (2.0 pt) Fill in blank (d).

(d) (6.0 points)

Suppose Roman now makes a scatterplot of *Revenue* against *Profit*. He notices that instead of a linear trend, there is a nonlinear trend.

He decides to try to predict *Revenue* from *Profit* using a nonlinear regression curve for which the prediction equation has the form:

$$\text{revenue}_{\text{predicted}} = \text{slope} \times \text{profit}^{\text{exponent}} + \text{intercept}$$

For example, if *slope* is 4, *exponent* is 2 and *intercept* is 50, the nonlinear regression will predict the following for a *profit* of 10:

$$4 \times 10^2 + 50 = 4 \times 100 + 50 = 450$$

To find the optimal values of *slope*, *intercept* and *exponent*, Roman writes an `rmse` function.

Complete the `rmse` function, which returns the root mean squared error of the nonlinear regression for any given values of the intercept, slope and exponent:

```
def rmse(slope, intercept, exponent):
    x = performance.column('Profit')
    y = performance.column('Revenue')

    y_predicted = _________
                     (a)

    return (_________) ** 0.5
               (b)
```

i. **(4.0 pt)** Fill in blank (a) **without** using `performance` in your solution. (You may use `x` and `y`.)

ii. **(2.0 pt)** Which of these could fill in blank (b)?

○ `np.mean((x-y) ** 2)`

○ `np.mean((x-y_predicted) ** 2)`

○ `np.mean((y-y_predicted) ** 2)`

○ `np.mean(sum(x-y) ** 2)`

○ `np.mean(sum(x-y_predicted) ** 2)`

○ `np.mean(sum(y-y_predicted) ** 2)`

(e) **(12.0 points)**

Roman notices that businesses in the `'Tech'` sector have a different *Advice* distribution than those in `'Entertainment'`. Siobhan thinks the differences observed in the sample are only due to chance.

i. **(3.0 pt)** Which of these are **alternative** hypotheses that Roman could use to assess his claims?

Select all that apply

□ Businesses in `'Tech'` have a different *Advice* distribution than those in `'Entertainment'`.

□ Businesses in `'Tech'` have a different *Advice* distribution than those in non-`'Tech'`.

□ Businesses in `'Tech'` have a different distribution than those in `'Entertainment'`.

□ `'Buy'` businesses have a different *Sector* distribution than `'Sell'` businesses.

□ `'Buy'` businesses have a different *Sector* distribution than non-`'Buy'` businesses.

□ None of the above.

ii. **(3.0 pt)** Which of the following test statistics could Roman use to assess his claims?

Select all that apply

□ The total variation distance between the *Sector* distribution of `'Buy'` businesses and the *Sector* distribution of `'Sell'` businesses.

□ The total variation distance between the *Advice* distribution of `'Tech'` businesses and the *Advice* distribution of `'Entertainment'` businesses.

□ The total variation distance between the *Advice* distribution of `'Tech'` businesses and the *Advice* distribution of non-`'Tech'` businesses.

□ Absolute difference of mean *Advice* for `'Tech'` businesses minus mean *Advice* for `'Entertainment'` businesses.

□ Absolute difference of mean *Advice* for `'Tech'` businesses minus mean *Advice* for non-`'Tech'` businesses.

□ None of the above.

iii. **(3.0 pt)** Roman simulates 1,000 values of the test statistic under the null and stores these in an array called `test_stats`. Suppose the observed value of the test statistic is `0.52`.

Write a Python expressions that returns the p-value for this hypothesis test.

iv. **(3.0 pt)** Roman finds that `percentile(96, test_stats)` evaluates to `0.5`.

If his *p*-value cutoff is **5%** and the observed test stastic is `0.52`, which of the following can he conclude?

Select all that apply.

□ The data are consistent with the null hypothesis.

□ The data are consistent with the alternative hypothesis.

□ The null hypothesis is true.

□ The null hypothesis is false.

□ There is not enough information to make any of these conclusions.

7. (32.0 points) AFC Richmond

Rebecca Welton, owner of the English football club AFC Richmond, is trying to use the team's past performance to make predictions about upcoming matches. She randomly samples the team's matches from the past 10 years and puts them in a table called `matches`. The first few rows are shown here:

Opponent	Home	Streak	Prior Goals	Goals	Outcome
Manchester	True	0	1.4	2	Draw
West Ham	True	3	2.2	4	Win
Everton	False	0	0.4	1	Lose

The table contains the following columns:

- *Opponent*: (string) the name of the opposing team
- *Home*: (bool) whether the match was played at home in the Richmond stadium
- *Streak*: (int) the number of matches that Richmond won in a row leading up to the match
- *Prior Goals*: (float) the average number of goals that Richmond scored in the prior 5 matches
- *Goals*: (int) the number of goals that Richmond scored in the match
- *Outcome*: (string) whether Richmond's outcome of the match was a win, loss or draw

(a) (2.0 pt) Suppose Rebecca would like to understand how *Prior Goals* varies between matches that were won and matches that were lost.

Which of the following would be most appropriate to visualize the relationship between these variables?

○ Scatterplot

○ Pivot Table

○ Total Variation Distance

○ Overlaid Histograms

○ Line Graph

○ Bar Chart

(b) (4.0 pt) Suppose Rebecca wants to understand how the distribution of *Outcome* varies between home games and away games.

Which of the following could be used to help understand the relationship between these variables?

Select all that apply.

□ Scatterplot

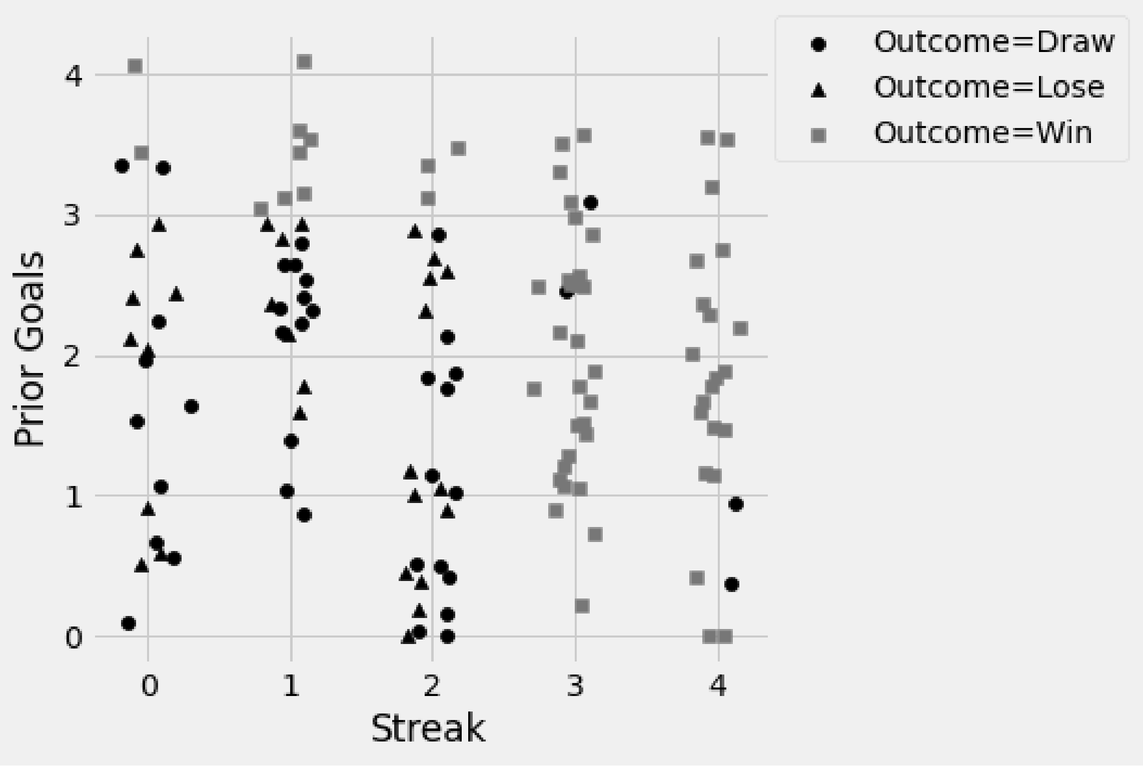
□ Pivot Table

□ Histogram

□ Line Graph

□ Bar Chart

(c) **(3.0 pt)** Rebecca creates the following scatterplot of *Streak* against *Prior Goals*, with the shape of each point representing the value of *Outcome*.



Note: The points on the plot have been horizontally jittered (i.e. every point has been moved horizontally by a small random amount) so that it's easier to view each individual point.

Suppose Rebecca want to use *Streak* and *Prior Goals* to create a classifier that can predict the outcome of a match. She writes the following partially completed code:

```
def classify(streak, prior_goals):
    if prior_goals > 3:
        return 'Win'
    elif _________:
        return 'Draw'
    else:
        return 'Lose'
```

Which of the following Python expressions, if used to fill in the blank above, would result in a classifier that never classifies a training point as `'Draw'` when the true class is `'Win'`.

Note: The classify function takes in unjittered points as input.

Select all that apply.

□ `prior_goals > 2`

□ `prior_goals > 2.5`

□ `streak < 2.5`

□ `streak < 2`

□ None of the above.

(d) **(2.0 pt)** Suppose Rebecca instead builds a k-nearest-neighbor classifier with $k = 5$ to predict the outcome of a Richmond match, using unjittered *Streak* and *Prior Goals* as its features.

Suppose AFC Richmond's next match is on Saturday and we know the following:

- The team currently has a win streak of 2
- The team average 1 goal during its last 5 matches

What would this nearest neighbor classifier predict as the *Outcome* for Saturday's match?

○ `'Win'`

○ `'Lose'`

○ `'Draw'`

○ There is not a majority class.

(e) **(2.0 pt)** What would a nearest neighbor with $k = 11$ predict for the *Outcome* in the question above?

○ `'Win'`

○ `'Lose'`

○ `'Draw'`

○ There is not a majority class.

(f) **(3.0 pt)** Out of the 127 matches in the table, Richmond won 25 times. If Rebecca were to build a k-nearest-neighbors classifier to predict the outcome of a match, which values of k are guaranteed to result in always predicting "Lose"?

Recall: The outcome of a match can be win, lose or draw.

Hint: To answer this question you **do not** need to use any information from the scatterplot above.

Select all that apply.

□ 127

□ 103

□ 100

□ 26

□ 25

□ There is not enough information to tell.

(g) **(2.0 pt)** True or False: If Rebecca uses $k = 9$ for the above Nearest Neighbor classifier, there is guaranteed to be a single majority class among the k nearest neighbors' outcomes for any possible values of *Streak* and *Prior Goals*. *Recall*: The outcome of a match can be win, lose or draw.

○ True

○ False

(h) **(8.0 points)**

Complete the `neighbors` function, which takes in the following arguments:

- `train`: A table with columns *Streak*, *Prior Goals*, and *Goals*. Each row represents a match.
- `new_match`: A length-two array containing Richmond's win streak and prior 5-game average goals. For example, `array([3, 1])` represents a win streak of 3 and a 5-game average of 1 goal.
- `k`: The value of k to use for k nearest neighbors.

It returns a table containing the k rows in `train` that are the nearest neighbors to `new_match` with respect to *Streak* and *Prior Goals* features.

```
def neighbors(train, new_match, k):
    streak_diffs = _________
                       (a)
    prior_goal_diffs = _________
                           (b)
    distances = _________
                    (c)
    train_dist = train.with_column(‘Distance’, distances)
    return train_dist.sort('Distance')._________
                                           (d)
```

i. **(2.0 pt)** Fill in blank (a).

ii. **(2.0 pt)** Fill in blank (b).

iii. **(2.0 pt)** Which of these could fill in blank (c)?

○ `streak_diffs - prior_goal_diffs`

○ `np.mean(streak_diffs - prior_goal_diffs)`

○ `np.mean((streak_diffs - prior_goal_diffs) ** 2) ** 0.5`

○ `(streak_diffs ** 2 + prior_goal_diffs ** 2) ** 0.5`

○ `np.mean(streak_diffs ** 2 + prior_goal_diffs ** 2) ** 0.5`

iv. **(2.0 pt)** Fill in blank (d).

(i) (6.0 points)

Suppose that Rebecca now wants to use k-nearest-neighbors to predict the number of *Goals* that Richmond will score in an upcoming match based on the team's *Streaks* and *Prior Goals* going into the match.

To generate a prediction for an upcoming match, Rebecca decides to compute the *root mean square* of the k-nearest-neighbors' *Goals.* The root mean square of a set of values is computed by squaring each value, taking the average of the squares, and finally taking the square root of the average.

For example, if $k = 3$ and the 3 nearest neighbors have *Goals* of 0, 1, & 2, then the *root mean square* is:

$$RMS = \sqrt{\frac{0^2 + 1^2 + 2^2}{3}} = 1.29$$

Complete the `prediction` function that takes in the same arguments as the `neighbors` function (i.e., `train`, `new_match` and `k`) and returns the *root mean square* of the *Goals* values from among the new match's k nearest neighbors.

```
def prediction(train, new_match, k):
    neighbor_goals = neighbors(_________)._________
                                 (a)         (b)

    rms = (_________ / _________) ** 0.5
              (c)         (d)
    return rms
```

i. (2.0 pt) Fill in blank (a).

ii. (1.0 pt) Fill in blank (b).

iii. (2.0 pt) Fill in blank (c).

iv. (1.0 pt) Fill in blank (d).

8. (0.0 points) Optional

(a) If there was any question on the exam that you thought was ambiguous and required clarification to be answerable, please identify the question (including the title of the section, e.g., Experiments) and state your assumptions. Be warned: We only plan to consider this information if we agree that the question was erroneous or ambiguous and we consider your assumption reasonable.

(b) Prof. Sahai hasn't seen a single episode of one of the following shows. Which is it?

○ Succession

○ The Boys

○ Euphoria

○ The Morning Show

○ Ted Lasso

(c) Draw a picture or share a few words describing your experience in Data 8!