# DATA 8 Foundations of Data Science

Fall 2022

Final Exam

**INSTRUCTIONS**

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

- The exam is closed book, closed notes, closed computer, closed calculator, except the provided midterm and final study guides.
- 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 select exactly *one* choice.

○ You must choose either this option

○ Or this one, but not both!

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

□ You could select this choice.

□ You could select this one too!

**Preliminaries**

Exam structure:

- True or False - 32 points
- DataTok - 19 points
- Paper - 17 points
- Sandpiper - 39 points
- Adventure - 12 points
- Verse Jumping - 16 points
- Stonks - 25 points
- A Last Words section, where you can state any assumptions you made on the exam

You can complete and submit these questions before the exam starts.

**(a)** What is your full name?

**(b)** What is your student ID number?

**(c)** Who is your lab GSI? You may write *Unknown* if you don't know their name.

1. **(32.0 points)** **True or False**

**(a)** **(2.0 pt)** In order to build a $k$ Nearest Neighbors classifier that uses the whole training set, you need to know the class value of some, but not all of the training examples.

○ True

○ False

**(b)** **(2.0 pt)** A classifier is considered to be overfitting if it performs very well on both the training set and the test set.

○ True

○ False

**(c)** **(2.0 pt)** Bayes' Rule can be used to model subjective beliefs about events that involve randomness.

○ True

○ False

**(d)** **(2.0 pt)** In linear regression, slope is measured in the same units as the numerical attribute on the y-axis.

○ True

○ False

**(e)** **(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 median* is roughly normal.

○ True

○ False

**(f)** **(2.0 pt)** If we use linear regression to predict $y$-values based on our $x$-values, where both $x$ and $y$ are measured in standard units, the intercept is guaranteed to be 0.

○ True

○ False

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

○ True

○ False

**(h)** **(2.0 pt)** If you use k-nearest neighbors on a data set that has only 2 possible categories for class (e.g. 0 or 1) and use an odd value for $k$ (e.g. 5, 9), there is guaranteed to be a unique class that has the majority among the $k$ nearest neighbors in the training set.

○ True

○ False

**(i)** **(2.0 pt)** If we use linear regression to predict $y$-values based on our $x$-values, the average of our residuals will depend on whether $x$ and $y$ are measured in standard units.

○ True

○ False

**(j)** **(2.0 pt)** If you are sampling a numerical attribute that can only take on values of 0 or 1, the SD of your sample cannot be larger than 0.5.

○ True

○ False

**(k)** **(2.0 pt)** The total variation distance can be applied to categorical distributions in which there are only 2 possible categories (e.g. purple or white).

○ True

○ False

**(l)** **(2.0 pt)** The reason we shuffle labels in an A/B test is to ensure that our subjects are randomly assigned to treatment and control.

○ True

○ False

**(m)** **(2.0 pt)** Suppose a hypothesis test is proposed and we already know that the null hypothesis is true. If 500 researchers each independently collect a sample of the same size to carry out an experiment and they all use 1% as their p-value cutoff, we should expect around 5 of them to reject the null.

○ True

○ False

**(n)** **(2.0 pt)** The chance of two events A and B both happening can sometimes be greater than the chance of either A or B (or both) happening.

○ True

○ False

**(o)** **(2.0 pt)** The distance between two individuals can be zero if calculated using only 2 numerical atrributes, but greater than zero if calculated using 3 numerical attributes.

○ True

○ False

**(p)** **(2.0 pt)** Evaluating a machine learning algorithm on a test set that was not involved in the training phase is a way to estimate classification accuracy on the population.

○ True

○ False

## 2. (19.0 points) DataTok

The Data8 community has been trying to spread the word about the course online by posting content on TikTok. To evaluate the performance of these posts, Nicole and Will put together a table called `videos` that contains a random sample of 331 Data 8 related TikTok videos posted in the last year. The first few rows are shown here:

| Caption | ID | Views | #Data8 | Date |
|---|---|---|---|---|
| berkeley | UC6D1L2vxEAg | 1774659 | True | 08-24 |
| lies or loophole | UComP_epzeKz | 96356 | False | 11-30 |
| best part was the 2000+ person data 8 class | UC3IZKseVp | 12628 | True | 08-24 |

The table has the following columns:

- *Caption*: (string) the video's caption (excluding hashtags)
- *ID*: (string) the posting account's ID in TikTok's database
- *Views*: (int) the number of unique users who viewed the video
- *#Data8*: (bool) whether the video had '#Data8' in the caption
- *Date*: (string) the month and day the video was published

**(a) (3.0 pt)** Which of these Python expressions returns the date of the most watched video?

*Select all that apply.*

☐ `videos.sort('Views').row(0).item(4)`

☐ `videos.sort('Views').column(4).item(0)`

☐ `videos.pivot('Date', 'ID', 'Views', max).column(0).item(0)`

☐ `videos.sort('Views', descending=True).column('Date').item(0)`

☐ `videos.select('Date', 'Views').group(0, max).sort(1, descending=True).column(0).item(0)`

☐ None of the above.

**(b) (3.0 pt)** Which of these Python expressions returns a table that displays the average number of video views for each combination of *Date* and *#Data8*?

*Select all that apply.*

☐ `videos.pivot(['Date', '#Data8', 'Views', np.average])`

☐ `videos.pivot('Date', '#Data8', 'Views', np.average)`

☐ `videos.select('Date', '#Data8', 'Views').apply(np.average, 2)`

☐ `videos.select('Date', '#Data8', 'Views').group([0, 1], np.average)`

☐ `videos.select('Date', '#Data8', 'Views').group([1, 0], np.average)`

☐ None of the above.

**(c)** **(4.0 pt)** Which of these Python expressions visualizes the distribution of the average number of views for each *Date* in the table?

*Select all that apply.*

□ `videos.group('Date', np.average).hist('Views')`

□ `videos.select(2, 4).group(0, np.average).hist(1)`

□ `videos.column('Date', 'Views').group(0, np.average).hist(1)`

□ `videos.select('Date', 'Views').group('Date', np.average).hist('Views average')`

□ `Table().with_column('Avg Views', videos.select(4, 2).apply(np.average, 1)).hist(0)`

□ None of the above.

**(d)** **(9.0 points)**

While looking at a table of videos is helpful, Nicole notices that the table doesn't contain the account names of who posted the video.

She creates a separate table called `accounts` that contains account names for all 1,225,082,327 TikTok users. The first few rows are shown here:

| Identifier | Account |
|---|---|
| UC6D1L2vxEAg | withloverico |
| UComP_epzeKz | caltvofficial |
| UC3IZKseVp | toomuchtrunko |

The table has the following columns:

- *Identifier*: (string) the account's ID in TikTok's database
- *Account*: (string) the account's handle

Nicole suspects that 'caltvofficial' could be a good channel to partner with since their videos seem to generate many views.

She writes the following partially completed code, which assigns `caption` to the caption of the video that has the maximum number of views among all videos posted by 'caltvofficial' with hashtag #Data8.

```
combined = ______(a)______
caltv = _____(b)_____
caption = ______(c)______
```

*Recall*: The `videos` table has columns *Caption*, *ID*, *Views*, *#Data8*, and *Date*.

**i.** **(3.0 pt)** Which of the following Python expressions could be used to fill in blank (a)?

*Select all that apply.*

□ `videos.join(‘ID’, accounts, ‘Identifier’)`

□ `videos.join(‘ID’, accounts, ‘Account’)`

□ `accounts.join('Account', videos, 'ID')`

□ `videos.with_column('Account', accounts.column(1))`

□ `accounts.with_columns('Caption', videos.column(0), 'Views', videos.column(2))`

□ None of the above.

**ii. (3.0 pt)** Which of the following Python expressions could be used to fill in blank (b)?

*Select all that apply.*

□ `combined.where(‘Identifier’, 'caltvofficial')`

□ `combined.where(‘Account’, 'caltvofficial')`

□ `combined.where('caltvofficial', are.equal_to('Account'))`

□ `combined.where(‘Account’, are.containing('caltvofficial'))`

□ `combined.where(‘Account’, are.equal_to('caltvofficial'))`

□ None of the above.

**iii. (3.0 pt)** Which of the following Python expressions could be used to fill in blank (c)?

*Select all that apply.*

□ `caltv.sort('Views').column('Caption').item(0)`

□ `caltv.sort('Views', descending=True).column('Caption').item(0)`

□ `caltv.sort('Views', descending=True).where('#Data8', True).column('Caption').item(0)`

□ `caltv.where('#Data8', True).sort('Views', descending=True).column('Caption').item(0)`

□ None of the above.

**3. (17.0 points) Paper**

Jim and Pam are doing quarterly review of their employer, Dunder Mifflin, a company that sells paper to small & medium sized businesses. One area the company has been investigating is its order shipping times.

To undersand this better, Pam plans to randomly sample shipments from their warehouse's records.

**(a) (2.0 pt)** Suppose that Pam wants to randomly sample shipping times to create a **95%** confidence interval for the **population mean** of shipping time and she knows that the population SD is 30 hours.

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

**(b) (2.0 pt)** Suppose that Pam wants to randomly sample shipping times to create a **68%** confidence interval for the **population mean** of shipping time and she knows that the population SD is 4 days.

What is the minimum sample size she needs to create a confidence interval that has a width of 2 days?

**(c) (5.0 points)**

Suppose that Pam randomly samples 100 shipping times and uses bootstrapping to create a **95%** confidence interval for the **population median** shipping time.

For the following two questions, assume that the confidence interval she constructs is (94, 106) hours.

**i. (3.0 pt)** Which of the following can be concluded from the confidence interval above?

○ 95% of shipping times in the population are between 94 and 106 hours.

○ The median shipping time in Pam's sample was exactly 100 hours.

○ If Jim independently repeats Pam's process 500 times, exactly 95% of the intervals he creates will contain the true population median.

○ If Jim randomly samples 100 shipments without replacement, he can expect roughly 95% of the shipping times to be between 94 and 106 hours.

○ None of the above.

**ii. (2.0 pt)** Pam suspects that the Dunder Mifflin ships paper slower than the median 2 days (48 hours) that Amazon Prime claims it takes to ship its paper orders.

Based on the above **95%** confidence interval of (94, 106) hours, if her p-value cutoff is **5%**, what should Pam conclude?

○ The data are consistent with the hypothesis that Dunder Mifflin ships paper slower than Amazon Prime does.

○ The data are consistent with the hypothesis that the distribution of the paper shipping times is the same for both Dunder Mifflin and Amazon Prime.

○ The data are consistent with the hypothesis that Dunder Mifflin ships paper faster than Amazon Prime does.

○ There is not enough information to make a conclusion of any kind.

**(d)** **(2.0 pt)** Suppose that Jim creates his own random sample of 100 shipping times. He observes a sample average of 96 hours for response time and she also knows that the population SD is 30 hours.

What is his **95%** confidence interval for the true **population mean** of shipping time (in hours)?

○ (87, 105)

○ (90, 102)

○ (93, 99)

○ (94, 106)

○ (95.4, 96.6)

○ (95.1, 96.9)

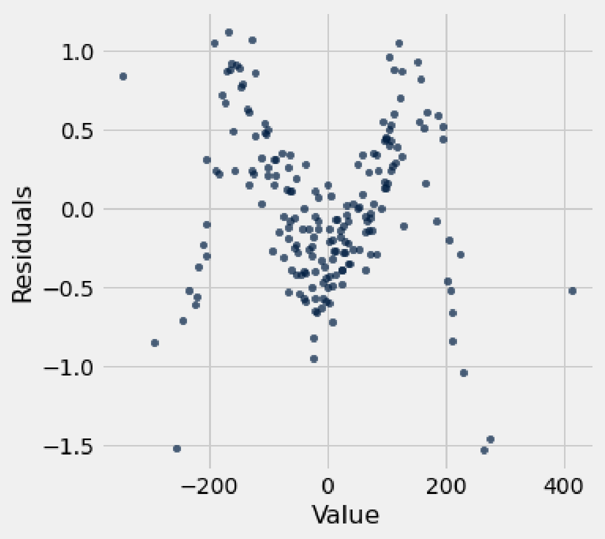
○ None of the above.

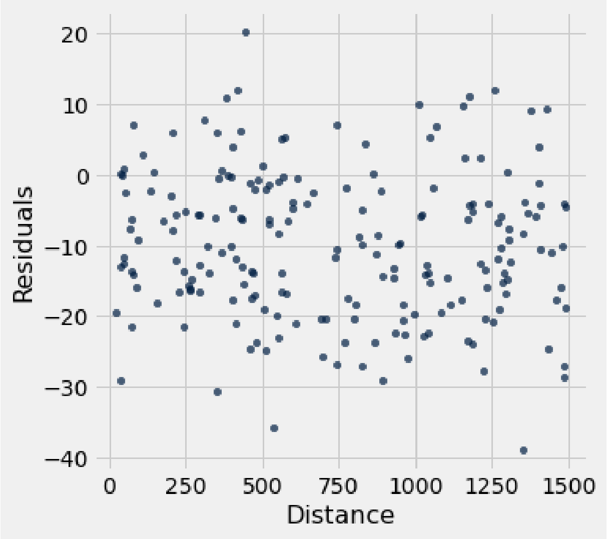
**(e)** **(6.0 points)**

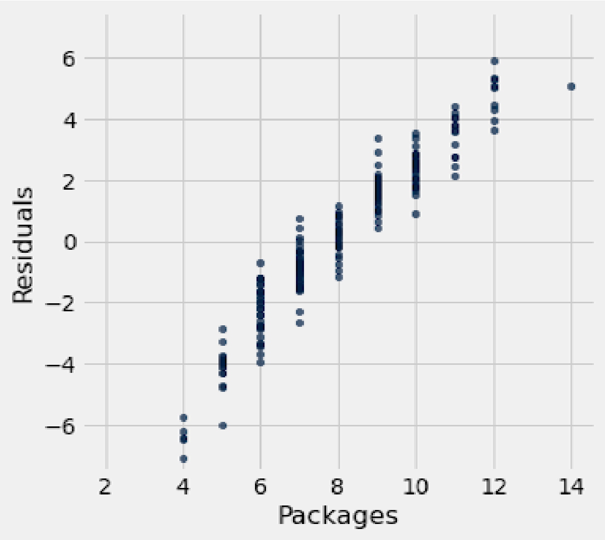
Suppose Pam tries to predict Dunder Mifflin's shipping time from each of the following four different variables:

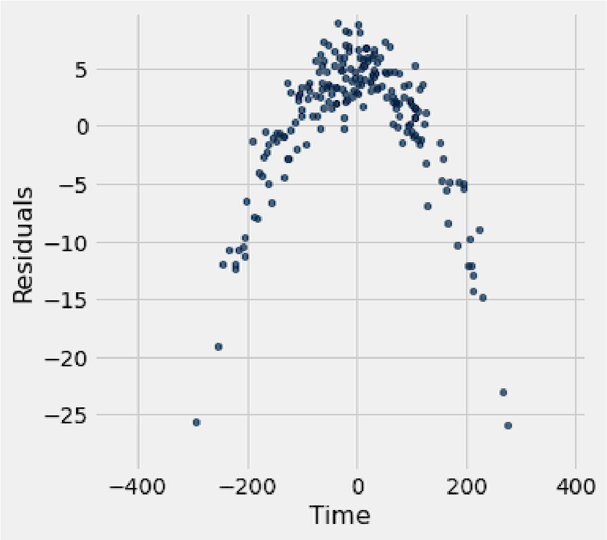
- *Value*: (float) the purchase value of the order (in dollars)
- *Distance*: (float) the distance from the delivery address to Dunder Mifflin's headquarters (in miles)
- *Packages*: (int) the number of packages required to deliver the order
- *Time*: (float) the time (in minutes) when the order was placed, after 5pm ET (e.g. 4:30pm would have a time of -30.0)

To assess her predictions from each variable, she creates four residual plots (two are on the next page):









**i.** **(2.0 pt)** Which of the plots above indicate that linear regression is a good fit?

*Select all that apply.*

□ Value

□ Distance

□ Packages

□ Time

□ None of the above.

**ii.** **(2.0 pt)** Pam 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.*

□ Value

□ Distance

□ Packages

□ Time

□ None of the above.

**iii.** **(2.0 pt)** For which of the above plots should Pam try a quadratic equation?

*Select all that apply.*

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

□ Value

□ Distance

□ Packages

□ Time

□ None of the above.

**4. (39.0 points) Sandpiper**

Sandpiper is a nursing home for the elderly that is being sued for overcharging its residents for basic necessities like toiletries, meals and medications.

Jimmy McGill is the attorney representing the residents in this case. He's put together a table called `statements`, which contains information from 263 randomly sampled annual statements that were given to the residents. Here are the first few rows:

| Name | Year | Cost | Charge | Category | Method |
|---|---|---|---|---|---|
| Irene Landry | 1999 | 116.3 | 218.9 | Meals | Digital |
| Alma May Urbano | 1996 | 142.8 | 316.3 | Beverages | Manual |
| Amos Lydecker | 2001 | 150.9 | 280.2 | Medications | Digital |
| Harold Goodman | 2000 | 161.7 | 348.7 | Toiletries | Manual |

The table contains the following columns:

- *Name*: (string) the name of the resident
- *Year*: (int) the year in which the statement was made
- *Cost*: (float) the fair market cost in dollars of all items in the statement
- *Charge*: (float) the amount that was charged in dollars to the resident for all items in the statement
- *Category*: (string) the category with the largest total charges in the invoice
- *Method*: (string) the method in which the statement was originally prepared ('Digital' or 'Manual')

**(a) (6.0 points)**

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

- the *Charge* column has a mean of 200 and a standard deviation of 10
- the *Cost* column has a mean of 100 and a standard deviation of 5
- the correlation between the *Charge* and *Cost* columns is 0.75

**i. (2.0 pt)** What is the **intercept** of a regression line fit to predict *Charge* from *Cost*?

**ii. (2.0 pt)** For a statement that has a *Cost* of 200, what would be the predicted *Charge* for a regression line fit to predict *Charge* from *Cost*?

**iii. (2.0 pt)** What are the units for the **slope** in the above regression?

○ Years

○ Dollars

○ Shrute Bucks

○ Dollars per Year

○ None of the above

**(b) (8.0 points)**

Jimmy decides to construct a confidence interval for the true slope of the regression line that predicts *Charge* from *Cost* by bootstrapping the regression line 10,000 times.

He first creates a `correlation` function, which returns the correlation between two numerical arrays. Then, he writes `slope_interval`, which takes in a confidence level (e.g. `90` would be inputted to generate a 90% confidence interval) and returns the confidence interval for the true slope as a two-element array:

```
def slope_interval(level):
    slopes = make_array()
    for i in np.arange(10000):
        boot_data = ______(a)______
        boot_x = boot_data.column('Cost')
        boot_y = boot_data.column('Charge')

        slope = ________(b)_________
        slopes = np.append(slopes, slope)
    left = _______(c)______
    right = _______(d)______  # Note: there is no question corresponding to this blank
    return make_array(left, right)
```

*Recall*: The `statments` table has columns *Name*, *Year*, *Cost*, *Charge*, *Category* and *Method.*

**i. (2.0 pt)** Which of the following could be used to fill in blank (a)? *Select all that apply.*

□ `statements`

□ `statements.sample()`

□ `statements.sample(level)`

□ `statements.sample(statements.num_rows)`

□ `statements.sample(with_replacement=False)`

□ None of the above

**ii. (3.0 pt)** Which of the following could be used to fill in blank (b)? *Select all that apply.*

□ `correlation(boot_x, boot_y)`

□ `correlation(np.std(boot_x), np.std(boot_y))`

□ `correlation(boot_x, boot_y) * np.std(boot_y) / np.std(boot_x)`

□ `correlation(boot_y, boot_x) * np.std(boot_y) / np.std(boot_x)`

□ `correlation(boot_x, boot_y) * np.std(boot_x) / np.std(boot_y)`

□ None of the above

**iii. (3.0 pt)** Which of the following could be used to fill in blank (c)? *Select all that apply.*

□ `percentile(0.05, slopes)`

□ `percentile(1-level, slopes)`

□ `percentile(level * 0.5, slopes)`

□ `percentile(0.5 - level/2, slopes)`

□ `percentile((1 - level) * 0.5, slopes)`

□ None of the above

**(c) (3.0 pt)** Suppose Jimmy instead creates the following function, which defines overcharging as a percentage of cost:

```
def overcharge(cost, charge):
    return (charge - cost)/cost * 100
```

Which of the following lines of code will return an array that contains the amount of **overcharging** for all of the statements in the table?

*Recall*: The `statments` table has columns *Name*, *Year*, *Cost*, *Charge*, *Category* and *Method.*

*Select all that apply.*

□ `statements.apply(overcharge)`

□ `statements.apply(overcharge, 'Cost', 'Charge')`

□ `statements.select('Cost','Charge').apply(overcharge)`

□ `statements.drop('Name','Year','Category', 'Method').apply(overcharge)`

□ `overcharge(statements.column('Cost'), statements.column('Charge'))`

□ None of the above

**(d) (10.0 points)**

Suppose Jimmy now adds the array from the previous question into the `statements` table as a new column called ***Overcharge***.

He then makes a scatterplot of ***Overcharge*** against ***Year*** and notices that there is a nonlinear pattern.

Jimmy decides to try to predict *Overcharge* from *Year* using a nonlinear regression whose prediction equation looks like the following:

$$\text{overcharge}_{\text{predicted}} = 2^{\text{slope}\times(\text{year}-1990)} + \text{intercept}$$

For example, if `slope` is 0.5 and `intercept` is 60, this nonlinear regression will predict the following for the *year* 2000:

$$2^{0.5*10} + 60 = 92$$

To find the optimal values of `slope` and `intercept`, Jimmy writes the following partially completed function, which returns the root mean squared error of the nonlinear regression for any given values of the slope and intercept:

```
def rmse(slope, intercept):
    x = ______(a)______
    y = ______(b)______
    y_predicted = ______(c)______
    return _______(d)________
```

*Recall*: The `statments` table now has columns *Name*, *Year*, *Cost*, *Charge*, *Category*, *Method* and *Overcharge*.

**i. (2.0 pt)** Which of the following lines of code could be used to fill in blank (a)?

*Select all that apply.*

□ `statements.column(1)`

□ `statements.column(2)`

□ `statements.column(6)`

□ `statements.column('Year')`

□ `statements.column('Overcharge')`

□ None of the above

**ii. (2.0 pt)** Which of the following lines of code could be used to fill in blank (b)?

*Select all that apply.*

□ `statements.column(1)`

□ `statements.column(6)`

□ `statements.column('Overcharge')`

□ `statements.column(3) - statements.column(2)`

□ `statements.column('Charge') - statements.column('Cost')`

□ None of the above

**iii. (3.0 pt)** Which of the following lines of code could be used to fill in blank (c)?

*Select all that apply.*

□ `slope * x + intercept`

□ `2 ** (slope * x) + intercept`

□ `2 * slope * (x - 1990) + intercept`

□ `2 ** (slope * (y - 1990)) + intercept`

□ `2 ** (slope * (x - 1990)) + intercept`

□ None of the above

**iv. (3.0 pt)** Which of the following lines of code could be used to fill in blank (d)?

*Select all that apply.*

□ `np.sqrt((y-y_predicted)**2)`

□ `np.sqrt(np.average(y - y_predicted))`

□ `np.sqrt(np.average(y_predicted - y))`

□ `np.sqrt(np.average((y_predicted - y) ** 2))`

□ `np.sqrt(np.average((y - y_predicted) ** 2))`

□ None of the above

**(e) (12.0 points)**

Jimmy notices that overcharges for statements prepared by a `'Manual'` method are typically higher than those prepared by a `'Digital'` method.

The nursing home's defense argues that any differences observed in the sample are only due to chance.

*Recall*: The `statments` table has columns *Name*, *Year*, *Cost*, *Charge*, *Category*, *Method* and *Overcharge*.

**i. (3.0 pt)** Which of the following is an alternative hypothesis that Jimmy could use to assess his claims?

*Select all that apply.*

□ Statements prepared by a `'Digital'` method have a lower *Overcharge* on average than statements prepared by a `'Manual'` method.

□ Statements prepared by a `'Manual'` method have a higher *Overcharge* on average than statements prepared by a `'Digital'` method.

□ Statements prepared by a `'Manual'` method have a lower *Overcharge* on average than statements prepared by a `'Digital'` method.

□ Statements prepared by a `'Manual'` method have a higher *Charge* on average than statements prepared by a `'Digital'` method.

□ Statements prepared by a `'Manual'` method have a higher *Cost* on average than statements prepared by a `'Digital'` method.

□ None of the above.

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

*Select all that apply.*

□ The total variation distance between the *Overcharge* distribution of `'Manual'` statements and the *Overcharge* distribution of `'Digital'` statements.

□ The mean *Charge* among `'Manual'` statements minus the mean *Charge* among `'Digital'` statements.

□ The mean *Overcharge* among `'Digital'` statements minus the mean *Overcharge* among `'Manual'` statements.

□ The mean *Overcharge* among `'Digital'` statements.

□ The mean *Cost* among `'Manual'` statements plus the mean *Cost* among `'Digital'` statements.

□ None of the above.

**iii. (3.0 pt)** Jimmy simulates 1,000 values of the test statistic and stores these in an array called `test_stats`. Suppose the observed value of the test statistic is `0.68`.

Which of the folllowing Python expressions returns the p-value for this hypothesis test?

*Select all that apply.*

□ `np.median(test_stats >= 0.68)`

□ `np.average(test_stats >= 0.68)`

□ `np.sum(test_stats <= 0.68)/len(test_stats)`

□ `np.sum(test_stats >= 0.68)/len(test_stats)`

□ `np.count_nonzero(test_stats >= 0.68)/len(test_stats)`

□ There is not enough information to answer because it depends on which test statistic was chosen.

**iv.** **(3.0 pt)** Jimmy uses a $p$-value cutoff of **5%** and determines that the observed test statistic falls very near the center of the distribution of simulated test statisics.

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.

□ There is a 5% chance that the null hypothesis is true.

□ There is a 5% chance that the alternative hypothesis is true.

□ ‘Manual’ statements have higher overcharges than ‘Digital’ statements.

□ There is not enough information to make a conclusion of any kind.

**5. (12.0 points) Adventure**

Troy & Abed plan to randomly select three missions **without replacement** from a list of **14** possible adventures to go on. There are 3 paintball fights, 5 crime investigations, and 6 time travel adventures to choose from.

*Note: Assume that each individual adventure is equally likely to be selected.*

**(a) (3.0 pt)** What is the probability that a paintball fight is selected first, a crime investigation second, and a time travel adventure third?

○ $\frac{3}{14} \times \frac{5}{14} \times \frac{6}{14}$

○ $\frac{3}{14} \times \frac{5}{13} \times \frac{6}{12}$

○ $\frac{3}{14} + \frac{5}{14} + \frac{6}{14}$

○ $\frac{3}{14} + \frac{5}{13} + \frac{6}{12}$

○ None of the above.

**(b) (3.0 pt)** If we know that a paintball fight and a crime investigation have been selected, what is the probability that no time travel adventures are selected?

○ $\frac{6}{12}$

○ $\frac{6}{14}$

○ $\frac{8}{14}$

○ $\frac{8}{14} \times \frac{8}{14} \times \frac{8}{14}$

○ $\frac{3}{14} \times \frac{5}{13} \times \frac{6}{12}$

○ None of the above.

**(c) (3.0 pt)** What is the probability that only paintball fights are selected for the 3 missions?

○ $3 \times \frac{1}{14}$

○ $\frac{1}{14} \times \frac{1}{13} \times \frac{1}{12}$

○ $\frac{3}{14} \times \frac{2}{13} \times \frac{1}{12}$

○ $3 \times \frac{1}{14} \times \frac{1}{13} \times \frac{1}{12}$

○ $1 - \frac{3}{14} \times \frac{2}{13} \times \frac{1}{12}$

○ None of the above.

**(d) (3.0 pt)** What is the probability that at least one paintball fight or crime investigation is selected?

○ $\frac{8}{14}$

○ $\frac{3}{14} + \frac{5}{14}$

○ $\frac{8}{14} \times \frac{7}{13} \times \frac{6}{12}$

○ $1 - \frac{6}{14} \times \frac{5}{13} \times \frac{4}{12}$

○ $1 - (\frac{11}{14})^3 + 1 - (\frac{9}{14})^3$

○ None of the above.

**6. (16.0 points) Verse Jumping**

Evelyn Wang has built a way to travel to parallel universes, each of which contains an alternate variant of our planet Earth (for example, in Earth-241, humans have fingers that look like hotdogs).

There is a small percentage of these parallel Earths (exactly *1%*) in which there is a man named Waymond.

Evelyn wants to find such an Earth, but she doesn't have time to explore every Earth variant.

Suppose Evelyn knows that if an Earth variant has a man named Waymond, there is a 90% chance that it has a woman named Joy. If an Earth variant doesn't have a man named Waymond, there is only a 5% chance that it has a woman named Joy.

**(a) (3.0 pt)** If Evelyn randomly selects an Earth variant to visit, what is the probability that it has a woman named Joy?

○ 0.01

○ 0.05

○ 0.93

○ $0.9 + 0.05$

○ $0.01 \times 0.9$

○ $0.01 \times 0.9 + 0.99 \times 0.05$

○ $0.01 \times 0.1 + 0.99 \times 0.95$

○ None of the above.

**(b) (3.0 pt)** If Evelyn randomly selects an Earth variant to visit, what is the probability that it has a man named Waymond **and** does not have a woman named Joy?

○ 0.1

○ 0.95

○ $0.01 \times 0.1$

○ $0.01 \times 0.9$

○ $0.01 \times 0.95$

○ $0.01 \times 0.1 + 0.99 \times 0.95$

○ $0.5 \times 0.1 + 0.5 \times 0.95$

○ None of the above.

**(c) (3.0 pt)** Suppose Evelyn's random selection leads her to visit an Earth variant that does **not** have a woman named Joy in it.

Given this information, what is the probability that this Earth variant has a man named Waymond?

○ 0.05

○ 0.95

○ $\dfrac{0.01 \times 0.1}{0.01 \times 0.1 + 0.99 \times 0.95}$

○ $\dfrac{0.01 \times 0.95}{0.01 \times 0.95 + 0.99 \times 0.05}$

○ $\dfrac{0.99 \times 0.95}{0.99 \times 0.95 + 0.01 \times 0.1}$

○ None of the above.

**(d) (4.0 pt)** Suppose Evelyn's random selection leads her to visit Earth-353, in which humans drive around in autonomous cars built by a company called Waymo-nd.

The information above makes Evelyn believe there is a 50% chance that Earth-353 has a man named Waymond.

Suppose Evelyn then discovers that Earth-353 has a woman named Joy. What is Evelyn's subjective probability that Earth-353 has a man named Waymond?

○ $0.5 \times 0.9 + 0.5 \times 0.05$

○ $\dfrac{0.01 \times 0.9}{0.01 \times 0.9 + 0.99 \times 0.05}$

○ $\dfrac{0.01 \times 0.9}{0.01 \times 0.9 + 0.99 \times 0.95}$

○ $\dfrac{0.5 \times 0.9}{0.5 \times 0.9 + 0.5 \times 0.05}$

○ $\dfrac{0.5 \times 0.05}{0.5 \times 0.9 + 0.5 \times 0.05}$

○ None of the above.

**(e) (3.0 pt)** Which of the following concepts are relevant to Bayes' Rule?

*Select all that apply.*

□ Anterior

□ Likelihood

□ Posterior

□ Prior

□ Superior

□ None of the above.

**7. (25.0 points) Stonks**

Marty Byrde is an investor who is trying to use the stock market's past performance to make predictions about future prices.

He randomly samples 240 public company stocks from the past year and puts them in a table called `stonks`. The first few rows are shown here:

| Name | Revenue Growth | Old Price | New Price | Rating |
|---|---:|---:|---:|---:|
| Twitter | 120 | 34.42 | 42.69 | Buy |
| Waystar | 50 | 25.38 | 31.23 | Sell |
| Dunder Mifflin | 150 | 12.45 | 11.21 | Hold |
| Hooli | 300 | 73.74 | 82.27 | Hold |

The table contains the following columns:

- *Name*: (string) the name of stock
- *Growth*: (int) the percentage growth in the company's revenue from 2020 to 2021
- *Old Price*: (float) the price of the stock at the end of 2020
- *New Price*: (float) the price of the stock at the end of 2021
- *Rating*: (string) whether Wall St. analysts recommend investors should buy, sell or hold the stock

**(a)** Choose the most appropriate data visualization for each piece of information

**i. (2.0 pt)** How *Revenue Growth* varies across all stocks with a `Buy` rating.

○ Scatterplot

○ Pivot Table

○ Histogram

○ Line Graph

○ Bar Chart

○ None of the Above

**ii. (2.0 pt)** How the relationship between *Revenue Growth* and *Old Price* varies based on *Rating.*

○ Colored Scatterplot

○ Pivot Table

○ Overlaid Histogram

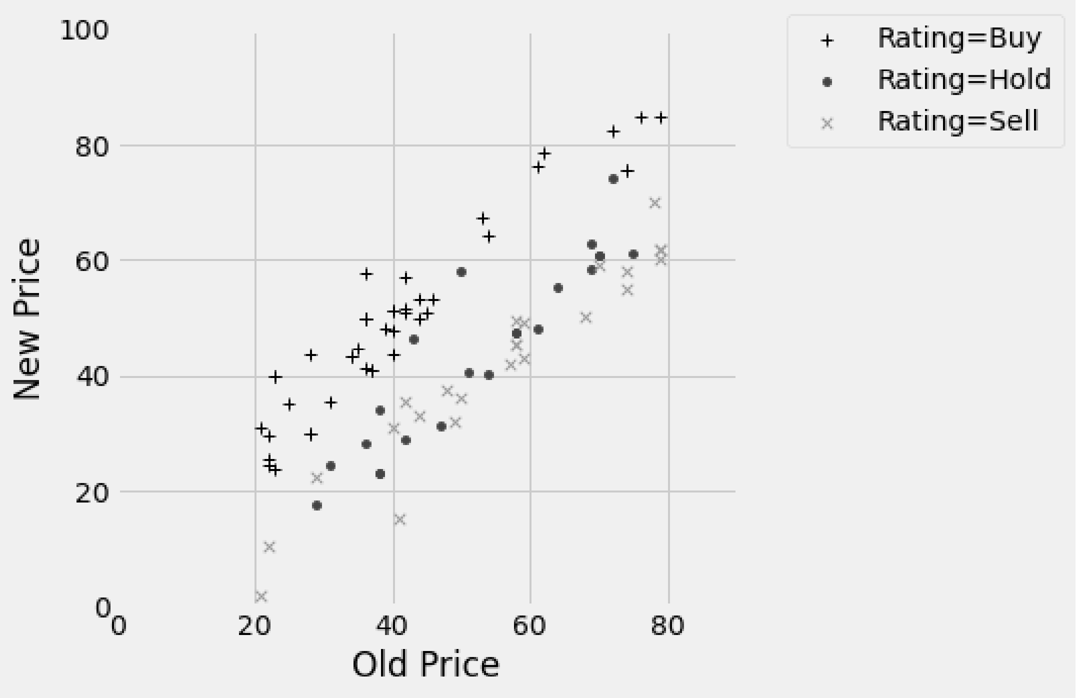
○ Line Graph

○ Bar Chart

○ None of the Above

**(b) (9.0 points)**

Marty creates the following scatterplot of *New Price* against *Old Price*, with color representing the value of *Rating*.



**i. (3.0 pt)** Suppose Marty want to use *Old Price* and *New Price* to create a classifier that can predict the rating of a stock. He writes the following partially completed code:

```
def classify(old_price, new_price):
    if _________:
        return 'Buy'
    elif old_price < 50:
        return 'Hold'
    else:
        return 'Sell'
```

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 `'Sell'` when the true class is `'Buy'`.

*Select all that apply.*

□ `True`

□ `False`

□ `new_price >= 50`

□ `new_price > old_price`

□ `np.random.choice(make_array(0,1)) >= 0`

□ None of the above.

**ii. (2.0 pt)** Suppose Marty instead builds a $k$-nearest-neighbor classifier with $k = 4$ to predict the rating of a stock, using *Old Price* and *New Price* as its features.

Suppose a new stock is proposed and we know the following:

- The stock had an old price of 80
- The stock had a new price of 60

What would this nearest neighbor classifier predict as the *Rating* of this stock?

○ `'Buy'`

○ `'Sell'`

○ `'Hold'`

○ There is not a majority class.

**iii. (2.0 pt)** Suppose a new stock is proposed and we know the following:

- The stock had an old price of 60
- The stock had a new price of 40

What would the above nearest neighbor with $k = 5$ predict as the *Rating* of this stock?

○ `'Buy'`

○ `'Sell'`

○ `'Hold'`

○ There is not a majority class.

**iv. (2.0 pt)** If Marty uses $k = 11$ for the above Nearest Neighbor classifier, there is guaranteed to be a single majority class among the $k$ nearest neighbors' ratings of a stock, regardless of its old and new prices.

*Recall*: The rating of a stock can be buy, sell or hold.

○ True

○ False

○ There is not enough information to tell.

**(c) (12.0 points)**

Suppose that Marty now wants to use $k$-nearest-neighbors to predict the *New Price* of a stock based on its *Revenue Growth* and *Old Price* (i.e., he's now doing a regression instead of classification).

**i. (6.0 points)**

To use the $k$-nearest-neighbors to perform this regression, Marty needs to first find the $k$ nearest neighbors of the stock with respect to *Revenue Growth* and *Old Price.*

He writes a `neighbors()` function, which takes in the following arguments:

- `train`: A three-column table in which the first column is labeled *Revenue Growth*, the second column is labeled *Old Price*, and the third column is labeled *New Price.* Each row of the table represents a stock in the training set.
- `new_stock`: An array of length two containing a stock's revenue growth and old price. For example, `array([120, 50])` corresponds to a stock that grew its revenue 120% and had an old price of 50.
- `k`: The value of $k$ to use for $k$-nearest-neighbors.

The function returns a table containing the $k$ neighbors in `train` that are closest to `new_stock`. It is shown, partially completed, here:

```
def neighbors(train, new_stock, k):
    growth_diffs = ________(a)________  # Note: There is no question about this blank
    old_price_diffs = _________(b)________
    distances = (growth_diffs ** 2 + old_price_diffs ** 2) ** 0.5
    train_dist = train.with_column(‘Distance’, distances)
    return ________(c)________
```

**A. (3.0 pt)** Which of the following lines of code could be used to fill in **blank (b)**?

*Select all that apply.*

□ `train.column(2) - new_match.item(2)`

□ `train.column(1) - new_match.item(1)`

□ `train.column('2020 Revenue') - new_match.item(2)`

□ `train.column('2020 Revenue') - new_match.item(1)`

□ `train.column('2020 Revenue') - new_match.column('2020 Revenue')`

□ None of the above

**B. (3.0 pt)** Which of the following lines of code could be used to fill in **blank (c)**?

*Select all that apply.*

□ `train_dist.sort(‘Distance’).take(np.make_array(0,k,1))`

□ `train_dist.sort(‘Distance’).take(np.arange(k))`

□ `train_dist.sort(‘Distance’).take(np.arange(0,k,1))`

□ `train_dist.sort(‘Distance’, descending=True).take(np.arange(k))`

□ `train_dist.sort(‘Distance’, descending=False).take(np.arange(k))`

□ None of the above

**ii. (6.0 points)**

To generate a prediction for a new stock price, Marty decides to compute the *harmonic mean* of the $k$-nearest-neighbors' *New Price* values. The harmonic mean is calculated by inverting each value, adding up the inverted values, and finally dividing the number of neighbors by the sum.

For example, if $k = 2$ and the 2 nearest neighbors have *New Price* values of 40 and 60, then the *harmonic mean* is:

$$\frac{2}{\frac{1}{40} + \frac{1}{60}} = 48$$

Marty writes a `prediction()` function that takes in the same arguments as the `neighbors()` function (i.e., `train`, `new_stock` and `k`) and returns the *harmonic mean* of the *New Price* values from among the stock's $k$ nearest neighbors. It is shown, partially completed, here:

```
def prediction(train, new_stock, k):
    neighbors_new_price = ________(a)________
    return ________(b)_________
```

**A. (3.0 pt)** Which of the following lines of code could be used to fill in blank (a)?

*Select all that apply.*

□ `neighbors(train, new_stock, k)`

□ `neighbors(train, new_stock, k).item(3)`

□ `neighbors(train, new_stock, k).select(3)`

□ `neighbors(train, new_stock, k).column(3)`

□ `neighbors(train, new_stock, k).column('New Price')`

□ None of the above

**B. (3.0 pt)** Which of the following lines of code could be used to fill in blank (b)?

*Select all that apply.*

□ `3/np.sum(1/neighbors_new_price)`

□ `np.average(neighbors_new_price)`

□ `np.average(1/neighbors_new_price)`

□ `1/np.average(1/neighbors_new_price)`

□ None of the above

**8. (0.0 points) The End**

**(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) (0.0 pt)** Prof. Sahai hasn't seen which of the following shows or movies?

○ Ozark

○ Everything Everywhere All at Once

○ Euphoria

○ Better Call Saul

○ Community

**(c) (0.0 pt)** Draw a picture or meme describing your experience in Data 8!