Data C8, Final Exam

Summer 2023

Name: _____

Email: _____@berkeley.edu

Student ID: _____

Name of the student to your left: _____

Name of the student to your right: _____

Instructions:

Do not open the examination until instructed to do so.

This exam consists of **80 points** spread out over **4 questions** on **14 pages** and must be completed in the **110 minute** time period on August 11, 2023, from 10:10 AM to 12:00 PM unless you have pre-approved accommodations otherwise.

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

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

Honor Code [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: _____

1 Barbenheimer Returns [18 Points]

Rotten Tomatoes, a movie review website, is measuring which of the two movies – Oppenheimer or Barbie – has higher reviews among Berkeley students. They believe that Berkeley students will give higher reviews to the Oppenheimer movie.

Researchers at Rotten Tomatoes randomly sample 1000 Berkeley students and show **each** student **both** movies under identical viewing conditions. Immediately after watching each movie, every student is asked to rate that movie on an integer scale from 1 (worst) up to, and including 10 (best). The reviews are collected in a table named `reviews`; shown below are the first few rows.

movie	review
Oppenheimer	8
Barbie	9
Oppenheimer	6
Barbie	8

... (1996 rows omitted)

(a) [2 Pts] Which of the following is a correct **null** hypothesis that Rotten Tomatoes should use to assess their claim? **Select one**.

- ◯ The Oppenheimer movie has **a different distribution of reviews** than the Barbie movie among the given sample of Berkeley students.
- ◯ The Oppenheimer movie has **the same distribution of reviews** as the Barbie movie among the given sample of Berkeley students.
- ◯ The Oppenheimer movie has **a different distribution of reviews** than the Barbie movie among Berkeley students.
- ◯ **The Oppenheimer movie has the same distribution of reviews as the Barbie movie among Berkeley students.**

(b) [2 Pts] Please state a clear and complete **alternative** hypothesis that Rotten Tomatoes should use to assess their claim.

Solution:

The Oppenheimer movie has higher reviews than the Barbie movie among Berkeley students

(c) [3 Pts] Rotten Tomatoes uses the **difference of means** as their test statistic. Complete the function below so that it returns the difference of mean reviews between the two movies. Larger values of the test statistic should favor the alternative hypothesis.

Note: Assume that the `reviews_table` argument resembles the `reviews` table above.

Hint: The `group` function will return a table that is sorted alphabetically based on the values in the column used for grouping.

```
def test_statistic(reviews_table):
    means_col = _____(A)_____
    return _____(B)_____
```

(i) Fill in the blank `(A)`

Solution: `reviews_table.group(0, np.mean).column(1)`

(ii) Which of the following options is most appropriate for blank `(B)`

◯ `means_col.item(0) - means_col.item(1)`

◯ `means_col.item(1) - means_col.item(0)`

(d) [3 Pts] Which of the following may be used to create simulations under the null hypothesis? **Select all that apply.**

☐ **Shuffle the values of only the `movie` column.**

☐ **Shuffle the values of only the `review` column.**

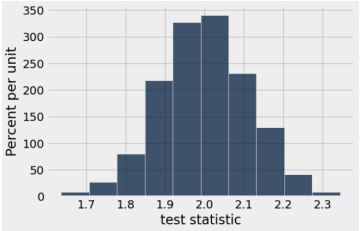
☐ **Shuffle the values of the `movie` column, then shuffle the values of the `review` column.**

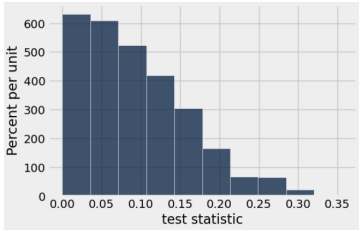
☐ Randomly sample all of the rows of the `reviews` table **with replacement**.

☐ Randomly sample all of the rows of the `reviews` table **without replacement**.

☐ None of the above.

(e) [2 Pts] Suppose we simulate $10,000$ values of the test statistic under the null hypothesis. Which of the following will our distribution of simulated test statistics most closely resemble?





◯ **Graph 1** ◯ Graph 2 ◯ Graph 3

(f) [3 Pts] You obtain a p-value of 0.37 from your experiment above. Which of the following statements are true? **Select all that apply.**

Note: Recall that larger values of your test statistic should favor the alternative hypothesis.

- ☐ **Your observed test statistic lies at the 63^{rd} percentile of the distribution of test statistics simulated under the null hypothesis.**
- ☐ 37% of the test statistics simulated under the null hypothesis were as, or less extreme than the observed test statistic.
- ☐ The Barbie movie has higher reviews than the Oppenheimer movie among Berkeley students.
- ☐ **With a p-value cutoff of 5%, our data are consistent with the null hypothesis.**
- ☐ None of the above.

(g) [3 Pts] Which of the following statements are true? **Select all that apply.**

- ☐ **If Rotten Tomatoes repeats the same experiment, but instead, they sample $10,000$ Berkeley students, the observed test statistic will more accurately reflect whether Oppenheimer is reviewed higher than Barbie among Berkeley students.**
- ☐ **If Rotten Tomatoes repeats the same experiment, but instead, they sample $10,000$ Berkeley students, the distribution of test statistics simulated under the null hypothesis will have a smaller standard deviation.**
- ☐ If Rotten Tomatoes repeats the same experiment, but instead, **they simulate** 1000 **values of the test statistic under the null hypothesis**, the distribution of these simulated test statistics will have a larger standard deviation.
- ☐ None of the above.

2 California Loves Transit [22 Points]

You've just been hired as a data scientist for the City of San Francisco! Your team is interested in studying public transportation, so you begin analyzing data from the widely-used BART train system and the AC Transit bus services during 2022.

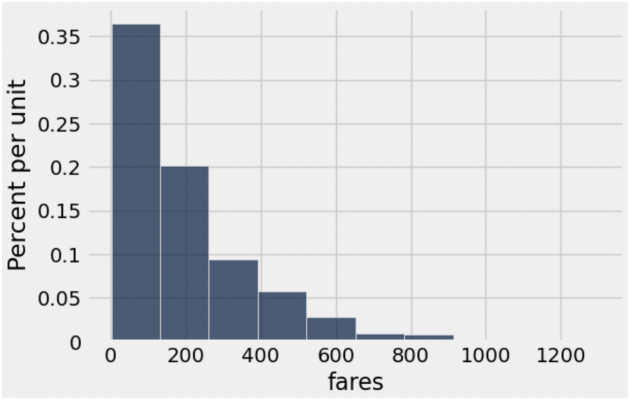
Unfortunately, there is so much data from 2022 that it will overwhelm your computer, so instead, your team gives you a large random sample of 1000 riders in a table called `transport`. Displayed below are the first few rows.

- `id` (**integer**): identification (id) of the rider.
- `transfer` (**boolean**): whether that particular rider transferred between a BART train and an AC Transit bus at least once during 2022.
- `fares` (**float**): total amount that particular rider spent on fares in 2022, measured in dollars.

id	transfer	fares
32849	True	12.5
29490	False	62
81305	False	131.75
70654	False	43

... (996 rows omitted)

(a) [2 Pts] Given below is the distribution of the `fares` column from the `transport` table. Which of the following conclusions can you draw from the plot? **Select all that apply**.



☐ **The distribution of the `fares` column in `transport` is right-skewed.**

☐ The distribution of the `fares` column in `transport` is left-skewed.

☐ **The median of the `fares` column in `transport` is less than the mean.**

☐ The median of the `fares` column in `transport` is **greater than** the mean.

(b) [2 Pts] Which of the following statements must be true? **Select all that apply**.

- ☐ The distribution of **fare spending** among all riders is approximately normal.
- ☐ **The distribution of sample means of fare spending is approximately normal for large random samples of data.**
- ☐ **The distribution of sample sums of fare spending is approximately normal for large random samples of data.**
- ☐ The distribution of **sample medians of fare spending** is approximately normal for large random samples of data.
- ☐ None of the above.

Your team is interested in estimating the proportion of all riders who had transferred between a BART train and an AC Transit bus at least once. You decide to use your sample of 1000 riders to estimate this unknown population parameter.

(c) [4 Pts] Fill in the blanks to generate a visualization of $10,000$ bootstrapped proportions of riders who transferred between a BART train and an AC Transit bus at least once.

```
resample_props = make_array()

for i in np.arange(10000):
    resamp = _____(A)_____
    resamp_prop = _____(B)_____
    _____(C)_____

Table().with_column("resample_props", resample_props).hist()
```

Fill in the blank `(A)`

Solution: `transport.sample()`

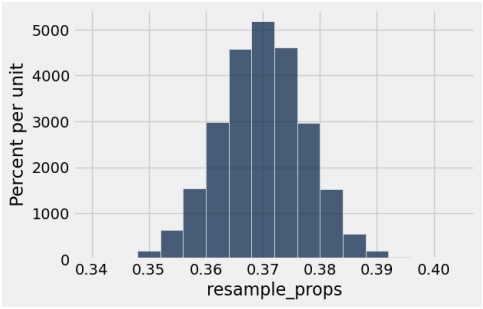
Fill in the blank `(B)`

Solution: `np.mean(resamp.column("transfer"))`

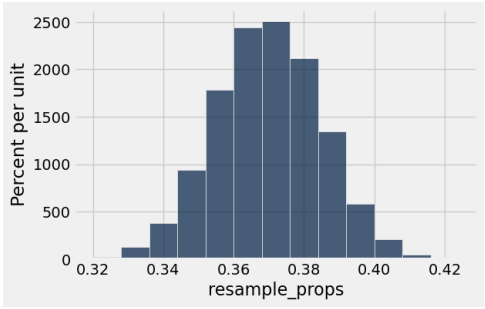
Fill in the blank `(C)`

Solution: `resample_props = np.append(resample_props, resamp_prop)`

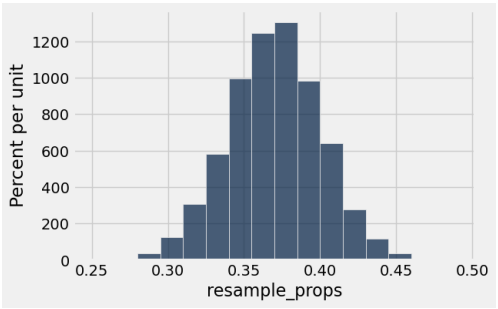
(d) [2 Pts] You find that the mean and standard deviation of your bootstrapped proportions, `resample_props` is 0.37 and 0.015, respectively. Which of the following most closely resembles the distribution of `resample_props`?



◯ Graph 1



◯ **Graph 2**



◯ Graph 3

(e) [3 Pts] Write a mathematical expression that evaluates to the probability that the first row in `transport` is included **at least once** in a single bootstrap re-sample of size 1000. **Please do not simplify.**

Solution: P(the first row is included at least once in one bootstrap sample)

$1 - $ P(the first row is not chosen for any of the bootstrap's 1000 rows)

$1 - $ P(the first row is not chosen)^{1000}

$1 - (\frac{999}{1000})^{1000}$

(f) [2 Pts] Fill in the blanks so that `interval` contains the left and right endpoints of a 95% confidence interval for the proportion of riders in the population who transferred at least once.

Note: You may used variable names defined from previous sub-parts in your code.

```
left = _____(A)_____
right = _____(B)_____
interval = make_array(left, right)
```

Fill in the blank `(A)`

Solution: `percentile(2.5, resample_props)`

Fill in the blank `(B)`

Solution: `percentile(97.5, resample_props)`

(g) [3 Pts] Which of the following conclusions can you draw using your 95% confidence interval **in part (f)**? **Select all that apply**.

- ☐ If someone takes the BART train, there is a 95% chance that they transfer to an AC Transit bus.
- ☐ **If you make confidence intervals from many large random samples from the population, you can expect that roughly 95% of the intervals you create will contain the true population proportion.**
- ☐ There is a 95% chance that the population's true transfer proportion is within the interval generated in **part (f)**.
- ☐ There is a 95% chance that the sample's true transfer proportion is within the interval generated in **part (f)**.
- ☐ None of the above.

(h) [4 Pts] Your team has one last request. They want your 95% confidence interval to be no wider than 5%. Using the maximum standard deviation of a $0-1$ population, what is the smallest sample size that satisfies this requirement? **Express your answer as an integer.**

Solution: Recall that the maximum standard deviation of a $0-1$ population is 0.5.

$0.05 = 4 * \frac{0.5}{\sqrt{samplesize}}$

$\sqrt{samplesize} = 4 * \frac{0.5}{0.05}$

$\sqrt{samplesize} = 4 * 10$

$samplesize = 1600$

3 Breaking Batter: Fried Chicken Edition [23 Points]

Walter and Jesse own a fried chicken restaurant, where they track various details about their food quality. They store this information in a table called `data`; displayed below are the first few rows. Every row corresponds to a distinct order of fried chicken, and the data was collected randomly. Assume that larger values on a $1 - 10$ scale are considered better (and smaller values worse).

- `chicken_quality` (**float**): quality of the raw chicken (**scale:** $[1.0 - 10.0]$)
- `cooking_temp` (**integer**): cooking temperature of the fried chicken, in degrees Fahrenheit
- `seasoning_amount` (**integer**): amount of seasoning in the fried chicken, in grams
- `resting_time` (**float**): resting time of the fried chicken before serving, in minutes
- `customer_score` (**float**): customer satisfaction rating of fried chicken (**scale:** $[1.0-10.0]$)

chicken_quality	cooking_temp	seasoning_amount	resting_time	customer_score
8.5	160	22.5	10.5	9.2
7.7	160	14.175	6.25	7.8
9.6	165	18.25	15	9.9

(a) [2 Pts] Walter calculates a correlation $r = 0.6$ between the two variables `customer_score` and `chicken_quality`. Which of the following conclusions can he draw from this correlation? **Select one.**

- ◯ **Fried chicken made from higher quality chicken generally tends to have higher customer satisfaction scores than fried chicken made from lower quality chicken.**
- ◯ In the `data` table, the `customer_score` values generally deviate less from their average than the `chicken_quality` scores deviate from their average.
- ◯ Fried chicken made from the highest quality chicken also has the highest customer satisfaction score.
- ◯ The use of better quality chicken in the fried chicken recipe causes higher customer satisfaction scores.

(b) [2 Pts] Given the correlation of 0.6 between `chicken_quality` and `customer_score`, mark the following as **True or False**.

(i) The correlation between `chicken_quality` in **standard units** and `customer_score` in **standard units** is 0.6.

◯ **True** ◯ False

(ii) The correlation between `chicken_quality` in **standard units** and `customer_score` in **original units** is 0.6.

◯ **True** ◯ False

Walter wants to predict the `customer_score` from `chicken_quality`. For the following parts, you may assume that:

- The `chicken_quality` column has a mean of 8.4 and a standard deviation of 0.7
- The `customer_score` column has a mean of 8.6 and a standard deviation of 0.5
- The correlation between `chicken_quality` and `customer_score` is 0.6

(c) [4 Pts] What are the **slope** and **intercept** of the regression line in **original units**? You do not need to simplify; you may write your answer as a mathematical expression.

Note: In your expression for the intercept in **part (ii)**, you may use the word "slope" to represent the value of the slope in **part (i)**.

(i) Slope:

Solution: `0.6 * (0.5)/(0.7)`

(ii) Intercept:

Solution: `8.6 - (slope * 8.4)`

(d) [2 Pts] The restaurant receives an exceptional shipment of raw chicken. This shipment has a `chicken_quality` that is 2 standard deviations above the mean. What is the predicted satisfaction score **in standard units** that customers will give the fried chicken made from this new shipment? **Please simplify your answer.**

Solution: `1.2`

```
y_su = r * x_su
y_su = 0.6 * 2 = 1.2
```

(e) [2 Pts] The first order of fried chicken made from the shipment in **part (d)** was cooked poorly, leading to a below average `customer_score`. If Walter adds this order to his `data` table and fits a new regression line on all the orders in `data`, will the slope of the line increase or decrease as compared to the regression line in **part (c)**? **Select one.**

- ◯ Increase
- ◯ **Decrease**
- ◯ Not enough information

(f) [3 Pts] To verify Walter's calculations, Jesse uses an optimization approach to find the least squares line that predicts `customer_score` from `chicken_quality`. Fill in the blanks so that `parameters` evaluates to an array of the slope and intercept of the least squares line that minimizes **root mean squared error**.

```
def rmse(slope, intercept):
    y_predicted = _____(A)_____
    return _____(B)_____

parameters = _____(C)_____
```

Fill in the blank `(A)`

Solution: `slope * data.column("chicken_quality") + intercept`

Fill in the blank `(B)`

Solution:

```
np.mean((data.column("customer_score") - y_predicted)**2)**0.5
```

Fill in the blank `(C)`

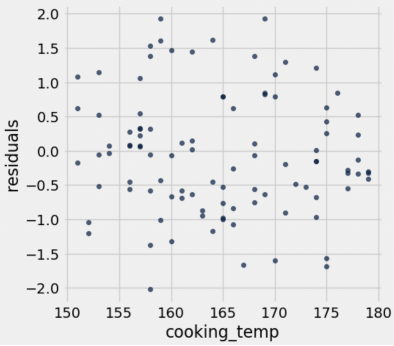
Solution: `minimize(rmse)`

(g) [2 Pts] **[Fill in the Blank]:** The slope and intercept that Jesse finds from his optimization approach will be _____ Walter's slope and intercept values from his regression approach.

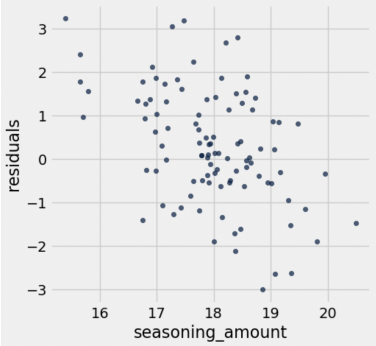
- ◯ greater than
- ◯ less than
- ◯ **equal to**
- ◯ Not enough information

Walter now attempts to predict `customer_score` from each of the other variables in the `data` table: `cooking_temp`, `seasoning_amount`, and `resting_time`.

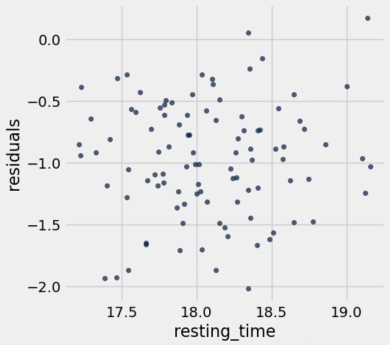
Jesse hands him three scatter plots and claims that these are the residual plots from the regression line that predicts `customer_score` from each of the three variables above.



Plot 1



Plot 2



Plot 3

Do each of the plots indicate that Jesse used the **regression line** to predict `customer_score`? If you answer **No**, explain in **one sentence** how you know that Jesse did not use the regression line. Please do not write anything if you answer **Yes**.

(i) [2 Pts] **Plot 1**: `cooking_temp` vs `customer_score`

◯ **Yes** ◯ No

Solution: N/A

(j) [2 Pts] **Plot 2**: `seasoning_amount` vs `customer_score`

◯ Yes ◯ **No**

Solution: Walter should not see association in the residual plot.

(k) [2 Pts] **Plot 3**: `resting_time` vs `customer_score`

◯ Yes ◯ **No**

Solution: Walter should not see residuals centered at the value of -1 – rather they should be centered at 0.

4 It's Always Meme Friday [16 Points]

As you may know, Kevin likes to share memes before the start of lecture, but he is concerned that students don't appreciate them. He presents 200 randomly selected lecture memes to all Data 8 students in hopes of understanding whether they like each meme or not. He records the data in a table called `meme_data`. Each row represents a meme, and the columns are as follows:

- `category` (**string**): the category of the meme, which is either an "image" or a "video".
- `insta_num` (**integer**): the number of times that meme has been shared on Instagram.
- `time` (**integer**): the duration of the meme, in seconds. Images will have a `time` value of 0.
- `nontext_percentage` (**float**): the percentage of the meme that is non-textual content (**scale:** $[0.0 - 100.0]$).
- `rating` (**float**): the percentage of Data 8 students who liked the meme (**scale:** $[0.0 - 100.0]$).

(a) [4 Pts] Choose which single technique is the most appropriate for answering each scenario. **Select one answer choice for each subpart.**

Note: Please select the **"None of the above"** option if the scenario cannot be answered from the `meme_data` table alone.

(i) Kevin wants to estimate the mean `rating` for all his memes among all Data 8 students.

◯ Linear Regression ◯ A/B Testing ◯ None of the above
◯ **Bootstrapping** ◯ Classification

(ii) Kevin wants to create a model that predicts the `rating` of a meme from the number of times it has been shared on Instagram.

◯ **Linear Regression** ◯ A/B Testing ◯ None of the above
◯ Bootstrapping ◯ Classification

(iii) Kevin wants to use the `time` column to predict what `category` a meme belongs to.

◯ Linear Regression ◯ A/B Testing ◯ None of the above
◯ Bootstrapping ◯ **Classification**

(iv) Kevin wants to use the number of times a given meme has been shared on Instagram to predict whether or not some particular Data 8 student will like the meme.

◯ Linear Regression ◯ A/B Testing ◯ **None of the above**
◯ Bootstrapping ◯ Classification

Kevin is interested in building a classification model that uses the numerical features in the `meme_data` table to predict whether a meme will be "popular" or not. Here, a "popular" meme is one that is liked by more than 50% of the Data 8 students.

(b) [2 Pts] Please complete the code below so that `meme_popular` is a copy of `meme_data` with an additional column called "`popular`". The "`popular`" column should include boolean values that indicate whether a meme is popular (`True`) or not (`False`).

```
pop_arr = _____(A)_____
meme_popular = _____(B)_____
```

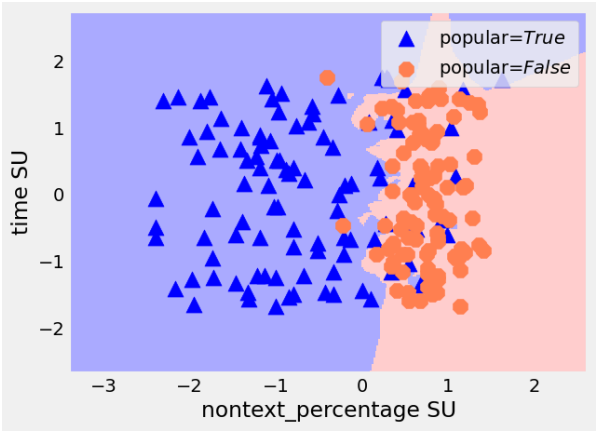
Fill in the blank `(A)`

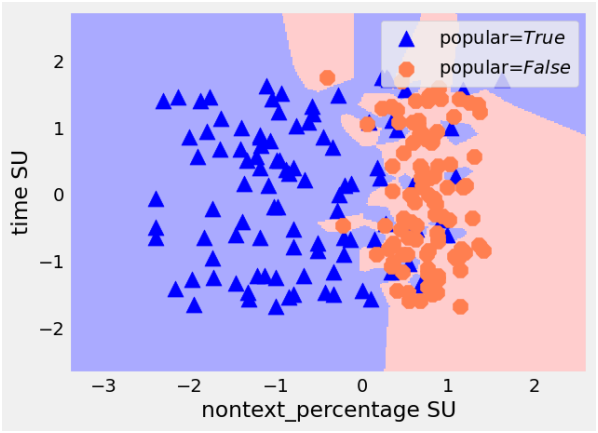
Solution: `meme_data.column('rating') > 50`

Fill in the blank `(B)`

Solution: `meme_data.with_column('popular', pop_arr)`

(c) [2 Pts] Kevin converts the `time` and `nontext_percentage` columns to **standard units** and creates two k-NN classifiers, each with a different value of k: k=3 and k=9. Which of the following plots corresponds to the 3-NN classifier?





◯ **Visualization A** ◯ Visualization B

(d) [4 Pts] Kevin divides his data into a training and testing data set. After training a 1-NN classifier, he notices that only 10% of the memes in the training data are popular, compared to 50% of memes in the testing data. He finds that this imbalance is due to an error in his code.

After correcting the error and re-distributing the data to restore the balance of popular memes, Kevin re-trains a 1-NN classifier. How would you expect the training and testing performance to change after re-balancing the data?

Training Accuracy

- ◯ Increases
- ◯ **Remains the same**
- ◯ Decreases

Testing Accuracy

- ◯ **Increases**
- ◯ Remains the same
- ◯ Decreases

(e) [4 Pts] Before using Kevin's classifier, a GSI guesses whether a meme from the test set is popular among Data 8 students. The GSI is accurate 75% of the time. For memes that the GSI predicts correctly, Kevin's model's accuracy is 82%; otherwise, Kevin's model's accuracy is 45%. Suppose we randomly sample a meme from the test set and Kevin's model predicts its class correctly. What is the probability that the GSI's prediction is right? **Write your answer as a mathematical expression.**

Solution: $\dfrac{0.75 * 0.82}{0.75 * 0.82 + 0.25 * 0.45}$

5 Congratulations [0 Pts]

Congratulations! You have completed the Final Exam.

- **Make sure that you have written your student ID number on *each page* of the exam.** You may lose points on pages where you have not done so.
- Also ensure that you have **signed the Honor Code** on the cover page of the exam for 1 point.

[Optional, 0 pts] Draw a picture (or graph) describing your experience in Data 8.