Fall 2023 Practice Final Exam

Foundations of Data Science

Name

Instructions

- Make sure you have a copy of the Final Exam Reference Guide, and rip off the Table Reference page attached at the end of the exam.
- Select the correct response(s) or provide a written response depending on the question type. If a prompt implies you should write code, then you can provide your own code or use the provided template. Try to provide your responses in the spaces provided. If you find that you need additional space, write your extended response(s) on a blank sheet of paper and number them, so we can connect your response to the question.
- You can assume the following code has been run, when you are writing your responses for Section B:

```
from datascience import *
import numpy as np
import matplotlib+
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use(’fivethirtyeight’)
```

- Once you have made an attempt on the exam, upload a PDF of your attempt to Canvas for a Complete score. Sample solutions will be released after the submission deadline.

This practice exam contains questions from past DATA 8 final exams in addition to a few additional or modified questions.

Section A

1. The median income from one group of US residents (Group 1) is $58,600 while the median income for another group of US residents (Group 2) is $42,500. What can we say for sure based on these statistics? Select all that apply.

   □ Being identified with Group 2 will cause you to earn less than if you were identified with Group 1.

   ✓ **There is a measurable association between being identified with the groups and median incomes in the US.**

   □ There is a statistically significant difference between the median incomes of these two groups.

2. What is a purpose of Bayes' Rule? Select all that are apply.

   ✓ **To quantify the impact of subjective probabilities on our predictions.**

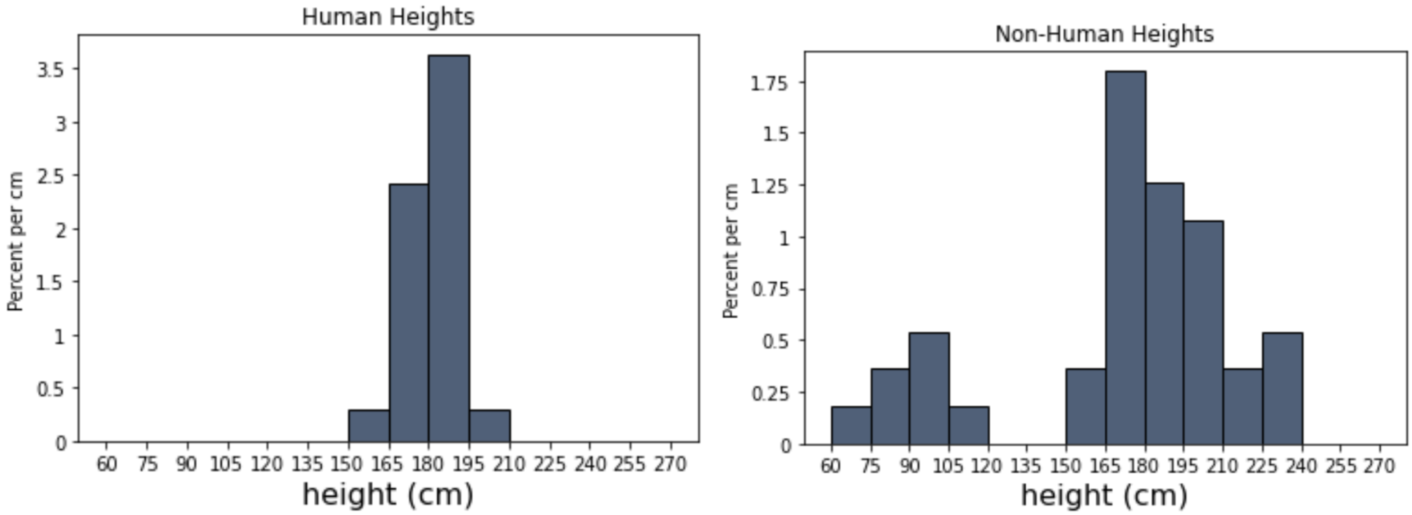
   □ To test whether there is a causal relationship between two variables.

   □ To evaluate the accuracy of a machine learning model on the population.

   □ To determine what percentage of our data lies within a certain number of standard deviations from the mean.

   ✓ **To update our predictions with new information.**

3. The following histograms display the distribution of the heights of humans and non-humans based on the Star Wars character data found in the table `characters`. Respond to the following prompts based on these visuals.



   (a) Based on the plot shown above, between 3% and 4% of humans have a height between 180 cm and 195 cm (not including 195 cm). Select one.

   ○ True

   ✓ **False**

   ○ This is not possible to determine from the graphic.

(b) Based on the plot shown above, there are more humans with a height between 180 cm and 195 cm (not including 195 cm) than non-humans. Select one.

○ True
○ False
√ **This is not possible to determine from the graphic.**

(c) According to this visualization, the standard deviation of non-human heights is larger than the standard deviation of human heights. Select one.

√ **True**
○ False
○ This is not possible to determine from the graphic.

4. 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 the table called `matches` that contains 6 columns. The columns `Opponent` and `Outcome` has string values, the column `Home` has Boolean values, and the rest of the columns have numerical values.

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
⋮	⋮	⋮	⋮	⋮	⋮

(a) 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? Choose one.

○ Scatterplot
○ Pivot Table
○ Total Variation Distance
√ **Overlaid Histograms**
○ Line Graph
○ Bar Chart

(b) 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**

Section B

5. Define a function `count_elem` that takes two arguments: an array `a` and a value `x`. It should return the number of times that `x` appears in `a`.

   For example, `count_elem(make_array('cat', 'cat', 'dog'), 'cat')` should return 2.

```
def count_elem(a, x):
    ______________ np.______________(______________)
```

Sample Solution:

```
def count_elem(a, x):
    return np.count_nonzero(a == x)
```

6. Dunder Mifflin is a paper company with multiple branch locations in the Northeastern United States. Its employees at the Scranton branch are starting their own TikTok channel, but they don't know what content to post.

   They're inspired by trending challenges (i.e., hashtags, expressions, sounds) so they collect data about videos posted by other Dunder Mifflin branches (e.g., Utica, Stamford) as part of a company-wide challenge.

   The data are put into a table called `TIKTOK_TBL` (previewed in the Table Reference Section).

   (a) Write a Python expression that will return the three branches that earned the largest total number of likes.

   Hint: Your answer should return an array.

Sample Solution:

```
branch_likes = TIKTOK_TBL.select("Branch", "Likes")
sum_by_branch = branch_likes.group("Branch", np.sum)
top_3_tbl = (sum_by_branch
             .sort("Likes sum", descending = True)
             .take(np.arange(3)))
top_3_arr = top_3_tbl.column("Branch")
top_3_arr
```

   (b) Write a Python expression that will visualize the branch distribution of videos whose hashtags include `\#TAG"`.

Sample Solution:

```
TAG = TIKTOK_TBL.where("Hashtag", are.containing("#TAG"))
by_branch = TAG.group('Branch')
by_branch.barh('Branch')
```

(c) Rather than focusing on the trending challenges, Dwight's colleague, Jim, thinks a better strategy is to understand which branch has grown the most since its first post.

He defines a branch's *growth* as the difference between the number of views on their first video and the number of views on their latest video.

Suppose Jim writes the following partially completed code, which returns the name of the branch that has experienced the most *growth*:

```
def growth(branch):
    latest_views = ______(1)______
    first_views = ______(2)______
    return latest_views - first_views
branches = TIKTOK_TBL.group("Branch").select("Branch")
growths = ______(3)______
branches.with_column("Growth", growths).______(4)______
```

i. Write a Python expression that should go in blank (1).

Sample Solution:

```
branch = TIKTOK_TBL.where("Branch", branch)
branch.sort("Day", descending = True).column("Views").item(0)
```

ii. Write a Python expression that should go in blank (2).

Sample Solution:

```
branch = TIKTOK_TBL.where("Branch", branch)
branch.sort("Day").column("Views").item(0)
```

iii. Write a Python expression that should go in blank (3).

Sample Solution:

```
branches.apply(growth, "Branch")
```

iv. Write a Python expression that should go in blank (4).

Sample Solution:

```
sort("Growth", descending = True).column("Branch").item(0)
```

Section C

7. 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`; previewed in the Table Reference section.

   (a) 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) Please state a clear and complete alternative hypothesis that Rotten Tomatoes should use to assess their claim.

   Sample Solution: The Oppenheimer movie has higher reviews than the Barbie movie among Berkeley students.

(c) 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)`

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

ii. Which of the following options is most appropriate for blank `(B)`? **Choose one.**

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

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

(d) 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.**

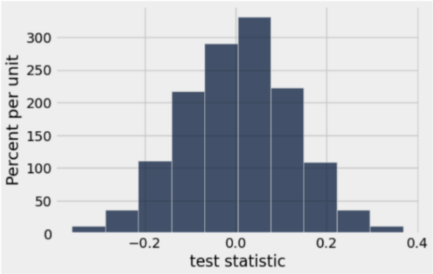
✓ **Shuffle the values of the `'movie'` column, then shuffle the values of the `'review'` column.**

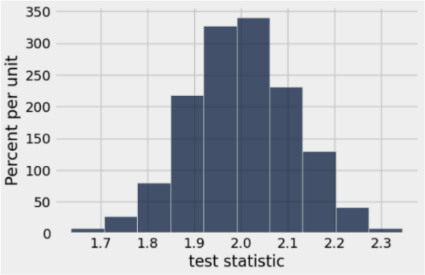
□ Randomly sample all of the rows of the reviews table **with replacement**.

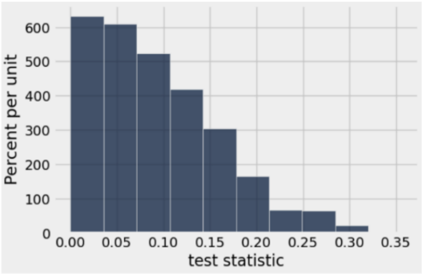
□ Randomly sample all of the rows of the reviews table **without replacement**.

□ None of the above.

(e) 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? **Choose one.**







✓ **The left graph**

○ The middle graph

○ The right graph

(f) 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 63rd 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) 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.

8. Christina is interested in learning more about the duration of songs on Spotify. She collects a random sample of 400 songs listed on the platform and stores the data in the table `songs`, which has one column labelled "Duration". The average song duration in the sample is 185 seconds and the standard deviation is 25 seconds. Christina wants to use this sample of songs to make some estimates about the population of songs and their durations.

   (a) Define a function `song_ci` that constructs a 95% confidence interval for the population mean as follows and returns it as an array. The function takes in the argument `reps`, the number of bootstrap repetitions wanted.

```
def song_ci(reps):
    stats = ________(a)________
    for ______(b)________:
        resample = _______(c)________
        new_mean = _______(d)________
        stats = _________(e)________
    left_end = ________(f)________
    right_end = ________(g)________
    return __________(h)________
```

Sample Solution:

```
def song_ci(reps):
    stats = make_array()
    for i in np.arange(reps):
        resample = songs.sample()
        new_mean = np.mean(resample.column(0))
        stats = stats.append(stats, new_mean)
    left_end = percentile(2.5, stats)
    right_end = percentile(97.5, stats)
    return make_array(left_end, right_end)
```

   (b) Christina creates an interval by using `song_ci(10000)`. To get a more accurate estimate at the same level of confidence, Christina would like to create a new 95% confidence interval that is half as wide as this one. Which one of the following do you think is the best advice for her?

   ○ She should use `song_ci(20000)`.
   ○ She should use a sample of size 800.
   ○ She should use `song_ci(40000)`.
   √ **She should use a sample of size 1600.**

Section D

9. 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` (previewed in the Table Reference section), which contains randomly sampled public information about the various businesses' performance over the last 40 years.

   (a) 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. Suppose Siobhan wants to predict profit from year and decides to fit a regression line. Which of the follow is the regression line for this data? Select one

   ○ `predicted_profit = 1 * year - 2050`
   √ `predicted_profit = 1 * year - 1950`
   ○ `predicted_profit = 1 * year`
   ○ `predicted_profit = 1 * year + 2050`
   ○ `predicted_profit = 1 * year + 2050`
   ○ None of the above

   ii. For Waystar businesses in 2008, what would this regression line predict as the profit? Select one.

   ○ 50 ○ 52 ○ 54 √ **58** ○ 66 ○ None of the above

   iii. Should Siobhan make a prediction using this regression line for the year 2023? They decide to use a policy that the model should not be used for prediction if the input value is more than 3 standard deviations above or below the average for that value. Select one.

   ○ Yes, the year is within 3 standard deviations of the sampled years.
   ○ Yes, the year is not within 3 standard deviations of the sampled years.
   ○ No, the year is within 3 standard deviations of the sampled years.
   √ **No, the year is not within 3 standard deviations of the sampled years.**

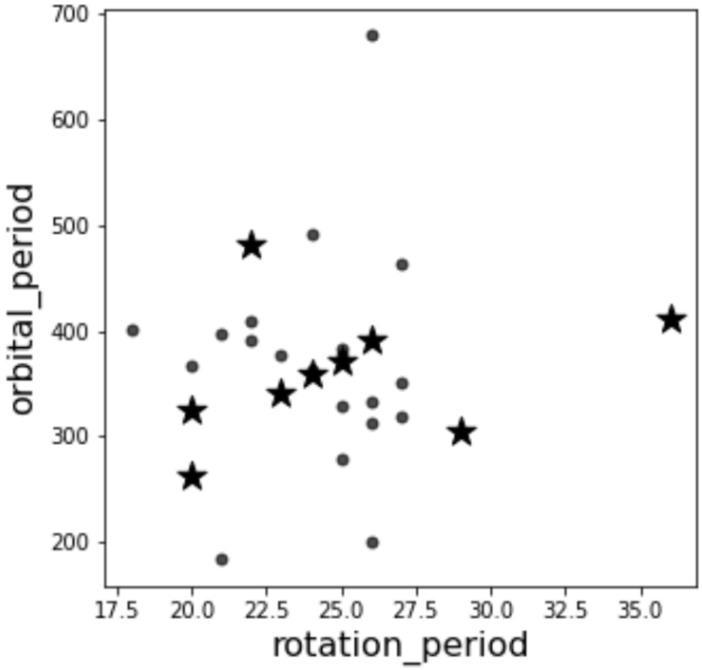
(b) To forecast Waystar's performance, Siobhan wants to understand what the profits for the businesses might be in future years. She first creates a function called `correlation`, which returns the correlation between two numerical arrays. Complete her `predict_profit` function, which takes in a `year_of_interest` (`int`), `years` (`np.array`), `profits` (`np.array`), and returns the predicted profit for that provided year of interest. Note that the arrays years and profits can be assumed to have come from a table like `performance` (same columns and data types).

```
def predict_profit(year_of_interest, years, profits):
        r = correlation(___(a)___)
        slope = r * np.std(profits) / np.std(years)
        intercept = ___(b)___
        predicted_profit = ___(c)___
        return predicted_profit
```

Sample Solution:

```
def predict_profit(year_of_interest, years, profits):
        r = correlation(years, profits)
        slope =  r * np.std(profits) / np.std(years)
        intercept = np.mean(profits) -  slope * np.mean(years)
        predicted_profit = intercept + slope * year_of_interest
        return predicted_profit
```

10. Use a k-NN classifier to predict whether or not a planet in the Star Wars universe contains a temperate climate. The following scatter plot indicates data points associated with planets that contain a temperate climate with a solid dot and data points associated with planets without a temperate climate with a star. Each planet's rotation and orbital periods were used to plot the dots and stars.



(a) The planet Dorin has a rotation period of 22 and an orbital period of 409. The planet Endor has a rotation period of 18 and an orbital period of 402. Based on these values, write an arithmetic expression that Python can evaluate to calculate the distance between these two data points. (Optionally, provide a short explanation to help us consider partial credit in scoring.)

Sample Solution: `np.sqrt((18 - 22) ** 2 + (402 - 409) ** 2)`

(b) What `has_temperate_climate` label (True corresponds with a dot, False corresponds with a star) would a k-NN classifier with `k = 5` assign to a planet with a rotation period of 22.5 and an orbital period of 600?

√ `True`

○ `False`

(c) The tables `training_planets` and `testing_planets` are randomly created from all the available planets that have a column labeled as temperate with values `True` or `False`. The data in the `training_planets` table is visualized in the above scatter plot. All of the test data is shown in the `testing_planets` table.

The k-NN classifier with `k = 5` predicated a `True` label (predicting that they would have a temperate climate on at least part of the planet) for all the planets in the provided test data. What would be the accuracy of the classifier in this case? Express your answer as a fraction or decimal that Python can evaluate. (Optionally, provide a short explanation to help us consider partial credit in scoring.)

Sample Solution: `6 / 10`

(d) Which of the following reflections about this classification process are correct? Select all that apply.

- □ Using a larger value for `k` will guarantee a higher accuracy for the classifier.
- □ Since the orbital period data and the rotational period data are of different magnitudes, then standardizing the data will guarantee a higher accuracy for the classifier.
- √ **None of the above.**

11. A classifier is considered to be overfitting if it performs very well on the training set, but not very well on the test set. **Select one.**

- √ **True**
- ○ False

12. In the year 2031, Rebecca and Sarah are excited to see Frank Ocean perform at a festival. Based on historical data, they know there's a 30% chance he will be on time, a 60% chance he will be late, and a 10% chance he doesn't show up at all.

Thankfully, they have found a website, willfrankbeontime.com, that attempts to predict the outcome of Frank Ocean's arrival. It returns one of the following:

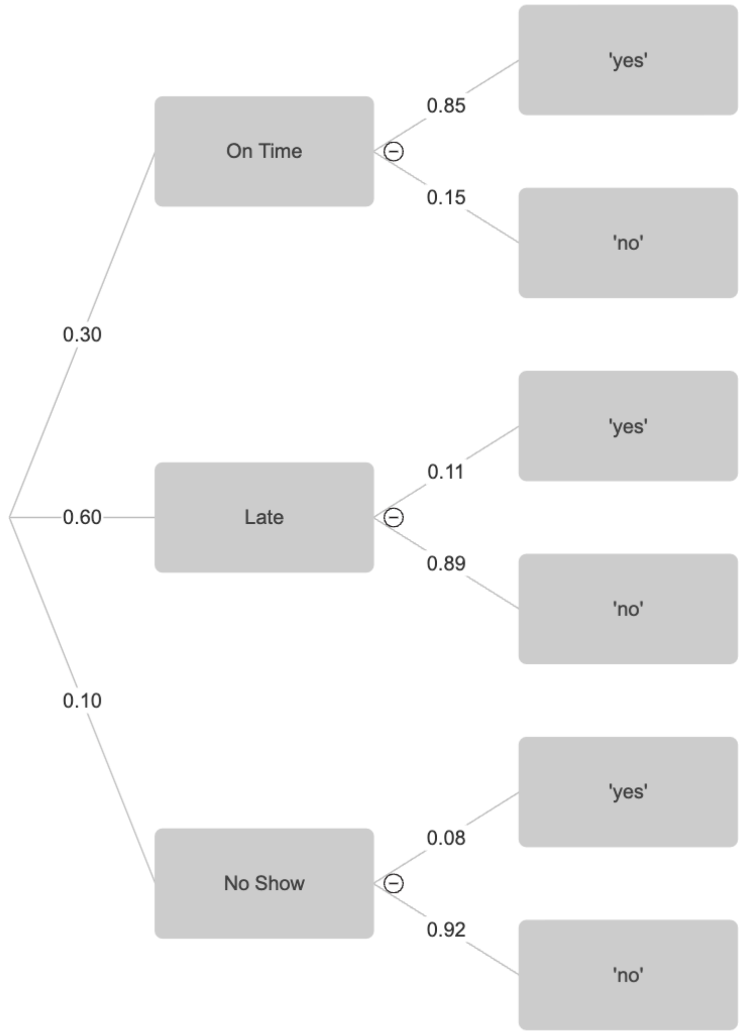
- 'yes', if Frank will show up on time
- 'no', if Frank will show up late or will not show up at all

If he is on time, the website returns 'yes' 85% of the time. If he is late, the website returns 'yes' 11% of the time. If he does not show up, the website returns 'yes' 8% of the time.

Hint: it will help to draw a tree diagram.

(a) Draw a tree diagram for this situation relating Frank Oceans 3 arrival possibilities and the 2 results from the website.

Sample Solution:



(b) What is the probability that Frank will be on time or show up late?

You can leave your answer as a mathematical or python expression.

Sample Solution: $0.3 + 0.6$

(c) What is the probability that Frank is on time and the website returns 'no'?

You can leave your answer as a mathematical or python expression.

Sample Solution: $0.3 * 0.15$

(d) Suppose the website returns "yes". Which of the following represents the probability that Frank was on time, given that the website returned 'yes'? **Select one.**

○ 0.30×0.85

○ $0.30 \times 0.85 \times 0.11 \times 0.08$

✓ $\frac{0.30 \times 0.85}{(0.30 \times 0.85)+(0.60 \times 0.11)+(0.11 \times 0.08)}$

○ 0.89

○ $\frac{0.85 \times 0.11 \times 0.08}{(0.30 \times 0.85)+(0.60 \times 0.11)+(0.11 \times 0.08)}$

(e) Suppose Frank shows up late. What is the probability that the website correctly predicted that he was not on time?

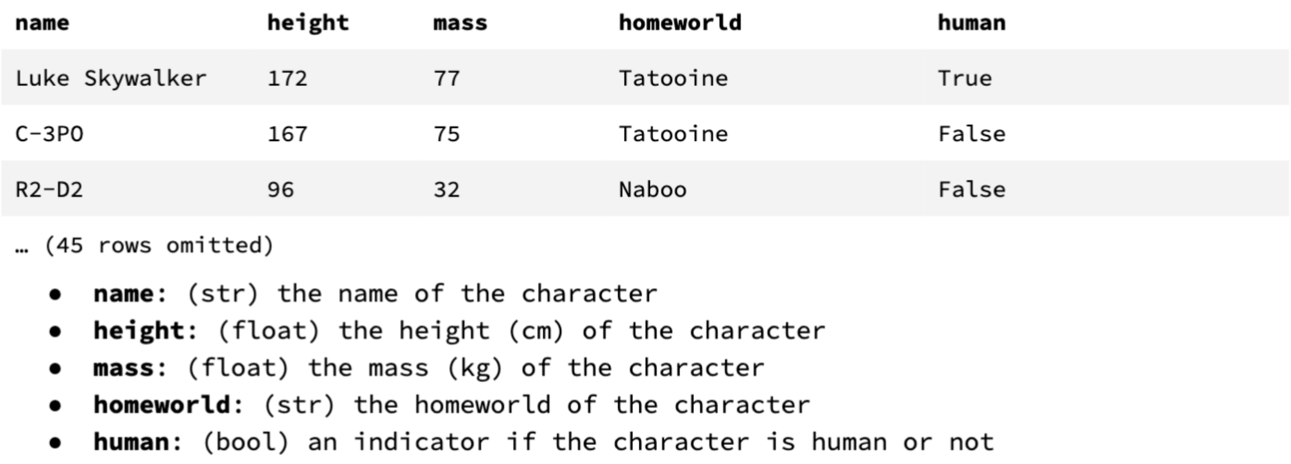
You can leave your answer as a mathematical or python expression.

Sample Solution: 0.89

Table Reference

characters

Here is a preview of the table `characters`:



name	height	mass	homeworld	human
Luke Skywalker	172	77	Tatooine	True
C-3PO	167	75	Tatooine	False
R2-D2	96	32	Naboo	False

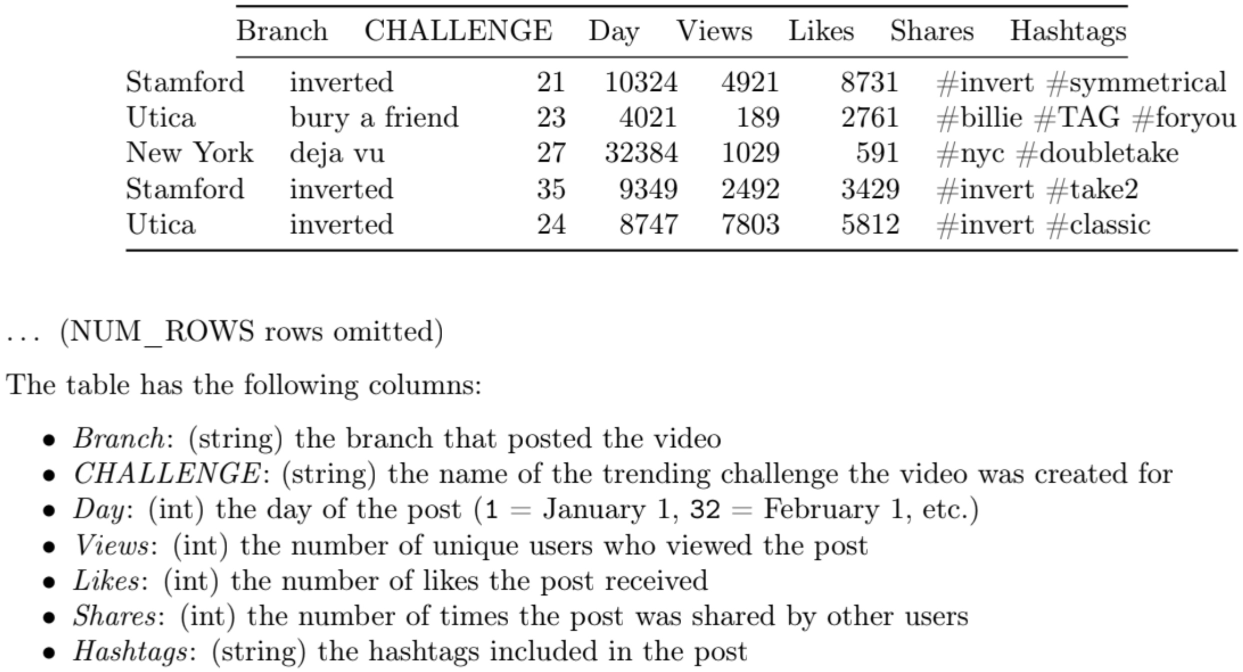
… (45 rows omitted)

- **name:** (str) the name of the character
- **height:** (float) the height (cm) of the character
- **mass:** (float) the mass (kg) of the character
- **homeworld:** (str) the homeworld of the character
- **human:** (bool) an indicator if the character is human or not

Table Reference

TIKTOK_TBL

Here is a preview of the table `TIKTOK_TBL`:



	Branch	CHALLENGE	Day	Views	Likes	Shares	Hashtags
	Stamford	inverted	21	10324	4921	8731	#invert #symmetrical
	Utica	bury a friend	23	4021	189	2761	#billie #TAG #foryou
	New York	deja vu	27	32384	1029	591	#nyc #doubletake
	Stamford	inverted	35	9349	2492	3429	#invert #take2
	Utica	inverted	24	8747	7803	5812	#invert #classic

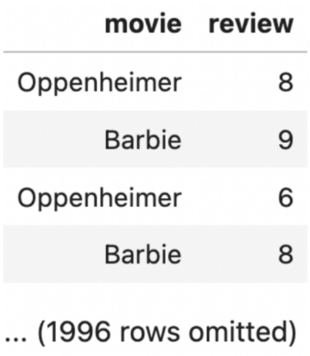
. . . (NUM_ROWS rows omitted)

The table has the following columns:

- *Branch*: (string) the branch that posted the video
- *CHALLENGE*: (string) the name of the trending challenge the video was created for
- *Day*: (int) the day of the post (1 = January 1, 32 = February 1, etc.)
- *Views*: (int) the number of unique users who viewed the post
- *Likes*: (int) the number of likes the post received
- *Shares*: (int) the number of times the post was shared by other users
- *Hashtags*: (string) the hashtags included in the post

reviews

Here is a preview of the table `reviews`:



movie	review
Oppenheimer	8
Barbie	9
Oppenheimer	6
Barbie	8

... (1996 rows omitted)

performance

Here is a preview of the table `performance`:

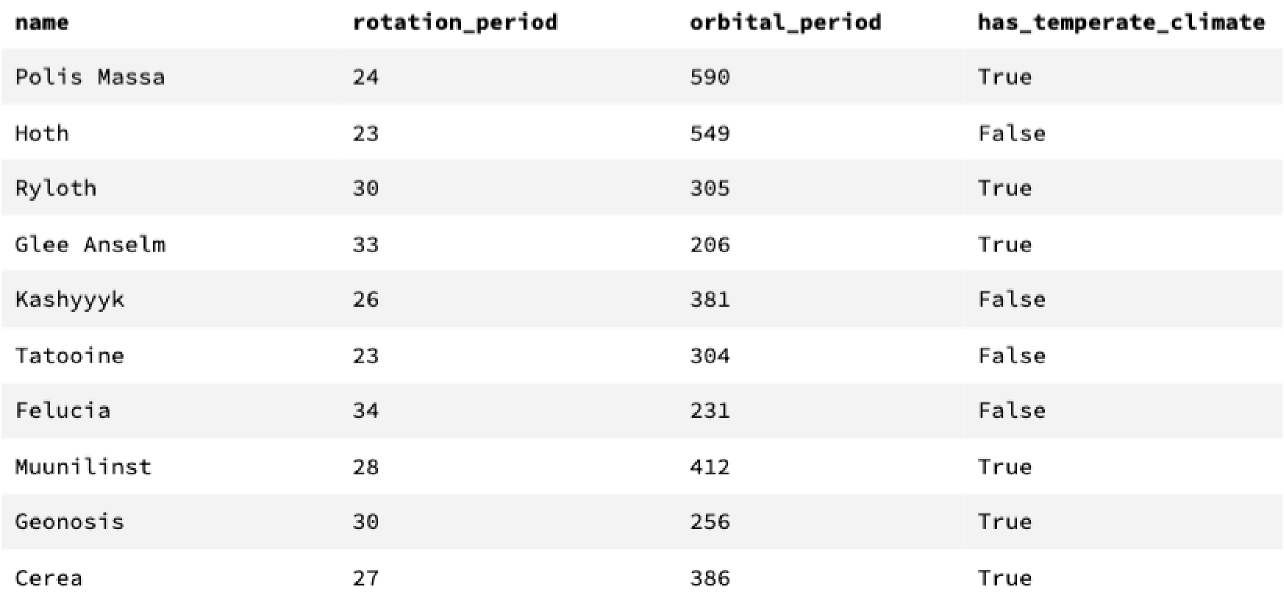
Name	**Year**	**Revenue**	**Profit**	**Sector**	**Advice**
Brightstar Cruises	2019	52.3	18.9	Entertainment	Hold
Adventure Parks	2018	42.8	16.3	Entertainment	Sell
Vaulter	2021	150.9	80.2	Tech	Buy

… (160 rows omitted)

- **Name**: (str) 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**: (str) the business's sector in the industry
- **Advice**: (str) Wall Street's stock purchase advice ('Buy', 'Hold', 'Sell')

testing_planets

Here is a preview of the table `testing_planets`:



name	rotation_period	orbital_period	has_temperate_climate
Polis Massa	24	590	True
Hoth	23	549	False
Ryloth	30	305	True
Glee Anselm	33	206	True
Kashyyyk	26	381	False
Tatooine	23	304	False
Felucia	34	231	False
Muunilinst	28	412	True
Geonosis	30	256	True
Cerea	27	386	True