# Spring 2023 Final Exam

Foundations of Data Science

Name

Total Score: __________ of 100 Points

## Instructions

- Remove the last page of this exam that contains the list of Tables to reference while you complete the exam, and make sure you have a copy of the provided Final Reference Guide.
- Select the correct response(s) or provide a written response depending on the question type. If a prompt asks you to 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 one of the provided blank sheets 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 plots
plots.style.use(’fivethirtyeight’)
```

- The Multiple choice questions (○) and multiple answer questions (□) will be scored like in Canvas.
- The open response questions will be graded as:
  - 3 Points: The response is correct and may contain a very very small error.
  - 2 Points: The response will be correct with a few small edits.
  - 1 Point: A reasonable attempt was demonstrated at providing a response.
  - 0 Points: No reasonable attempt was provided.
- Once you are finished, turn in your exam and you are welcome to leave. Thank you being a part of the class!

# Section A - 14 Points

1. (2 points) 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. (2 points) What is a purpose of Bayes' Rule? Select all that are apply.

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

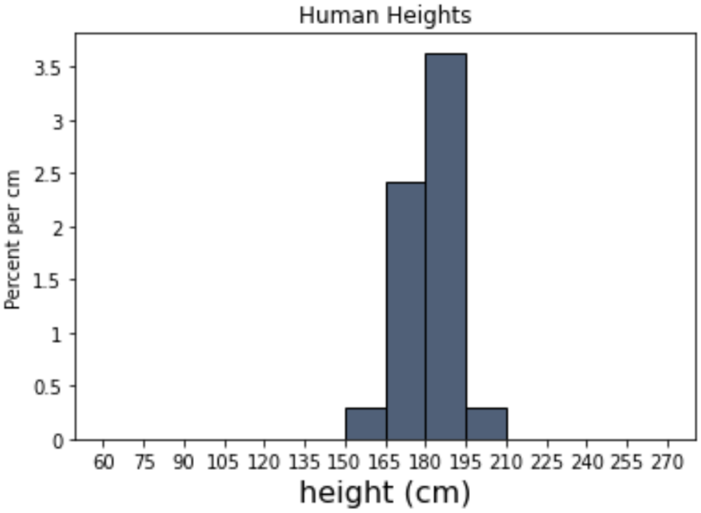
   □ To test whether there is a causal relationship between two variables.

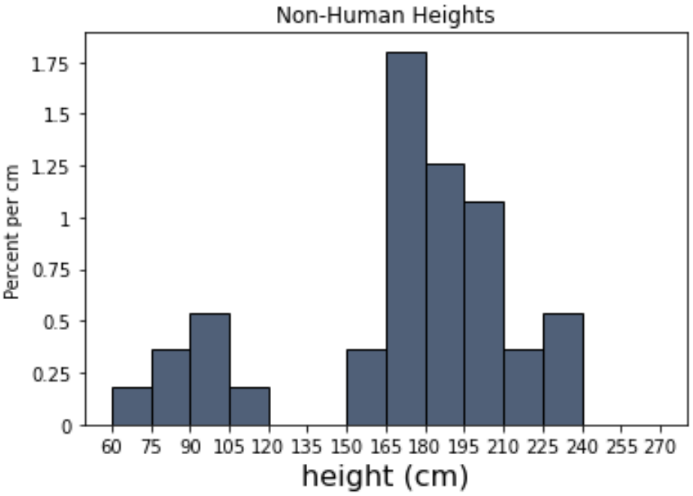
   □ To evaluate the accuracy of a machine learning model on the population.

   □ 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) (2 points) 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) (2 points) 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) (2 points) 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) (2 points) 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) (2 points) 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 - 24 Points

The following tasks focus on the Star Wars information stored in the `characters` and `planets` tables.

5. (3 points) Write code that outputs the name of the planet with the largest orbital period.

```
planets.___(a)___.___(b)___.item(0)
```

**Sample Solution:**

```
planets.sort(’orbital_period’, descending=True).column(’name’).item(0)
```

6. (3 points) Write code that would produce a table showing the average rotation period and the average orbital period for the planets in `planets` that have a temperate climate and those that don’t.

```
planets.___(a)___.___(b)___
```

**Sample Solution:** `planets.drop(‘name’).group(’has_temperate_climate’, np.average)`

7. (3 points) Write code that outputs a histogram of only the orbital periods for the planets that contain a temperate climate.

```
planets.___(a)___.___(b)___
```

**Sample Solution:**

```
planets.where(’has_temperate_climate’, True).hist(’orbital_period’)
```

8. (3 points) Write code that produces a scatterplot showing the relationship between the orbital period and the rotation period for the planets in the `planets` table. The horizontal axis should reflect the rotation period values.

**Sample Solution:** `planets.scatter(’rotation_period’, ’orbital_period’)`

9. (3 points) Complete the following code that produces a horizontal bar chart showing the counts of humans vs non-humans in the `characters` table.

```
characters.___(a)___(___(b)___).___(c)___(___(d)___)
```

**Sample Solution:**

```
characters.group(’human’).barh(’human’)
```

10. (3 points) Complete the following code that produces a table by combining the planets data with the characters data based on the characters’ homeworld values. The first few rows of the `character_homeworlds` table that your code should produce are in the Table Reference.

```
character_homeworlds = ___(a)___.___(b)___(___(c)___, ___(d)___, ___(e)___)
```

**Sample Solution:**

```
character_homeworlds = characters.join(’homeworld’, planets, ’name’)
```

11. (3 points) Write a function called `weight` that outputs the weight (in Newtons) of a character as if they were on Earth based on their mass. The input should be the mass (in kg) of the character as a float. To find that weight, you would multiply the mass (in kg) times the acceleration of gravity on Earth (9.8 Newtons/kg). For example, `weight(84.2)` would produce an output of `825.16`.

**Sample Solution:**

```
def weight(mass):
    return mass * 9.8
```

12. (3 points) Use the `'weight'` function to complete the following code that will output a table that has the same labels as characters, but with an extra column `'weight_on_Earth'`. Assume the `weight` function you defined previously works as intended.

```
weights = characters.___(a)___
characters.with_column(___(b)___, ___(c)___)
```

**Sample Solution:**

```
weights = characters.apply(weight, 'mass')
characters.with_column('weight_on_Earth', weights)
```

# Section C - 33 Points

13. A computer script randomly selects a word from the following array:

    `animal_array = make_array(‘mouse‘, ‘dog‘, ‘dog‘, ‘dog‘, ‘cat‘, ‘cat‘)`

    (a) (2 points) If `np.random.choice(animal_array)` is run, the chance that the resulting string is `’mouse’` is `1 / 6`.

    √ **True**

    ○ False

    (b) (3 points) If `np.random.choice(animal_array)` is run, what is the chance that the resulting string is `’rabbit’`? Write your answer as a valid Python expression such as `(3 / 16) ** 2`. (Optionally, provide a short explanation of your result to help us consider partial credit in scoring.)

    **Sample Solution:** `0`

    (c) (3 points) If `np.random.choice(animal_array, 3)` is run, what is the chance that the resulting strings are all `’cat’`? Write your answer as a valid Python expression such as `(3 / 16) ** 2`. (Optionally, provide a short explanation of your result to help us consider partial credit in scoring.)

    **Sample Solution:** `(2 / 6) ** 3`

    (d) (3 points) If `np.random.choice(animal_array, 3, replace=False)` is run, what is the chance that the resulting strings are all `’dog’`? Write your answer as a valid Python expression such as `(3 / 16) ** 2`. (Optionally, provide a short explanation of your result to help us consider partial credit in scoring.)

    **Sample Solution:** `(3 / 6) * (2 / 5) * (1 / 4)`

14. There is a medical test for a disease that affects 1% of the people in a certain population. The test has high accuracy:

    - For a person who has the disease, the test returns a positive result with a chance of 98%.
    - For a person who does not have the disease, the test returns a negative result with a chance of 99%.

    (a) (2 points) If a randomly selected person from the population, takes the test, and the test result comes back positive, what is the probability that they actually have the disease? Select one.

    ○ 0.01<br>
    ○ 0.98<br>
    ○ 0.99<br>
    ○ 0.01 * 0.98<br>
    ○ 0.01 * 0.99<br>
    ✓ **(0.01 * 0.98) / (0.01 * 0.98 + 0.99 * 0.01)**<br>
    ○ (0.01 * 0.99) / (0.01 * 0.99 + 0.99 * 0.98)

    (b) (2 points) Consider a slightly different situation where a person has symptoms of the disease and their doctor recommends that they take the test. If the test result comes back positive, what is the probability that they actually have the disease? Select one.

    ○ Less than the probability in part a.<br>
    ○ The same as the probability in part a.<br>
    ✓ **Greater than the probability in part (a).**

15. (2 points) According to the Central Limit Theorem, if a sample is large, and drawn at random from the population without replacement, then the probability distribution of the sample average is roughly normal. Select one.

    ○ True<br>
    ✓ **False**

16. (2 points) If a hypothesis test results in a p-value of 0%, then we know for sure that the null hypothesis is false and the alternative hypothesis is true. Select one.

    ○ True<br>
    ✓ **False**

17. (2 points) An A/B Test can be used to determine if two sets of sample data were generated from the same population. Select one.

    ✓ **True**<br>
    ○ False

18. The table `chocolate` contains information on about 1,600 chocolate bars. You may assume that this is a simple random sample from a much larger population of chocolate bars. Belgium and Switzerland are two European countries known for their chocolate. Suppose we want to know which country's chocolate bars are rated higher on average in the population. In our sample, Belgium has a mean rating of 3.09 while Switzerland has a mean rating of 3.34, giving a mean difference of 0.25 in the sample, which we can use as an estimate of the mean difference in the population. We will use the bootstrap method to help us quantify the uncertainty of this estimate. The code below defines a function `ave_rating` that takes in a table `tbl` with the same column labels as chocolate, and a location loc, and computes the mean rating for all chocolate bars whose Location value is equal to loc. For example, `ave_rating(chocolate, ’Belgium’)` should return `3.09`.

```
def ave_rating(tbl, loc):
    loc_tbl = tbl.where(’Location’, are.equal_to(’loc’))
    ratings = loc_tbl.column(’Rating’)
    return np.mean(ratings)
```

(a) (3 points) Write code to generate 5,000 bootstrap samples, compute the mean difference between the ratings of the Swiss and Belgian chocolate bars (that is, Swiss average minus Belgian average) in each bootstrap sample, and store all of the results in the array `boot_diffs`.

```
boot_diffs = make_array()
for ___(a)___:
    boot_table = chocolate.___(b)___
    ave_switz = ___(c)___
    ave_belgium = ___(d)___
    diff = ___(e)___
    boot_diffs = np.append(boot_diffs, ___(f)___)
```

**Sample Solution:**

```
boot_diffs = make_array()
for __ in np.arange(5_000):
    boot_table = chocolate.sample()
    ave_switz = ave_rating(boot_table, ’Switzerland’)
    ave_belgium = ave_rating(boot_table, ’Belgium’)
    diff = ave_switz - ave_belgium
    boot_diffs = np.append(boot_diffs, diff)
```

(b) (3 points) Write code that uses `boot_diffs` to compute an approximate 95% confidence interval for the population mean difference between the ratings of Swiss and Belgian chocolate bars. After the code is executed, left should store the left endpoint of the interval and right should store the right endpoint. You may assume that `boot_diffs` has been computed correctly.

```
left = ___(a)___
right = ___(b)___
```

**Sample Solution:**

```
left = percentile(2.5, boot_diffs)
right = percentile(97.5, boot_diffs)
```

(c) (2 points) Suppose that we are testing whether or not Belgian and Swiss chocolate bars have the same mean ratings in the population. Also, suppose that the numerical values of `left` and `right` turned out to be -0.03 and 0.56, respectively, so the confidence interval developed in part (b) is (-0.03, 0.56). In this particular situation, which of the following can we conclude from this hypothesis test? Select one.

○ Belgian and Swiss chocolate bars have the same mean ratings.

√ **Belgian and Swiss chocolate bars could have the same mean ratings.**

○ Belgian and Swiss chocolate bars have different mean ratings.

(d) (2 points) Which of the following are true based on the confidence interval above? Select all that apply.

□ If we repeat this process many times under the same conditions, we can expect that roughly 95% of the intervals that are created will contain the true population mean rating for Belgian chocolate bars.

□ If we randomly sample 1,000 chocolate bars without replacement, he can expect roughly 95% of the mean differences of chocolate bar ratings for the two countries to be between -0.03 and 0.56.

□ 95% of mean differences of chocolate bar ratings for the two countries in the population are between -0.03 and 0.56.

√ **None of the above.**

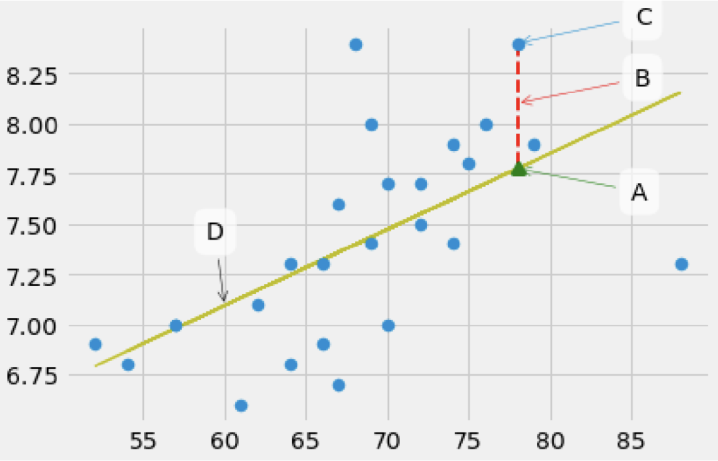
(e) (2 points) There is a 95% chance that this hypothesis test will incorrectly reject the null hypothesis.

○ True

√ **False**

# Section D - 29 Points

19. The following scatter plot shows the relationship between two numerical variables with the least-squares regression line drawn through the Points. Labels A (the triangle on the line), B (the dashed line), C (a dot), and D (the solid diagonal line) indicate various parts of the visualization. Match each term to its label/letter on the graph or "Not Pictured". Some letters may be used more than once or not at all, but there should be one correct response for each of the following parts.



(a) (2 points) A predicted value

√ **A** ○ B ○ C ○ D ○ Not Pictured

(b) (2 points) A residual

○ A √ **B** ○ C ○ D ○ Not Pictured

(c) (2 points) The RMSE

○ A ○ B ○ C ○ D √ **Not Pictured**

(d) (2 points) An observed value

○ A ○ B √ **C** ○ D ○ Not Pictured

20. 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.

(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. (2 points) 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. (2 points) 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. (2 points) 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.**

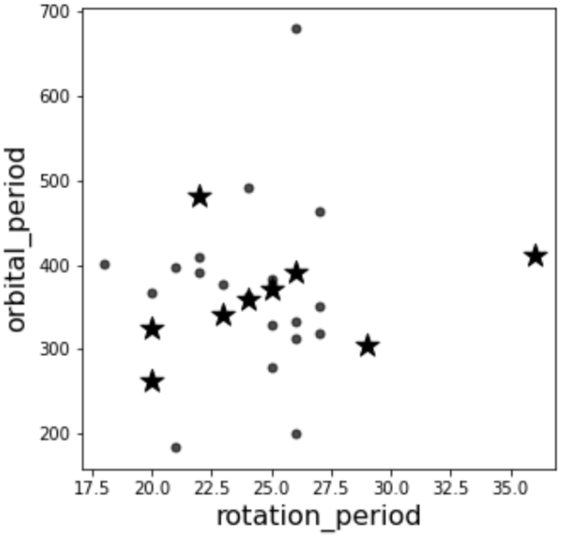
(b) (3 points) To forecast Waystar's performance, Siobhan wants to understand what the profits for the businesses might be in future years. She first creates a function called `correlation`, which returns the correlation between two numerical arrays. Complete her `predict_profit` function, which takes 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.

```
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
```

21. 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) (3 points) 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) (2 points) 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) (3 points) 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) (2 points) 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.**

22. (2 points) 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