Fall 2023 Midterm Exam

Foundations of Data Science

Name

Total Score: _____ of 100 Points

Instructions

- Make sure to rip off the Table Reference and Midterm Reference Guide 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 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 plt
plt.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:
  - Full Points: The response is correct and may contain a very very small error.
  - Partial Points: A reasonable response was provided. The partial point value will depend on your response.
  - No Points: No reasonable attempt was provided.
- Once you are finished, turn in your exam and you are welcome to leave.

Section A - 30 Points

1. (2 points) Suppose we have discovered an association between two variables in a dataset. Which of the following would be the best way to test whether it is causal? Choose one.

   ◯ Brainstorm some potential confounding factors and test whether any of them has an association with both variables.

   √ **Run a randomized controlled experiment.**

   ◯ If both variables are numerical, use a scatter plot to check for a trend.

2. (4 points) Which of the following must be true, for an experiment to count as a randomized controlled experiment? Select all that apply.

   √ **There is a control group.**

   ☐ The experimenters control who is selected to participate in the experiment.

   ☐ Each participant is informed whether they are in the treatment group or not.

   ☐ The distribution of ages of the participants in the experiment are representative of the distribution of ages in the population at large.

   √ **Randomness is used to determine whether each participant will be part of the control group(s) or treatment group(s).**

3. (4 points) Suppose you are curious about the financial situations of recent Berkeley graduates. You have data on 200 recent graduates. Included in the data set are the starting salaries for each of the graduates (`'Income($1K)'`) and their unpaid student debt (`'Debt($1K)'`). In order to understand how a graduate's debt might be associated with their salary, you make the following scatterplot:

   Which of the following are valid conclusions that can be drawn from this graph above? Choose all that apply.

   ☐ There is a positive association between student debt and salary.

   √ **There is a negative association between student debt and salary.**

   ☐ There is no association between student debt and salary.

   ☐ There are no Berkeley graduates with a debt greater than $100K.

   √ **Among the graduates surveyed, 3 of them have debt greater than $80K.**

   ☐ Among the graduates surveyed, higher debt caused them to have lower starting salaries.

4. A real estate company has a dataset of all their buildings, with three attributes for each building: its size (in square feet), its type (residential or commercial), and its estimated value (sale price) if sold (in dollars).

   (a) (3 points) Select all that are correct:

   √ **The size attribute is a numerical variable.**

   ☐ The type attribute is a numerical variable.

   √ **The value attribute is a numerical variable.**

   (b) (2 points) The standard visualization to understand the distribution of building types is: (choose one)

   √ **A bar chart**

   ◯ A line plot

   ◯ A scatter plot

   ◯ A histogram

   ◯ Two histograms, overlaid

   (c) (2 points) The standard visualization to understand the distribution of building sizes is: (choose one)

   ◯ A bar chart

   ◯ A line plot

   ◯ A scatter plot

   √ **A histogram**

   ◯ Two histograms, overlaid

   (d) (2 points) The standard visualization to check for an association between building size and building value is: (choose one)

   ◯ A bar chart

   ◯ A line plot

   √ **A scatter plot**

   ◯ A histogram

   ◯ Two histograms, overlaid

   (e) (2 points) The standard visualization to check for an association between building size and building type is: (choose one)

   ◯ A bar chart

   ◯ A line plot

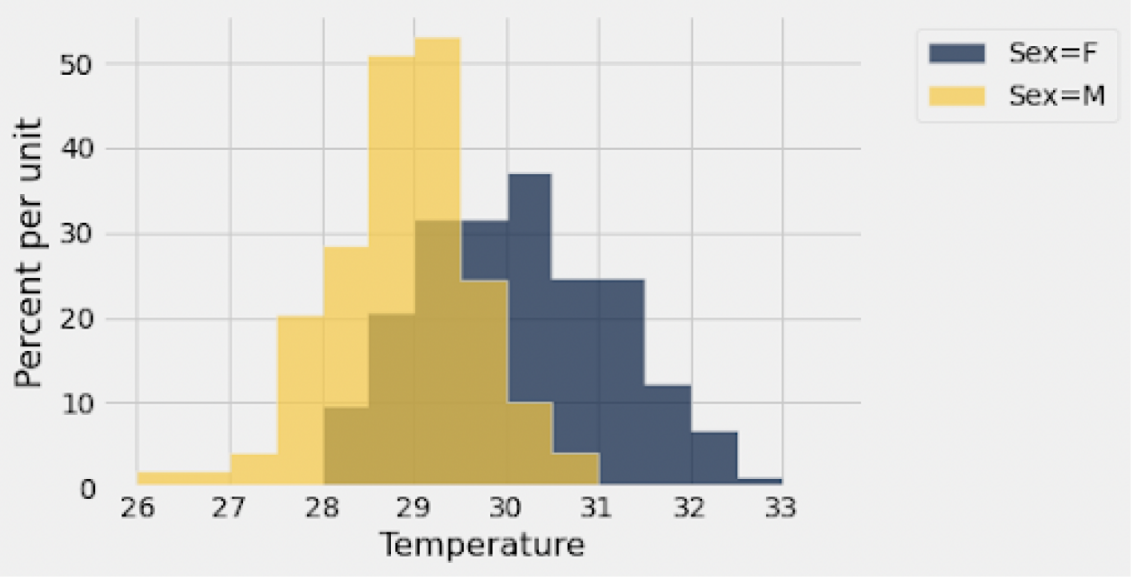
   ◯ A scatter plot

   ◯ A histogram

   √ **Two histograms, overlaid**

5. When hatching a baby turtle from an egg, we incubate the egg at some temperature. A researcher read that the temperature an egg is incubated at influences whether or not the turtle that hatches will be male or female. They randomly sample turtle eggs, and record the incubation temperature (in Celsius) and the sex of the turtle that hatches. The following histogram shows the distribution of temperatures based on the sex of the turtle.

You can assume that 100% of the data is captured in this visualization.



(a) (3 points) In the sample, more than 50% of the male turtles were incubated at a temperature between 29.5 and 30.0 degrees.

◯ True

√ **False**

◯ This is not possible to determine based on the provided information.

(b) (3 points) In this sample, the number of male turtles with incubation temperatures between 29.5 and 30 degrees is the same as the number of female turtles incubated between 30.5 and 31 degrees.

◯ True

◯ False

√ **This is not possible to determine based on the provided information.**

(c) (3 points) If the bins used to form the histogram for female turtles were replaced with a single bin from 28 to 33, how tall would the resulting bar be? Make sure to include the units in your answer.

Sample Solution: The width of the bin would be 33 - 28 = 5 degrees. The bin would create 100% of the female turtle data. Together, this means that the height of the resulting bar would be 100% / 5 degrees = 20 percent per degree.

Section B - 43 Points

For this section, your goal is to provide Python code that could be run in our notebooks that will produce the answer to the questions asked. In most cases, we have provided a template to get you thinking. You can alternatively ignore the template and write your own code from scratch.

6. In San Francisco, the Existing Buildings Energy Performance Ordinance (Environment Code Chapter 20) requires that each non-residential building with at least 10,000 square feet of conditioned (heated or cooled) space and each residential building with at least 50,000 square feet of conditioned space must be benchmarked using Energy Star Portfolio Manager annually. Each non-residential building specified above is also required to undergo an energy audit or retrocommissioning at least once every 5 years.

   The table `building_data` contains relevant San Francisco building information and 2021 energy use (measured in thousands of BTUs (British thermal units)). On the Table Reference page, you can see a preview of this table.

   (a) (4 points) How many `'Commercial'` buildings are there in `building_data`.

   ```
   commercial_buildings = _____(a)_____.____(b)______(_____(c)_____, ____(d)______)
   commercial_buildings._____(e)_____
   ```

   Sample Solution:

   ```
   commercial_buildings = building_data.where('property_type', 'Commercial')
   commercial_buildings.num_rows
   ```

(b) (4 points) What is the address for the building with the largest floor area? You can assume there is a unique building with the largest floor area.

```
sorted_data = ____(a)______.____(b)______(_____(c)_____, _____(d)_____)
____(e)______._____(f)_____(____(g)______)._____(h)_____
```

Sample Solution:

```
sorted_data = building_data.sort('floor_area', True)
sorted_data.column('building_address').item(0)
```

(c) (3 points) You've received a CSV file called `zip_code.csv`. Write code that will create a table called `zip_codes` from that CSV file that contains all the information in the `zip_code.csv` file. On the Table Reference page, you can see a preview of what `zip_codes` looks like. Zip codes and postal codes are equivalent in this context.

Sample Solution:

```
zip_codes = Table.read_table('zip_codes.csv')
```

(d) (3 points) Use the join method to create a table called `building_data_geo` that adds the latitude, longitude, and population estimate information from `zip_codes` to the data in `building_data`. You do not need to do any additional sorting or re-ordering beyond using the join method. On the Table Reference page, you can see a preview of what `building_data_geo` should look like.

Sample Solution:

```
building_data_geo = building_data.join('postal_code', zip_codes, 'zip')
```

(e) (4 points) When reading the data, it seems that Python assumed the postal code (zip code) values were numerical. Write code that will check if the data type of the values in the `postal_code` column of `building_data_geo` is float. Your code should output the `bool` value `True` or `False`. As a hint, `type(2.0)` would evaluate to be float.

Sample Solution:

```
type(building_data_geo.column('postal_code').item(0)) == float
```

(f) (4 points) The postal codes in `building_data_geo` are actually float values, but they need to be strings. Create a function called `float_to_str` that takes a float and returns a string version of the float ignoring any decimal part.

For example, `float_to_str(94118.0)` should return `'94118'`.

Hints: `str(94118.0)` would create the string `'94118.0'`, not `'94118'`.

Sample Solution:

```
def float_to_str(a_float):
    return str(int(a_float))
```

(g) (3 points) Use the `float_to_str` function to create an array called `postal_codes` of the postal codes formatted as strings.

Sample Solution:

```
postal_codes = building_data_geo.apply(float_to_str, 'postal_code')
```

(h) (3 points) Update the `building_data_geo` table such that the values in the `'postal_code'` column are strings, not floats.

Hint: Remember that `postal_codes` is an array of the postal codes as strings.

Sample Solution:

```
building_data_geo = building_data_geo.with_column('postal_code', postal_codes)
```

(i) (4 points) Create a bar chart of the distribution of the postal codes in the `building_data_geo` table. Make sure the bars are in order such that the longest bars are at the top of the visualization.

```
by_zip =_____(a)_____._____(b)_____(_____(c)_____)
by_zip_sorted = _____(d)_____._____(e)_____(_____(f)_____, _____(g)_____)
_____(h)_____
```

Sample Solution:

```
by_zip = building_data_geo.group('postal_code')
by_zip_sorted = by_zip.sort('count', True)
by_zip_sorted.barh('postal_code')
```

(j) (4 points) Create a table with two columns showing the median energy use for 2021 for each postal code based on the data in `building_data_geo`. Your table should have a row for each postal code showing the median energy use for the buildings with that postal code.

```
reduced_data =  _____(a)_____._____(b)_____(_____(c)_____, _____(d)_____)
_____(e)_____._____(f)_____(_____(g)_____, _____(h)_____)
```

Sample Solution:

```
reduced_data = building_data_geo.select('postal_code', 'energy_use_2021')
reduced_data.group('postal_code', np.median)
```

(k) (3 points) Using the data in `building_data_geo`, create a visualization to show the relationship between the floor area of a building and its energy usage in 2021.

Sample Solution:

```
building_data_geo.scatter('floor_area', 'energy_use_2021')
```

7. (4 points) Which of the following functions correctly returns the number of occurrences of a specific value in a given array? For example, `count_arr_occurences(make_array(0,1,0,5,1), 1)` should evaluate to 2 and `count_arr_occurences(make_array("a", "b", "c"), "c")` should evaluate to 1. Select all that apply.

☐
```
def count_arr_occurences(arr, value):
    count = 0
    for i in np.arange(value):
        if arr.item(i) == value:
            count = count + 1
    return count
```

√
```
def count_arr_occurences(arr, value):
    count = 0
    for x in arr:
        if x == value:
            count = count + 1
    return count
```

☐
```
def count_arr_occurences(arr, value):
    return arr == value
```

√
```
def count_arr_occurences(arr, value):
    return np.sum(arr == value)
```

Section C - 27 Points

8. In a game called September, players take turns selecting tokens and making moves based on the selected tokens. DDuring each player's turn, they randomly select two tokens from a container, make a play based on the two tokens, and then put all the tokens back in the container for the next player. The distribution of tokens is:

   - Earth Token: 21 Tokens
   - Wind Token: 12 Tokens
   - Fire Token: 1 Token

   (a) (3 points) What is the probability that a player will select no Wind tokens when it is their turn?

   Sample Solution: `(22 / 34) * (21 / 33)`

   (b) (3 points) What is the probability that a player will select 2 Fire tokens when it is their turn?

   Sample Solution: `0`

   (c) (3 points) What is the probability that a player will select at least one Wind token when it is their turn?

   Sample Solution: `1 - (22 / 34) * (21 / 33)`

9. According to a recent survey, 28% of surveyed adults in the United States use LinkedIn. For the sake of this question, assume that the chance of a randomly sampled adult in the United States being a LinkedIn user is 28% (independently of all others).

   (a) (2 points) For which sample size below is there a higher chance that a random sample of that size will contain a percent of LinkedIn users of more than 50%?

   √ **20**
   ◯ 1,000

   (b) (3 points) According to the Law of Large Numbers (Law of Averages), with a smaller sample size the percentage of surveyed adults in that sample that use LinkedIn is more likely to be closer to 28% than a larger sample size.

   ◯ True
   √ **False**

10. In the game of Wordle, a player guesses up to 6 words until they correctly guess the secret word of the day or run out of guesses. Their guess count is either the guess number that was correct, 1 through 6, or X if all 6 guesses were incorrect. For all 1,000 students who played Wordle yesterday, we have collected the proportion of students with each guess count. These proportions appear in the table below and an array called `students`.

1	2	3	4	5	6	X
0.0	0.17	0.33	0.27	0.20	0.02	0.01

```
students = make_array(0.0, 0.17, 0.33, 0.27, 0.20, 0.02, 0.01)
```

Wordle's creator, Josh Wardle, sent us the proportion of guess counts for all players who tried to guess yesterday's word in an array called `everyone`.

1	2	3	4	5	6	X
0.0	0.09	0.25	0.32	0.28	0.03	0.03

```
everyone = make_array(0.0, 0.09, 0.25, 0.32, 0.28, 0.03, 0.03)
```

(a) (2 points) What best describes the table for the students? Choose one.

◯ Probability Distribution

√ **Empirical Distribution**

(b) (3 points) What is one way to simulate randomly selecting 1,000 individuals from the population of individuals that played Wordle yesterday? Choose one.

◯ `sample_proportions(1000, students)`

√ `sample_proportions(1000, everyone)`

◯ `sample_proportions(1000, make_array('1', '2', '3', '4', '5', '6', 'X'))`

◯ `sample_proportions(1000, make_array(1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7))`

(c) (3 points) If we assume the distribution provided by Josh Wardle is similar for tomorrow, what is the chance that a randomly selected Wordle player will guess the word in less than 4 guesses?

Sample Solution: `0.00 + 0.09 + 0.25`

11. (5 points) Create a function called `roll` with arguments `k`, `n`, and `trials` that simulates trials (the number of trials) rolls of n fair 6-sided dice, and each time counts how many of those dice show `k` or higher, and then displays an empirical histogram of those counts.

    For example, if `k` is 5, `n` is 3, and rolling 3 dice results in a 6, a 4, and a 5, then 2 of the 3 dice are 5 or larger (the 6 and the 5). So, `roll(5, 3, 10_000)` would output a histogram created by repeating simulation 10,000 times.

```
def ___(a)____(__(b)__, __(c)__, __(d)__):
    """Repeatedly roll n dice and check how many results are k or larger."""

    outcomes = make_array()
    possible_results = np.arange(1, 7)

    for _____________(e)_____________
        rolls = _____________(f)_____________
        outcomes = _______(g)________(outcomes, np.count_nonzero(rolls >= ___(h)__))

    Table().with_column('Outcomes', ______(i)______).____(j)____(bins=np.arange(30))
```

Sample Solution:

```
def roll(k, n, trials):
    """Repeatedly roll n dice and check how many results are k or larger."""

    outcomes = make_array()
    possible_results = np.arange(1, 7)

    for i in np.arange(trials):
        rolls = np.random.choice(possible_results, n)
        outcomes = np.append(outcomes, np.count_nonzero(rolls >= k))

    Table().with_column('Outcomes', outcomes).hist(bins=np.arange(30))
```

Table Reference

The table `building_data` contains 9 columns. The values in the columns `parcel_s`, `building_name`, `building_address`, `property_type`, and `energy_audit_due_date` have a `str` data type. The values in the rest of the columns `int` or `float` data types.

parcel_s	building_name	building_address	postal_code	floor_area	property_type	year_built	energy_audit_due_date	energy_use_2021
0010/001	2801 Leavenworth Street	2801 LEAVENWORTH ST	94109	133675	Commercial	1907	2024-04-01T00:00:00.000	6.21001e+06
0010/002	Argonaut Hotel-SV	495 JEFFERSON ST	94109	180000	Commercial	1907	2025-04-01T00:00:00.000	7.34107e+06
0011/008	Anchorage Garage	500 BEACH ST	94133	198525	Commercial	1974	2024-04-01T00:00:00.000	1.88699e+06

... (590 rows omitted)

The `zip_codes` table contains 4 columns. All the values in this table are either `float` or `int` data type.

zip	latitude	longitude	irs_estimated_population
94102	37.78	-122.42	21610
94103	37.77	-122.41	22940
94104	37.79	-122.4	1720

... (48 rows omitted)

At some point, you are asked to create the table `building_data_geo`. It should look like:

postal_code	parcel_s	building_name	building_address	floor_area	property_type	year_built	energy_audit_due_date	energy_use_2021	latitude	longitude	irs_estimated_population
94102	0296/001	449 Powell Street	449 POWELL ST	34173	Commercial	1913	2024-04-01T00:00:00.000	2.08193e+06	37.78	-122.42	21610
94102	0296/005	Chancellor Hotel	433 POWELL ST	46800	Commercial	1914	2021-04-01T00:00:00.000	3.01398e+06	37.78	-122.42	21610
94102	0296/006	400 POST ST	400 POST ST	61807	Commercial	1909	2020-04-01T00:00:00.000	9.32405e+06	37.78	-122.42	21610

... (590 rows omitted)