Fall 2024 Midterm Exam

Foundations of Data Science

Name: Sample Solutions

Question:	1	2	3	4	5	6	7	8	9	10	11	12	Total
Points:	3	4	3	6	6	35	4	9	4	9	5	12	100
Score:													

Instructions

- Make sure you have a copy of the Midterm Exam Reference Guide with the Table Reference included.
- 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 template blanks or boxed 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 Python code:

```
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.

1. (3 points) In an effort to investigate whether a new treatment is effective at reducing hip pain, researchers randomly sampled 100 patients from a medical group that used that new treatment, as well as other older treatments. They asked each patient whether or not they had received the new treatment or an older treatment, as well as whether or not they had experienced a reduction in pain. Overall, the results showed that those who received the new treatment observed a significant reduction in pain.

   ☐ This an experiment.

   ☐ This is a randomized controlled experiment.

   √ **This is an observational study.**

   ☐ The new treatment causes a reduction in pain.

   √ **There is a significant association between the new treatment and pain reduction.**

2. Assume the following code has been run:

```
array_1 = make_array(1, 2, 3, 4)
array_2 = np.arange(10, 14)
array_3 = np.arange(4)
```

   For the following expressions:

   - If the code will run without producing an error, provide the output in the provided box.
   - If the code will produce an error, describe the error in the provided box. We don't expect you to provide the technical error type (syntax error, type error, etc.); just describe what is wrong in general terms.

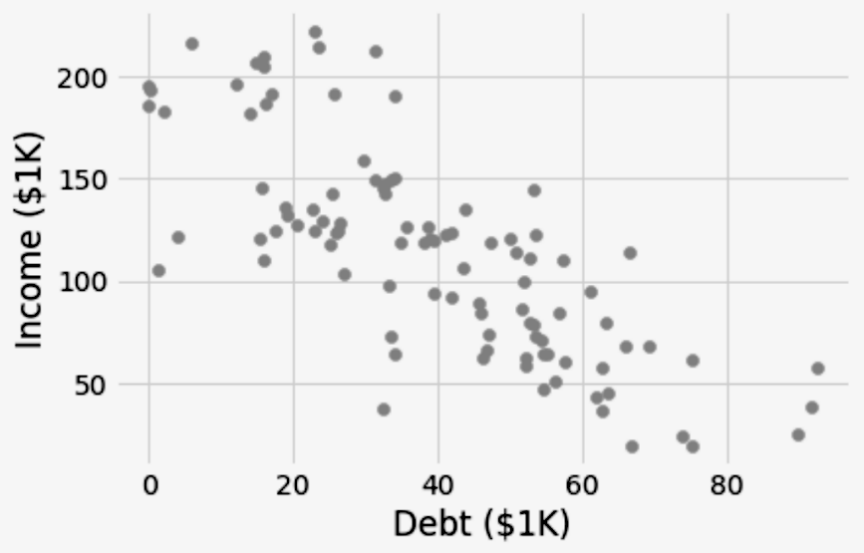
   (a) (2 points) `np.sum(array_1 + array_2)`

   Sample Solution: This code will produce the output `56`, an integer.

   (b) (2 points) `np.sum(array_1 + array_3)`

   Sample Solution: This code will produce the output `16`, an integer.

3. (3 points) Suppose you are curious about the financial situations of recent data science graduates. You have data from a sample of 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 graduates in the population with a debt greater than \$100K.

√ **Among the graduates surveyed, at least 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) (2 points) The standard visualization to understand the distribution of building types is: (choose one)

   √ **A bar chart** ◯ A line plot ◯ A scatter plot ◯ Overlaid histograms

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

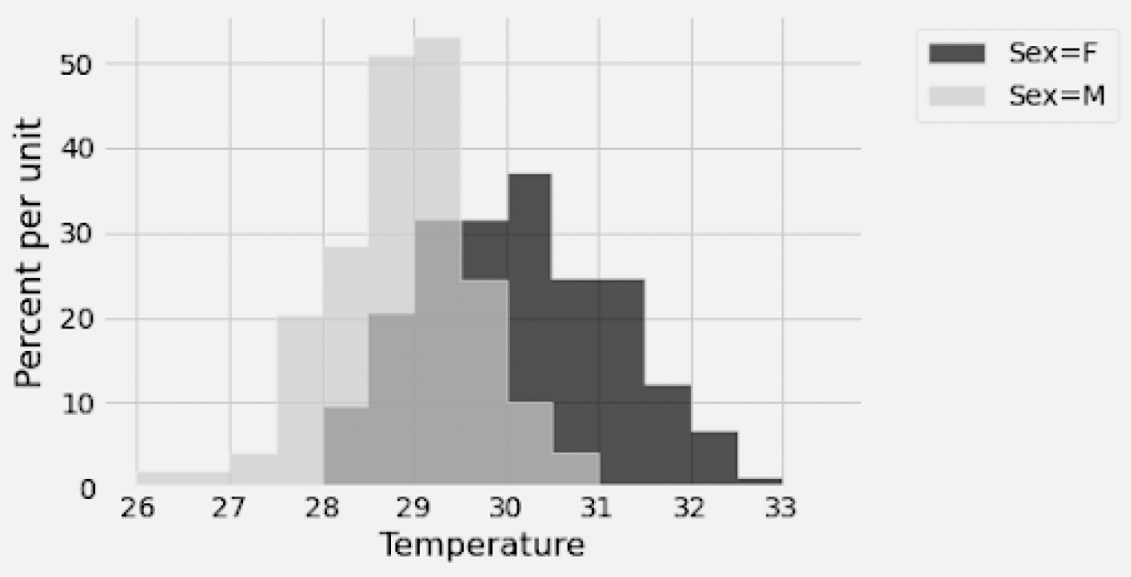
   ◯ A bar chart ◯ A line plot √ **A scatter plot** ◯ Overlaid histograms

   (c) (2 points) The standard visualization to compare the distributions of the estimated values (sale price) of the two types of buildings is: (choose one)

   ◯ A bar chart ◯ A line plot ◯ A scatter plot √ **Overlaid histograms**

5. When hatching a baby turtle from an egg, we incubate the egg at some temperature. A researcher read that the temperature of which an egg is incubated influences whether or not the turtle hatches male or female. To test this, 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) (2 points) In this sample, more than 50% of the male turtles were incubated at a temperature between 29.0 and 29.5 degrees.

◯ True

√ **False**

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

(b) (2 points) In this sample, there are at least 20 female turtles incubated between 30.5 and 31 degrees.

◯ True

◯ False

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

(c) (2 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.

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 = ____________.____________(____________, ____________)
   commercial_buildings._____________
   ```

   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 = ____________.____________(____________, ____________)
   ____________.____________(____________).____________
   ```

   Sample Solution:

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

(c) (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')
```

(d) (3 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
```

(e) (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))
```

(f) (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')
```

(g) (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)
```

(h) (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 =____________.____________(____________)
by_zip_sorted = ____________.____________(____________, ____________)
_______________________________
```

Sample Solution:

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

(i) (4 points) Create a table with two columns showing the mean 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 mean energy use for the buildings with that postal code.

```
reduced_data =  ____________.____________(____________, ____________)
____________.____________(____________, ____________)
```

Sample Solution:

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

(j) (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__ == 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)
```

8. In a game called September, players take turns selecting tokens and making moves based on the selected tokens. During a player's turn, they randomly select one token from a container and keep it; then randomly select another token from the container and keep it; 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 of the same kind of tokens when it is their turn?

   Sample Solution: `(21 / 34) * (20 / 33) + (12 / 34) * (11 / 33)`

   (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) (2 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 either correctly guess the secret word of the day or run out of guesses. Their guess count is either the number of guesses needed to guess the correct word (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) (2 points) You would like to test whether the distribution of guess counts among the 1,000 students who played Wordle is different from the distribution provided by Josh Wardle. What test statistic could be used to run this hypothesis test?

   Sample Solution: The TVD can be used for this test.

(d) (2 points) If you 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 _________(______, ______, ______):
    """Repeatedly roll n dice and check how many results are k or larger."""

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

    for _________________________________
        rolls = _________________________________
        outcomes = ________________(outcomes, np.count_nonzero(rolls >= _______))

    Table().with_column('Outcomes', _____________).__________(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))
```

12. According to a March 2024 Statista study, 6.48% of all US adults (18+) "trust a great deal in AI ability to make ethical decisions". You are on a research team that wants to see if the percentage is higher for those in the Bay Area, so you collect a random sample of 100 adults (18+) around the the Bay Area and ask them the same question. You found that 8 of them responded that they trust a great deal in AI ability to make ethical decisions. In order to decide between these two positions, the data scientists will conduct a hypothesis test.

(a) Consider the following statements:

A: In the population of San Francisco adults, 8% of them trust a great deal in AI ability to make ethical decisions.

B: In the population of San Francisco adults, more than 8% of them trust a great deal in AI ability to make ethical decisions.

C: In the sample of San Francisco adults, 8% of them trust a great deal in AI ability to make ethical decisions.

D: In the sample of San Francisco adults, more than 8% of them trust a great deal in AI ability to make ethical decisions.

E: The percent of adults in San Francisco who trust a great deal in AI ability to make ethical decisions is the same as the percent of adults in the US trust a great deal in AI ability to make ethical decisions. Any difference we see is due to random chance.

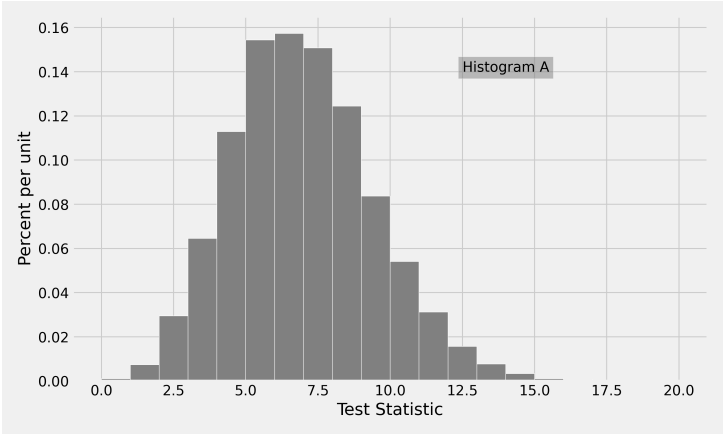
Fill in the following blanks with the letter of the statement that best completes the provided statement.

i. (2 points) Statement **E** is the most appropriate null hypothesis considering the goal of the test and available data.

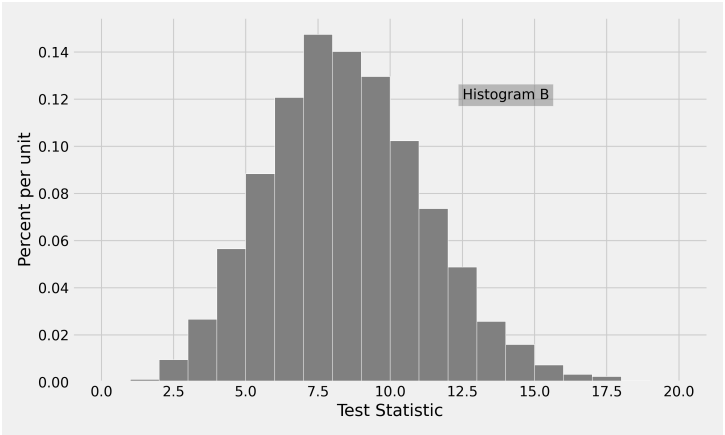
ii. (2 points) Statement **B using 6.48%** is the most appropriate alternate hypothesis considering the goal of the test and available data.

(b) (2 points) In order to decide between the two hypotheses, the data scientists have picked an appropriate test statistic and simulated it 10,000 times under appropriate conditions. Which of the following visuals shows the distribution of their simulated values? Select one.

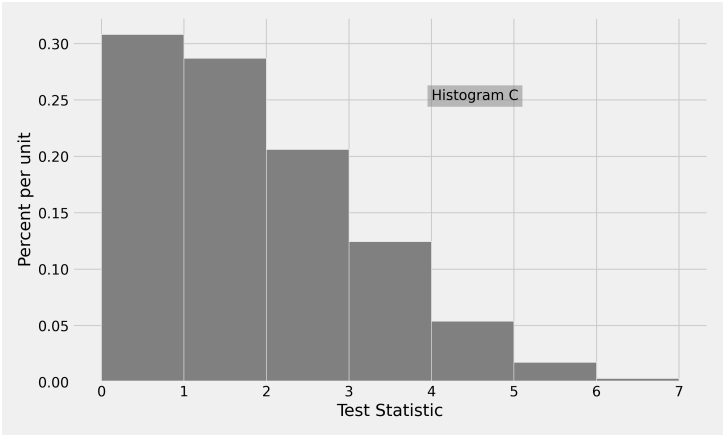
√



○



○



(c) (3 points) The data scientists store the 10,000 simulated test statistics in an array called `simulated_test_stats`, and the observed test statistic in a variable called `observed_stat`. Write code that will compute the p-value for this hypothesis test

```
p_value = ____________________________________________________________________
```

Sample Solution: One possible solution is:

```
p_value = np.mean(simulated_test_stats >= observed_stat)
```

(d) (3 points) The p-value turns out to be 0.328. If the p-value cutoff for this test is 5%, what is an appropriate conclusion to this test? (Select all that apply.)

- √ **The data are consistent with the null hypothesis.**
- ☐ The data are consistent with the alternative hypothesis.
- √ **There is evidence to support the claim that the percentage of San Francisco adults who trust a great deal in AI ability to make ethical decisions is the same as the percentage of US adults.**
- ☐ There is evidence to support the claim that the percentage of San Francisco adults who trust a great deal in AI ability to make ethical decisions is more than 6.48% (or the percentage of US adults).
- ☐ The percent of adults in San Francisco who trust a great deal in AI ability to make ethical decisions is more than 6.48%.
- ☐ The percent of adults in San Francisco trust a great deal in AI ability to make ethical decisions is the same as the percent of adults in the US who report being online almost constantly. Any difference we see is due to random chance.

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)