HW01: Coding Portion

PSTAT 100: Spring 2024 (Instructor: Ethan P. Marzban)

MEMBER 1 (NetID 1) MEMBER 2 (NetID 2)
MEMBER 3 (NetID 3)

Coding Portion: Health Inspections

> **Tip**
>
> Don't try to answer the sub-questions within any one question in list format - write your answers narratively, referencing code output wherever necessary. Additionally, think of the Coding Portions of the Homework Assignments as Mini-Mini-Projects. Specifically, some of the questions that are asked of you may be open-ended, which is by design! Feel free to stop by office hours (either the Instructor's or the TAs') to discuss!

The County of Los Angeles Department of Public Health routinely publishes results of environmental health inspections for several types of businesses (e.g. restaurants, apartments, etc.) at this link. In this part of our homework, we will investigate some of the results of these health inspections; the specific dataset we will be using can be found in the `data` subdirectory, with the file name `safety_ratings.csv`, and includes the following variables:

- `Facility`: the name of the facility being reviewed
- `Last Routine Inspection`: date of the last routine inspection (as of early March 2024)
- `Score`: score of the last routine inspection
- `Address`: address of the facility being reviewed
- `City`: city of the facility being reviewed (for those unaware, the county of Los Angeles is comprised of several smaller cities; e.g. Burbank, Santa Monica, etc.)

Also included in the `data` subdirectory is a file called `city_info.csv`, which contains selected information about the various cities included in the County of Los Angeles (data accessed and modified from this source). Specifically, the `city_info.csv` dataset contains the following variables:

- `City_Name`: the name of the city
- `Supervisorial_District`: the Supervisorial District of the city
- `Class`: the class of the city
- `Population_2010`: the population of the city in 2010
- `Inc_Yr`: the year of Incorporation of the city
- `Inc_Month`: the month of Incorporation[1] of the city
- `Inc_Day`: the day of month of Incorporation of the city

[1] **Incorporation**, in an urban geography context, refers to the act of officially forming a city.

Part 1: Exploring the Cities

Let's start off by exploring the cities included in the County of Los Angeles (i.e. by exploring the `city_info.csv` file)

! Question 1

- According to the `city_info.csv` file, how many cities are located in the county of Los Angeles?
- What was the total (aggregated) population of cities in Los Angeles in 2010?
- What was the most recent city to be Incorporated in the County of Los Angeles?
- What was the oldest city to be Incorporated in the County of Los Angeles?

ANSWERS TO QUESTION 1:

Replace this line with your answers

As a Data Scientist, it is important that we understand as many of the variables in our dataset as possible (which sometimes involves drawing on **domain knowledge.**) Google is a great resource for this! For example, it's not entirely obvious (from our dataset alone) what the "`Class`" of a city refers to.

! Question 2

- What are the different classes of cities?
- Use Google to look up what differences in these classes of cities, and write down a few.

ANSWERS TO QUESTION 2:

Replace this line with your answers

Similarly, not all of us may know what the different Supervisorial Districts of Los Angeles are.

! Question 3

- Use Google to look up how many Supervisorial Districts there are in the County of Los Angeles, and write down their names.

ANSWERS TO QUESTION 3:

Replace this line with your answers

Alright, let's flex our statistical knowledge a bit.

! Question 4

- Generate a barplot of Incorporation Year, and identify which year/s saw the greatest number of cities incorporated.
- Does there appear to be a month in which Incorporations typically occur? Answer this question using a graph.

ANSWERS TO QUESTION 4:

Replace this line with your answers

We can also flex our `tidyverse` skills.

! Question 5

- Use the `group_by()` function to group the `city_info` dataset by Supervisorial Districts, and compute the total (aggregate) population within each Supervisorial District.

ANSWERS TO QUESTION 5:

Replace this line with your answers

Part 2: Exploring the Restaurants and Ratings

Alright, let's turn our attention to the restaurants that were reviewed.

! Question 6

- How many restaurants were included in the dataset?

ANSWERS TO QUESTION 6:

Replace this line with your answers

If you skim through the dataframe, you might notice several restaurants located at 380 World Way.

! Question 7

- What major building is located at 380 World Way? (Use Google!) Why does it make sense that there might be many restaurants listed as having this location?
- How many restaurants are located at this address?

ANSWERS TO QUESTION 7:

Replace this line with your answers

Okay, that's enough preliminary exploration (for now). Let's turn our attention to the heart of this dataset: the safety ratings!

! Question 8

- Group the `safety_ratings` dataframe by city, and compute the median safety rating within each city.
- Use this to produce a graph with city name on the y-axis and average (**median**) score on the x-axis. Play around with axis text size and figure margins to make the figure as long as possible.

ANSWERS TO QUESTION 8:

Replace this line with your answers

Now, the graphic we produced in the question above is a bit misleading, because we know that not all cities have the same number of restaurants! As such, number of restaurants surveyed might be a confounding variable that artificially inflates (or deflates) a city's average safety rating.

! Question 9

- Re-do your plot from the previous question, this time scaling each point according to the number of restaurants that were included in the city. You may find it useful to apply a scaling transformation (e.g. square root) to the point sizes.

ANSWERS TO QUESTION 9:

Replace this line with your answers

Now, this plot is actually revealing something else about our dataset. Note, for example, that our plot includes a city called "(213) 385-9900 ___". This is clearly a mis-input.

! Question 10

- What was the name of the restaurant whose `City` was listed as `(213) 385-9900 __`?
- Use Google to look up this restaurant, and find which city it is really located in. Then, replace its `City` value (in the safety ratings dataframe) with the correct city.

ANSWERS TO QUESTION 10:

Replace this line with your answers

Additionally, note that our plot contains both a city called "Woodland Hills" and a city called "Wkoodland Hills". This is *also* clearly a mis-input!

! Question 11

- List out the unique values of the `City` variable as they appear in the safety ratings dataframe. Identify which values you believe to be typos (e.g. "Wkoodland Hills"); write down a list of these misspelled cities.
- Replace the misspelled city values with their correct spelling (e.g. replace all instances of "Wkoodland Hills" with "Woodland Hills", etc.)

ANSWERS TO QUESTION 11:

Replace this line with your answers

Finally, note that there is a city called "California" in our dataset, that has a suspiciously small point on our plot (indicating that there is a suspiciously small amount of restaurants included in this city).

! Question 12

- How many restaurants have a `City` value of `"California"`?
- Use Google to look up each of these restaurants; replace their `City` value with their correct city locations (as identified by Google).

ANSWERS TO QUESTION 12:

Replace this line with your answers

Part 3: Further Exploration of Ratings

Do more populous cities seem to have different average safety rantings than less populous cities? This is the main question we're going to try and answer in this part, by using plots.

! Question 13

- Merge the safety ratings and cities dataframes. As a hint: you may need to use the `toupper()` function somewhere in this step. Display the first few rows of the merged dataframe.

ANSWERS TO QUESTION 13:

Replace this line with your answers

Now that we have both the safety rating information as well as the populations in a single dataframe, it's time to begin formatting our dataframe into a format that `ggplot()` will recognize.

First, notice that not all cities included in the safety ratings dataframe appear in the cities dataframe. (This is largely because the safety ratings dataframe includes *neighborhoods* and a few neighboring cities of LA, whereas the cities dataframe includes only cities that were formally incorporated into the county of LA). To simplify our considerations, let's here on out focus only on cities that were formally incorporated into the county of LA.

! Question 14

- Make a dataframe that includes only the following variables: `City`, `Supervisorial_District`, `Score`. Group this dataframe by `City` and `Supervisorial_District`, and compute the average rating of each city/supervisorial district combination along with the population of the underlying city. Remove all cities with a missing `Supervisorial_District` value. The first few rows of your final table should look something like this:

City	Supervisorial_District	med_score	pop
AGOURA HILLS	3	97	23387
ALHAMBRA	5	95.5	89501
ARCADIA	5	97	56719
ARTESIA	4	96	17608

Hint: When displaying the population values, think about what summarizing metric you might be able to use to extract out the desired population value. (You could also consider simply appending the population column from the original `cities` dataframe.)

ANSWERS TO QUESTION 14:

Replace this line with your answers

Okay, this is looking pretty good! Let's start making some plots.

! Question 15

- Use your dataframe from the above question to create a scatterplot of median safety ratings (on the y-axis) and population (on the x-axis). Color your plot based on supervisorial district.

ANSWERS TO QUESTION 15:

Replace this line with your answers

The different supervisorial districts are getting a bit muddled - coloring might not have been the best choice. When it comes to displaying variations across categories, another option available to us is **facetting**.

! Question 16

- Make another scatterplot of median safety ratings (on the y-axis) and population (on the x-axis); this time, use the `facet_wrap()` function to facet based on supervisorial district.

ANSWERS TO QUESTION 16:

Replace this line with your answers

Finally, as mentioned many times throughout this course, interpreting our plots is a key part of being a good datascientist.

! Question 17

- Does there appear to be a relationship between median safety ratings and population? Does the nature of the relationship appear to change across supervisorial districts?

ANSWERS TO QUESTION 17:

Replace this line with your answers
