4.1 Introduction

In this Sprint you will be required to do some initial data exploration on the data set that you put together in Sprint #2 with respect to the questions that you set up in Sprint #1.

In order to complete this assignment, you need to use Pandas and SQL in order to generate a set of charts and plots regarding your data. Before going into specific requirements, three grading guidelines will be discussed. These guidelines should provide information on how the report (and code) will be graded.

1. **Reproducability:** The code needs to run and all plots and graphs from your report need to be generated by the code. If I run your Sprint #2 code and then your Sprint #3 code I should be able to produce all charts. There are two important aspects of this.
   (a) Your code in Sprint #3 should start from the relational database tables you built in Sprint #2. Note that if you see code that you would like to update in Sprint #2, or if there is additional data cleaning that needs to be done in the script, you should update those functions in your Sprint #2 code.
   (b) The plots and graphs in your report should be 100% generated from the code. This means that your code needs to save some type of image file which is then placed in the resulting PDF document.
2. **Interpretability:** All plots and graphics should be interpretable. This means that they should convey information (useful), they should be easy to read and make a point.[1]
3. **Insightfullness:** A major purpose of doing data exploration is bring insight into your problem. Plotting this for no reason (or without outlining why they are important to look at), is not useful and should be avoided.

4.2 Requirements

To complete this assignment, your need to complete three tasks and submit both code and a write-up (PDF file).

1. Clean up your code from Sprint #2. Now that you have a few moments, please go through your previous submitted code verifying that it is clean and easy to read. There should be no hard coded paths in functions, there should be documentation describing what sections of code do and the code should, once again, be able to be run by others. Address any concerns raised during the grading process.

   Implement proper permissions on the tables within your code. In this case, make sure that after your loading completes it allows all users within the group STUDENTS to access the table.

   Finally, provide a short description of those changes at the start of your write-up.

[1] There are a lot of good websites that show visualization design techniques. This one (https://paldhous.github.io/ucb/2016/dataviz/week2.html) for example, has a lot of good informatation and examples.

2. For each of your questions from Sprint #1, please write-up a plan on how you intend to solve it. These should be short descriptions (1-2 paragraphs) for each and should answer the "how" question around your analysis.

   As an example, lets assume that I'm using the BART data and one of my questions is to determine which stations are "most" utilized by BART riders. Than I would write the following:

   > I plan on looking at zip-code level population data and determining for each bay area zip code what the closest BART station is. There are roughly 200 zip codes, so I can do this by hand. I'll then download zip code population data from the US census to determine the number of potential riders of in that zip code.
   >
   > Once this data is collected and the map created, I'm going to rank the stations by the number of riders who enter that station each day divided by the population in applicable zip codes. If a zip code is close to more than one station, I'll divide the population equally between the stations. Using this information I'll be able to create a ranking of the station penetration, showing which ones are used most frequently by the community.

   For this, go as technical as you know, but don't worry too much about hitting every detail. If you have a problem where regression makes sense and you've taken a regression class, then mention it! We will be talking in more depth around different algorithms in class coming up, so you will have plenty of opportunity to learn more. Just make sure that if you say something it makes sense. Don't misrepresent your knowledge.

3. Please identify the top 4-5 variables that you are going to use to answer your Sprint #1 questions and plot them. Specifically identify 3 interesting and detailed plots (they each will have more than one variable) which also provide insight into your problem.

   While digging into specific plotting libraries is beyond the scope of this class there are currently three libraries that I see commonly used with Python: plotly, matplotlib and ggplot. If your data has geographic information then I *strongly* recommend you find a way to plot information related to this.

Your submission should consist of **three** components: (1) your updated Sprint #1 code (2) the code that generates your plots and (3) your write up (in **PDF** format). The same coding standards that applied in Sprint #1 also apply here: no notebooks, easy to read, etc. The code should run from start to finish, generating and saving the plot as image files in the current directory.