This part of the project will analyze the LA Parking Dataset by focusing on location. It will include visualizations such as Bar charts and HeatMaps utilizing Google Maps API and it will run statistical analysis and testing on zip codes, fine amounts and fine frequencies using Chi-square Test and ANOVA.

I first imported all dependencies that I felt were necessary to run an analysis. I set my api key to gkey and entered my google api key into a ‘config.py’ file. This file was placed in a .gitignore file to hide the key when posting to Git, GitHub and the internet.

Find Data:

We searched Kaggle for datasets that we would be interested in working with. We found a dataset that was “Kaggle maintained” so we felt that this would be a good dataset to work with. We spent the first class discussing with the group what questions we could ask with the data.

Clean Data:

The data set was so large and for this portion of the project I would be working with Google gmaps API I decided to cut the data into a sample set of 1,000 `head -n 1000 large\_data.csv > small\_data.csv` command. This way I would only have to make one call to the api once for the data.

After looking through the smaller data set I noticed that the latitude and longitude that we once thought would give us coordinates weren’t incorrect. They were both out of the range of acceptable coordinates so I used the api call to grab the coordinates. I continued cleaning the data by using a Pandas command `.apply` and a lambda function to add “Los Angeles, CA” onto the address string after seeing that it was calling for coordinates far outside of the Los Angeles, CA area when calling for generic named addresses such as “525 S. Main St.”

I built a new dataframe with the data we wanted and added columns for the data we would retrieve. After creating and running a `For` Loop to retrieve the data with an API call utilizing a `Try` and `Except` to catch any errors, I next moved to the zip code data. I used the api response and Regular Expression to manipulate the `format address` string to parse out the zip code. I think stored the zip into the data frame. After changing the blanks to NaN’s I dropped the NaN’s, ready to move on to my analysis of the data.

Analyze Data:

I used gmaps to build a HeatMap based on location address and “Fine Amount” therefore giving higher weight to higher fines. This visualized The higher fines areas in LA. I next researched how to put a boundary area for the LA city limits. This was not something that was built in to gmaps so I had to go retrieve the data from another api call, bring in that data to a list of tuples then bring that list of tuples into gmaps. Now I had my heatmap and a boundary. I used gmaps layers to place one on top of the other for my final map visualization.

I build two bar chart for visualization of zips codes and fine amounts and fine frequencies. These two listed all 85 zip codes and the total amount of fine while the other displayed the frequencies. It showed that visually certain zip codes have more ticket counts and some have more fine totals. I added an average mean line for both barchart to see the variation.

*Chi-square test*

Next I asked the question “Are you more likely to get tickets in certain zip codes than others?”

The Null hypothesis was: “There is no difference in zip codes and frequencies of tickets.”

I used a Chi-square test to test this hypothesis. I took the average of all the zip code and ticket frequencies which amounted to 11.28 I wanted to see if the fine count in each of the zip codes was give purely to chance. My Chi-square statistic was 3,090.78 well higher than my critical value at the 95% confidence interval with 84 degrees of freedom (85 zip codes minus 1) of 105.28. Therefore I reject the null hypothesis that this could be given to chance. I concluded the result were statistically significant.

*ANOVA*

My final test was of the question: Are fines greater in certain zip codes for the top five most ticketed zip codes?

My null hypothesis was: There is no difference in fine amounts of the top five zip codes. So parking in one zip code will result in the same fine if given a ticket.

I used an ANOVA test for this analysis. I took the top 5 zip code ticket frequencies and the average of the fines to run the analysis. I built a list of the top 5 zip code frequencies to filter out only those zip codes to a dataframe. I built a boxplot for a visual aid. It shows the range of the fine amounts and the average of the fine amounts for the top 5 zip codes. Just taking a quick look at the visual aid gives us some idea but to make sure I ran an ANOVA. I built 5 groups for the top 5 zip codes and ran the f\_oneway test. The p-value was 0.000015 therefore I rejected the null hypothesis and concluded there is a statistical difference for fine amounts for the top 5 zip codes.

Things that I learned:

* You don’t always need to have a program to run and do it for you. Sometimes you need to go out and create the data or the program yourself. I saw this first hand trying to find if gmaps could put a boundary around the city limits. I ended up creating this on my own from another data source and feeding the list to gmaps

Things I would do if I had more time:

* Zip code list to better highlight in the heatmap the zones for the individual zip codes
* Compare types of violations to heatmap and zip codes
* Compare sample and population statistics using T-Tests