Lyft Bay Wheel Exploration:

Where to build the next station?

[Codes Repository]

Students:

KyuRi Kim

Tianhao Wu

Professor:

Marta Gonzalez

GSI:

Weixing Li

1. Introduction

The project is to look into people's behavior on using bikes and to analyze underused/overused shared bike stations and the reason for it. We intend to help Lyft decision makers to find where to build the next stations optimally. For the project, we focused on the bay area's bike stations and used BayWheel data provided by Lyft. This data set contains trip duration start/end time, start/end station ID, start/end long & lat, bike ID, and user type. No trajectories of trips are included. From this data, it is possible to show the trip distribution of bikes, mostly used stations, and usage time distribution.

There were two papers that are related to our project. Both "Micromobility evolution and expansion: Understanding how docked and dockless bikesharing models complement and compete - A case study of San Francisco" and "Enhancing equitable service level: Which can address better, dockless or dock-based Bikeshare systems?" used user data of Ford GoBike, a station-based bikesharing system, and JUMP, a dockless electric bikesharing system. These studies were based on San Francisco.

The objective of "Micromobility evolution and expansion:..." was to compare data of dockless bikesharing system and docked based bikesharing system to find out personal behavior. To find out people's behavior on using both systems, using the data, they have created images that show usage demand during AM and PM, graphs that show trip duration and distance, and graphs showing average daily trips and miles per bike for both systems. Also, they have looked into users' destination choices. After doing such analysis, they have found out that trip distance and trip duration was longer for JUMP. Also, users used GoBike to commute between major public transit transfer stations and used JUMP to travel to comparably lover-density neighborhoods.

The paper "Enhancing equitable service level:..." tried to find out whether dockless bikeshare systems are eligible to replace dock-based systems. In order to come to the conclusion, they primarily compared given conditions for both bikesharing systems and also considered both systems' service levels for communities of concern (CoCs). Then they formed images showing service areas of Ford GoBike and JUMP with MTC areas to compare each system's service level. Also, they have created images showing trip origins and destinations for both systems to see

mostly used areas. Through these images and analysis, they have reached the conclusion that dockless bikeshareing systems are better for CoCs and can replace dock-based systems.

Like these two papers, we will use data of bikes located in the bay area to analyze usage of bike stations and people's behavior. We will form some images and graphs about trip distribution, trip durations, and usage of bike stations.

2. Data and Methods

2.1 Data Inputs

The main data we used is BayWheels' Trip data, which is an open dataset released by Lyft. It includes monthly trip data of bikes from the year 2017 to the present. We've selected a recent month, October 2021, for exploratory data analysis. The complementary data/library is CensusGeocode python library, which outputs a geocode of given longitude and latitude pairs as input. Another complementary data is shape data of California census tracts. Raw polygon shape files are provided. We merged them into our flow and tessellation dataframes.

2.2 Data Cleaning

First Step of data cleaning is to clean out invalid and blank data entries. The raw dataset has 212,512 entries of data, we dropped 65,699 entries with missing values, which is 30.9% of the total data. We have 146,813 entries of data to work with.

Since geolocations of records are on the station decks level, which are unnecessarily precise, we calculate out one geolocation for each bike station by averaging out the longitude and latitude pairs of decks. This results in a tessellation data frame of bike stations, with inflow and outflow data.

Since such bike stations data are scattered, and not representing community areas in reality, we integrate all bike stations according to census tract, resulting multiple bikestations aggregate into one census tract. In doing so, we utilize CensusGeocode python library, which queries the GEOID of given longitude and latitude pairs. GEOID contains census tract id as last 6 digits. As a result, we have a tessellation dataframe of census tract outflows.

2.3 Data Exploration

Based on the Scikit-Mobility python library, we may easily output maps as we have flow dataframes. Knowing the start and end time of each trip, we are also able to analyze travel behaviors. We've analyzed trip duration, start and end time distribution for each day, and plotted the kernel density estimate probability plots for such behaviors. Figures will be shown in the results section.

3. Results and Discussions

3.1 Trip Distribution

Station to Station bike trips data provided by Lyft raw data can be visualized as a flow diagram utilizing sci-kit mobility library, shown in Figure 1. Red lines shows flows, blue dots are for bike stations, which are averaged coordinates of respective bike decks.

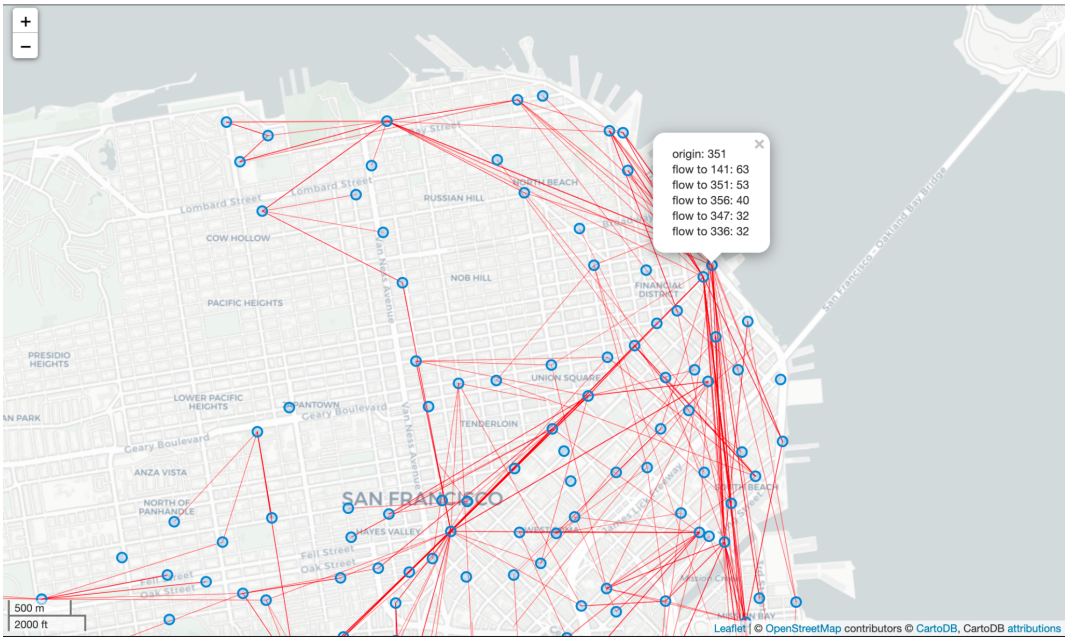

Figure 1. Bike Trip Distribution - Station to Station

Since stations are clustered in regions from the heatmap Figure A.1 in Appendix, we decide to aggregate stations into census tracts based on methods we mentioned in the data and methods section. By doing so, we can further analyze the trip data combining census tracts data queried from ACS. Resulting census tracts are shown as colored tessellation and flows are denoted by red lines, shown in Figure 2.

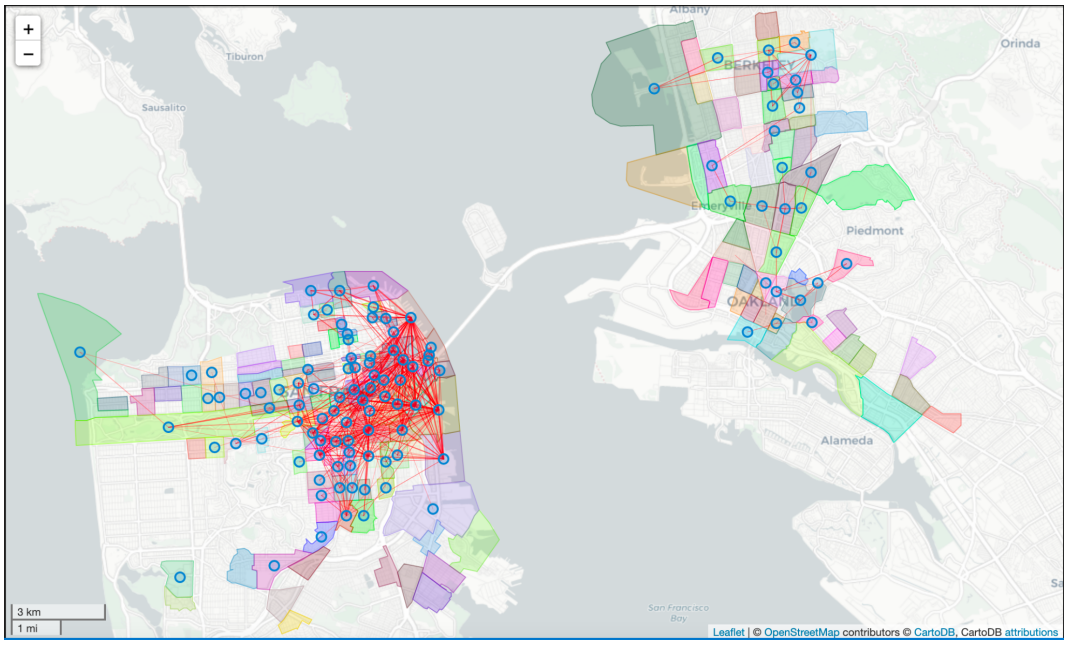

Figure 2. Bike Trip Distribution - Tract to Tract

3.2 Clustering Census Data

Utilizing Cenpy library, we queried American Community Survey (ACS) variables including total population, travel time to work, vehicles available, vehicles used in commuting, about census tracts involved in our analysis. We processed ACS variables into meaningful variables, stated as: average number of vehicles owned, average travel time, average number of vehicles used in commuting, land area per person. We also included the start and end frequency of bike stations into our clustering variables.

Next, we use elbow method and silhouette scores to decide how many clusters we should use in clustering census data. We decide to use 4 clusters as the optimal number of clusters, as shown in plots below, Figure 3 and 4.

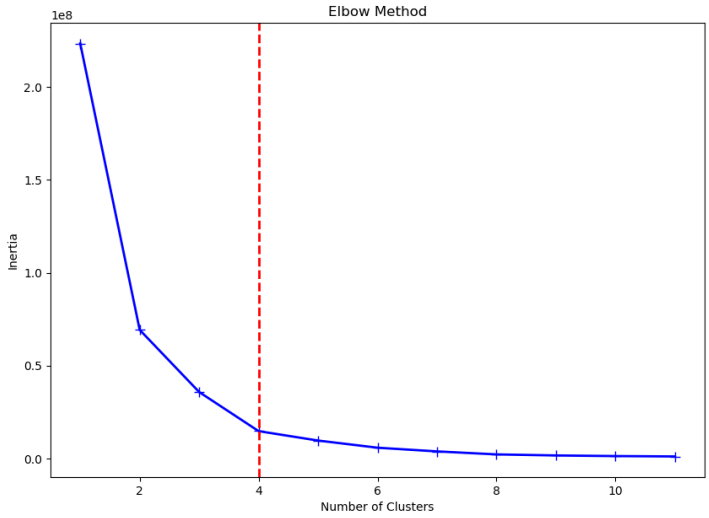

Figure 3. Inertia vs Number of Clusters

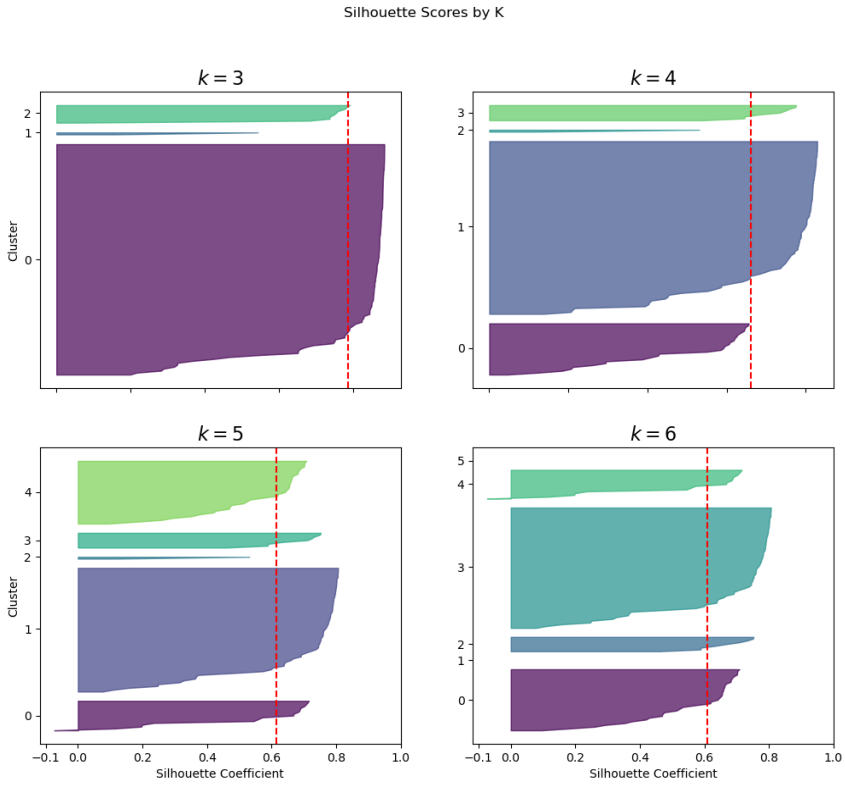

Figure 4. Silhouette Scores by different number of clusters

Following Figure 5 is clustered visualization of San Francisco and Berkeley census tracts based on processed variables. We may imply the reason for clusters by analyzing how census tracts are clustered by variables based on Figure A.2 in Appendix. Purple color area, cluster 3, are downtown SF, and we may imply that they have high frequency of bike usage, low commuting time, low land area per capita, and low average number of vehicles. Green color area, cluster 2, represents the majority of the census tracts, with the highest land area sum aggregated. They have low shared bike usage, higher land per capita, longer travel time and higher percentage of vehicles owned.

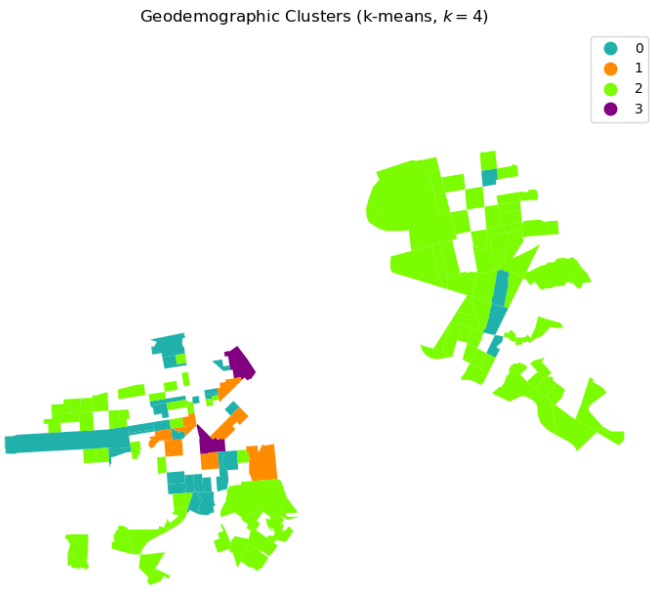

Figure 5. Clustering Visualization (K=4)

3.3 Network Analysis

After seeing how data are clustered, we did network analysis to see how they are connected. We assigned stations as nodes and each bike trip as an edge. Then, we got a histogram, Figure 6, that shows the number of nodes for each range of degrees.

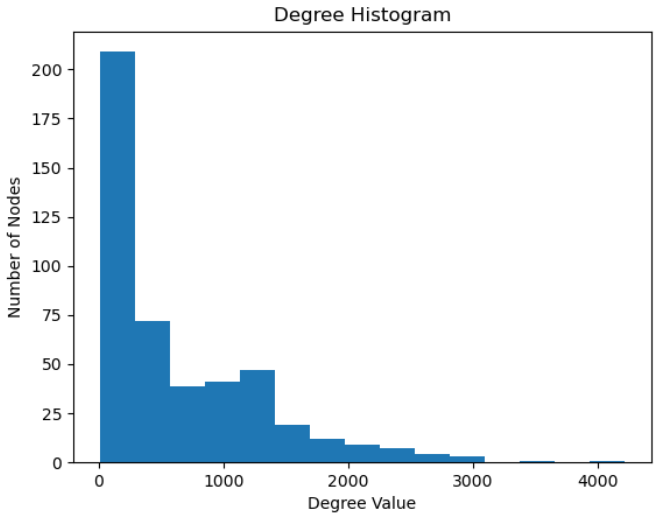

Figure 6. Degree Distribution Histogram

We also did a robustness test to see if the network is well connected. This test allows us to see the network breaking down as we remove connected nodes. For random failures, randomly selected stations that are full or empty will be chosen to be removed. However, for the attacks, the stations that are having peak usage will be selected and be removed. The result of this robustness test is depicted as a graph, Figure 7. Surprisingly, two graphs were having similar shape, but it was possible to see that removing stations with higher usage are causing the network to fail faster.

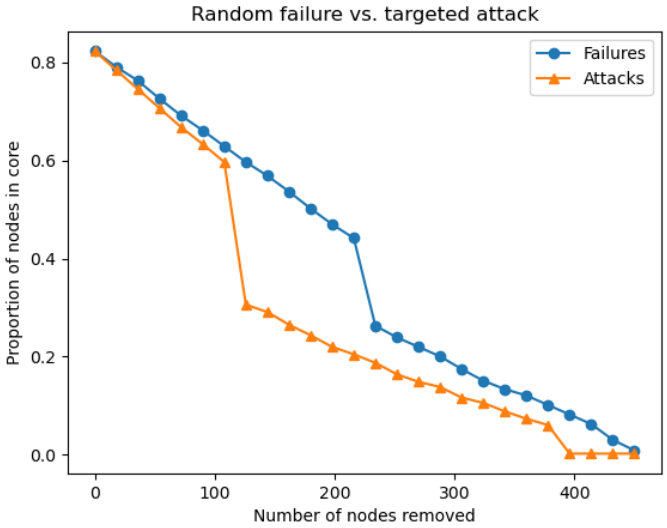

Figure 7. Robustness Test Result

3.4 Travel Behavior

To find out the personal behavior, we started with forming a graph that shows the usage density of the bike according to hours using its start and end time, shown in Figure 8. Two lines have similar shape but the line for end time is little behind the line for start time.

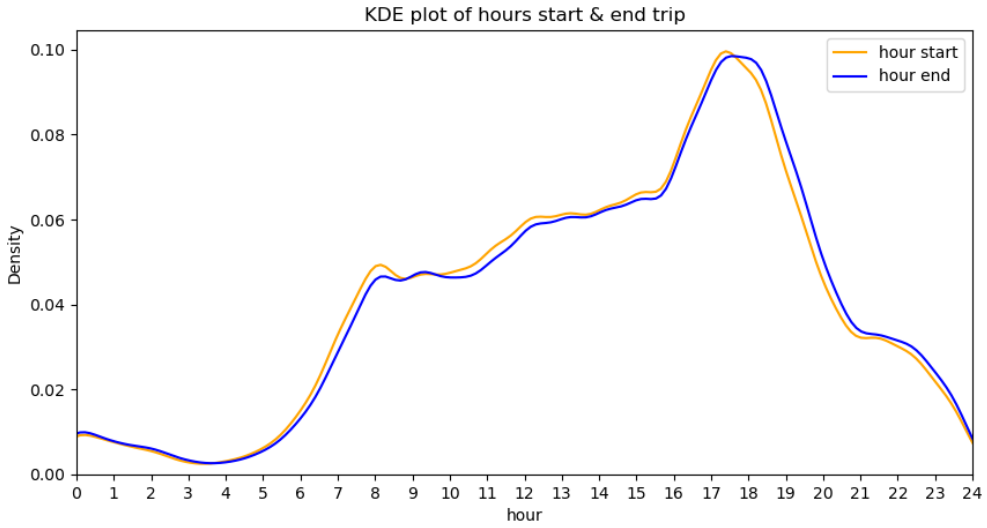

Figure 8. Usage Density Distribution for Start and End Time

To clearly see the trend and personal behaviors, we formed a graph for each day, shown in Figure 9. This graph showed that there are two different trends, one for weekdays and another one for weekends. To see these trends more clearly, we combined data for weekdays and weekends to form two lines, shown in Figure 10. From this graph, it could be easily seen that there are two peaks for weekdays. These peaks are formed at commute time. We could know people are using bikes to go to work and to come back from work. For weekends, we could interpret bike usage increases during daytime. This is possibly because people do not go to work, but people living in bay areas and people visiting bay areas are using bikes to go to other locations during daytime.

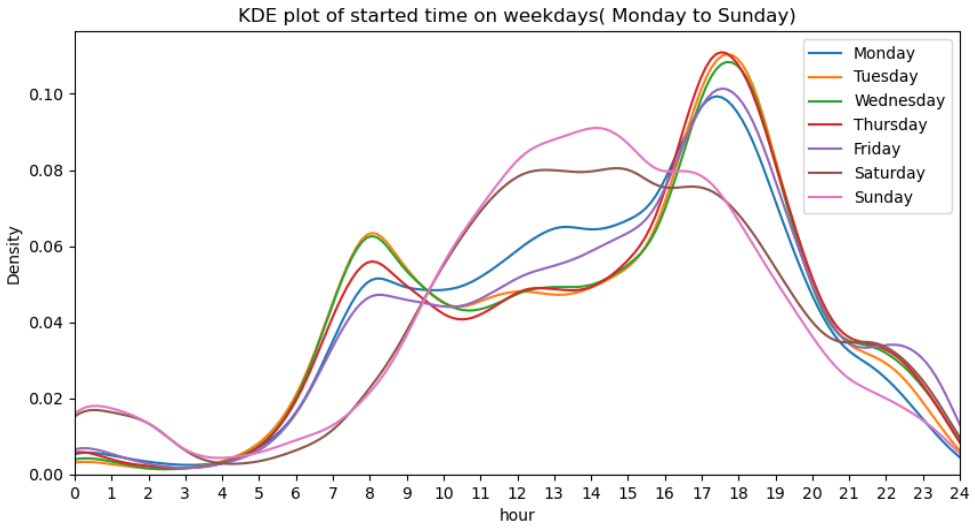

Figure 9. Usage Density Distribution for a Week

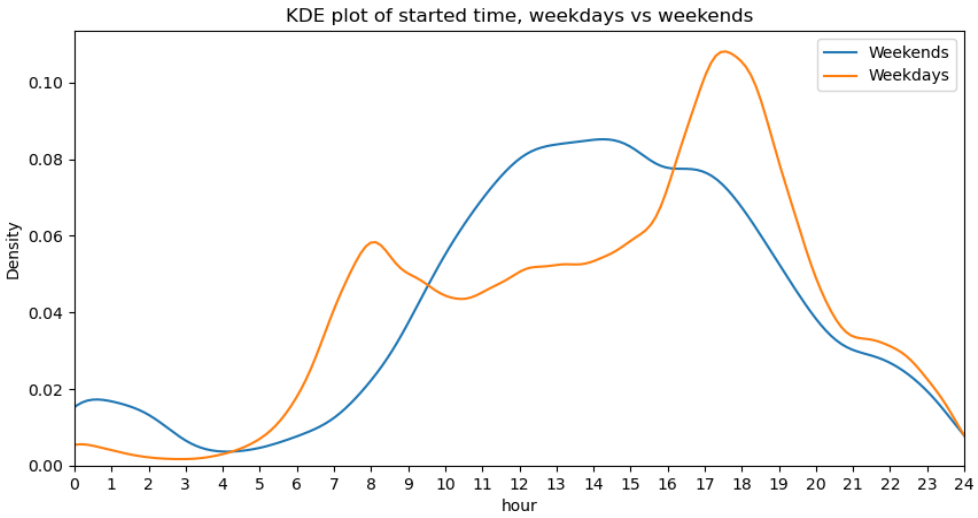

Figure 10. Usage Density Distribution, Weekdays vs. Weekends

Then we got trip durations using start and end time. To see the average value of trip duration, we formed a trip duration distribution graph, shown in Figure 11. The mean value of trip duration is 46.4 minutes, having 11.6 minutes of standard deviations.

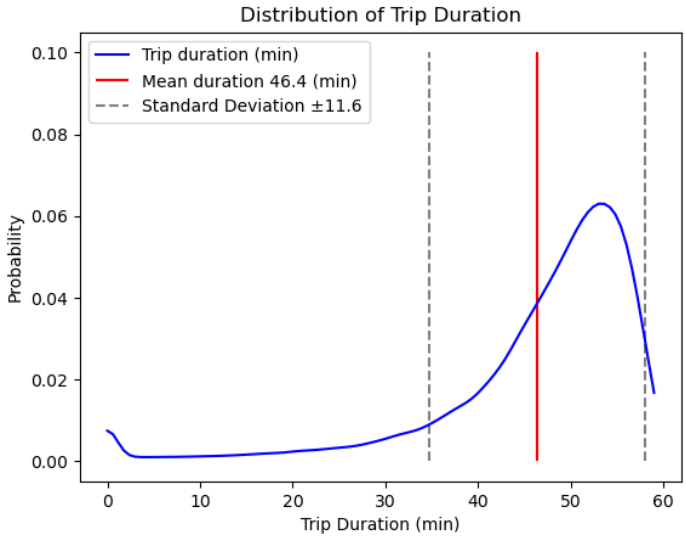

Figure 11. Distribution of Trip Duration

3.5 Station Serviceability

One of main concerns while analyzing the data was about dock usage. We wanted to see whether or if docks are underused/overused. To identify this matter, we formed a bar graph, Figure 12, that shows frequency of dock usage. To form a graph, starting data was used. The formed graph shows that most stations are used less than 250 times a month and there are a really small number of stations that are used more than 1500 times a month. These stations are found to be station ID of SF-J23-1 and SF-G27, and the station names of Market St at 10th St and Powell St BART Station (Market St at 4th St) respectively.

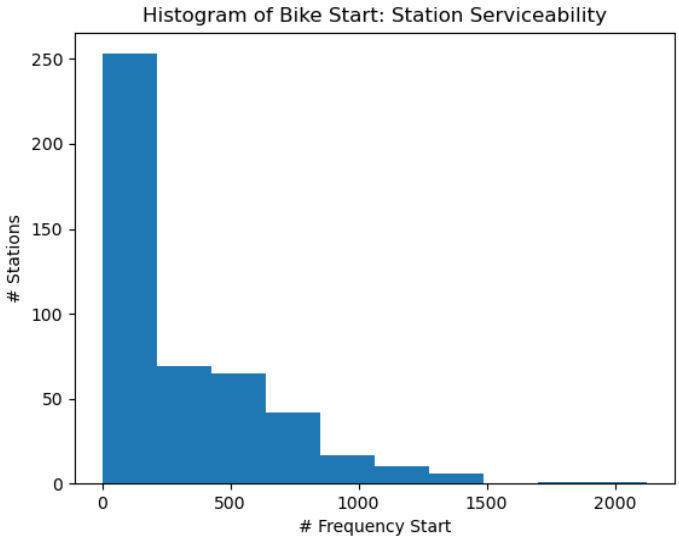

Figure 12. Frequency of Station Usage vs. Number of Stations

4. Future Work and Conclusions

The resulting data and graphs are saying that the bike usage is affected by people's everyday life. The usage increases during commute time for weekdays and increases during daytime for weekends because there are many visitors in bay areas. Also, most of the bike stations are used less than 250 times a month, but there are two outliers. These outliers are used more than 1500 times a month. Also by looking at the location of these stations, it is possible to know high usage demand is reasonable. Since two stations, Market St at 10th St and Powell St BART Station, are located in downtown San Francisco, they are getting overused. Therefore, to improve accessibility, we recommend ensuring these stations have enough bikes at all times for people to rent around these areas.

Also, the areas that are having high usage are not connected nor close to each other. This also means there are certain areas that are highly used. By knowing this, we want to recommend building more bike sharing stations in highly used areas. Building near overused stations is preferred.

For now, we are not joining much census data about tracts. In the future, analysis can query the cenpy library, and join more data about demographic information such as population, income level, education level, mode of transportation, travel characteristics, housing structures,

etc. The potential difficulty will be data query and cleaning of census data, and visualization of overlaying on maps.

References

[1] Lazarus, Jessica, et al. "Micromobility evolution and expansion: Understanding how docked and dockless bikesharing models complement and compete–A case study of San Francisco." *Journal of Transport Geography* 84 (2020): 102620.

[2] “System Data: Bay Wheels.” *Lyft*, https://www.lyft.com/bikes/bay-wheels/system-data.

[3] Qian, Xiaodong, Miguel Jaller, and Debbie Niemeier. "Enhancing equitable service level: Which can address better, dockless or dock-based Bikeshare systems?." *Journal of Transport Geography* 86 (2020): 102784.

Appendix

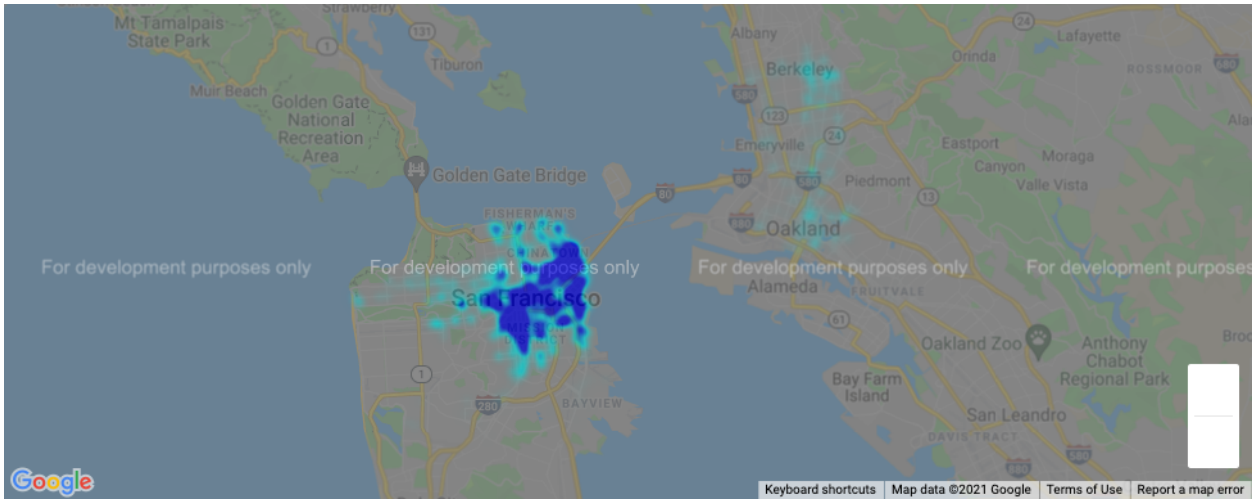

Figure A.1 Heatmap of bike starts

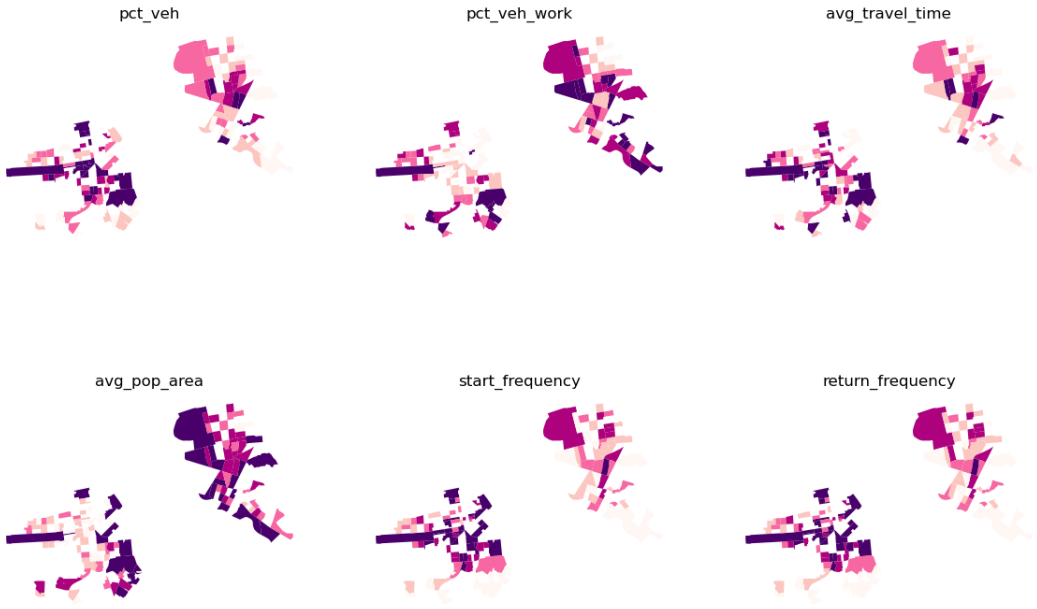

Figure A.2 Clustered by variables