Report - Bike Sharing Systems

Travel Patterns and Correlation Study

Team Member: Zhaohong Niu, Sunny Kulkarni and Chenxi Cui

Project Idea -

The objective of our report is to do a base study of the existing bike sharing systems and identify patterns towards solving the bike and spots imbalance problem.

As part of our study, we -

a. Analysed the citibike station location and its proximity to the subway stations.
b. Studied the riders trip data for other bike sharing systems for different locations in US to gain insights into the bike rider's behavior.

Motivation / Background -

Bike Sharing System (BSS) is an eco-friendly alternative to travel short distance in today's humming cities. Bikes stations are located throughout a city and Subscribers can take a bike from any station and return it to any other. The number of BSS and the riders are growing at an exponential rate. As of August 2014, there are over 500 systems with more than half of a million bikes. Citibike ridership increased to 10 million bike rides in year 2015, a 24 percent annual increase. (Refer Figure 1)

With this success comes a set of management problems. The biggest issue faced is system imbalance; bikes become clustered in certain areas leaving other places devoid of bikes. Trucks are Courier bikes are used to mitigate this problem. Courier bikes have trailers attached that can move four or five bikes at a time. Majority of trips occur during morning and evening rush hours. Moving bikes around a city is expensive and the cost per bike moved can climb over a dollar straining the very limited resources of the bike sharing system operators.

Also, study of historical rides from different perspectives would be the first step towards the solution. Identifying areas where bikes tend be clustered leaving surrounding places empty would be the focal point of attention and solution.

Existing Research -

Various research is done in this field especially to address the imbalance problem. Stations were clustered with similar usage patterns. Optimized planning phase also helps to solve the problem of what is the best number of bikes and docking spots at specific locations across the city. The existing research is segregated into -

a. Planning Phase - Identify best number of bikes and spots for each location to meet the surge and availability of docking spots during rush hours.
b. Pre-balancing - Utilize night hours and off-rush timing to move the bikes around the city. It is easier for trucks to move during night hours. Courier bikes that can carry up to 5 bikes are used to balance bikes during day time.
c. Mid-Rush balancing - Optimally move bikes between the stations during the rush hours to ensure availability bikes and spots for the riders. This requires drastically different approach than Overnight Pre-balancing as in-use bike riders are continuously affecting the target solution making the problem very dynamic to solve.
d. Self-Balancing stations - Identify the stations that balance on their own and identify the demographics, geolocation that is promoting it. These are the optimal stations that only need to be monitored and studied to build replicable solutions.
e. Incentive scheme - to complement rebalancing. Provide incentives to riders to shift their behavior to benefit the system as a whole.

[Ref. Research Section Document a)]

Optimization techniques such as IP formulation, tractable integer programming formulation and study of topography is applied to all the above scenarios to understand the solve the imbalance issue.

Using k-means clustering algorithm for Citibike data, three distinct clusters emerge from the stations; stations that accumulate, loose bikes stations that stay close to zero net flow. Analyzing both the morning and evening rush hours, these cluster of stations are heavily correlated to the residential versus business districts. (Refer Figure 3)

To augment this research we considered the study of trips and the delay trend. Delays, in large numbers, will impact the optimization algorithms.

Methods & Data -

- Identify proximity to subway stations

  We want to find out whether the proximity of subway and Citi Bike stations influence the use of Citi Bike. As Citi Bike company expanding its scale in New York City, there are more than 600 stations in year 2016. We ran simple python groupby functions using September 2016 Citi Bike data to updated the newest station list. With a shapefile of all subway entrances in New York City, we were able to create buffer around each entrances, and intersect with Citi Bike stations that are within the buffer. As (Refer Figure 2) shows below, blue dots are all Citi Bike stations that have more proximity than grey dots, which are further from subway stations. We set the buffer radius as 0.2 miles, because it's the reasonable walking distance users are willing to use Citi Bike to commute to subway entrances.

  To compare the difference between the two categories, we visualized the average daily use of two types of stations by extracting and combining start station ID of each rides. (Refer Figure 4) shows that stations within walking distance to subway stations have higher average use than stations that are not within walking distance. (Refer Figure 5 and Figure 6) to support our conclusion.

  As we successfully separate stations by proximity to subway stations, and identify difference between the categories, we want to find out if subway use has a correlation with Citi Bike use. MTA turnstile data has the enters and exits of each passenger. We picked up two stations to examine the pattern. As (Refer Figure 7) suggests, because the number differs too little by day, it's not visualized correlation between subway rides and Citi Bike use.

- Results

  Finally, we ran a multivariate regression methods to see if some demographics factors, e.g. neighborhood population, household units or population of children, have influence on citibike usage. As (Figure 8 and Figure 9) indicates, the R-square is too high because some strong multicollinearity between demographic attributes. Citi Bike use gets higher when its station is close to subway stations, while no significant statistical evidence to support our initial result.

- Analysis of Divvy Bike Sharing System in Chicago
  We analyzed the trend for Chicago's Divvy Bike Sharing System by correlating the data for rider's age and the duration of the trip. (Refer Figure 13). This plot tells us the primary riders age is between 15 and 75 years. Mostly the bikes are rented for 30 minutes, but there is set of users that continue to use the bikes and pay the premium fee. Divvy charges $1.50 for first delay after 30 minutes and later $6 for each half hour after 90 minutes of delay. (Refer Figure 14 and 15)

- Analysis of Bay Area Bike Sharing System in San Francisco
  We analyzed the annual bike ridership for period September 2014 to August 2015 and plotted a histogram to find the trend of riders across their trip duration. (Refer Figure 16). Though each bike ride is available for up to 30 minutes, before surcharge is applied, on an average most bikes are dropped within approximately 15 minutes. That tells us the average distance a user is willing to consider a bike as a commute option. This definitely combines with the reach of public transit system as the riders would prefer to ride than walk to remote places covering the distance in a much shorter time.

- Neighborhood Area Analysis
  We were curious about the future pattern of bike usages, especially about the trip duration and the most popular start and end stations. Thus, we made a prediction model based on the data from CitiBike system. We chose data from Aug. 2016 as our dataset, ran different classification algorithms to predict destination neighborhood and scored each on accuracy and precision. Each destination neighborhood is either correct or incorrect, so accuracy is an effective metric.

  In order to build model, first we analyzed and got the summery about our dataset which including trip duration, most and least popular stations (shows as Figure 10) . Then we extracted data features by using information from Citi Bike open data and GeoJSON file of NYC boundaries to assign each station to its associated neighborhood. This way 329 stations could be reduced to 37 neighborhoods. And by using extracted features we could get the average trip duration for each neighborhood and borough easily. (Refer Figure 11 and 12)

Graphs

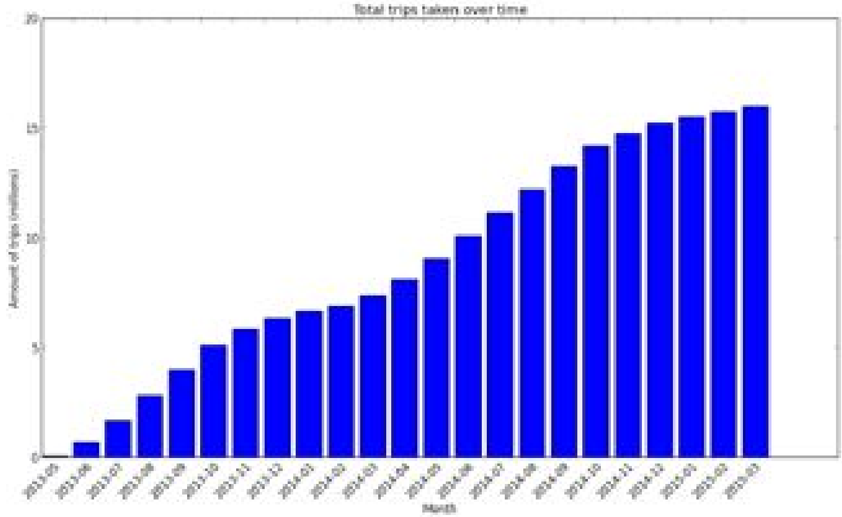


Figure 1 - NYC’s Citibike Cumulative Increase of Trips over Years [Ref. Research Section Document a)]

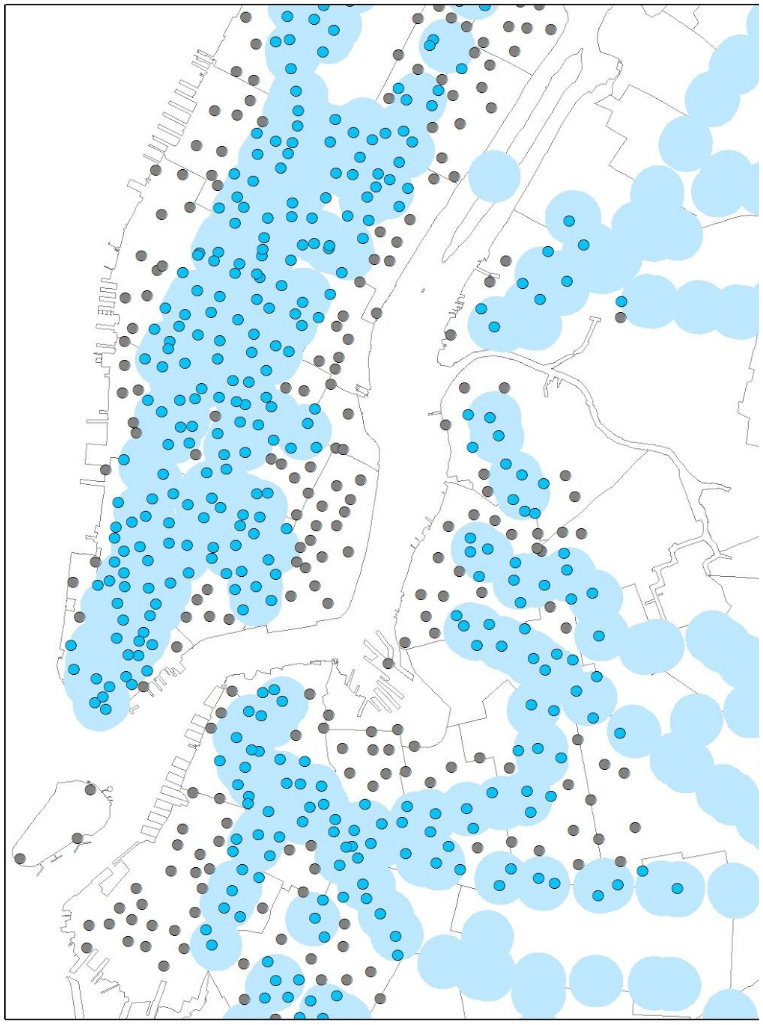


Figure - 2 - Vicinity of New York Subway Stations and the Citi Bike Stations.

Black dot => Citibike stations, Blue Circle center => Subway stations. Circle denotes 0.2 miles radius around the subway stations. We assumed 0.2 miles as a good walking distance to consider Citibike as a transport medium.

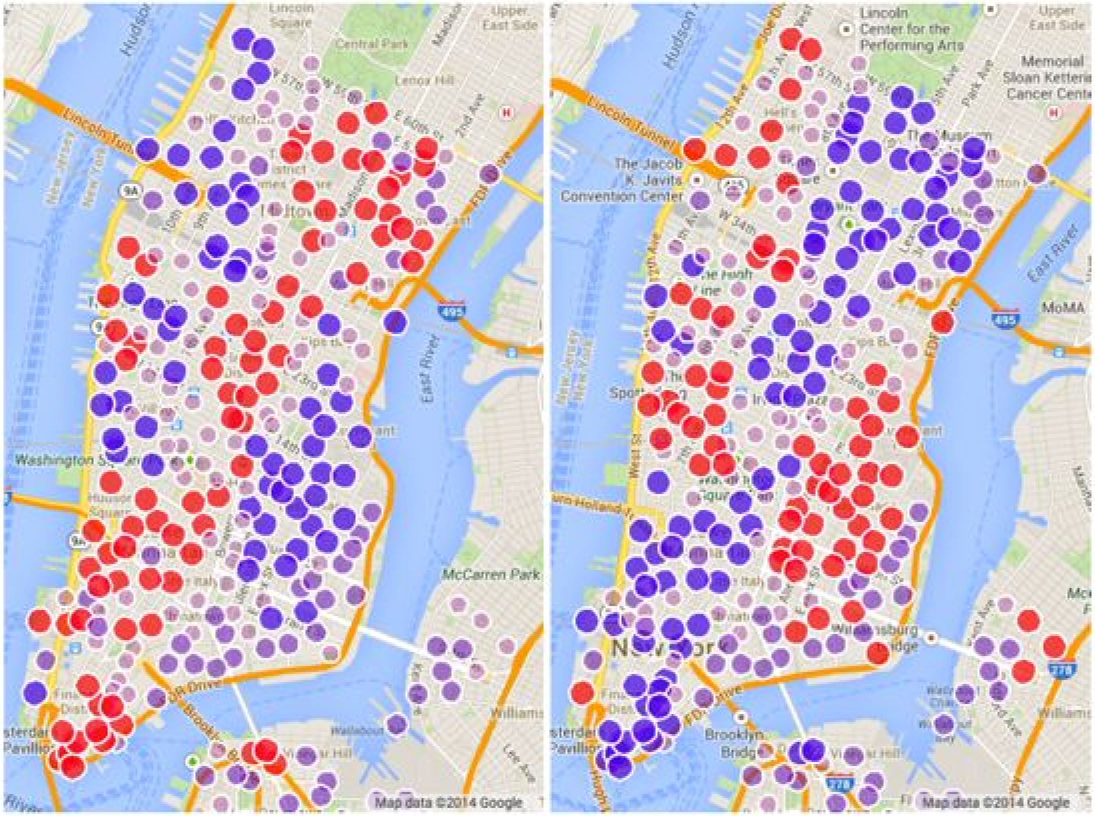


Figure 3 - NYC's Citibike Usage Trend - Morning (left) and Evening (right). Blue => Consumers, Red => Producers and Purple => Self-balancing. [Ref. Research Section Document a)]

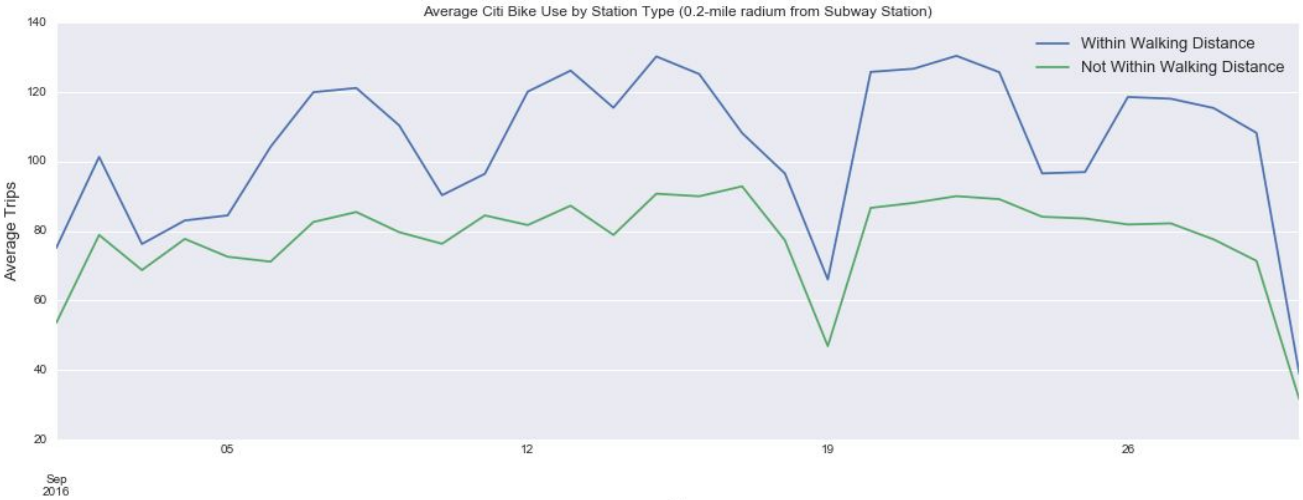


Figure 4 - Average Citibike use for MTA Station that have bike stations within 0.2 miles versus stations that do not have them.

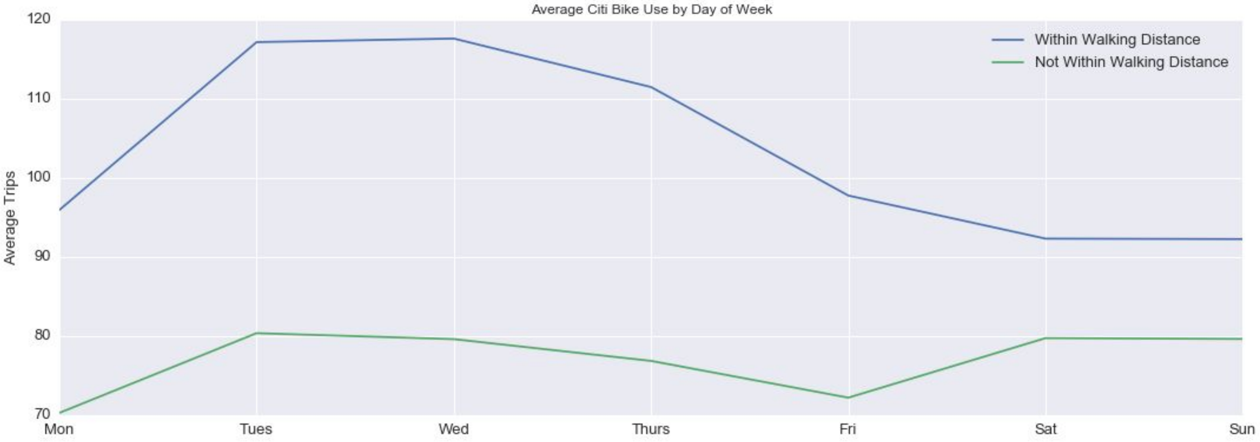


Figure 5 - Average Citibike use for MTA Station that have bike stations within 0.2 miles versus stations that do not have them - by <u>Day of Week (September, 2016)</u>.

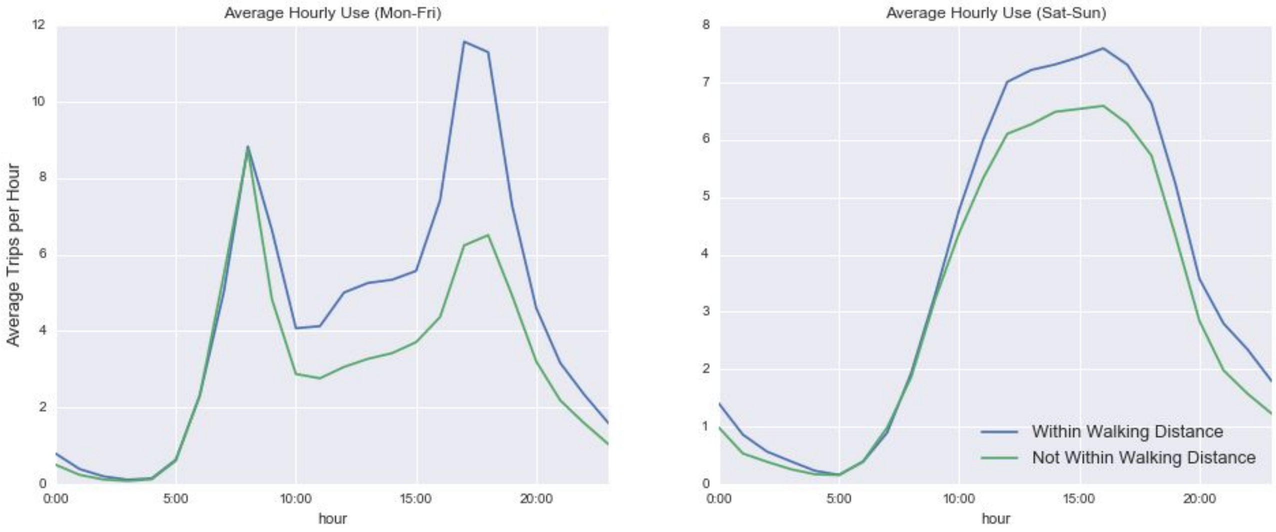


Figure 6 - Average trips per hour by Average Hourly use of Stations. Segregated by stations that have bike stations within 0.2 miles versus stations that do not have one (September 2016).

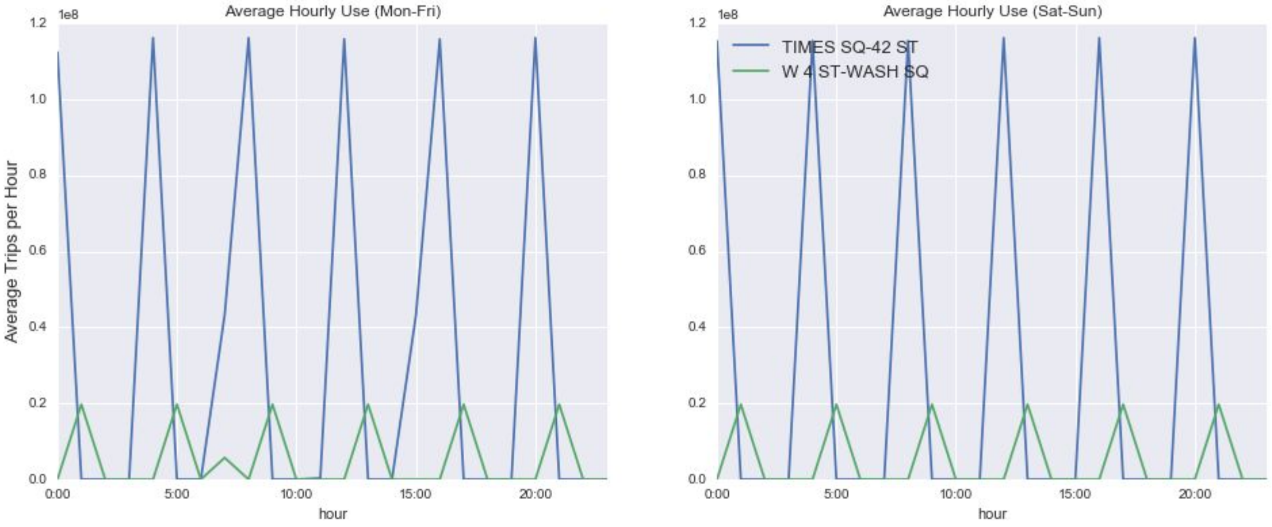
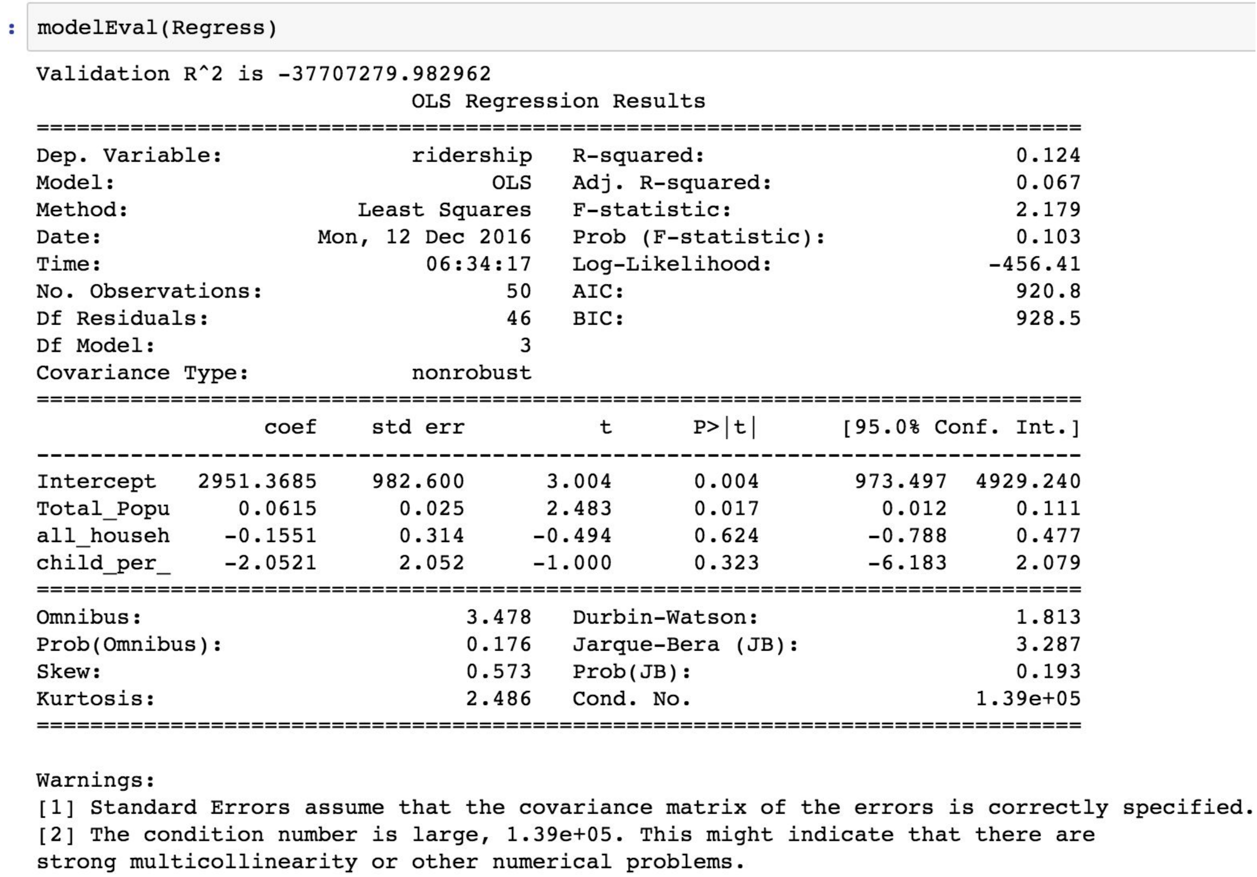


Figure 7 - Average hourly subway use (exits) of Weekdays and Weekends (September 2016).



```
: modelEval(Regress)

Validation R^2 is -37707279.982962
                            OLS Regression Results
==============================================================================
Dep. Variable:              ridership   R-squared:                       0.124
Model:                            OLS   Adj. R-squared:                  0.067
Method:                 Least Squares   F-statistic:                     2.179
Date:                Mon, 12 Dec 2016   Prob (F-statistic):              0.103
Time:                        06:34:17   Log-Likelihood:                -456.41
No. Observations:                  50   AIC:                             920.8
Df Residuals:                      46   BIC:                             928.5
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept   2951.3685    982.600      3.004      0.004       973.497  4929.240
Total_Popu     0.0615      0.025      2.483      0.017         0.012     0.111
all_househ    -0.1551      0.314     -0.494      0.624        -0.788     0.477
child_per_    -2.0521      2.052     -1.000      0.323        -6.183     2.079
==============================================================================
Omnibus:                        3.478   Durbin-Watson:                   1.813
Prob(Omnibus):                  0.176   Jarque-Bera (JB):                3.287
Skew:                           0.573   Prob(JB):                        0.193
Kurtosis:                       2.486   Cond. No.                     1.39e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.39e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 8 - Multivariate regression model, Average Ridership from starting Citi Bike Station ~ neighborhood population, household units and population of children

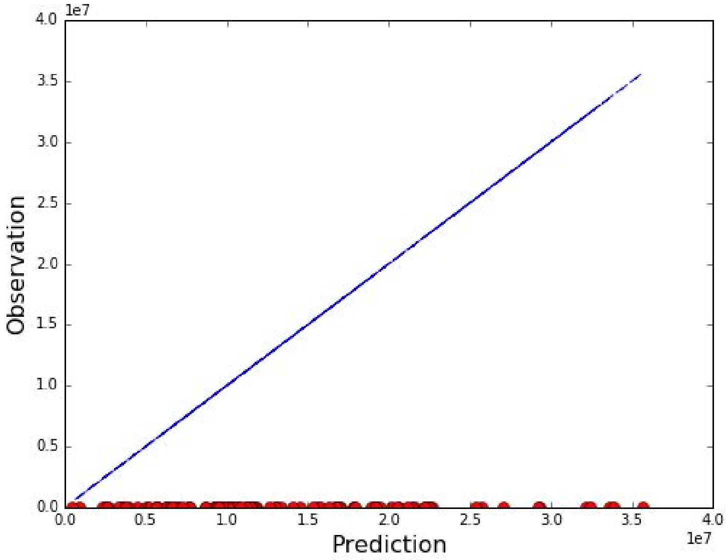
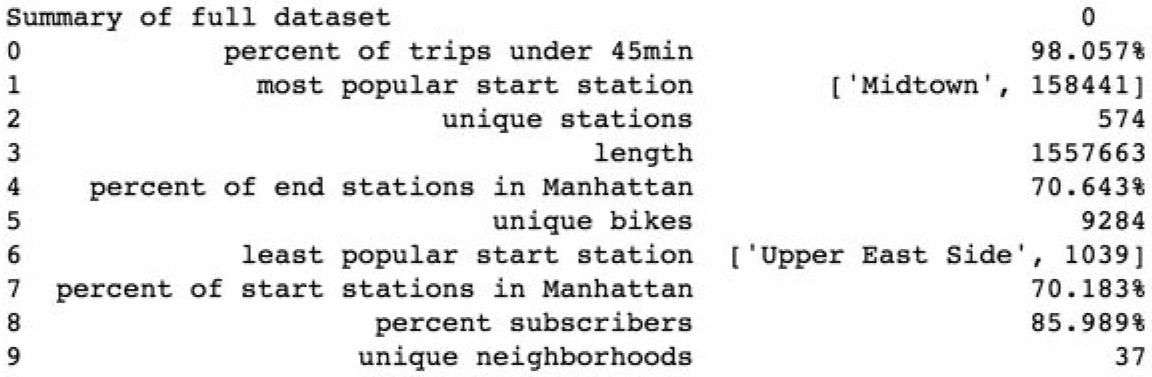


Figure 9 - Observed versus Predicted data of the multivariate regression model, Average Ridership from starting Citi Bike Station ~ neighborhood population, household units and population of children



```
Summary of full dataset                                            0
0             percent of trips under 45min                   98.057%
1               most popular start station       ['Midtown', 158441]
2                          unique stations                       574
3                                   length                   1557663
4    percent of end stations in Manhattan                    70.643%
5                             unique bikes                      9284
6              least popular start station  ['Upper East Side', 1039]
7  percent of start stations in Manhattan                    70.183%
8                      percent subscribers                   85.989%
9                     unique neighborhoods                        37
```

Figure 10 – Summary of Citi Bike dataset

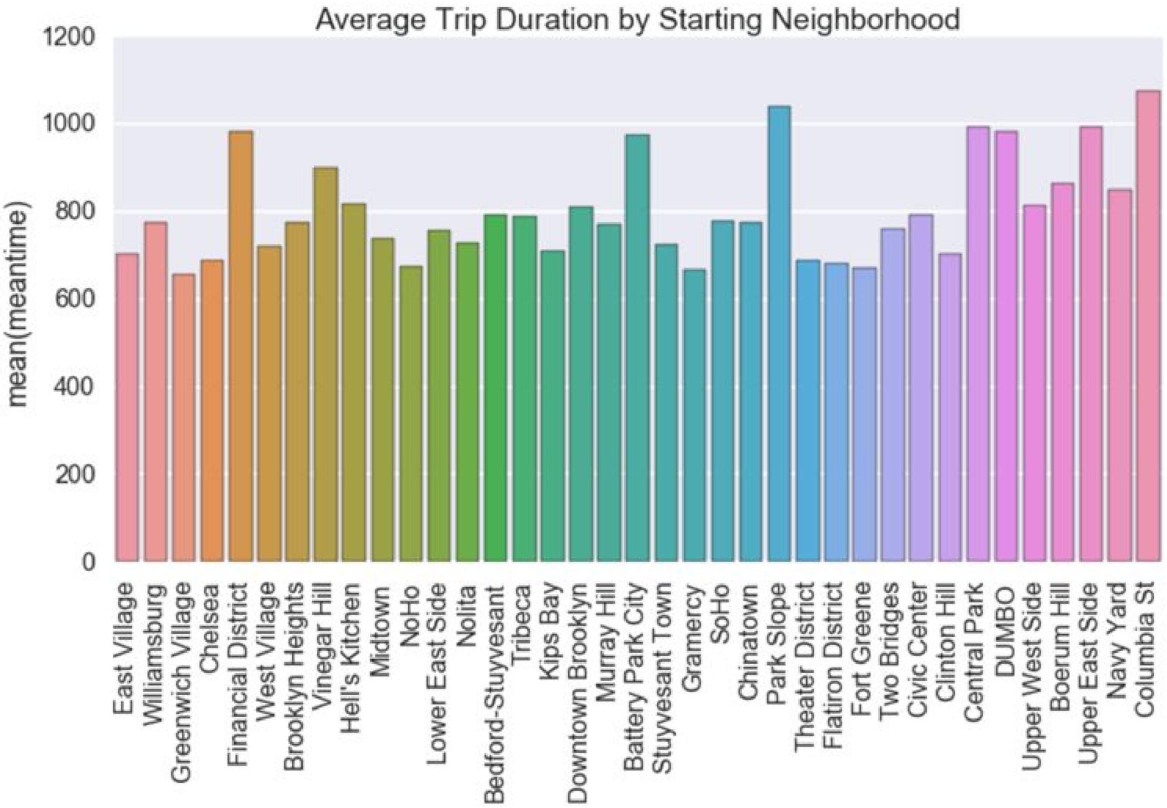


Figure 11 – Summary of Citi Bike dataset

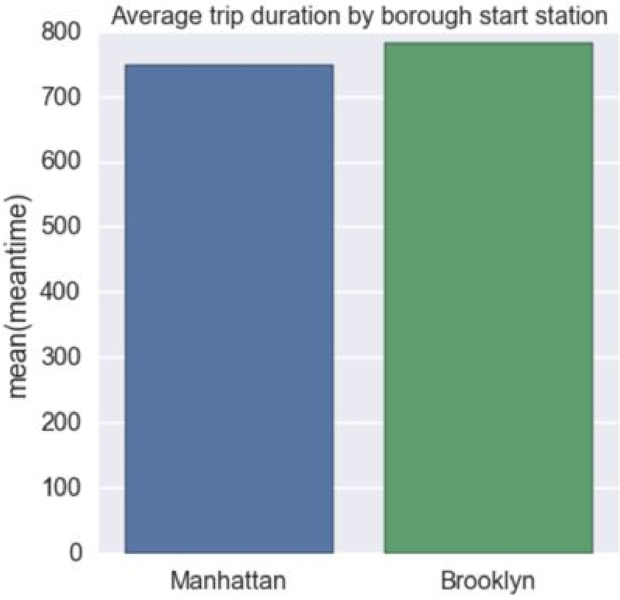


Figure 12 – Average trip duration by start station - Neighborhood and Borough

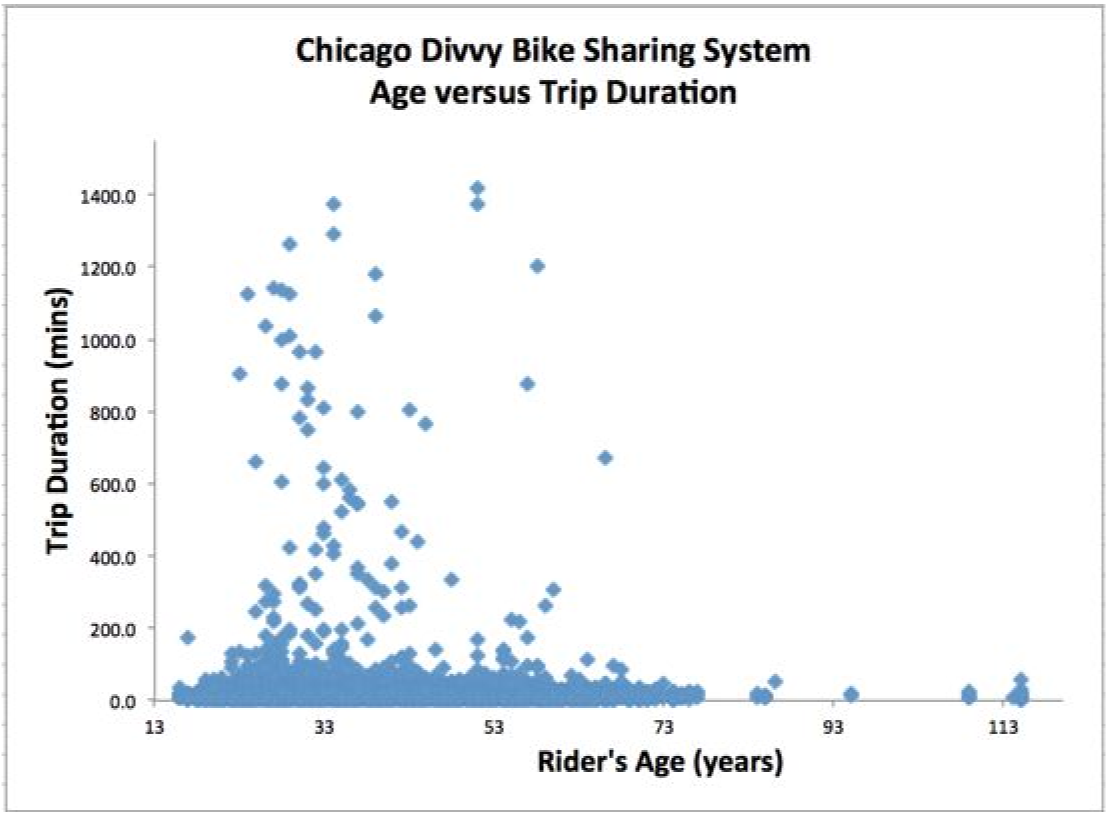


Figure 13 - Plot to understand correlation between Age and Trip duration.

Sample Trips data for Annual Subscribers for date range January to June 2015.

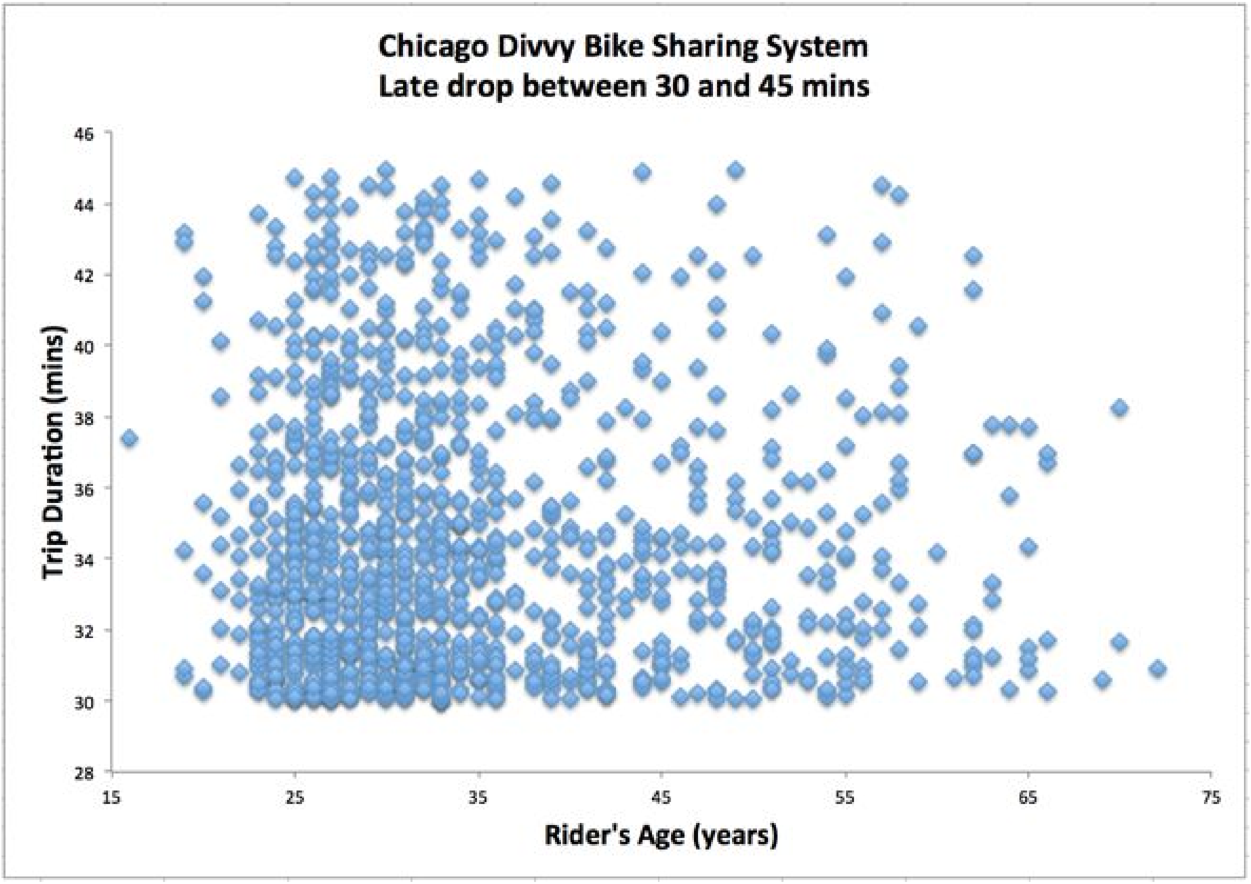


Figure 14 - Plot to understand correlation between Riders Age and Late drop of Bikes between 30 to 45 mins. Sample Trips data for Annual Subscribers and for date range January to June 2015.

As seen here, the delays occur mainly between age group 23 and 36 years for a delay seen here as approximately 6 mins.

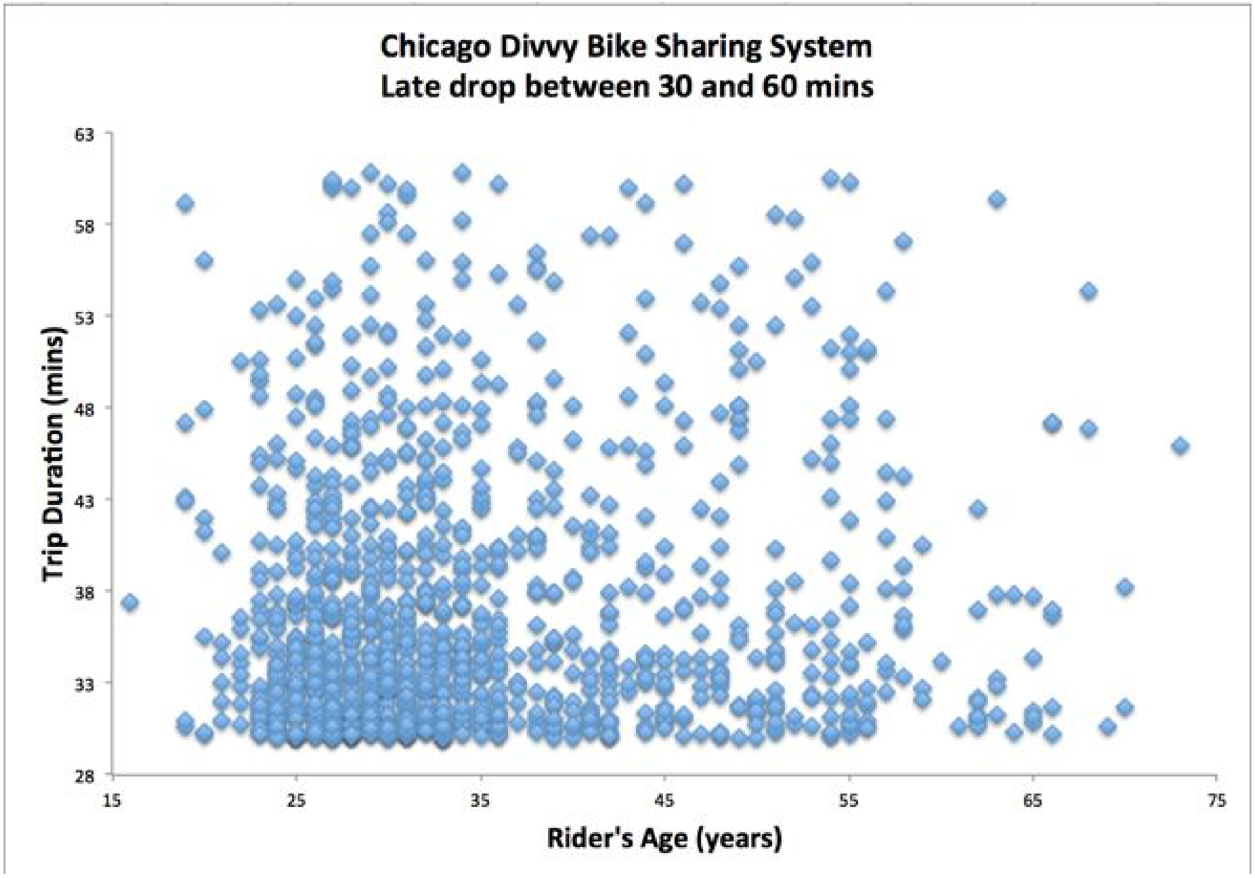


Figure 15 - Plot to understand correlation between Riders Age and Late drop of Bikes between 30 to 60 mins. Sample Trips data for Annual Subscribers and for date range January to June 2015.

Divvy charges additional $1.50 if bike is dropped between 30 and 60 mins.

As seen here, since riders still drop their bikes within the first few mins past due time, it indicates delay in bike drop.

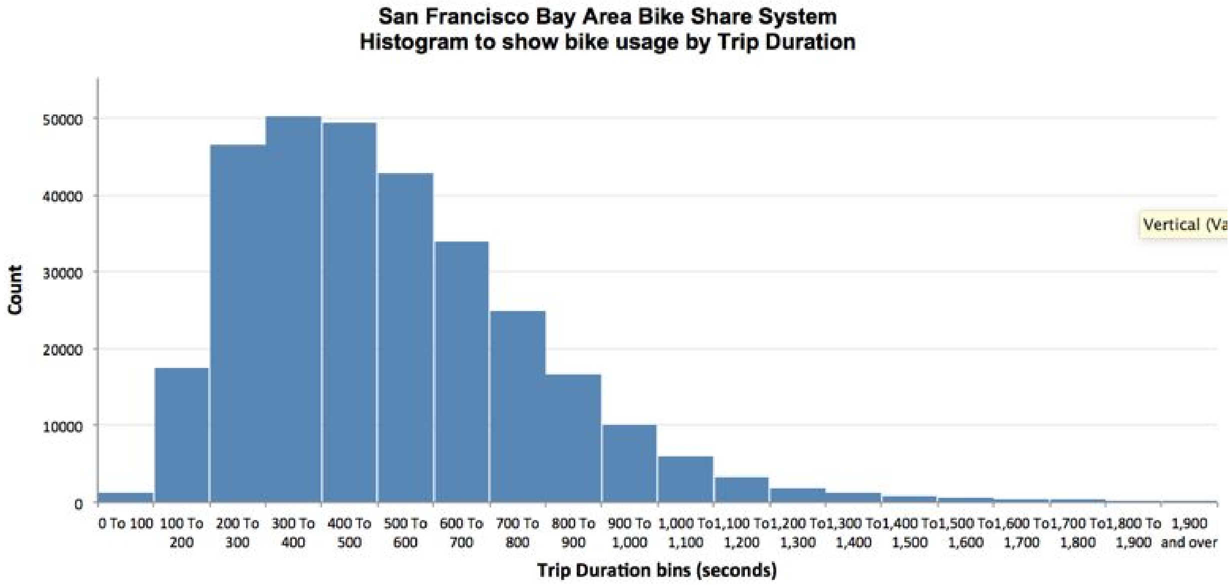


Figure 16 - Histogram to understand the rider's trip duration trend for San Francisco Bay Area bike sharing system riders.

References

Data -

a. BetaNYC, Bike Share Data best practices (link) (https://github.com/BetaNYC/Bike-Share-Data-Best-Practices/wiki/Bike-Share-Data-Systems)

b. Bay Area Bike Share, Open Data (link) (http://www.bayareabikeshare.com/open-data)

c. Chicago Bike Share, Divvy (link) (https://www.divvybikes.com/system-data)

d. NYC Bike Share, Citibike (link) (https://www.citibikenyc.com/system-data)

e. MTA turnstile data (link) (http://web.mta.info/developers/turnstile.html)

Research -

a. Smarter Tools For (Citi)Bike Sharing (link) (https://ecommons.cornell.edu/handle/1813/40922)

b. A Tale of 22 Million Citi Bike Rides: Analyzing the NYC Bike Share System (link) (http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/)

Project Effort Details -

a. Zhaohong Niu conducted the analysis of *Identify proximity to subway stations* in Methods & Data section
b. Sunny Kulkarni did the study for Existing Research and analyzed the bike sharing systems trip data for
    i. Chicago's Divvy bikes
    ii. Minneapolis Nice Ride bikes and
    iii. San Francisco's Bay Area bikes.
c. Chenxi conducted research of *Neighborhood Area Analysis* in Methods & Data section