NYC Citi Bike Visualization Analysis

15 May 2019

Hongdou Li

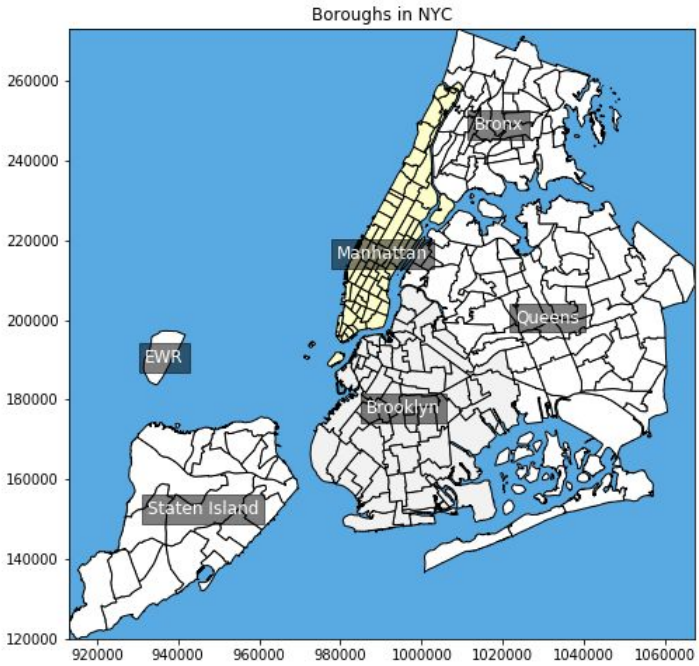
Introduction

My data set is the Citi Bike Trip History data in New York from March to August 2018. The dataset contains information of Citi Bike Usage, including Trip Duration, Start Time and Date, Stop Time and Date, Start Station Name, End Station Name, Station, ID, Station Lat/Long, Bike ID, User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member), Gender (Zero=unknown; 1=male; 2=female), Year of Birth of users.

This dataset can be accessed from CitiBike open data. Citi Bike is a privately owned public bicycle sharing system serving New York City. I am particularly interested in this dataset because I am curious about people's transportation habits in a metropolitan like New York. Questions I'd like to answer include Where do Citi Bikers ride? When do they ride? How far do they go? Which stations are most popular?

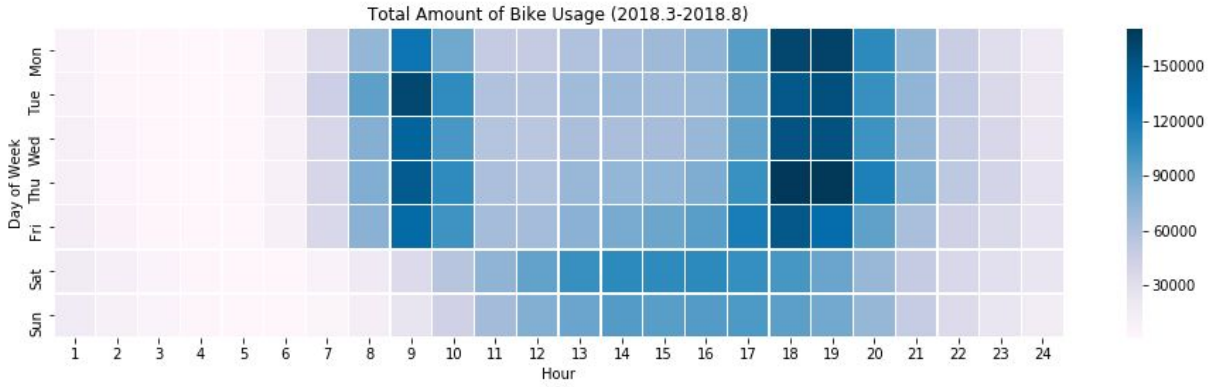
Summary of Data

I. Chloropleth Map



First, let's look at the general distribution of CitiBike Usage from March to August. Above is a map of new york city with different Boroughs. the Color of Boroughs indicates the total usage amount of that borough during 6 months period. We can tell from the graph that most of the usages happen in Manhattan and some in Brooklyn. The usage in other boroughs is neglectable.

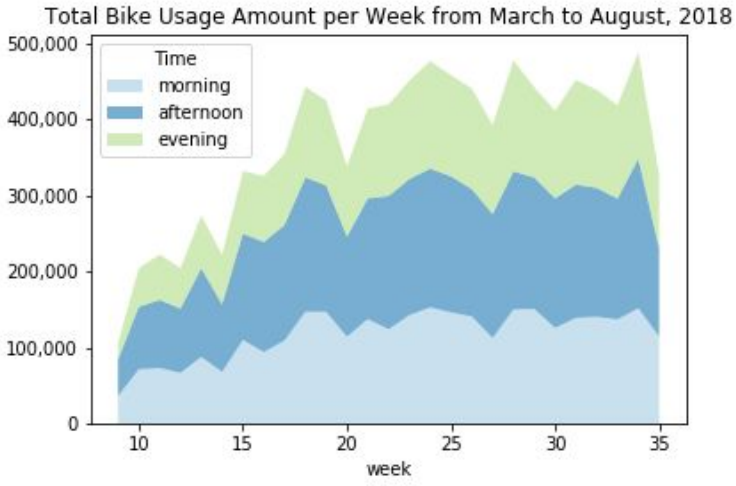
II. Heat Map



Above is a heatmap of Citibike usage amount at each hour of each day. It is interesting to point out that usage habits are quite different at weekdays and weekends. At weekdays, the peak hours appear at 7-10 am and 5-7 pm, which indicates that majority users choose Citi bike as a vehicle to go to work or go home after work. The peak hour during the weekend lags at around 11 am. Maybe some users will choose to stay up late and take the bike to explore the city.

It is also worth mention that the usage amount at Wednesday is relatively lower than other weekdays, A possible explanation might be many people choose to work from home on Wednesday.

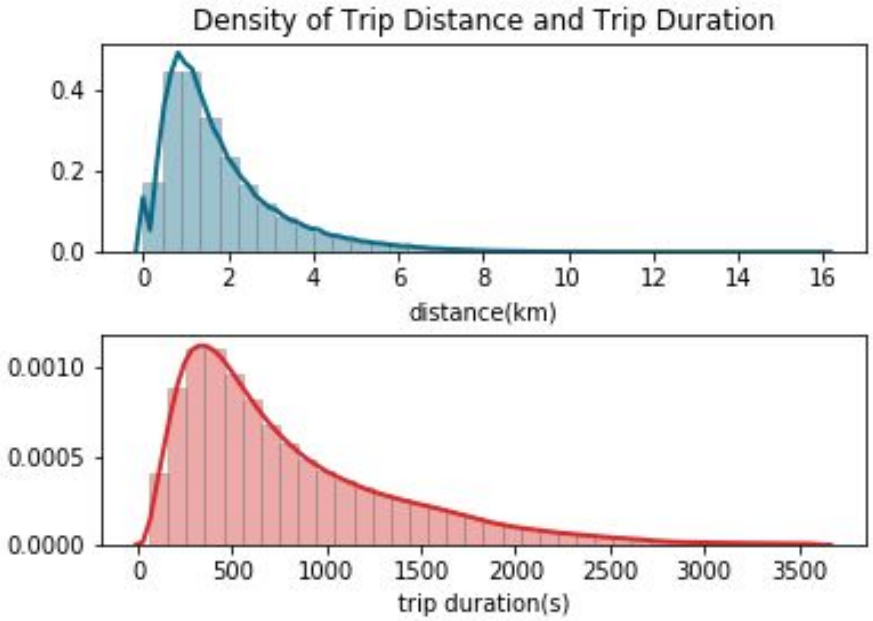
III. Stacked Graph



Above is a stacked graph of CitiBike usage in each week for 6 months. It starts with week 9 and ends with week 34. It is interesting to notice that with the weather become warmer, especially at week 15, which is the end of April, the usage of CitiBike increases.

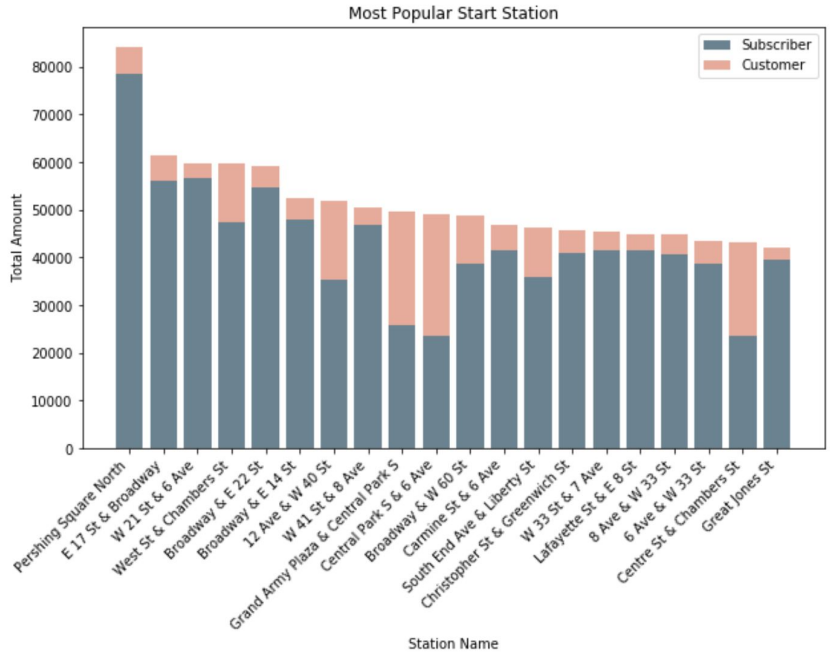
I also divide the usage into the morning (before 12 pm), Afternoon(before 18 pm) and evening(after 18 pm). There is a noticeable increase in evening usage with warm weather.

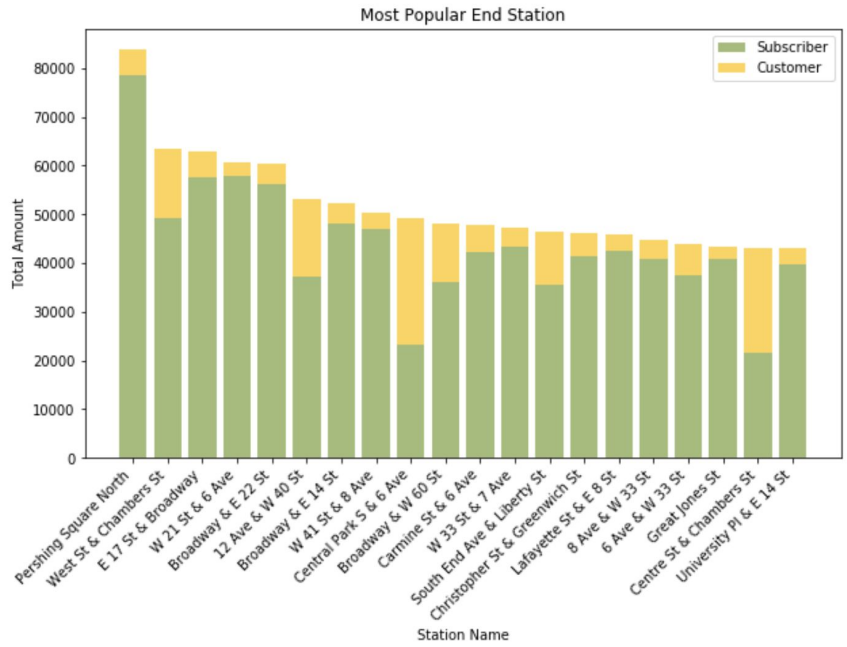
IV. Histogram



Above is Histograms showing the distribution of trip duration and trip distance. As is shown in the graph above, most of the trips are about 1km and last about 480s, which is 8 minutes. Thus, it may indicate that the main purpose of using Citybike is for short distance trips.

V. Bar Plot

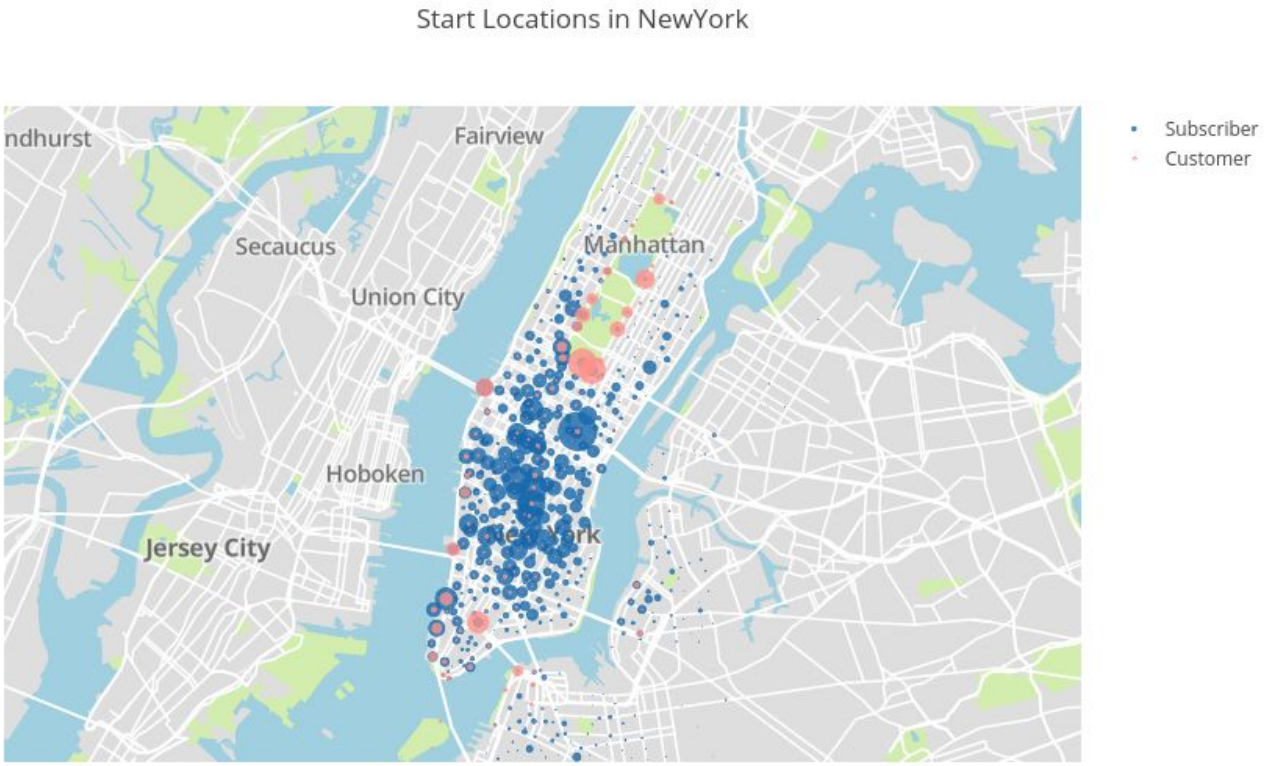




As for which are the most popular stations, the above two bar plots gives the answer. As you can see, there are overlaps between popular start stations and end stations. The most popular station is Pershing Square North. This result is not surprising because this is the closest station near Grand Central Terminal.

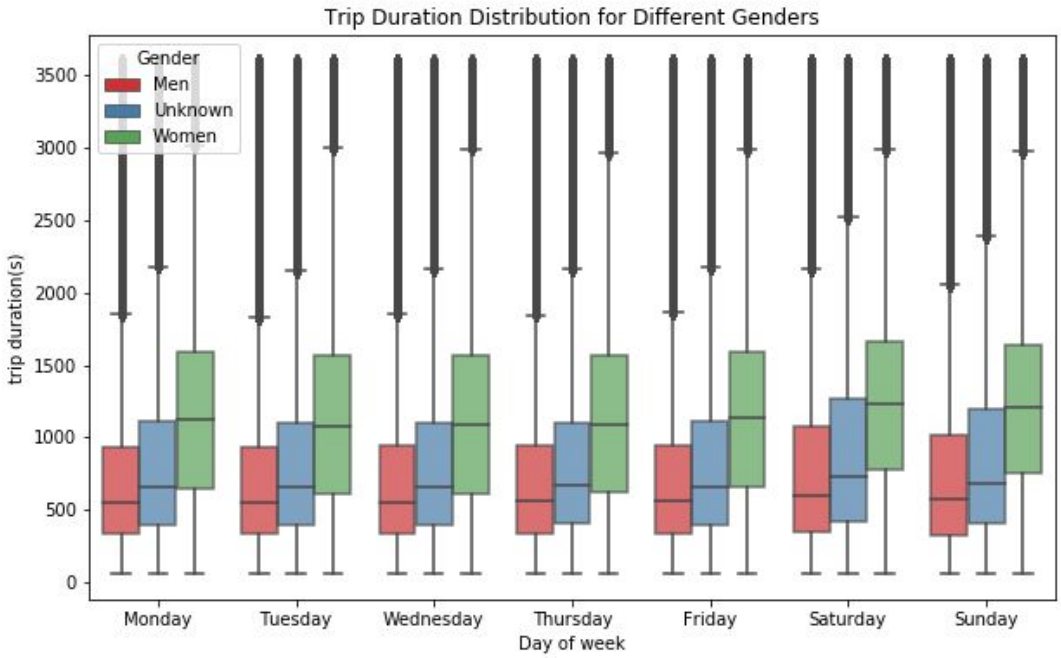
It is also worth mention that among those popular stations, most of the users are subscribers, which means that they bought a year pass. For several stations near central park, however, the ratio of the customers (who uses daily pass) and subscribers is almost 1. A possible explanation is that users in these stations are mostly visitors who come to New York for tourists and short-term vacation. For other stations, the users are people who work nearby.

VI. Bubble Map



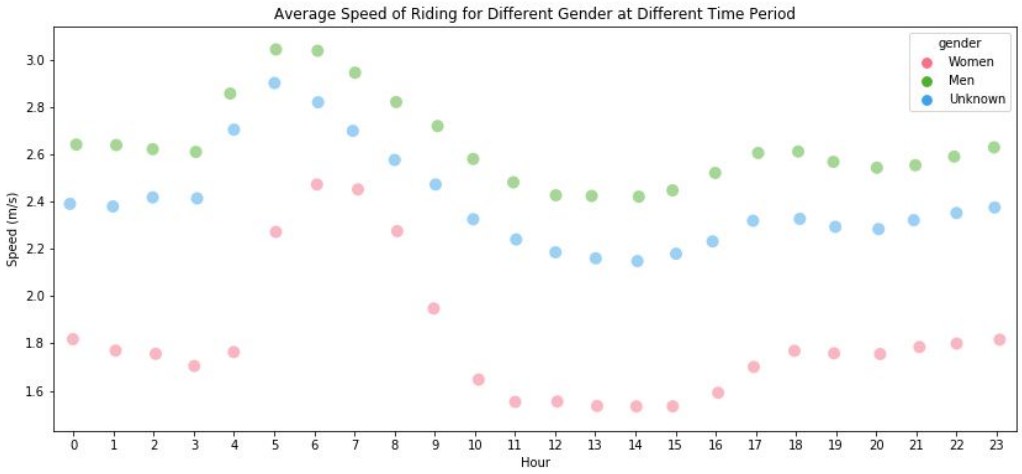
Following the question Above, I draw a bubble map showing the usage habits of different stations. Pink bubbles indicate customers while blue bubbles indicate subscribers. The size of bubbles showing the total amount of usage in this station. Just as we assumed, customer users gathered around central park, while Subscribers are mainly in downtown.

VII. Boxplot



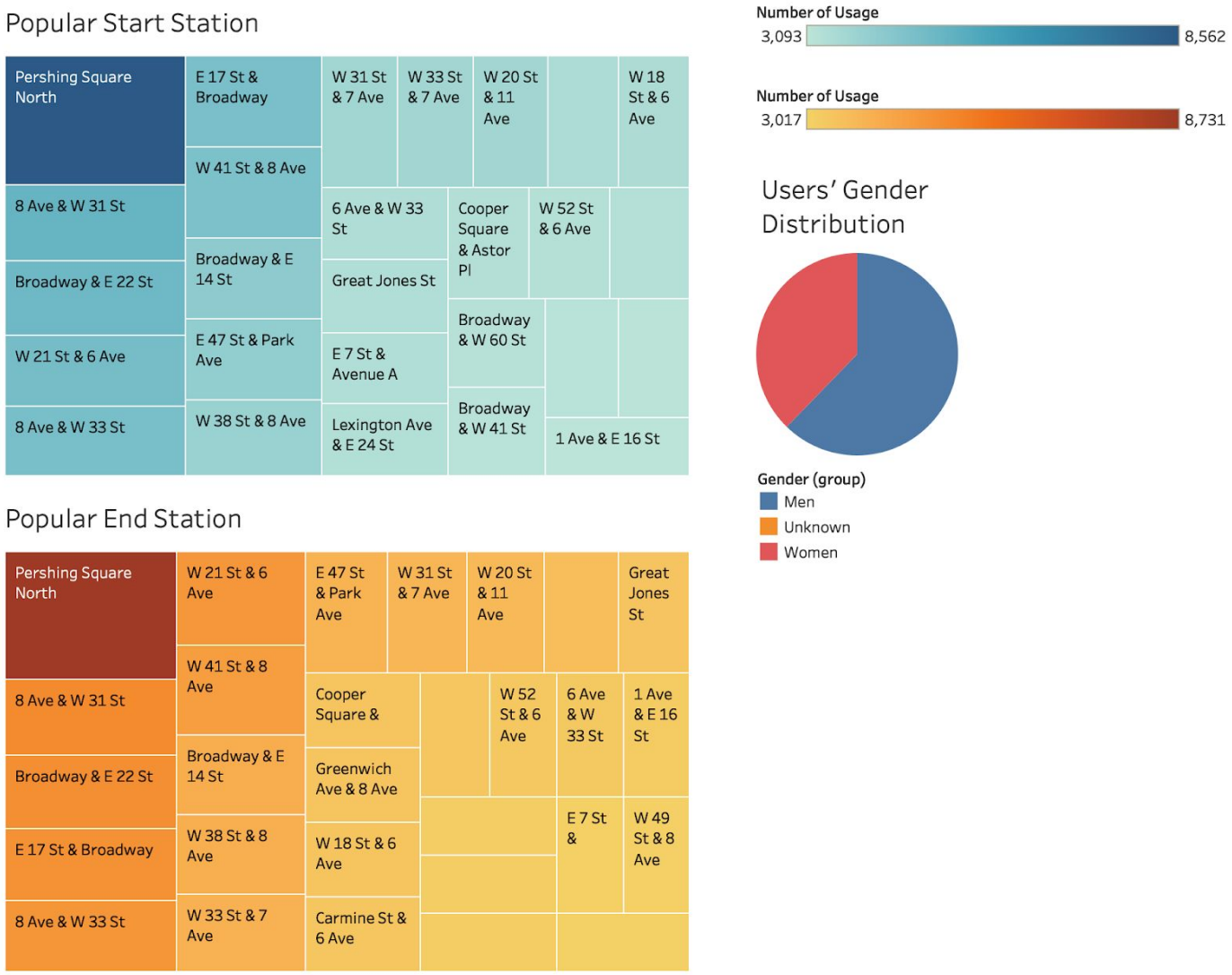
Above is the box plot showing the distribution of Trip duration for different genders. In general, Female users have longer trip compared with male users. Meanwhile, Although the distribution is quite stable for each day of the week. We can still notice slightly longer trips at weekends.

VIII. Scatterplot



Above is a scatter plot showing the average speed of users at different time periods of a day. We can see that the speed peaks at 6 am when there is not much traffic on the road. In addition, female riders seem to ride slower with more caution than male riders.

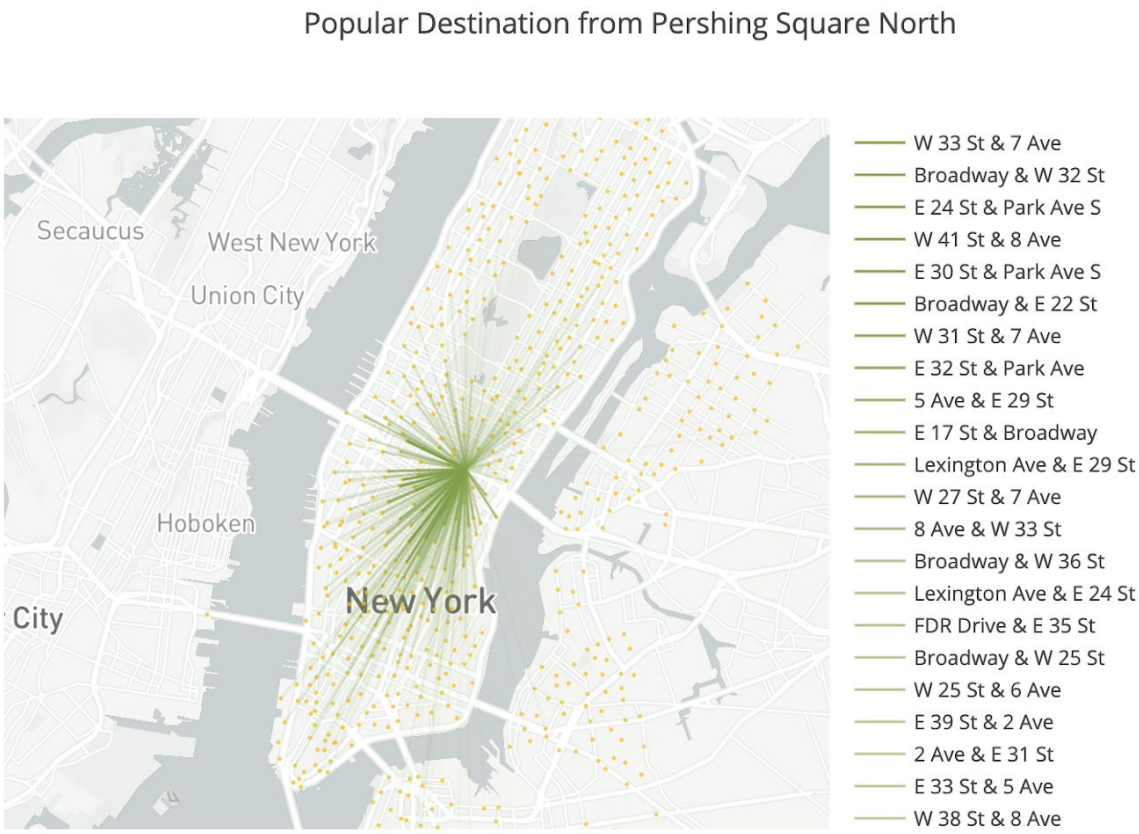
IX. Treemapping(Interactive plot)



Above is a treemapping showing the most popular start station and end station. The color indicates the total amount of usages. As we showed before, Pershing Square North is the most popular station due to its location.

This Interactive plot can show the station popularity between male and female users. By clicking the part chart at the right. For both male and female users, the most frequently visited station is Pershing Square North.

X. Connection Map(Interactive plot)



Since Pershing Square North is the most popular station, I'm curious about what are users' destination from Pershing Square North. As is shown in the plot above, The most frequent destination is W33 st & 7 Ave (Where is the Manhattan Mid Town). Some popular destinations are listed at the right, ordered by its popularity. Also, the greener the line, the more users choose this route.

This is also a interactive plot and users can toggle to the route from Pershing Square North to the destination by clicking the name.

Conclusion

In summary, we can answer the questions we have previously. Citibike users concentrated at Manhattan. The usage peaks at rush hours when people go to work or go home from work. The major purpose of users choosing Citibike is for short trips from work to major transition station.

Appendix

Please see all the code in Github.

GitHub

https://github.com/helenali323/NYC_Citi_Bike.git

Citations

1. Seaborn
2. Matplotlib
3. Plot.ly