Springboard—Intermediate Data Science with Python

Final Report for Capstone Project – Bike Share Advertising in New York City

Andrew Harris

September 2019

Table of Contents

List of Figures

List of Tables

1.0 Statement of Problem

Citi Bike is a public bicycle sharing system operating in New York City. Citi Bike have over 800 bicycle sharing stations located throughout the city which have space for small billboard advertising. Citi Bike wishes to profile their users so that they can determine the value of their advertising space and identify the products and services which should be marketed at a given station.

Citi Bike has collected data for every trip taken from July 2013 onward. The data contains information such as the start time, end time, start station and end station. The data also contains rider specific information such as the gender, birth year and whether or the not the rider is a subscriber (Annual Member) or customer (24-hour pass or 3-day pass user).

This trip data can be segmented based on rider type (customer vs. subscriber), gender, age, start station and end station so that the station user demographics can be determined for individual stations and used to support marketing activities.

2.0 Description of Client

The client for this problem is Citi Bike itself. Citi Bike has over 800 active bicycle sharing stations located throughout New York City. These stations currently have payment machines which have small billboards on the sides perpendicular to the payment machine user interface. One of the two billboards is typically a map of the surrounding area while the other billboard is an advertisement for Citi Bike itself or for services offered by Citi Bank.

The segmentation described in the "Statement of Problem" section will allow the client to determine the user demographic for each station. Citi Bike could use this information in two ways:

1. Banking services offered by Citi Bank which are geared toward specific demographics can be advertised at stations which are most frequently used by that demographic.
2. If Citi Bike decided to lease the advertising space to a third-party organization, they would have data supporting the user demographic of each station. User demographic data would allow Citi Bike to determine the value of the advertising space and identify the organizations to which they should market the advertising space.

3.0 Data Set

The source of data for these models is the Citi Bike website at https://www.citibikenyc.com/system-data. The data set itself is called "tripdata" and is reported on a monthly basis with a .csv file for each month. Each line in the data is an individual trip with 15 attributes. The 15 attributes are as follows:

- start time, start station id, start station name, start station latitude and start station longitude.
- end time, end station id, end station name, end station latitude and end station longitude.
- trip duration, bike id, user type, birth year and gender.

Citi Bike has processed the data to remove any trips that were taken by staff as part of service and inspection activities or were taken between "test" stations. Citi Bike has also processed the data to remove any trips below 60 seconds in length as such trips are potentially false starts or users attempting

to re-dock a bike to verify that it is secure. Thus, the trips present in the data set are actual paid trips taken by users.

Six months of data (January 2019 to June 2019) containing over nine million trips taken between over 800 bicycle stations will be used for the analysis.

4.0 Data Wrangling

The six months of data (January 2019 to June 2019) were downloaded from the Citi Bike website as .csv files and uploaded into Jupyter Notebook as Pandas data frames. The six data frames were appended to create one data frame with 9,054,981 rows. The 'starttime' and 'stoptime' columns were converted to datetime format and the data frame was sorted in ascending order based on 'starttime'.

Prior to performing a column by column assessment of the data, a search for NaN entries was performed. 37 rows containing NaN entries were identified. In each of these rows, the 'start station id', 'start station name', 'end station id' and 'end station name' all contained NaN. These 37 rows were deleted immediately as there was no way to determine the start location or end location of these trips.

Next, a column by column assessment was performed in the following order:

'bikeid' Column Assessment:

This column represents the ID number for the bike that was used for the trip. A count of the number of unique values in this column found that there were 16969 ID numbers which is not an unreasonable figure.

'usertype' Column Assessment:

Citi Bike has two types of users: "Subscriber" who are users with annual memberships and "Customer" who are users with short term passes such as 24-hour or 3-day durations. A count of the number of occurrences of each value in this column confirmed that all entries were either "Subscriber" or "Customer".

'gender' Column Assessment:

Citi Bike has three categories for 'gender': 0 for unknown, 1 for male and 2 for female. A count of the number of occurrences of each value in this column confirmed that all entries were either 0, 1 or 2.

'birth year' Column Assessment:

This column reports the year in which the rider taking the trip was born. The minimum reported 'birth year' was 1857. This 'birth year' is clearly a false entry as this person would be 161 years old. The minimum 'birth year' was 2003 indicating that the youngest user was 16 years old which is in alignment with the minimum age policy on the Citi Bike website. As life expectancy in USA is around 79, all trips with a 'birth year' before 1939 were removed so that the maximum age of the riders was 80. This deletion removed 11,526 rows from the data set.

'tripduration' Column Assessment:

This column reports the duration of the trip in seconds. The 'tripduration' values were compared to the calculated difference between the 'stoptime' and 'starttime' values to see if the results matched. There were 111 rows where the difference between 'stoptime' and 'starttime' was 3600 seconds greater than the 'tripduration' value. These discrepancies were corrected by setting the 'tripduration' column equal to the difference between the 'stoptime' column and 'starttime' column for the entire data set.

The next task was to assess the maximum trip duration which should be retained. The Citi Bike website contains the following statement:

"Lost Bike Charge - If you have kept a bike out for more than 24 hours at a time, it is considered lost or stolen and there is a fee of $1200 (+ tax). This can also happen if you did not dock your bike properly, so your ride stayed open, and someone else took the bike and has not returned it."

Based on this "Lost Bike Charge" policy and the docking issue described, all durations greater than 24 hours were removed. This deletion removed 3228 rows from the data set.

In the remaining 'tripduration' data, it was found that 80% of all trips are 20 minutes or less and 99% of all trips are 57.5 minutes or less.

The Citi Bike website contains the following information regarding allowable trip durations and late charges:

- For 24-hour and 3-day pass holders, 30 minute rides are allowed. If a ride exceeds 30 minutes, late charges of $4.00 for each additional 15 minutes are applied.
- For the annual subscribers, 45 minute rides are allowed. If a ride exceeds 45 minutes, then late charges of $2.50 for each additional 15 minutes are applied.

Based on this fee structure, a bike trip with a two hour duration would result in $12.50 and $20.00 in late charges for "Subscriber" and "Customer" user types respectively. Due to the availability of bikes and the close proximity of stations to each other, the user is better off swapping bikes to initiate a new trip rather than incurring substantial late charges.

As the reported duration of the trip increases, the likelihood that the bike was incorrectly docked and then used by a second user increases. Thus, it is suspected that many trips beyond one hour are the result of improperly docked bikes being taken by a second user and then returned to a different station.

All trips with a duration in excess of one hour were removed. This deletion removed 81,627 rows from the data set.

'starttime' and 'stoptime' Column Assessment:

All 'starttime' values were found to be within the six month time frame that the data is supposed to cover (January 1, 2019, 00:00:00 to June 30, 2019, 23:59:59). First and last 'starttime' are very close to the start and end of the range that the data set is supposed to cover. When assessing the 'stoptime' column, it was observed that some of the trips have a 'stoptime' outside of the six month range.

For simplicity in describing the dataset, trips that end outside of the six month range were deleted. The dataset can be described as "all trips less than or equal to one hour starting and ending between January 1, 2019 and June 30, 2019 inclusive". This deletion resulted in the removal of 248 rows.

'start station id', 'start station name', 'start station latitude', 'start station longitude' Column Assessments:

These four columns were assessed to ensure that a given ‘start station id’ corresponded to a single ‘start station name’ and a single set of ‘start station latitude’ and ‘start station longitude’ coordinates. The number of unique values in each of these columns was counted. There were 820 'start station id' values. However, there were 821 'start station name' values and 822 'start station latitude' and 'start station longitude' values.

The unique combinations of 'start station id', 'start station name', 'start station latitude' and 'start station longitude' were isolated into a separate data frame which was found to contain 825 rows. The repeated values in these four columns were analyzed. Station ID 243.0 and 3727.0 were found to have two sets of 'start station latitude' and 'start station longitude' coordinates. Two new station ID numbers 4444 and 5555 were created. Station ID 4444 was assigned to the rows containing one of the two sets of coordinates for 243.0. Station ID 5555 was assigned to the rows containing one of the two sets of coordinates for 3727.0. It is now the case that all unique ‘start station id’ numbers have only one set of latitude and longitude coordinates.

'end station id', 'end station name', 'end station latitude', 'end station longitude' Column Assessments:

These four columns were assessed to ensure that a given ‘end station id’ corresponded to a single ‘end station name’ and a single set of ‘end station latitude’ and ‘end station longitude’ coordinates. The number of unique values in each of these columns was counted. There were 837 'start station id' values. However, there were 838 'start station name' values and 839 'start station latitude' and 'start station longitude' values.

The unique combinations of 'end station id', 'end station name', 'start station latitude' and 'start station longitude' were isolated into a separate data frame which was found to contain 842 rows. The repeated values in these four columns were analyzed. Station ID 243.0 and 3727.0 were found to have two sets of 'end station latitude' and 'end station longitude' coordinates. This is the same issue that was discovered when assessing the start station information. The station ID numbers 4444 and 5555 that were used to correct this issue.

The list of unique end stations was then compared to the list of unique start stations to identify stations that were present in one list but not in the other. All 822 ‘start station id’ values were found to be present in the list of 839 end station ids. The 17 ‘end station id’ values which were not found in the ‘start station id’ list were searched in the Citi Bike 'Find a station' search engine (https://member.citibikenyc.com/map/). One of these stations was found to be in Brooklyn, another one didn’t return a result and the remaining 15 stations were located on the west side of the Hudson River in New Jersey. The number of trips ending at these 17 stations was only 31. These stations were deleted as it is likely that they are missing trip data due to the low trip volume reported and the fact that no trips start at these stations. This deletion removed 31 rows of data.

5.0 Data Visualization and Descriptive Statistics

5.1 Station Volume Assessment

The number of trips starting and ending at each station were summed to get the ‘total volume’.

The following summary statistics were calculated for ‘total volume’:

- Minimum 'total volume' is 10.
- Maximum 'total volume' is 141,192.
- Median 'total volume' is 14,408.
- Mean 'total volume' is 21,796.

The following percentile statistics were calculated for ‘total volume’:

- 25% of the stations (205 stations) had a 'total volume' that is less than or equal to 6578.
- The top 25% of the stations (206 stations) had a 'total volume' that is greater than 32,147.
- The top 10% of the stations (83 stations) had a ‘total volume’ greater than 51,341.
- The top 1% of the stations (9 stations) had a total volume greater than 86,810.

These statistics indicate that station volume varies substantially for the 822 stations and that a small number of stations are responsible for a very large portion of the trip volume.

The latitude and longitude coordinates for all 822 stations were plotted on Google Maps using the gmap API. The 'total volume' percentile for each station is color coded as follows:

- The 25th percentile stations for ‘total volume’ are plotted as red circles.
- The stations bounded by the 25th and 75th percentile are plotted as yellow circles.
- The stations above the 75th percentile (Top 25% stations for 'total volume') are plotted as blue circles.

See Figure 1: Total Volume – All 822 Stations on page 9.

The following observations were made:

- The blue circles are located primarily in Midtown Manhattan and Lower Manhattan. There are 8 blue circles in the Upper West Side and 6 blue circles in the Upper East Side. There are no blue circles in Queen's and only 9 blue circles in Brooklyn.
- There are 18 yellow circles in Queen's. Brooklyn is largely comprised of yellow circles. The regions of Manhattan above Midtown Manhattan are largely comprised of yellow circles. There is a cluster of yellow circles in the east side of Midtown Manhattan and along the south coast of Lower Manhattan.
- There are less than 30 red circles on all of Manhattan Island and they are primarily concentrated in Harlem and East Harlem. The majority of all circles present in Queens are red. In Brooklyn, there are clusters of red circles typically representing the bike stations furthest from Manhattan.

From these observations, it is clear that the highest 'total volume' stations are primarily in Midtown and Lower Manhattan. 'The total volume' of the stations decreases as one moves north away from Midtown Manhattan or away from Manhattan through Queens or Brooklyn.

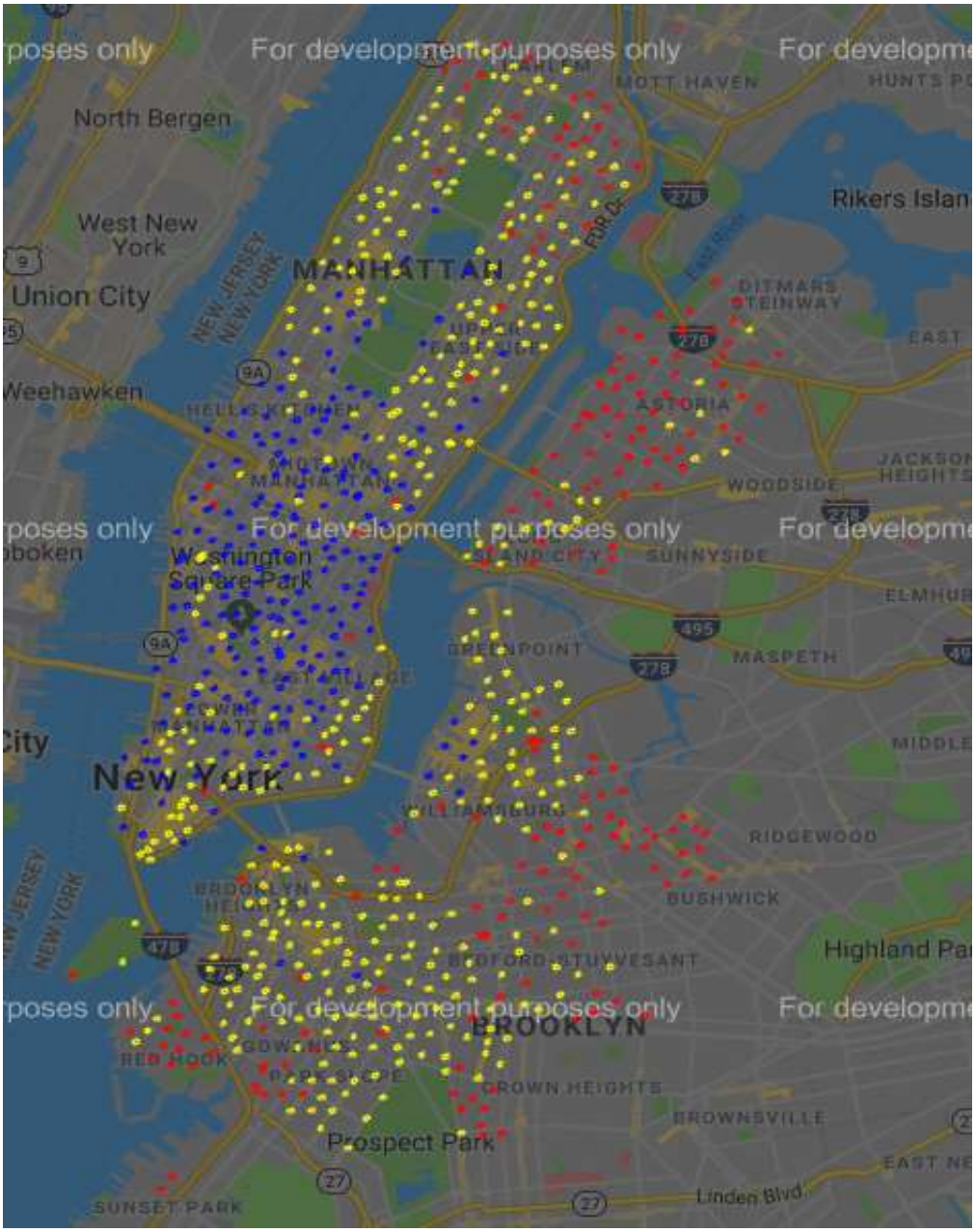

Figure 1: Total Volume – All 822 Stations

5.2 Trip Frequency Assessment

The frequency for each unique trip was calculated and reported as the 'total frequency' for that trip. For this calculation, trips between stations are grouped together based on having the same start station and the same end station. Thus, a trip starting at station "A" and ending at station "B" is not the same as a trip starting at station "B" and ending at station "A".

The following summary statistics were calculated for 'total frequency':

- The total number of unique trips taken over the six month time frame is 304,919.
- Minimum 'total frequency' is 1.
- Maximum 'total frequency' is 4040.
- Median 'total frequency' is 5.
- Mean 'total frequency' is 29.

The following percentile statistics were calculated for 'total frequency':

- 25% of the trips (~76,230 trips) were taken only 1 or 2 times.
- 50% of the trips (~152,460 trips) were taken anywhere from 3 to 22 times.
- 25% of the trips (~76230 trips) were taken 23 or more times.
- The top 1% of trips for 'frequency' (3049 trips) were taken more than 354 times.
- The top 0.1% of trips for 'frequency' (305 trips) were taken more than 935 times.

These statistics indicate that the frequency for individual trips varies substantially and that a small number of unique trips are responsible for a very large portion of the trips taken.

The 305 most frequently taken trips (Top 0.1%) were plotted on Google Maps using the gmap API. Each of these trips was taken in excess of 935 times. For the trips being plotted, the following color coding has been applied:

- Stations that are "start stations only" for the group of trips are plotted as pink circles.
- Stations that are "end stations only" for the group of trips are plotted as cyan circles.
- Stations that are both "end stations" and "start stations" are plotted as purple circles.
- A blue line is drawn from start station to end station to show where a trip begins and ends.
- Purple circles without any blue lines originating from them represent trips that started and ended at the same station.

See Figure 2: The 305 Most Frequently Taken Trips on page 11.

The following observations were made:

- The majority of these trips occur inside of Manhattan. There are a large number of trips within Midtown Manhattan and the East Village Region of Lower Manhattan.
- There are many trips inside of Central Park and along the waterfront from Battery Park to as far north as 495 bridge to New Jersey.
- The trips in Brooklyn tend to appear in small isolated groups.

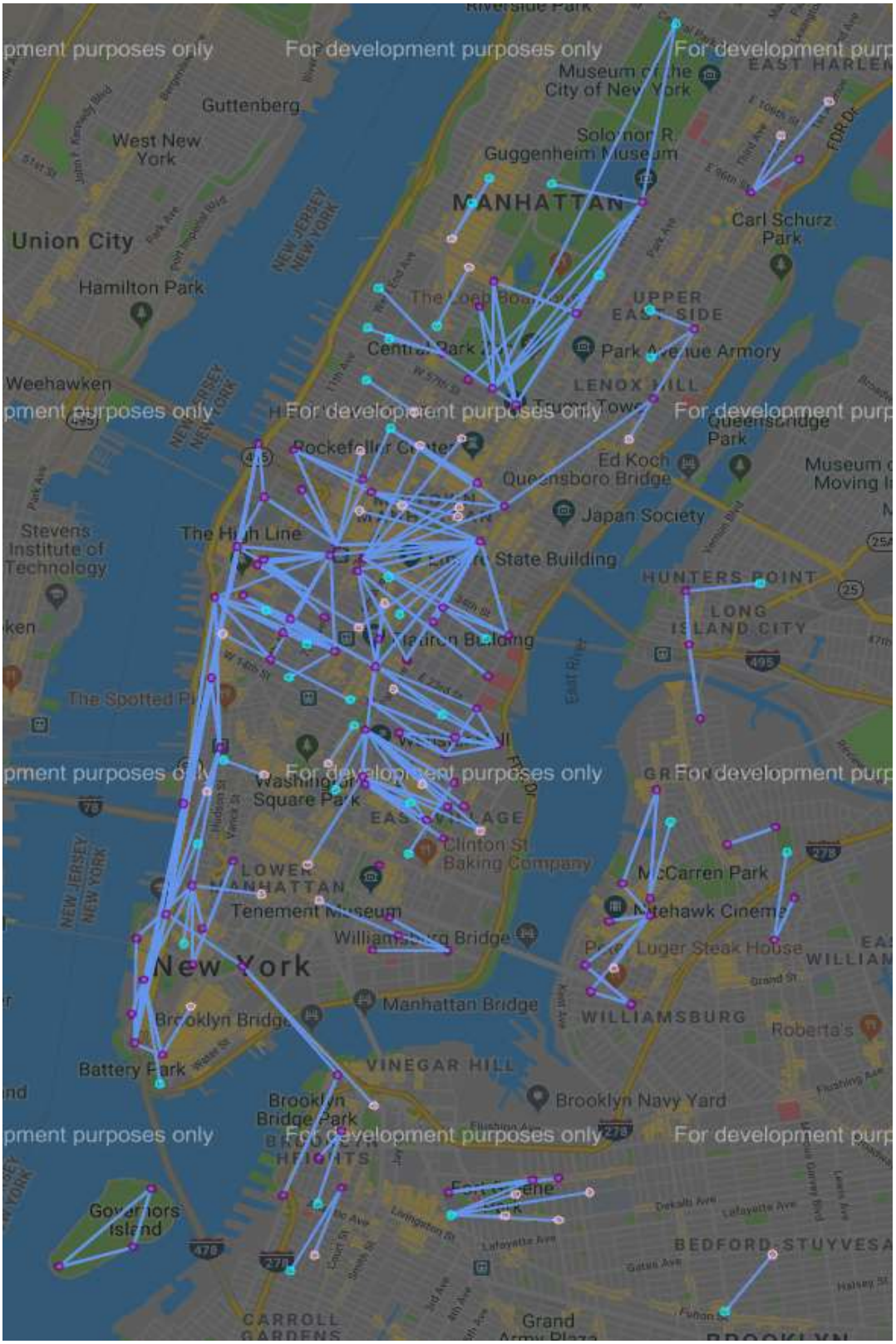

Figure 2: The 305 Most Frequently Taken Trips

5.3 Station Volume and Trip Frequency Assessments for Customers and Subscribers.

The data set was split into one data set consisting of the trips taken by the “Customer” user type and another data set consisting of the trips taken by the “Subscriber” user type. It was observed that 88% of all trips were taken by subscribers and 12% of all trips were taken by customers.

For both data sets, the number of trips starting and ending at each station were summed and reported as the ‘total volume’. A comparison of the top 100 stations for ‘total volume’ of customers and the top 100 stations for ‘total volume’ of subscribers shows that these two groups only have 42 stations in common.

The top 100 'total volume' stations of customers and the top 100 'total volume' stations of subscribers were plotted on Google Maps using the gmap API. The following color coding has been applied:

- The top 100 subscriber stations are plotted as sky blue circles.
- The top 100 customer stations are plotted as magenta circles.
- The stations in common between the top 100 subscriber stations and the top 100 customer stations are plotted as orange circles.

See Figure 3: Top 100 'total volume' Stations for Customers and Subscribers on page 13.

The following observations were made:

- The blue circles are concentrated primarily in Lower Manhattan and Midtown Manhattan.
- In Manhattan, pink circles are located primarily along the boundaries of Central Park and along the Hudson River.
- In Brooklyn, pink circles are located near the Brooklyn bridge, in the Williamsburg neighborhood and the upper boundary of Prospect Park.
- The orange circles for "common stations" are located primarily in Lower Manhattan and Midtown Manhattan.

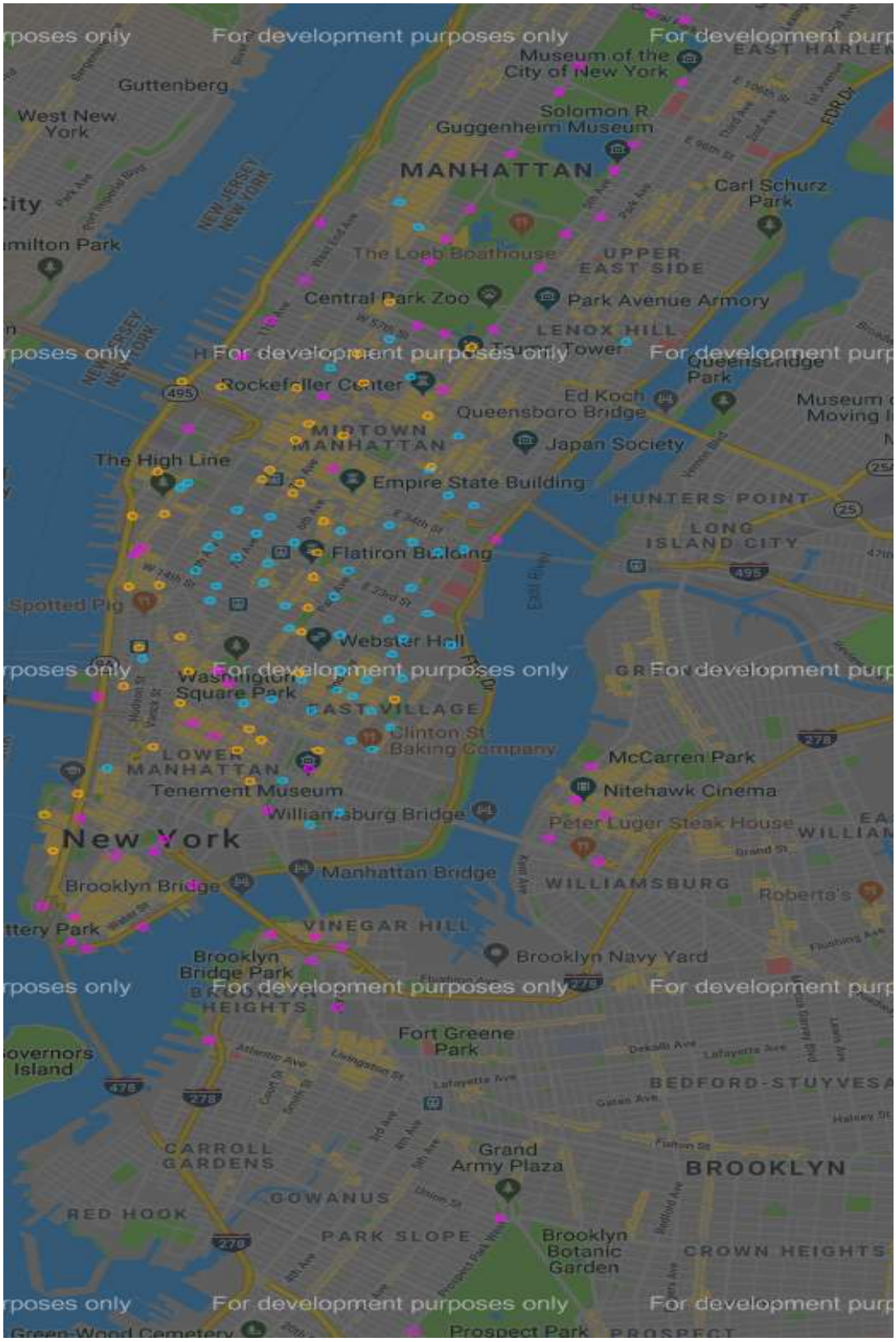

Figure 3: Top 100 'total volume' Stations for Customers and Subscribers

For both data sets, the frequency for each trip was calculated and reported as the 'total frequency' for that trip.

The top 200 most frequently taken customer trips were plotted on Google Maps using the gmap API. Each of these trips was taken in excess of 258 times. For the trips being plotted, the following color coding has been applied:

- Stations that are "start stations only" for the trips are plotted as pink circles.
- Stations that are "end stations only" for the trips are plotted as cyan circles.
- Stations that are both "end stations" and "start stations" are plotted as purple circles.
- A blue line is drawn from start station to end station to show where a trip begins and ends.
- Purple circles without any blue lines originating from them represent trips that started and ended at the same station.

See Figure 4: Top 200 Trips taken by Customers on page 15.

The following observations were made:

- The majority of the trips are trips across the Brooklyn Bridge, inside of Central Park and along the waterfront from Battery Park to as far north as the 495 bridge to New Jersey.
- Other trips include the Williamsburg Bridge, Prospect Park and Governors Island.
- There are no trips across Lower Manhattan or Midtown Manhattan.

The top 200 most frequently taken subscriber trips were plotted on Google Maps using the gmap API. Each of these trips was taken in excess of 967 times.

See Figure 5: Top 200 Trips taken by Subscribers on page 16.

The following observations were made:

- The majority of these trips are inside of Manhattan.
- There are a large number of trips within Midtown Manhattan and the East Village Region of Lower Manhattan.
- There are some trips along the waterfront from Battery Park to as far north as the 495 bridge to New Jersey similar to the customer trips in Figure 4.
- There no trips through Central park.
- There are small isolated clusters of trips in Brooklyn.

Comparing Figure 4 to Figure 5, it appears that the customers are taking recreational trips in areas that are popular with tourists while the subscribers are taking trips within Lower Manhattan or Midtown Manhattan which are likely more pragmatic in nature.

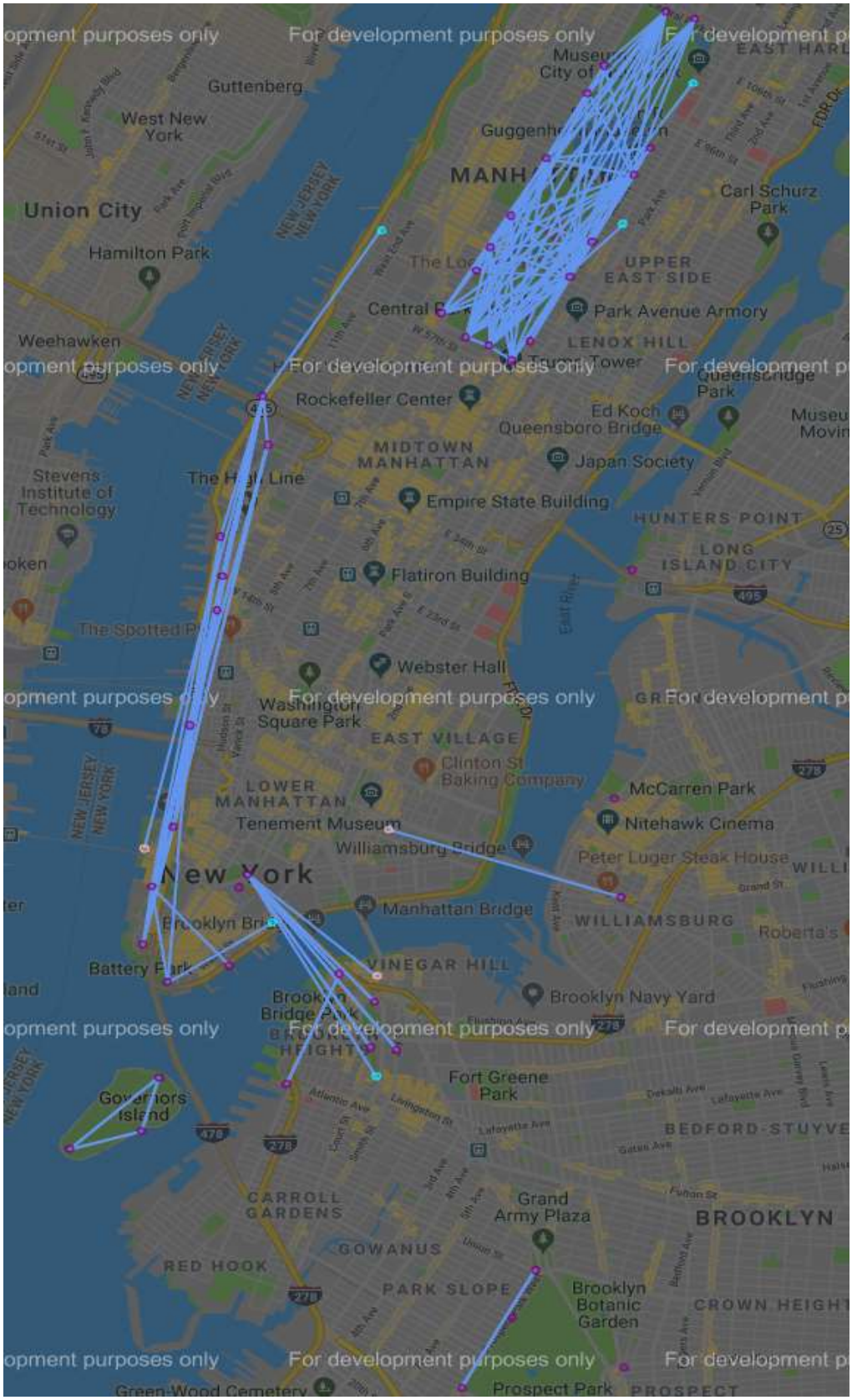

Figure 4: Top 200 Trips taken by Customers

Figure 5: Top 200 Trips taken by Subscribers

5.4 Station Volume and Trip Frequency Assessments for Males and Females.

The data set was split into one data set consisting of the trips taken by the “Male” gender and another data set consisting only of the trips taken by the “Female” gender.

The following summary statistics were calculated:

- 70% of all trips were taken by males, 23% of all trips were taken by females and 7% of all trips were taken by riders of ‘unknown’ gender.
- 94% of the male trips were taken by subscribers and 6% of the male trips were taken by customers.
- 90.5% of the female trips taken by subscribers and 9.5% of the female trips were taken by customers.

For both data sets, the number of trips starting and ending at each station were summed and reported as the ‘total volume’. A comparison of the top 100 stations for male ‘total volume’ and the top 100 stations for female ‘total volume’ shows that there are 76 stations in common.

The top 100 'total volume' stations for males and the top 100 'total volume' stations for females were plotted on Google Maps using the gmap API. The following color coding has been applied:

- The top 100 male stations are plotted as sky blue circles.
- The top 100 female stations are plotted as magenta circles.
- The stations common to both the top 100 male stations and the top 10 female stations are plotted as orange circles.

See Figure 6: Top 100 'total volume' Stations for Males and Females on page 18.

The following observations were made:

- Blue circles are found primarily in Midtown Manhattan.
- Magenta circles are found in the Upper West Side and Upper East Side near Central Park, and in Williamsburg and in Lower Manhattan.
- Orange circles are found in Midtown Manhattan and Lower Manhattan.

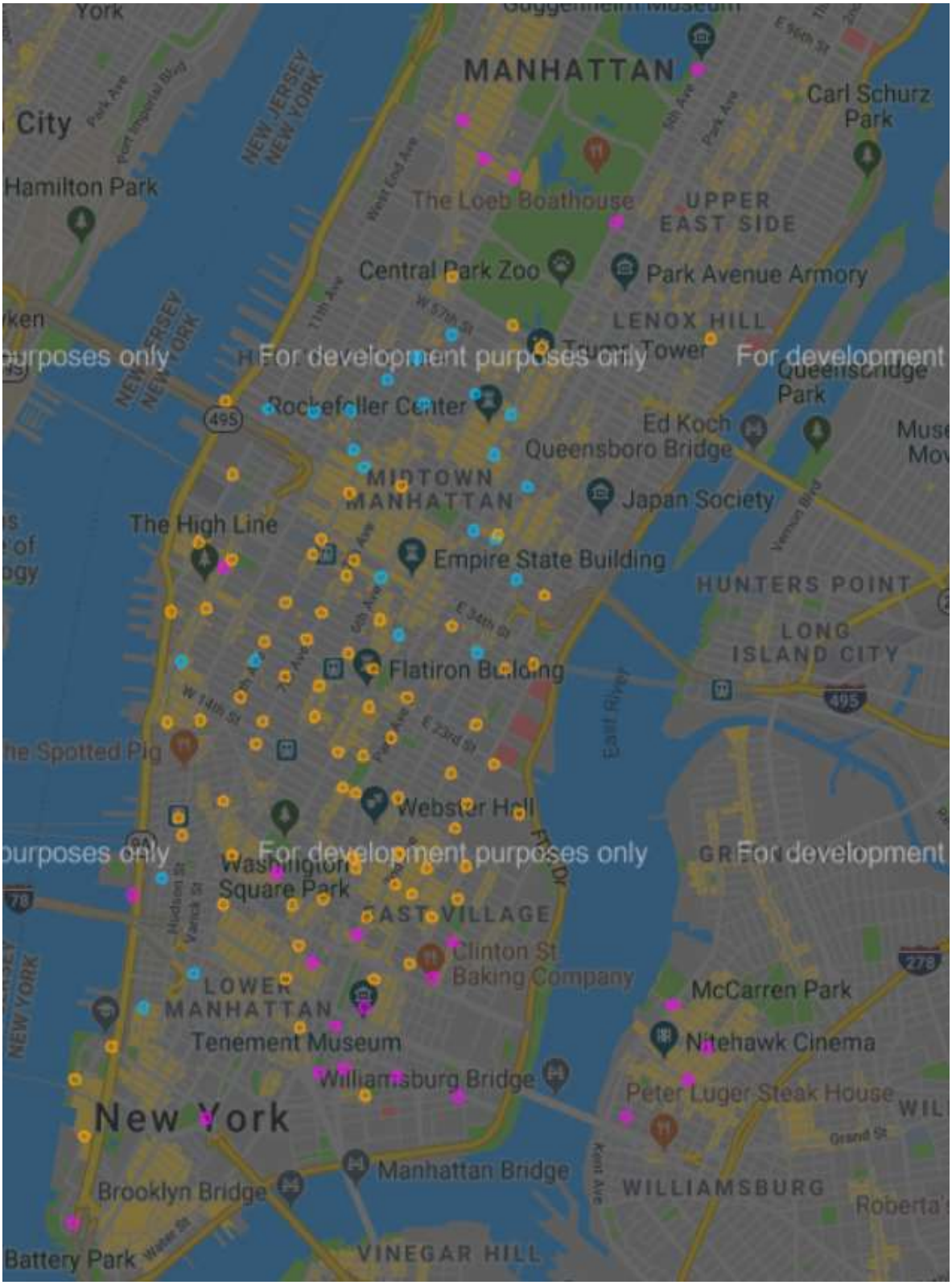

Figure 6: Top 100 'total volume' Stations for Males and Females

For both data sets, the frequency for each unique trip was calculated and reported as the 'total frequency' for that trip. The top 200 most frequently taken male trips were plotted on Google Maps using the gmap API. Each of these trips was taken in excess of 799 times.

For the trips being plotted, the following color coding applies:

- Stations that are "start station only" for the trips are plotted as pink circles.
- Stations that are "end station only" for the trips are plotted as cyan circles.
- Stations that are both "end stations" and "start stations" are plotted as purple circles.
- A blue line is drawn from start station to end station to show where a trip begins and ends.
- Purple circles without any blue lines originating from them represent trips that started and ended at the same station.

See Figure 7: Top 200 Most Frequently Taken Male Trips on page 20.

The following observations were made:

- The majority of these trips are inside of Manhattan.
- There are a large number of trips within Midtown Manhattan and the East Village Region of Lower Manhattan.
- There is a cluster of trips around the Battery Park Area.
- There are small isolated clusters of trips in Brooklyn.

The top 200 most frequently taken female trips were plotted on Google Maps using the gmap API. Each of these trips was taken in excess of 256 times. See Figure 8: Top 200 Most Frequently Taken Female Trips on page 21.

The following observations were made:

- The most common female trips include inside of Central Park, the waterfront from Battery Park to 495 bridge to New Jersey, the East Village and Williamsburg.
- The female trips do not occur in Midtown Manhattan the way the male trips do in Figure 7.

The male top 200 trips shown in Figure 7 closely resembles the top 200 subscriber trips shown in Figure 5. The top 200 female trips in Figure 8 shows some resemblance to the top 200 customer trips in Figure 4, particularly with regard to the central park trips and waterfront trips from Battery park to Hudson Yards.

Figure 7: Top 200 Most Frequently Taken Male Trips

Figure 8: Top 200 Most Frequently Taken Female Trips

5.5 Station Availability Assessment

An assumption was made that if a station has no "trip starts" or "trip stops" on a given day that the station was “unavailable” for service on that day. Using this assumption, it was found that 620 out of 822 stations (75%) are available every day during the six month time frame and 202 out of 822 stations (25%) have at least one day of being unavailable.

For the 202 stations that are unavailable for at least one day:

- The median unavailability is 19.50 days. Thus, 101 stations have unavailability of 19.50 days or less. These 101 stations are unavailable less than 10% of the time (> than 90% availability).
- The 75th percentile of unavailability is 107 days. Thus, ~50 stations have unavailability of greater than 19.50 days but less than or equal to 107 days. These ~50 stations are unavailable between 10% and 59% of the time (41% to 90% availability).
- The remaining ~50 stations have unavailability greater than 107 days but less than or equal to 178 days. These stations are unavailable between 59% and 98% of the time (2% to 41% availability).

The latitude and longitude coordinates for all 822 stations were plotted on Google Maps using the gmap API. The stations were color coded based on availability as follows:

- The stations available 100% of the time are plotted as blue circles.
- The stations in the 50th percentile of unavailability (>90% availability) are plotted as cyan circles.
- The stations bounded by the 50th to 75th percentile of unavailability (41% to 90% availability) are plotted as yellow circles.
- The stations above the 75th percentile of unavailability (2% to 41% availability) are plotted as red circles.

See Figure 9: Station Availability Plot on page 23.

The following observations were made:

- There is a cluster of red and yellow in to the east of Williamsburg on the edge of the Citi bike service area.
- There is a cluster of cyan circles with a few red and yellow circles in Manhattan.
- Aside from these two clusters, almost all cyan, red and yellow circles have adjacent circles that are blue.

Station unavailability as assumed in this section is unlikely to impact most users. The vast majority of stations had either 100% availability or greater than 90% availability during the six month time frame. The majority of stations with unavailability less than 90% (yellow and red) are usually adjacent the stations to stations which had 100% availability. Thus, it is unlikely that service interruptions had any impact on the trips taken.

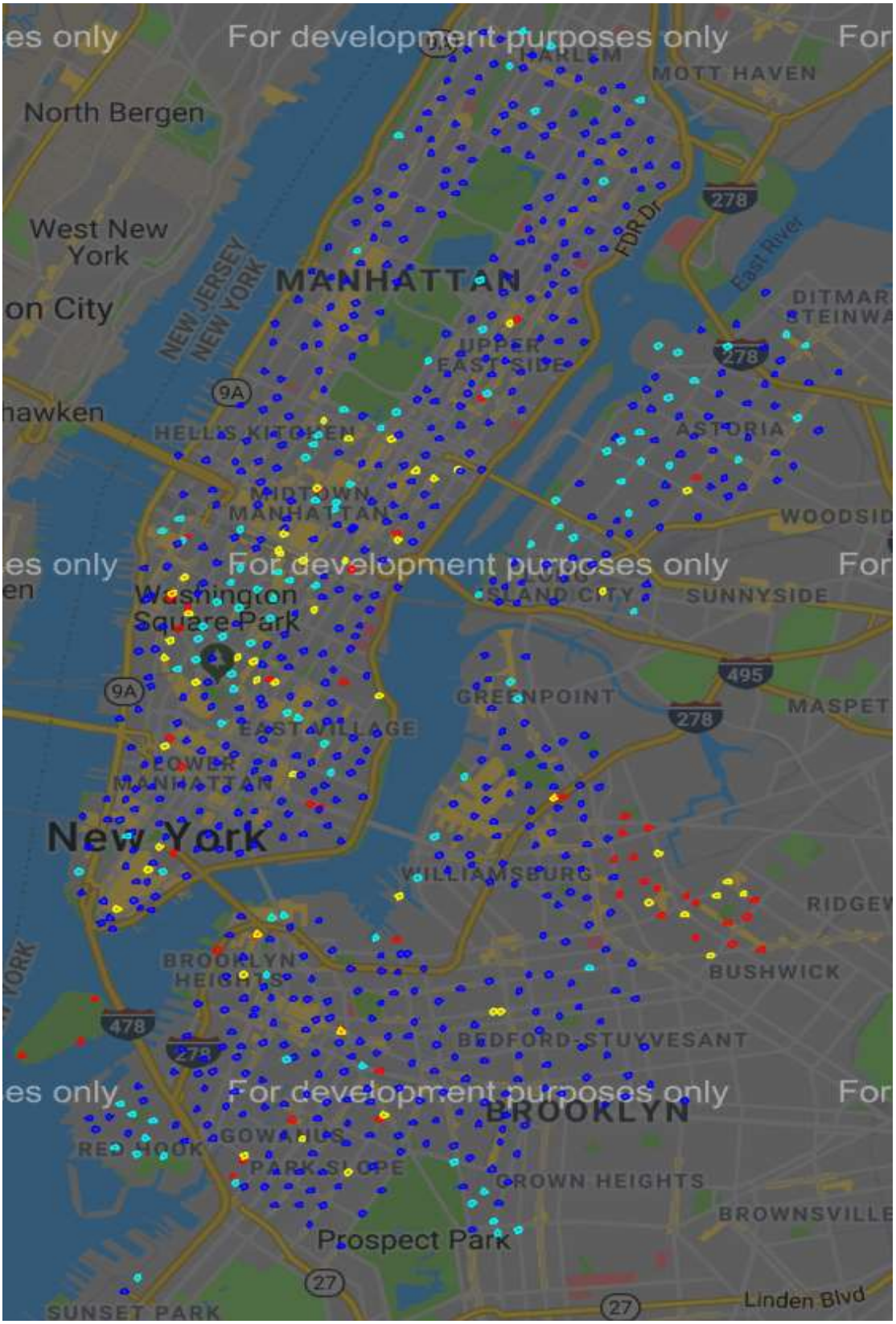

Figure 9: Station Availability Plot

6.0 Inferential Statistics

Two sets of comparisons were performed using inferential statistics:

1. Individual trips were tested to see if the groups of riders taking a given trip had mean ages and mean durations that were statistically the same. This test was performed on each of the 300 most frequently taken trips.
2. The 30 stations with the highest volume were compared against each other to see if the groups of riders had mean ages that were statistically the same.

All statistical comparisons were performed using Welch's two sample t-test. Testing was only performed for instances were both groups of riders being compared had a volume or frequency of at least 30 as a large sample size is a requirement of this t-test method.

The hypothesis for the Welch's two sample t-test are as follows:

- H0: means are the same.
- HA: means are not the same.

The results of this two-tailed test were evaluated at the 95% confidence level.

6.1 Age Data Assessment

Prior to performing these three sets of tests, the age data was assessed by plotting histograms. It was discovered that the age data contained a disproportionately large number of trips taken by people who are 50 years of age.

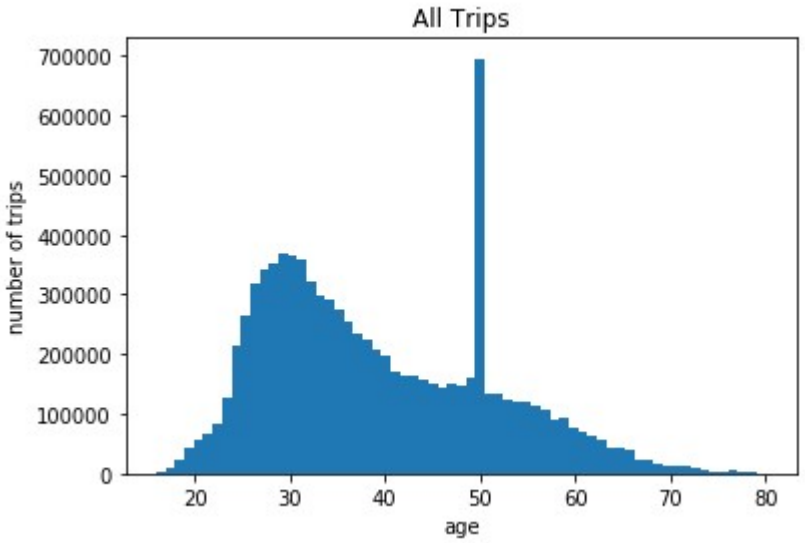

Figure 10: Number of Trips vs. Age - All Trips

The histogram in Figure 10 shows that there are just under 700,000 trips taken by people who are 50 years of age. However, the numbers of trips taken by each other age group is always below 400,000.

Histograms of the six gender and user type combinations were plotted to determine which subsets of the data had excessive numbers of trips taken by riders who are 50 years of age.

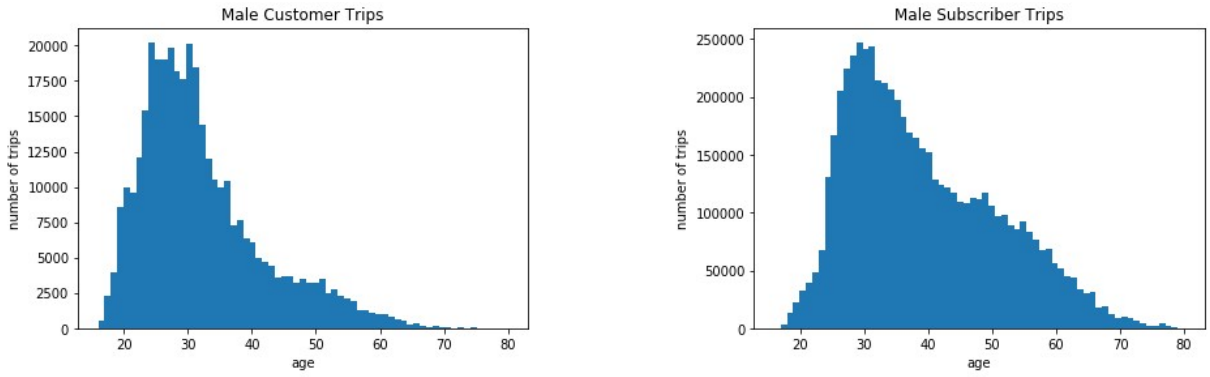

Figure 11: Number of Trips vs. Age – Male User Type Comparison

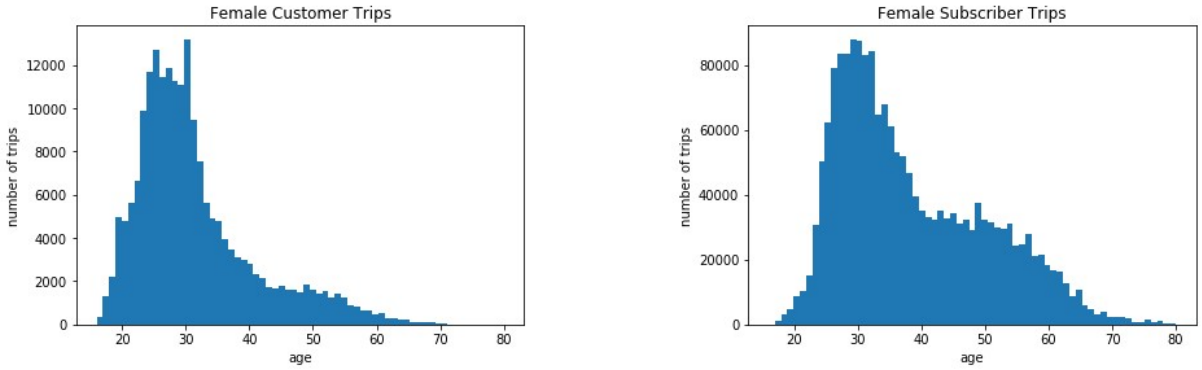

Figure 12: Number of Trips vs. Age - Female User Type Comparison

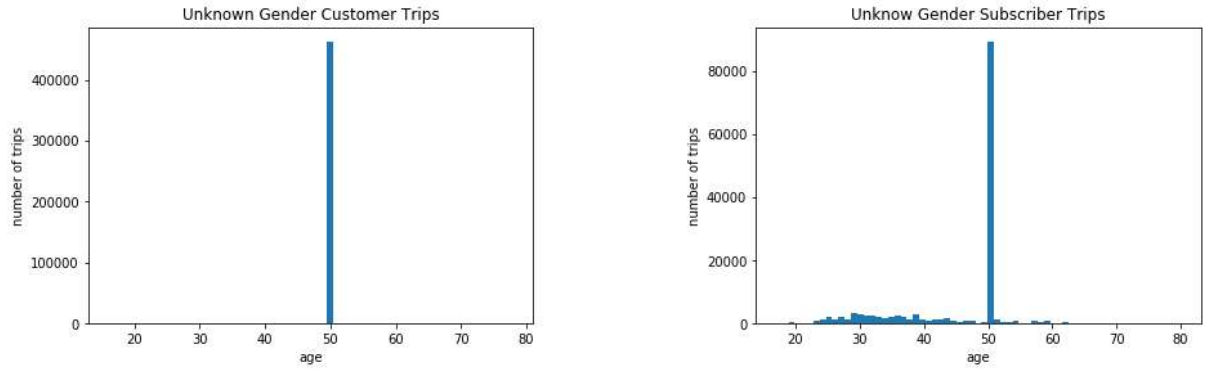

Figure 13: Number of Trips vs. Age - Unknown Gender User Type Comparison

From these six histograms, it is clear that the disproportionately large number of trips taken by people who are 50 years of age is limited to the unknown gender customers and subscribers.

It was calculated that 62% of unknown gender Subscriber trips and 98% of unknown Gender customer trips were taken by people who are 50 years old. As per the histograms shown in Figure 11 and Figure 12, no other age group is this highly represented in the Male Subscriber, the Female Subscriber, Male Customer or Female Customer data.

A decision was made to remove all 'Unknown' Gender Customers and Subscribers from the data set and from further analysis as it is suspected that age has not been correctly reported. Also, given that the gender is unknown, these trips are of very limited value for addressing the business problem.

6.2 Mean Age and Mean Duration Testing Between Rider Groups

Individual trips were tested to see if the groups of riders taking a given trip had mean ages and mean durations that were statistically the same. This testing was performed on each of the 300 most frequently taken trips. The results of these t-tests are shown below in Table 1.

For many of these trips, the number of times the trip was taken by either female customers, male customers or all customers is less than 30 which is too small of a sample to be used for Welch's t-test. Thus, the testing was not performed. The column "Number of trips tested" indicates the number of trips which had sufficient data to perform Welch's two sample t-test for the comparison pair of rider groups being tested.

Comparison Pair Tested	Number of trips tested	Mean Age: "Do not reject H0"	Mean Duration: "Do not reject H0"	Both Tests: "Do not reject H0"
Customers to Subscribers	118	6.78%	17.80%	0.85%
Males to Females	295	41.36%	36.61%	16.61%
Male Subscribers to Male Customers	80	15.00%	20.00%	1.25%
Female Subscribers to Female Customers	38	7.89%	7.89%	0.00%
Male Subscribers to Female Subscribers	292	41.44%	39.73%	18.15%
Male Customers to Female Customers	37	97.30%	75.68%	72.97%
Male Subscribers to Female Customers	38	2.63%	2.63%	0.00%
Female Subscribers to Male Customers	77	32.47%	33.77%	14.29%

Table 1: Testing of Mean Age and Mean Duration between Rider Groups for Top 300 Trips.

Interpretation of the t-test results shown in Table 1:

1. The number of t-test comparisons made between rider groups varies from 37 to 295 out of 300 trips.
2. Based on “customer to subscriber” and “male to female” comparison results, males and females are more likely to have same mean age or mean duration for a given trip than customers and subscribers.
3. Based on “male subscriber to male customer”, “female subscriber to female customer”, “male customer to female customer” and “male subscriber to female subscriber” comparison results, rider groups of same user type and different gender are more likely to have same mean age or same mean duration for a given trip than rider groups of a different user type and same gender.
4. Based on “female subscribers to male customers”, “female subscriber to female customer”, and “male subscriber to male customer” comparison results, female subscribers and male customers are more likely to have the same mean age or same mean duration for a given trip than rider groups with the same gender and different user type.
5. Male subscribers and female customers are the least likely to have the same mean age or same mean duration for a given trip.
6. Male customers and female customers are the most likely to have the same mean age or same mean duration for a given trip.

6.3 Mean Age Testing between High Volume Stations

The 30 stations with the highest volume were compared against each other to see if the groups of riders using these stations had mean ages that are statistically the same. The results are shown in Table 2:

Demographic	Tests performed	Do Not Reject H0	Reject H0 for HA
All Riders	435	12	423
Customers	435	103	332
Subscribers	435	15	420
Males	435	18	417
Females	435	71	364
Male Subscribers	435	20	415
Female Subscribers	435	129	306
Male Customers	435	61	374
Female Customers	435	154	281

Table 2: Testing of Mean Age Between 30 Stations with Highest Total Volume

Observations made from the t-test results shown in Table 2:

1. Volume at these stations is sufficiently high across all nine rider groups so that there are no instances of tests not being performed.
2. Females customers, Female subscribers and Females had the most instances of "do not reject H0" when comparing mean age between stations.
3. All Riders, Subscribers, Males and Male Subscribers all had twenty or less instances of "do not reject H0" when comparing mean age between stations.

4. Customers had almost seven times as many instances of "do not reject H0" than the subscribers when comparing mean age between stations.

Table 3 below shows the frequency at which the number of “Do not Reject H0” occurred when comparing different trips:

Instances of "Do not Reject H0"	Number of Comparisons
0	198
1	71
2	63
3	59
4	22
5	16
6	2
7	3
8	1
9	0

Table 3: Frequency Table for Instances of "Do Not Reject H0" When Comparing Between Stations

Observations made from the t-test results shown in **Table 3: Frequency Table for Instances of "Do Not Reject H0" When Comparing Between Stations**Table 3:

1. 198 of 435 of these station comparisons had no instances of “Do Not Reject H0” for any of the 9 rider groups compared.
2. 193 of 435 of these station comparisons had an instance of “Do Not Reject H0” for either 1, 2, or 3 of the 9 rider groups compared.
3. 44 of these station comparisons had an instance of “Do Not Reject H0” for either 4, 5, 6, 7 or 8 of the 9 rider groups compared.
4. None of these station comparisons had an acceptance of the null hypothesis for all 9 of the rider groups compared.

Based on the observations made from Table 2**Error! Reference source not found.** and Table 3, it is clear that the mean age of rider groups using these high volume stations is generally not statistically the same when comparing between stations. The groups of riders most likely to have a mean age that is statistically the same are customers and groups with the female gender.

7.0 Baseline K-Means Clustering Model

7.1 Model Development

The following steps were performed to prepare a k-means clustering model:

- The latitude and longitude coordinates for each of the 822 stations were converted to radians.
- The Haversine distance between each pair of stations was calculated and converted to kilometers.
- Multidimensional Scaling was used to assign numbers to each station using the Haversine distances as a dissimilarity measure.
- The numbers obtained from multidimensional scaling were normalized to range from 0 to 1.
- The columns 'start station mds' and 'end station mds' were created to contain the normalized multidimensional scaling numbers for each trip.
- The 'usertype' was set to 0 for customer and 1 for subscriber.
- The 'gender' was set to 0 for male and 1 for female.
- The 'age' data was scaled to range from 0 to 1.

The K-means clustering algorithm was then run for values of k from k=2 to k=20 to cluster the trips based on start station mds, end station mds, user type, gender and age.

7.2 Results

The SS value for each k value was calculated and plotted in the graph below:

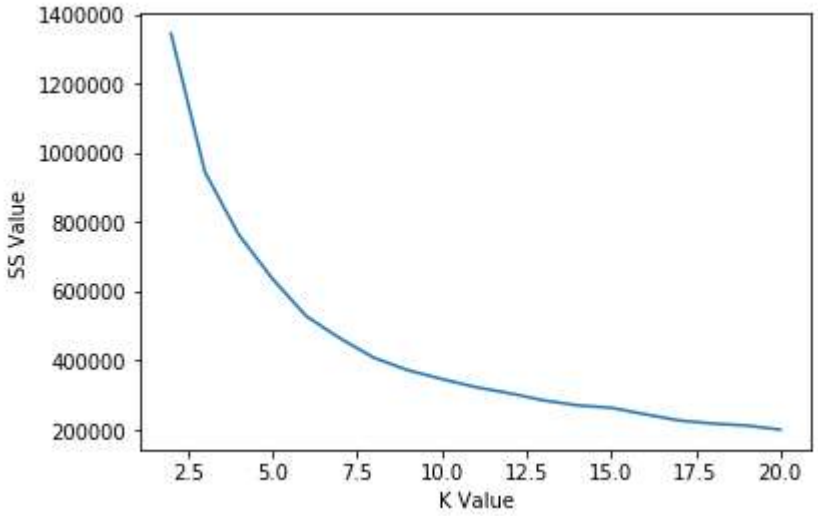

Figure 14: SS Value vs. K Value Plot – Baseline K-Means

The following observations are made based on Figure 14:

- Elbow Points appear to occur at k = 3 and k = 6.
- From K = 10 onward, the graph is very smooth and almost a straight line.

The cluster results for k = 3, k = 6 and k = 10 are shown below in Table 4, Table 5 and Table 6. In these tables, the "Primary Location" column is used to describe the main location of trip occurrences when it is the case that more than one cluster has the same gender/user type combination.

Cluster	Gender	User Type	Min Age	Max Age	Mean Age	Trips	Percent	Primary Locations
0	Male	Subscriber	16	80	39.2	5901493	70.7%	N/A
1	Female	Subscriber	16	80	38.4	1881309	22.5%	N/A
2	Both	Customer	16	80	31.7	560959	6.7%	N/A

Table 4: Cluster Description Baseline K-Means (k = 3)

When k = 3, the customers of both genders are in one cluster together while the male subscribers and female subscribers are in different clusters.

Cluster	Gender	User Type	Min Age	Max Age	Mean Age	Trips	Percent	Primary Locations
0	Male	Subscriber	43	80	53.8	1664146	19.9%	1. Midtown Manhattan.
1	Male	Subscriber	16	42	31.4	2712464	32.5%	1. Midtown Manhattan.
2	Female	Subscriber	16	80	38.4	1881309	22.5%	N/A
3	Male	Subscriber	16	80	37.2	1524883	18.3%	1. Lower Manhattan 2. Brooklyn.
4	Female	Customer	16	80	31.2	197919	2.4%	N/A
5	Male	Customer	16	80	32	363040	4.4%	N/A

Table 5: Cluster Description Baseline K-Means (k = 6)

When k = 6, male customer, female customers and female subscribers each have their own cluster. The male subscribers are split over three clusters. One of the Male subscriber clusters contains trips primarily in Lower Manhattan and Brooklyn taken by riders ranging from 16 to 80 years of age. The other two male subscriber clusters contain trips that are primarily in Midtown Manhattan. One of these clusters is for trips taken by riders aged 43 to 80 and the other cluster is for trips taken by riders aged 16 to 42.

Cluster	Gender	User Type	Min Age	Max Age	Mean Age	Trips	Percent	Primary Locations
0	Male	Subscriber	16	48	32.9	1103492	13.2%	1. Midtown Manhattan.
1	Female	Subscriber	42	80	53.7	496500	6.0%	1. Lower Manhattan. 2. Midtown Manhattan.
2	Male	Customer	16	80	32	363040	4.4%	N/A
3	Male	Subscriber	16	74	36.6	760559	9.1%	1. Lower Manhattan 2. Brooklyn.
4	Female	Customer	16	80	31.2	197919	2.4%	N/A
5	Male	Subscriber	42	80	53	987435	11.8%	1. Midtown Manhattan.
6	Male	Subscriber	42	80	53.9	887893	10.6%	1. Lower Manhattan. 2. Midtown Manhattan
7	Female	Subscriber	16	79	36.4	515360	6.2%	1. Lower Manhattan 2. Brooklyn.
8	Male	Subscriber	16	43	30.9	2162114	25.9%	1. Lower Manhattan. 2. Midtown Manhattan
9	Female	Subscriber	16	43	30.8	869449	10.4%	1. Lower Manhattan. 2. Midtown Manhattan.

Table 6: Cluster Description Baseline K-Means (k = 10)

When k = 10, male customers and female customers each have their own cluster but all other gender/user type combinations have multiple clusters.

Male subscribers are clustered as follows:

- Cluster 0 – ages 16 to 48 with trips primarily in Midtown Manhattan.
- Cluster 5 – ages 42 to 80 with trips primarily in Midtown Manhattan.
- Cluster 8 – ages 16 to 43 with trips primarily in Lower Manhattan and south end of Midtown Manhattan.
- Cluster 6 – ages 42 to 80 with trips primarily in Lower Manhattan and south end of Midtown Manhattan.
- Cluster 3 – ages 16 to 74 with trips primarily in Brooklyn and the south end of Lower Manhattan.

The top 200 most frequently taken trips in each of these clusters are shown in Figure 15, Figure 16 and Figure 17.

Female subscribers are clustered as follows:

- Cluster 1 – ages 42 to 80 with trips primarily in Lower Manhattan and south end of Midtown Manhattan.
- Cluster 9 – ages 16 to 42 with trips primarily in Lower Manhattan and south end of Midtown Manhattan.
- Cluster 7 – ages 16 to 79 with trips primarily in Brooklyn and the south end of Lower Manhattan.

The top 200 most frequently taken trips in each of these clusters are shown in **Figure 18** and **Figure 19**.

Based on the clustering results from k = 3, k = 6 and k = 10, the subscribers for each gender are getting subdivided based on location and age as the value of k increases.

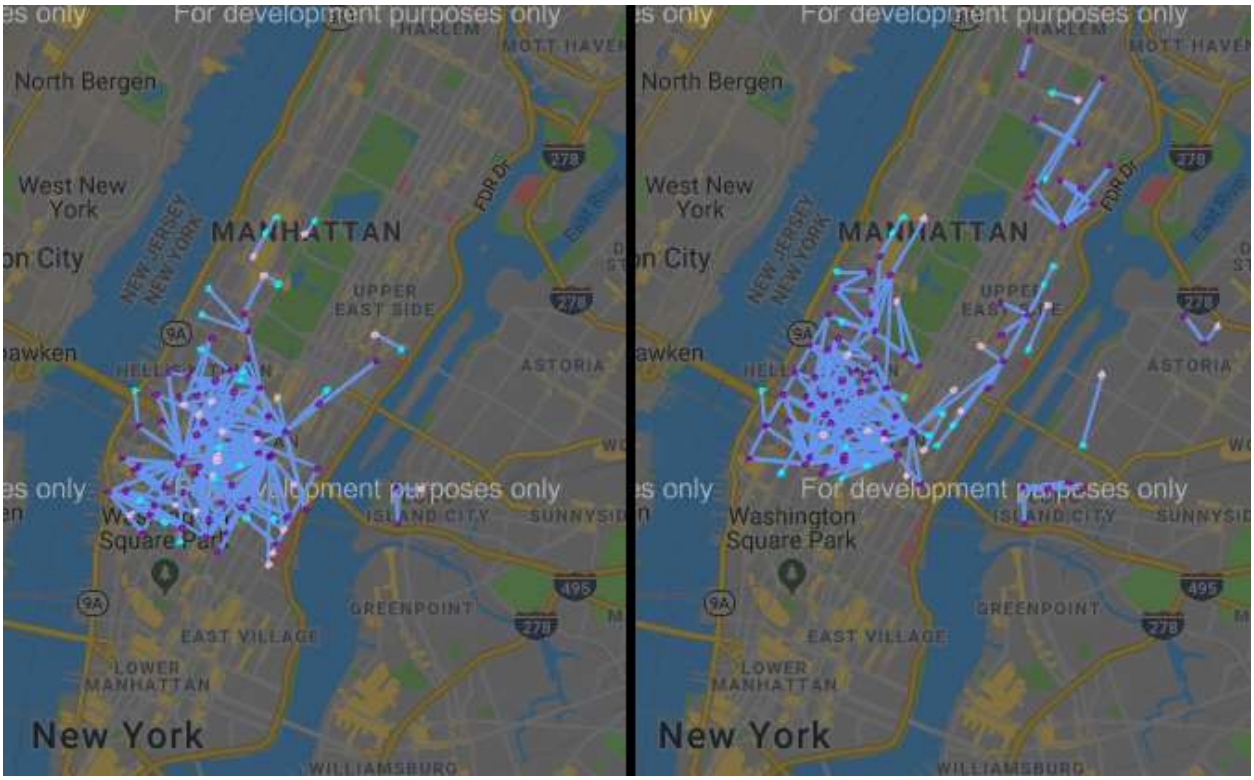

Figure 15: Cluster 5 – Ages 42 to 80 (left) and Cluster 0 – Ages 16 to 48 (right) Male Subscriber Trips in Midtown Manhattan

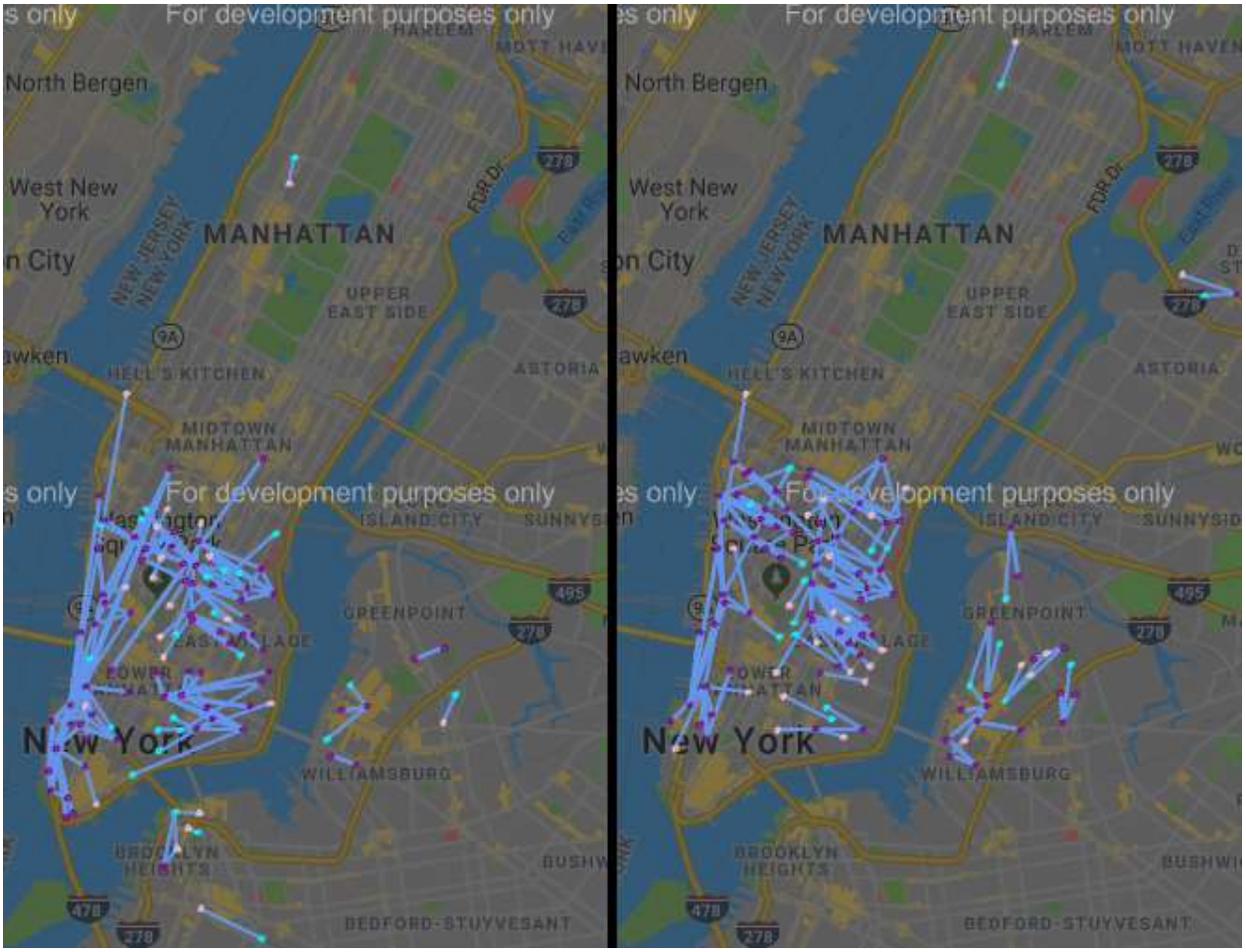

Figure 16: Cluster 6 – Ages 42 to 80 (left) and Cluster 8 – Ages 16 to 43 (right) Male Subscriber Trips in Midtown and Lower Manhattan

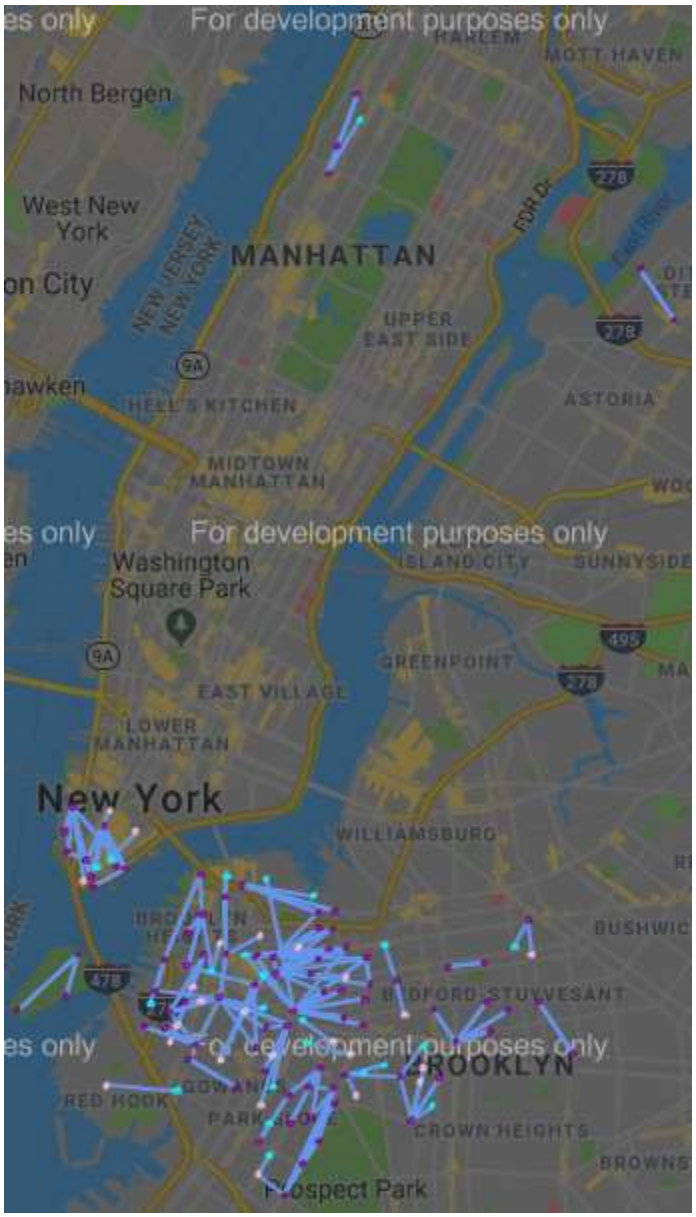

Figure 17: Cluster 3 – Ages 16 to 74 Male Subscriber Trips in Brooklyn and Lower Manhattan

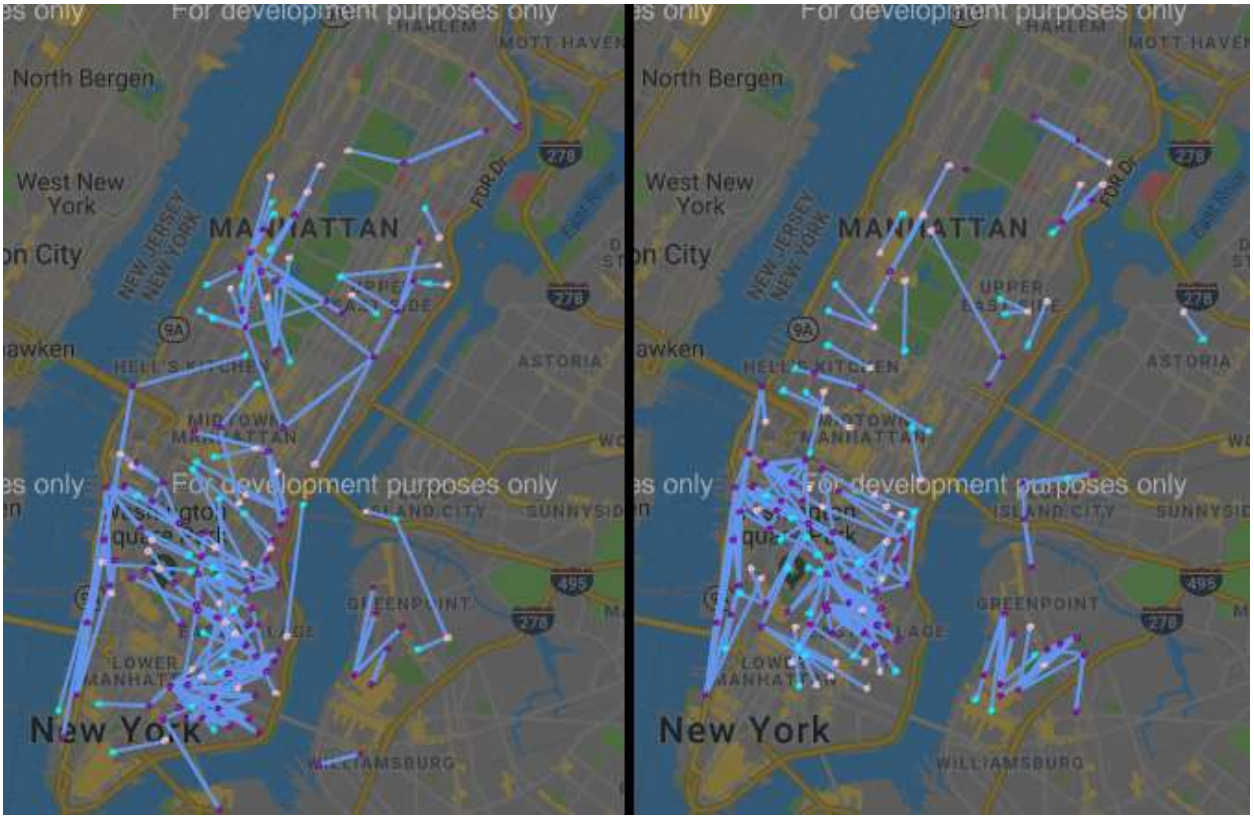

Figure 18: Cluster 1 – Ages 42 to 80 (left) and Cluster 9 – Ages 16 to 42 (right) Female Subscriber Trips in Midtown and Lower Manhattan

Figure 19: Cluster 7 – Ages 16 to 79 Female Subscriber Trips in Brooklyn and Lower Manhattan

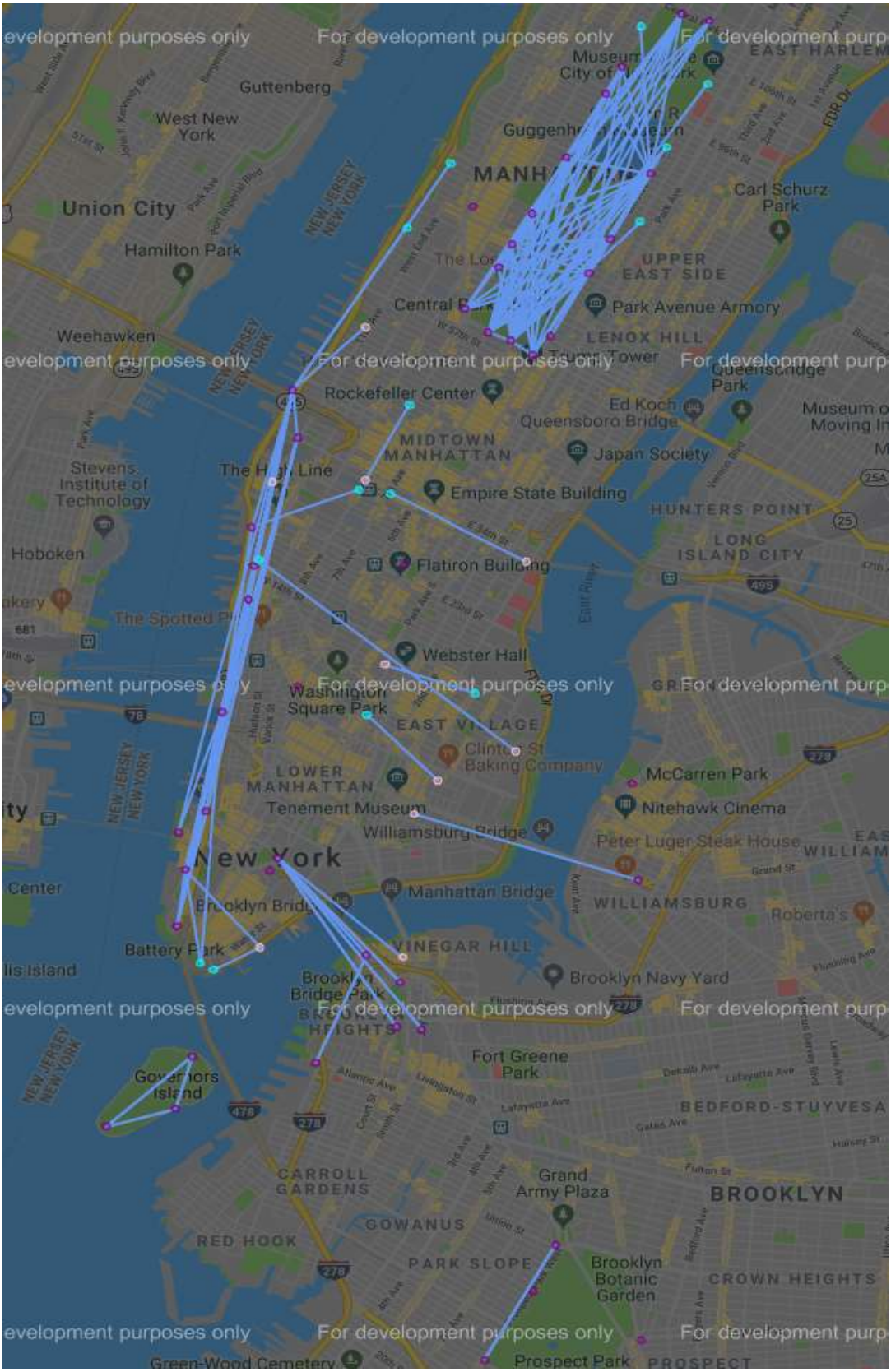

Figure 20: Cluster 2 – Male Customer Trips 16 to 80 years of age

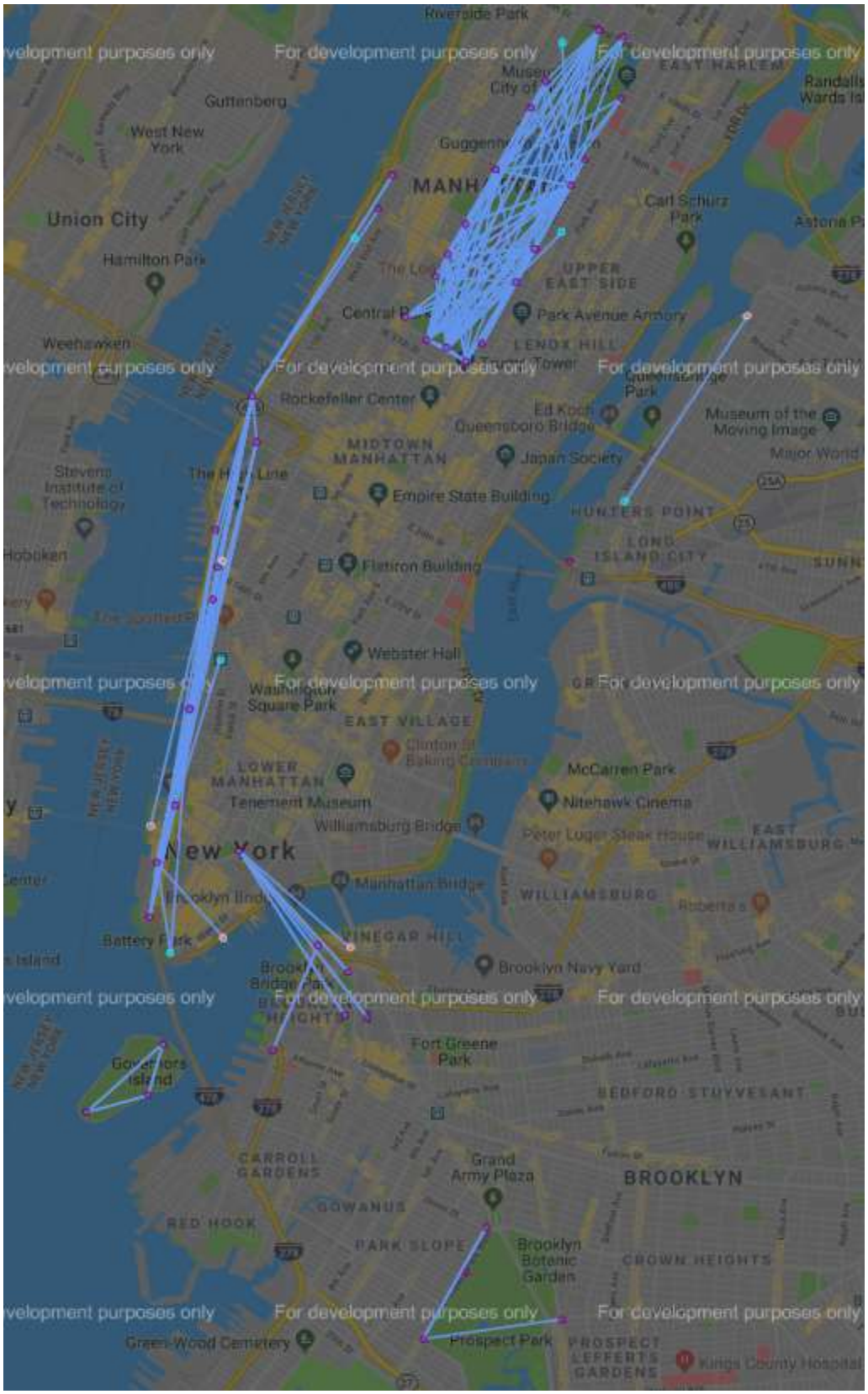

Figure 21: Cluster 4 – Female Customer Trips 16 to 80 years of age

8.0 K-Means Clustering Model with Start Time and End Time

8.1 Model Development

In addition to the steps performed in Section 7.1 to create the baseline K-means model, the following steps were performed to incorporate start time and end time:

- The time portion of the timestamps present in the 'starttime' and 'stoptime' columns were extracted, saved as strings and then converted to time objects formatted as hours-minutes-seconds. The new columns were labelled as 'start' and 'end' respectively.
- The times in 'start' and 'end' was converted to seconds by multiplying the hours by 3600 and the minutes by 60 and then adding these two quantities to the number of seconds.
- The 'start' and 'end' values were both converted into a sine component and a cosine component using the following formulae:

$$\text{sin\_comp} = \sin\left(2 \times \pi \times T \div (60 \times 24 \times 60)\right)$$

$$\text{cos\_comp} = \cos\left(2 \times \pi \times T \div (60 \times 24 \times 60)\right)$$

  where T = 'start' or T = 'end'. This two-component representation of time captures the cyclical nature of the time of day.
- The two sets of time components for 'start' and 'end' were normalized to range from 0 to 1.

The K-means clustering algorithm was then run for values of k from k=2 to k=20 to cluster the trips based on start station mds, end station mds, user type, gender, age, sin_start, cos_start, sin_end and cos_end.

8.2 Results

The SS value for each k value was calculated and plotted in the graph below:

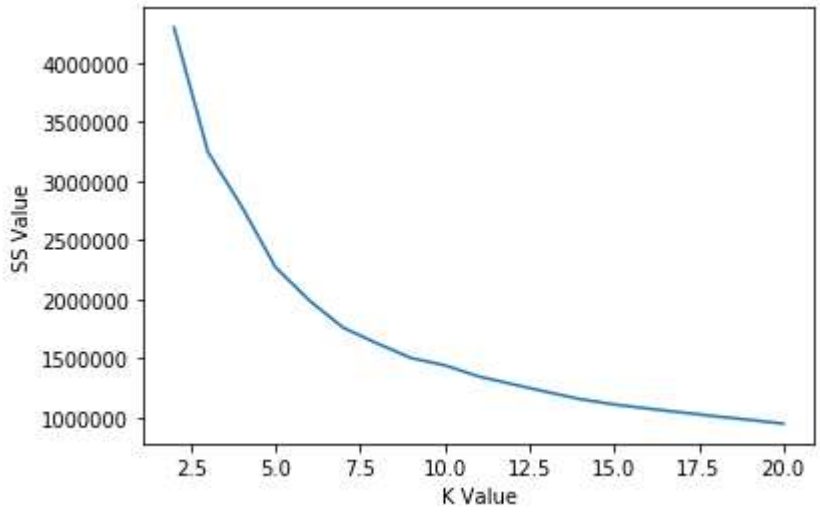

Figure 22: SS Value vs. K Value Plot – K-Means with Start Time and End Time

The following observations were made based on **Figure 1**Figure 22:

- Elbow Points appear to occur at k = 3, k = 5 and k=7.
- From k = 10 onward, the graph is very smooth and almost a straight line.

The cluster results for k = 3, k = 7 and k = 10 are shown in**Table 1**Table 7, Table 8 and Table 9.

Cluster	Gender	User Type	Min Age	Max Age	Mean Age	Earliest Start Time	Latest Start Time	Earliest End Time	Latest End Time	Trips	Percent
0	Male	Both	16	80	39.7	1:45:05	14:14:12	2:11:24	14:41:43	2686240	32.2%
1	Female	Both	16	80	37.7	0:00:00	23:59:59	0:00:00	23:59:59	2079228	24.9%
2	Male	Both	16	80	38	0:00:00	23:59:59	0:00:00	23:59:59	3578293	42.9%

Table 7: Cluster Description K-Means with Start Time and Stop Time (k = 3)

Cluster	Gender	User Type	Min Age	Max Age	Mean Age	Earliest Start Time	Latest Start Time	Earliest End Time	Latest End Time	Trips	Percent
0	Male	Subscriber	16	80	39.1	15:09:25	20:04:07	15:35:21	20:29:49	2004786	24.0%
1	Male	Both	16	80	39.9	2:48:50	11:06:20	3:11:53	11:34:14	1671700	20.0%
2	Female	Subscriber	16	80	38	0:00:00	23:59:59	0:00:00	23:59:59	1038937	12.5%
3	Female	Both	16	80	38.6	1:44:33	14:04:14	2:12:26	14:33:20	868864	10.4%
4	Male	Subscriber	16	80	40.3	10:35:26	15:41:59	11:01:58	16:11:27	1369540	16.4%
5	Male	Both	16	80	36.1	0:00:00	23:59:59	0:00:00	23:59:59	873767	10.5%
6	Both	Customer	16	80	31.5	0:00:00	23:59:59	0:00:00	23:59:59	516167	6.2%

Table 8: Cluster Description K-Means with Start Time and Stop Time (k = 7)

Cluster	Gender	User Type	Min Age	Max Age	Mean Age	Earliest Start Time	Latest Start Time	Earliest End Time	Latest End Time	Trips	Percent
0	Male	Subscriber	16	80	36.1	0:00:00	23:59:59	0:00:00	3:59:59	880478	10.6%
1	Male	Subscriber	16	80	40.3	10:35:01	15:41:51	11:01:58	16:09:12	1365189	16.4%
2	Male	Both	16	80	40	2:48:50	11:05:59	3:11:53	11:34:14	1653192	19.8%
3	Female	Subscriber	16	80	36.1	0:00:00	23:59:59	0:00:00	23:59:59	290378	3.5%
4	Female	Subscriber	16	80	39.5	10:37:10	15:31:29	11:03:29	15:57:27	448459	5.4%
5	Male	Subscriber	16	80	39.2	15:07:56	20:03:54	15:35:21	20:29:49	2002768	24.0%
6	Female	Subscriber	16	80	38.5	14:56:12	19:42:40	15:25:29	20:00:45	596064	7.1%
7	Female	Customer	16	80	31.2	0:00:00	23:59:59	0:00:00	23:59:59	197880	2.4%
8	Male	Customer	16	80	32	0:00:00	23:59:59	0:00:00	23:59:59	362906	4.3%
9	Female	Both	16	80	38.5	2:39:35	11:06:24	3:02:23	11:34:41	546447	6.5%

Table 9: Cluster Description K-Means with Start Time and Stop Time (k = 10)

When k = 3, there is one cluster of females containing both user types and two clusters of males containing both user types.

The male clusters are as follows:

- Cluster 0 – ages 16 to 80 with trips starting and ending between 1:45 and 14:41.
- Cluster 2 – ages 16 to 80 with trips starting and ending between 00:00 and 23:59.

The 200 most frequently taken trips in Cluster 0 and the 200 most frequently taken trips in Cluster 2 were plotted on Google Maps using the gmap API and placed side by side below in Figure 23:

Figure 23: Cluster 0 (left) and Cluster 2 (right) - 200 Most Frequently Taken Trips for Males

The figure above clearly shows that Cluster 0 and Cluster 2 are nearly identical with regard to the 200 most frequently taken trips. Both figures are covering the same geographic area.

Histograms of the start time and end time values for the trips in Cluster 2 were plotted in Figure 24:

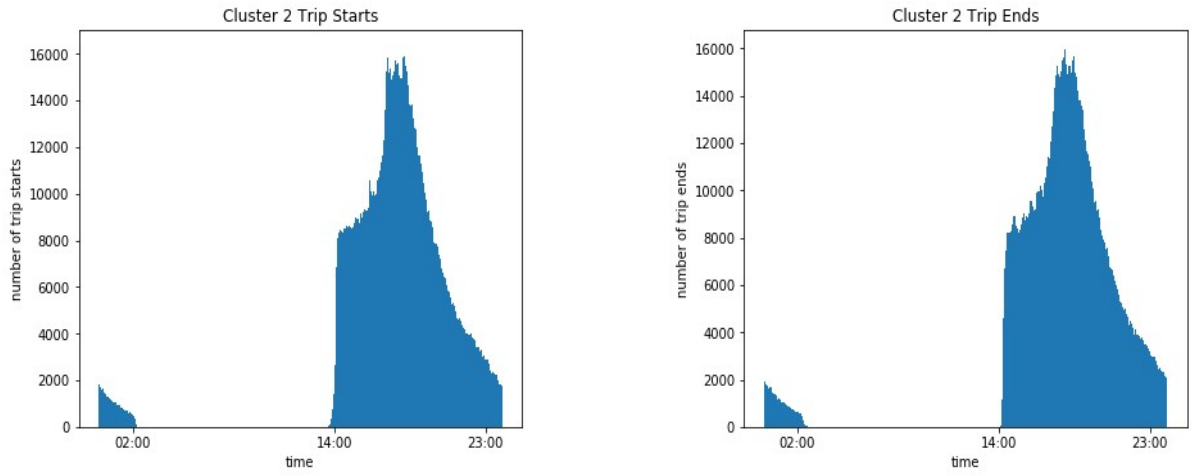

Figure 24: Cluster 2 Histogram of Trip Starts and Trips Ends

The histograms above clearly show that there are no trip starts or trips ends between approximately 2:00 and 14:00. This break in the Cluster 2 data approximately corresponds to the range of trip starts and trip ends present in the Cluster 0 data (14:45 to 14:41) with overlap at the ends of the range.

When k = 7, the clustering is as follows:

There are two clusters of male subscriber trips:

- Cluster 0: ages 16 to 80 with trips starting and ending between 15:09 and 20:29.
- Cluster 4: ages 16 to 80 with trips starting and ending between 10:35 and 16:11.

There are two clusters of male trips containing both user types:

- Cluster 1: ages 16 to 80 with trips starting and ending between 02:48 and 11:34. This cluster consists of 1,653,487 subscribers and 18,213 customers.
- Cluster 5: ages 16 to 80 with trips starting and ending between 00:00 and 23:59. This cluster consists of 873,680 subscribers and only 87 customers.

There are two clusters of female trips:

- Cluster 2: subscribers aged 16 to 80 with trips starting and ending between 00:00 and 23:59.
- Cluster 3: subscribers and customers aged 16 to 80 with trips starting and ending between 01:44 and 14:33. This cluster consists of 842,372 subscribers and 26,492 customers.

There is one cluster of customers containing both genders:

- Cluster 6: ages 16 to 80 with trips starting and ending between 00:00 and 23:59.

Histograms of the trip starts and trip stops were plotted similar to Figure 24 to investigate the ranges of the clusters which had a start time and end time range of 00:00 to 23:59. The following observations were made:

- There is a break in the Cluster 5 data which corresponds to the combined time ranges of Cluster 0, Cluster 4 and Cluster 1.
- There is a break in the Cluster 2 data which corresponds to the time range of Cluster 3 “Subscribers”.
- There is a break in the Cluster 6 data which corresponds to the time ranges of Cluster 1 “Customers” and Cluster 3 “Customers”.

The 200 most frequently taken trips in each of these seven clusters were plotted on Google maps using the gmap API and the results were reviewed. Each of these clusters was found to be covering the whole geographic area serviced by Citi Bike.

When k = 10, the clustering is as follows:

There are three clusters of male subscriber trips:

- Cluster 0: ages 16 to 80 with trips starting and ending between 00:00 and 23:59.
- Cluster 1: ages 16 to 80 with trips starting and ending between 10:35 and 16:09.
- Cluster 5: ages 16 to 80 with trips starting and ending between 15:07 and 20:29.

There is one cluster of male trips containing both user types:

- Cluster 2: ages 16 to 80 with trips starting and ending between 02:48 and 11:34. This cluster consists of 1,653,058 subscribers and only 134 customers.

There are three clusters of female subscriber trips:

- Cluster 3: ages 16 to 80 with trips starting and ending between 00:00 and 23:59.
- Cluster 4: ages 16 to 80 with trips starting and ending between 10:37 and 15:57.
- Cluster 6: ages 16 to 80 with trips starting and ending between 14:56 and 20:00.

There is one cluster of female trips containing both user types:

- Cluster 9: ages 16 to 80 with trips starting and ending between 02:39 and 11:34. This cluster consists of 546,447 subscribers and only 39 customers.

There are two clusters which only contain customers:

- Cluster 7: Females aged 16 to 80 with trips starting and ending between 00:00 and 23:59.
- Cluster 8: Males aged 16 to 80 with trips starting and ending between 00:00 and 23:59.

Histograms of the trip starts and trip stops were plotted similar to Figure 24 to investigate the ranges of the clusters which had start times and end times ranging from 00:00 to 23:59. The following observations were made:

- There is a break in the Cluster 0 data which corresponds to the combined time ranges of Cluster 1, Cluster 2 “Subscribers” and Cluster 5.

- There is a break in the Cluster 3 data which corresponds to the combined time ranges of Cluster 4, Cluster 9 “Subscribers” and Cluster 6.
- There are no breaks in the data for Cluster 7 and Cluster 8.

The top 200 most frequently taken trips in each of these clusters were plotted on Google maps using the gmap API. These plots are shown in Figure 25 to Figure 29 on pages 47 to 51.

Based on the clustering results from k = 3, k = 7 and k = 10, the subscriber trips for each gender are getting subdivided based on start time and end time as the value of k increases. Each cluster covers the entire geographic range of the service area and the entire age range of the riders. The geographic and age “splitting” that was observed to occur in the baseline K-Means model in Section 7.0 does not appear to be occurring at all.

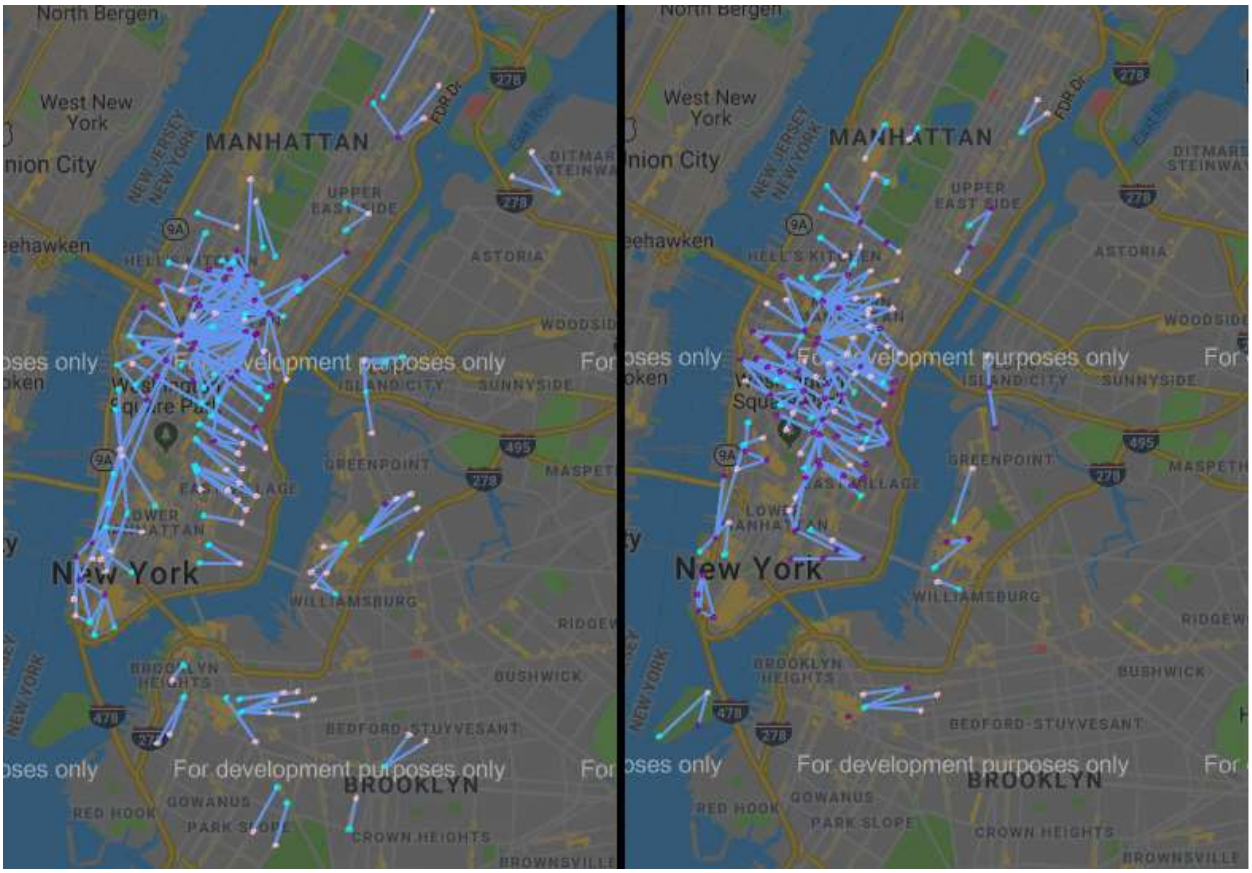

Figure 25: Cluster 2 – Male Subscriber and Customer Trips 02:48 to 11:34 (Left) and Cluster 1 - Male Subscriber Trips 10:35 to 16:09 (Right)

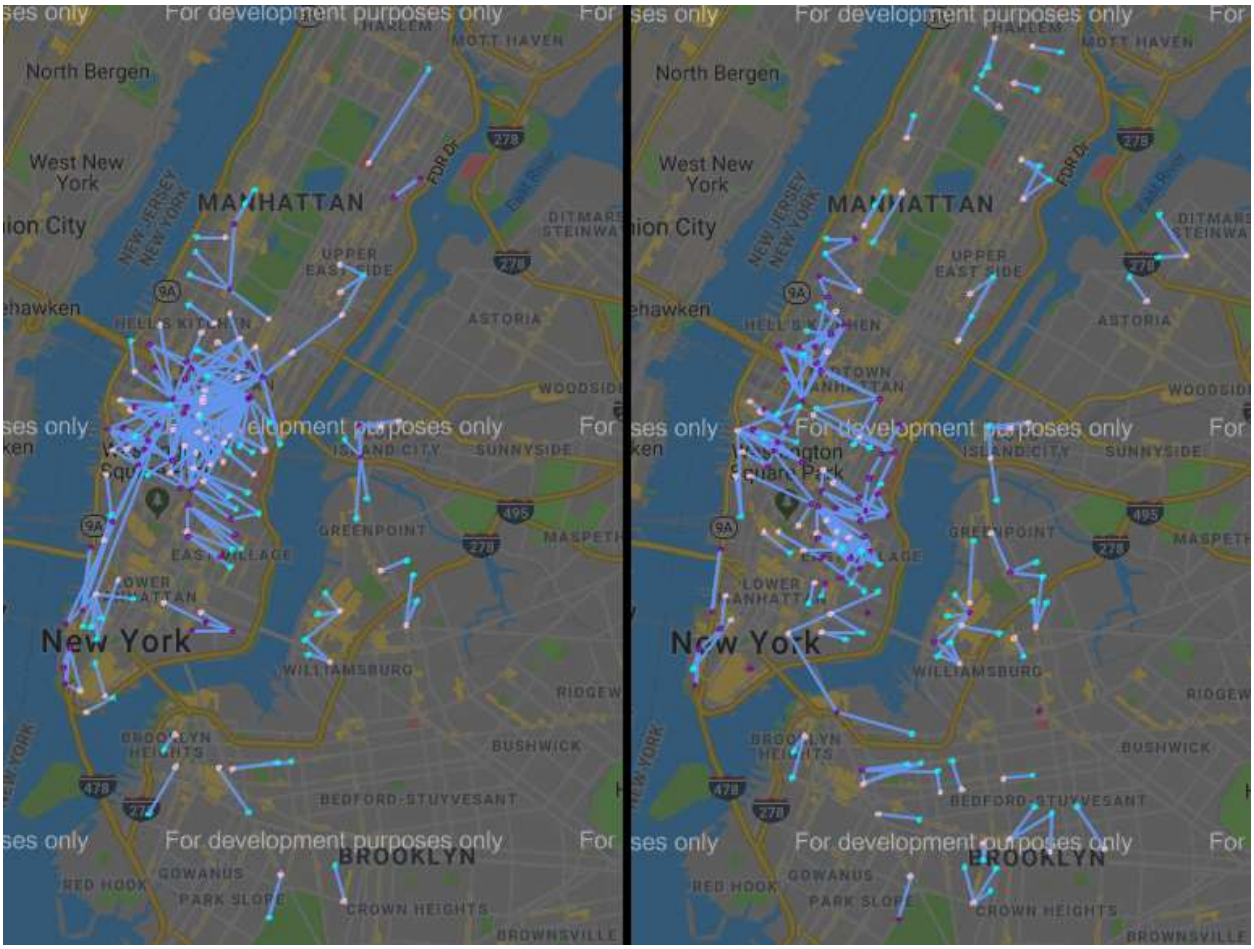

Figure 26: Cluster 5 – Male Subscriber Trips 15:07 to 20:29 (left) and Cluster 0 – Male Subscriber Trips 00:00 to 23:59 (right).

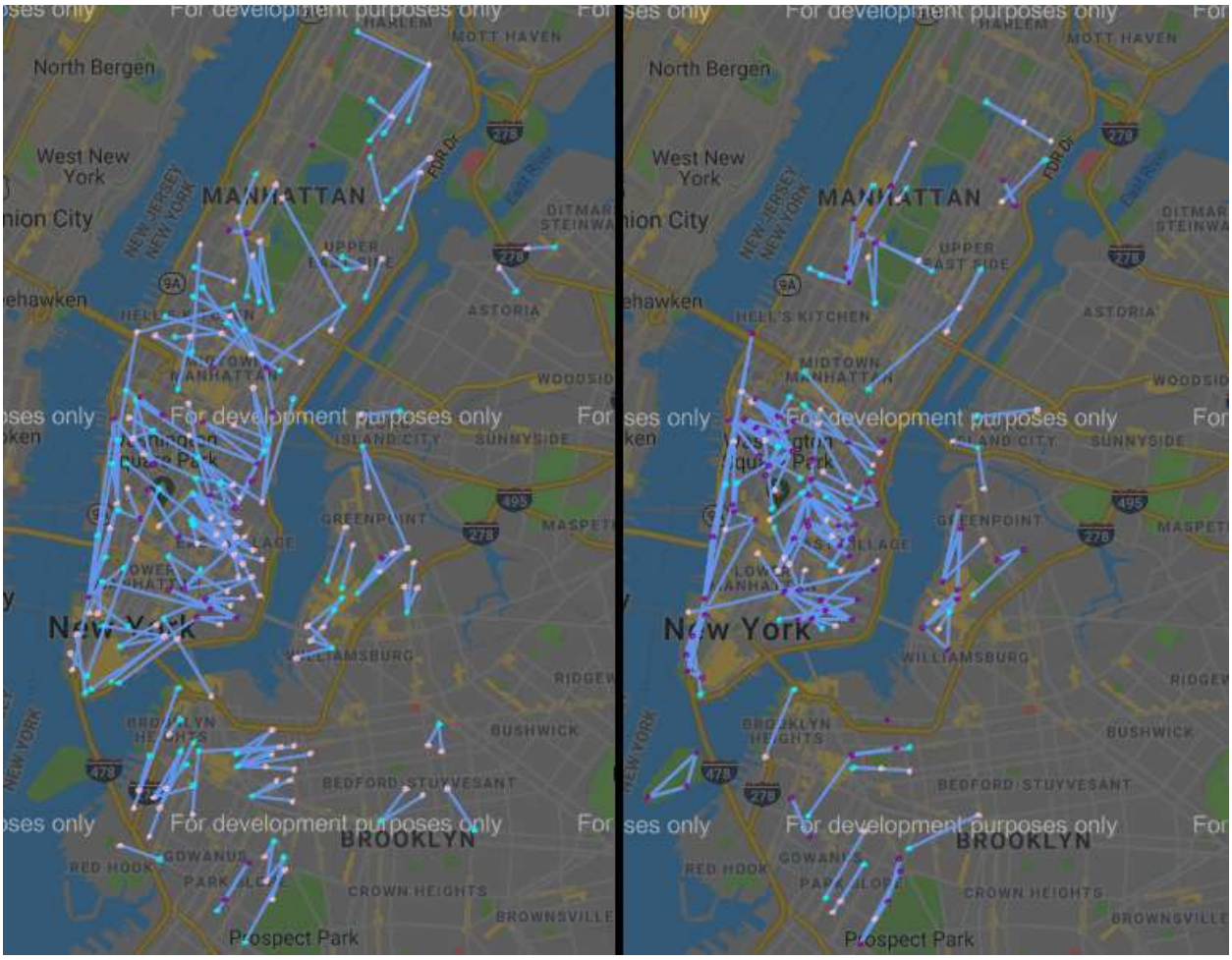

Figure 27: Cluster 9 - Female Subscriber and Customer Trips 02:39 to 11:34 (left) and Cluster 4 - Female Subscriber Trips 10:37 to 15:57 (right).

Figure 28: Cluster 6 - Female Subscriber Trips 14:56 to 20:00 (left) and Cluster 3 - Female Subscriber Trips 00:00 to 23:59 (right).

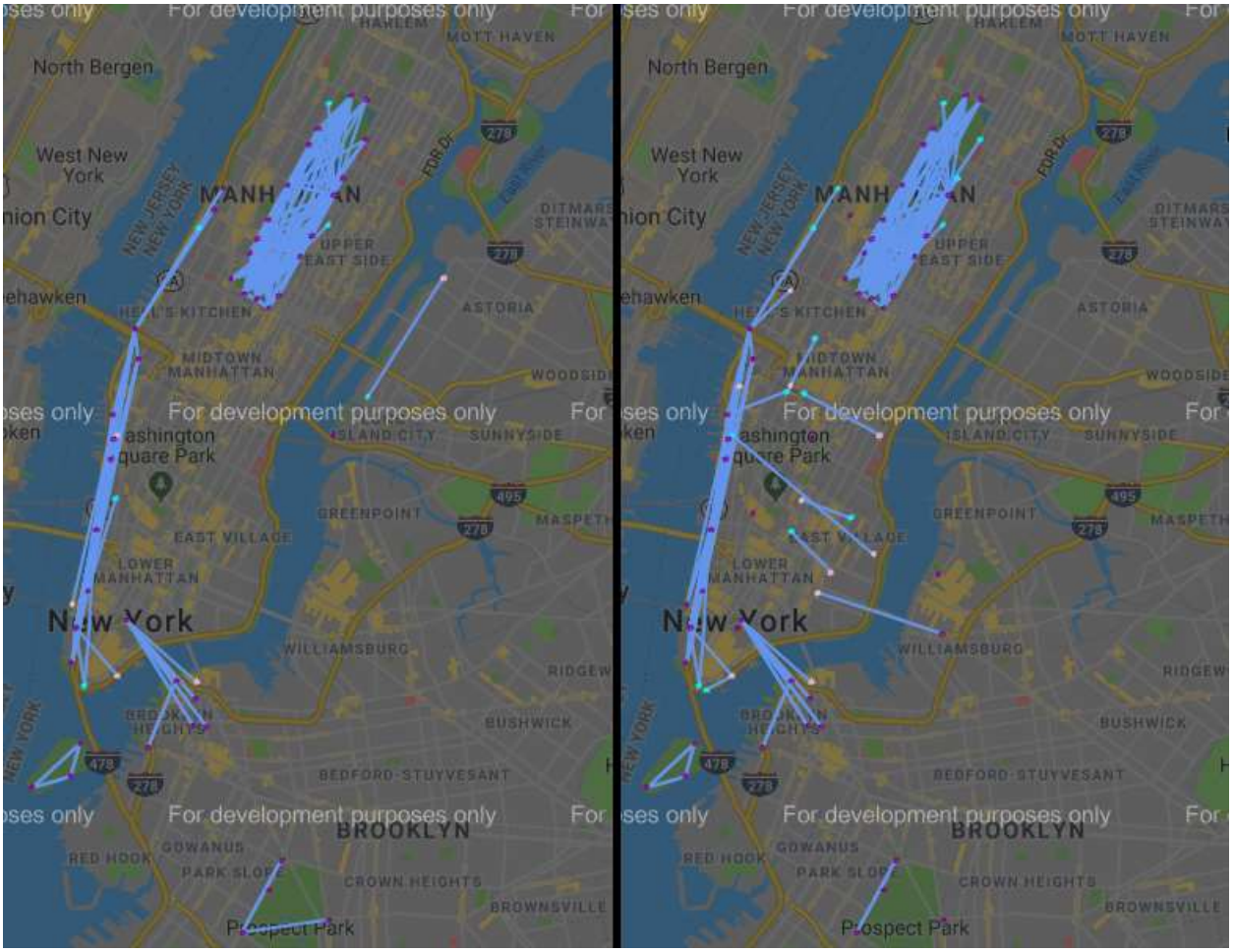

Figure 29: Cluster 7 – Female Customer Trips 00:00 to 23:59 (left) and Cluster 8 – Male Customer Trips 00:00 to 23:59 (right).

9.0 Conclusions

The purpose of this analysis was to segment a data set of Citi Bike trips based on rider type, gender, age, start station and end station so that station user demographics could be determined and used to support the placement of demographically targeted bill board advertisements at select stations.

From the data visualization and descriptive statistics, it was determined that:

- Station volume varies substantially between stations and a small number of stations are responsible for a very large portion of the rider volume.
- The frequency at which unique trips are taken varies substantially and a small number of trips are responsible for a very large portion of the total number of trips taken.
- The stations with the highest volume are concentrated in Midtown and Lower Manhattan. Station volume decreases as one moves north away from Midtown Manhattan or away from Manhattan through Queens or Brooklyn.
- Customers are taking trips in areas that are popular with tourists (Brooklyn Bridge, Central Park and waterfront from Battery Park to the 495 bridge) while the subscribers are primarily taking trips within Lower Manhattan or Midtown Manhattan.

From the inferential statistics, it was determined that:

- When comparing two different groups of riders taking the same “high frequency” trip, the mean age of the two groups was generally not statistically the same.
- When comparing the same rider group between two different “high volume” stations, the mean age of the two groups was generally not statistically the same.

A baseline K-means clustering model with the start station location, end station location, age, gender and user type was developed. As the number of clusters was increased, this model subdivided the trips taken by subscribers of each gender based on location and age. These clusters provide insight into solving the business problem as they identify stations where advertisements can be placed to target specific demographic groups of riders.

A K-means clustering model with start time and stop time for each trip was also developed. As the number of clusters was increased, this model subdivided the trips taken by subscribers of each gender into time bands. Each of these clusters covered the whole Citi Bike service area and the entire age range. These clusters do not provide insight into solving the business problem as they do not break up the trips based on age or location so that stations can be selected to target specific demographic groups of riders.

10.0 Recommendations for the Client

Based on the results of this analysis, the following three recommendations for addressing the business problem have been made:

<u>Recommendation 1 – Focus on Male Specific Advertising.</u>

From the descriptive statistics and data visualization in Section 5.4, it was observed that:

- 70% of all trips were taken by males and 23% of all trips were taken by females. The remaining 7% of all trips were taken by riders of unknown gender.
- A comparison of the top 100 stations for male total volume and the top 100 stations for female total volume showed that there are 76 stations in common.

From the clustering performed in Section 7.0 for k = 10, it was observed that:

- The male subscriber trip clusters (5, 0, 6, 8 and 3) shown in Figures 15, 16 and 17, cumulatively cover the same geographic areas as the female subscriber trip clusters (1, 9, and 7) shown in Figures 18 and 19.
- The male customer trip cluster 2 shown in Figure 20 covers the same area as the female customer trip cluster 4 shown in Figure 21.

Based on these two sets of observations, it is clear that the majority of the trips are taken by males and that the female trips are starting and ending in geographic areas where male trips are occurring. Thus, any station advertising targeted specifically at female users will most likely be at the expense of targeting a much greater number of male users. In the absence of a female specific advertising campaign, the default advertising strategy across all stations should be to target male users as they are the majority.

Recommendation 2 – Focus on Subscriber Specific Advertising.

From the descriptive statistics and data visualization in Section 5.3, it was observed that:

- 88% of all trips were taken by subscribers and 12% of all trips were taken by customers.
- A comparison of the top 100 stations for total volume of customers and the top 100 stations for total volume of subscribers showed that these two groups have 42 stations in common.
- Each of the 200 most frequently taken subscriber trips were taken in excess of times 967.
- Each of the 200 most frequently taken customer trips were only taken in excess of 258 times.

From the clustering performed in Section 7.0 for k = 10, it was observed that:

- The customer trip clusters 2 and 4 shown in Figure 20 and Figure 21 respectively are comprised of trips across the Brooklyn Bridge, inside of Central Park and along the waterfront from Battery Park to as far north as the 495 bridge to New Jersey.

Based on these two sets of observations, it is clear that the vast majority of trips were taken by subscribers and that the customer trips are very highly concentrated in specific areas. However, many of the waterfront customer trips and central park customer trips overlap with trips present in the female and male subscriber clusters. The subscriber trips are taken at much higher frequencies than customer trips. Thus, it is recommended that all advertising be targeted toward subscribers as the vast majority of users are subscribers and the high frequency trips in the customer clusters often overlap with high frequency trips in the subscriber clusters.

Recommendation 3 – Demographic Targeting Strategies Using Clusters.

The results of the clustering performed in Section 7.0 for k = 10 can be used as follows to target specific user demographics in specific geographic areas.

Male Subscribers in Midtown Manhattan

Male Subscribers in Midtown Manhattan are divided into two clusters: Cluster 5 ages 42 to 80 and Cluster 0 ages 16 to 48. The following strategies can be used to target male subscribers:

- Males aged 42 to 80 can be targeted by placing advertisements at the stations which are start points and end points for the high frequency trips in cluster 5.
- Males aged 16 to 48 can be targeted by placing advertisements at the stations which are start points and end points for the high frequency trips in cluster 0.
- Males aged 16 to 80 can be targeted by identifying the high frequency trips that occur in both cluster 5 and cluster 0. Advertisements that appeal to this age range can be placed at the stations which are the start points and end points for these trips.

Male Subscribers in Lower Manhattan and south end of Midtown Manhattan

Male Subscribers in Lower Manhattan and the south end of Midtown Manhattan are divided into two clusters: Cluster 6 ages 42 to 80 and Cluster 8 ages 16 to 43. The following strategies can be used to target male subscribers:

- Males aged 42 to 80 can be targeted by placing advertisements at the stations which are start points and end points for the high frequency trips in cluster 6.
- Males aged 16 to 43 can be targeted by placing advertisements at the stations which are start points and end points for the high frequency trips in cluster 8.
- Males aged 16 to 80 can be targeted by identifying the high frequency trips that occur in both cluster 6 and cluster 8. Advertisements that appeal to this age range can be placed at the stations which are the start points and end points for these trips.

Female Subscribers in Midtown Manhattan and Lower Manhattan

Female Subscribers in Midtown Manhattan and Lower Manhattan are divided into two clusters: Cluster 1 ages 42 to 80 and Cluster 9 ages 16 to 42. The following strategies can be used to target female subscribers:

- Females aged 42 to 80 can be targeted by placing advertisements at the stations which are start points and end points for the high frequency trips in cluster 1.
- Females aged 16 to 42 can be targeted by placing advertisements at the stations which are start points and end points for the high frequency trips in cluster 9.
- Females aged 16 to 80 can be targeted by identifying the high frequency trips that occur in both cluster 1 and cluster 9. Advertisements that appeal to this age range can be placed at the stations which are the start points and end points for these trips.

Subscribers in Brooklyn and south end of Lower Manhattan

Subscribers in Brooklyn and the south end of Lower Manhattan are divided into two clusters: Cluster 7 Females ages 16 to 79 and Cluster 3 Males ages 16 to 74. The following strategies can be used to target subscribers in this particular area:

- Females can be targeted by placing advertisements at the stations which are start points and end points for the high frequency trips in cluster 7.
- Males can be targeted by placing advertisements at the stations which are start points and end points for the high frequency trips in cluster 3.

Customers

Customers across the entire service area are divided into two clusters: Cluster 2 Males and Cluster 4 Females. From the Figure 20 and Figure 21, it is clear that the trips taken are almost identical.

- Females can be targeted by placing advertisements at the stations which are start points and end points for the high frequency trips in cluster 2.
- Males can be targeted by placing advertisements at the stations which are start points and end points for the high frequency trips in cluster 4.

11.0 Future Work

The following recommendations for further analysis have been made:

1. Reduce the data set to the user type and gender of the target demographic. The clustering algorithm will focus on finding groups of similar trips within the target demographic which will likely provide more insight than the current clustering which was performed on the entire data set. The clustering algorithm should run faster as it will have two less attributes to handle. Reducing the data set further by reducing the age range of the trips could also be considered.

2. Remove low frequency trips and low volume stations from the data set. Removing low frequency trips and low volume stations will allow the clustering algorithm to focus on finding groups of similar trips within the trips that are most frequently taken. Thus, the analysis will be more focused on the trips and stations that are most relevant to the business problem.

3. Use a full year of data. The analysis presented in this report is based on the most recent six months of system data (January 2019 to June 2019). This data range was selected as it reflects the most recent locations of bike stations within the system and provides an amount of data that is easily manageable with available computing resources. However, this range omits six months of the year (July to December) which could have substantially different trip data than the six months of the year that were analyzed. Thus, it is recommended to use a full year of data.