Bike Share Advertising in New York City

Springboard—Intermediate Data Science with Python

Andrew Harris

Problem Statement

- Citi Bike is a public bicycle sharing system operating in New York City.
- Citi Bike has over 800 bicycle sharing stations which have space for small billboard advertising.
- Citi Bike has collected data for every trip taken from July 2013 onward.
- This trip data can be segmented based on rider type (short term customer vs. annual subscriber), gender, age, start station and end station so that the station user demographics can be determined.
- Station user demographics can be used to support marketing activities.

Client Description

- The client for this problem is the Citi Bike organization.
- Citi Bike can use the station user demographics determined from segmentation 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 regarding the user demographic of each station.

Data Set

- Available online at https://www.citibikenyc.com/system-data.
- Each line of data is an individual trip with 15 attributes:

➢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 (Customer or Subscriber), birth year and gender (male or female or unknown).

- Six months of data (January 2019 to June 2019) containing over nine million trips was used as the "starting point" for this analysis.

Data Wrangling

- Rows with NAN values were removed.
- Birth years indicating an age greater than 80 were removed.
- Trip durations greater than one hour were removed.
- Trips durations were corrected as needed to match difference between start time and end time.
- Start station and end station descriptive columns (id, name, latitude, longitude) were checked to ensure each id had only one set of latitude coordinates.
- 17 stations that didn’t appear as both start stations and end stations were removed.

Data Visualization and Descriptive Statistics

- 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 away from Midtown Manhattan

Data Visualization and Descriptive Statistics

Gender:

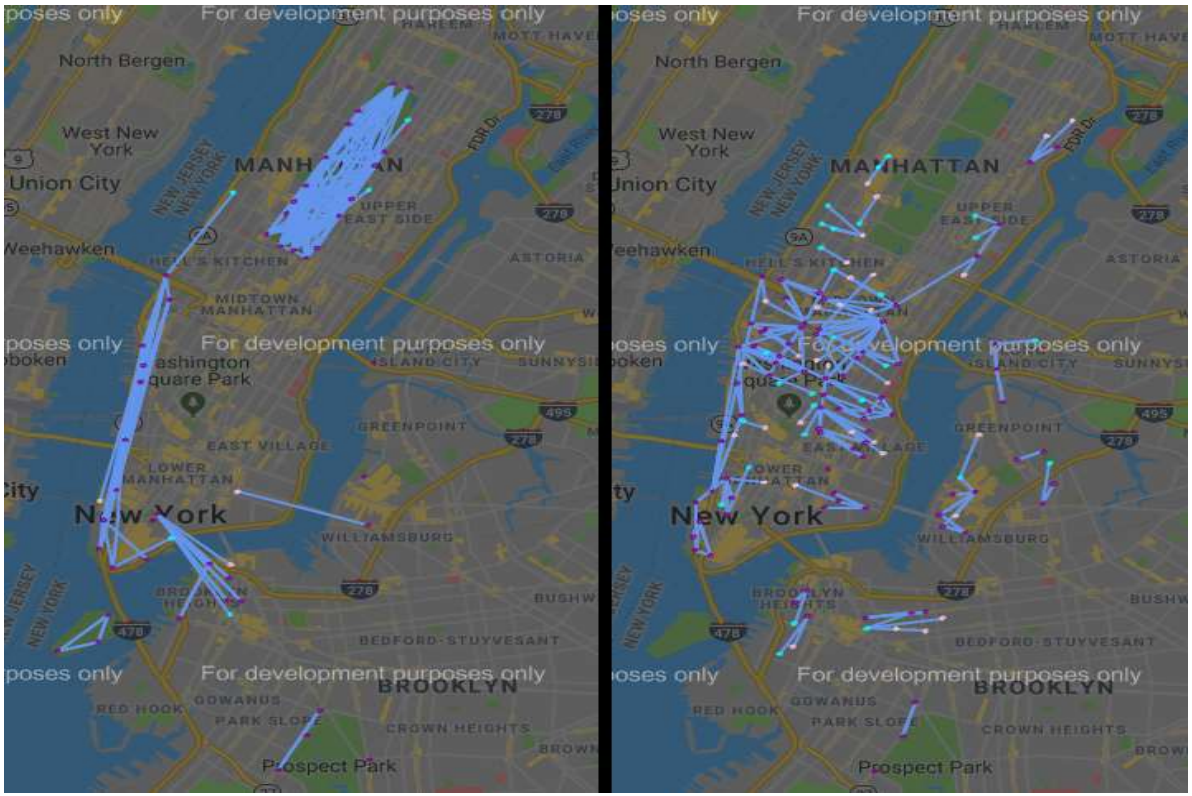
- 70% of trips were taken by males. 23% of trips were taken by females.
- 7% of trips were taken by riders of ‘unknown’ gender.
- ‘unknown’ gender trips removed from data set.

User Types - Customers and Subscribers:

- Customer – user with 24 hour pass or 3 day pass.
- Subscriber – user with annual pass.
- 88% of all trips were taken by subscribers and 12% of all trips were taken by customers.

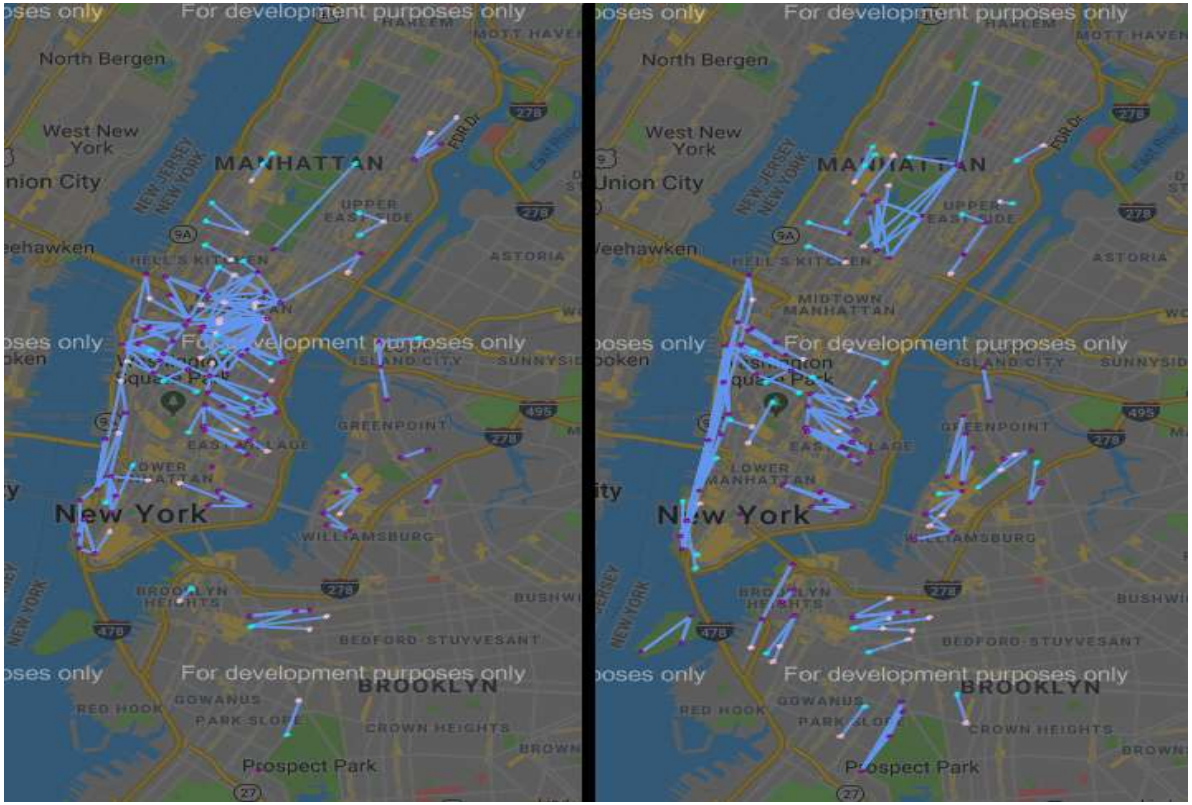
Data Visualization and Descriptive Statistics

Customer (left) & Subscriber (right) – 200 Most Frequently Taken Trips



Data Visualization and Descriptive Statistics

Male (left) & Female (right) – 200 Most Frequently Taken Trips



Machine Learning – K-Means Clustering

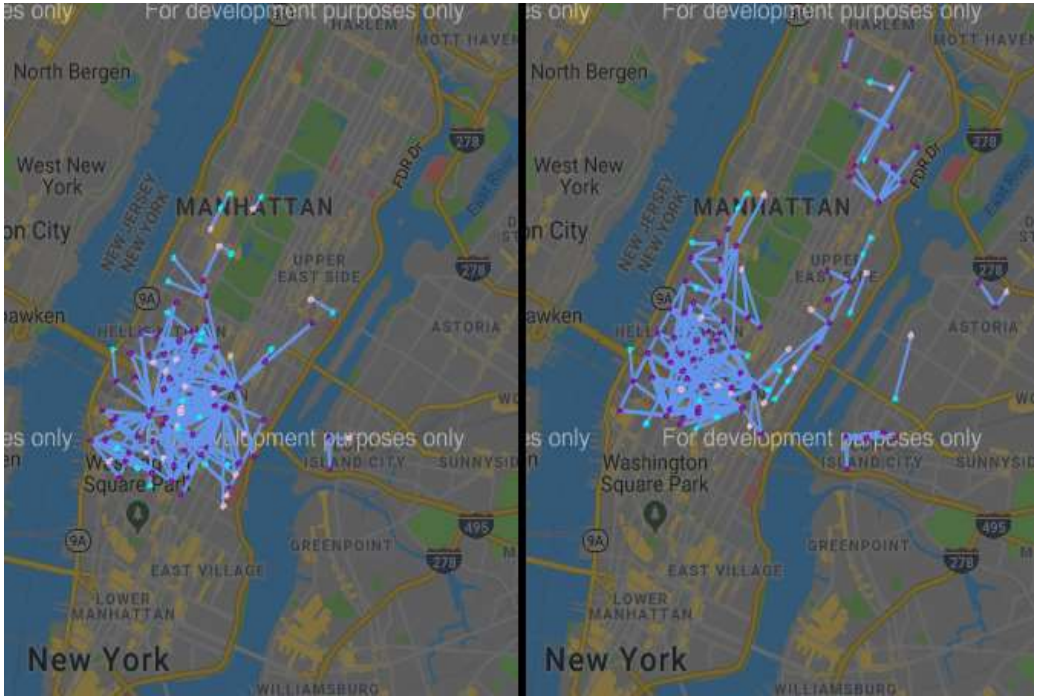
- Data set reduced to five attributes to create a “baseline” model: start station mds, end station mds, user type (Customer or Subscriber), gender (Male or Female) and age.
- User type and gender represented as 0 and 1.
- Age scaled to be a decimal number ranging from 0 to 1.
- Start station mds and end station mds – numbers were assigned to each station using multi-dimensional scaling with the Haversine distance between stations as a dissimilarity measure. These numbers were scaled to range from 0 to 1.
- K-Means clustering performed for k= 2 to k=20.
- 200 most frequently taken trips in each cluster were plotted.

Machine Learning – Results for k = 10

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.

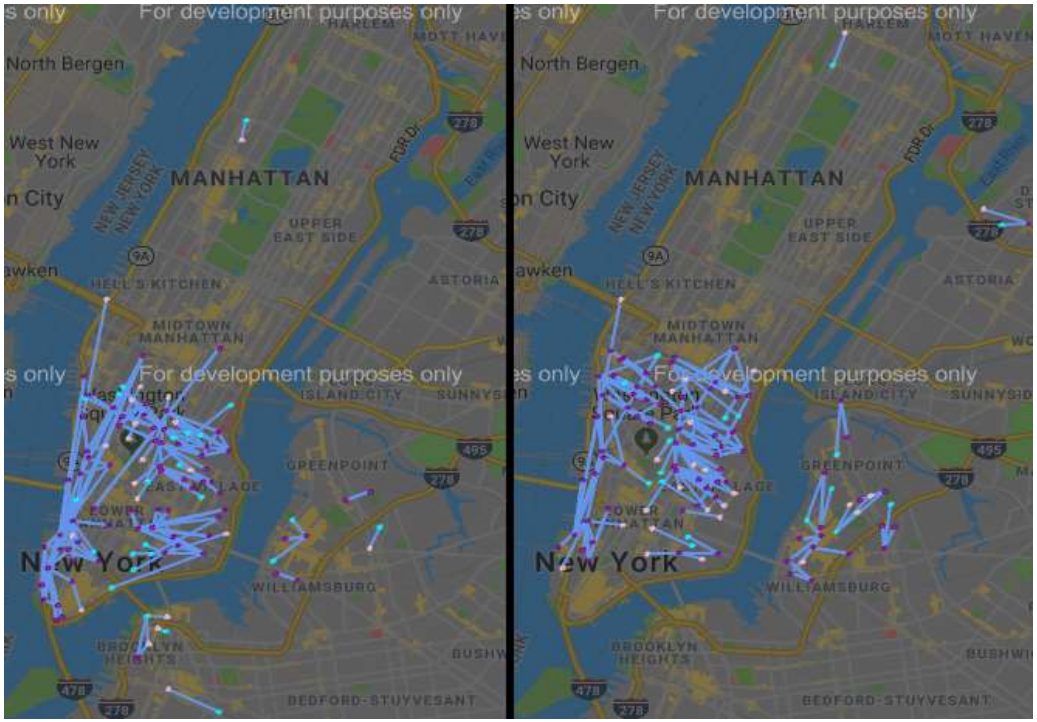
Machine Learning – Results for k = 10

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



Machine Learning – Results for k = 10

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

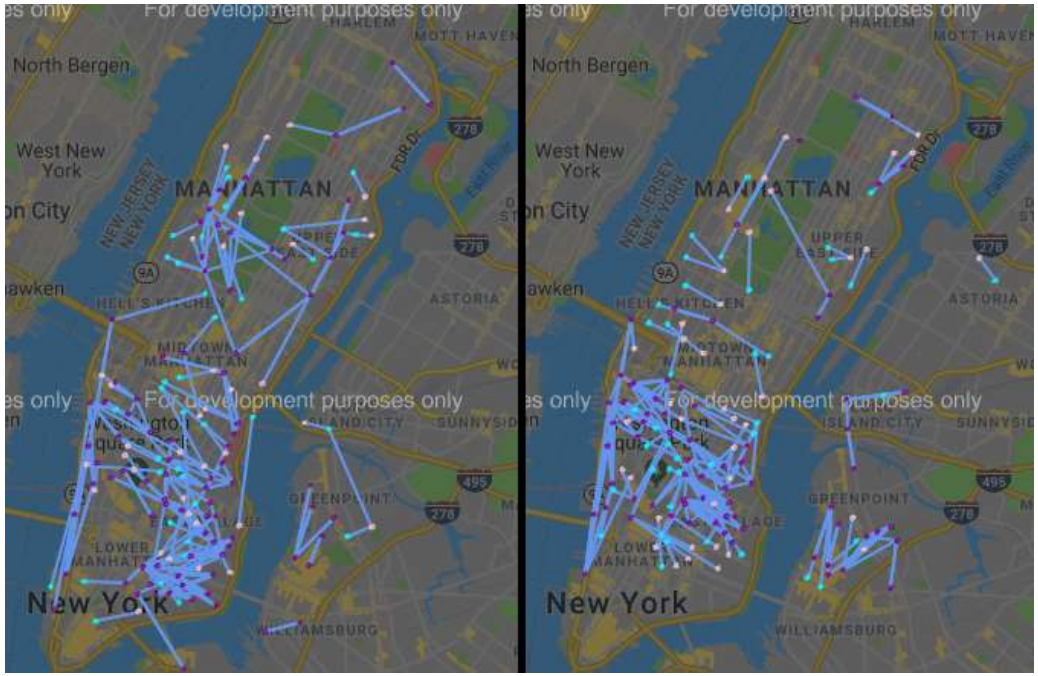


Machine Learning – Results for k = 10

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

Machine Learning – Results for k = 10

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

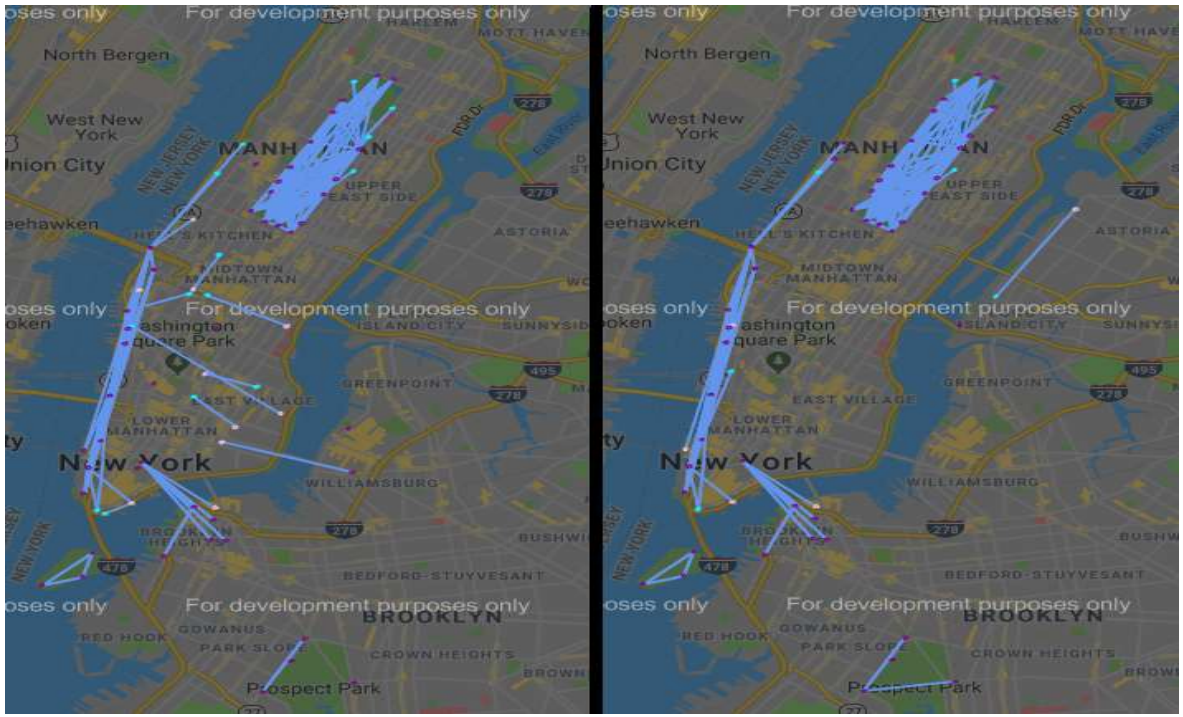


Machine Learning – Results for k = 10

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

Machine Learning – Results for k = 10

Cluster 2 – Male Customer Trips Ages 16 to 80 (left) and Cluster 4 – Female Customer Trips Ages 16 to 80 (right)



Machine Learning – K-Means Clustering

- Time model was also used for k =2 to k=20.
- start time and end time converted to “seconds past midnight”.
- Both times were converted to a two component representation using sine and cosine to reflect cyclical nature of time of day. Both components for both times were then scaled to range from 0 to 1.
- Result: subscriber trips for each gender were subdivided based on start time and end time as the value of k increased. Each cluster covered the entire geographic range of the service area and the entire age range of the riders.
- Not a useful result for solving the business problem.

Conclusions

Baseline K-means Clustering Model:

- As K 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.

K-means Clustering Model with Start Time and End Time:

- As K 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

Recommendations for Client

Recommendation 1 - Focus on Male Specific advertising:

- Majority of trips are taken by Males.

Recommendation 2 – Focus on Subscriber Specific Advertising:

- Majority of trips are taken by Subscribers.

Recommendation 3 – Demographic Targeting Strategies Using Clusters:

- Example: 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.

Future Work

There are three recommendations for further analysis:

1. Reduce the data set to the user type and gender of the target demographic.
2. Remove low frequency trips and low volume stations for the data set.
3. Use a full year of data so that trips taken reflect all twelve month of the year.