Analysis of Taxi Ride Migration to New Subway Lines

By: Abhishek Modoor,
Pranay Gupta

Overview

For our project, we decided to look at NYC taxi ride share data. We wanted to analyze the volume and demand of taxi rides across NYC in order to determine where we can implement new modes of transportation (ideal being subway). The intention was that it can reduce traffic for the roads, and also allow people to take cheaper modes of transport if their route allowed them do so.

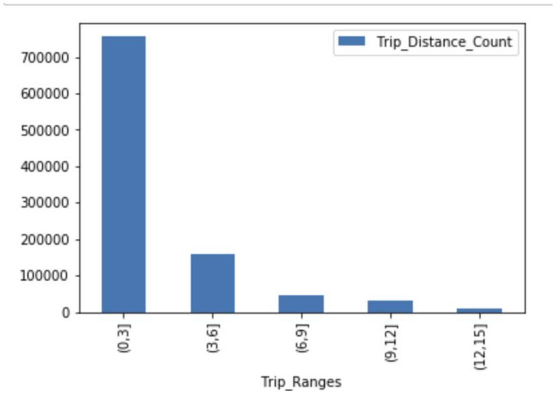
We extracted our datasets from kaggle.com and NY.gov. We used yellow taxi ride share data from March 2018, green taxi ride share data from 2015, and subway entrance and exit data. We focused our first half of the project with the goal of gaining key insights such as ideal time of day, volume of demand across different boroughs, average trip distance, and pickup vs drop-off statistics among others.

By taking these patterns, we wanted to implement a machine learning experiment to see if we could predict ideal locations within NYC where we could introduce new subway lines, add them to existing stations, or create other modes of transport within the area (bikes, buses, ride share programs).

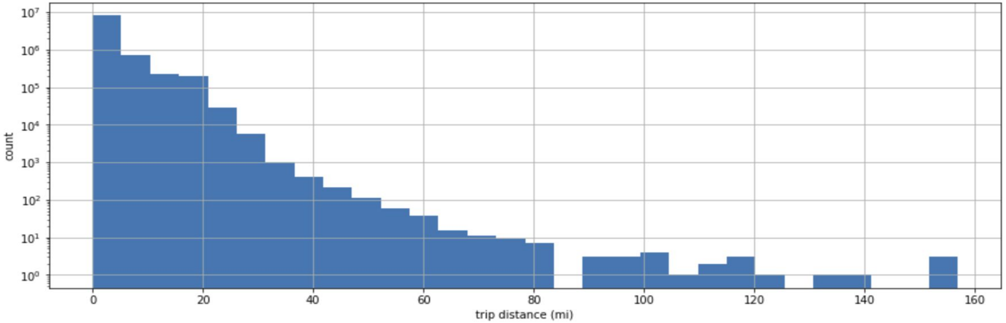
Key Insights

For the first plot, we look at the trip distances within 15 miles.

We see that most of the trips are within 0 and 3 miles. This tells us that we can create subway lines within NYC since it can be quicker to get to areas within 3 miles.



Overall, we see that the distribution of trip average in the second plot corroborates the pattern of the first.

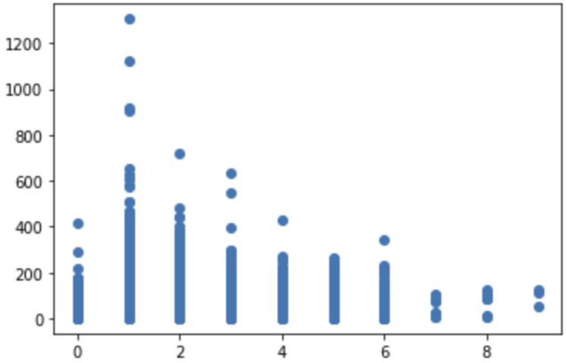


Key Insights

In the first plot, we see the variability of trip cost based on passenger count.

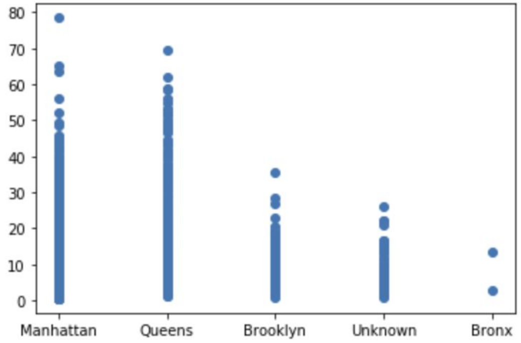
We see that one passenger represents the largest variability.

Creating subway lines would help because of the standard $2.75 MTA fare.



For the second plot, we see the fare rate per passenger based on the borough.

We see that Manhattan has some of the highest fare rates, so we want to focus our analysis to Manhattan over other boroughs.

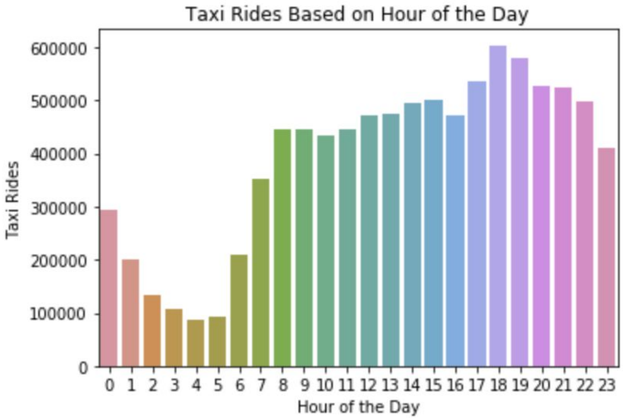


Key Insights

We saw that the times between 3 pm and 10 pm, had high volume of taxi rides.

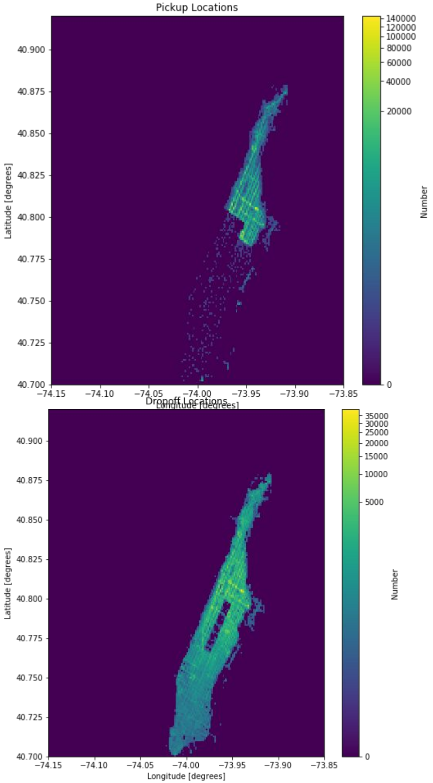
We decided to focus our analysis for the machine learning experiments during

the day-hours when people are more likely to take a subway.



Key Insights

This graph analyzes the volume of taxi activity across NYC with pickups as the first plot and drop-offs as the second plot. We focused our analysis to uptown, and specifically past 110th street.

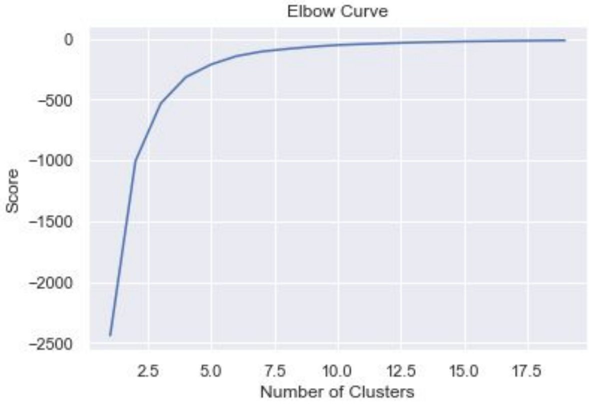


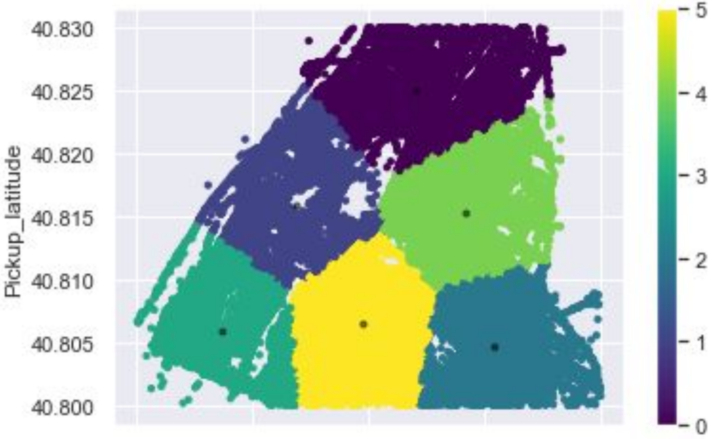
Key Machine Learning Insights

Due to our area of focus past 110th street, we implemented our machine learning analysis within the district of Harlem.

We performed a k-means clustering algorithm on the data points within Harlem to identify locations for potential subway lines or modes of transport.

After performing an elbow method (left plot 1), to identify 6 as our optimal cluster amount, we performed k-means. We identified 6 centroids and we inferred we would want to place our subway lines or modes of transport there. We saw that prominent subway lines were there as well so it could as viable alternatives.





Conclusions

After pursuing k-means within Harlem, we found 6 more potential locations to open subway lines. Ideally, we would choose centroids as our new potential subway lines because of their close proximity to the cluster area.

However, the bottleneck of selective subway routes can prove to make it harder to choose an ideal location. For example, B,1, and A only pass through select stations.

The algorithm doesn’t account for specific mobility patterns (selective subway lines). We would have a difficult time to ascertain whether or not an individual near an existing centroid/point would necessarily use that station. This would cause accuracy of predicting an arbitrary individual trajectory from pickup to drop-off to decrease. We would need more time to perform analysis to mitigate this bottleneck.

Acknowledgements

Libraries used:

Pandas
Matplotlib.pyplot
Seaborn
Numpy
Altair
Folium
Utils.py: (From Rutgers University, Computer Science, Introduction to Data Science CS 439, DeMelo, Pandas Assignment 2)
https://www.kaggle.com/muonneutrino/nyc-taxis-eda-and-mapping-position-to-borough#data