Table of Contents

1   Introduction

New York City (NYC) is the most populous city in the United States. It is also the most densely populated major city in the United States. Situated on one of the world's largest natural harbors, New York City consists of five boroughs. The five boroughs: Brooklyn, Queens, Manhattan, Bronx, and Staten Island – were consolidated into a single city in 1898.

The objective of this project is to understand the traffic in different boroughs of New York City and to try and categorise the various zones within various boroughs as being: office destinations, residential destination, popular brunch destination or party destination according to it's popularity given the time-range and the day of the week.

I have used data generated by Uber, an online transportation network company. This data is available at Kaggle under the heading: 'Uber Pickups in New York City', provided by user: 'FiveThirtyEight' who obtained the data from the NYC Taxi Limousine Commission (TLC) by submitting a Freedom of Information Law request on July 20, 2015.

The directory downloaded contains data on over 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015. I have used ® to extract knowledge and mine the dataset.

2   Problem Understanding

There is a lot of information stored in the traffic flow data of any city. This data when mined over location can provide information about the major attractions of the city, it can help us understand the various zones of the city such as residential areas, office/school zones, highways, etc. This can help governments and other institutes plan the city better and enforce suitable rules and regulations accordingly. For example, a different speed limit in school and residential zone than compared to highway.

The data when monitored over time can help us identify rush hours, holiday season, impact of weather, etc. This knowledge can be applied for better planning and traffic management. This can at a large, impact the efficiency of the people in the city and can also help avoid disasters, or at least faster redirection of traffic flow after accidents.

3   Data Understanding

The dataset downloaded contains, roughly, four groups of files:

1. Uber trip data from 2014 (April - September), separated by month, with location infor- mation (longitude - latitude values)

2. Uber trip data from 2015 (January - June), with less fine-grained location information

3. Non-Uber FHV (For-Hire Vehicle) trips. The trip information varies by company, but can include day of trip, time of trip, pickup location, driver's for-hire license number, and vehicle's for-hire license number.

4. Aggregate ride and vehicle statistics for all FHV companies (and, occasionally, for taxi companies)

However this project used two groups of files that was generated by Uber:

- Uber trip data from 2014 (April - September)

- Uber trip data from 2015 (January - June)

3.1 Uber trip data from 2014

There are six files of raw data on Uber pickups in New York City from April to September 2014 (uber-raw-data-xyz14.csv). The files are separated by month and each has the following columns:

- Date/Time : The date and time of the Uber pickup

- Lat : The latitude of the Uber pickup

- Lon : The longitude of the Uber pickup

- Base : The TLC base company code affiliated with the Uber pickup

3.2 Uber trip data from 2015

Also included is the file uber-raw-data-janjune-15.csvThis file has the following columns:

- Dispatching base num : The TLC base company code of the base that dispatched the Uber

- Pickup date : The date and time of the Uber pickup

- Affiliated base num : The TLC base company code affiliated with the Uber pickup

- locationID : The pickup location ID affiliated with the Uber pickup

3.3 Data for NYC Zone lookup

For coarse-grained location information from the pickups in 2015, the file taxi-zone-lookup.csv shows the taxi Zone (essentially, neighborhood) and Borough for each locationID contained in the Uber trip data from 2015.

4 Data Preparation

4.1 2014 trip data

- A batch of data, apr june14 is prepared. It contains 1880795 observations after binding rows from the file containing data from April, May and June of 2014.

- The data from July, August and September are put into a separate batch, july sep14. This batch contains 2653532 observations in total.

- Any NA values present were omitted from both the datasets prepared.

- The DateTime column is parsed using 'lubridate' package that makes it easier to work with date_time formatted data. The different parsed values (Year, Month, Date, Hour, Minute and Second) is then factored into the datasets to be used later.

4.2   2015 trip data

- The data for 2015 is contained in a single .csv file unlike the 2014 data.

- After reading the file, the Pickup date is parsed using 'lubridate' package. 'Month' of each observation is filtered to divide the dataset into two batches (Training Set and Test Set).

- A left-join is performed upon the locationIDin the 2015 data with the LocationIDfrom taxi-zone-lookup.csv to add columns Borough, Zone, service zone corresponding to each pickup location ID in the 2015 datasets.

- A batch of data, jan_mar15 is prepared. It contains 6477194 observations after binding rows from the file containing data from January, February and March of 2015. This is used as the Training Set later.

- The data from April, May and June are put into a separate batch, apr_june15. This batch contains 7793285 observations in total. This is used as the Test Set later.

- Any NA values present were omitted from both the datasets prepared.

5   Modeling

5.1   Borough Formation

Training Set: apr_june14
Test Set: july_sep14

The data in the columns Lat and Lon are the latitude and longitude of the pickup locations. This pair of data uniquely identify each location on the geographical coordinate system.
Since we have 1880795 locations, we can try to identify various Boroughs in New York City based on the density of pickup locations. In the project, Density-based spatial clustering of applications with noise (DBSCAN) and K-Mean Clustering were applied and tested for this purpose.

DBSCAN is a density-based clustering algorithm. Given a set of points in some space such as Latitude Longitude pair, it groups together points that are closely packed together (points with many nearby neighbors). It marks the point that lie alone/separate in low-density regions as outliers (whose nearest neighbors are too far away).

K-means is one of the simplest unsupervised learning algorithms for clustering. K-means clustering aims to partition n observations into k (fixed apriori) clusters in which each observa- tion belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

The output (the various clusters formed) of the training set is then matched against the clusters obtained from the test set to check for consistency, thus to determine the correctness of the model. We can also check the output against the real geographical map to see the efficiency of using these methods (DBSCAN, K-Mean).

5.2 Growth Rate

Training Set: Uber trip data from 2014 (April - September)
Test Set: Uber trip data from 2015 (January - June)

The data for the various months in 2014 are taken to visualise the growth in the number of Uber Pickup requests per month for each borough.

This growth is then compared against the growth for 2015 to see if we can draw any inferences based on this model.

5.3 Popular Zones

Training Set: jan mar15
Test Set: apr june15

The data is filtered according to various time durations and day of the week:

- Weekday (Monday - Thursday)

    - Morning: 07 - 10 A.M. – Residential Locations
    - Evening: 16 - 19 P.M. – Office Locations

- Weekend (Friday - Sunday)

    - Morning: 11 A.M. - 14 P.M. – Brunch Locations
    - Evening: 21 P.M. - 01 A.M. – Party Locations

The various boroughs are analysed to identify the top three boroughs that have the most significant amount of trips.
The data filtered according to the time and day of the week is then analysed to find the top three popular destinations/zones in the most significant boroughs and then categorised accordingly.

This model is then deployed to the test data and the result for the training and test data is checked for consistency to determine the correctness of the model proposed.

6 Evaluation

6.1 Model 1: Borough Formation

6.1.1 DBSCAN

DBSCAN seemed to work poorly on the Training Data to identify various boroughs of NY City.

Advantages of DBSCAN:

- We do not have to provide the number of clusters as an input parameter, unlike other clustering algorithm such as K-Nearest Neighbours.

- It can identify clusters of arbitrary shape.

- A data point with no close neighbours is assigned noise rather than its nearest cluster.

Disadvantages of DBSCAN:

- We have to provide 'eps' value, which is the neighbourhood radius. When this was taken to be a higher value, it identified all points to be mostly in the same cluster. However, when a smaller value was provided, it made many clusters (Figure 1: with eps = 0.009). Hard to define a good value for the same.

- It's performance for the size of the dataset we had was very slow. Hence, had to be tested on a dataset much smaller than the size of the training data. Thus, validating this on the test data did not seem to be justified given the time and results obtained.
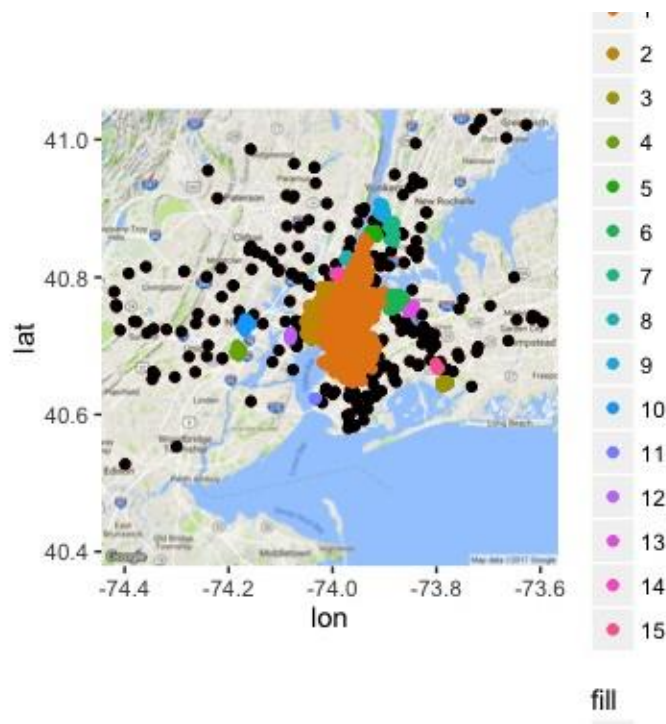
Figure 1: Boroughs in NYC using DBSCAN on 'apr14' data

6.1.2 K-Means Algorithm

Given that we know the number of boroughs already, K-Means Algorithm was a good can- didate for the modeling. Infact, the results from K-Means provided a good estimate of the boundary of various boroughs but with respect to the Uber Data on the training set. These boroughs were then matched against the real Boroughs and can be named accordingly.

The key features of k-means which make it efficient are often regarded as its biggest draw-backs:

- Euclidean distance is used as a metric and variance is used as a measure of cluster scatter.

- The number of clusters k is an input parameter: an inappropriate choice of k may yield poor results. However, we knew the target number of boroughs we were expecting.
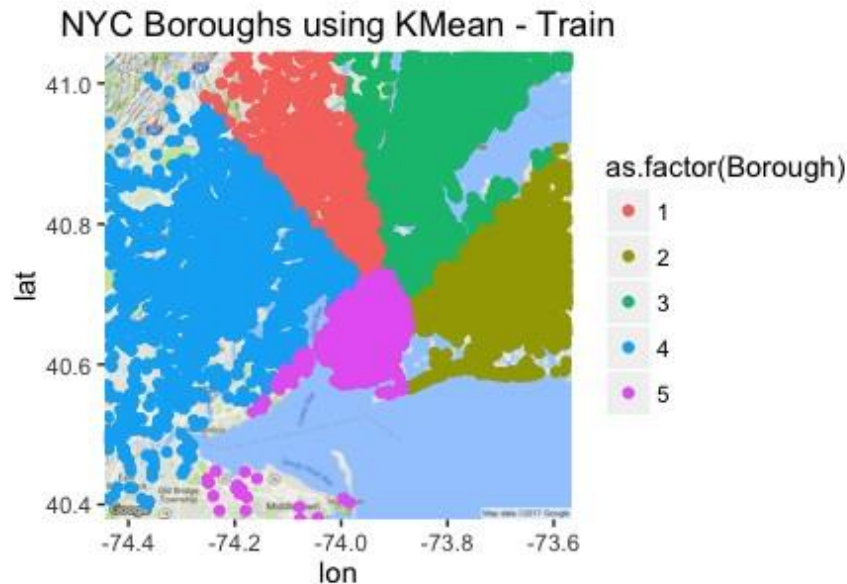

Figure 2: Boroughs in NYC using K-Means on 'apr june14' data

In Figure.2 the boroughs (clusters) formed is matched against the real boroughs and thus the cluster number corresponds to the following Boroughs:

1. Manhattan

2. Queens

3. Bronx

4. Staten Island

5. Brooklyn

6.2   Model 2: Growth Rate

For the boroughs identified above, the growth rate or demand rate for the months of 2014 (April - September) was plotted. As, seen in Figure.3, the growth has been rising at a consistent rate for Cluster 4 (Staten Island). However, Cluster 1 (Manhattan) seemed to have a steep fall from Month June to July. Here, the months starts from April (Month 1) upto September (Month 6). Similarly, Cluster 2 (Queens) shows a steep growth for the same months.
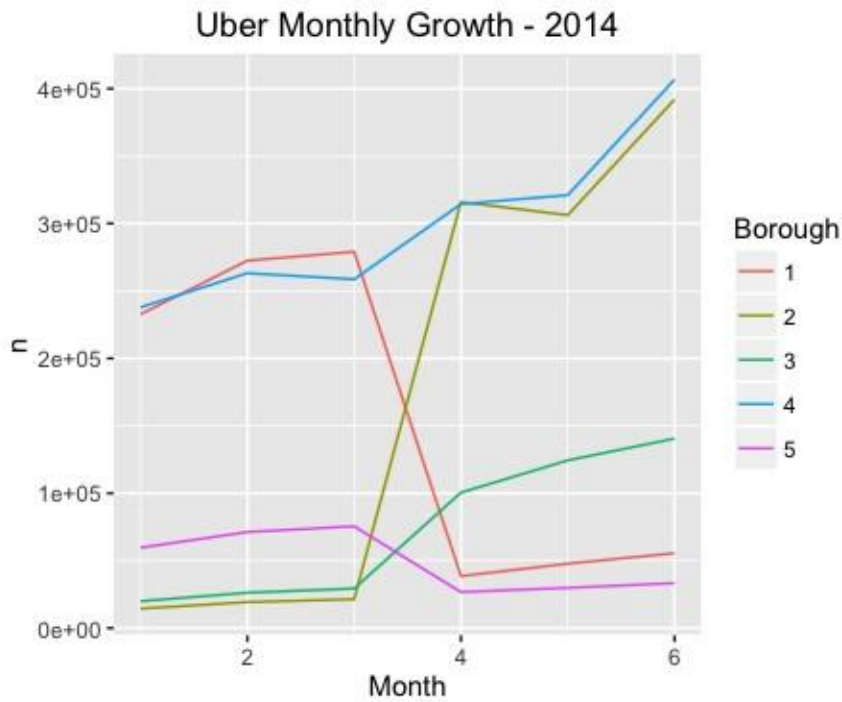
Figure 3: Growth Rate for 2014 data

6.3 Model 3: Popular Zones

6.3.1   Residential Locations

According to the analysis on the Training data, for the category of Residential Locations it is found that the Top 3 boroughs are (in decreasing order): Manhattan, Brooklyn, Queens. However, these may not always be accurate residential locations (like in case of Queens).

- Top 3 zones (Manhattan): East Village, TriBeca/Civic Center, Upper East Side North.

- Top 3 zones (Brooklyn): Park Slope, Williamsburg, Green Point.

- Top 3 zones (Queens): Laguardia Airport, JFK Airport, Astoria.

**See Appendix for further information on the same.

6.3.2   Office Locations

According to the analysis on the Training data, for the category of Office Locations it is found that the Top 3 boroughs are (in decreasing order): Manhattan, Brooklyn, Queens. However, these may not always be accurate Office locations (like in case of Queens) and could possibly also be school or other work locations.

- Top 3 zones (Manhattan): Midtown Centre, Union Square, Midtown East.

- Top 3 zones (Brooklyn): Williamsburg, Park Slope, Green Point.

- Top 3 zones (Queens): Laguardia Airport, JFK Airport, Long Island City.

6.3.3 Brunch Locations

According to the analysis on the Training data, for the category of Brunch Locations it is found that the Top 3 boroughs are (in decreasing order): Manhattan, Brooklyn, Queens.

- Top 3 zones (Manhattan): East Village, TriBeca/Civic Center, West Village.

- Top 3 zones (Brooklyn): Williamsburg, Park Slope, Williamsburg (South Side).

- Top 3 zones (Queens): Laguardia Airport, JFK Airport, Astoria.

**See Appendix for further information on the same.

6.3.4 Party Locations

According to the analysis on the Training data, for the category of Party Locations it is found that the Top 3 boroughs are (in decreasing order): Manhattan, Brooklyn, Queens.

- Top 3 zones (Manhattan): East Village, West Village, Civic Center.

- Top 3 zones (Brooklyn): Williamsburg (North Side), Williamsburg (South Side), Park Slope.

- Top 3 zones (Queens): JFK Airport, Laguardia Airport, Astoria.

**See Appendix for further information on the same.

7   Deployment & Conclusions

7.1   Model 1: Borough Formation

7.1.1   Deployment

Model 1 is deployed on the Test data for 2014 (july sep14) by using K-Means clustering to find 5 clusters which corresponded to the 5 Boroughs of NYC.

7.1.2   Conclusion

- The boroughs/clusters formed were consistent to the Training Data.

- The deployment of Model 1 using K-Means on both the training and test set provided consistent and satisfactory results.

- Even when compared to the real borough mapping, the results were impressive. The boroughs formed however covered some extra regions which is due to the categorisation of various areas under NYC by Uber such as bridges, outer regions, suburban areas which might not necessarily fall under the jurisdiction of NYC in real life.
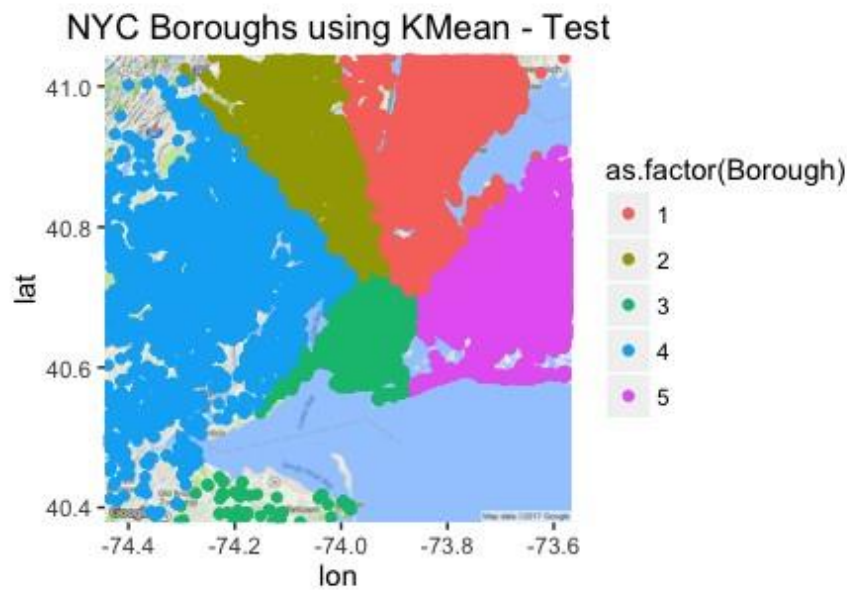
Figure 4: Boroughs in NYC using K-Means on 'july sep14' data

In Figure.3 the boroughs (clusters) formed is matched against the real boroughs and thus the cluster number corresponds to the following Boroughs:

1. Bronx

2. Manhattan

3. Brooklyn

4. Staten Island

5. Queens

7.2   Model 2: Growth Rate

7.2.1   Deployment

Model 2 is deployed on the Test data, the data for 2015. The growth rate is plotted for the different boroughs.

7.2.2   Conclusion

- Compared to the 2014 (Training data), the growth rate for 2015 is more stable in terms of growth rise and fall.

- The stability in the chart/rate could be due to the rise of the brand awareness of Uber amongst the people. This could be due to the publicity (negative and positive) received by Uber which given known market trends for 2015, most likely increased when compared to 2014.
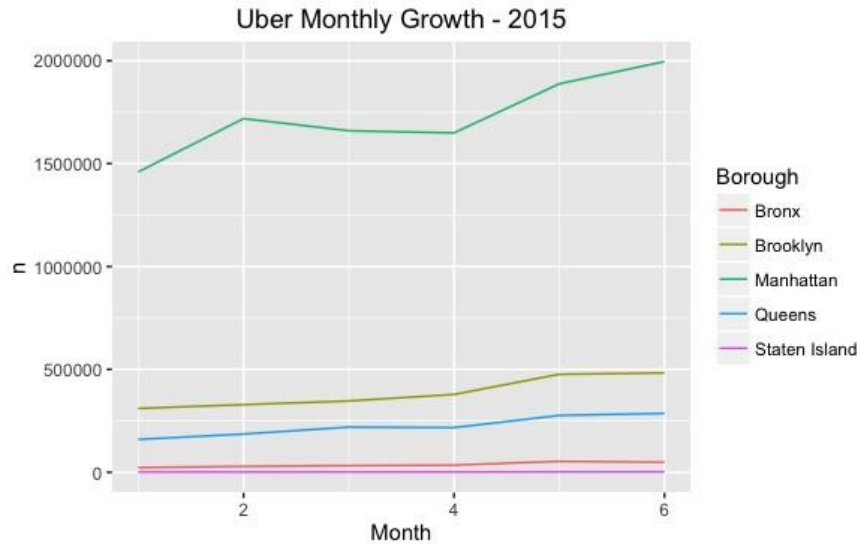
Figure 5: Growth Rate for 2015 data

7.3 Model 3: Popular Zones

7.3.1 Deployment

Model 3 is deployed on the Test data for 2015 (apr june15). The categorisation is done the same way as for the Training data and is checked to see if the model is able to predict consistent results for NYC.

7.3.2 Conclusion

- The model has impressive results on the prediction and consistency when compared to the training data results.

- There was only one inconsistency, for the category 'Brunch Locations' for the borough of Brooklyn. In the training data the Top 3 destination are: Williamsburg, Park Slope, Williamsburg (South Side). However, for the Test Data, it was found to be: Williamsburg, East Williamsburg, Park Slope.

- The inconsistency could point towards changing trends in the data. This on further investigation and support of more data could help to find the reason in the shift of trend. For eg; it could be cause of the rise of a certain neighbourhood due to local or Government initiative, or due to one or many upcoming businesses such as restaurants since it is in the popular Brunch Location.

**See Appendix for further information on the same.

Appendices

Figure 6: Top 3 zones (Manhattan): Residential Locations (Training Data)

Figure 7: Top 3 zones (Brooklyn): Residential Locations (Training Data)

Figure 8: Top 3 zones (Queens): Residential Locations (Training Data)

Figure 9: Top 3 zones (Manhattan): Office Locations (Training Data)

Figure 10: Top 3 zones (Brooklyn): Office Locations (Training Data)

Figure 11: Top 3 zones (Queens): Office Locations (Training Data)

Figure 12: Top 3 zones (Manhattan): Brunch Locations (Training Data)

Figure 13: Top 3 zones (Brooklyn): Brunch Locations (Training Data)

Figure 14: Top 3 zones (Queens): Brunch Locations (Training Data)

Figure 15: Top 3 zones (Manhattan): Party Locations (Training Data)

Figure 16: Top 3 zones (Brooklyn): Party Locations (Training Data)

Figure 17: Top 3 zones (Queens): Party Locations (Training Data)

Figure 18: Top 3 zones (Manhattan): Residential Locations (Test Data)

Figure 19: Top 3 zones (Brooklyn): Residential Locations (Test Data)

Figure 20: Top 3 zones (Queens): Residential Locations (Test Data)

Figure 21: Top 3 zones (Manhattan): Office Locations (Test Data)

Figure 22: Top 3 zones (Brooklyn): Office Locations (Test Data)

Figure 23: Top 3 zones (Queens): Office Locations (Test Data)

Figure 24: Top 3 zones (Manhattan): Brunch Locations (Test Data)

Figure 25: Top 3 zones (Brooklyn): Brunch Locations (Test Data)

Figure 26: Top 3 zones (Queens): Brunch Locations (Test Data)

Figure 27: Top 3 zones (Manhattan): Party Locations (Test Data)

Figure 28: Top 3 zones (Brooklyn): Party Locations (Test Data)

Figure 29: Top 3 zones (Queens): Party Locations (Test Data)