Capstone Project - The Battle of Neighborhoods

1. Introduction

1.1 Background

The New York metropolitan area is one of the world's most populous megacities with more than 8 million residents. The large population and high population density (27,751/sq mi) offers a huge market for wide range of businesses. Moreover, the iconic New York City Subway system is the largest rapid transit system in the world. This subway system is widely used by New Yorkers.

When setting up business a clear study of market is to be done to understand the risks and opportunties. Data analytics and Machine learning can be used to develop insights from data, thereby helping in better decision making.

NYC Foods is planning to set up a network of food trucks across NYC. They plan to capitalise on the high population and robust Subway network of the city. They plan to set up the food trucks near subway stations, and provide low cost fast food to busy subway users. The eventual plan is to have food trucks all across NYC. But to begin with they have decided to start in the borough of The Bronx.

1.2 The Problem

Our client NYC Foods has entrusted us with the project of narrowing down on the most suitable subway stations in The Bronx, where they can put up the first set of food trucks. The best stations will be the ones with low concentration of other food trucks, food stalls or fast food joints near the station as these are our client's direct competitors. The presence of restaurants, diners etc. does not matter as our client's target customers are those who prefer fast food at low cost. So, we need to use data and cluster the subway stations in The Bronx based on the concentration of our client's competitors near the station.

2. Data Description

2.1 Data Sources

The data that will be used and how it will be used in this project:

1. The details all subway stations in New York City is obtained from Spatial Data Repository of NYU. The .json file has names and coordinates of all the subway stations of New York.
2. Our client has decided to begin operations in the borough of The Bronx. The names of stations in The Bronx will be obtained from Metropolitan Transportation Authority.
3. The details of our competitors (fast food joints, food trucks and food stalls) near each subway station will be obtained using Foursquare API.

2.2 Data Cleaning

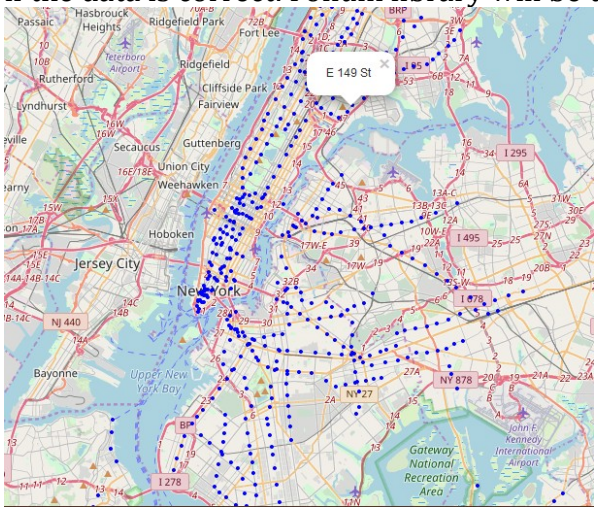
1. From the NYC station json file, a Data frame needs to be created which will have station name, logitude and latitude.
2. To obtain the names of stations in Bronx, the names from the website needs to be cleaned. It has to be done keeping in mind the names in the json file. The periods, hyphen, some names etc needs to corrected to match the json file. This cleaned file will used to narrow down only Bronx stations from the Data Frame.

3. Methodology

We begin by loading the NYC subway stations json file and store it in a Data Frame. Json library will be used for it. It will look like below:

	Stop	Latitude	Longitude
0	Van Cortlandt Park - 242 St	40.889248	-73.898583
1	238 St	40.884667	-73.900870
2	231 St	40.878856	-73.904834
3	Marble Hill - 225 St	40.874561	-73.909831
4	215 St	40.869444	-73.915279

We then use this Data Frame to do basic exploration by plotting them on New York map to see if the data is correct. Folium library will be used to do the plotting. The plot looks like below:



We see that the data looks good and plot reflects actual NYC subway map. So, now we can move on to narrowing down the data to only those subway stations in The Bronx borough. This is because out client has preferred to start their operation in The Bronx.

The names of the Bronx stations obtained has to be cleaned to match that in our original data frame. Once this is done we can narrow the data frame to reflect only Bronx stations. We gain use this data to plot Bronx stations to see if we have got our data correct. The plot looks like this:

We see that we have successfully narrowed down our dataset to the stations in Bronx. Now we need to find out the food joints, especially our client's competitors. This can be done using Foursquare API.

To speed up the data download and to be memory efficient we get details of only food joints. This is done using the foursquare food category code '4d4b7105d754a06374d81259' which will passed in API request.

API request will return details of all the food joints within 500 meters of the station. But our client is only bothered about their competitors which are fast food joints, Burger joints, food trucks, food stalls. We use the Foursquare category id of these to find if each returned food joint is a competitor or not. We categorise them as food truck if they are one and rest all competitors are categorised a fast food joint.

Finally we have data frame with all the competitor details near each subway station in The Bronx. We group this dataframe on station name and get the count of number of competitors (fast food, food truck) near each station. A snap shot of the dataframe looks like this:

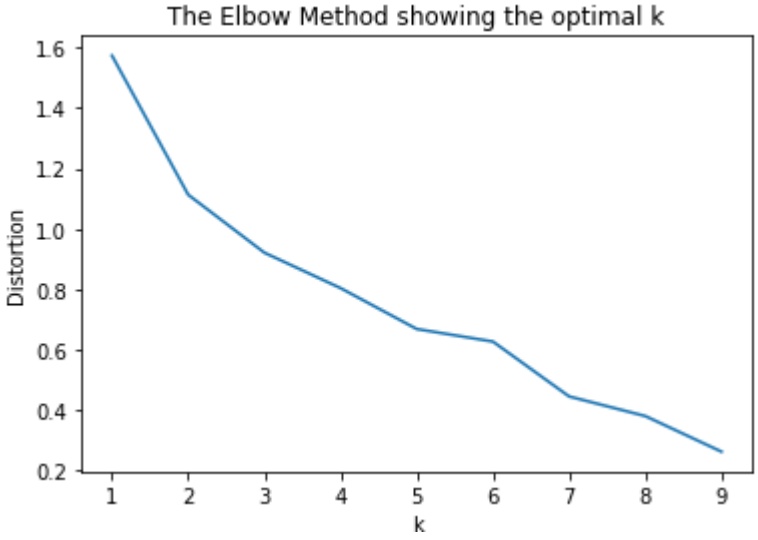
	Subway Station	fast food	food truck
0	138 St - Grand Concourse	1	0
1	149 St - Grand Concourse	4	1
2	161 St - Yankee Stadium	4	4
3	167 St	2	0
4	170 St	3	0
5	174 St	2	1
6	174-175 Sts	1	3
7	176 St	1	1
8	182-183 Sts	2	0
9	183 St	2	0

A copy of this data frame is created and merged with the data frame having all Bronx stations and coordinates. This will look like:

	Subway Station	fast food	food truck	Latitude	Longitude
0	138 St - Grand Concourse	1	0	40.810476	-73.926138
1	149 St - Grand Concourse	4	1	40.816109	-73.917757
2	161 St - Yankee Stadium	4	4	40.813224	-73.929849
3	167 St	2	0	40.818375	-73.927351
4	170 St	3	0	40.827905	-73.925651

This dataframe will be used later to plot cluster.

Now we use machine learning to find the best set of stations near which out client can set up the food trucks. We use K-Means clustering algorithm to accomplish our task. Before that we need to find the right K – the number of clusters. Elbow method is used and 5 is found to be the best value of K.



Now we run the Kmeans clustering. We fit out data into the model and each station will be clustered into one of the 5 categories. The above mentioned dataframe will be manipulated and cluster label of each station is entered into it.

:

	Subway Station	Cluster Labels	fast food	food truck	Latitude	Longitude
0	138 St - Grand Concourse	1	1	0	40.810476	-73.926138
1	149 St - Grand Concourse	0	4	1	40.816109	-73.917757
2	161 St - Yankee Stadium	2	4	4	40.813224	-73.929849
3	167 St	1	2	0	40.818375	-73.927351
4	170 St	0	3	0	40.827905	-73.925651

4. Results

We have clustered each station into one of the clusters based on number of fast food joints and food trucks near it. Lets plot this get a visual representation of it.

Colour	Cluster
Red	0
Blue	1
Cyan	2
Green	3

Orange	4

Lets have a look at each of the clusters:

Cluster 0:

	Subway Station	Cluster Labels	fast food	food truck	Latitude	Longitude
1	149 St - Grand Concourse	0	4	1	40.816109	-73.917757
4	170 St	0	3	0	40.827905	-73.925651
13	233 St	0	3	0	40.888022	-73.860341
15	3 Av - 138 St	0	3	0	40.893193	-73.857473
24	Burnside Av	0	4	0	40.853453	-73.907684
36	Intervale Av	0	4	1	40.822181	-73.896736

Cluster 1:

	Subway Station	Cluster Labels	fast food	food truck	Latitude	Longitude
0	138 St - Grand Concourse	1	1	0	40.810476	-73.926138
3	167 St	1	2	0	40.818375	-73.927351
5	174 St	1	2	1	40.833771	-73.918440
7	176 St	1	1	1	40.837288	-73.887734
8	182-183 Sts	1	2	0	40.845900	-73.910136

Cluster 2:

	Subway Station	Cluster Labels	fast food	food truck	Latitude	Longitude
2	161 St - Yankee Stadium	2	4	4	40.813224	-73.929849
6	174-175 Sts	2	1	3	40.839306	-73.913400
19	Bedford Park Blvd - Lehman College	2	1	2	40.873412	-73.890064
26	Cypress Av	2	0	2	40.805368	-73.914042
44	Mt Eden Av	2	1	3	40.844434	-73.914685

Cluster 3:

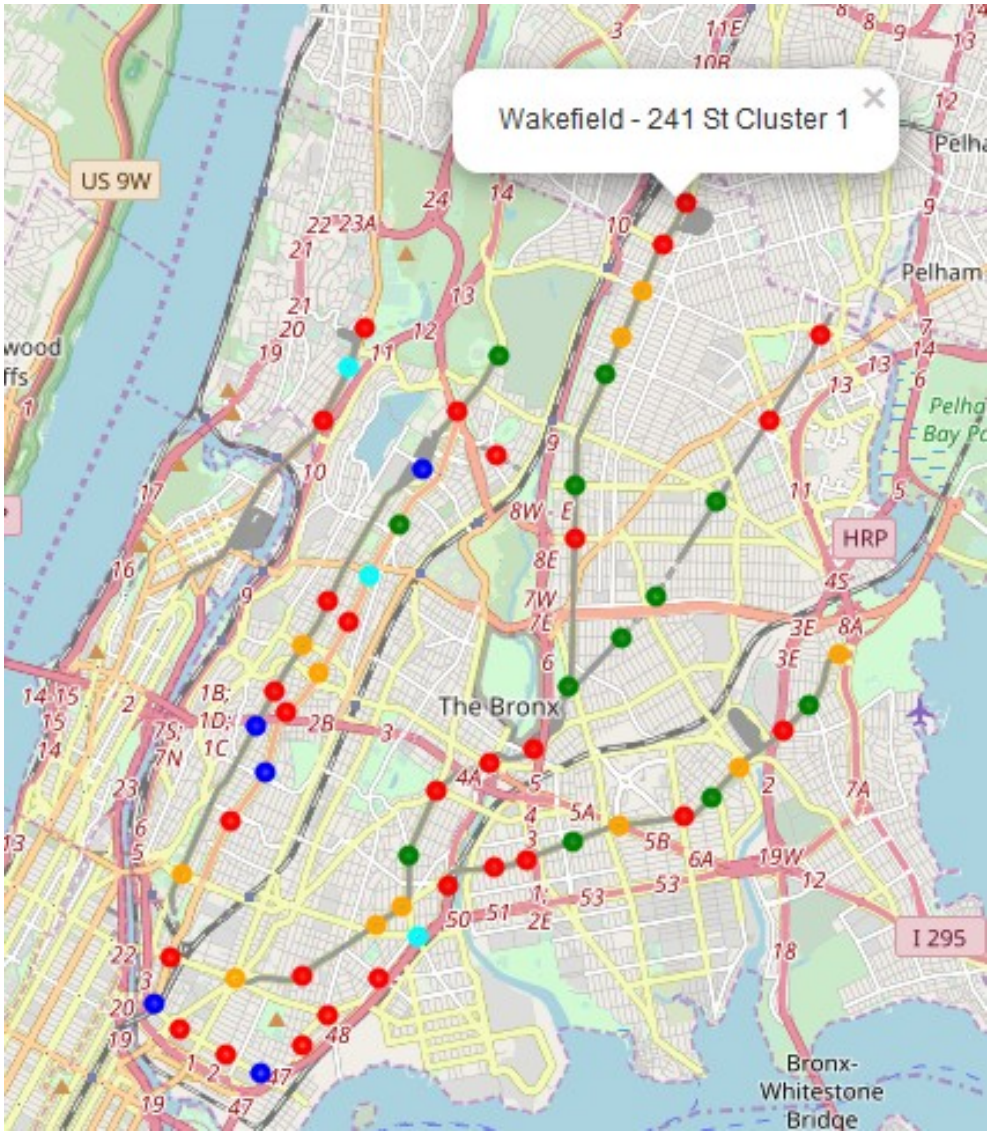
	Subway Station	Cluster Labels	fast food	food truck	Latitude	Longitude
16	3 Av - 149 St	3	7	0	40.884667	-73.900870
32	Fordham Rd	3	8	2	40.861296	-73.897749
35	Hunts Point Av	3	6	1	40.820948	-73.890549

Cluster 4:

	Subway Station	Cluster Labels	fast food	food truck	Latitude	Longitude
12	231 St	4	0	0	40.883895	-73.862633
20	Bronx Park East	4	0	0	40.848828	-73.868457
22	Buhre Av	4	0	0	40.846810	-73.832569
23	Burke Av	4	0	0	40.871356	-73.867164

From examining the cluster we arrive at following findings:

1. The best cluster is Cluster 4. They have very low or nill concentration of fast food joint and food trucks. This has almost no competitors for our client. This is the best set of stations to put up food trucks.
2. The second best cluster is Cluster 1. It has very low to nil concentration of food trucks with low concentration of fast food joints.
3. The third best cluster is Cluster 0. It has very low to nil concentration of food trucks with medium concentration of fast food joints.
4. The second least preferred cluster is Cluster 3. It has high concentration of fast food joint with low to medium concentration of food trucks.
5. The least preferred cluster is Cluster 2. It has high concentration of food trucks and low concentration of fast food joints.



Colour	Cluster	Preference Rank
Red	0	3
Blue	1	2
Cyan	2	5

Green	3	4
Orange	4	1

The above table and plot gives the final summary of our recommendations. The stations in orange cluster is best option followed by Blue and Red. Green and Cyan clusters are best avoided.

5. Discussion

The data analysis and machine learning algorithm was used to successfully accomplish the task entrusted by our client. We were able cluster the subway stations in The Bronx based on the concentration of our client's competitors.

The method the we used can be further expanded to include other boroughs and clustering of stations can be done for each borough. This will come in handy when our client NYC Foods plan to expand to other boroughs.

Furthermore, among the set of stations in the preferred cluster we can further narrow down and choose the best ones among them if we have passenger traffic data station wise. Higher traffic would mean higher the number of prospective consumers for our client.

6. Conclusion

Data and machine learning algorithms can be effectively used to solve business problems and drive value out of data.

As shown in this project, this not just helps businesses to do well it also adds value to consumers and makes available the best of the service.