Battle of the Neighborhoods

Capstone Project by Omer Ihsan
July 7th, 2020

1. Introduction

1.1 Background

A recent study has revealed that the **success of a business can be determined even before a business has started its operations**. A major determinant of business success is the selection of the location. It is common knowledge that starting a "Ferrari" showroom on a university campus will not yield the expected returns. A business needs to decide on the best location taking into account factors such as:

Customers - is the location convenient for customers?

Staff - are there sufficient numbers of local staff with the right skills willing to work at the right wage?

Support services - are there services offering specialist advice, training and support?

Cost - how much will the premises cost? Those situated in prime locations (such as city centers) are far more expensive to rent than edge-of-town premises.

Infrastructure and Economy of Scale - All businesses need proper infrastructure to function e.g. a telephone company would need an electricity meter, fast internet and customer access to function.

Logistics - In order to bring in raw materials and ship out finished goods, a business must have access to roads, rails and other means of transport.

Government Facilitation - It is desirable to select places where the government has announced tax holidays and rebates on conducting and starting businesses.

Resource - It is utterly important that the business has the services of the best minds to function and expand. A business must be started in a place where the required human capital is available abundantly.

1.2 Business Problem¶

Currently an aspiring business must undertake an extensive, time consuming and exhausting practice to decide on a location to start the operations. Many a times, the exercise is not undertaken seriously because of the effort involved. The result is a business, which has a promising product but due to a bad selection of site operations, the business fails.

1.3 Interest

The aim of this project is to address the issue of selection of the best location for starting a cafeteria business and reduce the time and effort that is put in addressing this issue. We will try to select a location keeping in mind all the above parameters. A location which satisfies most items can be selected to run a cafeteria business.

2. Data Acquisition and Wrangling

2.1 Data Sources

For the problem statement at hand, we are looking to open a restaurant that will give the highest possible return on investment. This means data required for this purpose are population density, purchasing power, competitors, and property affordability. In this project, data on population density, purchasing power (per capita income) and competitors (foursquare location data) has been used.

The following sources of data were used for this purpose

1. New York City Population by Borough, 1950 - 2040
   https://data.cityofnewyork.us/resource/xywu-7bv9.csv

2. Per Capita Income
   https://www.baruch.cuny.edu/nycdata/income-taxes/per_cap.htm

3. Property Valuation and Affordability
   https://www.bloomberg.com/graphics/property-prices/nyc/

4. New York City Cafes Data
   https://foursquare.com/explore?mode=url&near=New%20York%2C%20NY%2C%20United%20States&nearGeoId=72057594043056517&q=Food

2.2 Data Wrangling

New York population data provided by the source details the population count for individual boroughs and provides the predicted population count for two subsequent decades 2030 and 2040. The data also provides the population as a percentage of the total NYC population for each decade from 1950 to 2040.

The Per Capita Income data requires to be scraped from the website directly. The data consists of a lot of unwanted fields. The rows and columns need to be dropped and the indexes need to be reset. Finally, the column data is available in string formats which need to be converted to integer values to get the Per Capita Income values for all the boroughs in NYC.

2.3 Feature Engineering and Feature Selection

New York population count data and the population percentage data are a good indicator to judge the consumer density. The larger the population count, more the consumers for any business. Also, the data provides an extrapolation of population count for each NYC borough which makes it a good feature to select a location for any business. The population count is converted to 'millions' for easy exploration.

	Borough	Population 2020 (in millions)	2020 Population Percentage
0	NYC Total	8.550971	100.00
1	Bronx	1.446788	16.92
2	Brooklyn	2.648452	30.97
3	Manhattan	1.638281	19.16
4	Queens	2.330295	27.25
5	Staten Island	0.487155	5.70

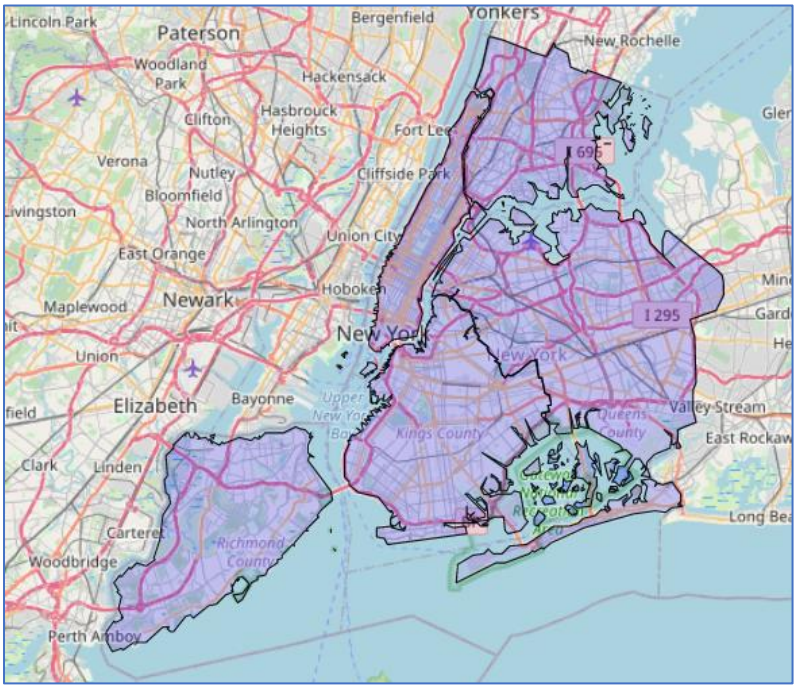
	Borough	Population 2040 (in millions)	2040 Population Percentage
0	NYC Total	9.025145	100.00
1	Bronx	1.579245	17.50
2	Brooklyn	2.840525	31.47
3	Manhattan	1.691617	18.74
4	Queens	2.412649	26.73
5	Staten Island	0.501109	5.55

The data provides values for Per Capita Income for different Geographical Areas \ Boroughs of NYC. The Per Capita Income of the Area is proportional to the spending power of the residents. A greater value of PCI indicates a greater spending power. The string dollar values will be converted to Integer values for Exploratory Data Analysis.

	Geographic Area	2013	2014	2015	2016	2017
0	Bronx	$30,647	$31,556	$32,778	$33,310	$35,564
1	Brooklyn (Kings)	$39,586	$41,399	$43,915	$45,629	$48,758
2	Manhattan	$145,231	$152,690	$155,779	$164,056	$175,960
3	Queens	$39,789	$40,997	$43,216	$44,031	$46,829
4	Staten Island (Richmond)	$46,219	$48,123	$50,894	$51,836	$54,908

According to Bloomberg.com the rental and property values vary in different boroughs of NYC. Properties in Manhattan are far more expensive and harder to get as compared to properties in Brooklyn and Queens. The commercial properties available for lease and rent in Manhattan Area according to Bloomberg.com is a “drawn-out affair”.

Python`s Folium library has an extensive database of maps and map layout functions. An overlay of the New York City map has been fetched to better understand the positioning of the boroughs. The positioning is important to understand because any business location is dependent on the surroundings. A location that has access to transport, public facilities and other commercial centers has a higher business value.



3. Exploratory Data Analysis

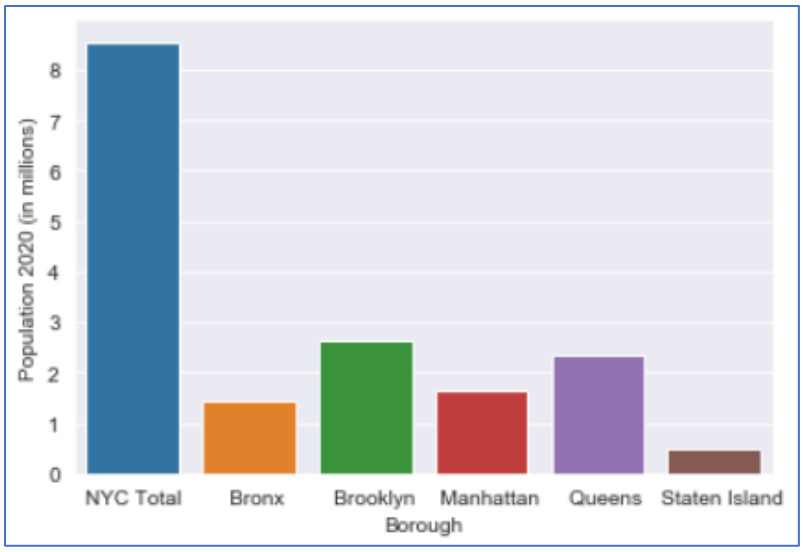
3.1 Target Variable

Since we are dealing with a classification problem, our target variable is a binary decision classified as a 'YES' that is a location for setting up a restaurant business or a 'NO' that is a location for setting up a restaurant business.

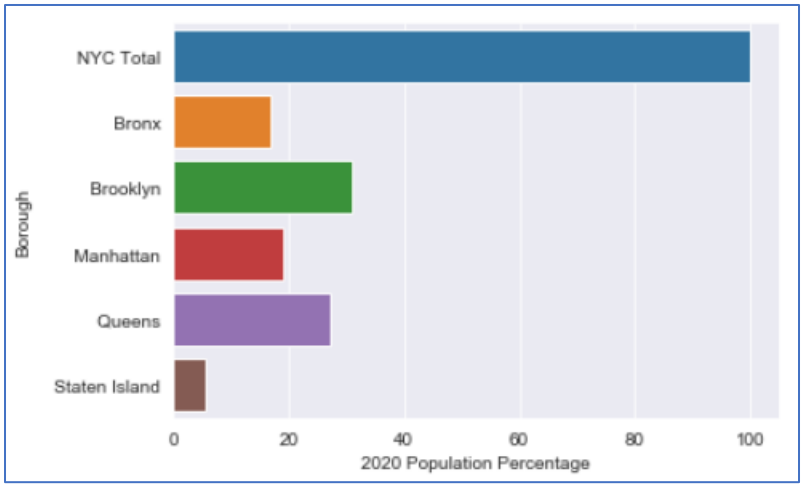
The classification will be based on inputs such as borough, per capita income, population (potential consumers), nearby restaurants\cafes (competitors) and access to facilities.

3.2 Population (Potential Consumers)

Population count which is directly proportional to an areas potential consumer of a service is an important consideration to start any business. New York City`s current population is approximately 8.55 million in 2020 as per the data source. Brooklyn has the highest population (about 2.64 million) followed by Queens and Manhattan. Staten Island has the least population (around 0.48 million). Population distribution in 2020 is represented in the bar plot.



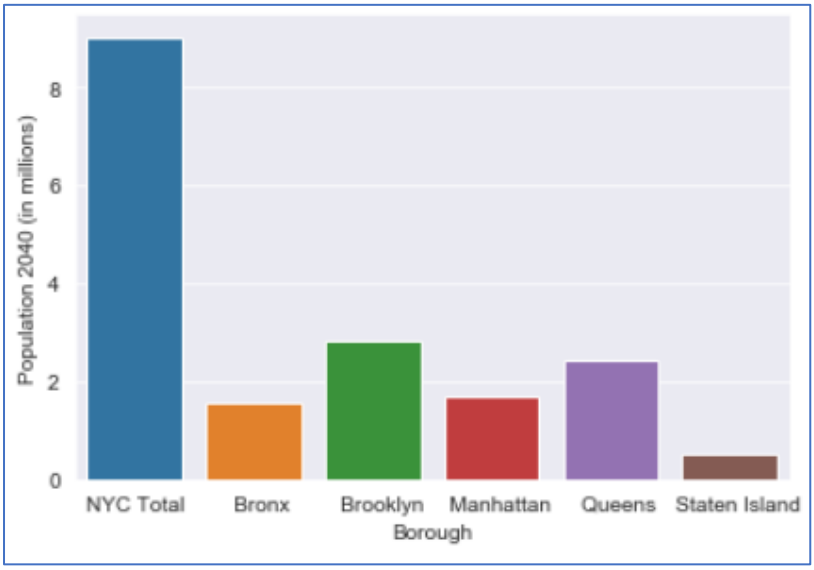
A similar representation of the population can be seen as a percentage distribution. Brooklyn in 2020 represents 31 percent of New York City`s total population.

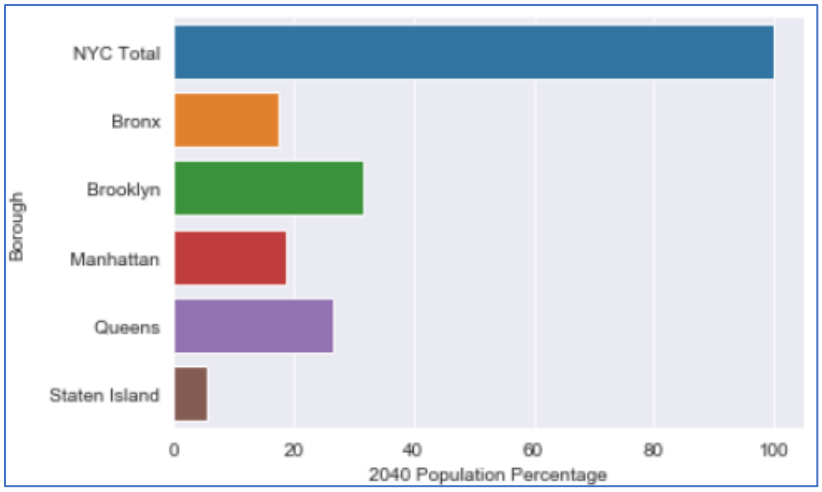


Since any business is interested to be accessible by the maximum number of people, three areas in New York City namely, Brooklyn, Manhattan and Queens may be shortlisted. Since population of Bronx and Staten Island Boroughs are low, we can leave them as these areas have fewer potential customers.

3.3 Future Population Center (Potential Consumers)

Extrapolation of data can give an understanding in to the future look out of businesses. As population was earlier discussed as an important factor in deciding a business location, we have extrapolated data for New York City till 2040. From the data the total population will grow to about 9.02 million in 2040 but the overall distribution does not change that much. Brooklyn remains to be the largest population center with almost 32 percent population.

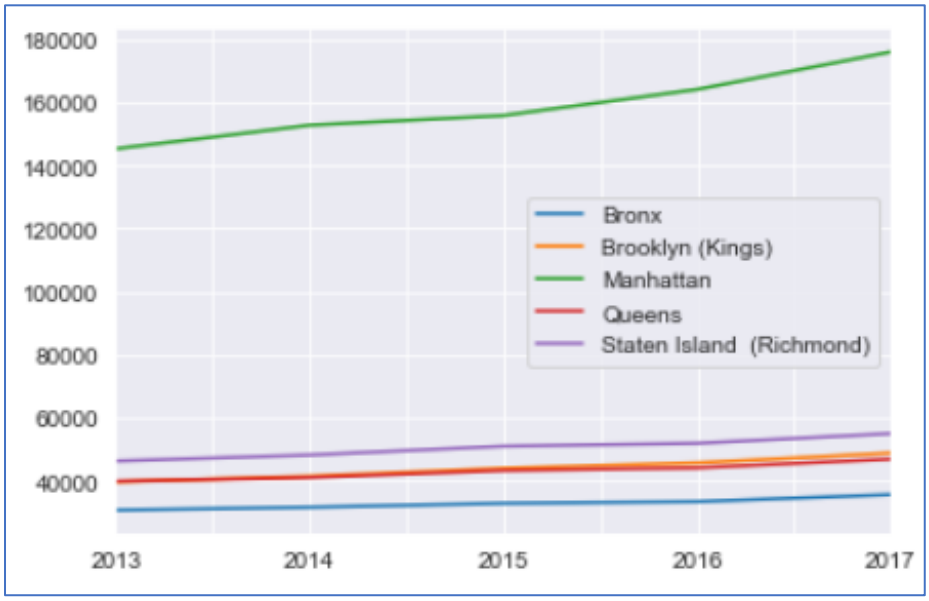




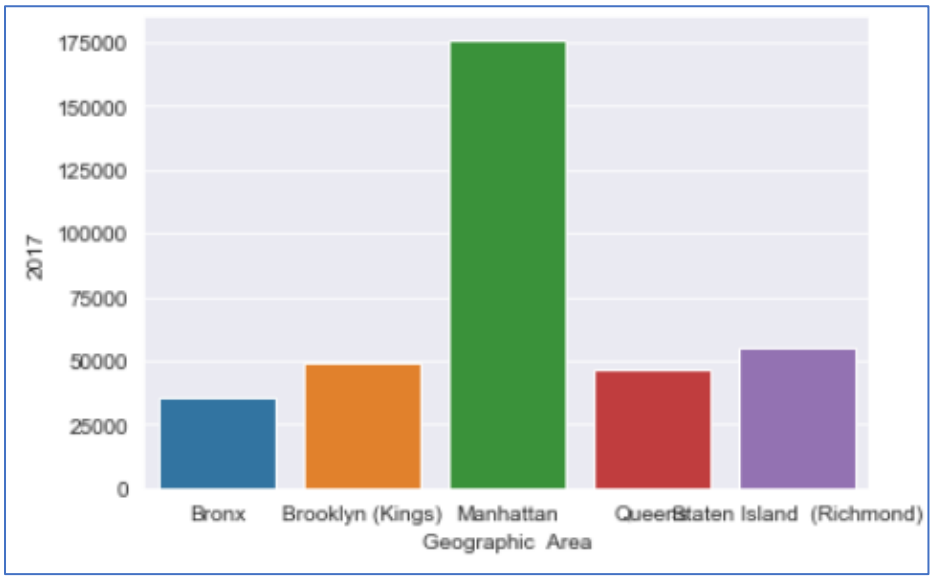
Looking at this data, we will focus on the boroughs having the highest population namely, Brooklyn, Manhattan, and Queens.

3.4 Per Capita Income

Manhattan has out-performed all other borough areas of NYC by a long way consistently for over the last decade. Data collected from the source shows the same. The presence of big businesses, financial headquarters of major businesses around the globe and Wall Street itself, supports the huge difference of Per Capita Income of Manhattan and the other boroughs.



A bar plot of the Per Capita Income for the year 2017 makes this clearer.

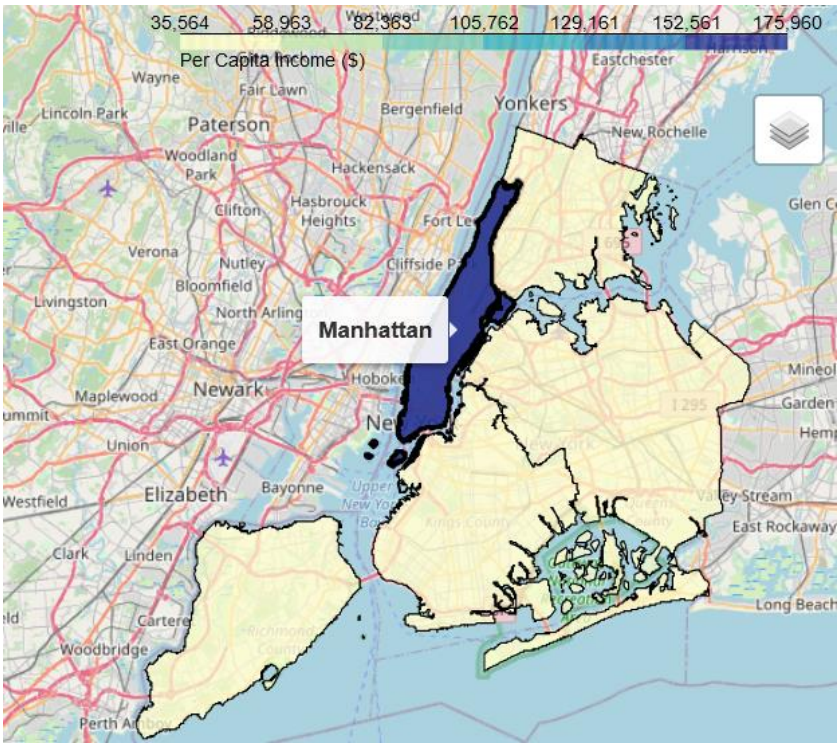


This practically shows that people living in Manhattan earn more and consequently spend more. The parameter is important for any business to consider because the more the people can spend, the more the businesses will thrive.

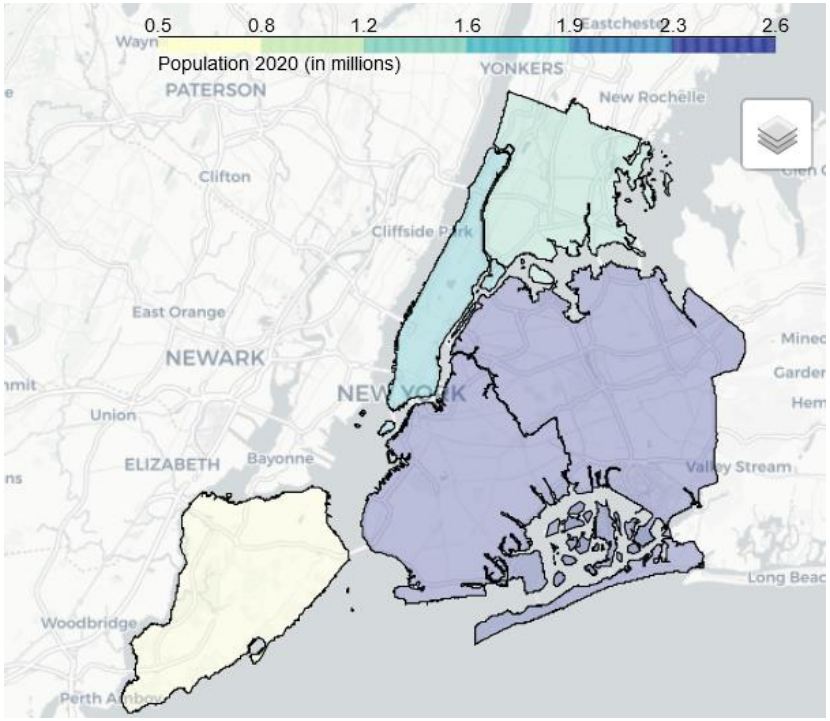
3.5 Choropleth Maps

A **choropleth map** is a thematic map in which areas are shaded according to the statistical variable being displayed on the map. This visualization is useful to represent the variability of a measurement across a region using a sequential color scheme (the higher the measurement the stronger the color).

Here we have used choropleth maps to quickly visualize three aspects of each borough of New York City. First is the Per Capita Income of the boroughs. The Per Capita Income of Manhattan is way more than that of the other boroughs. As we have seen that the per capita Income directly translates into the people`s spending power, Manhattan is the right place to invest for a posh and high-end restaurant. People can afford to spend lavishly for quality here. Definitely, the returns will be high and swift.



The population count for the boroughs Brooklyn and Queens is on the higher end. Manhattan has a lower population count as compared to all other boroughs except Staten Island. But at the same time, it can be seen that the other boroughs have a much larger area as compared to Manhattan. So a better way to visualize the impact of borough population is the population density.



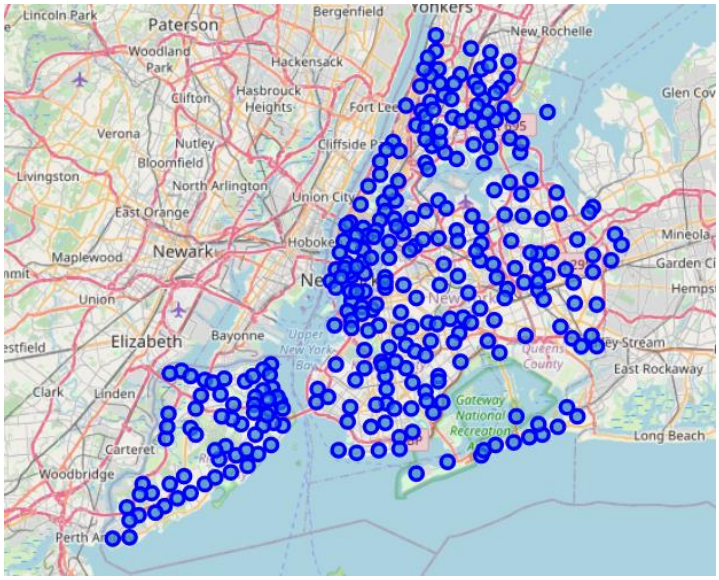
Looking at the population density choropleth map, it is clear that Manhattan has a high population density as compared to all other boroughs. What this practically means is that a greater number of people are cramped into a smaller space. So, business location is accessible to more number of people in a smaller radius distance.

Lastly, we have focused on the environment and aura the borough provides to a business. Manhattan is somewhat a narrow corridor of land with water all around. It has green spaces including the landmark New York Central Park. Moreover, the infrastructure presents more viable business options. Presence of major business centers and the Wall Street itself cannot be ignored. A choropleth will quickly show the water areas and green areas for Manhattan.

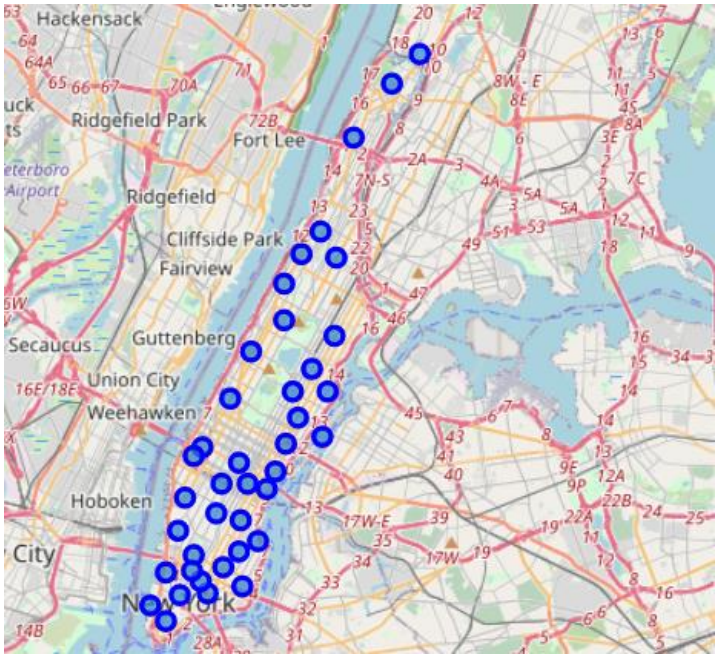
3.6 New York Neighborhoods

Though it is truly up to a business to decide, based on the exploratory data analysis we have selected Manhattan to be the area of further interest. Manhattan poses all the benefits for a business i.e. high population density, high per capita income and presence of major business and commercial centers.

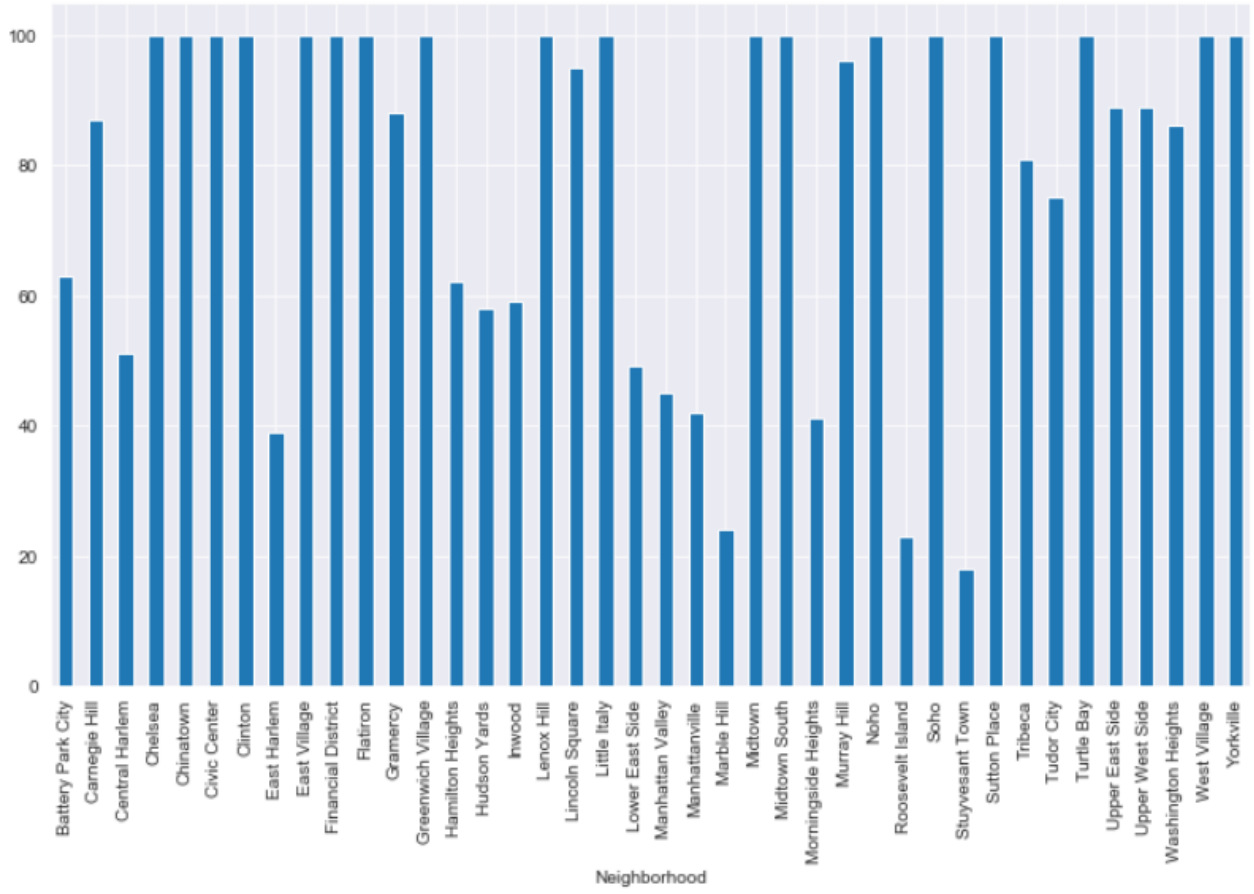
The State of New York has 5 boroughs and 306 Neighborhoods. We have used folium to mark the neighborhoods on New York boroughs.



Further we reduced the data set and will focus on Manhattan neighborhoods only.



Manhattan has 40 neighborhoods. These neighborhoods have different levels of business activities and the categories of commercial centers are highly varied. In fact, the number of unique categories of venues\commercial centers was found to be 327. A visual presentation of the number of venues returned against each Manhattan Neighborhood is as follows.



As we are interested in opening a restaurant, we should check which neighborhoods have a concentration of restaurants and cafes so that such localities can be avoided and omitted from further analysis.

4. Data Clustering

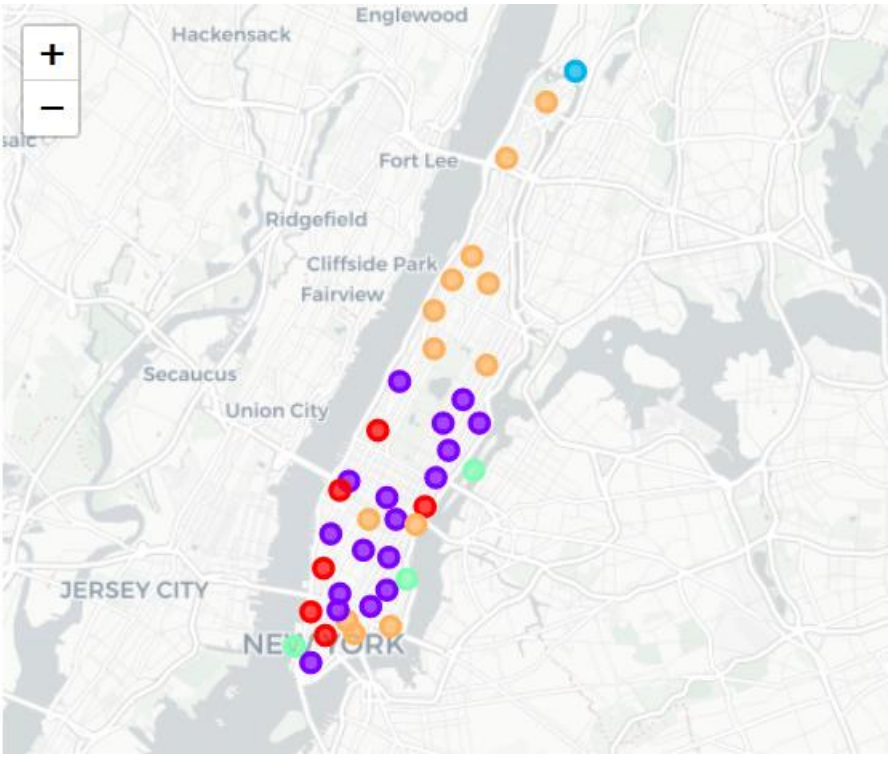
4.1 K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Unsupervised learning means that there is no outcome to be predicted, and the algorithm just tries to find patterns in the data. In k means clustering, we have the specify the number of

clusters we want the data to be grouped into. The algorithm randomly assigns each observation to a cluster and finds the centroid of each cluster. Then, the algorithm iterates through two steps: Reassign data points to the cluster whose centroid is closest. Calculate new centroid of each cluster. These two steps are repeated till the within cluster variation cannot be reduced any further. The within cluster variation is calculated as the sum of the Euclidean distance between the data points and their respective cluster centroids.

4.2 Manhattan Neighborhood Clusters

For the sake of simplicity and considering the randomness of the Venue Categories returned from the Foursquare API, we will set the 'Cluster Number' to 5 i.e. all the neighborhoods in the Manhattan data will be clustered in to 5 groups. This grouping will purely be based on the similarity of the data points i.e. the top ten venue categories of each neighborhood. The resulting clusters are mapped here. Each cluster is colored differently for easy identification.



4.3 Number of Cafes\Restaurants in Clusters

The number of cafes\restaurants presented as count in the top ten venues of each Neighborhood Cluster determines the density of existing competition. More the number of currently running restaurant businesses, harder it will be for a new entrant to capture market share.

```
The number of Cafes\Restuarants in Cluster 0 =  22
The number of Cafes\Restuarants in Cluster 1 =  55
The number of Cafes\Restuarants in Cluster 2 =  0
The number of Cafes\Restuarants in Cluster 3 =  2
The number of Cafes\Restuarants in Cluster 4 =  52
```

4.4 Proportion of Cafes\Restaurants in Top Venues of Clusters

The proportion of cafes and restaurants in the top venues of each neighborhood cluster gives us a good idea about the balance of activities in a neighborhood. For example, if the neighborhood is a majorly non-residential area, the likelihood of finding parks and other family outdoor activity venues will be slim. Similarly, the popularity of restaurants in an area signify that people eat out in these places. These could be the corporate neighborhoods around Wall Street where people are mostly eating out.

```
The proportion of Cafes\Restaurants in the Top Venues of Cluster 0 is 36.67 %
The proportion of Cafes\Restaurants in the Top Venues of Cluster 1 is 32.35 %
The proportion of Cafes\Restaurants in the Top Venues of Cluster 2 is 0.0 %
The proportion of Cafes\Restaurants in the Top Venues of Cluster 3 is 6.67 %
The proportion of Cafes\Restaurants in the Top Venues of Cluster 4 is 40.0 %
```

5. Results and Discussion

5.1 Manhattan

Out of the 5 New York boroughs, we selected Manhattan. Our focus is to invest to open a café in New York that will yield bigger and quicker profits. Manhattan`s population density and the per capita income of the inhabitants made it an ideal choice to start a café.

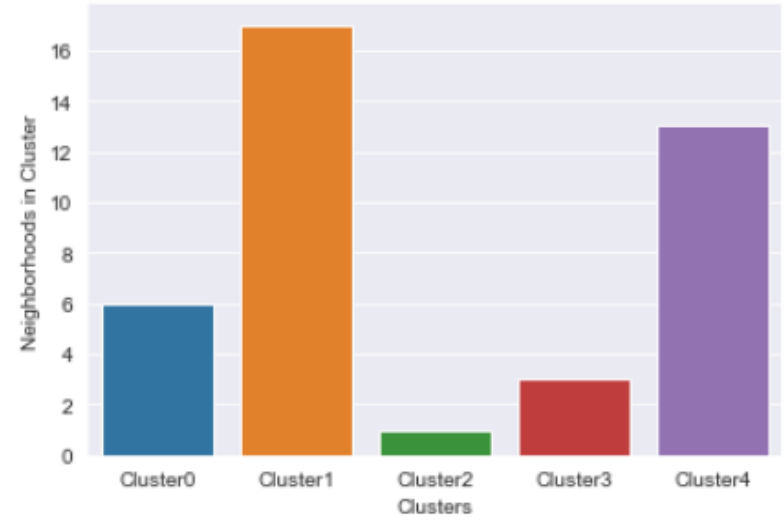
5.2 Manhattan Neighborhoods

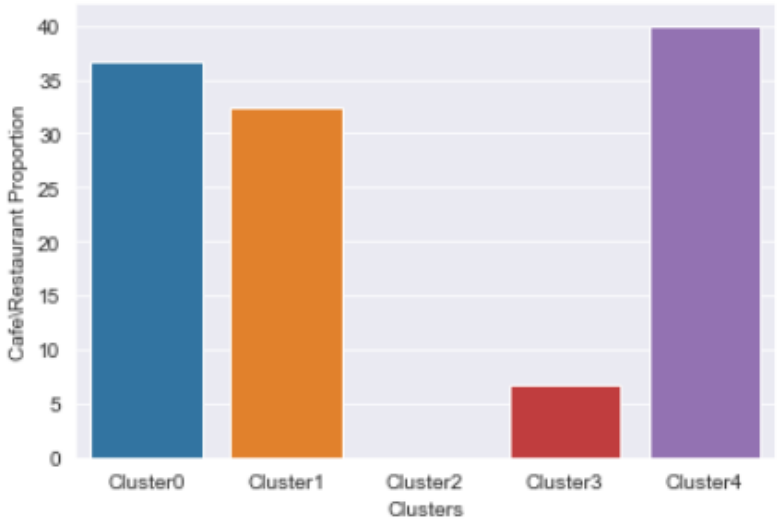
Manhattan neighborhoods were clustered in to 5 groups based on the similarities of the venue categories in each neighborhood. Clustering the data made the analysis for selecting a location easier. Cluster data provided insights into business competition.

5.3 Number and Proportion of Cafes in Neighborhood Clusters

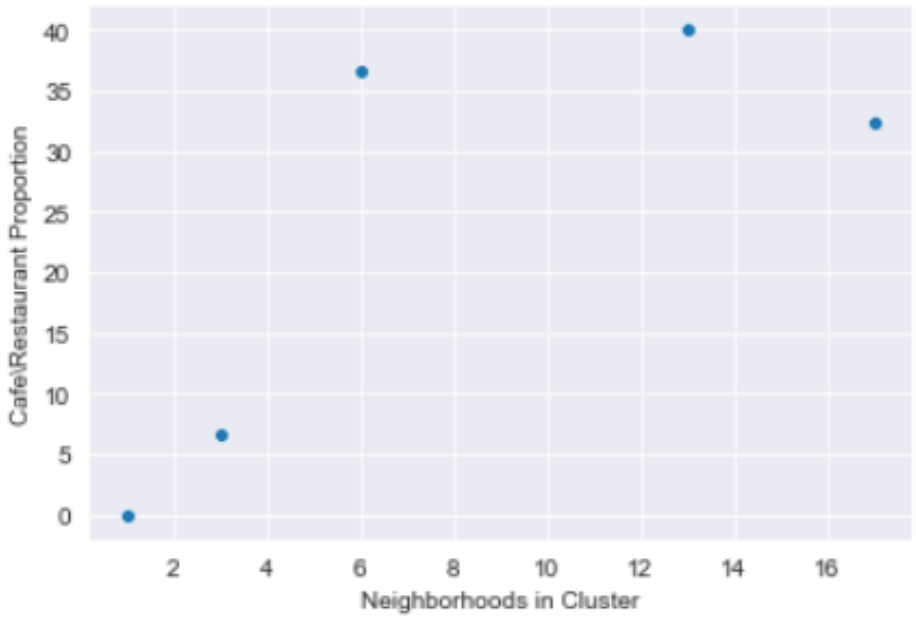
Each cluster has a different number of neighborhoods clubbed together. This clubbing is based on the similarities of the top ten venues of each neighborhood. As a result, each cluster has a different number of existing cafes\restaurants and the proportion is different.

The count and proportions are shown as bar plots below.





From a scatter plot between Number of Existing Cafes\Restaurants in the top venues list and the Proportion of the Number out of the total number of venues in the cluster shows a direct proportionality between them. Anyhow we can see that after a certain Number of Cafes the Proportion does not increase. This shows the saturation point of a neighborhood.



6. Conclusion

Parameters such as consumer density, spending power and competition are extremely important in deciding for a location to start any business. Anyhow they are not the only parameters. This study provided these parameters to potential investors who wanted to start a café\restaurant in New York. Though it strictly depends on the investor and the line of business, based on the analysis for consumer density and spending power of the people of New York, we were able to select Manhattan as the potential borough to start the café. Clustering of the data provided insight into business competitors. Within the neighborhood clusters, some clusters had a high percentage of popular cafes. It will not be wise to open a café in a saturated space such as the ones with a high proportion of cafes as the top venue categories.

	Clusters	Cafe\Restaurant Proportion	Neighborhoods in Cluster
0	Cluster0	36.67	6
1	Cluster1	32.35	17
2	Cluster2	0.00	1
3	Cluster3	6.67	3
4	Cluster4	40.00	13

This can be answered in two different approaches. First is to select the neighborhood clusters which have an extremely low proportion of cafes i.e. Cluster 2 and Cluster 3. Anyhow looking at the data we realize that the Number of neighborhoods in the Clusters are small i.e. 1 and 3 respectively. Thus, the business potential insight is not favorable.

The other approach is to select a neighborhood cluster which has the highest proportion of Cafes as the top venues' categories of the neighborhoods. The proportion shows that restaurant business is booming in the neighborhood. But we cannot be sure if there is space for more competitors. It might as well reached a saturation point.

Hence, we have finalized to suggest to our investor to go for an investment in Cluster 1 mainly because of two reasons. The number of neighborhoods in the cluster is the highest which suggests a larger area and bigger number of consumers. Secondly, the proportion being lower than the highest proportion of 40% suggests that there is room for competition. A new restaurant will surely find its space in one of the neighborhoods of Cluster 1.