BATTLE OF NEIGHBORHOODS

1.Introduction/Business Problem

The Background:

The City of New York also popularly known as NYC is one of the most populated cities in The United States. It is also known as the financial, media and cultural capital of the world. The city has beautiful sky-scraping buildings as well as known for many tourist places. It is a land of opportunities at the same time extremely competitive and fast paced. NYC is also known for its cultural diversity as it is the leading gateway for immigrant population. Manhattan's population density is the highest of any county in the United States. The city is basically made up of five boroughs namely - Brooklyn, Queens, The Bronx, Manhattan and Staten Island.

Manhattan is the place where one can find the city's top attractions like Empire State building, Central Park, Central Park, The Chrysler Building and many more. It is the most densely populated boroughs of all the five. It is home to the world's largest two stock exchanges - The NASDAQ and New York Stock Exchange.

As we can understand from the backdrop of the city, it is not just a tourist spot but also highly competitive. Since it is a financial hub and has top attractions for people to visit, it is but natural that the not just the standard of living but also the cost of a starting a business is extremely high. As I analyze the data and draw insights, the conclusion may lead to reduction of risk and give more returns to my client when they start their restaurant in Manhattan.

Description of the Problem:

Restaurant business is something where an entrepreneur makes lot of sacrifices, experiences certain amount of losses in the beginning but as the business picks up because of words of mouth of its customers and better marketing strategy the owner finally ends up making profits. But before all this, the choice of place where he/she can start the business is also very important. Location is something where one should be able to draw crowds, be easily accessible and have potential for growth.

As already described the population of Manhattan is diverse and includes people from all ethnicities like Asians, Jewish, Spanish, Brazilian, South Asians where Indian Population is the highest followed by Bangladeshis and Pakistanis and many other countries as well. Therefore, it is important to study the cuisine, the demographics etc. of the place.

My client a hotelier who has run successful Indian restaurants in places like Mumbai, Delhi, Bangalore now wants to explore the market of Manhattan. He wants to go international by opening his first of restaurant business in this densely populated and most competitive market of

Manhattan, NYC. His belief lies in his menu, the secret ingredients of the food, brand, marketing and most importantly his desire to see the customer happy and satisfied so that they return.

Yes, all the above aspects are important, but I feel that location is the boss for any business especially restaurants. Since I have been assigned the duty of analyzing the location now it is upon me to offer a good suggestion to my client.

Objective:

The main objective of this data science project is to analyze as to whether Manhattan as a location for an Indian restaurant is feasible or not as my client wants to open his restaurant 'The Foodies' right here. I hope to provide proper rationale for him with my data analysis.

2. Data Requirements:

The analytics method determines data content, format and representation that is the sources of initial data collection guided by domain knowledge. In this segment therefore there is a description about the source of data and the data that is used for the project.

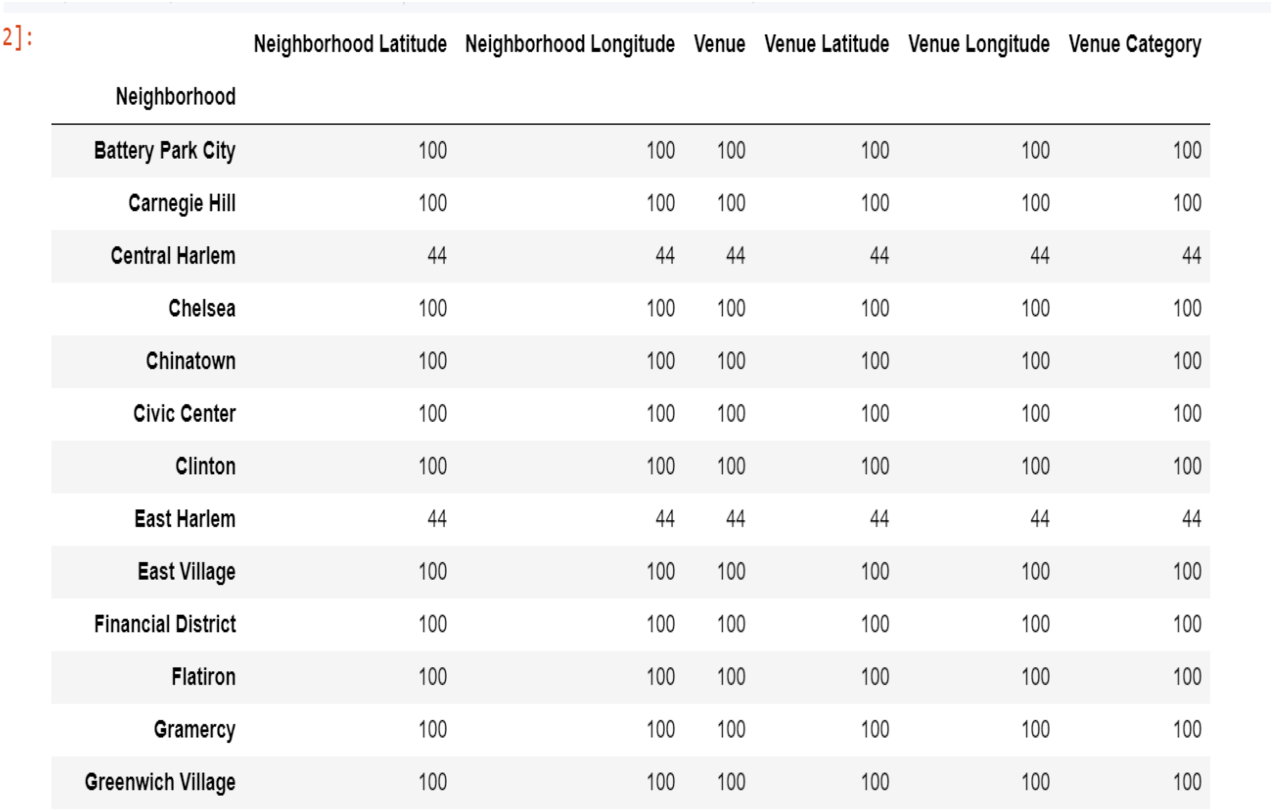
Data for this Project:

Data Science is the field of exploring, manipulating and analyzing data and then using this data to answer any questions or make recommendations. Data is the key ingredient for the preparation of a recipe called data analysis. In this project, the data of New York City will used. Further the data will be used to analyze the Borough Manhattan and its various neighborhoods. The segmentation of the neighborhoods of the borough Manhattan will be done later. From the neighborhoods, there will be segmentation of venues.

Source of the Data:

The data is sourced from the following link:https://geo.nyu.edu/catalog/nyu_2451_34572 The data frame has 5 boroughs and 306 neighborhoods. Later this data will be converted to a csv file for further analysis. Also, Foursqaure API is used here for gathering the data relating to nearby venues, restaurants in the borough by leveraging on the geographical coordinates of Manhattan. All the Necessary images, plots have been uploaded in the final report along with the analysis of the same.

The image below shows the data of Manhattan grouped by its Neighborhood along with venues and with the respective latitudes and longitudes.



2]:

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Battery Park City	100	100	100	100	100	100
Carnegie Hill	100	100	100	100	100	100
Central Harlem	44	44	44	44	44	44
Chelsea	100	100	100	100	100	100
Chinatown	100	100	100	100	100	100
Civic Center	100	100	100	100	100	100
Clinton	100	100	100	100	100	100
East Harlem	44	44	44	44	44	44
East Village	100	100	100	100	100	100
Financial District	100	100	100	100	100	100
Flatiron	100	100	100	100	100	100
Gramercy	100	100	100	100	100	100
Greenwich Village	100	100	100	100	100	100

3. Methodology

Business Understanding:

Business understanding is the core question that is places at the beginning of the methodology. It gives clarity around the problem to be solved and allows us to determine which data will be used to answer the core question. A clearly defined question directs the analytical approach that will be required to solve the problem. As stated earlier the main question here is to analyze whether Manhattan as an area is feasible to open a restaurant or not.

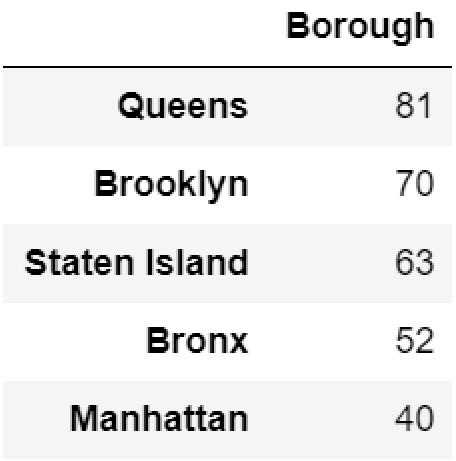
Analytic Approach:

Selecting the right approach is depends on the question being asked. Once the problem is defined, then approach can be selected. This means identifying what types of patterns will be needed to address the question effectively. For the purpose of this project Exploratory Data Analysis was done.

Exploratory Data Analysis:

New York City has a total of 5 Boroughs and 306 neighborhoods. In this project first the boroughs of the city are identified. Then the segmentation of Manhattan is done.

The image below shows the boroughs of NYC.

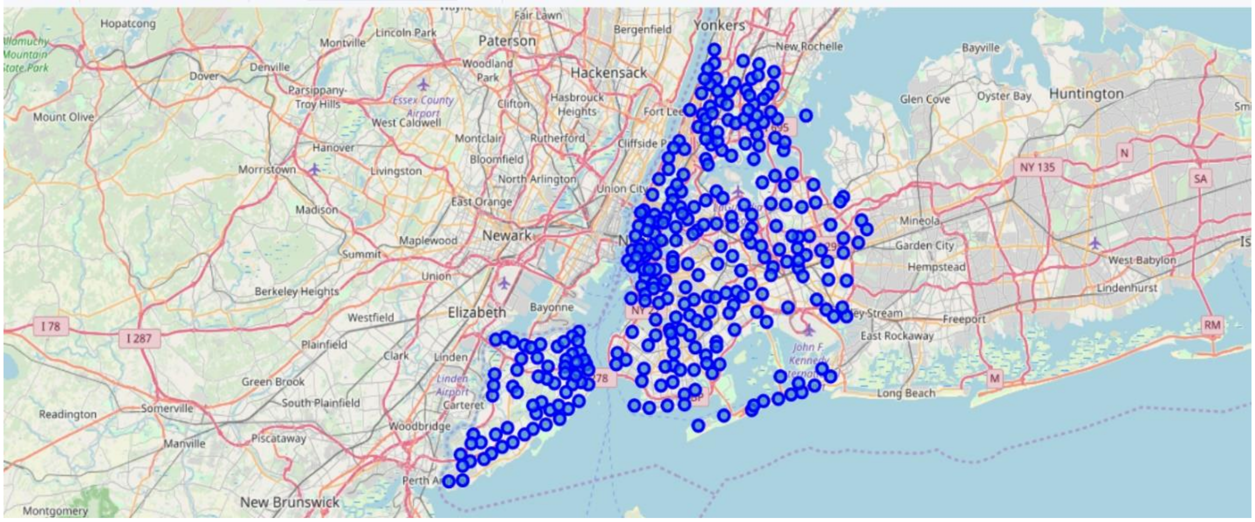


	Borough
Queens	81
Brooklyn	70
Staten Island	63
Bronx	52
Manhattan	40

In this stage, techniques such as predictive or descriptive statistics and visualization can be applied to the data set to assess the content, quality and offer initial insights about the data.

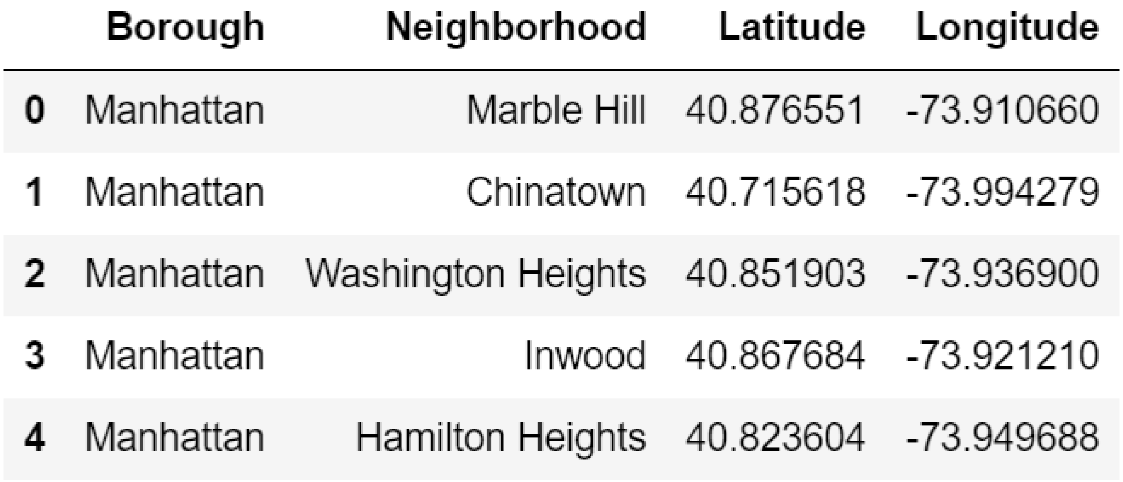
For the purpose of this project the boroughs of NYC were drawn from the data and then visualization of NYC with these boroughs was done. The latitude and longitude of NYC was also gathered.

The image below shows the map of NYC with all the boroughs:



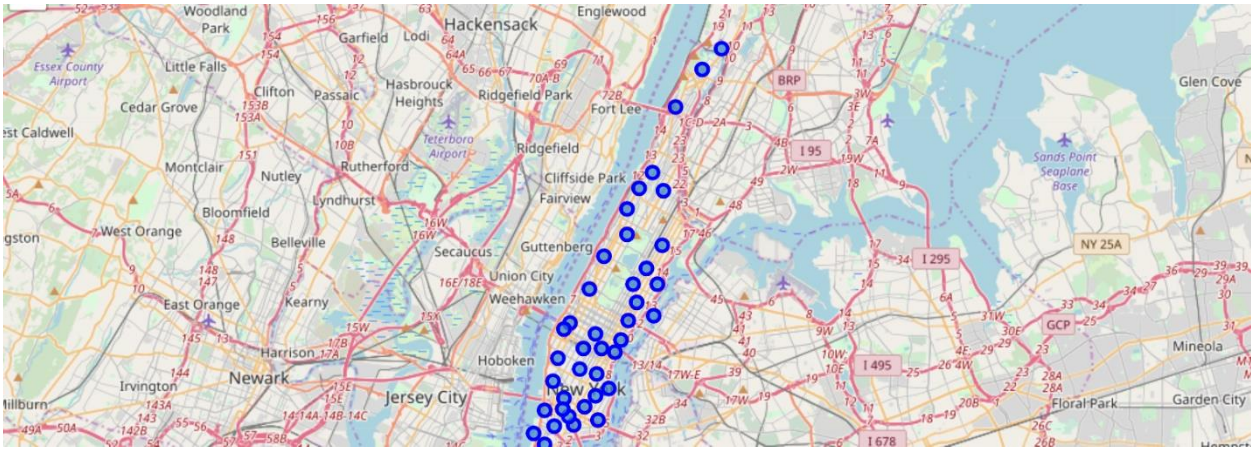
Once the data relating to NYC were gathered, then the borough Manhattan around which this project revolves was chosen. After getting the geographical coordinates of Manhattan, its segmentation was done based on its neighborhoods.

The following image gives a glimpse of its neighborhoods:



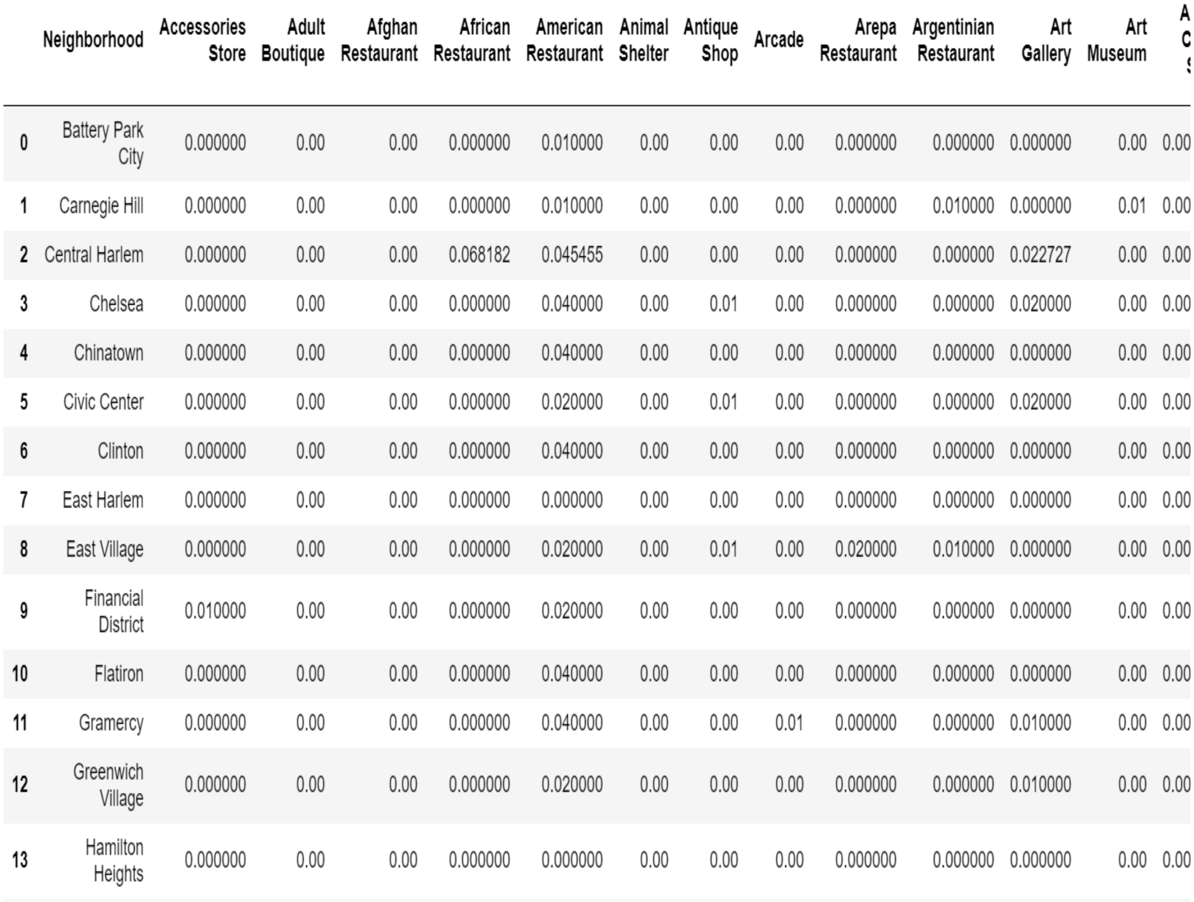
	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

After the segregation, the map of Manhattan was created using the folium module in Python and the next image shows the map of Manhattan based on its neighborhoods:



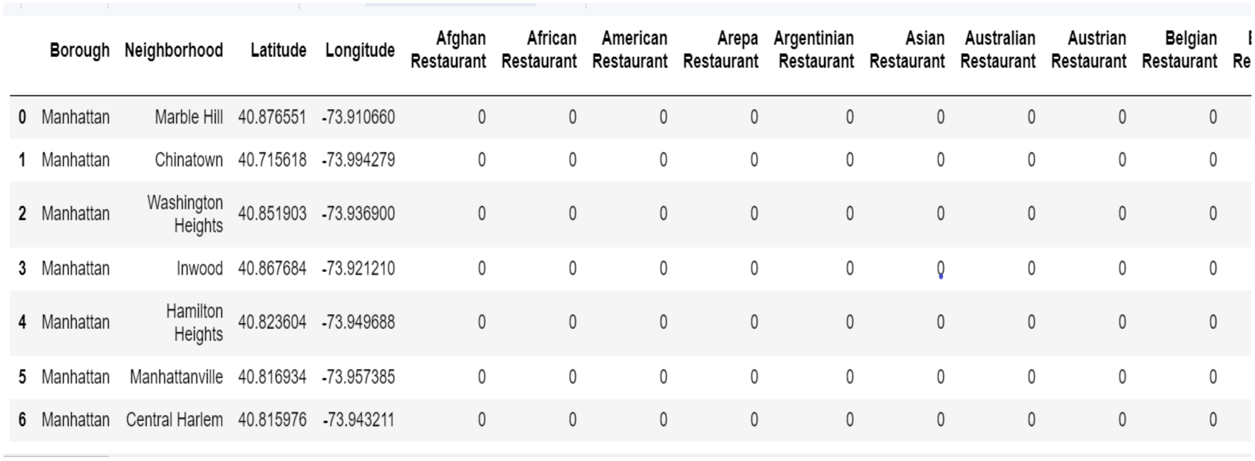
Once this was visualized, then Foursquare was used to generate the various venues and their categories in the neighborhoods of Manhattan. All in all, 3,302 venues with 331 venue categories were generated for the 40 of its neighborhoods.

The next image gives a glimpse of the venues with its categories and neighborhoods:



	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	A C S
0	Battery Park City	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00
1	Carnegie Hill	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.00	0.000000	0.010000	0.000000	0.01	0.00
2	Central Harlem	0.000000	0.00	0.00	0.068182	0.045455	0.00	0.00	0.00	0.000000	0.000000	0.022727	0.00	0.00
3	Chelsea	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.01	0.00	0.000000	0.000000	0.020000	0.00	0.00
4	Chinatown	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00
5	Civic Center	0.000000	0.00	0.00	0.000000	0.020000	0.00	0.01	0.00	0.000000	0.000000	0.020000	0.00	0.00
6	Clinton	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00
7	East Harlem	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00
8	East Village	0.000000	0.00	0.00	0.000000	0.020000	0.00	0.01	0.00	0.020000	0.010000	0.000000	0.00	0.00
9	Financial District	0.010000	0.00	0.00	0.000000	0.020000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00
10	Flatiron	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00
11	Gramercy	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.01	0.000000	0.000000	0.010000	0.00	0.00
12	Greenwich Village	0.000000	0.00	0.00	0.000000	0.020000	0.00	0.00	0.00	0.000000	0.000000	0.010000	0.00	0.00
13	Hamilton Heights	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.00

The next image gives a glimpse of the restaurants in Manhattan based on the neighborhoods of the borough:



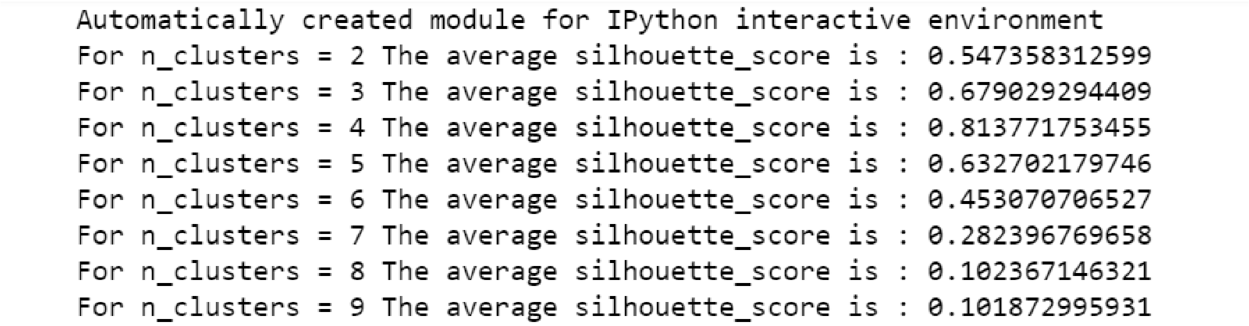
	Borough	Neighborhood	Latitude	Longitude	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	Belgian Restaurant
0	Manhattan	Marble Hill	40.876551	-73.910660	0	0	0	0	0	0	0	0	0
1	Manhattan	Chinatown	40.715618	-73.994279	0	0	0	0	0	0	0	0	0
2	Manhattan	Washington Heights	40.851903	-73.936900	0	0	0	0	0	0	0	0	0
3	Manhattan	Inwood	40.867684	-73.921210	0	0	0	0	0	0	0	0	0
4	Manhattan	Hamilton Heights	40.823604	-73.949688	0	0	0	0	0	0	0	0	0
5	Manhattan	Manhattanville	40.816934	-73.957385	0	0	0	0	0	0	0	0	0
6	Manhattan	Central Harlem	40.815976	-73.943211	0	0	0	0	0	0	0	0	0

Machine learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using instructions but to rely on patterns and inference instead. Machine learning has various categories which are Supervised Learning, Semi-Supervised learning and Unsupervised Learning.

For the purpose of this project I have used Unsupervised Learning as means for analyzing the data collected. Unsupervised learning builds a mathematical model from a set of data which contains only inputs and no desired output labels. It is basically used to find structure in data like clustering or grouping of data.

Silhouette (Clustering) refers to a method of interpretation and validation of consistency within clusters of data. The Silhouette value is a measure of how similar an object is like its own cluster compared to other clusters. Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster. Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

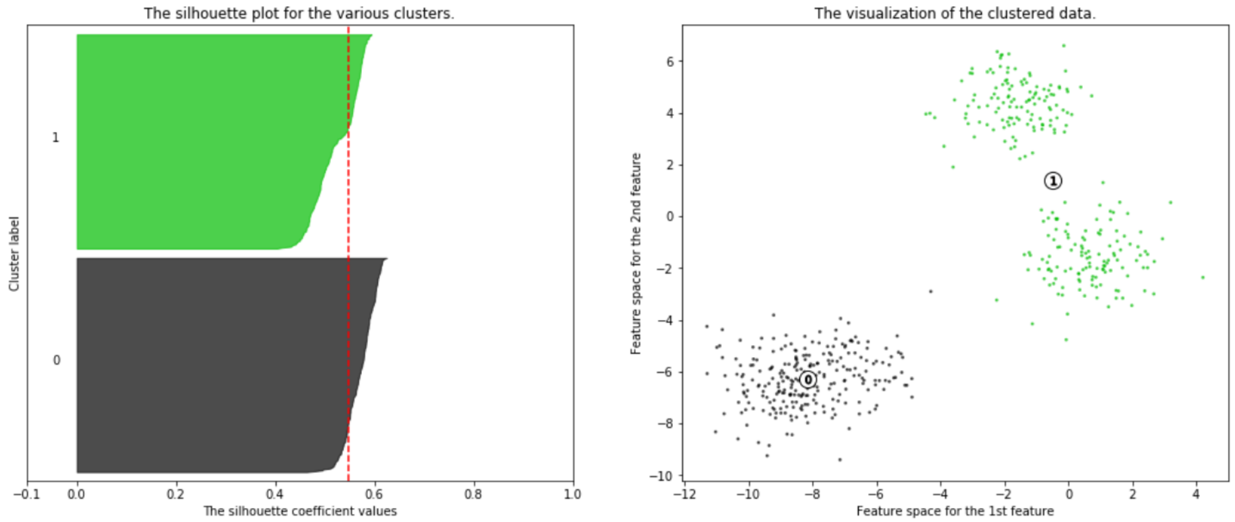
I used the silhouette score to analyze the clusters. In this analysis 9 clusters have been used and the scores of each of the clusters have been provided.



```
Automatically created module for IPython interactive environment
For n_clusters = 2 The average silhouette_score is : 0.547358312599
For n_clusters = 3 The average silhouette_score is : 0.679029294409
For n_clusters = 4 The average silhouette_score is : 0.813771753455
For n_clusters = 5 The average silhouette_score is : 0.632702179746
For n_clusters = 6 The average silhouette_score is : 0.453070706527
For n_clusters = 7 The average silhouette_score is : 0.282396769658
For n_clusters = 8 The average silhouette_score is : 0.102367146321
For n_clusters = 9 The average silhouette_score is : 0.101872995931
```

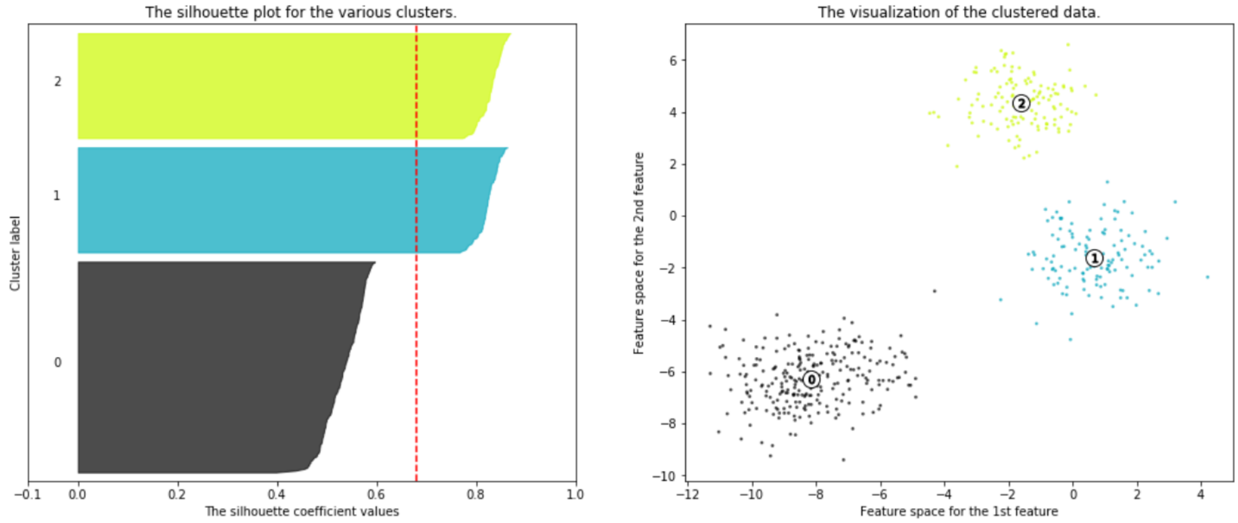
This Silhouette shows which object lies within their cluster, which are somewhere between the clusters. It can be inferred that for n_clusters 2,6,7,8,9 the silhouette scores are below average, and they show wide fluctuations in the plot. The silhouette scores for n_clusters 3 and 5 is slightly above average but not quite satisfactory. Out of all the best is n_cluster 4 where the score is the highest. The plot is of similar thickness and are of similar sizes.

Silhouette analysis for KMeans clustering on sample data with n_clusters = 2


The silhouette plot for the various clusters.

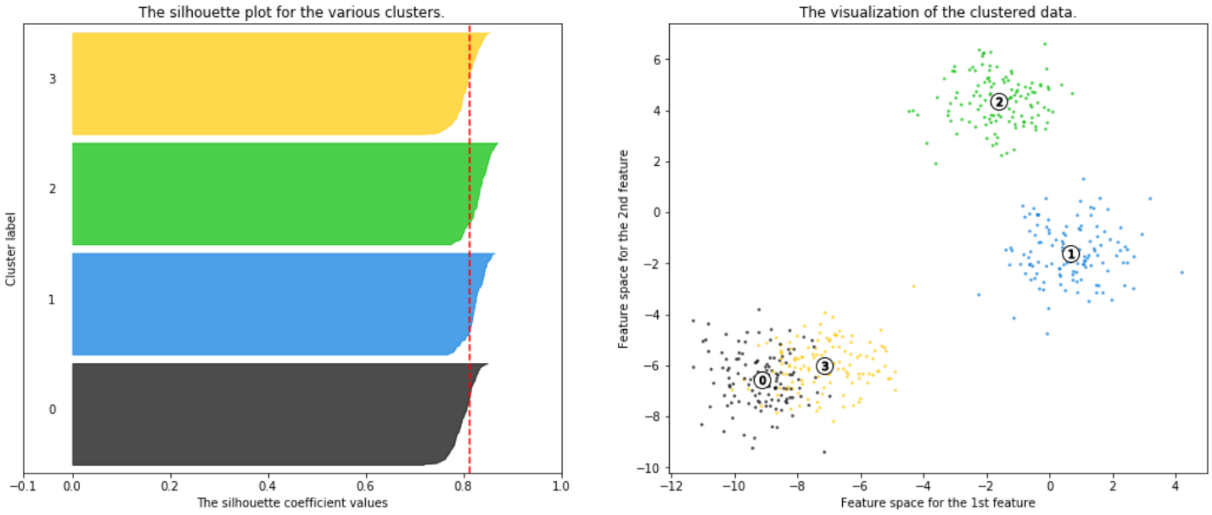
The visualization of the clustered data.

Silhouette analysis for KMeans clustering on sample data with n_clusters = 3


The silhouette plot for the various clusters.

The visualization of the clustered data.

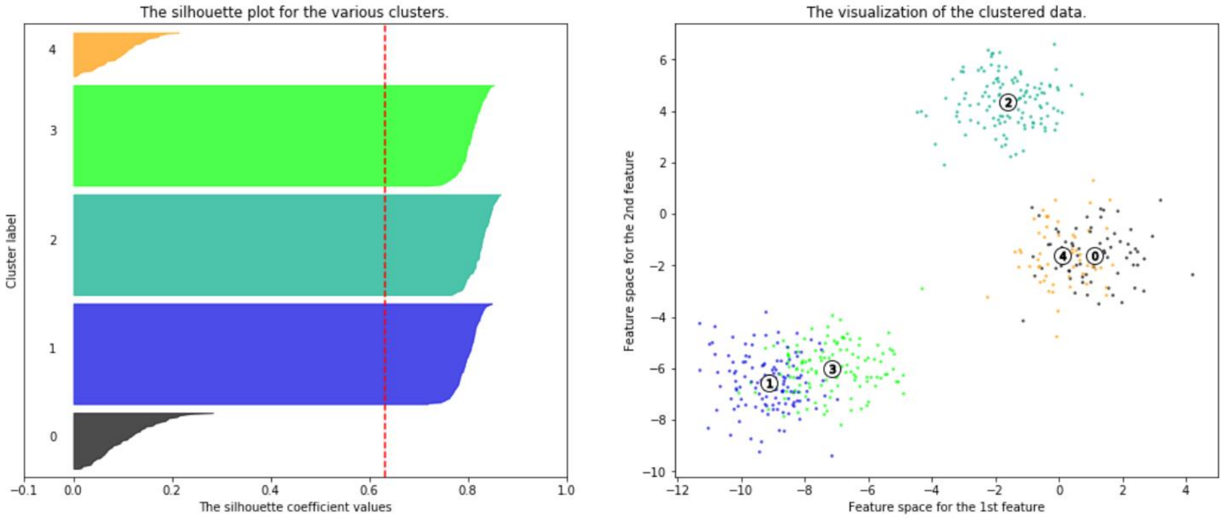
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



The silhouette plot for the various clusters.

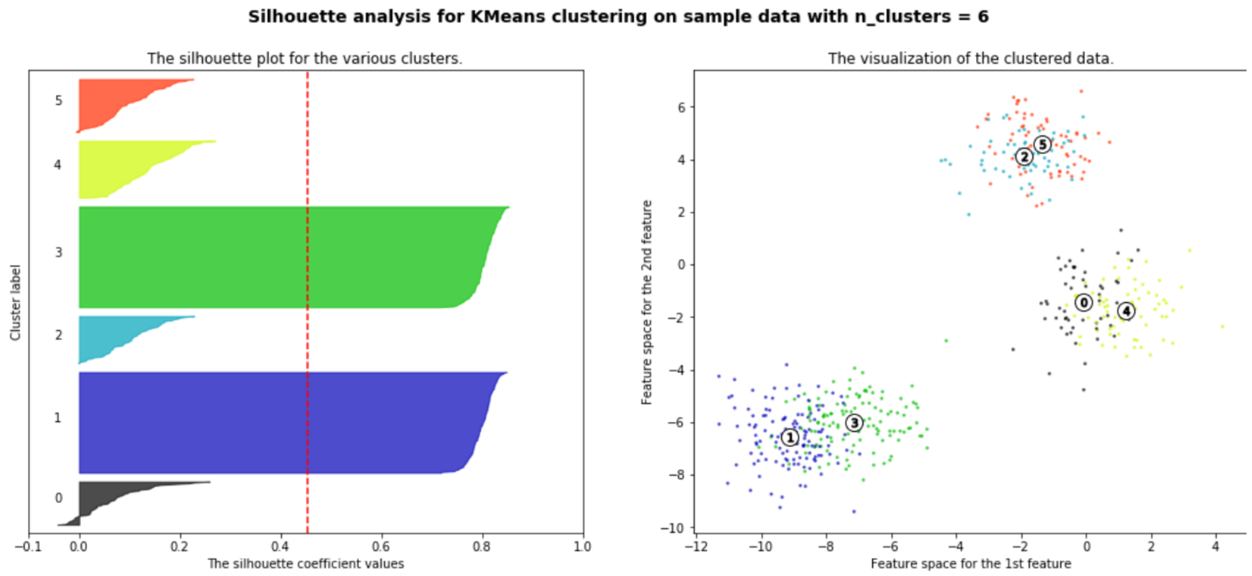
The visualization of the clustered data.

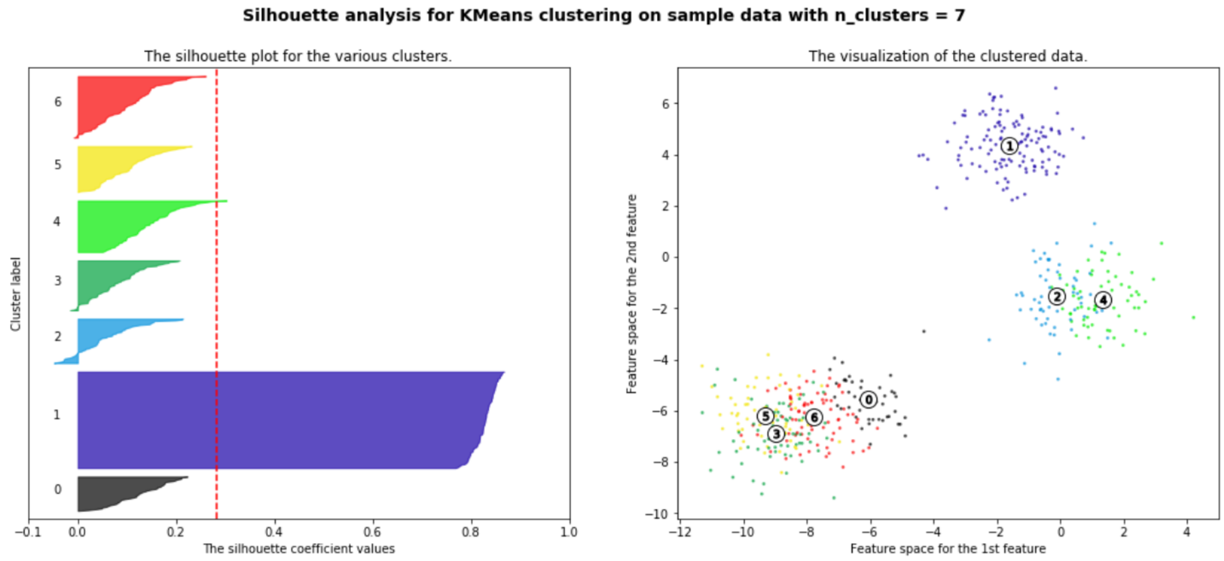
Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



The silhouette plot for the various clusters.

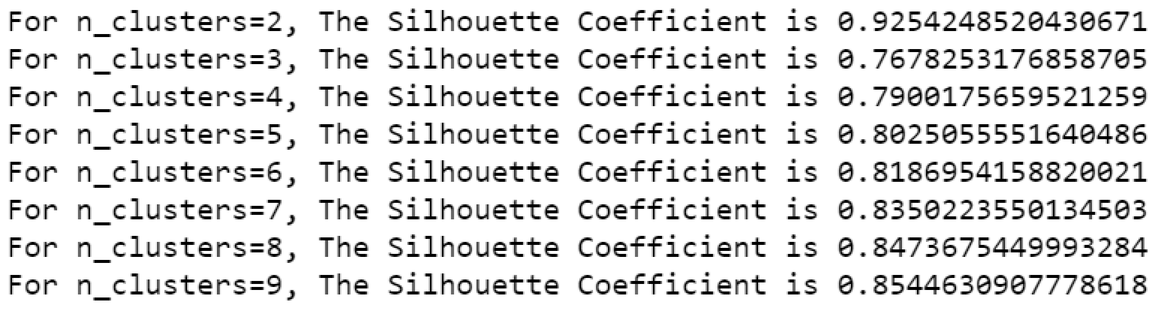
The visualization of the clustered data.





In this project Silhouette analysis shows the cluster quality and helps to find the K clusters through the means. In order to determine the optimal value of K for our dataset, I have used the Silhouette coefficient method.

The following image shows the Silhouette Coefficient with the coefficients for all the clusters and the optimal cluster has been chosen.



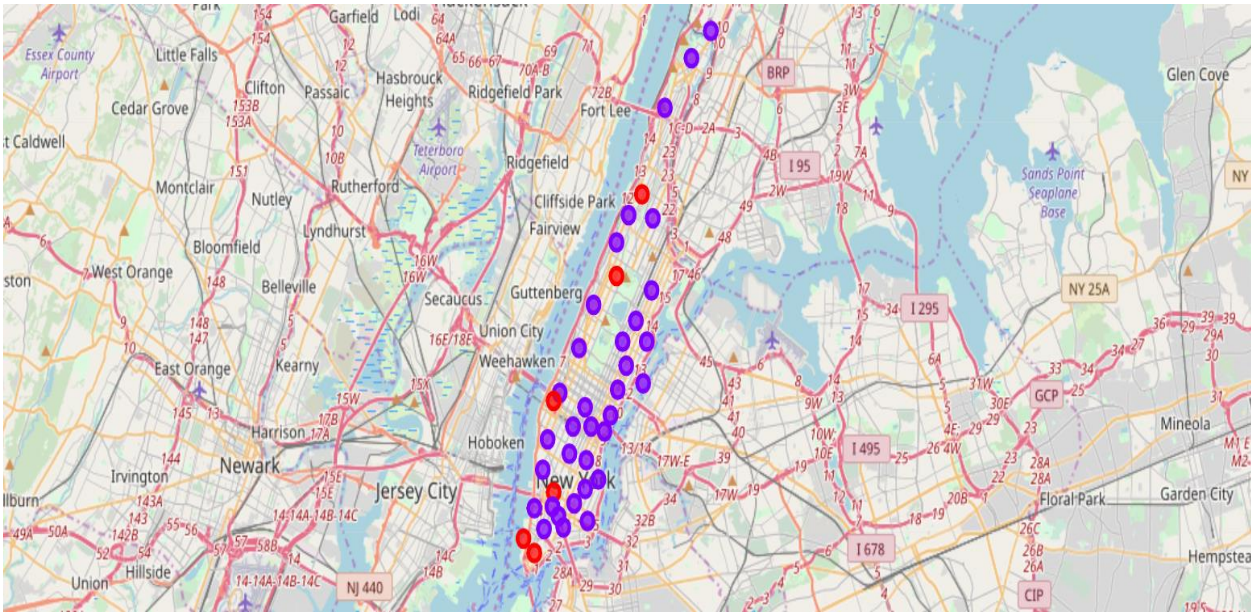
```
For n_clusters=2, The Silhouette Coefficient is 0.9254248520430671
For n_clusters=3, The Silhouette Coefficient is 0.7678253176858705
For n_clusters=4, The Silhouette Coefficient is 0.7900175659521259
For n_clusters=5, The Silhouette Coefficient is 0.8025055551640486
For n_clusters=6, The Silhouette Coefficient is 0.8186954158820021
For n_clusters=7, The Silhouette Coefficient is 0.8350223550134503
For n_clusters=8, The Silhouette Coefficient is 0.8473675449993284
For n_clusters=9, The Silhouette Coefficient is 0.8544630907778618
```

As the maximum coefficient is for cluster 2 that is 0.9254248520430671. Therefore, 2 should be the optimal number of clusters. So, for the purpose of k means 2 clusters have been chosen.

4.**Results:**

From the venues data, the data of restaurants was then taken. The data pertaining to the restaurant s was taken in all the neighborhoods in Manhattan and the analysis was based on this. Clustering analysis was done based only on the restaurants in the neighborhoods of Manhattan.

After the clustering analysis was done using the k means, it is visible that the cluster 0 has a positive value while the cluster 1 has a negative value. But the cluster with positive value doesn't have a very high value. It can be stated that the market in Manhattan is not very saturated. It is visible in this map which shows the various clusters of Manhattan based on neighborhoods.



5. Discussion:

As we have seen from the analysis, there are different types of restauarants in Manhattan which offer tastes of various countries from around the globe. Since Manhattan as a location is not very saturated as we can see from the analysis, a restauarant with great menu and tasty cusines can be opened.

6. Conclusion:

I have performed the analysis on a limited data. But based on this data, it can be safely concluded that an Indian restaurant can be opened in this borough. It can also be recollected that there are different types of restauarants which offer different types of cuisines from around the globe in Manhattan. This means that there is demand for different tastes and it is evident that people want to treat themselves to good food.

Undoubtedly, there is competition but competition is a feature of a healthy market condition. Since my client has a goodwill for his brand in India, he can use this to his advantage. As I had already mentioned that location is the boss for any business, Manhattan seems to satisfy that condition. Therefore, I can conclude that the combination of the location plus my client's brand and his cuisine can together create a success story here in Manhattan as well.