Finding a Perfect Location for opening a restaurant in New York City

1. Introduction / Business Problem:

1.1 Background:

New York City is the most densely populated major city in the United States. It is located at the southern tip of the U.S. state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world. So the need of the restaurants is very obvious. There are lots of restaurants already in the New York City and many more are opening very frequently.

1.2 Problem:

Suppose you want to start your own business by opening a restaurant in the New York City, but you have no idea which will be the best location to open it. Suppose you want to open a Pizza restaurant, you would want to open in such a location that it is easily accessible to the customers. Also you would want that there is no other restaurant of the same category nearby. So to find a perfect location to open a new restaurant we can use the power of Data Science to solve our problem.

So this project aims to solve the following question :

If a person wants to open a new Pizza restaurant, then what will be the perfect location to open it?

1.3 Target Audience:

The persons or a business man who is looking to open a Pizza Restaurant in New York.

2. Acquiring data:

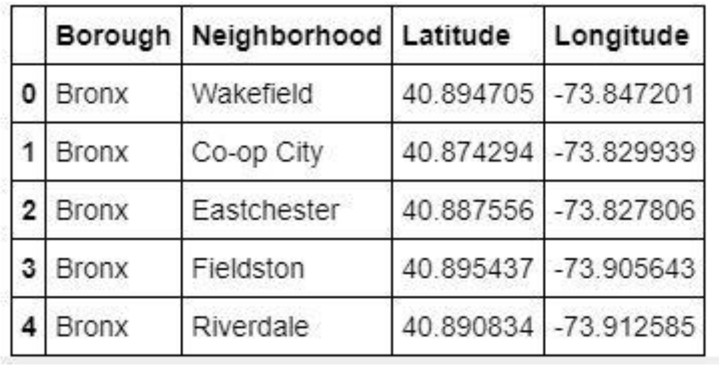
For the above problem to solve, we will need the data of the Boroughs and Neighbourhoods of the New York City. Also we will be requiring the latitudes and longitudes of each neighbourhood. The data of the Boroughs and Neighbourhood can be easily found from the internet. We can scrap the data from various online sources and then combine the data and clean it. Also we will be using the Four Square API to explore the various neighbourhood to get the details of restaurants that are present and how many. So with the help of these, we will try to solve our problem to find a perfect location for a restaurant to open.

3. Methodology:

In this section we will discuss the methodology of the project, i.e. how the all project was carried out. Firstly we will discuss and describe the exploratory data analysis that was done. Also we will be discussing about the machine learning algorithm that was used.

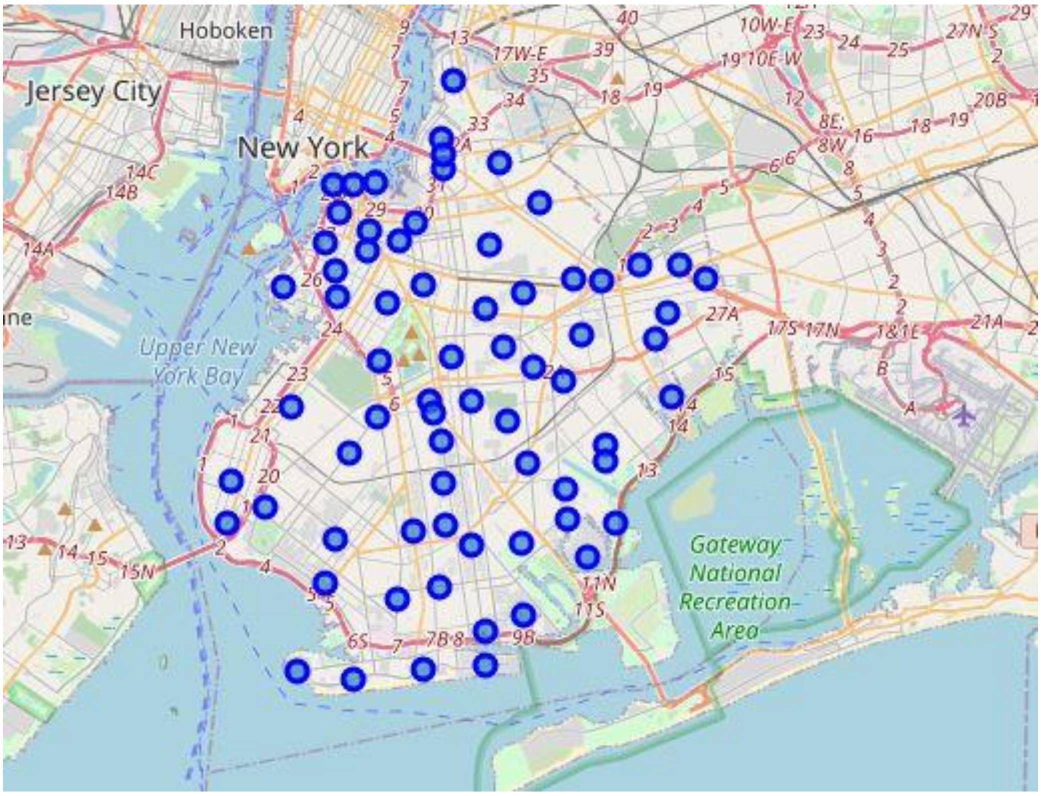
3.1 Exploratory Data Analysis:

The first part was to collect the data and create a Data Frame, containing the details of the Borough and Neighbourhoods of the Ney York. All the details were found in a Json file. We read that Json file and converted the details to a Pandas Data Frame. Here is how the initial data looked. It consisted the list of Boroughs and neighbourhoods and also there Coordinates.



	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

For this project, we assumed that the person wants to open a new Restaurant in Brooklyn, one of the 5 borough of New York. We then filtered out the data which consisted details of Brooklyn borough. Also I used the Python folium library to visualize the neighbourhoods on the map of New York. Here is how it looked:

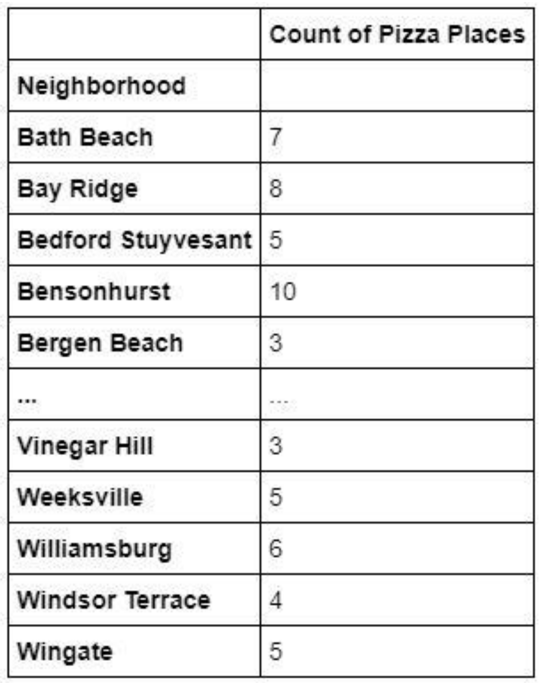


3.2 Four Square API:

When the data of the Brooklyn borough was filtered out, the Four Square API was used to explore each neighbourhoods in the Brooklyn. FourSquare is a location provider company. We can use the Foursquare to explore various things at any location. So using this, we found out all the venues which were present in each neighbourhood. Since our main focus is on the places of Pizza, we filtered out the venues whose category were as Pizza place. Doing this we got the details of the Pizza places in each neighbourhoods.

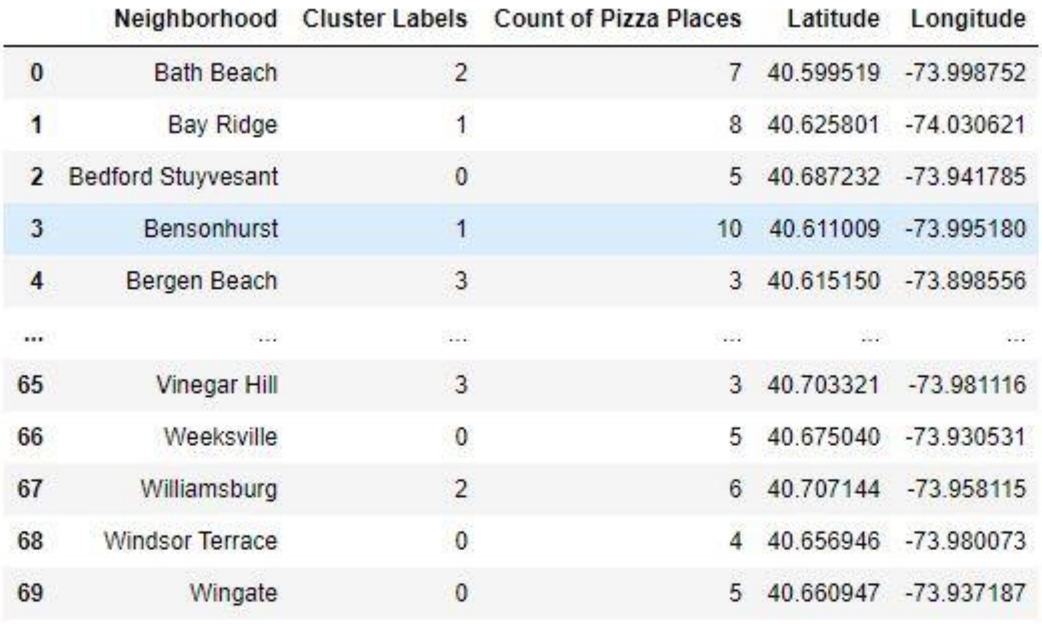
3.3 K-means Clustering:

K-means clustering is a type of unsupervised learning. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. A Data Frame was created consisting the name of neighbourhoods and the number of pizza places in it. The data looked like this :



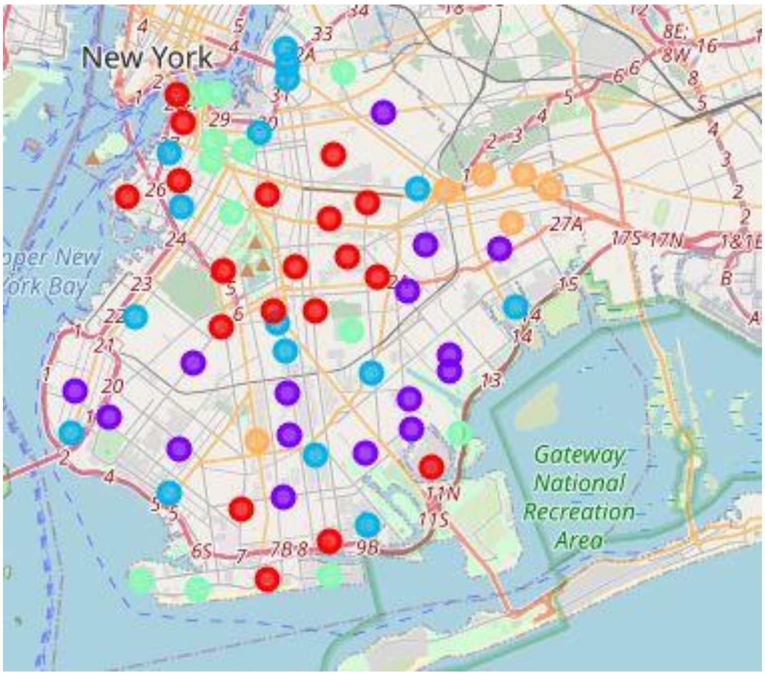
	Count of Pizza Places
Neighborhood	
Bath Beach	7
Bay Ridge	8
Bedford Stuyvesant	5
Bensonhurst	10
Bergen Beach	3
...	...
Vinegar Hill	3
Weeksville	5
Williamsburg	6
Windsor Terrace	4
Wingate	5

We then used the K-means Clustering to cluster the neighbourhoods. We created 4 clusters. By clustering the neighbourhoods, we were able to find the group of neighbourhoods having the least number of pizzas. So by using the Four square API, the problem was solved. This is how the data looked after clustering and adding the labels in the data. All the neighbourhoods were assigned a cluster label.



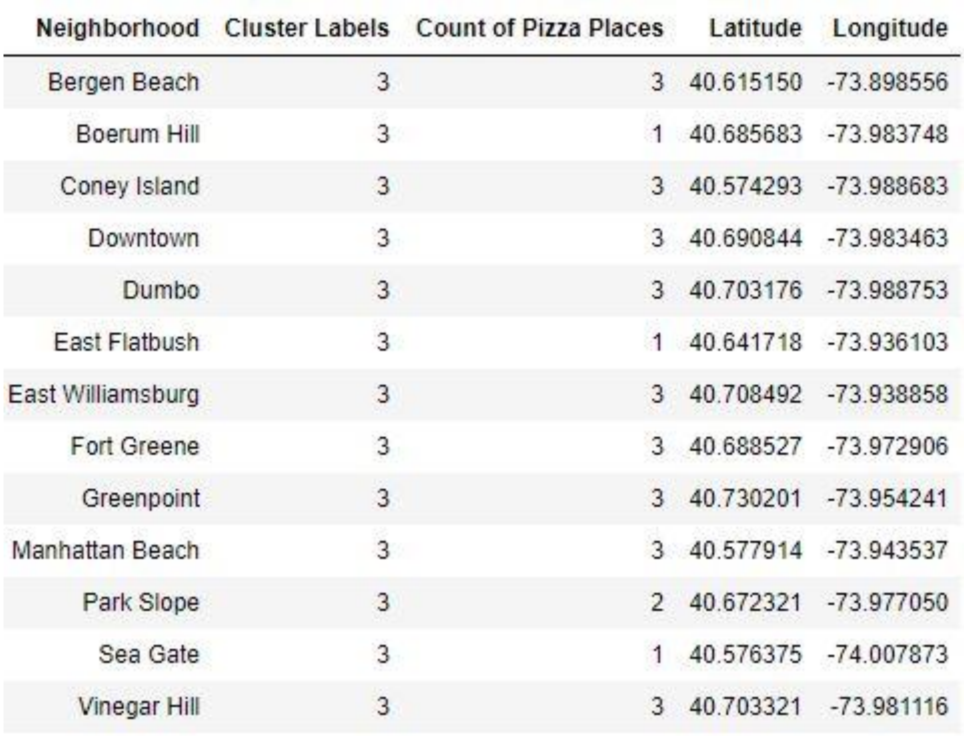
	Neighborhood	Cluster Labels	Count of Pizza Places	Latitude	Longitude
0	Bath Beach	2	7	40.599519	-73.998752
1	Bay Ridge	1	8	40.625801	-74.030621
2	Bedford Stuyvesant	0	5	40.687232	-73.941785
3	Bensonhurst	1	10	40.611009	-73.995180
4	Bergen Beach	3	3	40.615150	-73.898556
...	...	...	...	...	...
65	Vinegar Hill	3	3	40.703321	-73.981116
66	Weeksville	0	5	40.675040	-73.930531
67	Williamsburg	2	6	40.707144	-73.958115
68	Windsor Terrace	0	4	40.656946	-73.980073
69	Wingate	0	5	40.660947	-73.937187

Also, the clusters can be visualized on the map. Again, the folium library was used to visualize the cluster on the map. Different clusters are marked in different colors.



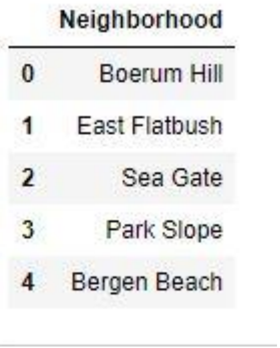
4. Result:

After clustering the neighbourhoods and analysing each cluster, we were able to tell the names of neighbourhood having the least number of Pizza places. We found that the Cluster 3 was the cluster having the neighbourhoods which had the least number of pizza places. Our main problem was to find the best location to open a new Pizza restaurant, for this the location should have less number of already opened pizza place. So i was able to find the names of neighbourhoods having the least number of Pizza places. Ideally these were the best location to open a new Pizza restaurant in the Brooklyn. This is how the data of the cluster 3 looked.



Neighborhood	Cluster Labels	Count of Pizza Places	Latitude	Longitude
Bergen Beach	3	3	40.615150	-73.898556
Boerum Hill	3	1	40.685683	-73.983748
Coney Island	3	3	40.574293	-73.988683
Downtown	3	3	40.690844	-73.983463
Dumbo	3	3	40.703176	-73.988753
East Flatbush	3	1	40.641718	-73.936103
East Williamsburg	3	3	40.708492	-73.938858
Fort Greene	3	3	40.688527	-73.972906
Greenpoint	3	3	40.730201	-73.954241
Manhattan Beach	3	3	40.577914	-73.943537
Park Slope	3	2	40.672321	-73.977050
Sea Gate	3	1	40.576375	-74.007873
Vinegar Hill	3	3	40.703321	-73.981116

As a result these are the top 5 neighbourhoods, where the restaurant can be opened.



	Neighborhood
0	Boerum Hill
1	East Flatbush
2	Sea Gate
3	Park Slope
4	Bergen Beach

4. Discussion:

As we analysed each cluster, we noted that the cluster3 were having the least number of pizza places. So we were able to find the top 5 neighbourhoods having the least number of Pizza places. Also we noted that the Cluster4 had the largest number of Pizza places, followed by cluster0.

4. Conclusion:

The solution to our problem was found. The top 5 places to open a new pizza restaurant were found. We used the data analysis, machine learning algorithms and the Four square API to solve our problem.