CAPSTONE PROJECT COURSERA : FINDING LOCATION IN BROOKLYN TO OPEN UP A BAKERY

By: Vaibhav Arora

INTRODUCTION

New York city is the financial capital of USA. On a daily basis many new business keep on opening and closing. A lot of factors need to be kept in mind when opening a business. One of them is the appropriate location because numerous business fail due to poor location choice.

PROBLEM STATEMENT

My client wishes to open up a Bakery in New York particularly in the Brooklyn area. As Brooklyn has a lot of neighbourhoods, he wishes to get the most appropriate location to open up his bakery.

After careful analysis we came upon a decision that we need to find the location that does not have many bakery but that that alone is not sufficient. For a bakery to perform successfully the area must show potential that opening a bakery there would be profitable. This potential of the areas will be determined by the top venues of that area. We want the top venues of that area to be something in the line of bakery or compliments it. So we will select those areas which have Café, Coffee shops and bars as their top venues keeping in mind that there must not be lot of bakery in that area as it would result in a tough competition.

DATA

DATA SOURCE

To perform this analysis, we will need the following data:

- A dataset that contains information about neighbourhoods of Brooklyn
- Venues's information of each neighbourhood.

Neighbourhoods data of Brooklyn will be obtained from dataset link: https://geo.nyu.edu/catalog/nyu_2451_34572

Top venues data will be obtained from Foursquare through an API.

DATA EXTRACTION AND CLEANING

The data collected from New York csv file will be converted to a dataframe using pandas. Then from that dataframe a new dataframe will be extracted which contains neighbourhood information of Brooklyn.

Once it is done the co-ordinates of each neighbourhood will be extracted using geolocator and a dataframe will be created having the neighbourhood of Brooklyn and its latitude and longitude.

	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

Then we will use Foursquare API to extract top venues of each neighbourhood and merge it with the dataframe containing neighbourhood information of Brooklyn.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bath Beach	Chinese Restaurant	Pizza Place	Pharmacy	Italian Restaurant	Bubble Tea Shop	Fast Food Restaurant	Gas Station	Park	Cantonese Restaurant	Surf Spot
1	Bay Ridge	Spa	Italian Restaurant	Greek Restaurant	Pizza Place	Pharmacy	American Restaurant	Chinese Restaurant	Bar	Bagel Shop	Mediterranean Restaurant
2	Bedford Stuyvesant	Pizza Place	Coffee Shop	Bar	Café	Deli / Bodega	Park	Discount Store	Cocktail Bar	New American Restaurant	Tiki Bar
3	Bensonhurst	Chinese Restaurant	Italian Restaurant	Sushi Restaurant	Donut Shop	Ice Cream Shop	Liquor Store	Russian Restaurant	Cha Chaan Teng	Noodle House	Sporting Goods Shop
4	Bergen Beach	Harbor / Marina	Playground	Donut Shop	Athletics & Sports	Baseball Field	Fish Market	Fish & Chips Shop	Filipino Restaurant	Field	Fast Food Restaurant
5	Boerum Hill	Dance Studio	Coffee Shop	Bar	French Restaurant	Furniture / Home Store	Bakery	Sandwich Place	Arts & Crafts Store	Gym / Fitness Center	Spa
6	Borough Park	Bank	Pizza Place	Pharmacy	Fast Food Restaurant	Hotel	Grocery Store	Coffee Shop	Restaurant	Chinese Restaurant	Café

METHODOLOGY

From the data we will analyse venues for each neighbourhood. We will collect the venues and display them on the map and their distribution over the neighbourhoods will be visualized using bar graphs.

Once we do this we will have a detailed information of each neighbourhood. Then we will perform clustering of neighbourhoods based on the criteria that we initially discussed.

Finally the relevant neighbourhoods will be selected and presented to the client to make final decision.

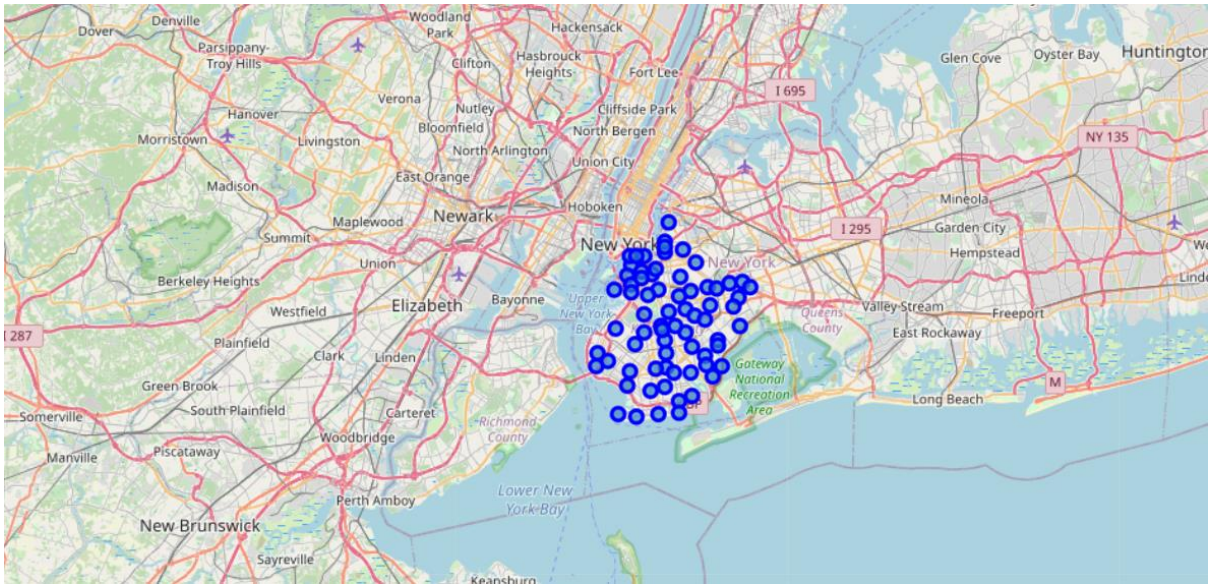
The tools and methods used in this project are:

Foursquare API places : This API allows us to access the venues based on the geographic co-ordinates. It also allows us to get detailed analysis of the venues like venue category, photos, ratings, reviews and much more.

K-means Clustering : This is a machine learning algorithm used to divide n observations into k clusters using the nearest mean distance of each observation from the cluster centroid.

DATA ANALYSIS

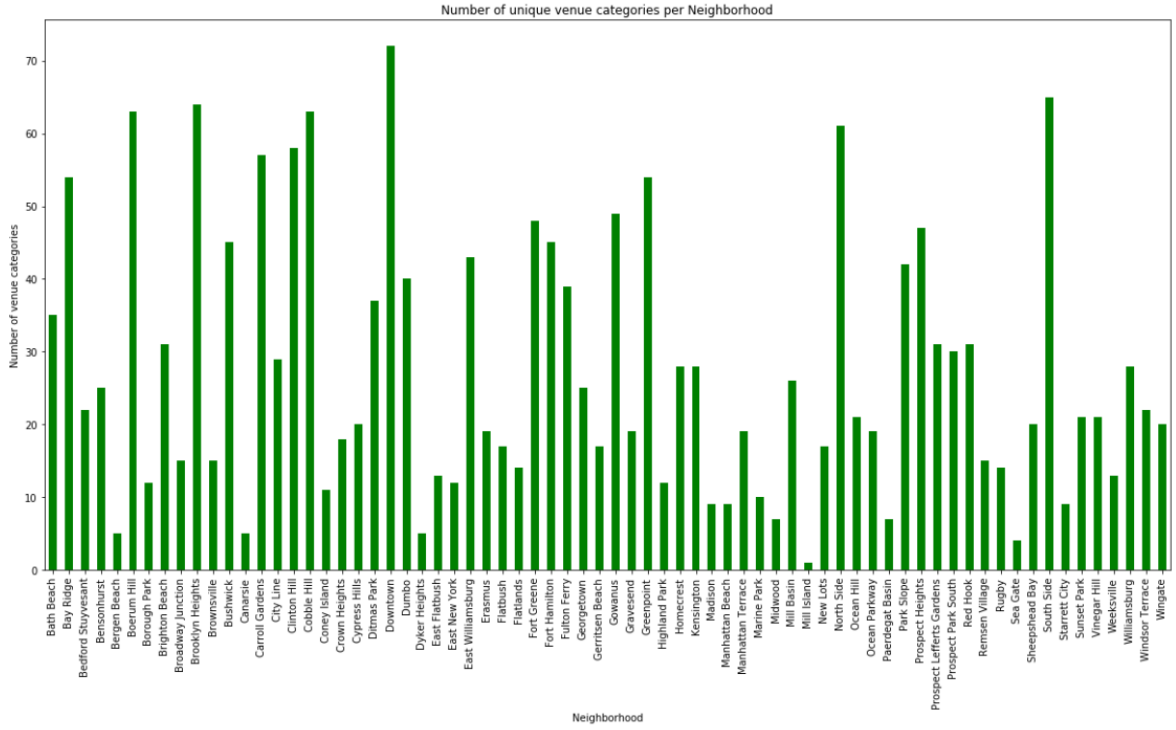
The data that was collected was cleaned using pandas and geolocator as discussed above. Then the data was used to map neighbourhoods of Brooklyn using folium



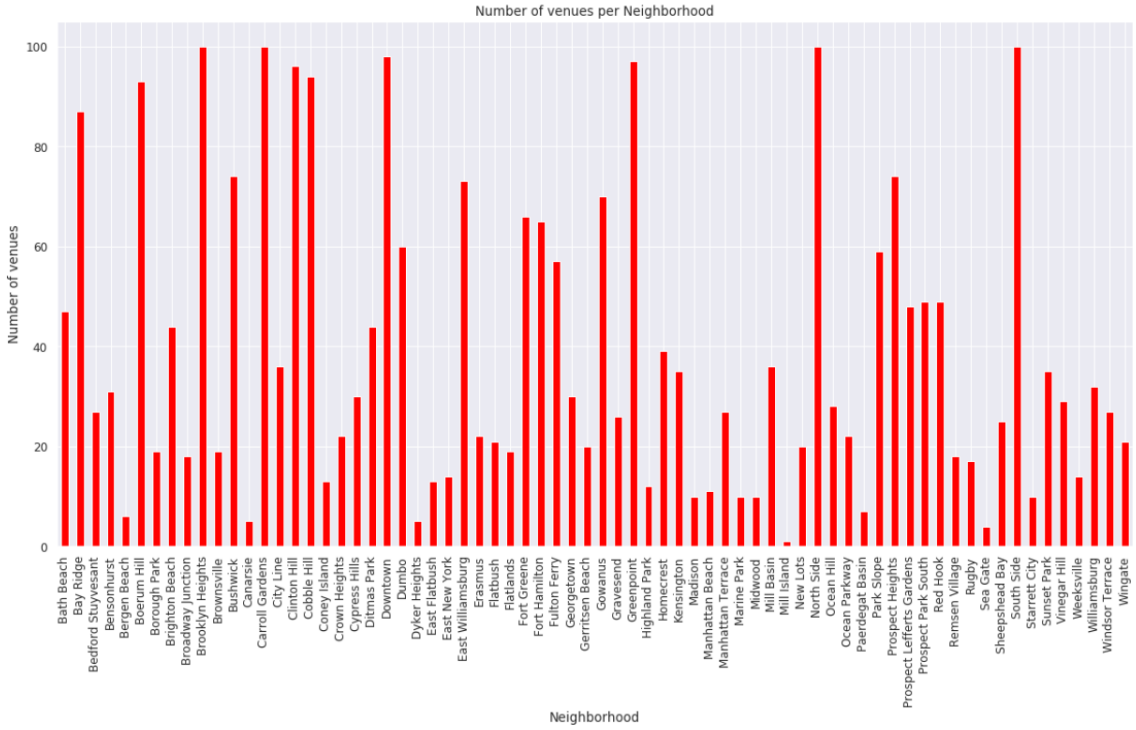
Once the data is mapped then we used Foursquare API to collect the venue details of each neighbourhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
1	Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
2	Bay Ridge	40.625801	-74.030621	Leo's Casa Calamari	40.624200	-74.030931	Pizza Place
3	Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
4	Bay Ridge	40.625801	-74.030621	The Bookmark Shoppe	40.624577	-74.030562	Bookstore

Then we make an analysis of neighbourhoods as to how many different venue categories are present in each neighbourhood. This is shown by a bar graph below.



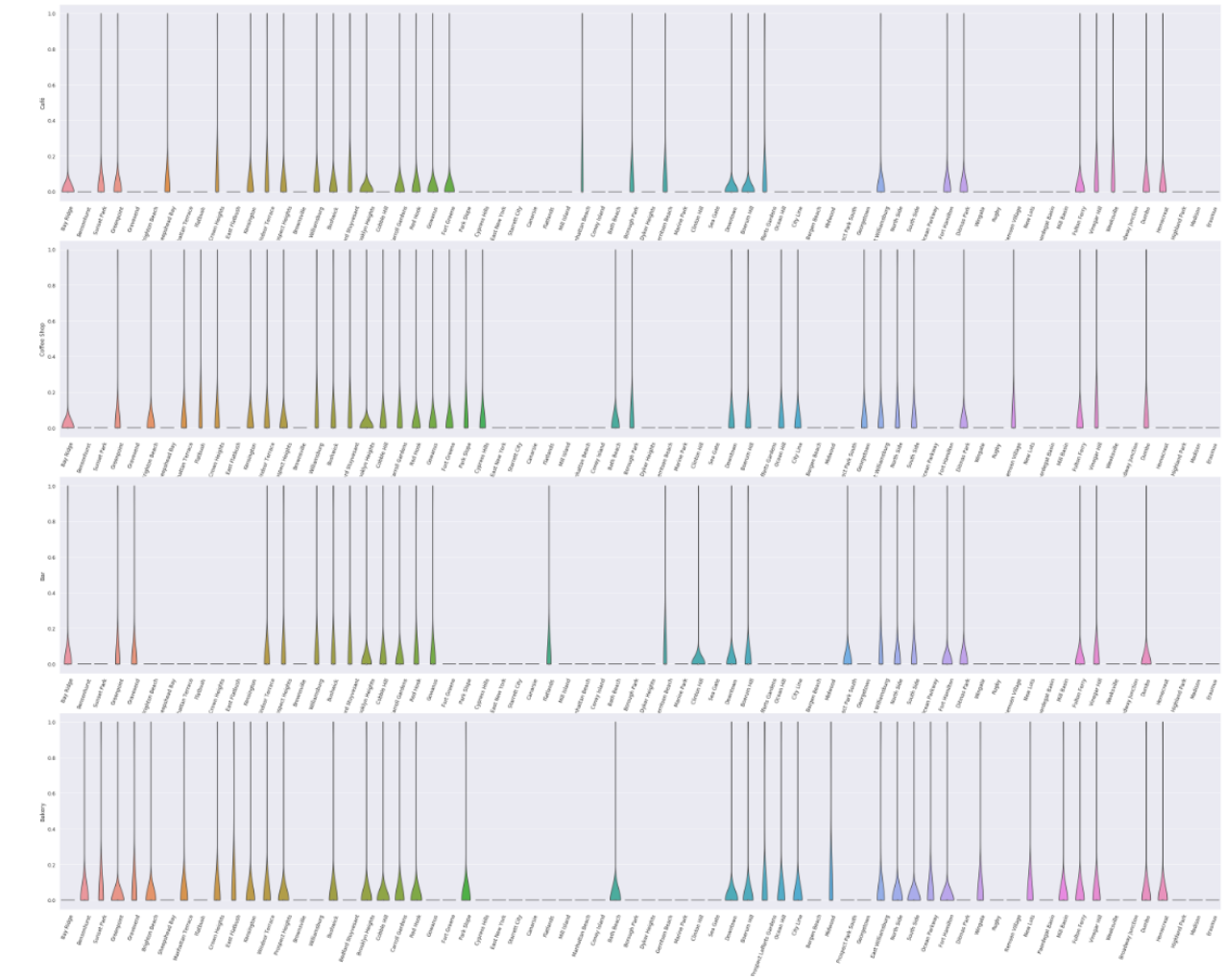
Then we had a look at how many different venues are there in each neighbourhood.



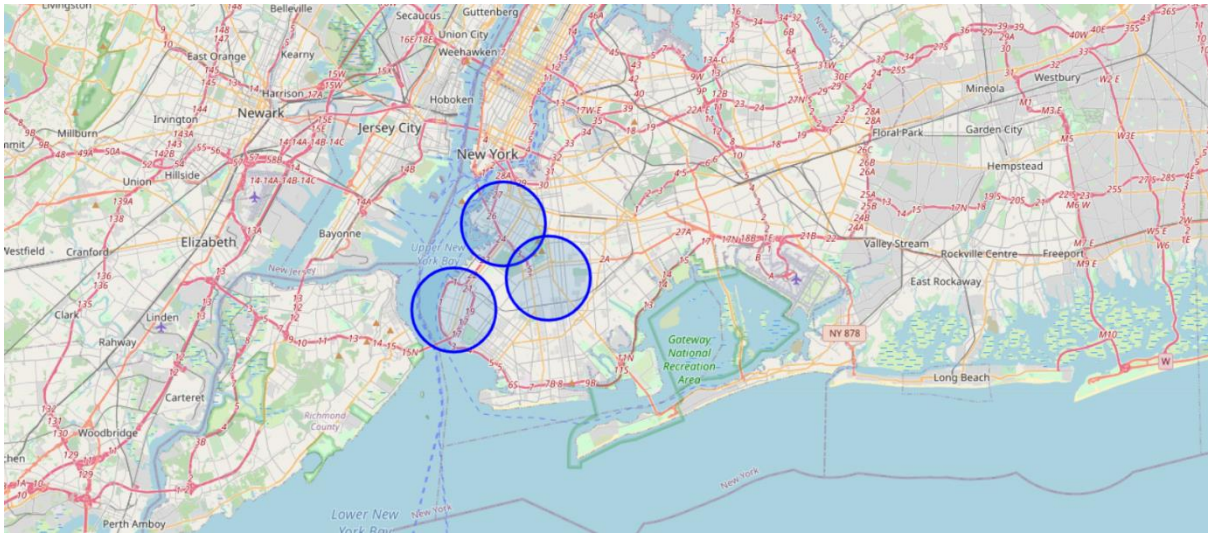
After that we created a dataframe containing the top 10 venues for every neighbourhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bath Beach	Chinese Restaurant	Pizza Place	Pharmacy	Italian Restaurant	Bubble Tea Shop	Fast Food Restaurant	Gas Station	Park	Cantonese Restaurant	Surf Spot
1	Bay Ridge	Spa	Italian Restaurant	Greek Restaurant	Pizza Place	Pharmacy	American Restaurant	Chinese Restaurant	Bar	Bagel Shop	Mediterranean Restaurant
2	Bedford Stuyvesant	Pizza Place	Coffee Shop	Bar	Café	Deli / Bodega	Park	Discount Store	Cocktail Bar	New American Restaurant	Tiki Bar
3	Bensonhurst	Chinese Restaurant	Italian Restaurant	Sushi Restaurant	Donut Shop	Ice Cream Shop	Liquor Store	Russian Restaurant	Cha Chaan Teng	Noodle House	Sporting Goods Shop
4	Bergen Beach	Harbor / Marina	Playground	Donut Shop	Athletics & Sports	Baseball Field	Fish Market	Fish & Chips Shop	Filipino Restaurant	Field	Fast Food Restaurant
5	Boerum Hill	Dance Studio	Coffee Shop	Bar	French Restaurant	Furniture / Home Store	Bakery	Sandwich Place	Arts & Crafts Store	Gym / Fitness Center	Spa
6	Borough Park	Bank	Pizza Place	Pharmacy	Fast Food Restaurant	Hotel	Grocery Store	Coffee Shop	Restaurant	Chinese Restaurant	Café

Then we created a violin graph to see the frequency of our criteria venue i.e. presence of Café, Coffee shop and bars along with bakery in every neighbourhood.

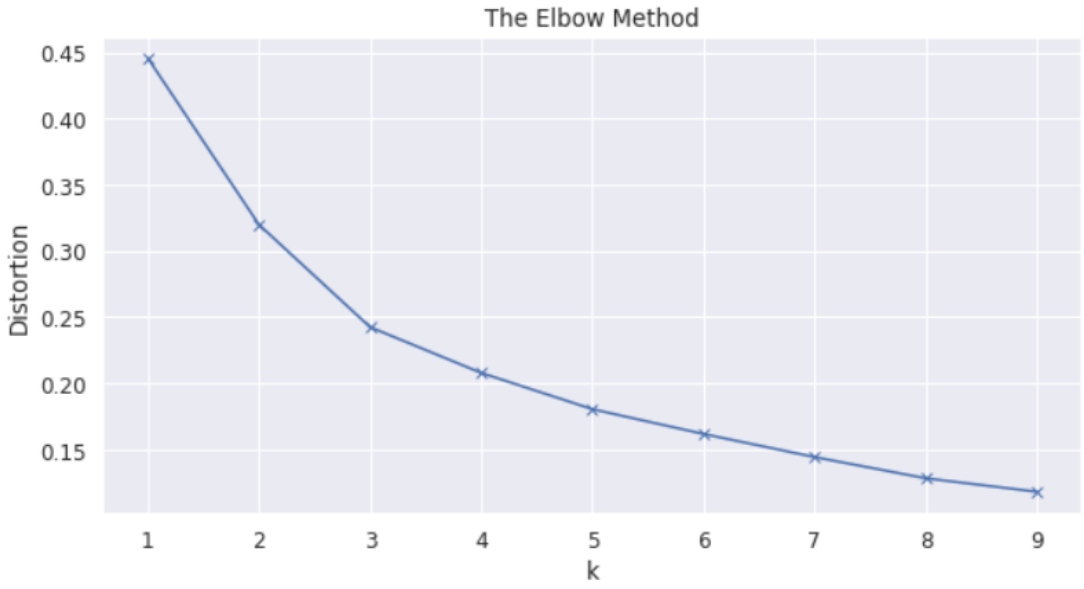


From this we made a preliminary analysis and selected the 3 best locations Brooklyn has to offer to our client and pointed them on the map.

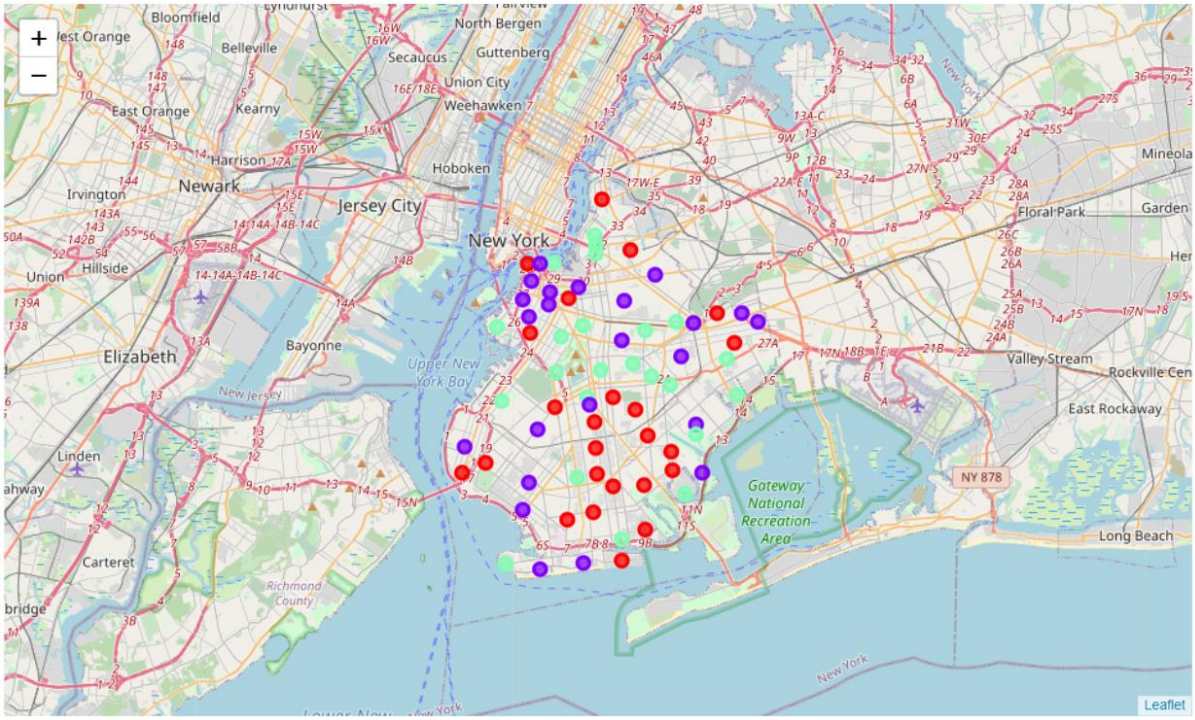


ALGORITHM

The Machine Learning algorithm used in this project is a clustering algorithm called K-means Algorithm. This algorithm uses the mean of distance between the points of cluster and its centroid. The nearest points are then put together in the cluster. But there is a problem with this algorithm that the number of cluster has to be set in the beginning. So in order to find the optimum number of clusters we will use what is called an elbow method. We will plot the distortion against number of clusters but we know as we increase the number of cluster the distortion will decrease. So what we look for here is a sudden change in distortion which forms an elbow like corner on the graph. In this case the optimum value of k=3.



Once the number of cluster is determined we apply k-means algorithm and determine the cluster of each neighbourhood and point it on a map.



CONCLUSION

In this project I have devised a clustering model that clusters the neighbourhood of Brooklyn according to the criteria devised in the business understanding. We have seen that the result in the preliminary analysis is in accordance with the clustering algorithm. Among the clusters formed the second cluster meets the criteria of our business and suggest large number of places that can be potentially good for opening up a bakery.

FUTURE SCOPE

In this project the primary criteria that was uses was the venue types of the neighbourhood. We can use more data like area demographics, area wise income of people to further reduce the number of results and increase the efficiency of the cluster formed.