The Battle of the Neighbourhoods (New York)

Table of contents

Introduction: Business Problem

In this Capstone project I am trying to figure out the most popular venue in New York City and to find out the optimal location/neighbourhood that is not already crowded with that popular venue. I am also trying to analyse the venue in each Borough in New York City. Specifically, this report will be targeted to stakeholders interested in looking most popular venue in New York City and to find appropriate location that is not already crowded.

Since there are lots of variety of venues in New York, we will try to detect **most popular venue and location that are not already crowded with that venue**. We would also prefer locations **as close to city as possible that covers most nearby neighbourhoods**, assuming that first two conditions are met.

Data

As per my problem description, below mentioned factors will influence my decision:

- Variety of venues in the neighbourhood.
- Number of different venue type in each Borough.
- Number of neighbourhoods in each Borough that is not already crowded with most popular venue.
- Location of each uncrowded neighbourhood to cluster them and find out highly dense cluster.

I will be using the following data sources to extract the required information:

- Raw data for New York City will be downloaded from external source.
- Borough, neighbourhood and their latitude and longitude data will be extracted from downloaded data.
- Number of venues, venue category and corresponding latitude and longitude of each venue in every neighbourhood will be obtained using **Foursquare API**.

So by using these information we will try to find out most common venue and optimal location for such venue. These neighbourhood data will be clustered by using ***K-mean Clustering*** method to identify optimal group of neighbourhoods which are not already crowded with most common venue.

After data download and preprocessing final data has the following attributes and below is the snapshot of the dataset:

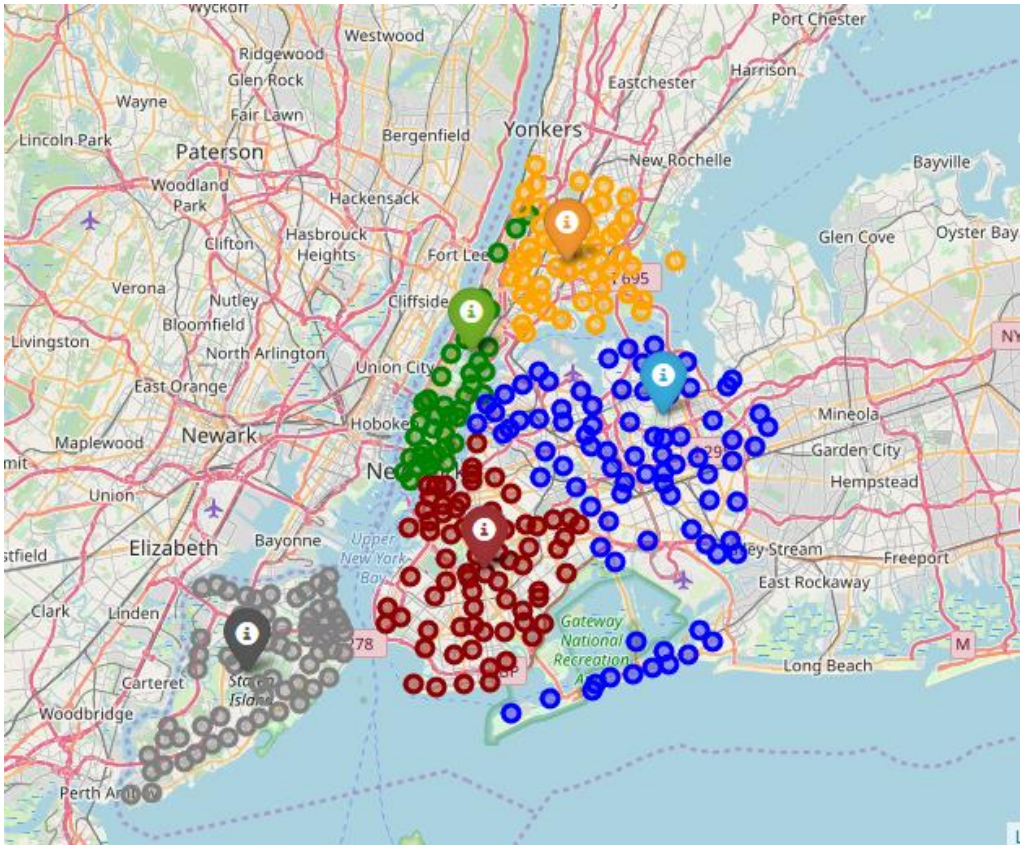
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

We extract 'Borough' and 'Neighbourhood' columns separately and examined that there are 5 borough and 306 neighbourhoods.

Let's graphically represent the each borough and their neighbourhoods.

Create a map of New York with each Borough & neighbourhoods

We create a map with each borough and their neighbourhoods. To distinguish each borough and their neighbourhoods, we used different colour for each borough. In the below figure we can clearly examined each borough and their neighbourhoods.



Next we are going to explore each neighbourhood and find different venue and its category. To explore neighbourhood and to get venue information, we will be using Foursquare API.

Exploring neighbourhood by using Foursquare API

We explored each neighbourhood and extract information of upto 100 venues in each neighbourhood and store them in a data frame. After preprocessing data, we get the venue data along with its category and geolocation data. Below is the snapshot of processed data:

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bronx	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Bronx	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Bronx	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Bronx	Wakefield	40.894705	-73.847201	Walgreens	40.896687	-73.844850	Pharmacy
4	Bronx	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Methodology

In this capstone project our main objective is to explore each borough and find out the **most popular venue** and then figure out the **optimal location** which is not crowded with that popular venue and cluster the neighbourhoods to identify highly dense cluster of neighbourhood which have not already crowded.

In first step we have collected the required data from external source and then explored each neighbourhood by using Foursquare API to find venues information in each neighbourhood.

Second step in our analysis will be to find the top 5 most common venues in each borough.

In the third and final step we will filter the original data with neighbourhoods which has no top most common venue. Then visualize each neighbourhood. Our main focus is to find out most promising area/zone that meet some basic requirements established in discussion with stakeholders. We will prompt map of all such locations and also create clusters (using **k-means clustering**) of those locations to identify highly dense neighbourhoods which would be target for optimal venue location by stakeholders.

Analysis

To find out the top most common venue in each Borough, first we aggregate the data by borough column and sum the number of different venues in each borough as below is the snapshot of aggregated data.

	Borough	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Entertai
0	Bronx	1	0	0	2	0	10	0	1	1	0	2	1	0	
1	Brooklyn	1	0	0	0	1	48	7	0	2	3	18	0	11	
2	Manhattan	4	2	1	2	0	82	1	1	2	5	30	4	4	
3	Queens	2	0	2	0	0	22	0	0	4	2	1	1	3	
4	Staten Island	0	0	0	0	0	13	0	1	0	0	2	1	1	

After that we aggregate the data by borough and neighbourhood combined and sum the number of different venues in each borough and neighbourhood combinations. We aggregate the data by this way to identify the neighbourhoods which are having most common venue and neighbourhoods which are not having common venue. So that we can cluster neighbourhoods which are not having most common venue and figure out optimal dense cluster/zone.

	Borough	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Cr S
0	Bronx	Allerton	0	0	0	0	0	0	0	0	0	0	0	0	
1	Bronx	Baychester	0	0	0	0	0	0	0	1	0	0	0	0	
2	Bronx	Bedford Park	0	0	0	0	0	0	0	0	0	0	0	0	
3	Bronx	Belmont	0	0	0	0	0	1	0	0	0	0	0	0	
4	Bronx	Bronxdale	0	0	0	0	0	0	0	0	0	0	0	0	

After some processing steps, we examined the top 5 common venues in each borough. Below is the snapshot of top 5 common venues in each borough.

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bronx	Pizza Place	Deli / Bodega	Donut Shop	Pharmacy	Supermarket
1	Brooklyn	Pizza Place	Coffee Shop	Bar	Bakery	Deli / Bodega
2	Manhattan	Coffee Shop	Italian Restaurant	American Restaurant	Café	Pizza Place
3	Queens	Pizza Place	Deli / Bodega	Chinese Restaurant	Bakery	Donut Shop
4	Staten Island	Bus Stop	Pizza Place	Italian Restaurant	Deli / Bodega	Bagel Shop

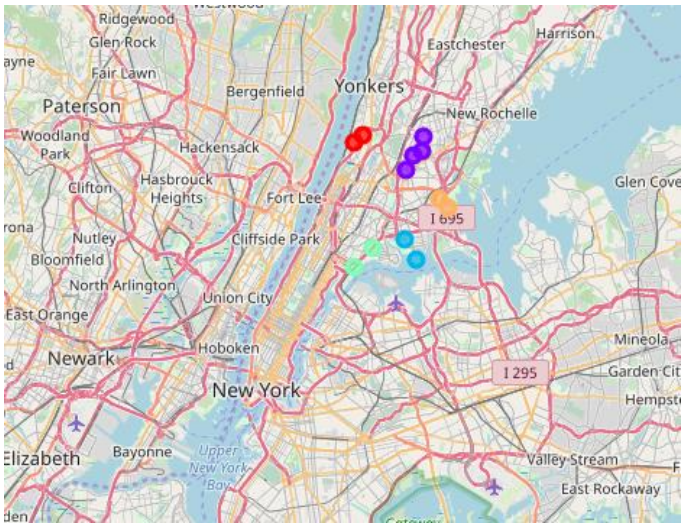
Here we figure out that each borough has pizza place as most common venue. So we start analysing each borough and cluster non-pizza place neighbourhoods by ***k-mean*** clustering.

Analyse Each Borough

We extracted data by borough wise and analyse each borough and its neighbourhoods.

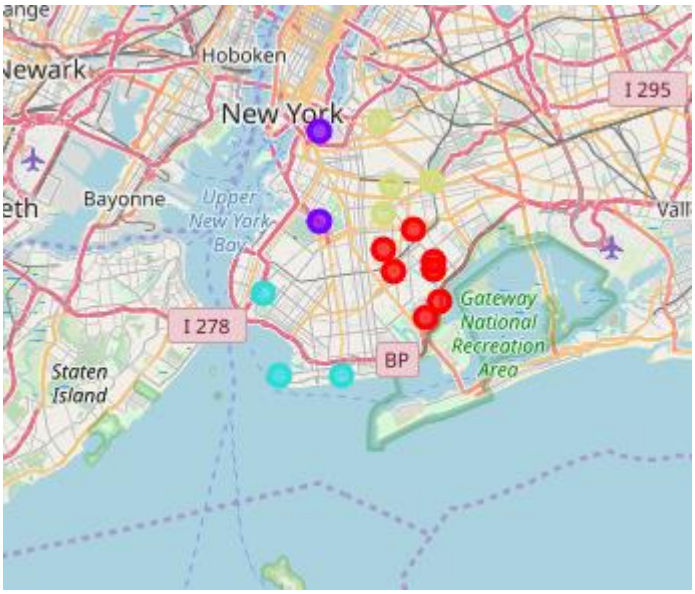
Cluster Analysis for Bronx

We found that there are 40 pizza places and 12 non pizza places in Bronx borough. Here we clustered neighbourhoods which are not having pizza place into optimal number of clusters.



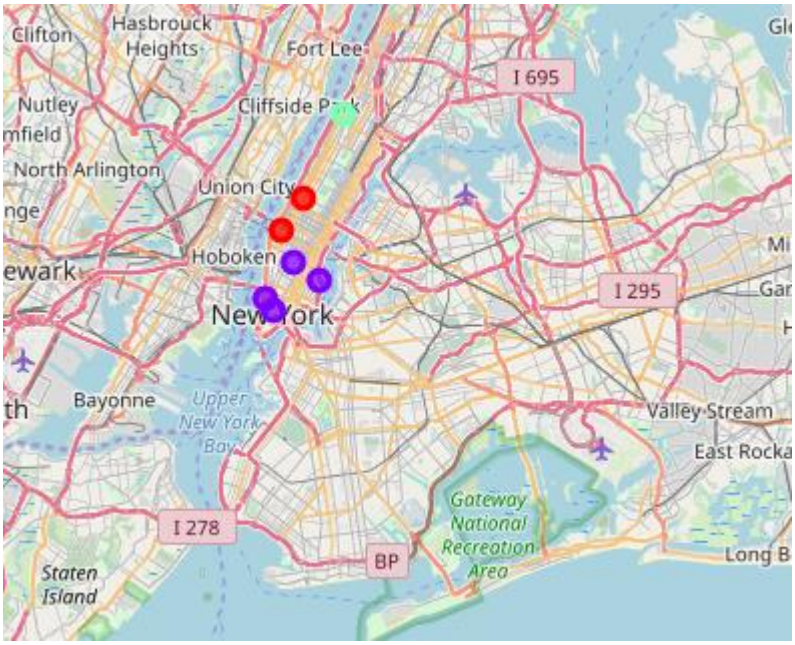
Cluster Analysis for Brooklyn

In Brooklyn borough there are 54 pizza places and 16 non pizza places. Below is the cluster visualization of Brooklyn and its neighbourhoods which are not having pizza places.



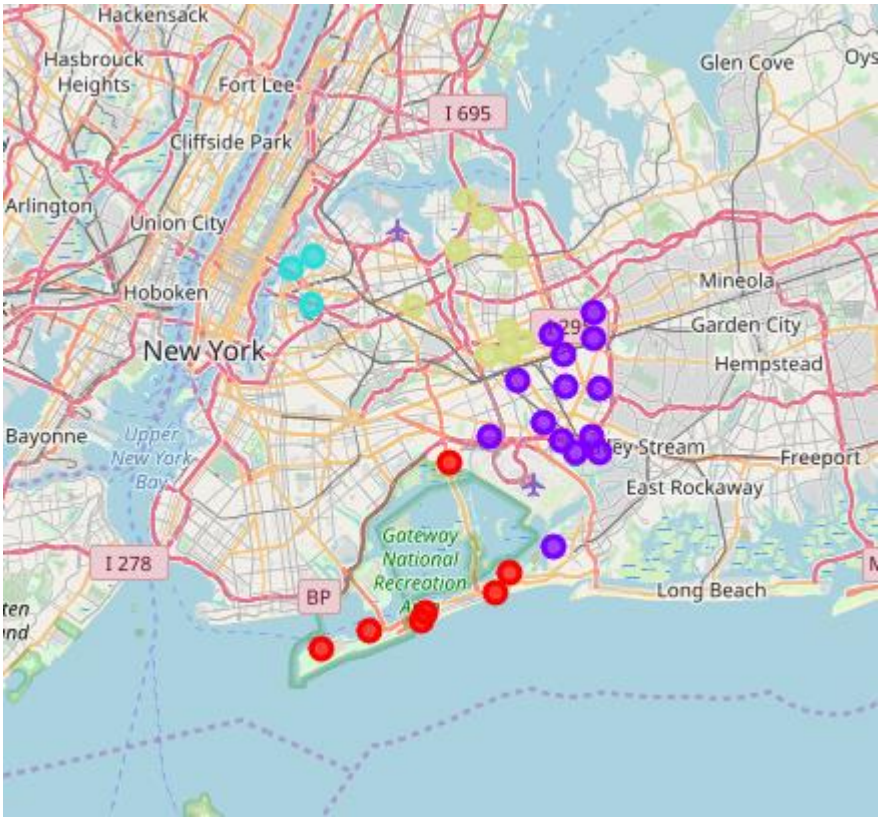
Cluster Analysis for Manhattan

In Manhattan Borough there are 33 pizza places and 7 non pizza places.



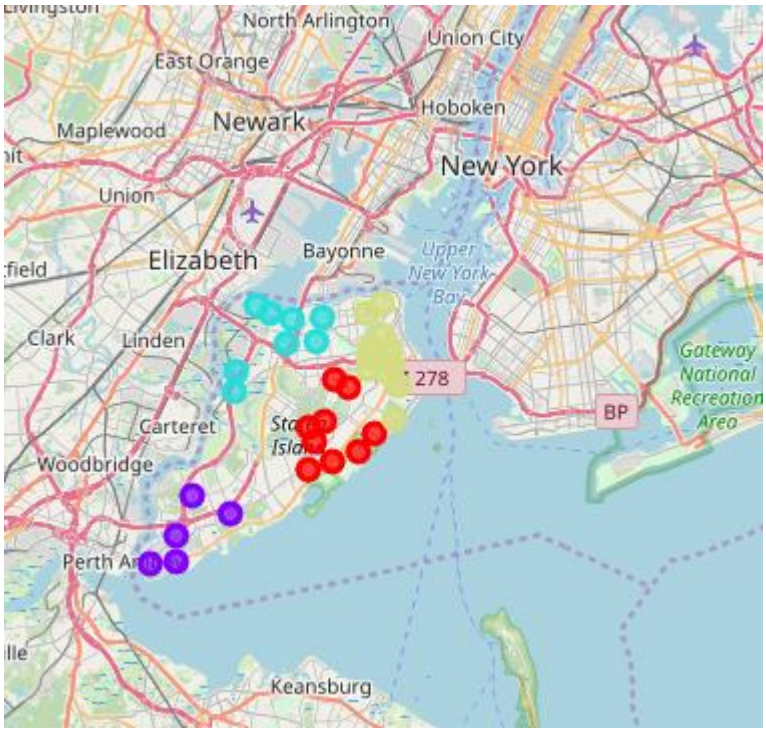
Cluster Analysis for Queens

In queens borough there are 48 pizza places and 33 non pizza places.



Cluster Analysis for Staten Island

In Staten Island there are 29 pizza places and 33 non pizza places.

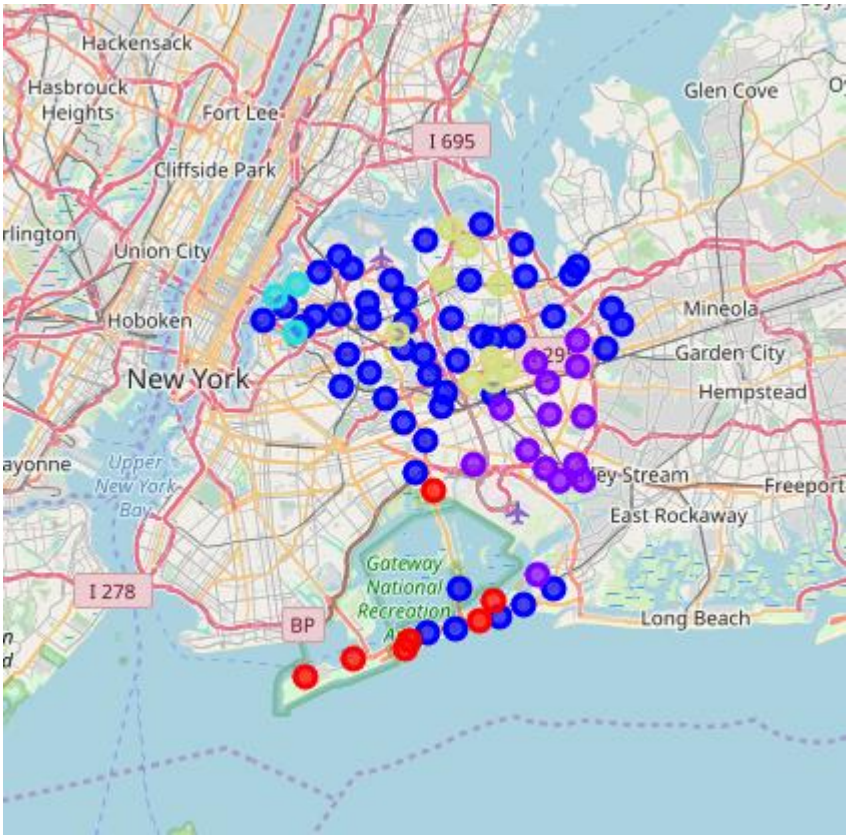


From the above visual representations we can see that **Queens and Staten Island** both are not much crowded with pizza places. In Queens' **cluster 1** consists majority of neighbourhoods so it can be considered for optimal

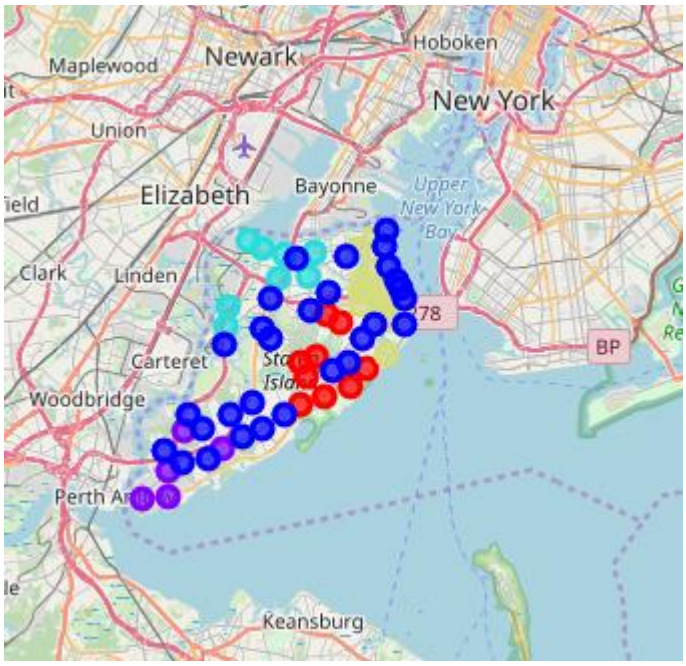
location. Similarly in Staten Island we can see that all clusters are highly dense and **cluster 3** is more appropriate cluster to be considered for optimal location.

Let's see which cluster is optimal for consideration by adding neighbourhoods which are having pizza place in Queens and Staten Island. In below figures blue circle represents neighbourhoods which are having pizza places.

Queens Cluster with pizza places:

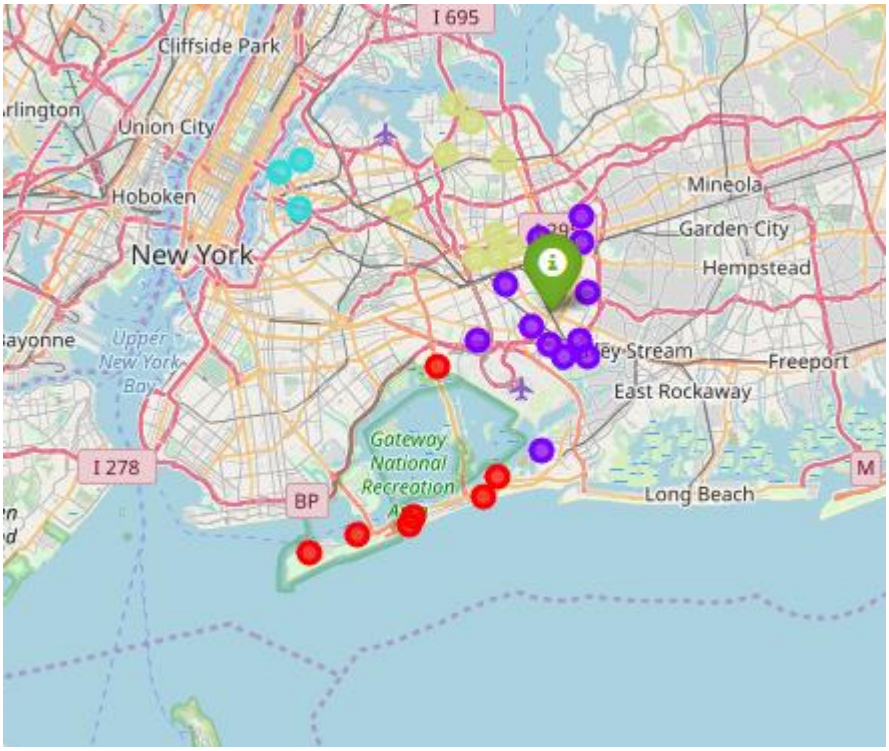


Staten Island cluster with pizza places:



Here we can clearly see that in Staten Island all clusters are more closely surrounded by pizza place. While in Queens' cluster 1 is not as much crowded with pizza place. So we can consider **Queens' cluster 1** for optimal location.

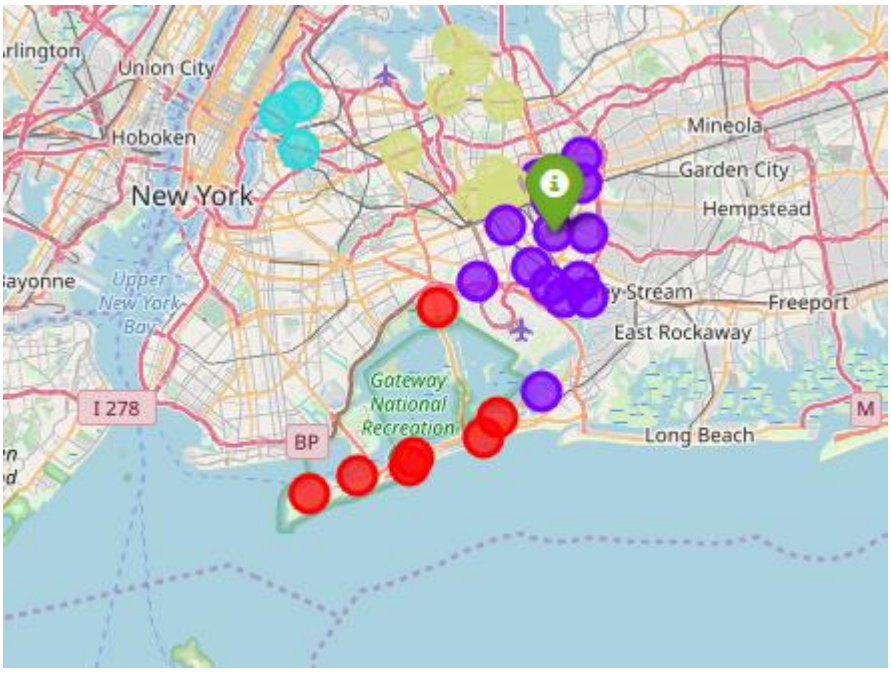
Now let's calculate the centre of this cluster which is the mean of latitude & longitude of each neighbourhoods separately and visualize the cluster along with centre.



Next we calculated the distance of each neighbourhood from centre in Queens' cluster 1. We sort this cluster information by distance in ascending order. So we get the first row as shortest distance neighbourhood from centre.

Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Pizza Place	Cluster Labels	Distance
Queens	St. Albans	40.694445	-73.758676	0	1	0.714652

Visualizing that optimal neighbourhood again:



Results and Discussion

As our analysis and graphical visualization shows that there are number of different venues in each borough. But we have seen that each borough has mostly crowded with pizza places. And in each borough, there are neighbourhoods with pizza place and non-pizza place. Their ratio are mostly fair in each borough but geographical location is different. In some borough non pizza places are surrounded with pizza places and in some borough they are isolated. So our attention was focused on such areas which are isolated from pizza places and covers major neighbourhood of non-pizza places.

Focusing on the objective we first group the data by borough and find out the number of occurrence of each venue types. And then sort the venue type by its number of occurrence which shows the top most common venue which is pizza place.

We again sort the data by borough & neighbourhood combined to see that which neighbourhood has top most venue in each borough. We filter these neighbourhood which has no pizza place.

These location were then clustered to create zones of interest which contain large number of neighbourhood not already crowded with pizza place.

Result of all this shows that there are two clusters which can be considered for optimal location for pizza place. One cluster is from Queens's borough and other is from Staten Island. Both clusters fairly contain large number of neighbourhoods. But in Staten Island cluster is already surrounded by pizza places and these places are more close to non-pizza places. While in Queens' cluster pizza places are far as compared to Staten Island. So we are considering Queens' cluster to find optimal location. For optimal location we first calculate the centre of that cluster and then calculate the distance of each neighbourhood in that cluster from centre. We sort the distance to find the closest location from centre which was **St. Albans** in **Queens borough**. This recommended location should therefore be considered only as a starting point. There may be other factors taken into account and other conditions may be introduced for optimal location.

Conclusion

Purpose of this capstone project was to identify an optimal location that is not crowded by most common venue in New York City. It gives stakeholders as a starting point for consideration. We calculated the number of occurrence of each venues in each borough which gives an idea about most common or popular venue in each borough. We found that pizza place is most common among all borough. Then clustering of those neighbourhoods where there were no pizza place gave us a zone of interest that meet some basic conditions.

We considered the cluster of neighbourhoods which was not already crowded/surrounded by pizza places.

Final decision on optimal location will be made by stakeholders based on specific characteristics of neighbourhoods and locations in every recommended cluster of neighbourhood, taking other factors and conditions into consideration.

.