The Battle of Neighborhoods

Sanjiv Tomar

November 8, 2020

1. Introduction

1.1. Background

New York is the most populous city in the United States and quite famous for the business and tourism. People do visit NY from different part of the world and enjoy different types of cuisine while their stay in New York. In order to offer different varieties of food, Mr. X is looking for to open a 'Caribbean Restaurant' in New York. There are 5 boroughs in NY, Mr. X is unsure as where to open the restaurant. His requirements for opening the restaurant are:

- As part of the business strategy, the area should be famous for different varieties of food.
- There should be more Caribbean Restaurants in the neighborhood so that Mr. X can give good competition to other restaurants.

The goal of the analytic solution is to explore all 5 boroughs in New York and help Mr. X in finding the most appropriate place to open the 'Caribbean Restaurant' which should meet the above conditions.

1.2. Problem

Data that might contribute to determining venue for Caribbean Restaurant might include attributes such as geospatial, venue categories, popularity etc. This project aims to identify the optimal place for opening the Caribbean Restaurant in New York.

2. Data acquisition and cleaning

2.1. Data Sources

2.1.1. Foursquare API: Foursquare location data is the key data requirement for solving this problem as this will help in exploring different boroughs in New York and understanding people choices of different varieties of food.

2.1.2. NY Boroughs and Zip Codes: The Zip codes for all 5 boroughs will be scraped from NYC health website: https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm. This will be the base data for exploring each ZIP code of all boroughs.

2.1.3. Geospatial data: Geographical coordinates (latitude & longitude) to be pulled from publicly shared data available at http://cocl.us/Geospatial_data

2.2. Data Cleaning

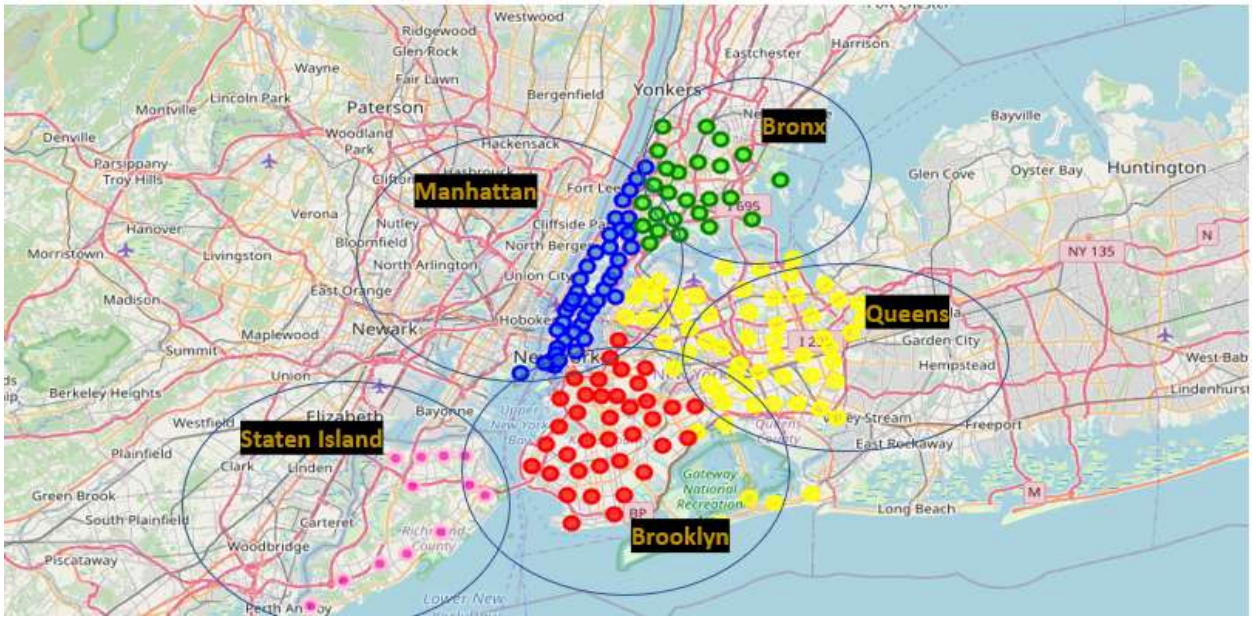
First, I pulled Borough, Neighborhood and Zip attributes of New York city from New York public health website. The data was not in very good shape as the contents were mixed amongst all three attributes. As part of data cleaning, I iterate over to each row in data frame and move the data to the right position and updated the correct data type such as Zip code should be integer instead of float.

Geospatial data was pulled from public open datasoft website and merged it with boroughs data using the Zip code. All NaN rows were identified and remove from the dataframe.

3. Exploratory Data Analysis/Modeling

The starting point of the analysis was to first visualize all 5 boroughs (Manhattan, Brooklyn, Queens, Bronx, Staten Island) on a map. This is to help identifying the location and density of neighborhoods for each borough. Compare to all other boroughs, it was seen that Staten Island has less density and since it's an island then might not be to good choice for opening the restaurant. So, I ruled out Staten Island in further analysis.

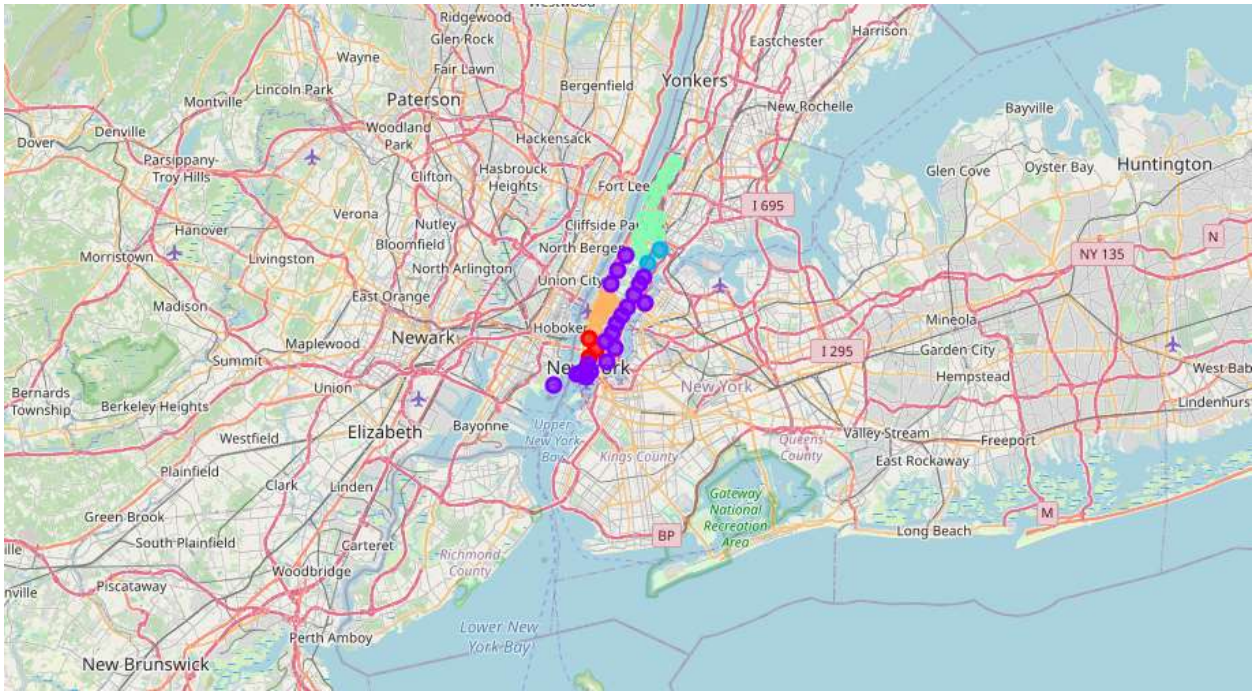
Visualizing all 5 boroughs of New York City:



In the next step of our analysis, I had extracted all the venues corresponding to Zip code then I defined the venue category and analyzed the outcome by top venues by popularity of each borough. next, I segmented the venues in cluster using k-Means and explored the outcome using Folium map.

Here is the output of K-Means clusters of each borough:

Manhattan:

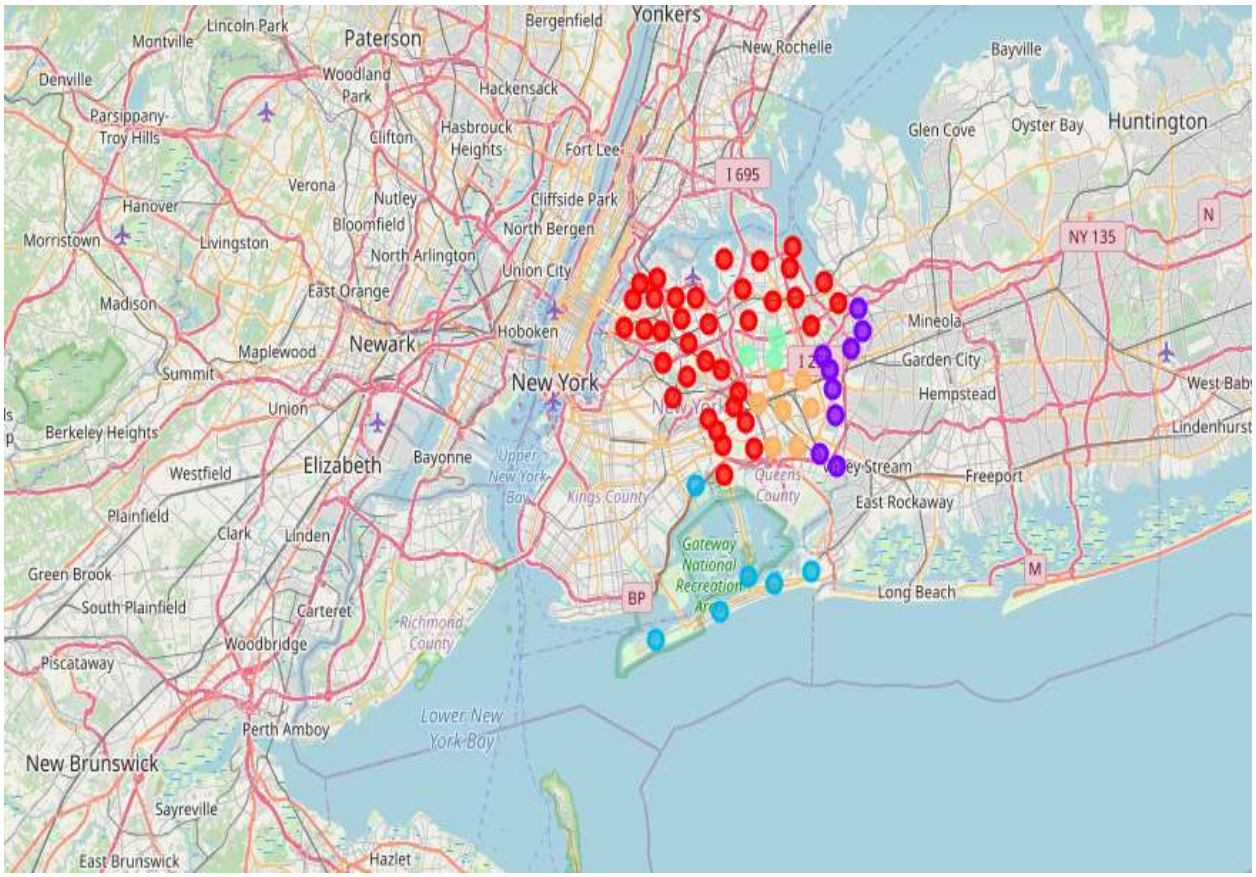


In Manhattan map, I learned about following clusters type:

- Cluster 1/Red: Italian, American restaurants and coffee shops.
- Cluster2/Purple: Park, gym, coffee shops and wine bars.
- Cluster 3/Blue: Mexican restaurants, pharmacy, super markets etc
- Cluster 4/Light blue: Chinese restaurants, Coffee Shops, Grocery Stores etc
- Cluster 5/Orange: Theater, Coffee Shops, Hotel etc

As you must have noticed, there is no Caribbean Restaurant exists in Manhattan, so it might not be a good choice to consider this borough as an option because it may take good amount of time to establish the restaurant in absence of other Caribbean restaurants. So we will rule out Manhattan from our analysis.

Queens:

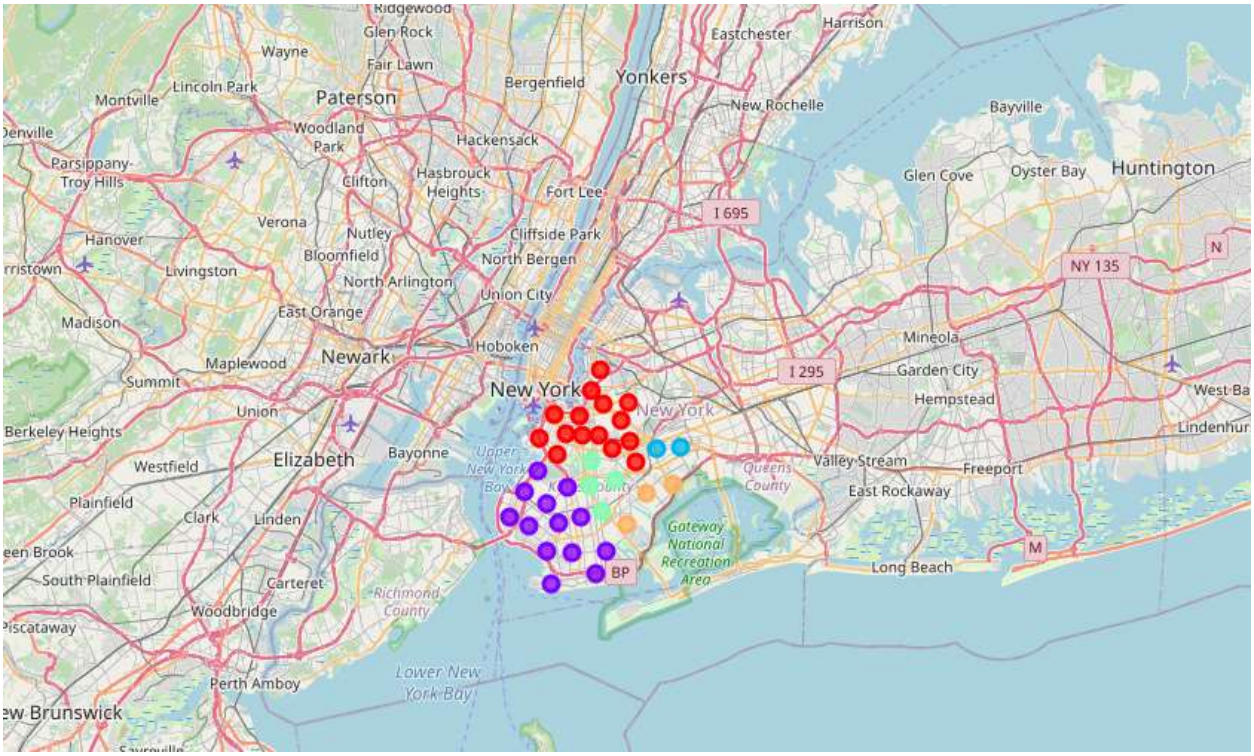


Below is the cluster output I got for Queens borough:

- Cluster 1/Red: Chinese & Korean Restaurants, Pizza/Bakery Shops, Pharmacy, Supermarket etc
- Cluster 2/Purple: Caribbean Restaurant, Fast Food Restaurant, Sandwich/Pizza Place Pharmacy etc
- Cluster 3/Blue: Beach, Deli/Bodega, Department/Discount Stores etc
- Cluster 4/Light blue: Sandwich/Pizza place, Ice Cream Shops, Bank etc
- Cluster 5/Orange: Chinese Restaurants, Donut Shops, Bus Stations etc

The second cluster identified Caribbean Restaurant which make Queens a candidate for further analysis and comparison with other venues.

Brooklyn:

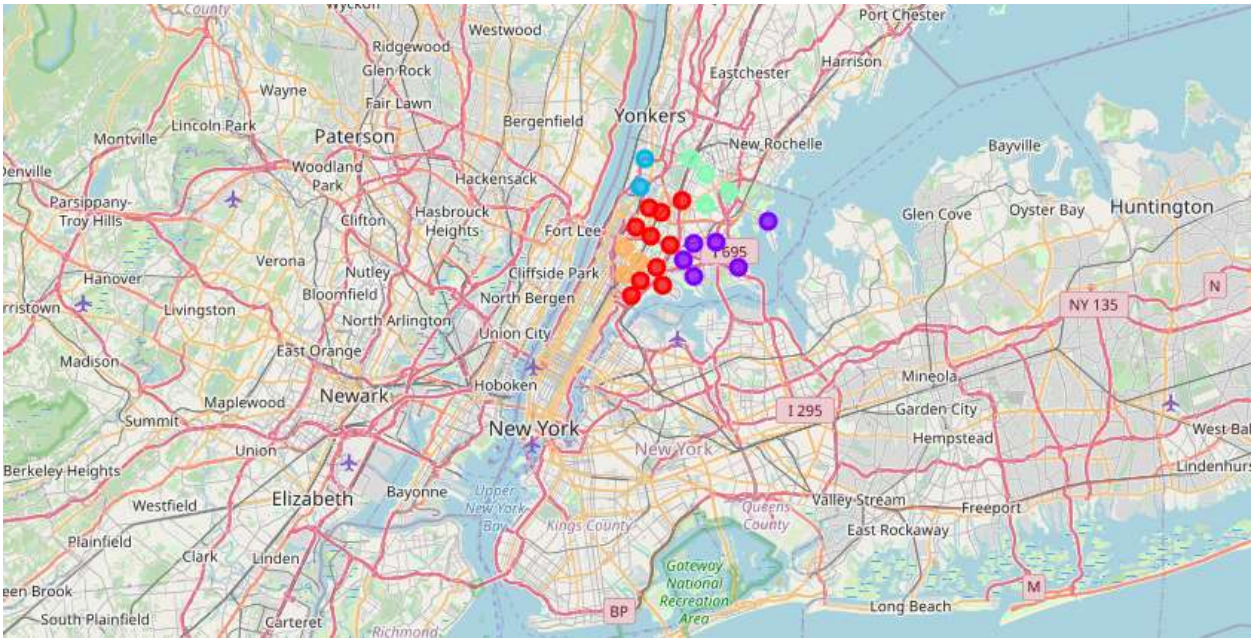


Below is the cluster output I got of Brooklyn:

- Cluster 1/Red: Bar, Caribbean Restaurant, Coffee Shop, Chicken Point, Mexican Restaurant etc
- Cluster 2/Purple: Pizza Place, Italian/Chinese Restaurant, Pharmacy, Bakery, Mobile Shops etc
- Cluster 3/Blue: Pizza Place, Supermarket, Discount Store, Playground etc
- Cluster 4/Light blue: Caribbean Restaurant, Chinese Restaurant, Bank, Mobile Shops etc
- Cluster 5/Orange: Chinese Restaurant, Pizza Place, Italian/American Restaurant, Super Market, Pharmacy etc

Similar to Queens, Brooklyn also has Caribbean Restaurants so it will be interesting to do the comparison later.

Bronx:

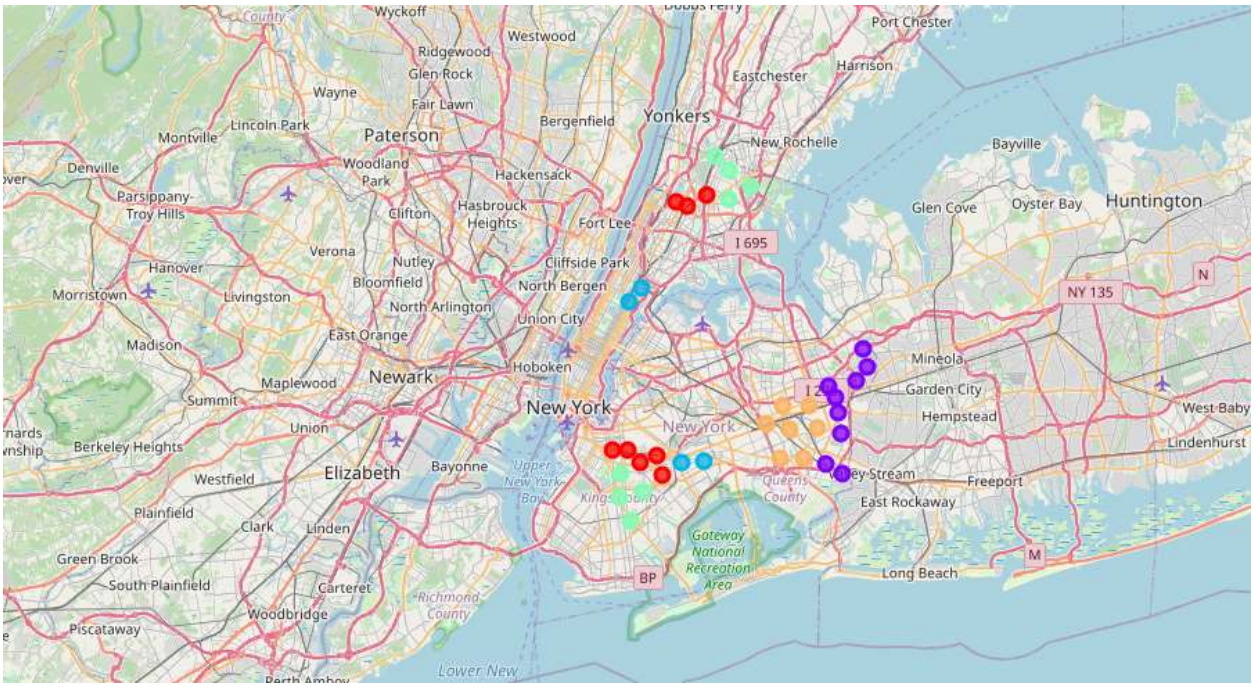


Below is the cluster output I got of Bronx:

- Cluster 1/Red: Pizza Place, Donut/Coffee Shop, Pharmacy, Latin/Chinese Restaurant, Bank, Caribbean/Chinese Restaurant etc
- Cluster 2/Purple: Latin American Restaurant, Bar, Pharmacy, Pizza/Donut Shop, Park etc
- Cluster 3/Blue: Pizza Place, Gym, Mexican/Spanish Restaurant etc
- Cluster 4/Light blue: Pizza Place, Caribbean Restaurant, Pharmacy, Chinese Restaurant, Gas Station etc
- Cluster 5/Orange: Fast Food Restaurant, Ice Cream Shop, Pizza Shop, Bar etc

Bronx also shows Caribbean Restaurant in the neighborhood.

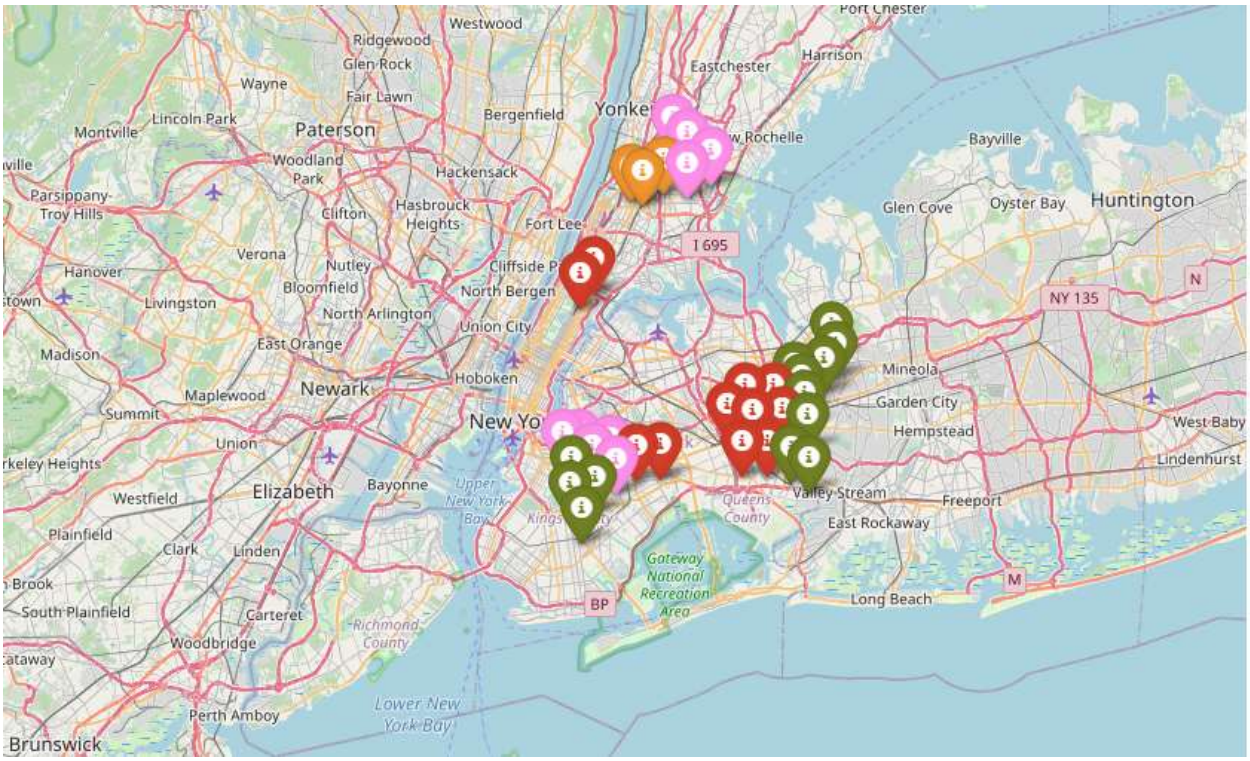
Now since I got the good understanding about types venues in each borough using k-means cluster analysis and gained understanding about Caribbean Restaurant location. Next step is to visualize the location of all Caribbean Restaurants on a single map.



As we can see Caribbean Restaurants are in Brooklyn, Queens and Bronx but not in Manhattan. Now as a next step, we would like to see the popularity/rank of venues where Caribbean Restaurants are located.

I sorted the data frame by most common venues and updated the ‘Rank’ field corresponding to popularity. I also defined color marker for each rank individually.

Post updating the Rank and color I looked at marker for each rank on the map. Here is the output:



In above map,

As per above map:

- Icon with Green color are most popular ones where Caribbean Restaurant is located
- Pink is the Second Rank and Orange is Third Rank
- The red icon indicates fourth rank which should be ruled out due to less popularity

Also, as you can see the green icon markers are shown in two different locations i.e. Queens and Brooklyn. Queens has more Caribbean Restaurants compared to Brooklyn.

Per assumption and past learning, high number of restaurants in similar category always give stiff competition. If we apply similar understanding to Queens and Brooklyn then we should consider Brooklyn over Queens because of less number of restaurants. So my recommendation will go with Flatbush neighborhood in Brooklyn borough for opening the Caribbean restaurant.

Here is the row level detail about Caribbean Restaurants located in Queens and Brooklyn.

Borough	Neighborhood	Cluster Labels	Latitude	Longitude	Zip
Queens	Southeast Queens	1	40.742944	-73.70956	11004
Queens	Southeast Queens	1	40.756983	-73.71480	11005
Queens	Southeast Queens	1	40.693538	-73.73574	11411
Queens	Southeast Queens	1	40.670138	-73.75141	11413
Queens	Southeast Queens	1	40.662538	-73.73514	11422
Queens	Southeast Queens	1	40.732239	-73.72108	11426
Queens	Southeast Queens	1	40.728235	-73.74782	11427
Queens	Southeast Queens	1	40.719981	-73.74127	11428
Queens	Southeast Queens	1	40.708833	-73.73903	11429
Brooklyn	Flatbush	3	40.649059	-73.93304	11203
Brooklyn	Flatbush	3	40.627946	-73.94552	11210
Brooklyn	Flatbush	3	40.662892	-73.95509	11225
Brooklyn	Flatbush	3	40.645256	-73.95553	11226

4. Conclusion

Purpose of this project was to identify the best space for opening the Caribbean Restaurant amongst 5 boroughs i.e. Manhattan, Brooklyn, Queens, Bronx and Staten Island. When we plotted the neighborhoods of all 5 boroughs in a single map, we found that Staten Island has relatively low population and less density so that is not the optimal choice. So we ruled out Staten Island from our analysis and focused on remaining 4 boroughs. We identified the top 10 venues of each borough and segmented them using K-Means clustering. The clustering output helped us to segment the boroughs by common venue categories. We identified the venues where Caribbean Restaurants located and analyzed them using Folium map. We finally identified Southeast Queens and Flatbush in Brooklyn where Caribbean Restaurants are popular however Queens has relatively large number of restaurants which may give stiff competition so Flatbush-Brooklyn is the optimal choice.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc