Report on 'The Battle of Neighborhoods'

1.Introduction and Business Problem

The City of New York, is the most populous city in the United States. It is diverse and is the financial capital of USA. It is multicultural. It provides lot of business oppourtunities and business friendly environment. It has attracted many different players into the market. It is a global hub of business and commerce. The city is a major center for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, theater, fashion, and the arts in the United States.

This also means that the market is highly competitive. As it is highly developed city so cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analysed carefully. The insights derived from analysis will give good understanding of the business environment which help in strategically targeting the market. This will help in reduction of risk. And the Return on Investment will be reasonable.

So there is a ABC company in New York who has many restruants in other cities of United States,has assigned me as Jr. Data Scientist to find a best locations in New York to open a new restruants and what cuisine people in that area prefer for better service.

Business Problem:

As the competition is high in the market it is very important to startegically plan. Various factors need to be studied inorder to decide on the Location such as :

1. New York Population
2. New York City Demographics
3. Are there any Farmers Markets, Wholesale markets etc nearby so that the ingredients can be purchased fresh to maintain quality and cost?
4. Are there any venues like Gyms, Entertainmnet zones, Parks etc nearby where floating population is high etc
5. Who are the competitors in that location?
6. Cuisine served / Menu of the competitors
7. Segmentation of the Borough and so on...

2. Data

We will be analyzing New York city.In this we will require the following datasets

Data 1:

Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segement the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and logitude coordinates of each neighborhood.

This dataset exists for free on the web. Link to the dataset is : https://geo.nyu.edu/catalog/nyu_2451_34572

:

	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

Data 2:

Second data which will be used is the DOHMH Farmers Markets and Food Boxes dataset. A farmers' market is often defined as a public site used by two or more local or regional producers for the direct sale of farm products to consumers. In addition to fresh fruits and vegetables, markets may sell dairy products, fish, meat, baked goods, and other minimally processed foods.

In this we will be using the data of Farmers Markets. https://data.cityofnewyork.us/dataset/DOHMH-Farmers-Markets-and-Food Boxes/8vwk-6iz2

:

	Borough	Market Name	Street Address	Latitude	Longitude	Days of Operation	Hours of Operations	Season Dates	Accepts EBT	Open Year-Round	Stellar Cooking Demonstrations	Food Activities for Kids	Location Point
0	Brooklyn	Woodhull Hospital Youthmarket	Broadway & Flushing Ave	40.700726	-73.941932	Wednesday	9 a.m. - 2 p.m.	07/10/2019-11/27/2019	Yes	No	No	No	(40.700726, -73.941932)
1	Manhattan	Mount Sinai Hospital Greenmarket	E 99th St bet Madison & Park Aves	40.789169	-73.952743	Wednesday	8 a.m. - 5 p.m.	06/12/19-11/27/19	Yes	No	No	No	(40.789169, -73.952743)
2	Bronx	170 Farm Stand	E 170th St & Townsend Ave	40.839882	-73.916783	Wednesday	2:30 - 6:30 p.m.	07/10/2019-11/27/2019	Yes	No	No	Yes	(40.839882, -73.916783)
3	Manhattan	Greenmarket at Oculus Plaza	Church & Fulton Sts, on Oculus Plaza	40.711535	-74.010464	Tuesday	7 a.m. - 7 p.m.	07/09/2019-11/30/19	Yes	Yes	No	No	(40.711535, -74.010464)
4	Queens	Ditmars Park Youthmarket	Steinway St bet Ditmars Blvd & 23rd Ave, at Di...	40.772854	-73.906061	Saturday	8 a.m. - 3 p.m.	07/13/2019-11/23/2019	Yes	No	No	No	(40.772854, -73.906061)

Data 3:
For the below analysis we will get data from wikipedia as given below :

1. New York Population
2. New York City Demographics
3. Cuisine of New York city
   https://en.wikipedia.org/wiki/New_York_City
   https://en.wikipedia.org/wiki/Economy_of_New_York_City
   https://en.wikipedia.org/wiki/Portal:New_York_City
   https://en.wikipedia.org/wiki/Cuisine_of_New_York_City
   https://en.wikipedia.org/wiki/List_of_Michelin_starred_restaurants_in_New_York_City

Data 4:
New York city geographical coordinates data will be utilized as input for the Foursquare API, that will be leveraged to provision venues information for each neighborhood. We will use the Foursquare API to explore neighborhoods in New York City.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
1	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Sam's Pizza	40.879435	-73.905859	Pizza Place
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

3. Methodology:

Business Understanding:

Our main goal is to find an optimum location in New York for ABC company to open an restaurant.

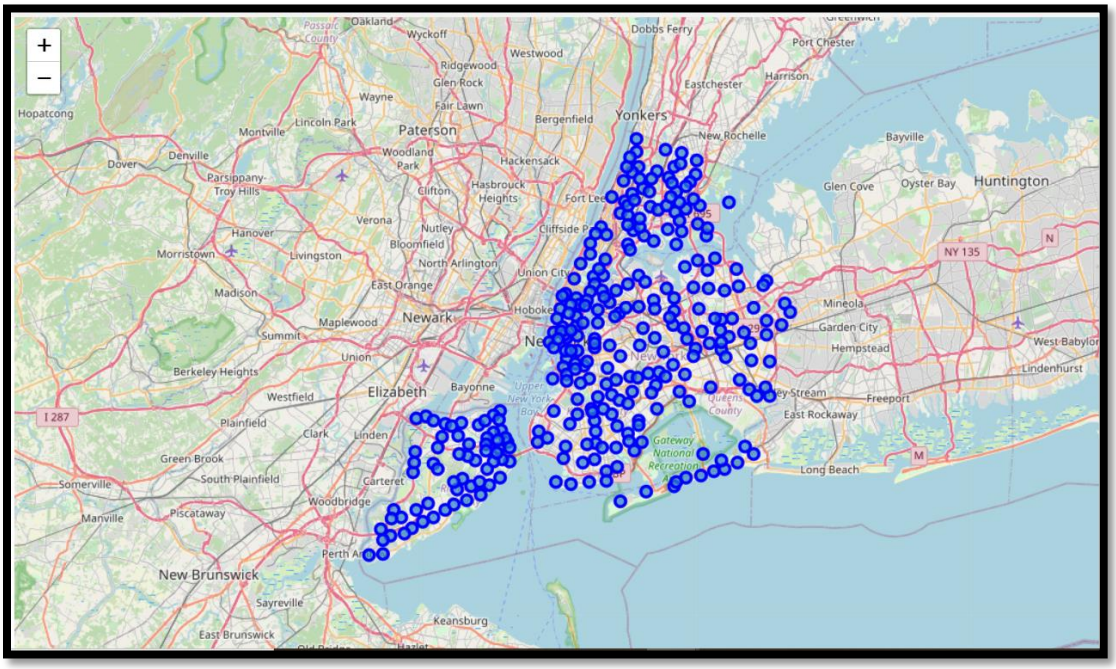
Analytical Approach:

New York city has 5 main borough and 306 neighborhoods, in this project we will have 2 parts, the 1st part will have the clustering of Manhattan and Brooklyn and 2nd part will have the clustering of Bronx, Queens and State Island.

Exploratory Data Analysis:

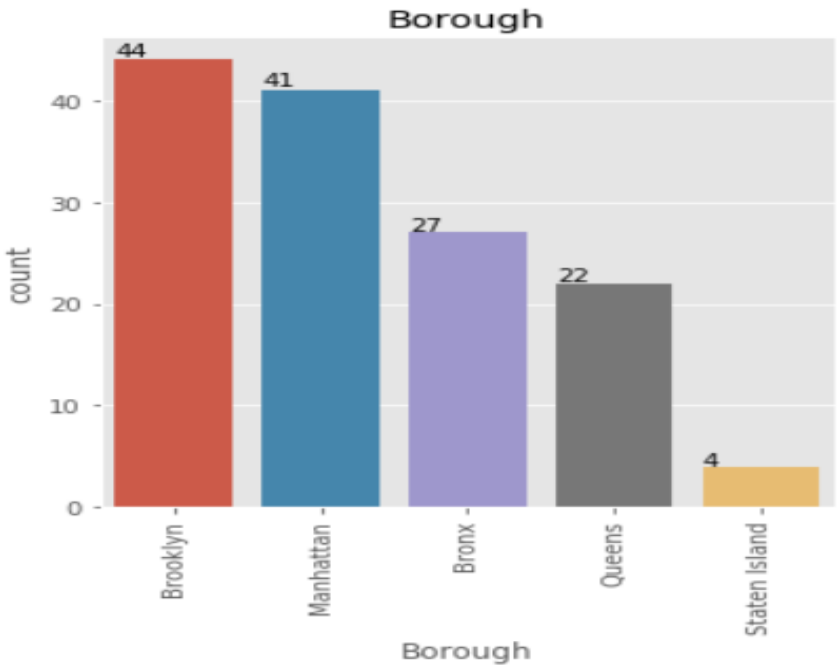
Data 1-New York city Geographical Coordinates Data

1. In this we load the data and explore data from **newyork_data.json** file.
2. Transform the data of nested python dictionaries into a pandas dataframe.
3. This dataframe contains the geographical coordinates of New York city neighborhoods.
4. This data will used to get venues data from foursquare.
5. We used geopy and folium libraries to create a map of New York city with neighborhoods superimposed on top.

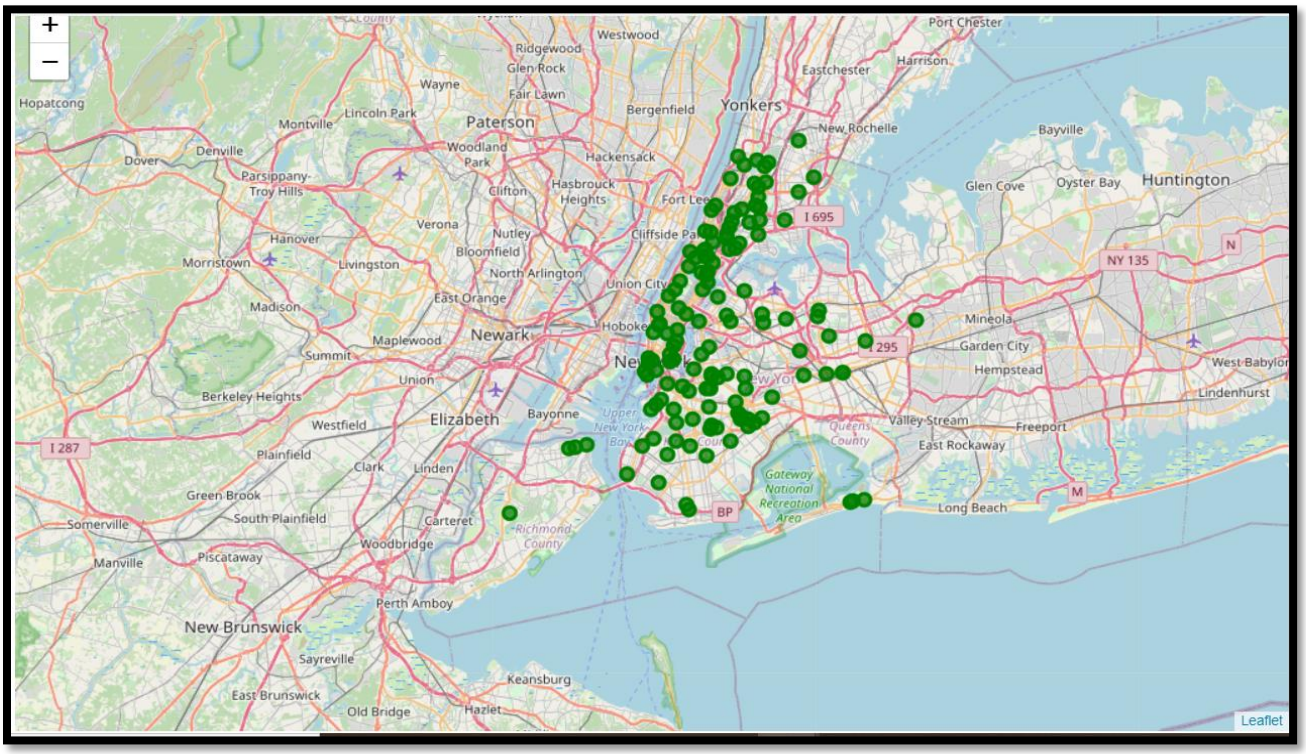


Data 2- Second data which is used is the DOHMH Farmers Markets and Food Boxes dataset. In this we will be using the data of Farmers Markets date.
There are totally 137 Farmers Market in New York city. Highest numbers are in Manhattan and Brooklyn. And lowest in Queens, Bronx and Staten Island.

Here is the bar diagram of the same:



We used geopy and folium libraries to create a map to visualize farmers markets of New York city.



Data 3: To analyize New York city population and cuisine, scrapped the data from Wikipedia pages given above in the data section. We used BeautifulSoup python library. BeautifulSoup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

1. **New York Population**: Insights from data

   - Manhattan is the geographically smallest and most densely populated borough.
   - Manhattan's population density is 368,000 per square mile.
   - Brooklyn on the western tip of long Island is the city's most populous borough.
   - Queens on Long Island north and east of Brooklyn, is geographically the largest borough.

	Borough	County	Estimate_2017	GrossDomesticProduct	square_miles	square_km	persons_sq_mi	squarekm	persons/sq.mi	persons/km2
0	The Bronx\n	\n Bronx\n	1,418,207\n	42.695\n	30,100\n	42.10\n	109.04\n	NaN	NaN	NaN
1	Brooklyn\n	\n Kings\n	2,559,903\n	91.559\n	35,800\n	70.82\n	183.42\n	NaN	NaN	NaN
2	Manhattan\n	\n New York\n	1,628,706\n	600.244\n	368,500\n	22.83\n	59.13\n	NaN	NaN	NaN
3	Queens\n	\n Queens\n	2,253,858\n	93.310\n	41,400\n	108.53\n	281.09\n	NaN	NaN	NaN
4	Staten Island\n	\n Richmond\n	476,143\n	14.514\n	30,500\n	58.37\n	151.18\n	NaN	NaN	NaN
5	City of New York	8,336,817	842.343	101,000	302.64	783.83	27,547	NaN	NaN	NaN
6	State of New York	19,453,561	1,731.910	89,000	47,214	122,284	412	NaN	NaN	NaN
7	Sources:[14] and see individual borough articl...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

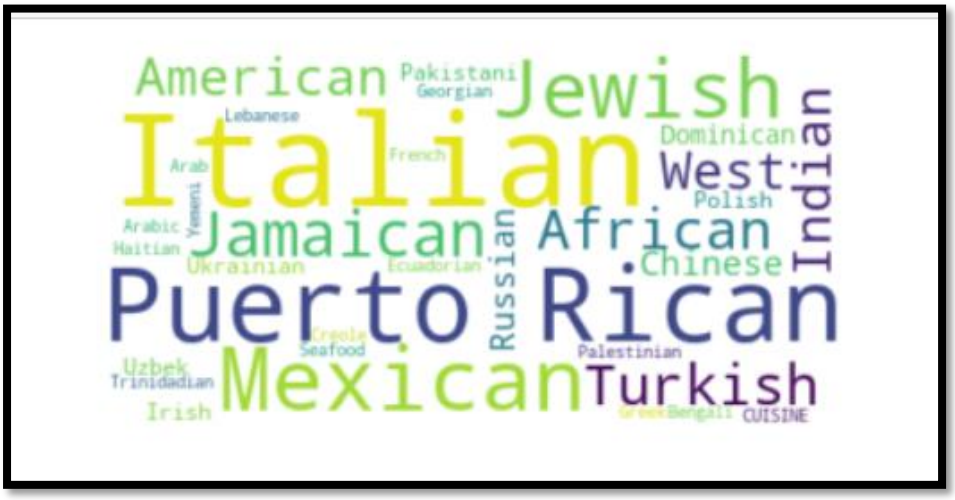
2. **Cuisine of New York city:**

   The data has been manually prepared. Data is taken from the Wikipedia page. Using this data we did word cloud.

   NEW YORK CITY CUISINE: Most preferred food in New York city is Italian, Mexican, Jewish, Indian, Puerto Rican and Irish.

BROOKLYN CUISINE:

Most preferred food in Brooklyn is Italian, Puerto Rican, Mexican and Jewish.

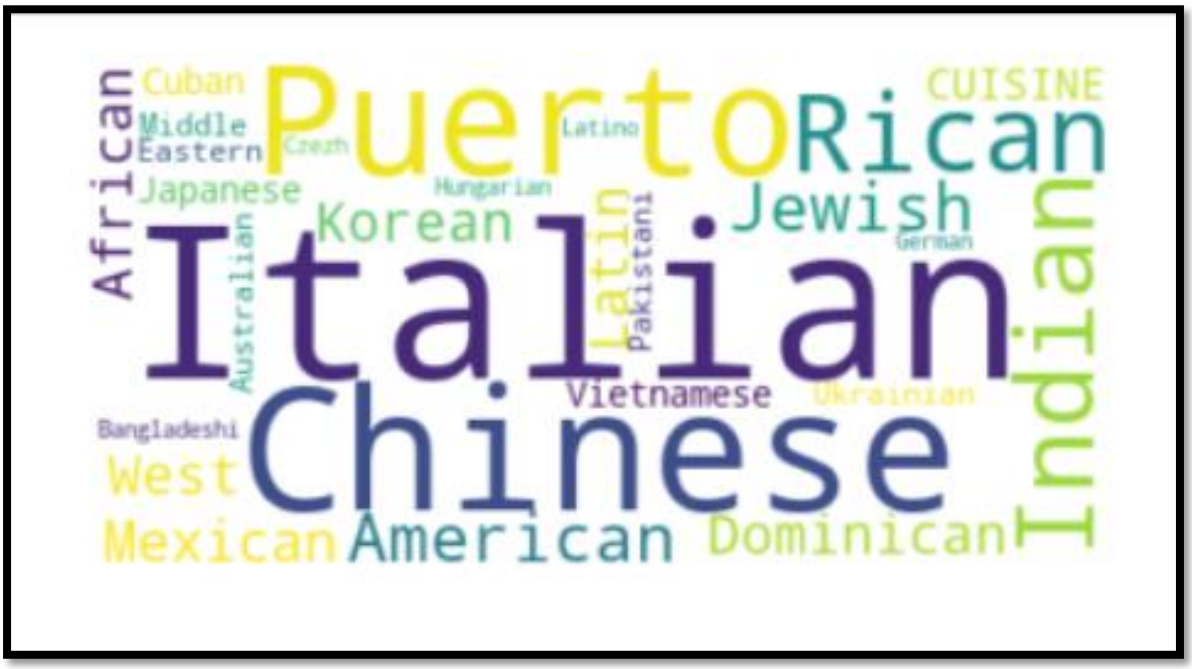


QUEENS CUISINE:

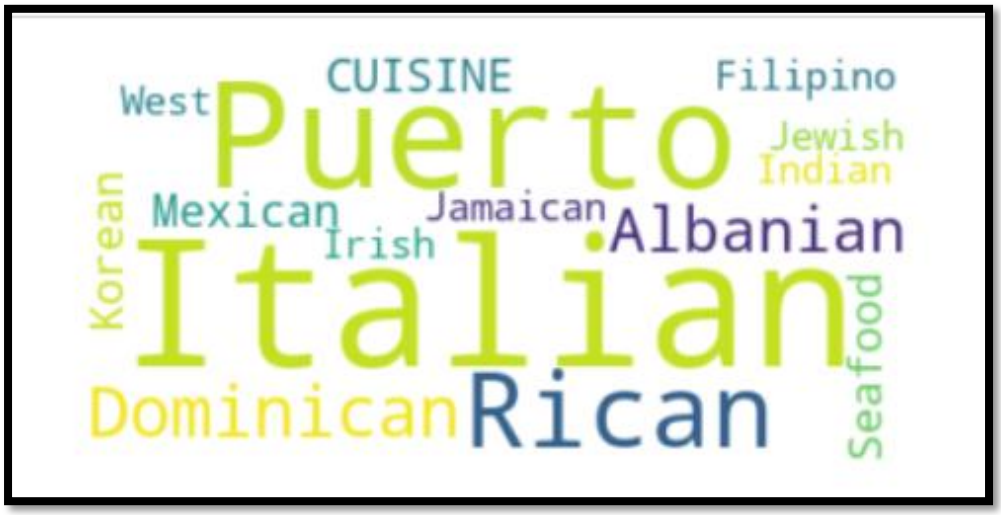
Most preferred food in Brooklyn is Italian, Irish, Mexican and Jewish.

MANHATTAN CUISINE:

Most preferred cuisine in Manhattan is Italian, Chinese, Indian, Jewish and Puerto Rican.



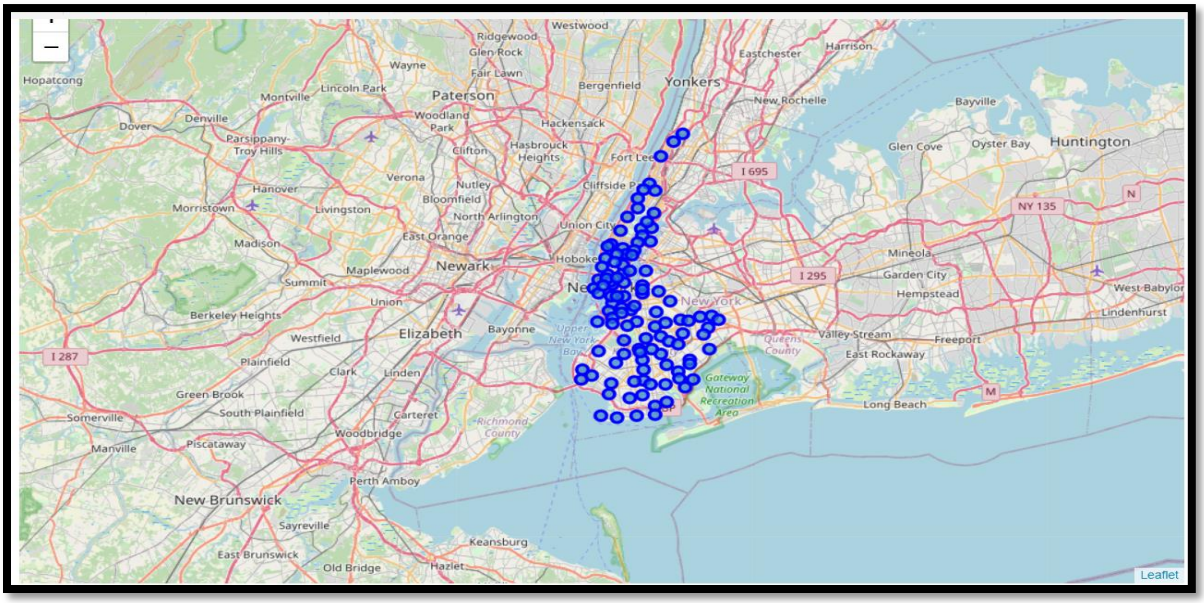
BRONX CUISINE:
Most preferred cuisine in Bronx is Italian, Dominican, Albanian and Puerto Rican.



There is very less data of cuisine relating to Staten Island so could not develop word cloud for it.

Data 4: New York city geographical coordinates data has be utilized as input for the Foursquare API, that has been leveraged to provision venues information for each neighborhood. We used the Foursquare API data to explore neighborhoods in New York city.

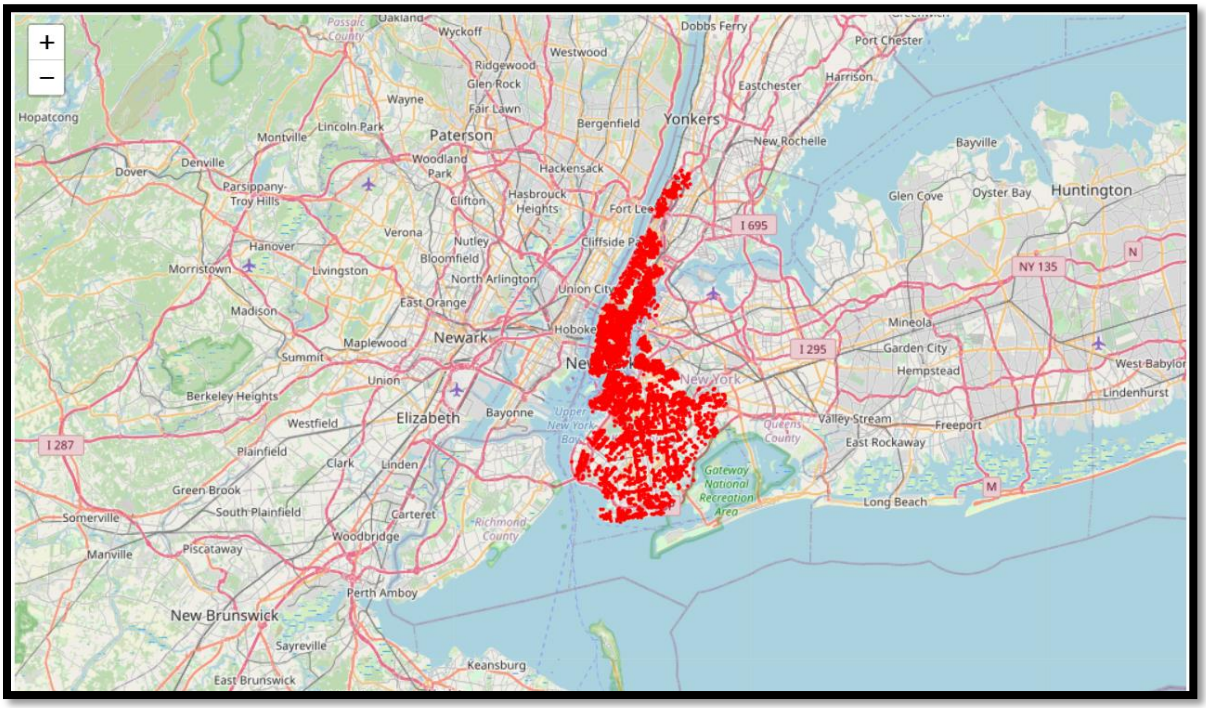
Brooklyn and Manhattan:



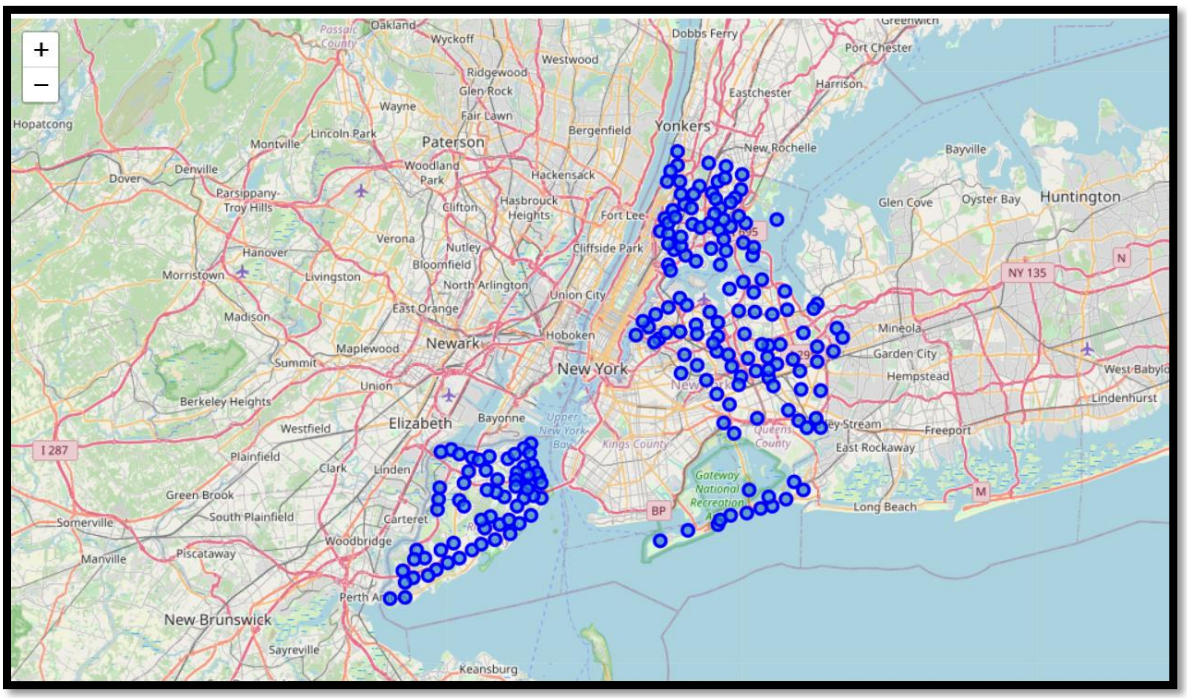
Using the geographical coordinates of each neighborhood foursquare API call are made to get top 200 venues in a radius of 1000 meters. The venues data is as follow:

	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
1	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Sam's Pizza	40.879435	-73.905859	Pizza Place
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

Brooklyn and Manhattan Venues Visualization:



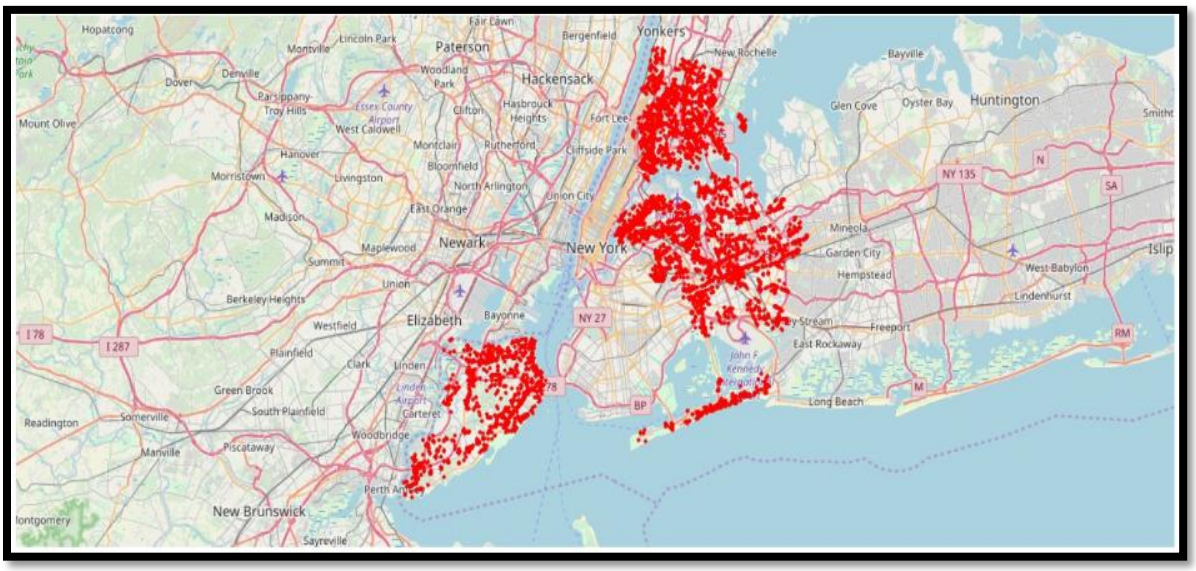
Bronx, Queens and Staten Island:



Bronx, Queens and Staten Island Venues Visualization:

	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Jackie's West Indian Bakery	40.889283	-73.843310	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Rite Aid	40.896521	-73.844680	Pharmacy

Bronx, Queens and Staten Island Venues Map Visualization:



4.Results:

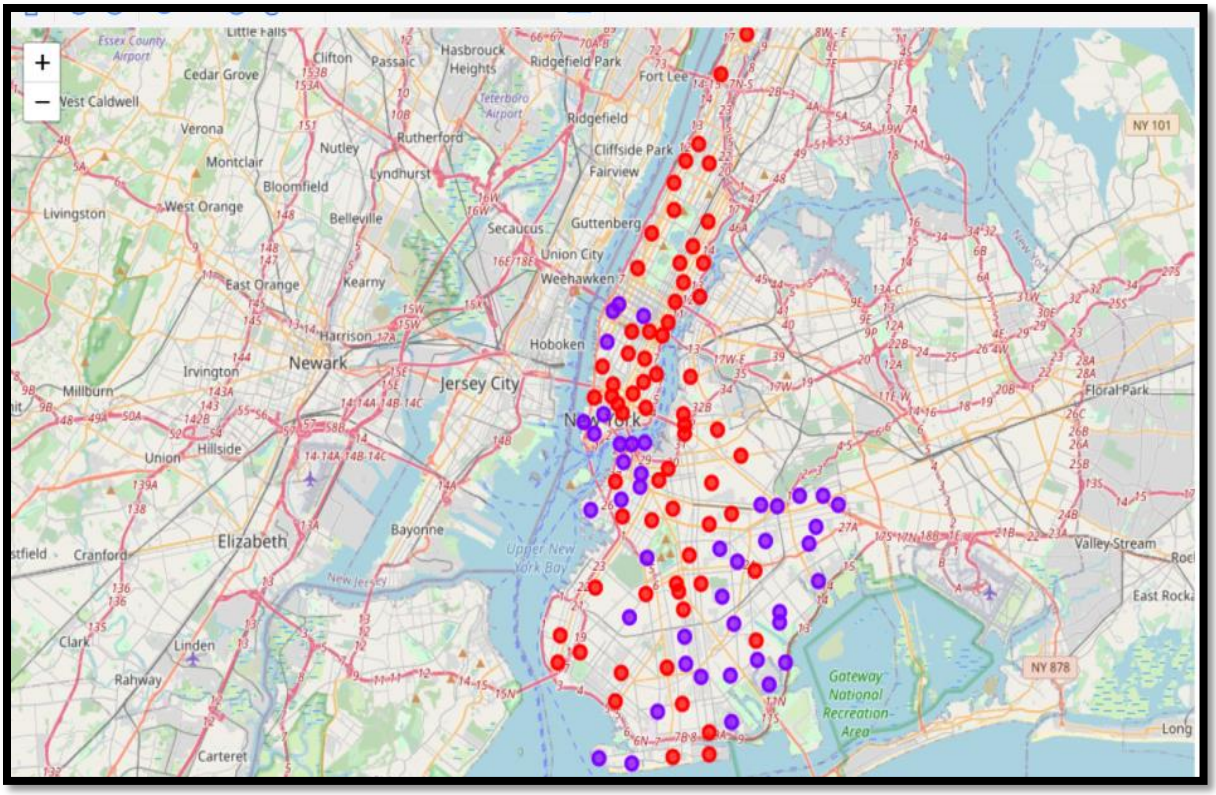
From this venues data we filtered and used only the restaurant data for Brooklyn & Manhattan clustering and Bronx, Queens and Staten Island clustering. As we have focused only on restaurant business.

Neighborhood K-Means clustering based on mean occurrence of venue category:

To cluster the neighborhoods into two clusters we used the K-mean clustering algorithm. K-means clustering aims to partition n obersvations into k clusters in which each observation belongs to the cluster with the nearest mean. It uses iterative refinement approach.

Brooklyn & Manhattan:

In the below Map visualization, we can see the different types of clusters created by using k-means for Brooklyn & Manhattan.

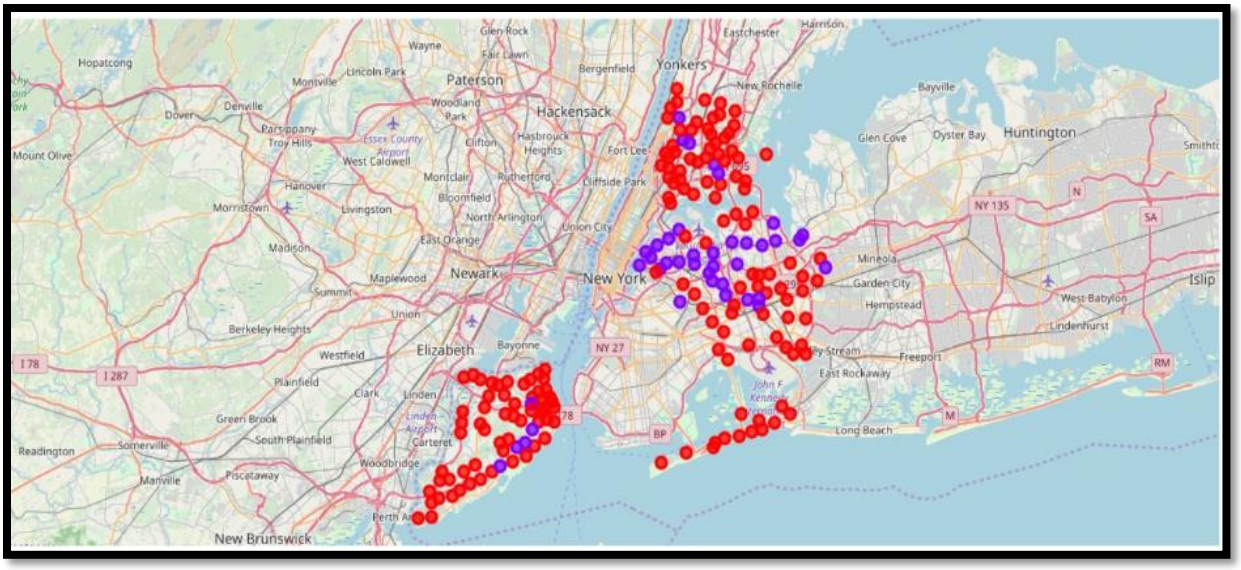


Cluster 0: the total and total sum of cluster0 has smallest value. It shows that the market is not saturated.

Cluster1: the total and total sum of cluster1 has the highest value. It shows that the markets are saturated and number of restaurants are very high.

Bronx, Queens and Staten Island:

In the below Map visualization, we can see the different types of clusters created by using k-means for Bronx, Queens and Staten Island.



Cluster0: the total and total sum of cluster0 has smallest value. It shows that the market is not saturated. There are untapped neighborhoods.

Cluster1: the total and total sum of cluster1 has highest value. It shows that the markets are saturated. Numbers of restaurants are high.

5. Discussion:

1. The farmers market can be started in Bronx, Queens and Staten Island.
2. The Italian and Puerto Rican restaurants can be started in Bronx, Queens and Staten Island as there numbers of restaurants are less.
3. If you are risk taker then you can start a restaurant in Brooklyn and Manhattan of cuisine like Italian, Irish, Jewish. The risk is high as there are many restaurants in the same area.

6. Conclusion:

This analysis is performed on limited data. This may be right or maybe not. But if good data is available there is scope to come up with much better result. If there are lots of restaurants then probably there is lot of demand and risk can be taken by open a restaurant in an area like Brooklyn and Manhattan.
As per neighborhood and demand Italian restaurant can be much preferred.