The Battle of the Neighborhoods – Final Report

Introduction

The City of New York, usually called either New York City (NYC) or simply New York (NY), is the most populous city in the United States and thus also in the state of New York. With an estimated 2017 population of 8,622,698 distributed over a land area of about 302.6 square miles (784 km^2), New York is also the most densely populated major city in the United States. Located at the southern tip of the state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass and one of the world's most populous megacities, with an estimated 20,320,876 people in its 2017 Metropolitan Statistical Area and 23,876,155 residents in its Combined Statistical Area. A global power city, New York City has been described as the cultural, financial, and media capital of the world, and exerts a significant impact upon gastronomy, commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. The city provides lot of business opportunities and business friendly environment. It has attracted many different players into the market. New York has emerged as a global node of creativity and and entrepreneurship, social tolerance, and environmental sustainability, and as a symbol of freedom and cultural diversity.

The City of New York is famous for its excellent entertainment, restaurants and for its diversity and is one of the symbols of freedom and cultural diversity. Throughout its history, the city has been a major port of entry for immigrants into the United States and approximately 37% of the city's population is foreign born and more than half of all children are born to mothers who are immigrants.

Problem and Purpose of this study

With the all attributes already described, the city also shows a very competitive market. 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 analyzed 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. As a Brazilian, I would like to explore the city and find the best place to build a new restaurant. The West 46th Street has historically been a commercial center for Brazilians living or visiting New York City. In 1995 the city officially recognized it as "Little Brazil Street" but now, I want to find a different area where I can start a successful business.

The objective is to locate in the city of New York the best location to build a new Brazilian Restaurant, analyzing New York's boroughs and its neighborhoods, and find the choice to start to build a Brazilian restaurant. The following factors will be considered in order to choose the best location:

1. Cost
2. City demographics
3. Competitors
4. Other venues in the same location
5. Etc.

Target Audience:

This can be interest for other entrepreneurs who wants to start a similar business in New York.

Success Criteria:

This project will be considered successful if I'm able to provide a good recommendation finding the best location (borough and Neighborhood) for a new Brazilian restaurant, considering the competition, the suppliers and the population.

Data

As mentioned in the document The Battle of the Neighborhoods_Week1-a, the city of New York and its boroughs will the object of this analyzes.

New York City is often referred to collectively as the five boroughs, and in turn, there are hundreds of distinct neighborhoods throughout the boroughs, many with a definable history and character to call their own. If the boroughs were each independent city, four of the boroughs (Brooklyn, Queens, Manhattan, and The Bronx) would be among the ten most populous cities in the United States (Staten island would be ranked 37th).

Data 1: The City of New York has a total of 5 boroughs and 306 neighborhoods. To segment all this data, a dataset from NYU (https://geo.nyu.edu/catalog/nyu_2451_34572) was utilized.

	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: Since we were going to on building a new restaurant, we should also analyze the suppliers. In this case, The farmers market was also an variable utilized in this study. The dataset utilized can be found on NYC OpenData webpage : https://data.cityofnewyork.us/dataset/DOHMH-Farmers-Markets-and-Food-Boxes/8vwk-6iz2

	FacilityName	Service_Category	Service_Type	Address	Address2	Borough	ZipCode	Latitude	Longitude	StartDate	EndDate
0	1 Centre Street	Farmers Markets and Food Boxes	Food Boxes	1 Centre Street	South Building, 9th Floor	Manhattan	11101	40.713028	-74.003753	NaN	NaN
1	125th Street Farmers Market	Farmers Markets and Food Boxes	Farmers Markets	125th St & Adam Clayton Powell Jr Blvd	NaN	Manhattan	10027	40.808981	-73.948327	6/13/17	11/21/17
2	170 Farm Stand	Farmers Markets and Food Boxes	Farmers Markets	170th St & Townsend Ave	NaN	Bronx	10452	40.840095	-73.916827	7/5/17	11/22/17
3	175th Street Greenmarket	Farmers Markets and Food Boxes	Farmers Markets	175th St bet Wadsworth Ave & Broadway	NaN	Manhattan	10033	40.845956	-73.937813	6/29/17	11/30/17
4	57th Street Greenmarket	Farmers Markets and Food Boxes	Farmers Markets	57th St & 9th Ave	NaN	Manhattan	10019	40.767925	-73.985716	5/17/17	12/23/17

Data 3: Other information related to population, demographics and from the city itself, wa extracted Wikipedia: https://en.wikipedia.org/wiki/New_York_City

	NewYorkCitysfiveboroughsvte	Jurisdiction	Population	GrossDomesticProduct	Landarea	Density	persons_sq_mi	persons_sq_km
0	The Bronx\n	\n Bronx\n	1,471,160\n	28.787\n	19,570\n	42.10\n	109.04\n	34,653\n
1	Brooklyn\n	\n Kings\n	2,648,771\n	63.303\n	23,900\n	70.82\n	183.42\n	37,137\n
2	Manhattan\n	\n New York\n	1,664,727\n	629.682\n	378,250\n	22.83\n	59.13\n	72,033\n
3	Queens\n	\n Queens\n	2,358,582\n	73.842\n	31,310\n	108.53\n	281.09\n	21,460\n
4	Staten Island\n	\n Richmond\n	479,458\n	11.249\n	23,460\n	58.37\n	151.18\n	8,112\n
5	City of New York	8,622,698	806.863	93,574	302.64	783.83	28,188	10,947\n
6	State of New York	19,849,399	1,547.116	78,354	47,214	122,284	416.4	159\n
7	Sources: [3] and see individual borough articl...	NaN	NaN	NaN	NaN	NaN	NaN	NaN

	Borough	Neighborhood	Cuisine	Latitude	Longitude	Zipcodes	Unnamed: 6	Unnamed: 7	Unnamed: 8
0	The Bronx	Bedford Park	Mexican, Puerto Rican, Dominican, Korean	40.870	-73.886	10458, 10468	NaN	NaN	NaN
1	The Bronx	Belmont	Italian, Albanian	40.855	-73.886	10457,10458, 10460	NaN	NaN	NaN
2	The Bronx	City Island	Italian, Seafood	40.848	-73.786	NaN	NaN	NaN	NaN
3	The Bronx	Morris Park	Italian, Albanian	40.852	73.853	10461, 10462	NaN	NaN	NaN
4	The Bronx	Norwood	Filipino	40.878	-73.878	10467	NaN	NaN	NaN

Data 4: The Venues information from each neighborhood was extracted from foursquare.com utilizing New York city geographical coordinates.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
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	Loeser's Delicatessen	40.879242	-73.905471	Sandwich Place

This project will also several Python packages as listed below:

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- JSON – Library to handle JSON files
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests

• Matplotlib – Python Plotting Module

• Sklearn – Python machine learning Library

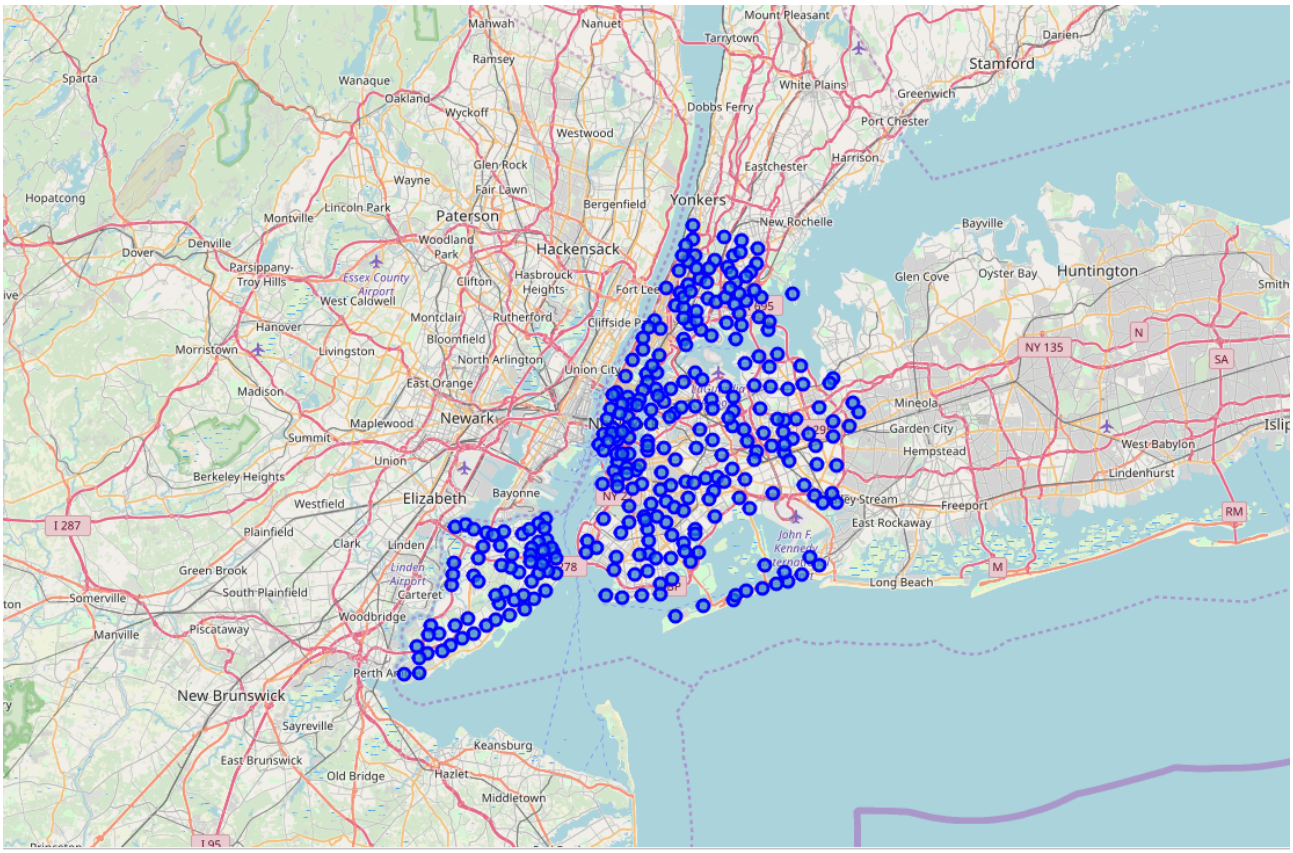
• Folium – Map rendering Library

Methodology:

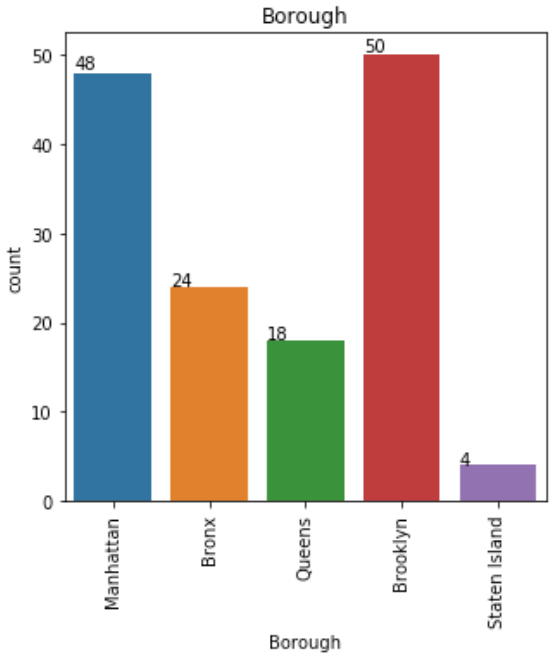
The main goal is to find an optimum location for a new Brazilian Restaurant in the city of New York and in order to accomplish it, the five boroughs and the 300 neighborhoods was analyzed. First, I cluster the boroughs of Manhattan and Brooklyn together. Manhattan and Brooklyn have the largest number of farmer markets. Then, I cluster Queens, Bronx and Staten Island.

Analyzing Data 1: New York geographical coordinates

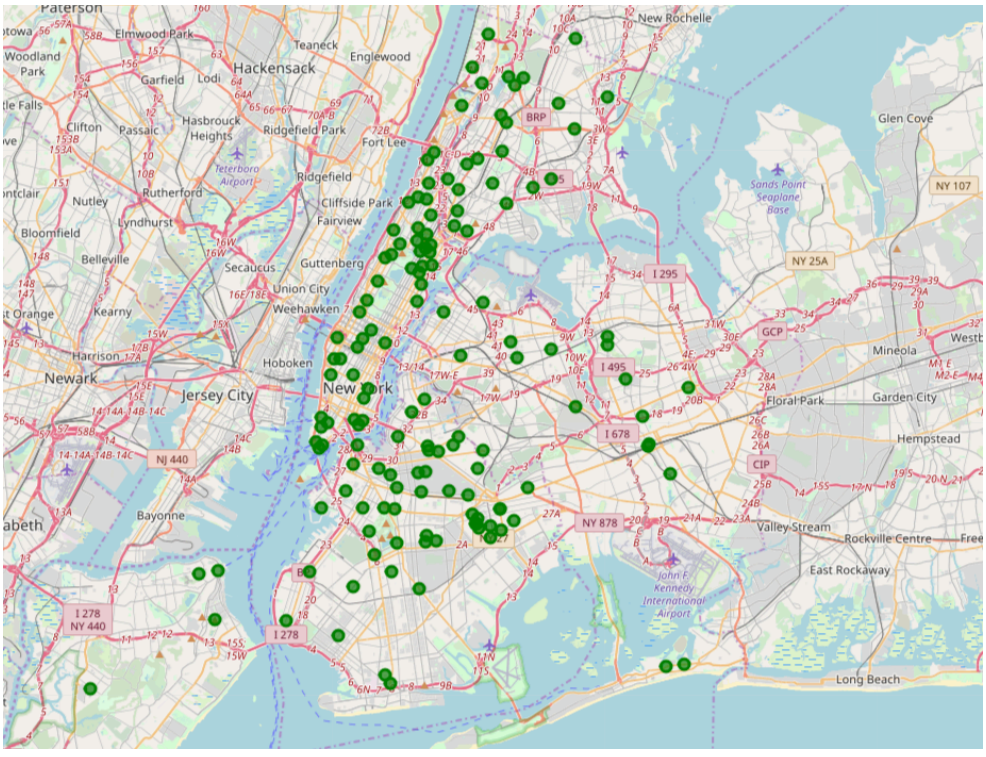
- Load and explore the data;
- Transform the data into a panda dataframe;
- Utilize the geopy and folium libraries to create a map of New York and its neighborhoods.



Analyzing Data 2: NYC OpenData webpage was utilized to collect a dataset for farmers market. A total of 144 Farmers market and 17 food boxes was found in New York city. Manhattan and Brooklyn has the highest numbers with 48 and 50.



Utilizing the geopy and folium libraries, a map was create showing this data.



Analyzing Data 3: I scrapped the from Wikipedia pages to collect the data utilized to analyze the New York City population, demographics and cuisines. With this information in hands, I utilized BeautifulSoup library. BeautifulSoup is a python package for parsin HTML and XML documents (including having malformed markup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

New York Population:

- Manhattan is the geographically smallest and most density populates borough;
- Queens is geographically the largest borough
- Staten Island has the lowest GDP and the lowest population
- Brooklyn has the second highest population between the boroughs

	Borough	County	Estimate_2017	GrossDomesticProduct	square_miles	square_km	persons_sq_mi	persons_sq_km
0	The Bronx	Bronx	1,471,160	28.787	19,570	42.10	109.04	34,653
1	Brooklyn	Kings	2,648,771	63.303	23,900	70.82	183.42	37,137
2	Manhattan	New York	1,664,727	629.682	378,250	22.83	59.13	72,033
3	Queens	Queens	2,358,582	73.842	31,310	108.53	281.09	21,460
4	Staten Island	Richmond	479,458	11.249	23,460	58.37	151.18	8,112
5		City of New York	8,622,698	93,574	806.863	302.64	783.83	28,188
6		State of New York	19,849,399	78,354	1,547.116	47,214	122,284	416.4
7	Sources: [3] and see individual borough articles							

New York demographics:

- New York city is the most populated city in the United States, with an estimated of 8,622,698 residents as of 2017, incorporating more immigration into the city than outmigration since the 2010 United States census.
- Racial composition can be seeing below

	Racial composition	2010[246]	1990[248]	1970[248]	1940[248]
0	White	44.0%	52.3%	76.6%	93.6%\n
1	—Non-Hispanic	33.3%	43.2%	62.9%[249]	92.0%\n
2	Black or African American	25.5%	28.7%	21.1%	6.1%\n
3	Hispanic or Latino (of any race)	28.6%	24.4%	16.2%[249]	1.6%\n
4	Asian	12.7%	7.0%	1.2%	–\n

New York Cuisine:

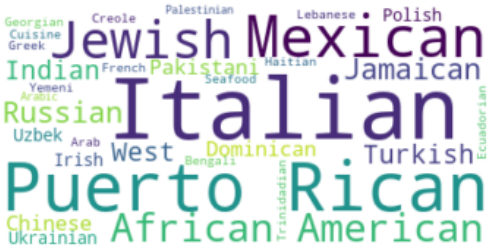
- This data was collected from Wikipedia
- Word cloud was created and the most preferred food for each borough was found.

	Borough	Neighborhood	Cuisine	Latitude	Longitude	Zipcodes
0	The Bronx	Bedford Park	Mexican, Puerto Rican, Dominican, Korean	40.870	-73.886	10458, 10468
1	The Bronx	Belmont	Italian, Albanian	40.855	-73.886	10457,10458, 10460
2	The Bronx	City Island	Italian, Seafood	40.848	-73.786	NaN
3	The Bronx	Morris Park	Italian, Albanian	40.852	73.853	10461, 10462
4	The Bronx	Norwood	Filipino	40.878	-73.878	10467

New York city most preferred food is Italian, Puerto Rican, Mexican, Jewish, Dominican.

Manhattan most preferred food is Italian, American, Puerto Rican, Indian.

Brooklyn most preferred food is Italian, Jewish, Mexican, Puerto Rica, American.



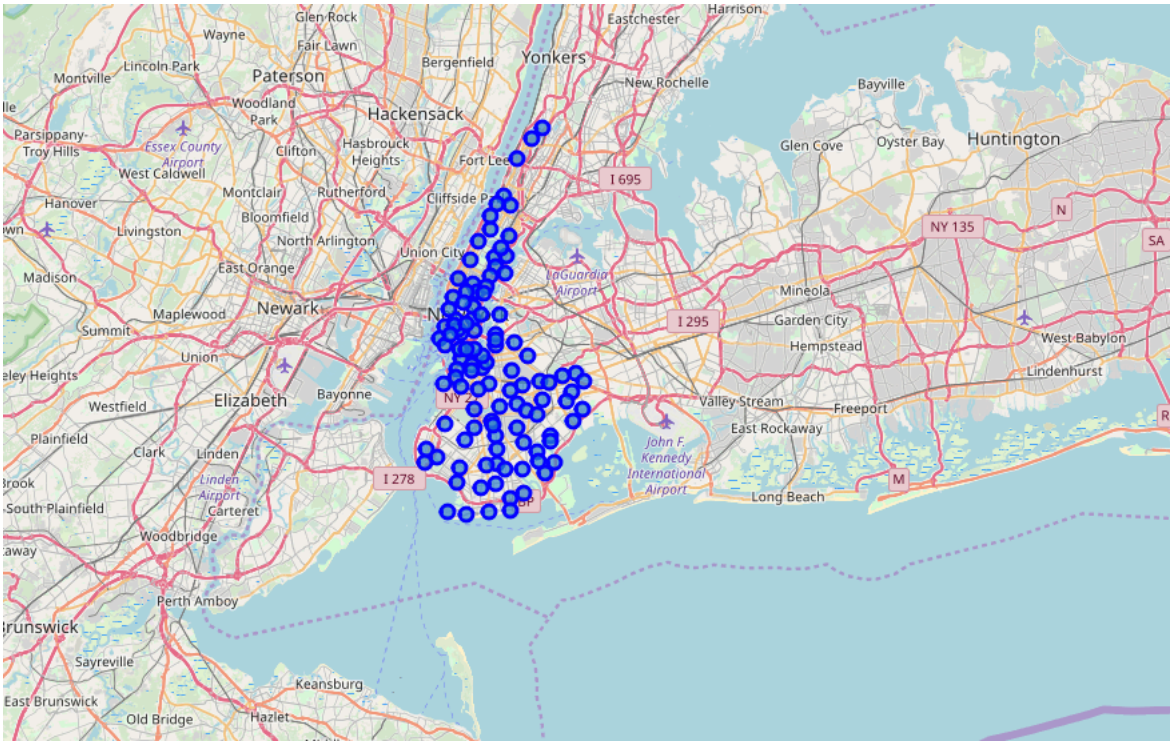
Queens most preferred food is Jewish, Italian, Indian, Irish, Mexican.

The Bronx most preferred food is Italian, Albanian, Dominican, Puerto Rican.

Staten Island most preferred food is Italian, Russian, Polish, Arab.

Analyzing Data 4: I have utilized the New York city geographical coordinates as input for Foursquare API to collect the venues information from each neighborhood. Foursquare was also utilized to explore these neighborhoods.

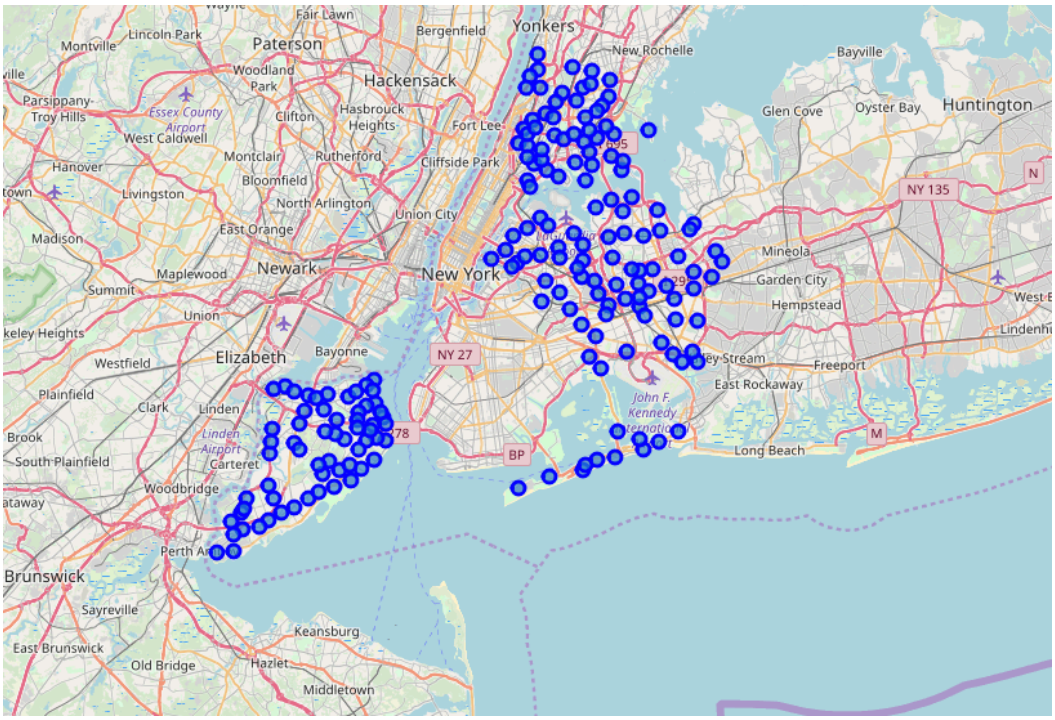
Clustering Brooklyn and Manhattan:



I utilized the geographical coordinates of each neighborhood and the Foursquare API calls were able to get 200 venues in a radius of 1000 meters. I was also able to look at the distribution of these venues. The dataframe had a total of 9679 venues, 403 different venue types and 110 neighborhoods. The sample of these venues can be seem below:

	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
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	Loeser's Delicatessen	40.879242	-73.905471	Sandwich Place

Clustering Queens, The Bronx and Staten Island:



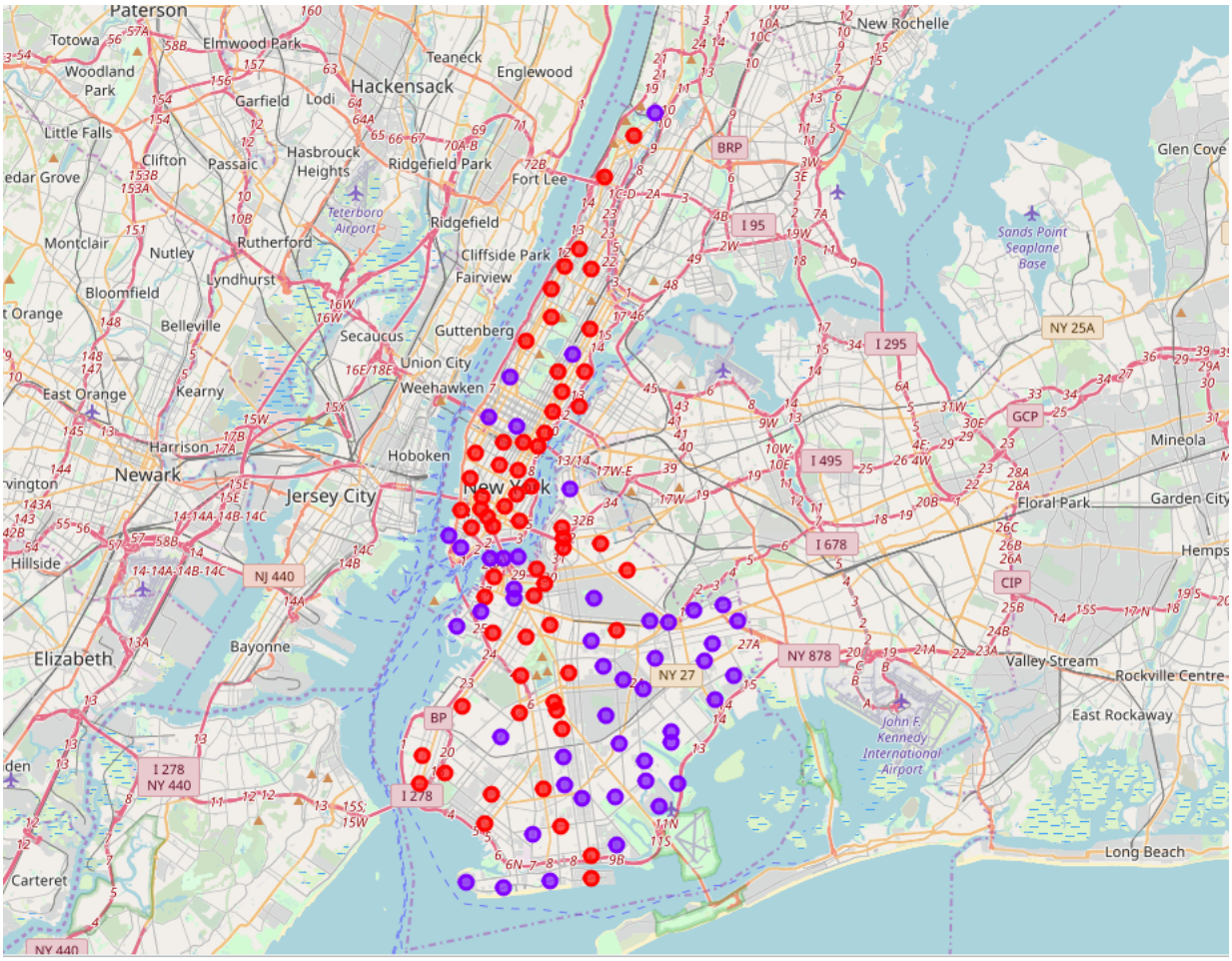
Same process was utilized for these three boroughs, the dataframe had over 10000 venues, 386 different venue types and 187 neighborhoods. The sample of these venues can be seem below:

	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	40.896649	-73.844846	Pharmacy

Outcomes

From all the venues provides by Foursquare, I filtered and analyzed only the restaurants based on the clustering I mentioned above: Manhattan and Brooklyn, and Queens, The Bronx and Staten Island.
In order to cluster the neighborhoods into two different clusters, I used k-means clustering algorithm. As we studied, *K*-means clustering is a type of unsupervised learning, it helped to find groups which have not been explicit labeled in the data. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group. It uses iterative refinement approach.

- Manhattan and Brooklyn: As we can see in the map below, the different types of clusters created by using k-means:



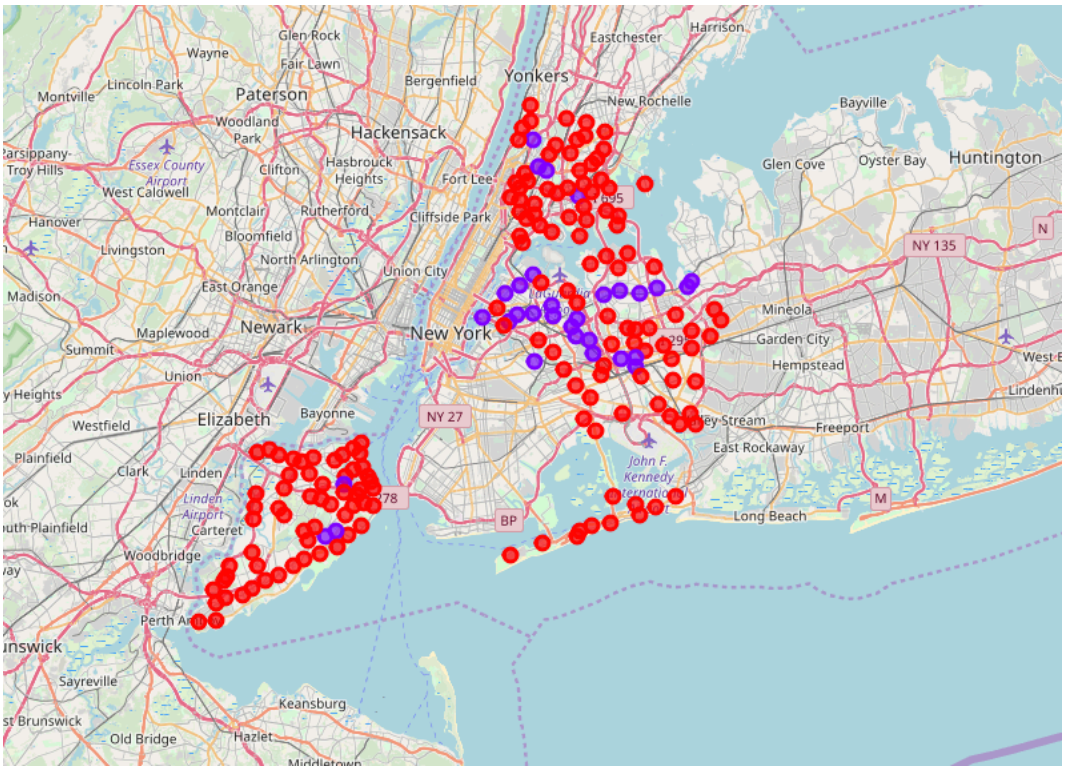
From the analyzes, I found that the Silhouette Coefficient was larger for n_clusters=2:

```
For n_clusters=2, The Silhouette Coefficient is 0.412112717046045
For n_clusters=3, The Silhouette Coefficient is 0.3122791507910412
For n_clusters=4, The Silhouette Coefficient is 0.21585547848103376
```

```
For n_clusters=5, The Silhouette Coefficient is 0.24785617739212407
For n_clusters=6, The Silhouette Coefficient is 0.23440371354029352
For n_clusters=7, The Silhouette Coefficient is 0.16607798928040174
For n_clusters=8, The Silhouette Coefficient is 0.17055248607274054
For n_clusters=9, The Silhouette Coefficient is 0.13106023041082165
```

Analyzing the clusters 0 and 1, I found that the Total and Total Sum of cluster0 had the smallest value and no saturated market. Cluster1, in the other hand, had Total and Total Sum had the highest value and saturated market, so number of restaurants here are very high. Didn't find any untapped markets.

- Queens, The Bronx and Staten Island: As we can see in the map below, the different types of clusters created by using k-means:



From the analyzes, I found that the Silhouette Coefficient was larger for n_clusters=2:

```
For n_clusters=2, The Silhouette Coefficient is 0.5678063752761374
For n_clusters=3, The Silhouette Coefficient is 0.3907046839713088
For n_clusters=4, The Silhouette Coefficient is 0.3452144862785586
For n_clusters=5, The Silhouette Coefficient is 0.3430292214931158
For n_clusters=6, The Silhouette Coefficient is 0.34463975575773415
For n_clusters=7, The Silhouette Coefficient is 0.3064770822185201
For n_clusters=8, The Silhouette Coefficient is 0.2986373492747028
For n_clusters=9, The Silhouette Coefficient is 0.29114748407468466
```

Analyzing the clusters 0 and 1, I found that the Total and Total Sum of cluster0 had the smallest value and no saturated market. Cluster1, in the other hand, had Total and Total

Sum had the highest value and saturated market, so number of restaurants here are very high. Untapped markets were found:

	Borough	Neighborhood	Latitude	Longitude	Total	Cluster_Labels
0	Bronx	Clason Point	40.806551	-73.854144	0	0
1	Staten Island	Todt Hill	40.597069	-74.111329	0	0
2	Staten Island	South Beach	40.580247	-74.079553	0	0
3	Staten Island	Port Ivory	40.639683	-74.174645	0	0
4	Staten Island	Woodrow	40.537453	-74.221351	0	0
5	Staten Island	Butler Manor	40.506082	-74.229504	0	0
6	Staten Island	Rossville	40.549404	-74.215729	0	0
7	Staten Island	Bloomfield	40.605779	-74.187256	0	0

Conclusion:

- Brooklyn and Manhattan have the higher concentration of restaurants business and is the most competitive market comparing to the other three boroughs.
- There are 112,700 hotel rooms in the city, a jump from 62,200 in 1988. Of the 80 new hotels built since 2015, two-thirds are outside Manhattan
- The Bronx, Queens and Staten Island also has a good number of restaurants but not the number required because its population. So, this are areas that can be explored.
- This dataset was used to predict an optimal place for a new investment in the food business (Brazilian Restaurant), but it can also for other businesses since the results contain a lot of useful information about other as well.