The Battle of the Neighborhoods

Eder Braz

May 25, 2020

1. Introduction:

1.1 Background

New York City (NYC) is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles (784 km^2), New York is also the most densely populated major city in the United States. New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

1.2 Problem Background

My client, a successful restaurant chain in Brazil is looking to expand in New York. They want to create a Brazilian steakhouse restaurant. However, he does not know in which neighborhood the restaurant would gain the most attention from customers and face the least competition from other steakhouses.

2. Data Description:

- New York Dataset that contains the 5 boroughs and all the neighborhoods that exists in each borough as well as the latitude and longitude coordinates of each

neighborhood: 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

- The Foursquare API will be used to explore neighborhoods in New York, more specifically, we will be using the explore function to get the most common venue categories in each neighborhood.

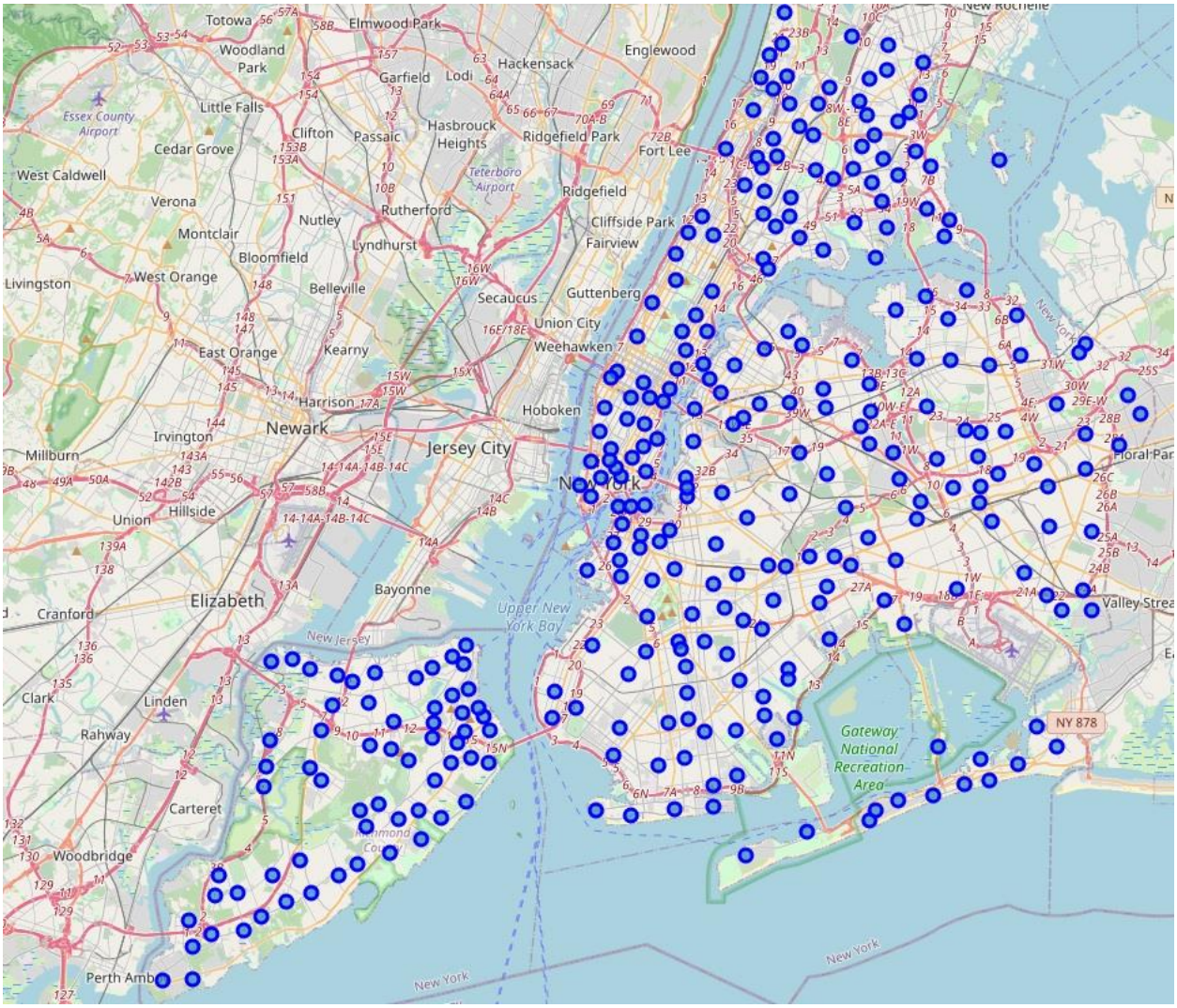
3. Methodology

3.1 Extracting Data and transforming

First, the dataset is downloaded from https://cocl.us/new_york_dataset. Then converted into a Pandas Dataframe.

	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

I used python Folium library to visualize geographic details of New York and its boroughs to create a map of New York with boroughs superimposed on top using latitude and longitude values to get the visual as below:



Utilizing the Foursquare API to explore the boroughs and segment them. I designed the limit as **100 venue** and the radius **500 meter** for each neighborhood from their given latitude and longitude information.

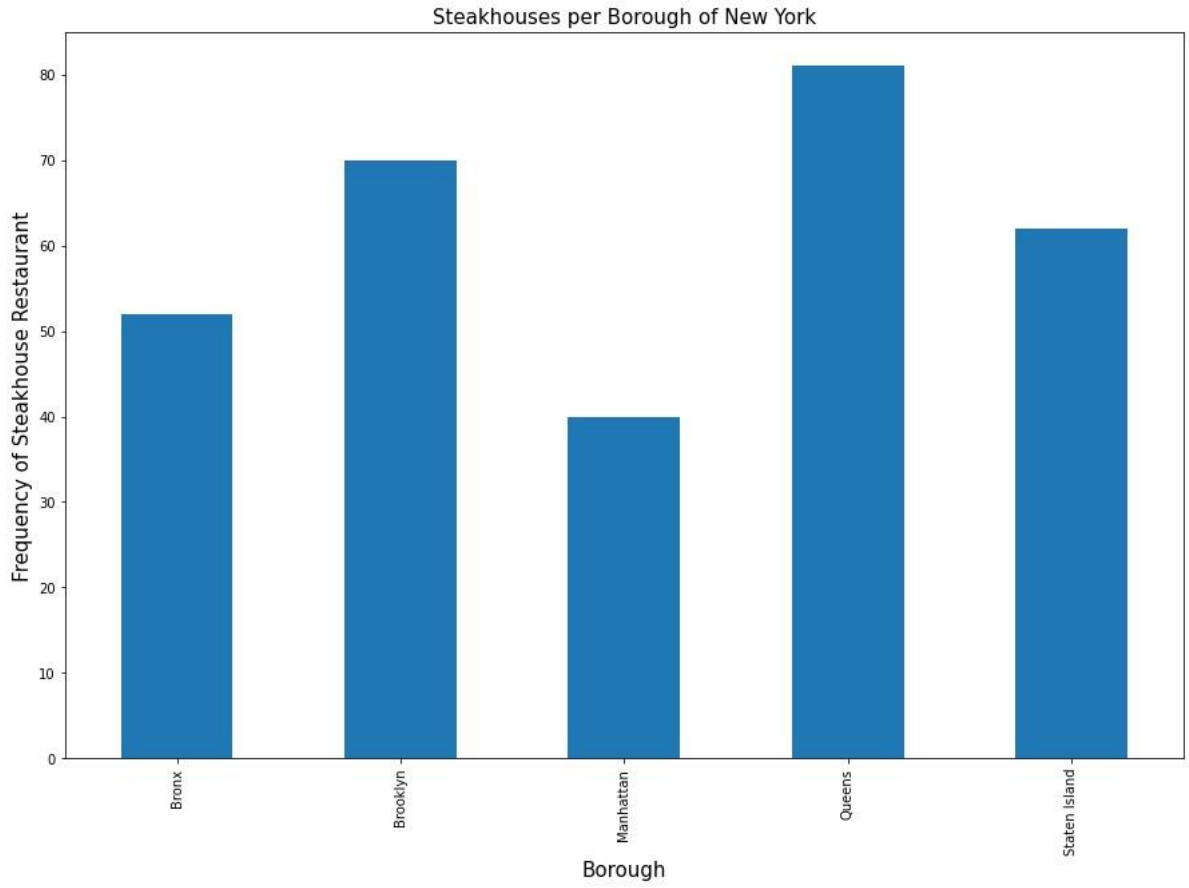
Here is the head of the list venues name, category, latitude and longitude information from Foursquare API:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

After extracting the information from FourSquare API. We are going to use the one hot encoding to transform the categorical values, group with the neighborhoods and use only the SteakHouse column.

	Borough	Neighborhood	Steakhouse	Latitude	Longitude
0	Bronx	Wakefield	0.0	40.894705	-73.847201
1	Bronx	Co-op City	0.0	40.874294	-73.829939
2	Bronx	Eastchester	0.0	40.887556	-73.827806
3	Bronx	Fieldston	0.0	40.895437	-73.905643
4	Bronx	Riverdale	0.0	40.890834	-73.912585

Let us see how many steakhouses each Borough have:



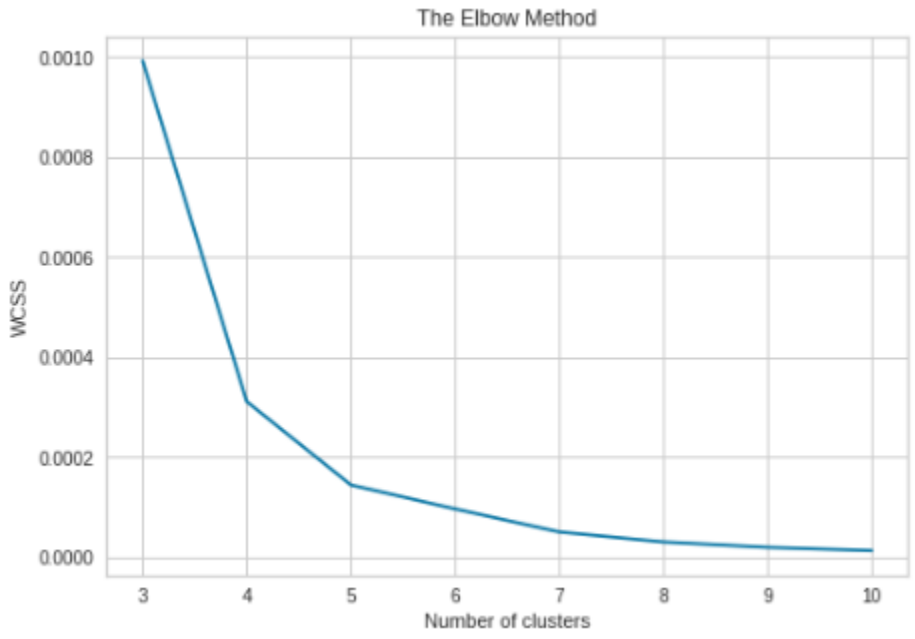
3.2 Clustering

In order to cluster the neighborhoods in New York, we use K-Means algorithm. In addition, we use the Elbow method to determine the optimal value of the number of clusters for K-means clustering.

To find out the optimal value K, we need plot the chart with the following features:

- Values for K on the horizontal axis
- The distortion or inertia on the vertical axis, which describe the values calculated by the cost function.

To determine the optimal number of clusters, we select the value of K at the “Elbow” in chart.

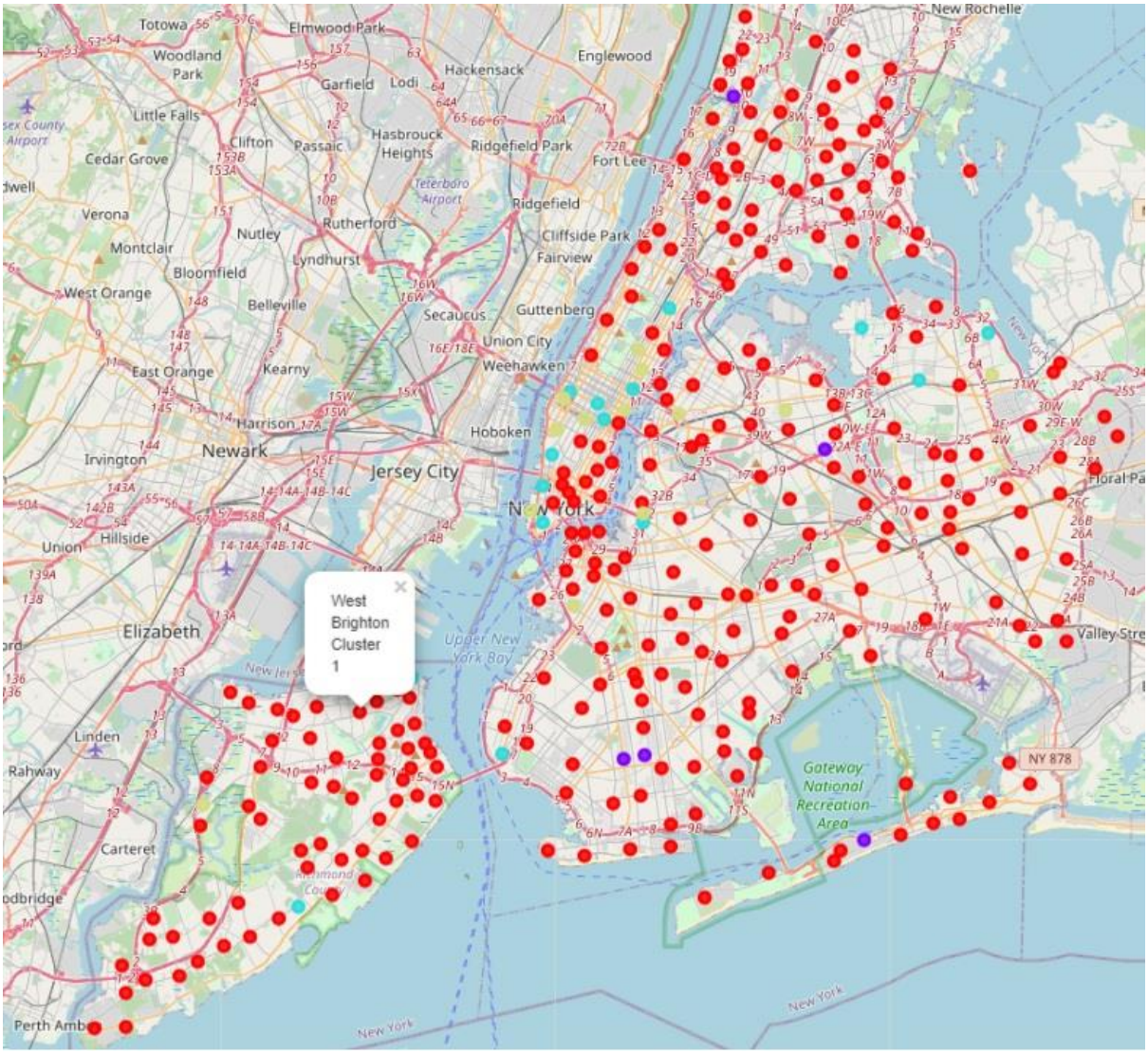


The best K value is 4. Now we need to train the model and merge the cluster labels with our dataset.

	Borough	Neighborhood	Latitude	Longitude	Steakhouse	Cluster Labels
0	Bronx	Wakefield	40.894705	-73.847201	0.0	0
1	Bronx	Co-op City	40.874294	-73.829939	0.0	0
2	Bronx	Eastchester	40.887556	-73.827806	0.0	0
3	Bronx	Fieldston	40.895437	-73.905643	0.0	0
4	Bronx	Riverdale	40.890834	-73.912585	0.0	0

4. Results

Plotting the clustered map:



Ultimately, let us analyze each cluster to determine which the most common venue category and steakhouse frequency in each venue.

Cluster 1:

	Borough	Neighborhood	Latitude	Longitude	Steakhouse	Cluster Labels
0	Bronx	Wakefield	40.894705	-73.847201	0.0	0
205	Staten Island	Grymes Hill	40.624185	-74.087248	0.0	0
204	Staten Island	West Brighton	40.631879	-74.107182	0.0	0
203	Staten Island	Rosebank	40.615305	-74.069805	0.0	0
202	Staten Island	Stapleton	40.626928	-74.077902	0.0	0
201	Staten Island	New Brighton	40.640615	-74.087017	0.0	0
199	Queens	Forest Hills Gardens	40.714611	-73.841022	0.0	0
206	Staten Island	Todt Hill	40.597069	-74.111329	0.0	0
198	Queens	North Corona	40.754071	-73.857518	0.0	0
196	Queens	Brookville	40.660003	-73.751753	0.0	0

Cluster 2:

	Borough	Neighborhood	Latitude	Longitude	Steakhouse	Cluster Labels
6	Manhattan	Marble Hill	40.876551	-73.910660	0.038462	1
53	Brooklyn	Manhattan Terrace	40.614433	-73.957438	0.040000	1
192	Queens	Lefrak City	40.736075	-73.862525	0.040000	1
194	Queens	Rockaway Park	40.580343	-73.841534	0.041667	1
98	Brooklyn	Ocean Parkway	40.613060	-73.968367	0.047619	1

Cluster 3:

	Borough	Neighborhood	Latitude	Longitude	Steakhouse	Cluster Labels
130	Manhattan	Financial District	40.707107	-74.010665	0.030000	2
61	Brooklyn	Williamsburg	40.707144	-73.958115	0.028571	2
125	Manhattan	West Village	40.734434	-74.006180	0.020000	2
152	Queens	College Point	40.784903	-73.843045	0.026316	2
115	Manhattan	Murray Hill	40.748303	-73.978332	0.024194	2
106	Manhattan	East Harlem	40.792249	-73.944182	0.023810	2
99	Brooklyn	Fort Hamilton	40.614768	-74.031979	0.028571	2
270	Manhattan	Sutton Place	40.760280	-73.963556	0.020619	2
178	Queens	Bay Terrace	40.782843	-73.776802	0.020833	2
179	Staten Island	Bay Terrace	40.553988	-74.139166	0.020833	2
116	Queens	Murray Hill	40.764126	-73.812763	0.024194	2
122	Manhattan	Tribeca	40.721522	-74.010683	0.028986	2
114	Manhattan	Midtown	40.754691	-73.981669	0.020000	2
113	Manhattan	Clinton	40.759101	-73.996119	0.020000	2
200	Staten Island	St. George	40.644982	-74.079353	0.027027	2

Cluster 4:

	Borough	Neighborhood	Latitude	Longitude	Steakhouse	Cluster Labels
107	Manhattan	Upper East Side	40.775639	-73.960508	0.011364	3
300	Manhattan	Hudson Yards	40.756658	-74.000111	0.017241	3
141	Queens	Long Island City	40.750217	-73.939202	0.013699	3
109	Manhattan	Lenox Hill	40.768113	-73.958860	0.010000	3
97	Brooklyn	South Side	40.710861	-73.958001	0.010000	3
68	Brooklyn	Gowanus	40.673931	-73.994441	0.015152	3
133	Queens	Jackson Heights	40.751981	-73.882821	0.012346	3
117	Manhattan	Chelsea	40.744035	-74.003116	0.009524	3
129	Manhattan	Battery Park City	40.711932	-74.016869	0.015873	3
272	Manhattan	Turtle Bay	40.752042	-73.967708	0.010000	3
118	Staten Island	Chelsea	40.594726	-74.189560	0.009524	3
154	Queens	Bayside	40.766041	-73.774274	0.013514	3
250	Manhattan	Midtown South	40.748510	-73.988713	0.010417	3

Finally, plot map with the best neighborhoods found:

5. Discussion

The stakeholders wanted to avoid unnecessary competition against existing steakhouses. They should avoid cluster 2 and cluster 3 that are the neighborhood with more steakhouses. Cluster 4 has little competition and a very interesting neighborhood

(Jackson Heights, Queens) with Latin and South American restaurants. Cluster 1 has 65 neighborhoods without competition.

This, of course, does not imply that those zones are actually optimal locations for a new restaurant! Purpose of this analysis was only provide info on areas not crowded with existing steakhouse. It is entirely possible that there is a very good reason for small number of steakhouses in any of those areas, reasons that would make them unsuitable for a new restaurant regardless of lack of competition in the area.

Recommended zones should therefore be considered only as a starting point for more detailed analysis, which could eventually result in location, which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

6. Conclusion

Purpose of this project was to identify New York areas close to center with low number of steakhouses in order to aid stakeholders in narrowing down the search for optimal location for a new restaurant. 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 traffic of each location, parking space, proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.