Applied Data Science: Capstone Project

The Battle of Neighborhoods – Week 2

Best Neighborhoods of NYC with Maximum Venues

By:

Ashkan Nejati

November 2019

List Contents

List of Figures

List of Tables

1. Introduction:

1.1 Problem Background:

The City of New York usually referred to as New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over a land area of about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States. It is 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 19,979,477 people in its 2018 Metropolitan Statistical Area and 22,679,948 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 commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. The city's fast pace has inspired the term New York minute. Home to the headquarters of the United Nations, New York is an important center for international diplomacy.

Situated on one of the world's largest natural harbors, New York City consists of five boroughs, each of which is a separate county of the State of New York. The five boroughs: Queens, Brooklyn, Manhattan, the Bronx, and Staten Island were consolidated into a single city in 1898. The city and its metropolitan area constitute the premier gateway for legal immigration to the United States. As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world. New York City is home to more than 3.2 million residents born outside the United States, the largest foreign-born population of any city in the world.

Tourism is a vital industry for New York City, which has witnessed a growing combined volume of international and domestic tourists, receiving an eighth consecutive annual record of approximately 62.8 million visitors in 2017. Tourism had generated an all-time high US$61.3 billion in overall economic impact for New York City in 2014, pending 2015 statistics. Approximately 12 million visitors to New York City were from outside the United States, with the highest numbers from the United Kingdom, Canada, Brazil, and China.

Figure 1: One of Tourist Attraction of New York City

Tourism industry nowadays is one of the most important resources for cities and countries to make money. Advertisement is one of the ways of attracting the tourists to different locations. Today's social media has the powerful role in this area. A social media influencer is a user who has established credibility in a specific industry has access to a huge audience and can persuade others to act based on their recommendations. An influencer has the tools and authenticity to attract many viewers consistently and can motivate others to expand their social reach. An influencer may be anyone from a blogger to a celebrity to an online entrepreneur. They must simply capitalize on a niche to attain widespread credibility. As a result, social media influencers can persuade people for visiting a specific area or city.

1.2 Problem Description:

Governor or Mayer of the City of New York has decided to make this city the first city destination for the tourists in the world in order to help different kind of businesses such as restaurants, coffee shops, gyms, entertainment zones, cinemas and others to be improved and increase their income through the tourist increment.

The city council has decided to invite the most famous Instagram Influencers to New York. Therefore, they can advertise different location and venues of this city by posting many pictures and influential comments on their profile. As mentioned above these people have many followers in their pages and this will become the best way of advertisement for NY venues.

The problem is where these influencers should be accommodated for their travel. The main purpose of this project is to find the best neighborhood with maximum venues nearby in order to access these venues very easily and in minimum time. Since the city of New York is the busy city and there are always too much traffic jams on its streets, we should find the location with maximum venues in minimum distances from them. Because it is better for the influencers to spend their time in restaurants, coffee shops and other tourist attraction location instead of wasting their time in the

traffic jam. We are going to use data science methodology, modelling and analysis tools to find some best neighborhood in NY City to solve this problem.

1.3 Target Audience:

To recommend the correct location, the city council of NY has appointed me to lead of the Data Science team. The objective is to locate and recommend to the governor which neighborhood of New York City will be best choice to accommodate the influencers. The Management also expects to understand the rationale of the recommendations made. This would interest anyone who wants to start find the specific location in New York City.

1.4 Success Criteria:

The success criteria of the project will be a good recommendation of borough/Neighborhood choice to the city council of NY based on lack of practical information in finding the best neighborhood with maximum venues. Other cities can also use the methodology and modelling technique of this project to find the specific location for their cities.

2. Data Acquisition and Cleaning

In this project, New York City will be analyzed. We will be using the below datasets for analyzing New York City.

2.1 Data Sources

Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the five boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. This dataset exists free on the web. Link to the dataset is:

https://geo.nyu.edu/catalog/nyu_2451_34572

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. The below is image of the Foursquare API data.

Each data point in this project is presented for the City's 306 neighborhoods, as well as for the City of Toronto as a whole. The data is sourced from several Census tables released by Statistics. For mapping the New York City neighborhoods shapes, we will use GeoJSON File.

2.2 Data Cleaning

Data downloaded or scraped from multiple sources were combined into tables. All the relevant data is in the features key, which is a list of the neighborhoods. Therefore, we have defined a new

variable that includes this data. Then the features data list is passed to pandas to create a Data Frame. We have checked the data frame to see it includes all 5 boroughs and 306 neighborhoods. Table below shows some data attracted from the above link:

Table 1: Some Boroughs & Neighborhoods of New York City

	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
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

We have used foursquare API for finding different kind of venues in different neighborhoods of New York City. In order to get the venues in the perimeter of each neighborhood, it is necessary to get the geographical coordinates (latitude and longitude) of each one of those and add them to the data frame. To explore the data returned by the Foursquare API, a maximum of 100 venues from the first neighborhood have been requested in a radius of 500 meters. After we got the data in features data format, we should clean the json file and structure it into a pandas data frame. Table below shows some data attracted from foursquare:

Table 2: Some Venues & Their Categories of New York City

	name	categories	lat	lng
0	Fish & Ting	Caribbean Restaurant	40.885539	-73.829151
1	Cozy Cottage Restaurant	Diner	40.886332	-73.827616
2	Mario's Pizza	Pizza Place	40.888628	-73.831260
3	Dyre Fish Market	Seafood Restaurant	40.889318	-73.831453
4	Taco Bell	Fast Food Restaurant	40.883029	-73.824901
5	Dyre Deli Grocery	Deli / Bodega	40.888235	-73.831282
6	HomeGoods	Furniture / Home Store	40.890814	-73.820849
7	Smashburger	Burger Joint	40.890172	-73.820584
8	St. Paul's Church National Historic Site	Historic Site	40.893482	-73.825328
9	Dunkin'	Donut Shop	40.885384	-73.828099

3. Exploratory Data Analysis

3.1 Calculation of distances

After we obtained the data from all neighborhoods of NYC, because we have the coordination of each neighborhood, we can display them in the map to see the scattering of them. Figure below shows the map of neighborhoods.

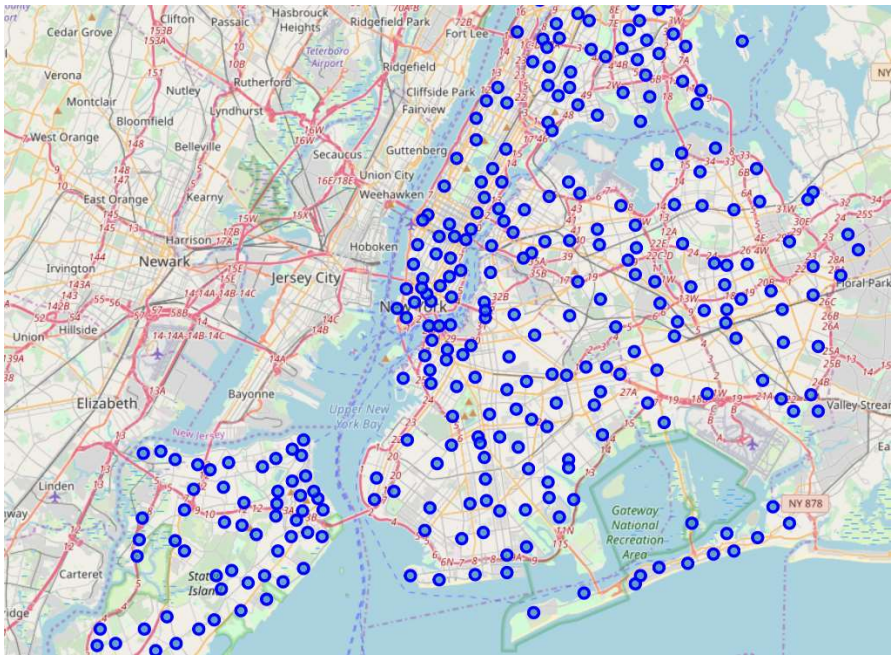

Figure 2: Some Neighborhoods of New York City

The map shows that the neighborhoods are not evenly spaced and the area cover by some of them, using a radius of 500 meters, overlaps. A different radius for each neighborhood results in a better venues search because that will avoid misrepresentation of the number of venues per neighborhood caused by too large or low radius values. To define the radius use with foursquare it is necessary to find the closest points for each neighborhood. The figure below shows the closest neighborhood to the first example in the data frame.

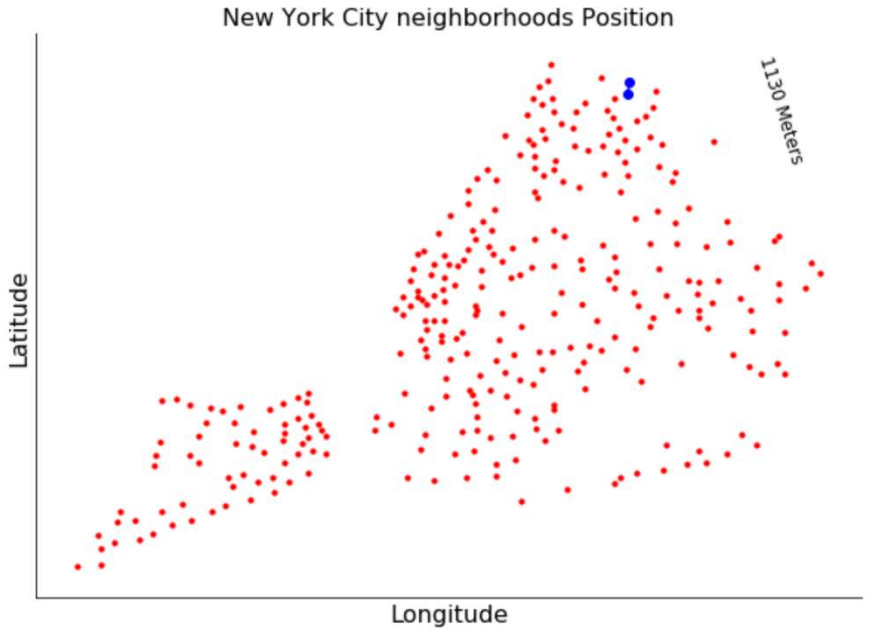

Figure 3: New York City Neighborhoods Positions

Table 3: Some Neighborhoods of New York City with Distances

	Borough	Neighborhood	Latitude	Longitude	Distance
0	Bronx	Wakefield	40.894705	-73.847201	565.0
1	Bronx	Co-op City	40.874294	-73.829939	481.0
2	Bronx	Eastchester	40.887556	-73.827806	742.0
3	Bronx	Fieldston	40.895437	-73.905643	388.0
4	Bronx	Riverdale	40.890834	-73.912585	388.0
5	Bronx	Kingsbridge	40.881687	-73.902818	436.0
6	Manhattan	Marble Hill	40.876551	-73.910660	384.0
7	Bronx	Woodlawn	40.898273	-73.867315	868.0
8	Bronx	Norwood	40.877224	-73.879391	468.0
9	Bronx	Williamsbridge	40.881039	-73.857446	439.0

A distance column is added to the data frame and is used as the radius cover for each neighborhood. The map is plotted using different radius for each neighborhood code. Now not only overlapping was avoided but also more area of the city is covered; consequently, more venues are retrieved. We have used half of distances to avoid overlapping.

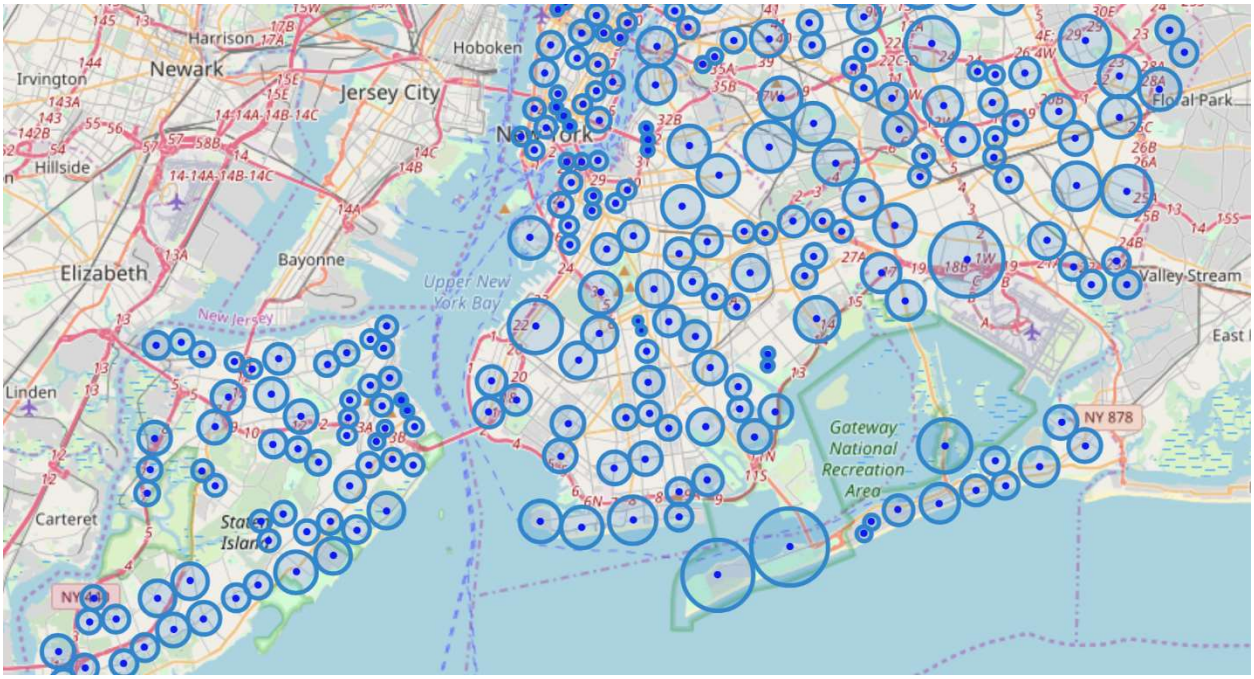

Figure 4: Some Neighborhoods of New York City with Their Distances

3.2 Adding Venues to Data Frame

We created a function to repeat the same process to all the neighborhoods in NYC for the venues and made a new data frame for them. Table below shows the data frame for the venues.

Table 4: Some Neighborhoods of New York City with Venues

	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	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop
5	Wakefield	40.894705	-73.847201	SUBWAY	40.890656	-73.849192	Sandwich Place
6	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898276	-73.850381	Caribbean Restaurant
7	Wakefield	40.894705	-73.847201	Koss Quick Wash	40.891281	-73.849904	Laundromat
8	Co-op City	40.874294	-73.829939	Dollar Tree	40.870125	-73.828989	Discount Store
9	Co-op City	40.874294	-73.829939	Rite Aid	40.870345	-73.828302	Pharmacy

We analyzed the data frame for the venue and found out the 5 neighborhoods have no venues around them.

3.3 Analyze venues data

In order to get a better sense of the best way of clustering the neighborhoods, it is necessary to analyze the venues data returned by Foursquare. Table below shows the number of venues for each neighborhoods.

Table 5: Some Neighborhoods of New York City with No. of Venues around them

	Neighborhood	No. of Venues	Distance
0	Allerton	32	353
1	Annadale	12	643
2	Arden Heights	4	694
3	Arlington	5	470
4	Arrochar	19	417
5	Arverne	18	515
6	Astoria	100	697
7	Astoria Heights	13	447
8	Auburndale	17	774
9	Bath Beach	46	656

The minimum amount of venues present on a neighborhood is 0, and the maximum is 146, expected given the limit of venues set on the request sent to the Foursquare API. 50% of the neighborhoods presents 22 or less venues. The venues Frequency Distribution of the number of venues is presented next.

Table 6: Some Statistics for the Venues

	No. of Venues
count	306.000000
mean	34.725490
std	32.207877
min	0.000000
25%	12.000000
50%	22.000000
75%	46.000000
max	146.000000

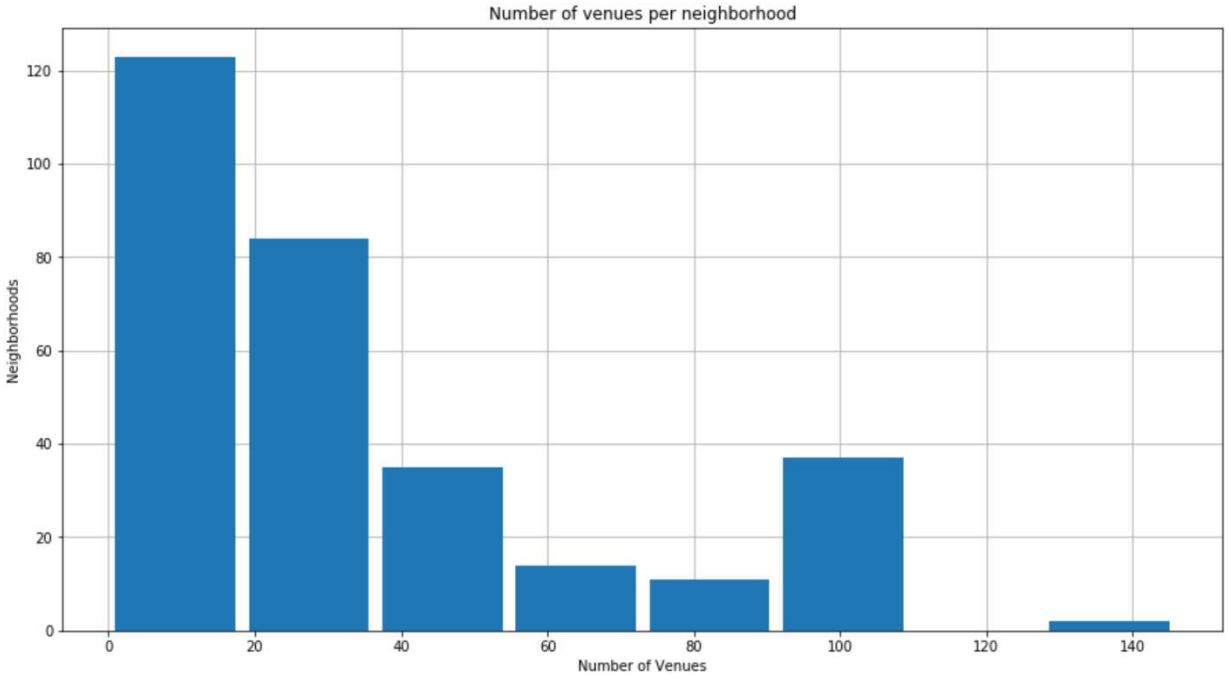

Figure 5: No. of Venues per Neighborhoods of New York City

Although the figure above shows the number of venues for the neighborhoods, but it is better to define another parameter like density to have a better sense for analyzing data. Because one neighborhood may have, large number of venues around it but on the other hand have the large area too. So given that each neighborhood has a different radius passed to the venues request, it's better to represent the venues per neighborhood in terms of density, that's venues per are cover for each neighborhood, in this case the area cover in the venues search defined by the distance to the closest neighborhood.

Table 7: Some Neighborhoods with Density of Venues

	Neighborhood	Density	Distance
0	Allerton	90	353.0
1	Annadale	18	643.0
2	Arden Heights	5	694.0
3	Arlington	10	470.0
4	Arrochar	45	417.0
5	Arverne	34	515.0
6	Astoria	143	697.0
7	Astoria Heights	29	447.0
8	Auburndale	21	774.0
9	Bath Beach	70	656.0

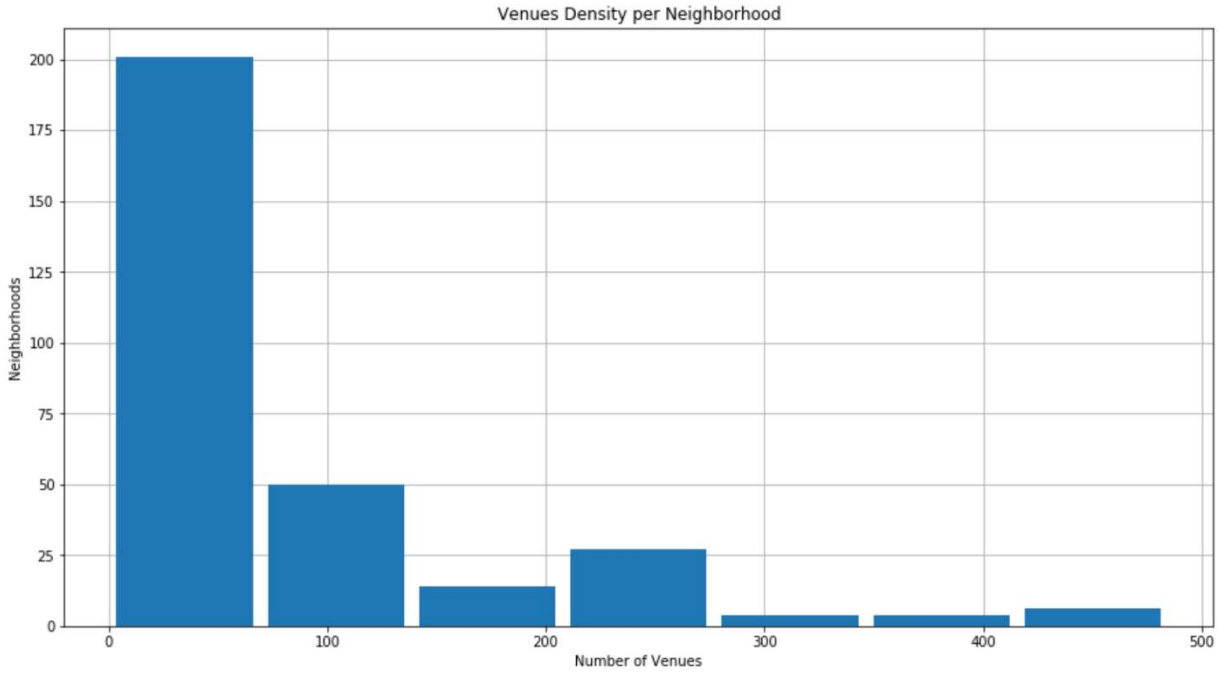

Figure 6: Venues Density per Neighborhoods of New York City

The diagram above shows that 50% of the neighborhoods presents a density between 0 and 45 venues per area (expressed as radius). That is expected given that NYC has a high population density. The last three bars on the plot have very low values.

4. Predictive Modeling

We have used K Means Clustering method for analyzing our data.

4.1 K-means Clustering

K-means clustering is a clustering algorithm that aims to partition n observations into k clusters. There are 3 steps:

1. Initialization – K initial "means" (centroids) are generated at random
2. Assignment – K clusters are created by associating each observation with the nearest centroid
3. Update – The centroid of the clusters becomes the new mean

Assignment and Update are repeated iteratively until convergence the result is that the sum of squared errors is minimized between points and their respective centroids. We have done this using scikit-learn.

4.2 Cluster Neighborhoods in NYC

The neighborhoods are clustered based on venues density. One important hyper parameter is the number of clusters and based on previous analysis a tentative value is five clusters. Next, the elbow method is used to have a better sense of the optimal number.

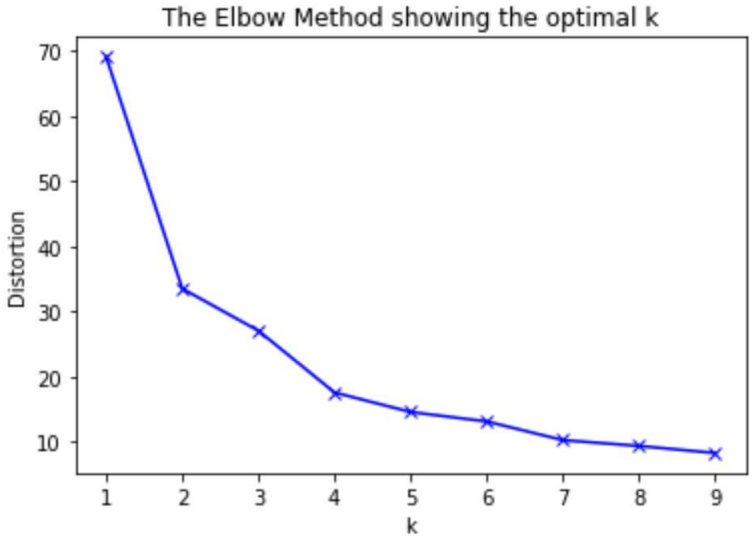

Figure 7: Using the Elbow method to calculate the optimum k

Using the elbow method, the optimal value of the number of cluster was defined as 7, which match with the value based on the histogram analysis. Then we have applied the clustering to the neighborhood data frame with density of venues as follows;

Table 8: Some Neighborhoods with their cluster number

	Neighborhood	Density	Distance	Cluster
0	Allerton	90	353.0	4
1	Annadale	18	643.0	0
2	Arden Heights	5	694.0	0
3	Arlington	10	470.0	0
4	Arrochar	45	417.0	3
5	Arverne	34	515.0	3
6	Astoria	143	697.0	6
7	Astoria Heights	29	447.0	3
8	Auburndale	21	774.0	0
9	Bath Beach	70	656.0	4

5. Results and Discussion

5.1 Examine Clusters

First, we have checked the centroids values of venues density and neighborhoods per cluster as follows;

Table 9: Clusters with their Centroid and Number of Neighborhoods

	Cluster	Centroid	Neighborhoods
0	0	14	98
1	3	38	79
2	4	74	61
3	6	129	19
4	1	213	25
5	5	278	15
6	2	442	9

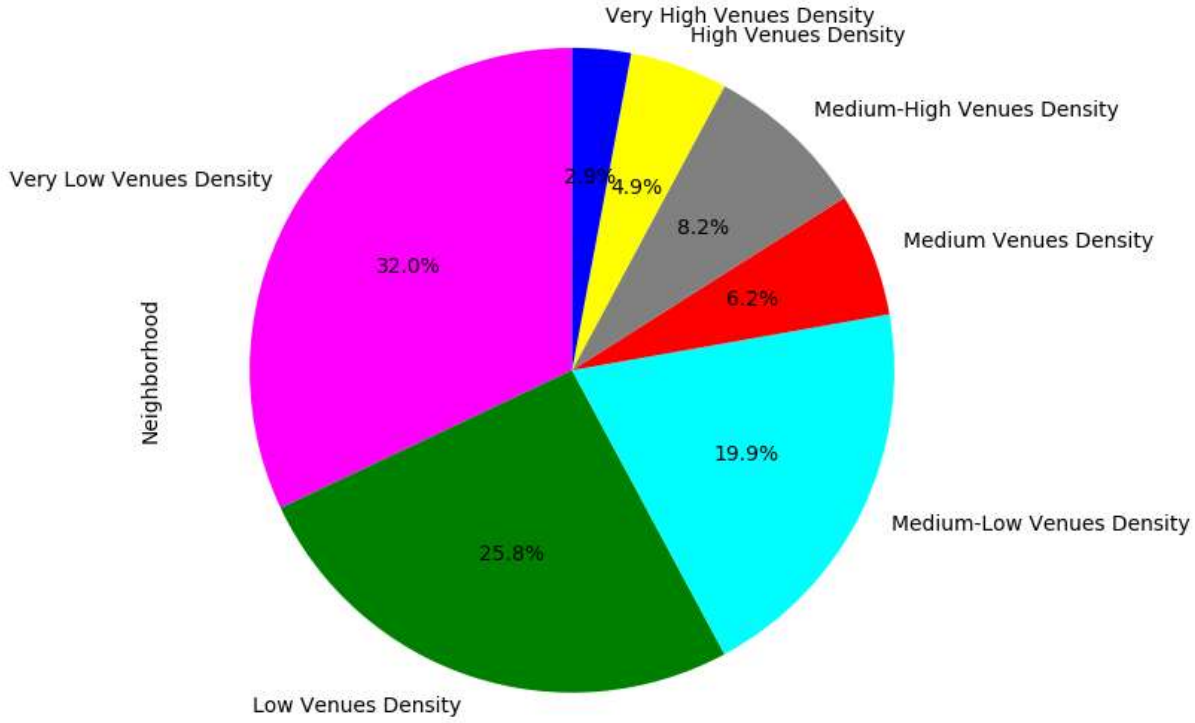

Figure 8: Percentage of Neighborhoods for Each Cluster

We put the neighborhoods with 7 categories;

- ✓ Very Low Venues Density: Centroid equal to 14
- ✓ Low Venues Density: Centroid equal to 38
- ✓ Medium-Low Venues Density: Centroid equal to 74
- ✓ Medium Venues Density: Centroid equal to 129
- ✓ Medium-High Venues Density: Centroid equal to 213
- ✓ High Venues Density: Centroid equal to 278
- ✓ Very High Venues Density: Centroid equal to 442

In addition, the figure above shows the percentage of neighborhoods for each categories.

5.2 Clustered Neighborhoods

We added the cluster column to the main data frame for the neighborhoods with their coordinate to show the in the map. Table below show the final data frame have been used for the map.

Table 10: Some Boroughs & Neighborhoods of with their Clusters

	Borough	Neighborhood	Latitude	Longitude	Distance	Cluster
0	Bronx	Wakefield	40.894705	-73.847201	565.0	4
1	Bronx	Co-op City	40.874294	-73.829939	481.0	0
2	Bronx	Eastchester	40.887556	-73.827806	742.0	0
3	Bronx	Fieldston	40.895437	-73.905643	388.0	0
4	Bronx	Riverdale	40.890834	-73.912585	388.0	3
5	Bronx	Kingsbridge	40.881687	-73.902818	436.0	3
6	Manhattan	Marble Hill	40.876551	-73.910660	384.0	6
7	Bronx	Woodlawn	40.898273	-73.867315	868.0	3
8	Bronx	Norwood	40.877224	-73.879391	468.0	0
9	Bronx	Williamsbridge	40.881039	-73.857446	439.0	4

The map below also shows the distribution of neighborhoods with 7 different colors belong to each cluster;

- ✓ Magenta: Very Low Venues Density
- ✓ Green: Low Venues Density: Centroid equal to 38
- ✓ Cyan: Medium-Low Venues Density: Centroid equal to 74
- ✓ Red: Medium Venues Density: Centroid equal to 129
- ✓ Gray: Medium-High Venues Density: Centroid equal to 213
- ✓ Yellow: High Venues Density: Centroid equal to 278
- ✓ Blue: Very High Venues Density: Centroid equal to 442

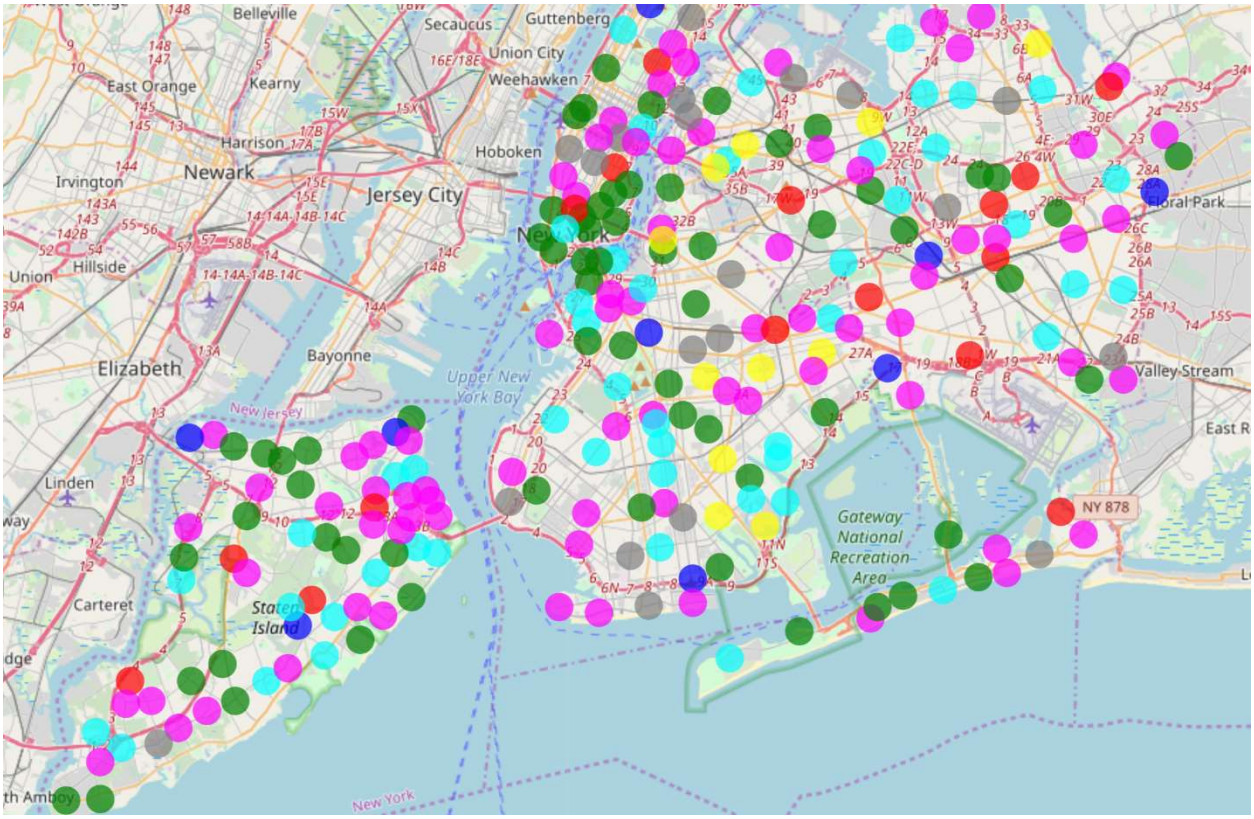

Figure 9: Neighborhoods of New York City with Their Categories

Table 11: Selected Neighborhoods of NYC with Maximum Number of Venues

No.	Borough	Neighborhood	Latitude	Longitude	Distance	Density
0	Manhattan	Chinatown	40.71562	-73.9943	242	413
1	Manhattan	Clinton	40.7591	-73.9961	216	462
2	Manhattan	Greenwich Village	40.72693	-73.9999	266	375
3	Manhattan	Hudson Yards	40.75666	-74.0001	216	379
4	Manhattan	Little Italy	40.71932	-73.9973	212	471
5	Manhattan	Murray Hill	40.7483	-73.9783	309	478
6	Manhattan	Soho	40.72218	-74.0007	212	471
7	Brooklyn	North Side	40.71482	-73.9588	223	448
8	Brooklyn	South Side	40.71086	-73.958	206	485

6. Conclusion

In this project, we clustered different neighborhoods of New York City according to number of density of venues to help us finding the best location with large number of venues around them to invite the Instagram Influencers for advertising these venues to their followers. As a result, the tourism industry can be improved in the NYC. Table above shows the location of these neighborhoods in the map. Other cities and countries can use this method to find the neighborhoods with large number of venues.