Clustering New York City Neighborhoods

Tony Celia

October 24, 2020

Introduction

New York City is massive with over 250 distinct neighborhoods. Moving there can be challenging as well. Most people will visit a few spots and then need to determine which neighborhood to settle in, though most people will not venture out of Manhattan.

In this report I will try to group neighborhoods into categories. The idea behind this is to guide someone to neighborhoods they would like based on the number and variety of not only venues but also the vast park system.

Data Used

There are two main sources for the data used.

First is NYC Open Data (https://opendata.cityofnewyork.us/). This site is hosted by NYC as a repository for data produced by the government. The data is free to use and there is a considerable amount hosted on the site.

Data can easily be pulled from this site via the Socrata Open Data (SODA) API. The data sets used in this analysis are:

- Neighborhood Names GIS
- Park Properties

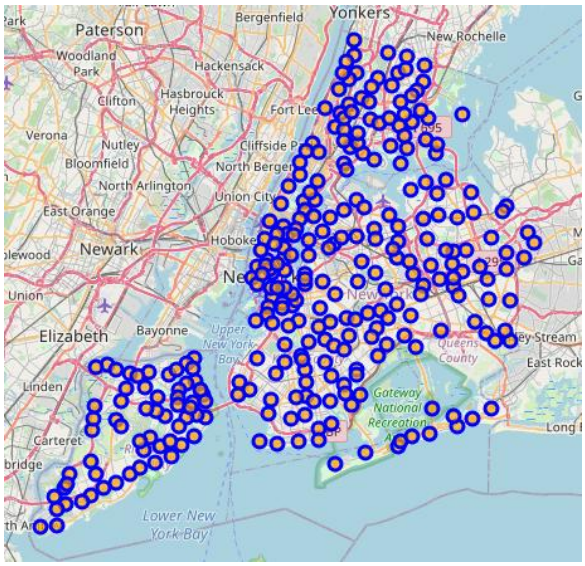
The second source for the data is Foursquare. Foursquare is a social networking site that allows consumers to discover and share information about businesses around them and allows the company to track movements and trending locations.

Both data sources offered clean data so no clean up was necessary.

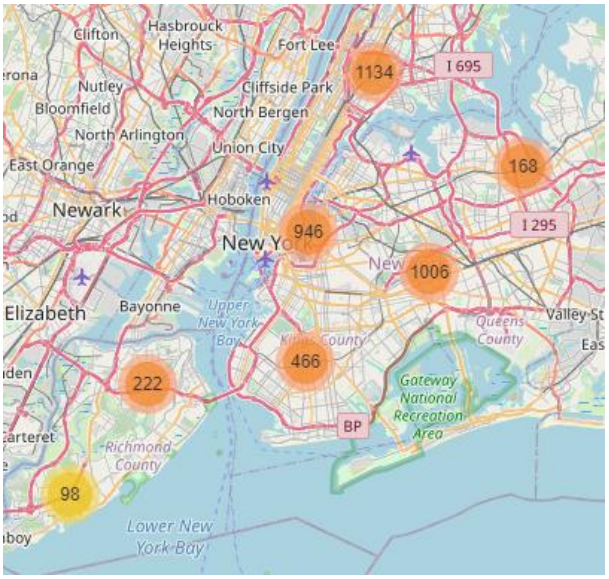
Data Analysis

While gathering and parsing through the data, the following observations were made:

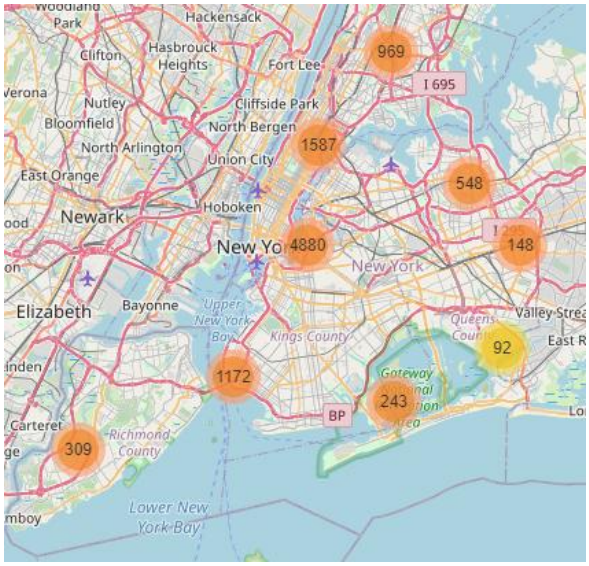
NYC is massive. Between all five boroughs, there are 299 neighborhoods. The map on the next page plots all 299 neighborhoods.



New York City also has a massive amount of parks (2,020). The map below contains them.



As for venues, as you would expect, NYC is loaded. The map on the next page contains almost 10,000 spots.



To analyze the data, I did the following:

Pulled a list of coordinates for each neighborhood and park as well as searched each neighborhood for up to 100 venues within 500 meters.

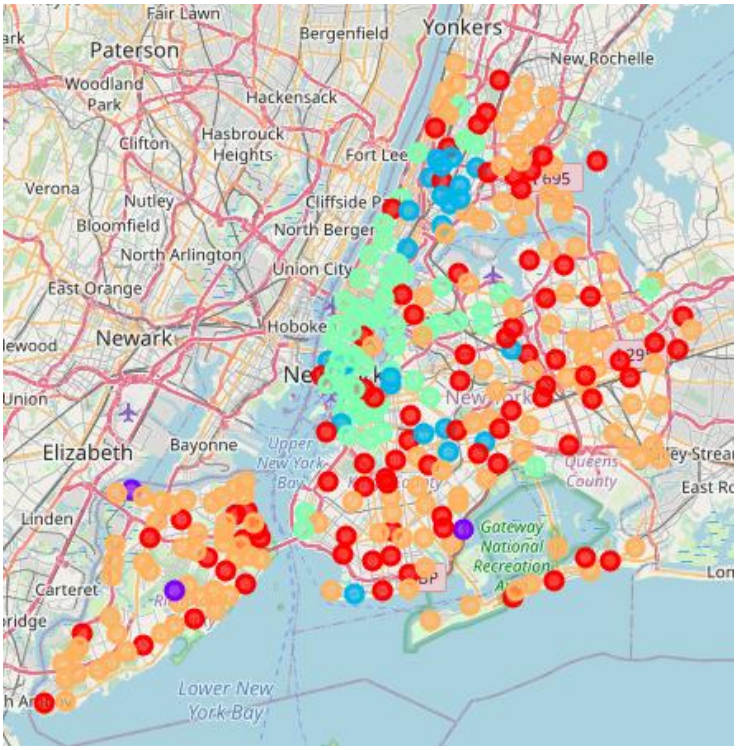
I went through each park and calculated the distance from each neighborhood. If the park was less than 0.33 miles, I added the number as well as the acreage of the park.

With a list of each neighborhood, the amount of parks, the size of the parks, the amount of venues, and the count of the unique types of venues, I was able to complete my analysis. After normalizing the data, I used k-means clustering to partition the neighborhoods into 5 categories. K-mean clustering attempts to categories data into like groups.

The categories had the following characteristics:

Labels	Count of Parks	Acres	Venue	Venue Category
0	4.756098	21.550268	32.719512	25.317073
1	2.666667	1495.167333	6.000000	5.333333
2	29.391304	25.666696	32.695652	23.913043
3	9.965517	18.102759	89.689655	54.258621
4	2.721805	24.127517	11.323308	9.639098

The map of the clusters are as follows:



Cluster	Color
0	Red
1	Purple
2	Blue
3	Green
4	Orange

Conclusion

We can see some trends from this data. The most obvious is with the Manhattan-like outer borough spots. Neighborhoods such as Williamsburg, Park Slope, Astoria, and Jackson Heights, to name a few, are very popular and have been growing with plenty of venues but overall limited outdoor space.

Someone looking for a more balanced life would be better served in red or blue as there are more outdoor/green spaces close but also plenty of venues, exactly what you would expect in the outer boroughs. Orange neighborhoods may have limited venues.

Purple neighborhoods have a massive amount of outdoor spaces but very limited venues to visit.