The Battle of Neighborhoods: NYC Families

Eric Jones

June 19, 2020

1. Introduction

1.1 Background

New York City, NY is the most populated city in the United States of America. As an international hub for travel, food, culture, art, business, and much more, a large number of people work and live in the area. As an area where so many people's lives happen, a growing amount of people are looking to bring families to live within the city's limits. However, with New York City being a largely urban environment, parents must determine neighborhoods in the area that cater do the needs or raising children in a city environment. Specifically, neighborhoods with nearby parks, civic centers, childcare facilities, and schools would be much more desirable for people living in the city with children. The identification of these neighborhoods would provide families the confidence to successfully raise children in the large-city environment of New York.

1.2 Problem

Living in New York City while raising kids requires finding an environment that is suitable for children and the responsibilities of the parents. A way to assess the suitability of a neighborhood in New York is to determine the number of parks, civic centers, childcare facilities, schools, and other similar venues that exist in the area. The goal of this project is to leverage the Foursquare location data to determine the best neighborhoods for families with children in New York City.

1.3 Audience

The target audience for this project is families moving to New York or living in New York already. The analysis performed in this project determines the best neighborhoods for these families to live with their children. Another stakeholder that could make use of this analysis is retail companies. Since these companies assist families finding a place to live within the city, the list of family-friendly neighborhoods determined by this project can help real estate agents determine the best locations for families.

2. Data

2.1 Data Sources

In order to determine the solution to the problem stated above, neighborhood and location data must be gathered to determine the venues that exist within a neighborhood. To do this, the Foursquare API is used to gather venue information based on latitudes and longitudes.

Data regarding the neighborhoods, including the latitude and longitude of each neighborhood, is obtained from an online data source: https://cocl.us/new_york_dataset

2.2 Data Usage

The dataset containing the New York neighborhoods will be used to generate the list of neighborhoods with their respective latitude and longitude. This data will be the information that is passed to the Foursquare API.

The Foursquare API will be used to generate venue category data in the neighborhoods in New York City. Using the explore function within the Foursquare API will enable the collection of all venue types within a certain radius of each neighborhood, based on the neighborhood's latitude and longitude. Once this data is collected, the results will be filtered to include only relevant venue types, namely childcare locations, parks, civic centers, and others. The frequency of each of the venue types will be determined for each neighborhood, and the neighborhoods will be clustered based on these frequencies. Once clustered, the set containing the highest frequency of these particular venues can be selected, thus determining the list of best neighborhoods for families in New York City.

3. Methodology

3.1 Exploratory Data Analysis

For the first part of exploratory data analysis, the generated dataframe from the dataset was examined. The dataframe was a table that included a list of neighborhoods with their respective borough, latitude, and longitude. To visualize the data, the Folium library was used to create a map with markers for each neighborhood, shown in Figure 1 below.

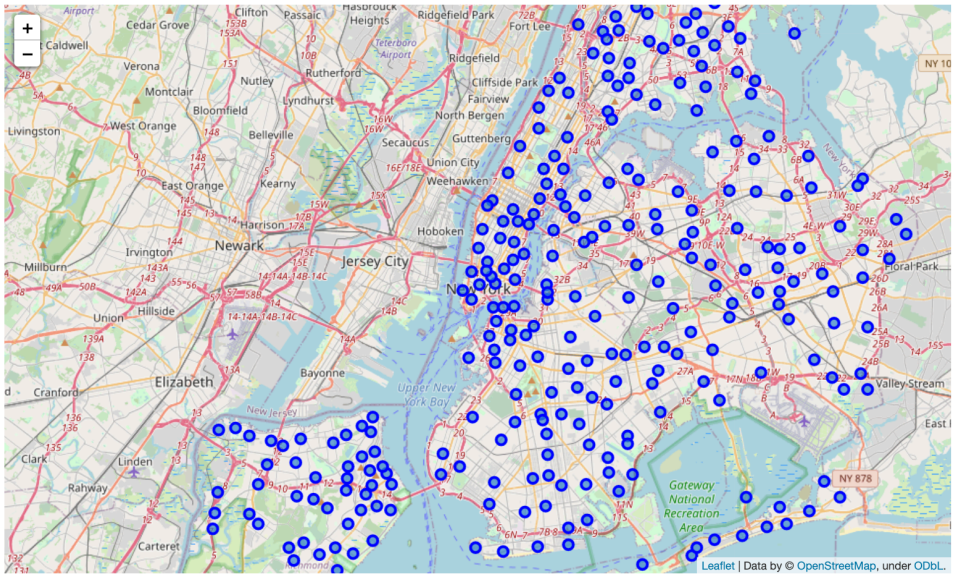

Figure 1: Map of New York City neighborhoods

After this, the matplotlib library was used to create a bar graph of the number of neighborhoods per borough in New York City. This enabled an understanding of how many neighborhoods were

in each borough, as this could not easily be understood from the map. The graph is in Figure 2 shown below.

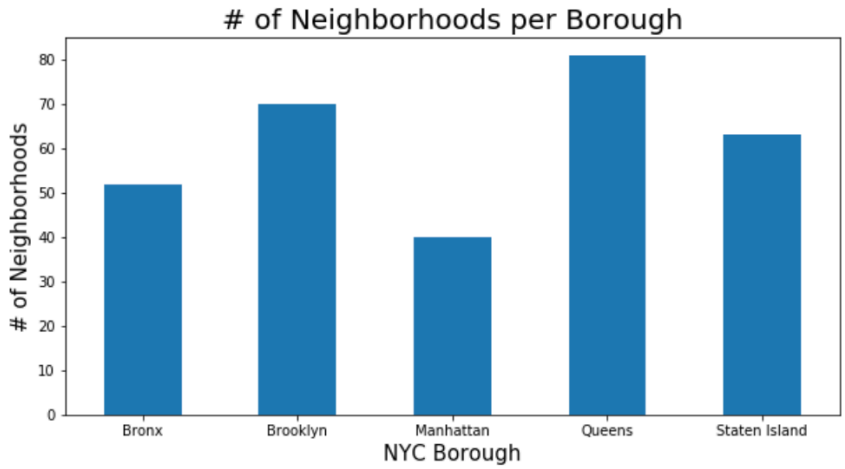

Figure 2: The number of neighborhoods per borough in New York

The next step was to use the Foursquare API to gather data regarding venues located within each neighborhood. For each neighborhood, the latitude and longitude points were used as the center points, and a radius of 500 meters around these center points was used to gather a list of venues. After writing a function to complete this process for every neighborhood, a dataframe was created with a complete list of the venues for all neighborhoods. The first few rows of this dataframe are shown in Figure 3 below.

	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	Walgreens	40.896528	-73.844700	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
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

Figure 3: Dataframe of all collected venues for the neighborhoods

Using this dataframe, all of the unique values of venue categories were listed. This list was sorted through to determine the venue categories that were most applicable to the problem involving families living in New York City. This resulted in the set of venue category types that would be used to reduce the dataframe: pharmacy, park, playground, kids store, baby store, school, daycare, and recreation center. Using this list as a filter, a reduced data frame with only the venue types of concern was created, shown in Figure 4.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Co-op City	40.874294	-73.829939	Rite Aid	40.870345	-73.828302	Pharmacy
3	Co-op City	40.874294	-73.829939	The Park	40.877645	-73.830836	Park
4	Riverdale	40.890834	-73.912585	Bell Tower Park	40.889178	-73.908331	Park

Figure 4: Dataframe of venues related to the problem for the neighborhoods

3.2 K-Means Clustering of the Neighborhoods

To analyze the data and answer the problem, the machine learning process of k-means clustering was used to group the data based on the venues in each neighborhood. The clustering strategy was used to identify a similar group of neighborhoods that contained the best venues in regard to the lives of families. One hot encoding was used with the gathered venue category data to identify the frequency of each venue type in a given neighborhood. The scikit-learn clustering algorithm was run to fit the encoded data. The determined cluster label and the venue categories in order of frequency were put into a dataframe. The dataframe is shown below in Figure 5.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	Borough	Latitude	Longitude
0	4	Allerton	Playground	Pharmacy	School	Recreation Center	Park	Kids Store	Daycare	Baby Store	Bronx	40.865788	-73.859319
1	4	Annadale	Pharmacy	Park	School	Recreation Center	Playground	Kids Store	Daycare	Baby Store	Staten Island	40.538114	-74.178549
2	0	Arden Heights	Pharmacy	School	Recreation Center	Playground	Park	Kids Store	Daycare	Baby Store	Staten Island	40.549286	-74.185887
3	0	Arrochar	Pharmacy	School	Recreation Center	Playground	Park	Kids Store	Daycare	Baby Store	Staten Island	40.596313	-74.067124
4	2	Arverne	Playground	School	Recreation Center	Pharmacy	Park	Kids Store	Daycare	Baby Store	Queens	40.589144	-73.791992

Figure 5: Dataframe of cluster labels with venue categories listed in order of frequency

The neighborhoods belonging to each cluster label were listed out to determine patterns in the venue types and assign labels. The neighborhoods in Cluster 0 all had the most popular venue type as pharmacies. The neighborhoods in Cluster 1 all had a park as the most common venue type. The neighborhoods in Cluster 2 all had a playground as the most common venue type. The neighborhoods in Cluster 3 all had a park and a playground as the two most common venue types and either a school or pharmacy as the third most common. Finally, the neighborhoods in Cluster 4 all had a pharmacy and either a playground or park as the common venue types. The list of cluster labels that was determined through an assessment of the neighborhoods within it is shown below in Figure 6.

Determined Cluster Labels:

CLUSTER 0: Pharmacy
CLUSTER 1: Park
CLUSTER 2: Playground
CLUSTER 3: Park & Playground
CLUSTER 4: Pharmacy & Playground/Park

Figure 6: The determined cluster labels after k-means clustering algorithm

4. Results

4.1 Assessment of Clustering

After the clustering algorithm, the neighborhoods in each cluster were compared based on the venue categories that were found in each. The neighborhoods in Cluster 3 were determined to be the neighborhoods containing the highest frequency of family-important venues. The reason for this is that the neighborhoods in this cluster all contained a playground and a park as the most common venue types. In addition to this, the neighborhoods contained either schools or

pharmacies as the most common venue types. Areas for children to safely spend time, such as parks and playgrounds, and schools and pharmacies are some of the most important venues to have access to as a family with children living in a city.

4.2 Map of Best Neighborhoods for Families

Using the Folium library in Python, a map of New York city was created with labels of the neighborhoods that were determined to be best fit for families. The map is shown in Figure 7 below.

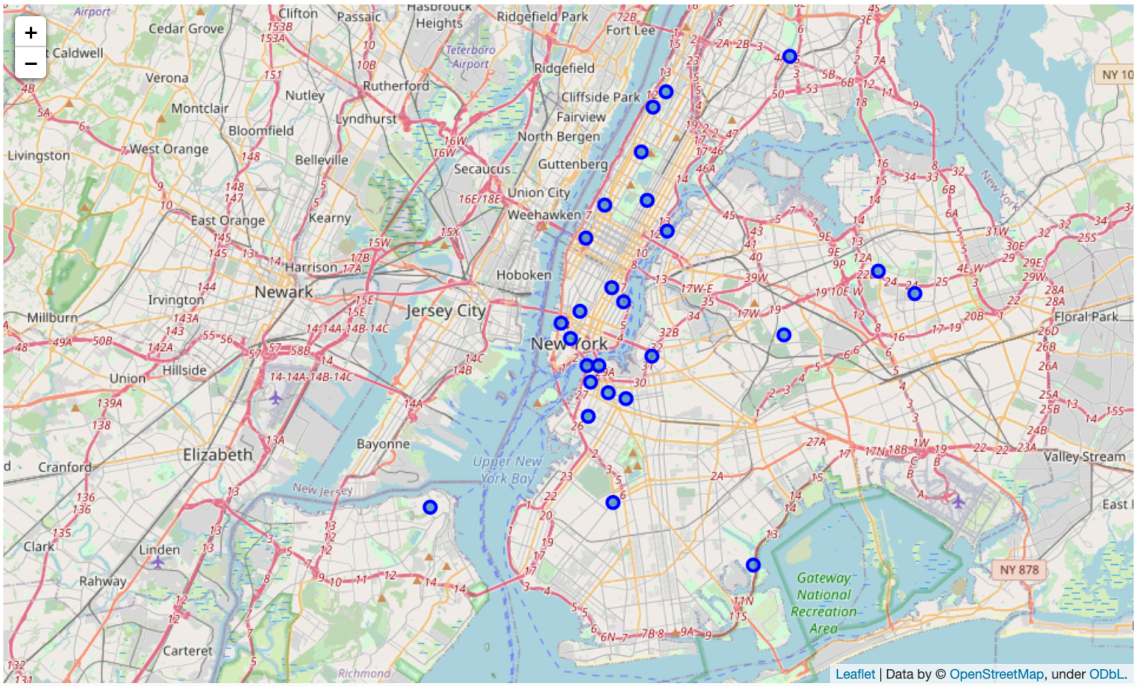

Figure 7: Map of neighborhoods determined to be best fit for families in New York City

4.3 List of Best Neighborhoods for Families

The list of the names of the neighborhood that were identified as the best for families in New York City is shown in Figure 8 on the next page.

	Cluster Labels	Neighborhood
19	3	Bergen Beach
25	3	Brooklyn Heights
31	3	Carroll Gardens
39	3	Civic Center
43	3	Clinton
57	3	Downtown
58	3	Dumbo
75	3	Fort Greene
79	3	Fulton Ferry
83	3	Gramercy
87	3	Greenwich Village
88	3	Hamilton Heights
98	3	Inwood
102	3	Kensington
110	3	Lincoln Square
115	3	Manhattan Valley
116	3	Manhattanville
122	3	Middle Village
134	3	New Brighton
148	3	Pomonok
153	3	Queensboro Hill
164	3	Roosevelt Island
180	3	Stuyvesant Town
188	3	Tribeca
192	3	Upper East Side
200	3	West Farms
203	3	Williamsburg
207	3	Woodlawn

Figure 8: List of neighborhoods in New York best fit for families

5. Discussion

5.1 Observations

During the filtering process of the data analysis, the number of neighborhoods being investigated decreased. This was the result of some neighborhoods not returning any venue types that aligned with the filtering categories. This could limit the ability to distinguish between the best and worst neighborhoods for families with children. Additionally, most neighborhoods featured multiple venue categories that were not present in that neighborhood. This limited the ability to distinguish the difference between the neighborhoods during the clustering process and the analysis of the clusters.

5.2 Recommendations

Based on these observations and the results, one recommendation is to determine a better way to handle the neighborhood searches commonly returning no venue types of a particular category. One simple way to do this would be to increase the radius of the Foursquare API search. This would capture a larger number of venues that were still close by to the center of each neighborhood.

A second recommendation to answer the problem of finding the best neighborhoods for families in New York would be to introduce more kinds of data into the analysis. For example, crime data

for the city of New York could be collected and analyzed. This would enable the determination of which neighborhoods have low crime rates, another important factor for the safety of families.

6. Conclusion

In this study, the neighborhoods within New York City were analyzed to determine the list of neighborhoods that were best for families living with children. The Foursquare API was used to gather data regarding nearby venue category types to understand which family-oriented venues existed in each neighborhood. The scikit-learn k-means clustering algorithm was used to group the neighborhoods based on the frequency of venue types. Using this analysis, neighborhoods that commonly featured playgrounds, parks, schools, and pharmacies were determined the best fit. This study proves to be relative to people with families either moving or living in New York City deciding where to live. Additionally, real estate companies can use this analysis and similar analyses to determine which neighborhoods to recommend to clients searching for homes within a specific city's limits.