Segmentation of Similar Neighborhoods in Brooklyn

1. Introduction

1.1 Background

Brooklyn is one of Ney York’s five boroughs and with 37,137 persons per square mile the most densely populated area in the United States besides Manhattan. Around 2.5 million people live within the limits of Kings County, NY. Many people move to Brooklyn, either from other boroughs, for example because of rising real estate prices in Manhattan, from other places in the United States (most coming from Chicago, Detroit or San Francisco) or even overseas.

1.2 Problem

When moving, people take several different features of potential new homes into account. This includes characteristics of the property itself as well as the location of the property. Characteristics of the property itself includes features like number of bed- and bathrooms, type of the floor, age of the property amongst many others. Characteristics of the location includes for example the distance to the workplace, what kind of people mostly live in the neighborhood or distance to all kind of venues. The characteristics of the property can be verified relatively easy with an inspection. Therefore, this project is taking a closer look at the location and aims to get people moving to Brooklyn a sense of in which kind of neighborhood they are moving based on the range of venues nearby.

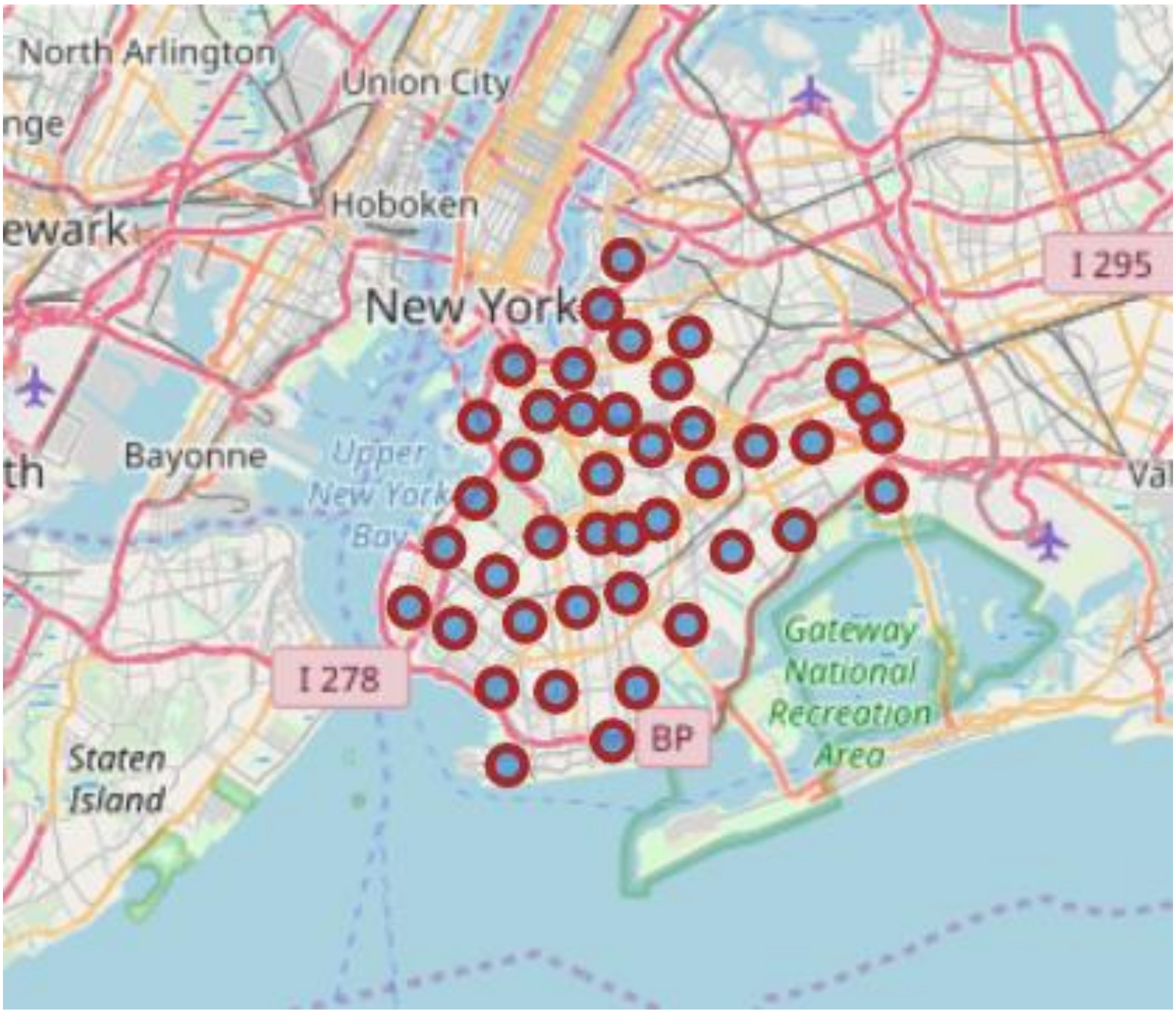
1.3 Interest

A large group of the people moving to Brooklyn right now are millennials. Since leisure activities are becoming increasingly important in this group and young people are also increasingly looking for a healthy work-life balance, it is very important for this group to have nearby venues to get to in their leisure time. Of course, Brooklyn has a lot of venues to offer, but while for some of the potential future residents it is more important to be closely located to bars or clubs, others prefer restaurants, parks or other venues. This is the gap this project is trying to close by dividing neighborhoods in Brooklyn based on the venues available, that these people get a sense for the venues available.

2. The data

One of the most important characteristics to achieve this goal is to get coordinates of different neighborhoods in Brooklyn. Neighborhoods in this approach are represented by zip code areas. The “US Zip Code Latitude and Longitude”-dataset from public.opendatasoft.com was used containing latitude and longitude coordinates for each zip code area in the United States. The “US Zipcode to County State to FIPS Look up“-dataset from data.world was used to retrieve county information for each zip code. To retrieve the venues for each zip code area, the name, location and type of venues were queried over the Foursquare Developer API.

All in all, 46 zip codes areas in Kings County were used. The map below was created using folium, showing the distribution of zip code areas on the map. 4200 venue locations and categories were retrieved from the Foursquare API.



3. Methodology

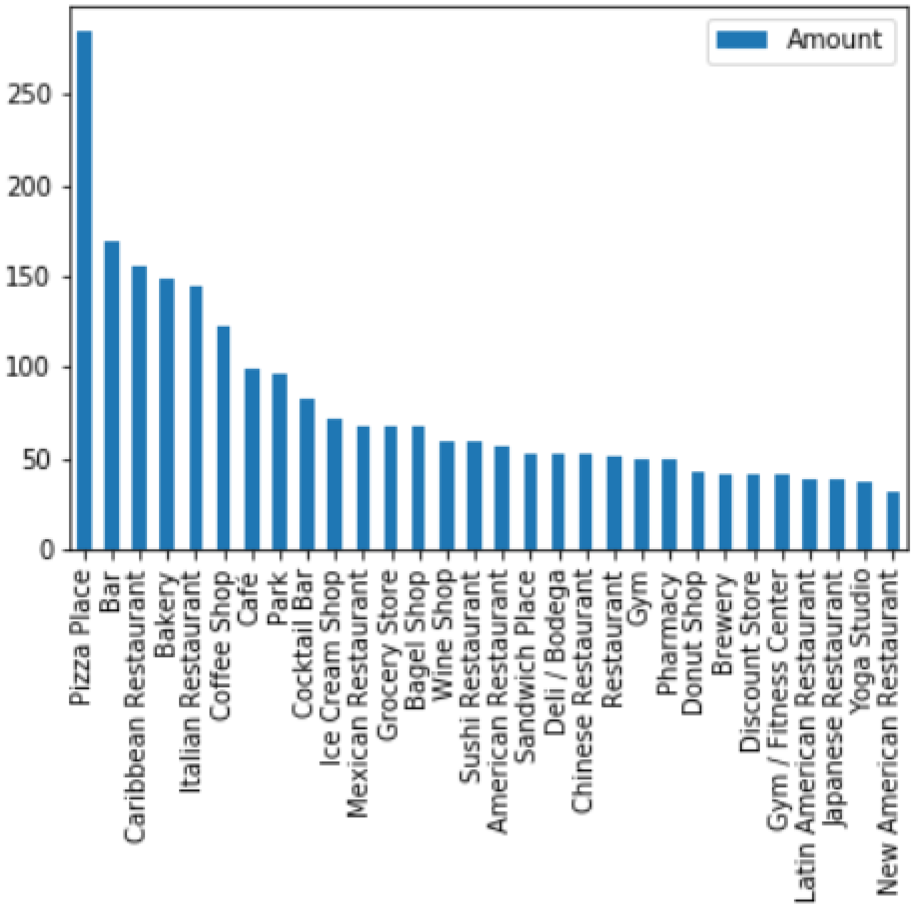
3.1 Data wrangling

The project itself was calculated within the IBM Cloud, using Jupyter Notebooks in IBM Watson Studio. First, unnecessary columns were deleted and some of the columns were renamed to make merging easier. To check which of the zip codes belong to Brooklyn (Kings County, New York), the datasets with latitude and longitude coordinates was merged with the dataset containing the county names based on the zip codes. Then, the state was initially restricted to New York (as other state have a “Kings County” as well) and the county to Kings County. As there are some duplicates in the data which would distort the clustering, duplicates were deleted. An algorithm returning nearby venues for each latitude and longitude combination was written, using the Foursquare

Developer API. The resulting venues were one-hot encoded to create dummies, that were aggregated by zip code area.

3.2 Exploratory Data Analysis

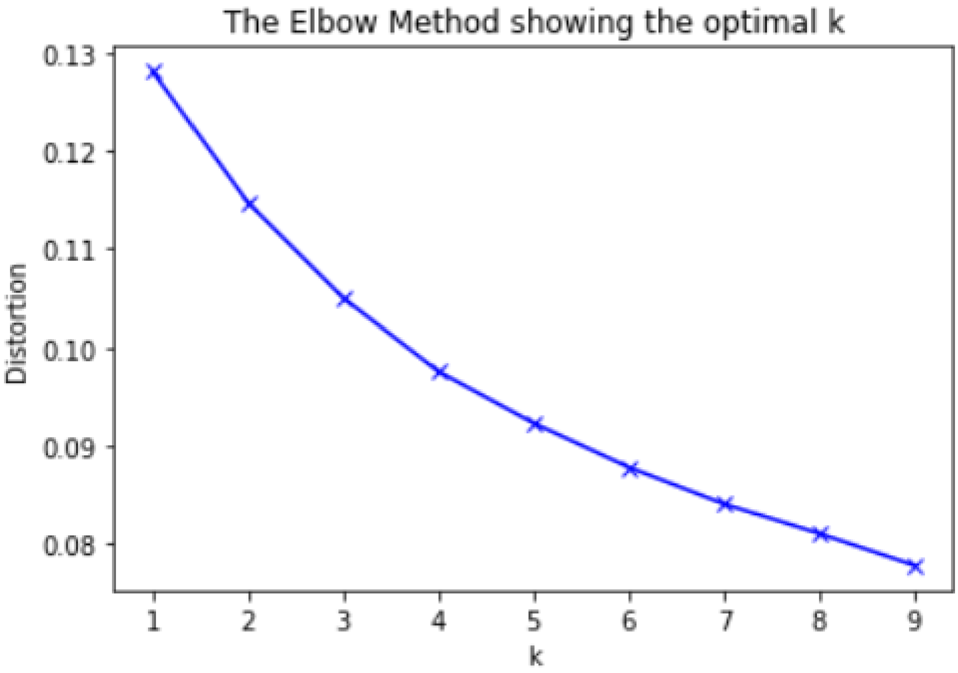
To get an overview over all venue categories and their distribution, the frequency of all venue category was calculated. 30 most common venue categories can be seen in the bar chart below:



3.3 Applying an k-means Algorithm

For segmentation of the dataset into clusters, the k-means algorithm was used, which is one of the most popular algorithms for unsupervised learning. It divides the data in k segments, that the sum of squared residuals of the cluster mean becomes minimal.

To examine the optimal amount of clusters, a Scree-Test was performed to see how much of additional variance would be explained by adding an additional cluster. The results are shown in the plot below:



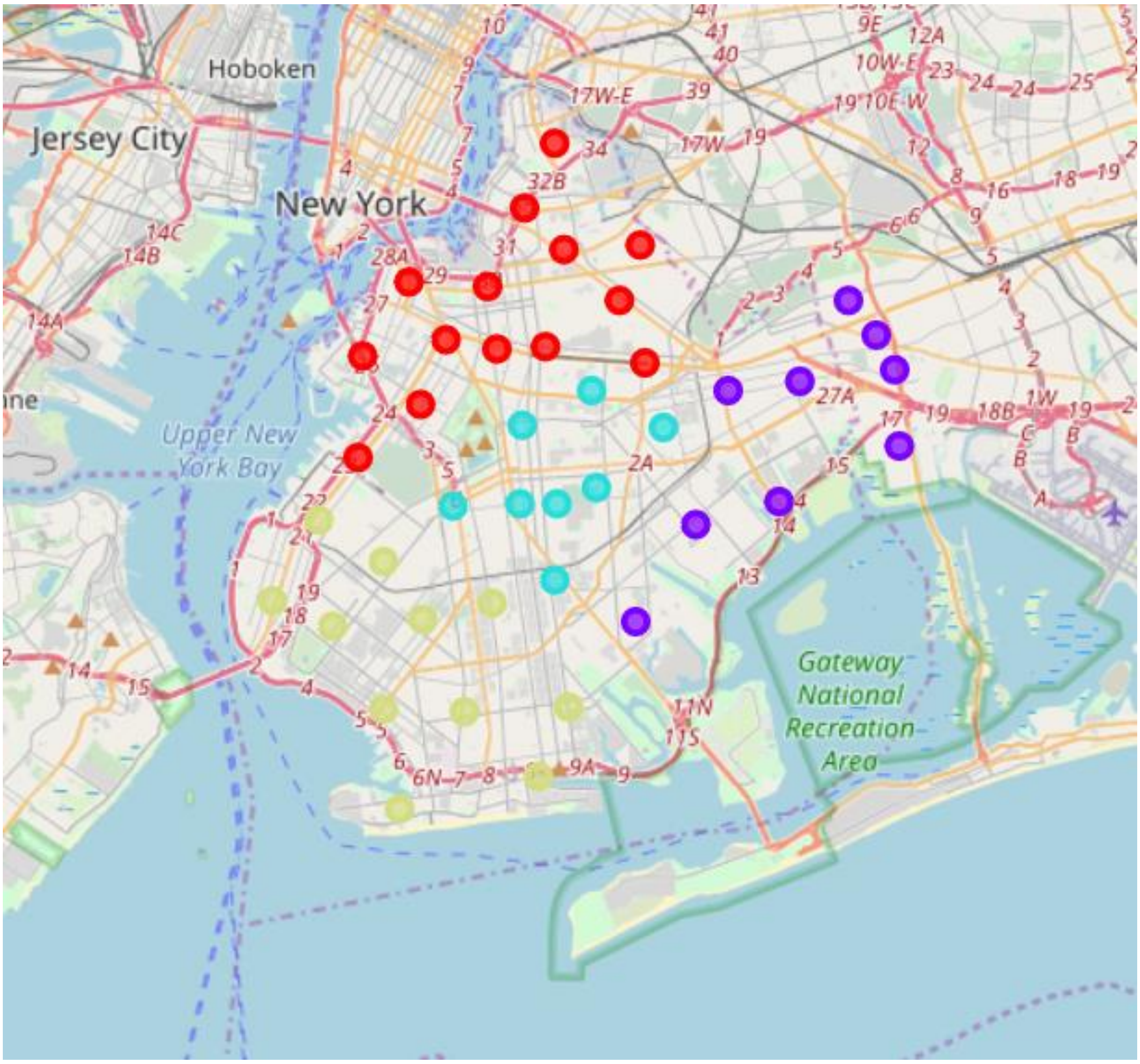
It becomes visible, that this is not an extremely clear result, as there is some room for interpretation over the optimal cluster amount. However, four clusters were chosen as the angle here is a little bit smaller. After applying the algorithm, the most common venue categories for each cluster were calculated (see bar charts in section 4.2).

4. Results

4.1 k-Means

When looking at the bar chart with frequencies of the most common venue categories, it becomes visible, that pizza places highly dominate in quantity, even more if you take into account that Italian restaurants are the fifth most common venue category. Most common places are related to food or drinks (restaurants, bars, cafés) or sport (gyms, yoga studios). A little bit unexpected is the relatively high amount of wine shops, which is even higher than pharmacies.

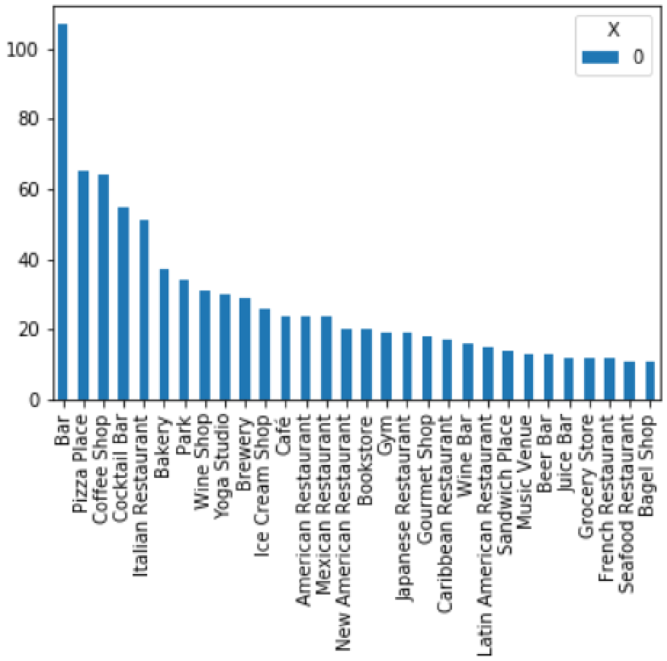
The k-means algorithm with four clusters comes to the following result:



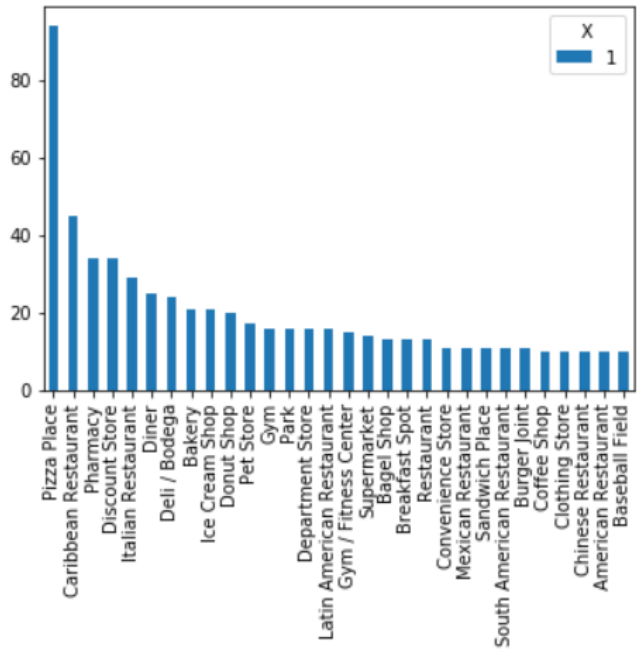
In this folium map it becomes visible that the segmentation divides the borough relatively clear into four cluster, each one in the direction of Manhattan (red – cluster 0), Queens (violet – Cluster 1) and Staten Island (green – Cluster 3) as well as one in the center (turquoise – cluster 2).

4.2 Most common venues for each cluster

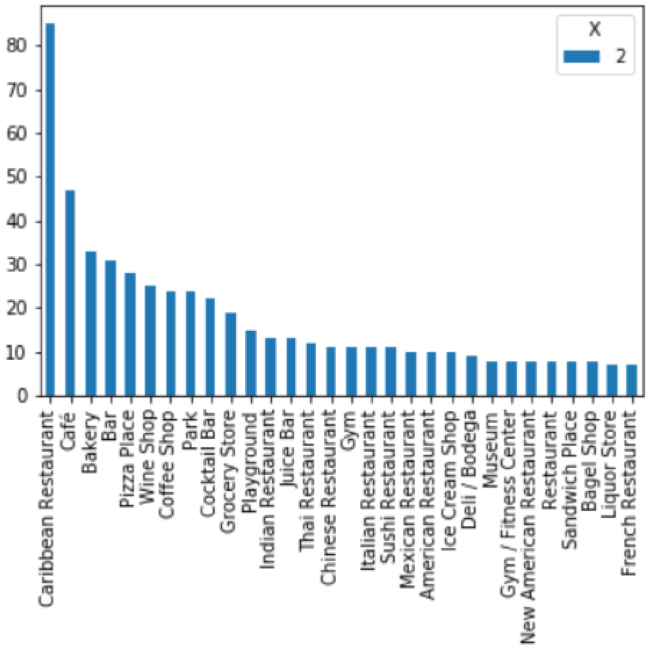
Lets take a deeper look at the most common venues in each cluster:



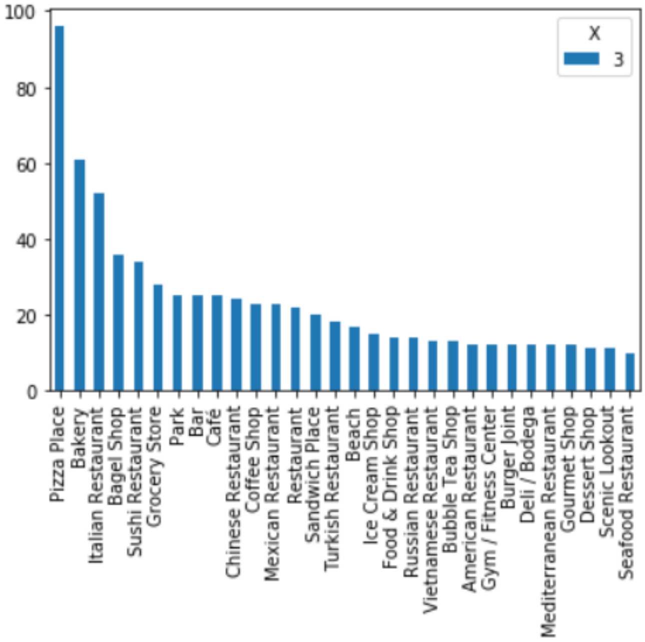
Cluster 0 seems to have a very high amount of bars, even more if you take into account that the fourth most common category is “Cocktail Bar”. There Pizza places and Italian restaurants are common as well, which is no surprise as they are amongst the most common categories in the whole dataset. Supermarkets are not as common as in the other clusters.



Cluster 1 seems to be extremely dominated by pizza places. Interesting is a high amount of pharmacies. There is a low amount of bars and cafes / coffee shops.



Cluster 2 has a high amount of Caribbean restaurants and a relatively high amount of cafes / coffee shops. There are also lots of playgrounds and wine shops.



Cluster 3 is dominated by pizza as well, but if you taker a closer look you find differences to cluster 1, e.g. lots of Caribbean restaurants, sushi restaurants and beaches.

5. Discussion & Conclusion

What information can we get from the results? First, it seems that frequencies of venue categories are distributed unevenly across Brooklyn – An even distribution would have resulted in cluster points being more mixed on the map. The borders of the clusters are extremely clear, which is a sign that the k-means is in retrospective a useful algorithm for the dataset used. The clear borders also indicate that there are large parts of Brooklyn that are kind of homogenous in terms of venue categories. An explanation for this could be the following: People from different countries tend to move to the same neighborhoods and there are large communities of e.g. Caribbean Americans or Asian Americans in Brooklyn (amongst many other groups). In neighborhoods with a high percentage of e.g. Caribbean American citizens it is more likely to have a high number of Caribbean restaurants. As the venue categories are strongly dominated by restaurants, this is resulting in a homogenous pattern over different zip code areas that are located to each other.

It seems that you should move to a neighborhood close to Manhattan (cluster 0) if you are into bars. This is no surprise, as neighborhoods like Williamsburg are located in this cluster. It seems to be more a place for people that like to party. You find categories like “bar”, “cocktail bar”, “beer bar”, “wine bar”, “wine shop” or “brewery” amongst the most popular ones. It is also good for people that like to be outdoors, as there are many parks. If you want to be near a supermarket or pharmacy, consider moving to another cluster. The cluster nears Queens (cluster 1) is the right place for people that like pizza or Caribbean food. It has If you are into cafes or bars, better avoid moving here. However, there are plenty of pharmacies and moderate amounts of convenience stores and clothing stores. The center cluster (cluster 2) has the advantage as well that it is near to all the other clusters. It is the perfect place for eating Caribbean food. There are also cafes / coffee shops and bars. It seems to be interesting for parents, as there are a lot of playgrounds and parks. Cluster 3 has a relatively high diversity of restaurant types (including sushi, Chinese, Turkish, Russian, Vietnamese), which seems to be interesting for foodies. A big plus is a high availability of grocery stores. It seems to be the only cluster with a moderate amount of beaches.

Based on these results, people that think about moving to Brooklyn can find the ideal neighborhood to move to, depending on what they prefer most. For future research, it should be considered to group all restaurants to one venue category or group venue categories in more extensive categories like “food”, “drinks” or “sports”.