**Neighborhood Clustering**

**Introduction:**

People become adapt to the neighborhood they are living in. They get accustomed to the way things work in that particular neighborhood and if they want to move to some other place majority of the people prefer the environment being similar to their original neighborhood.

1. **Business Problem:**

With housing schemes booming and a person looking for a new place to settle down, how will he know which scheme is best suited for him? In this project we will solve this problem using the clustering technique. We will compare different neighborhoods and match the neighborhood that best matches the current neighborhood of the individual. In this way he will have a better idea of which neighborhood he wants to switch to.

1. **Target Audience:**

The targeted audience of our application will be those people who want to move to a new place and want to keep the traditions and customs of their current neighborhood.

**Data:**

In this project we will compare the neighborhoods of Queens, New York. We will first download the new York data set from https://cocl.us/new\_york\_dataset.

As we know that a neighborhood is defined by a number of things including the facilities available in the surroundings. These facilities include

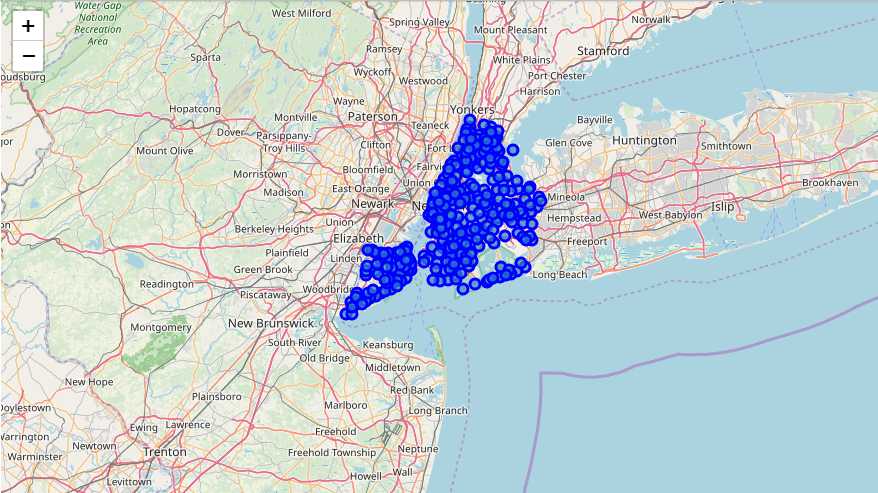
* Health Care
* Educational Institutions
* Shopping Malls
* Recreational Parks
* Restaurants

If the neighborhood has these facilities within walking distance than it is considered a good neighborhood.

In order to cluster different neighborhoods based on such facilities and venues we use the Foursquare API’s to get the relevant data and create a data frame for analysis. Foursquare will return the venues and facilities in each neighborhood and we can compare the neighborhoods based on the number, rating, visits to each of these facilities.

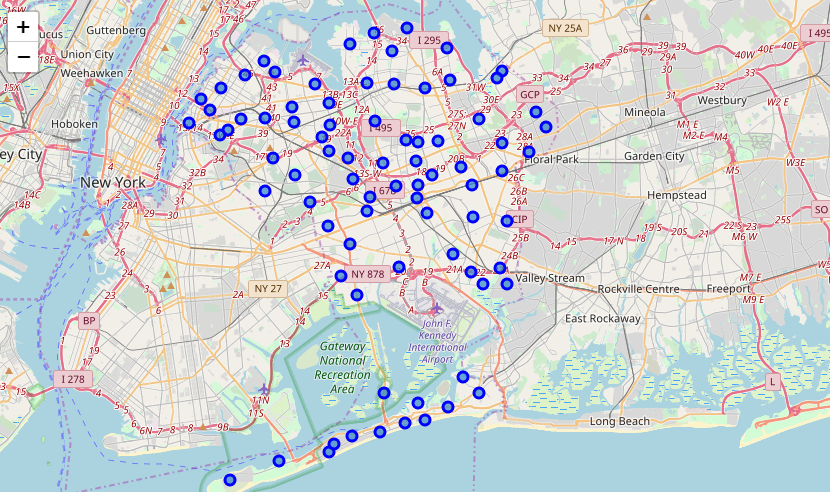
**Methodology:**

Once downloading the data, we will pre-process it. We will first create a panda’s data frame consisting of four columns namely Borough, Neighborhood, Latitude and Longitude. For each neighborhood in the dataset we will have a corresponding row in the data frame. After that we will plot the neighborhoods on the map using the Folium library. The initial neighborhoods come out to be like

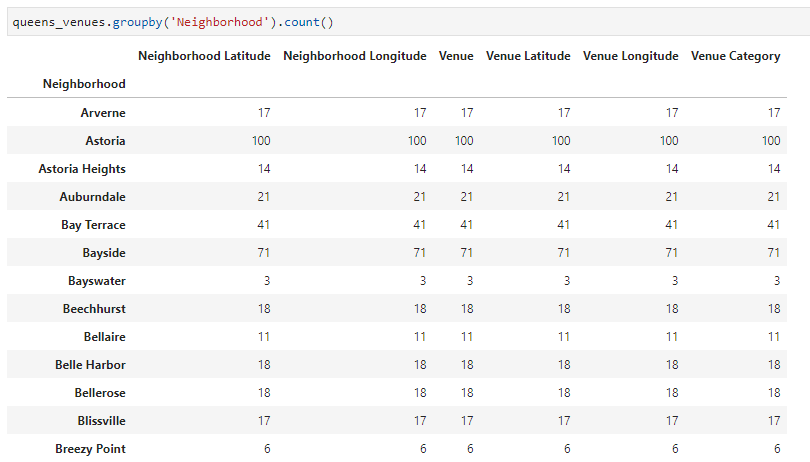


**Analyzing Queen’s Neighborhoods:**

From our panda’s data frame, we will extract the neighborhoods in Queens. After that we will retrieve the latitude and longitude value of Queens using the geolocator. We will now plot the neighborhoods of Queens which we are going to cluster later on.

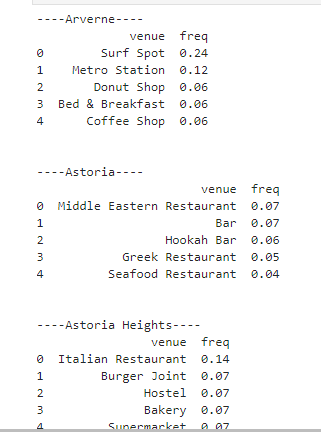


We will store all these neighborhoods in a data frame. Now its time to compare these neighborhoods and cluster them. For this we will use the Foursquare API’s. We will retrieve 100 venues in a radius of 500 for each neighborhood. The API returns a JSON file consisting of all the information of the venues. However, a lot of this information is going to be useless for our analysis hence we are first going to preprocess it. We will extract the required information namely Venue Name, Venue Location, Venue Category etc. Once we preprocess the data frame we will group the data frame based on the Neighborhood cell.



Our data frame looks a something like this. Now we will use the one hot encoding scheme to convert the values between 0-1 for more accurate analysis.

After that we will analyze the top 5 venues for each neighborhood and their occurrence. It will look something like

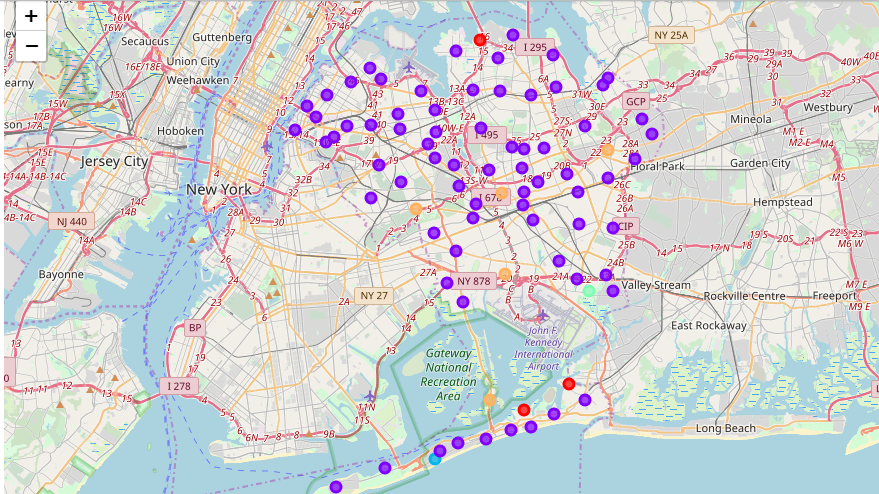


**K Clustering:**

Once we have retrieved the venues for each neighborhood, we will use the k neighbor clustering to cluster the neighborhoods. This will cluster the neighborhoods based on the venues and facilities. The optimal value of k comes out to be 5 hence we will divide the neighborhood into 5 clusters. In this way the use based on his current neighborhood will know if the neighborhood he is moving to is similar of dissimilar.

**Results:**

Clustering the neighborhoods using k = 5 gives us



This image shows us which neighborhoods in Queens are similar and which are dissimilar. Because of the scope of this project we chose to compare the neighborhoods of Queens only, else we can scale this to support neighborhoods of multiple boroughs at the same time.

**Conclusion:**

As we discussed above moving to a new neighborhood can be cumbersome especially when you are going to experience changes. With this solution a person can check out and opt for the neighborhoods that are similar to his existing neighborhood.