Hunting Best Neighborhood for Living in NYC

IBM Applied Data Science Capstone

By: Surya Gurung

Mar 2020

1. Background

New York City (NYC) is the most populous city in the United States (US) with over 8 million population spread over five boroughs, Brooklyn, Queens, Manhattan, Bronx, and Staten Island. It is well known as global capital of finance, media, and immigrant. NYC is also a global leader in entertainment, fashion, tourism, technology, education, arts, sports, politics, research, and many more industries. The World's largest stock exchanges operate from the famous financial center Wall Street. Many of the World's largest financial and media companies are based in NYC.

Time Square, NYC

In brief, NYC is center of opportunities to make the American dream come true. There is no other place like New York to start your living to pursue your American dream no matter where you come from. People around the World migrate here to raise and build better future for their family. Today, you can find the communities representing whole world, in fact more than 800 languages are spoken in NYC. But being the most populous and diverse city in US, it can be very intimidating to find right place to start living in NYC. It may seem easy to find the best place for your family's new life but people might easily get lost in the process and end up in wrong place. It requires careful and thorough research and decision to find the best place to live with happiness.

2. Business Problem

So, the problem every new city dweller or migrant faces is how to find the best neighborhood to rent a new apartment or to own a new house to start a happy and successful life in NYC? The main objective of this capstone is to develop a k-means clustering model using data science methodology and visualize the clusters in the NYC map by using NYC neighborhood

datasets along with Foursquare API to help the new city dwellers to select best possible NYC neighborhood quickly with greater precision. So, this project will greatly simplify the tedious process of hunting the new apartment or house in NYC.

3. Target Audience

Obviously, the new apartment hunters and home owners would be very interested to find the best neighborhoods in NYC to own a home or rent an apartment to start their NYC life successfully. Beside the new city dwellers, the real estate investors, financial institutions, and business investors are also going to be interested to find the best neighborhoods to own the real estate properties.

4. Data Requirements & Sources

To build the machine learning models, first we need the datasets representative of the business problem. For this project, we are going to use following datasets:

- We need dataset containing the neighborhoods name and its geographic locations. New York City neighborhood dataset contains list of neighborhood names and their geographic coordinates in five boroughs, Brooklyn, Bronx, Manhattan, Queens, and Staten Island. This dataset defines the scope of this project which is the neighborhoods in five boroughs of NYC.
- The Foursquare venue dataset doesn't contain the NYC schools information. Because the school information is one of the very important features when searching for apartment or house, I am going to use following NYC schools data sources:
    - NYC school locations dataset contains list of Elementary, K-8, High school, Junior High-Intermediate-Middle, Secondary School, K-12 all grade, and Early Childhood in five boroughs of NYC.
    - The performance of middle and elementary schools contains the average ratings and score of students and schools.
    - The high schools performance data contains the average rating and score of the student and schools
- The Foursquare venue dataset doesn't contain the crime data but it is critical information when making decision for living in new neighborhood. The NYC crime dataset contains all valid felony, misdemeanor, and violation crimes reported the New York City Police Department in 2019 and the geographic location of the crimes. This dataset doesn't contain the neighborhood name. So, I am going to add neighborhood

names to the NYC crime dataset based on the crime's geographic location and its closest neighborhood.

- We are also going to use the venue data returned by the Foursquare API based on the NYC neighborhoods dataset. This data will be used to perform clustering of the NYC neighborhoods based on the most important venues for living and to raise kids. By clustering the neighborhoods based on the most important venues, it will help to make the neighborhood selection process based on the criterions of individual dweller.

5. Data Collection & Preprocessing:

All of the datasets are retrieved using either the requests python library or pandas read_csv function. The important features are extracted and converted into a pandas dataframe. The NYC neighborhoods dataset is the main dataframe and all other dataframe are going to merge to it. For the consistency, the feature names are going to be lower case with underscore for space.

Four features, *borough, latitude, longitude, and neighborhood* are extracted from the NYC neighborhood dataset and converted into a pandas dataframe. There are 299 rows each representing one neighborhood in one of the five boroughs of NYC. As a note, four neighborhoods in different boroughs share same name and those are Sunnyside, Bay Terrace, Murray Hill, and Chelsea. The latitude and longitude values are going to be used as inputs for Foursquare API to retrieve the venues data around the neighborhoods.

Since Foursquare API doesn't return NYC schools dataset, I created a dataframe by extracting the school number, school name, school type, latitude, and longitude from the NYC schools location data source and neighborhood of the schools are extracted from the NYC neighborhood dataset by using geopy's distance function. The student ratings and scores are extracted from the NYC schools quality review dataset sources and then merged the dataframe with the school dataframe. The tidy the school dataframe contains 1638 rows with eight features, *dbn, school_name, school_type, student_rating, student_score, latitude, longitude, and neighborhood*. Then, a dataframe is created with school types as features with number of each type as its values using groupby() and unstack() functions. Similarly, another dataframe is created with student ratings as features with number of each rating type as its values. These two new dataframes are merged with the NYC neighborhood dataset.

Similarly, the NYC crime dataset is retrieved as json file and converted into a dataframe with *crime_id, borough, crime_level, latitude, and longitude* as features but doesn't contain neighborhood name. By using geopy's distance function, I calculated the distance between crime locations and the NYC neighborhood geographic locations to find out the neighborhood

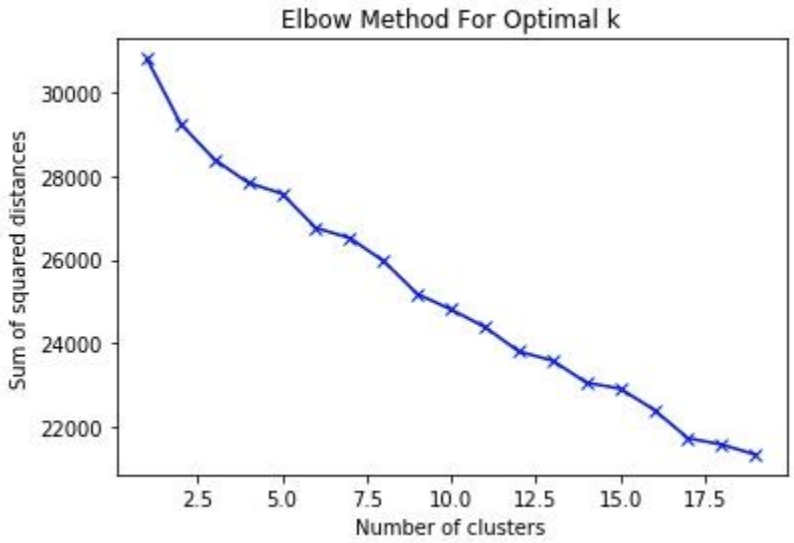
names of the crimes. Now, a dataframe with crime_level as features and number of crime level as value is created. Similarly, this dataframe is also merged with the NYC neighborhood dataset.

So far, we have added the school and crime data to the NYC neighborhood dataset. The final piece of data we need add to the main dataset is the Foursquare venue dataset. We are going to use the Foursquare API to get the venues around 1000 meters radius from each neighborhood geographic locations. This API returns many different venues from 10 major categories but we are not going to need all kind of venue data. I retrieved the venues from the categories Shop & Service, Professional & Other Places, Travel & Transport, Arts & Entertainment, and Outdoors & Recreation. Obviously, you can select more types of venue than I did. I did further selection of the venue types from these categories. Finally, a dataframe is created with venue types as features and number of each venue type in each neighborhood as value. Then it is merged with NYC neighborhood dataset.

The variances of all features were calculated. Since the variances are varied widely, data standardization method is applied to the NYC neighborhood dataset and created the training dataset.

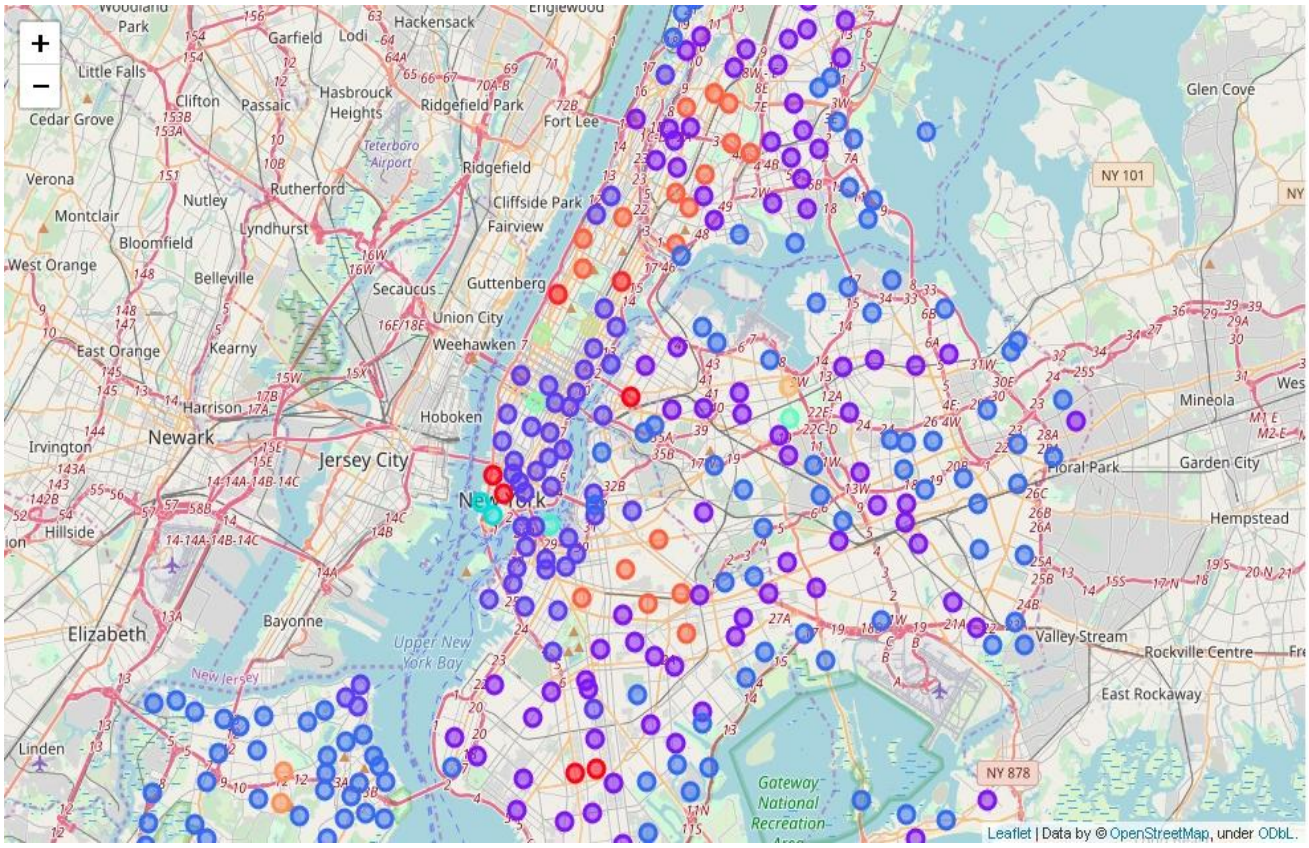
6. Modeling:

For this project I am going to perform k-means clustering but first, I did the hyperparameter tuning of the k-means clustering by using the Elbow method.



I performed 19 k-means clustering with the parameter n_clusters ranging from 1 to 19 and plotted the parameter n_clusters Vs Sum of Squared Distances. By observing the plot, I picked 16 as the optimal cluster number. But, with many trials, I also found that 8 numbers of clusters seems to give very good clusters of neighborhoods.

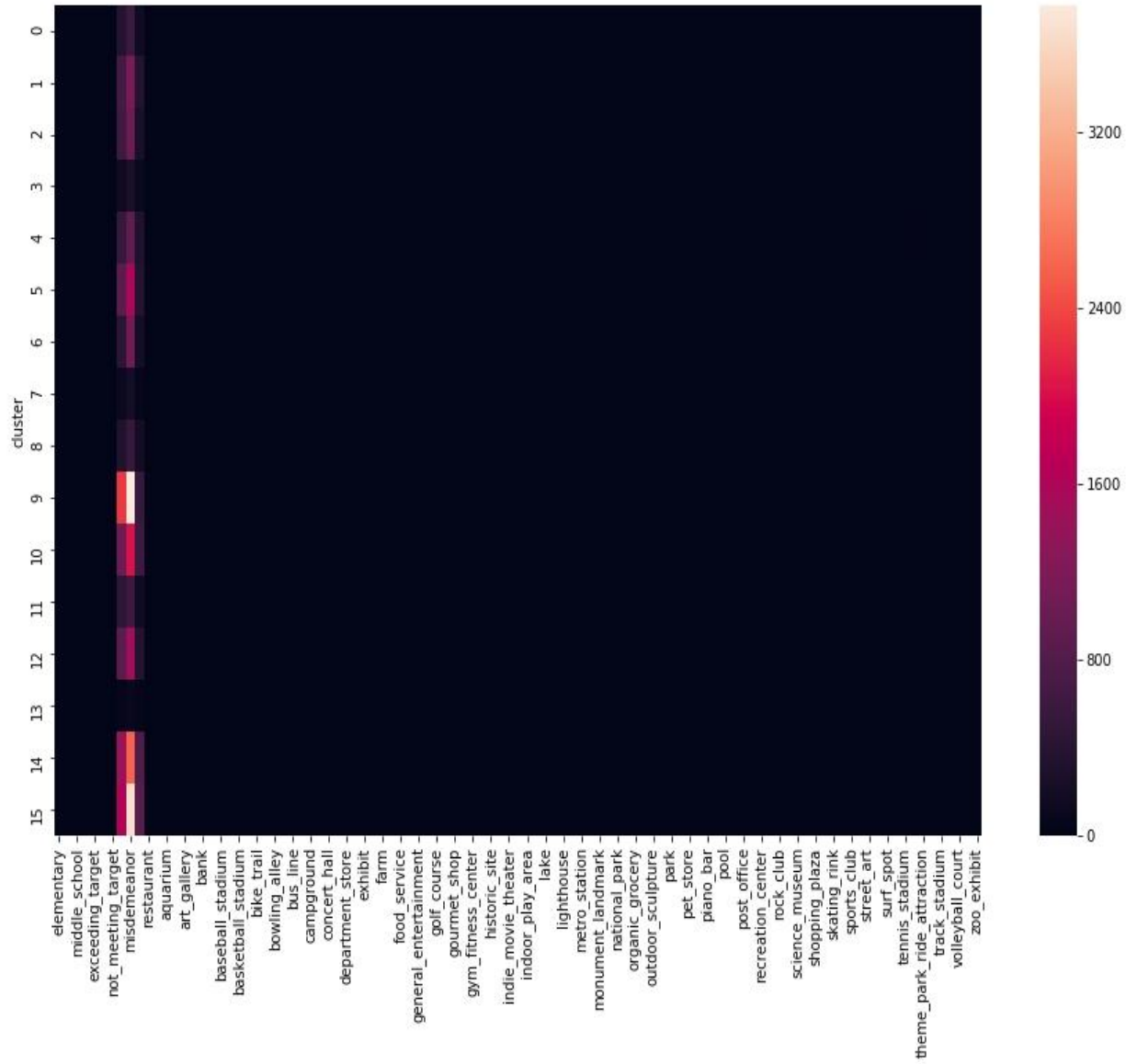
So, the optimal k-means clustering is performed with 16 clusters and the cluster labels are added to the NYC neighborhood dataset and now, it has 299 rows and 108 columns. Finally, the clusters of neighborhoods are visualized on Folium map of NYC.



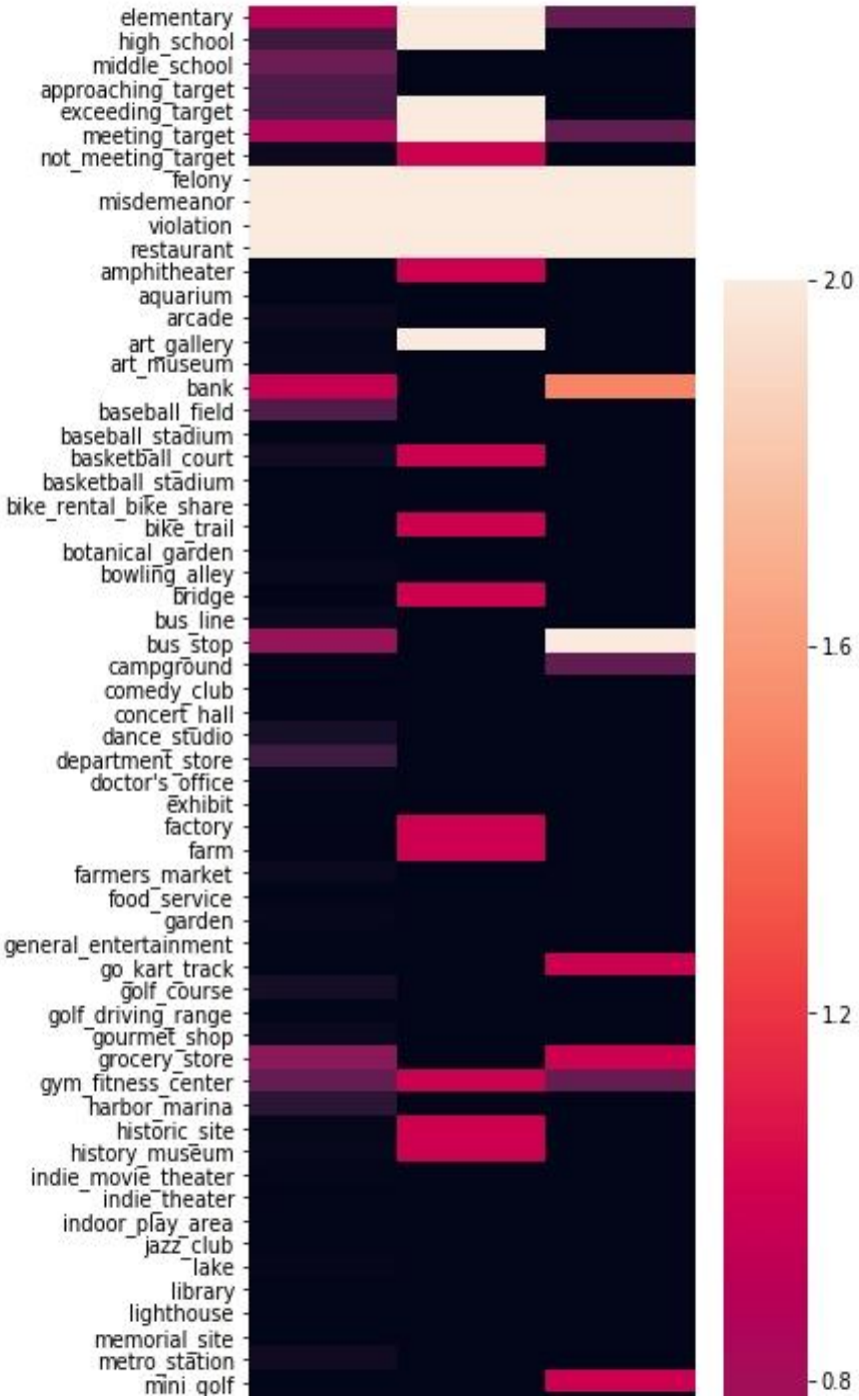
Clusters of NYC neighborhood.

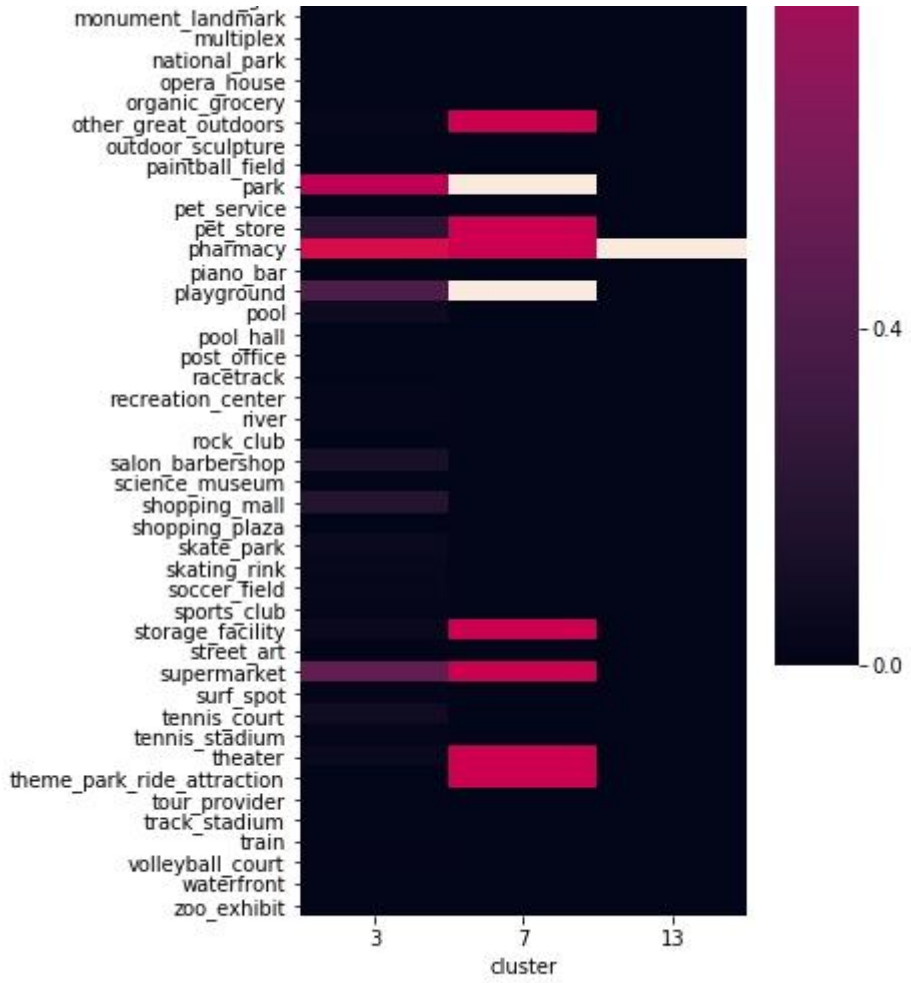
7. Analysis:

I created a dataframe by grouping the NYC neighborhood dataset by 'cluster' level and calculated the mean values of all features in each neighborhood clusters. Then I visualized the dataframe using the seaborn heatmap. From the observation of the heatmap, we can tell clusters 3, 7 and 13 have least average crimes. The clusters 0, 8, and 11 have little more crimes than the clusters with least crimes. But the clusters 9, 10, 12, 14, and 15 have most crimes. Least crimes would be obvious choice for anyone who would be looking to settle their life in new neighborhood. So, I would like to narrow down the clusters to cluster 3, 7 and 13 to further analyze the clusters visually.

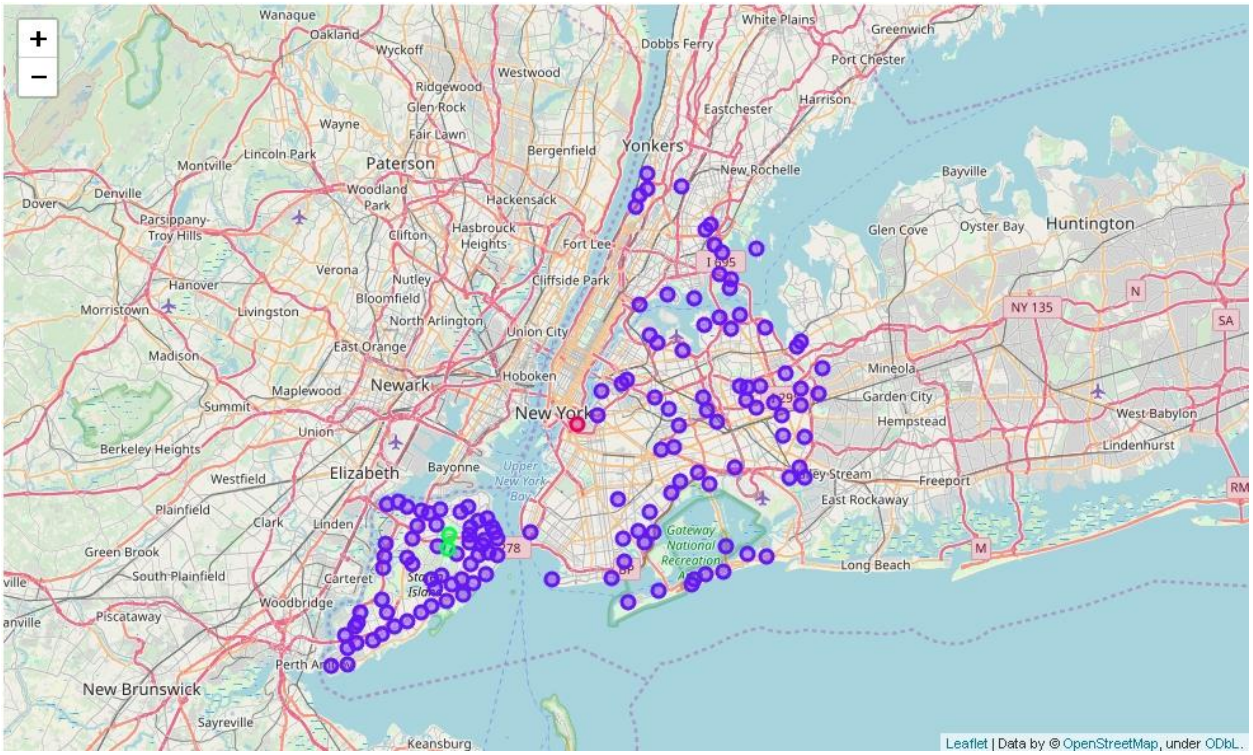


I filtered the clusters 3, 7 and 13 from the dataframe and then grouped it by 'cluster' labels and calculated the mean values of the features. Now, it is visualized using the heatmap. By observing the heatmap (it is divided into two vertical pieces to fit properly), we can tell the cluster 7 has either very bright or very dark colors. Cluster 3 has more different shades of colors. Cluster 13 has more very dark color. So, the cluster 7 is very good choice for living with less crimes, great elementary and high schools, lots of restaurants, art galleries, amphitheater, bike trail, gym, museum, park, playground and close proximity to city. The cluster 3 has more shades of colors because it offers more variety of features than cluster 7. Since, the cluster 13 has more very dark color it offer far less features than other two clusters.


elementary
high_school
middle_school
approaching_target
exceeding_target
meeting_target
not_meeting_target
felony
misdemeanor
violation
restaurant
amphitheater
aquarium
arcade
art_gallery
art_museum
bank
baseball_field
baseball_stadium
basketball_court
basketball_stadium
bike_rental_bike_share
bike_trail
botanical_garden
bowling_alley
bridge
bus_line
bus_stop
campground
comedy_club
concert_hall
dance_studio
department_store
doctor's_office
exhibit
factory
farm
farmers_market
food_service
garden
general_entertainment
go_kart_track
golf_course
golf_driving_range
gourmet_shop
grocery_store
gym_fitness_center
harbor_marina
historic_site
history_museum
indie_movie_theater
indie_theater
indoor_play_area
jazz_club
lake
library
lighthouse
memorial_site
metro_station
mini_golf
2.0
1.6
1.2
0.8



Finally, I plotted the neighborhoods in the clusters 3, 7 and 13 on the folium map. By looking at this map, we can tell most of the neighborhood in cluster 3 is located suburban area. One neighborhood in cluster 7 is at the edge of east river across the downtown Manhattan. The neighborhoods in cluster 13 are located in Staten Island.



8. Conclusion:

By observing the visualization of the clusters on NYC map, we can tell that clustering of neighborhoods looks pretty consistence with geographic, urban and suburban characteristics. The more suburban neighborhoods are clustered into cluster-3. Most of the city neighborhoods are clustered into cluster-2. Similarly, neighborhoods with high crimes are clustered into cluster-14 or cluster-15. Cluster-1 also looks very interesting. It contains most of densely populated residential neighborhoods located in between the city (Manhattan) and more suburban neighborhood.

From the first heatmap, we can tell clusters 3, 7 and 13 have least average crimes. Clusters 0, 8, and 11 looks to have little more crimes than the clusters 3,7, and 13. The clusters 1, 2 and 4 have more crimes than 0, 8 and 11. Clusters 9, 10, 12, 14, and 15 have most crimes. First, I am narrowing down the selection of neighborhoods based on the average crime numbers. So, I am focusing on the clusters with least average crimes only.

From the observation of heatmap of clusters 3, 7 and 13, we can tell cluster 7 is a great neighborhood for living but this cluster got only one neighborhood which is located **Vinegar Hill in Brooklyn**, just right across the east river from the down town, Manhattan. But cluster 3 is also looks a great collection of 136 neighborhoods. So, lots of choices. Looking at the second NYC map, we can tell this cluster contains most of the suburban neighborhoods and some urban neighborhoods. These neighborhoods got plenty of good schools, lots of natures, amenities, outdoors activities, variety of services and lesser average crimes.

So, based on this modeling, I would highly recommend the **Vinegar Hill** neighborhood, if you prefer to live in very close proximity of the city. But if you prefer to live little farther from the city, then the neighborhoods in cluster 3 are very good choice. You can view the neighborhood locations on the second NYC map in purple color and it would help you to make a rough estimation of your choice. Then take a look at the cluster_3 dataframe for more detail features of the neighborhoods.