Battle of Neighborhoods

Looking for new house to buy in New York

IBM Applied Data Science Capstone Project

Mohammad Charsooghi

January 2021

1. Introduction

1.1 Background

Moving to new cities and finding new house, especially in big and crowded cities, can be challenging. It would be even more challenging, when you don't want just buy a new property but both the condition of the neighborhood and the available services such as health care or schools are important for you. Moreover, maybe you like to take investment into account so you prefer invest in a property that you can sell easy in the future or even earn some money from it.

If the city is new to you, probably you don't know living in which neighborhood is the most fit with your desires and needs. And deciding would be very difficult. How we can fit the house that fit our desires. Data science can help answer to this question and in the continue we will develop a tools to help people find similar neighborhoods in a region based on their needs (link to Jupyter Notebook on my Github).

1.2 Problem statement

In this research we try to help people who want to buy a new house in New York to find the best neighborhood which is the most fit with their desires. The question is that in which neighborhood in New York should one with the following desires buy a new house.

Features we like to consider:

- Size of the building
- Price of the building
- Value growth in a few past years
- Available health care facilities, and Schools
- Available high-grade restaurant and healthy stores
- Safety and less likely of crime

These are just an imaginary assumption to conduct this project. One can select and analyze their owns.

Interests

Although we focused on a special city and also some features which may differs from one person to another, but it could easily be extended to other locations and tuned with others needs. It would be a great supporting tool to narrow down peoples choices when looking for the new house. This tool also give the real state agents lots of insights abouth different features of a property and similar locations in the city, which inturn help them give better advices to their costumers.

2. Data

2.1 Data sources and description

- Data of sales price of buildings in New York between 2016 and 2019 is used to extract price and size of the buildings in different neighborhoods link to data.
- **Forsquare API** is used to extract top venues in each neighborhood.
- A .geojson file is used for the geospatial information of the neighborhoods link to data.
- New York demographic info link to data.
- New York City Restaurant Inspection Results link to data.
- Schools point locations link to data and link to iZone school list.
- NYPD Complaint Data Historic link to data.
- Precinct information in New York link to data.
- Recognized Shop Healthy Stores link to data

I put a copy of .csv data files in my github repository beside this Jupyter Notebook, for easy access to data. However, I changed the name for feasibility and also for space and memory considerations, unnecessary columns (features) are removed before uploading. You can still access to full data through the provided links.

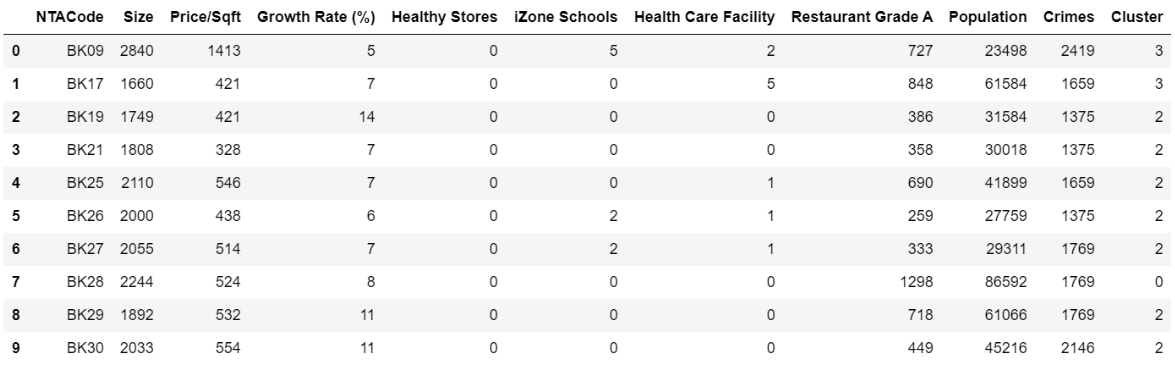
2.2 Data description

Each data on the source contains lots of information and many features. For this study we only focus on the following features:

- Buildings: Gross squre feet, sale price, year sold, nat name
- Forsquare API: 100 venues in the radius of 500 meters of the center of the neighborhood
- Restaurants: Grade and Zip code
- Health care facility: made by combining adult and child health care facility in different neighborhoods
- Schools: schools locations and also number of iZone schools in the area
- I used data of sectors of precincts in New York (link) to extract which neighborhoods each covers.
- I used data of zip code (link), google map, and geospatial data of neighborhoods (link) to conncet zip code regions to neighborhoods.

These data are downloaded from different sources. There are some missing values and also each of source has its own format. In the continue, we will load datafiles select proper features and convert them to desired format. After cleaning the dataset we will gather all data as a single data frame for further analysis and modeling.

The first top ten rows of the final data frame including desired features is shown:

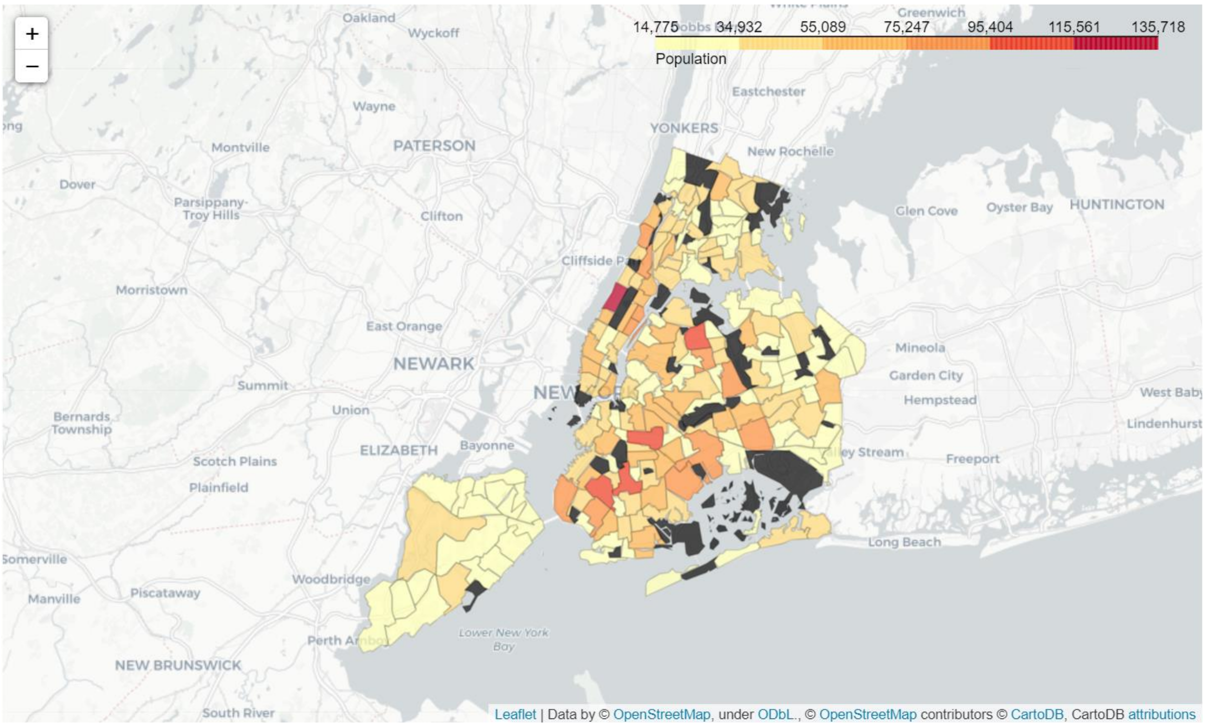


	NTACode	Size	Price/Sqft	Growth Rate (%)	Healthy Stores	iZone Schools	Health Care Facility	Restaurant Grade A	Population	Crimes	Cluster
0	BK09	2840	1413	5	0	5	2	727	23498	2419	3
1	BK17	1660	421	7	0	0	5	848	61584	1659	3
2	BK19	1749	421	14	0	0	0	386	31584	1375	2
3	BK21	1808	328	7	0	0	0	358	30018	1375	2
4	BK25	2110	546	7	0	0	1	690	41899	1659	2
5	BK26	2000	438	6	0	2	1	259	27759	1375	2
6	BK27	2055	514	7	0	2	1	333	29311	1769	2
7	BK28	2244	524	8	0	0	0	1298	86592	1769	0
8	BK29	1892	532	11	0	0	0	718	61066	1769	2
9	BK30	2033	554	11	0	0	0	449	45216	2146	2

3. Methodology and Exploratory Data Analysis

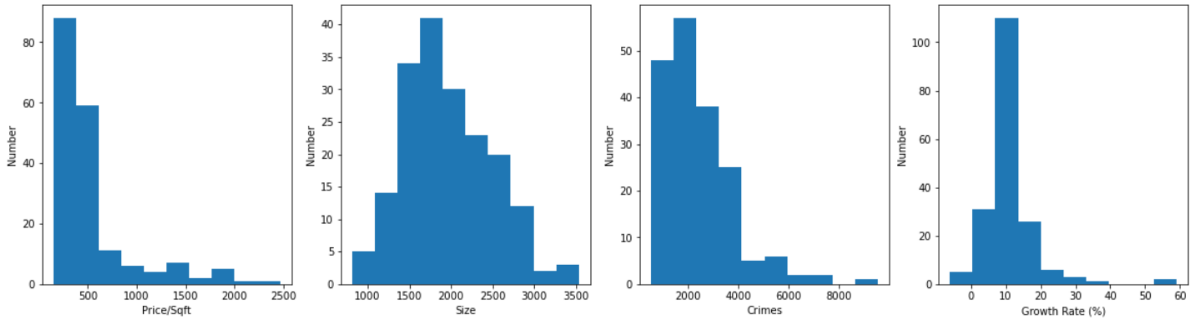
3.1 Visualize Neighborhood Segments in New York City

In this section we will explore data to find out useful information inside them. For the first step it is good to visualize the distribution of neighborhoods and their populations using Folium map.



3.2 Feature Analysis and their Effect on Sale Price

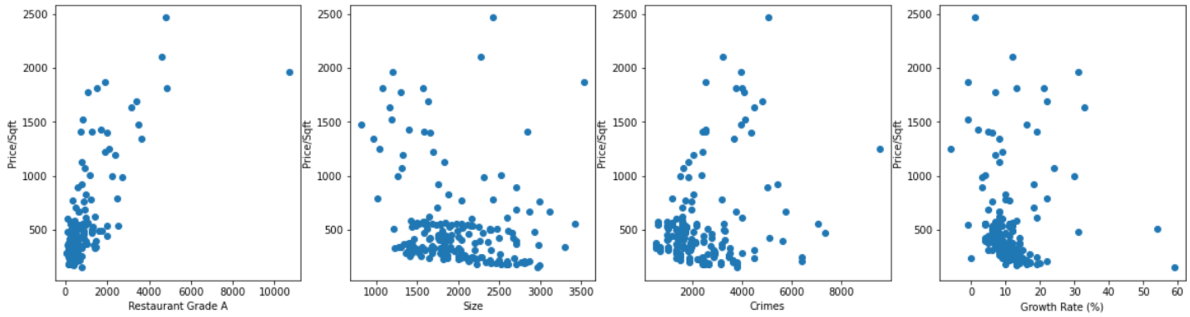
In this section we will plot histogram of some parameters. It is good to know about the range and distribution of the values.



As shown in the above histograms:

- Mostly, the price per square feet is less than 1000 $.
- Most of the buildings has the size around 1500-2500 square feet.
- In average the price of buildings increased by 10% over the past few years, annually.

The following figure shows scatter plot of some selected features and the price:

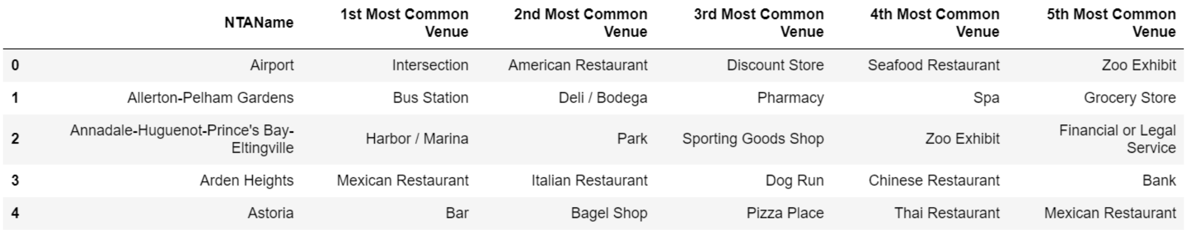


Restaurant with grade A has a positive correlation with the price but for the others there isn't any strong correlation.

3.3 Analysis Top Venues in Each Neighborhood

In the continue based on the Foursquare location data API we will extract 100 venues in the range of 500 of each neighborhood and will categorize them and extract top five venue category for each neighborhood.

The following table is just the top 5 rows of the most five common venues in different neighborhoods. For complete results, one can refers to Jupyter Notebook.

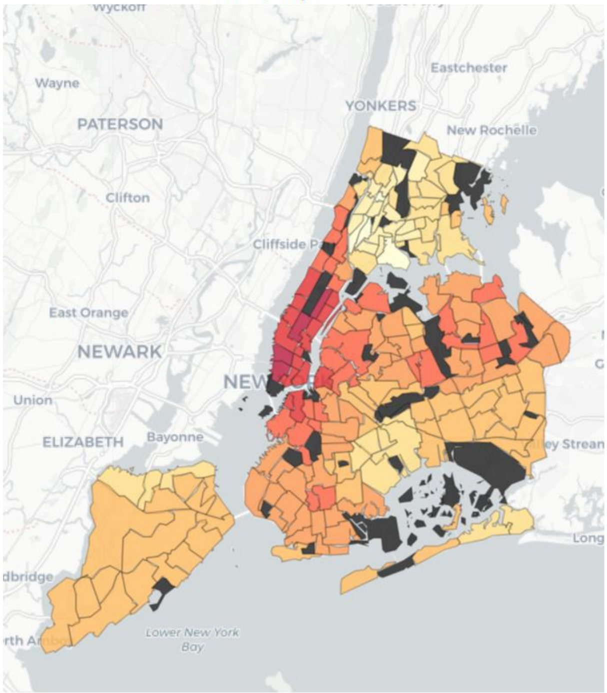


	NTAName	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Airport	Intersection	American Restaurant	Discount Store	Seafood Restaurant	Zoo Exhibit
1	Allerton-Pelham Gardens	Bus Station	Deli / Bodega	Pharmacy	Spa	Grocery Store
2	Annadale-Huguenot-Prince's Bay-Eltingville	Harbor / Marina	Park	Sporting Goods Shop	Zoo Exhibit	Financial or Legal Service
3	Arden Heights	Mexican Restaurant	Italian Restaurant	Dog Run	Chinese Restaurant	Bank
4	Astoria	Bar	Bagel Shop	Pizza Place	Thai Restaurant	Mexican Restaurant

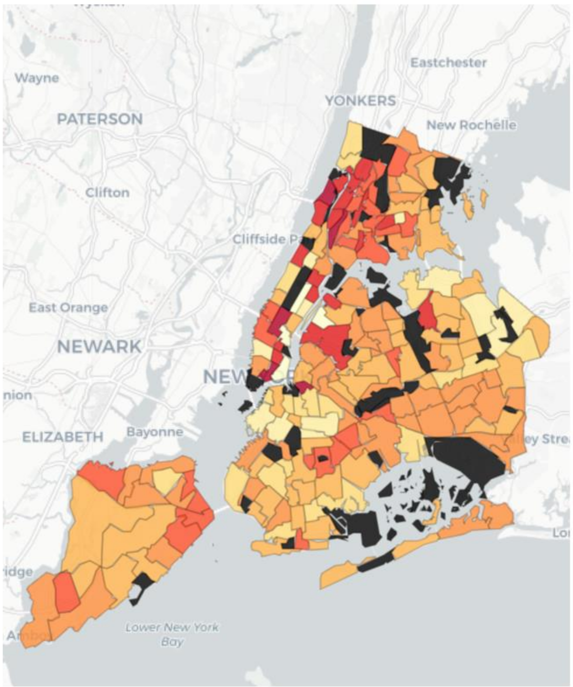
3.4 Map of Neighborhood's Features Distribution

The following snapshots show the distribution of different features in the neighborhoods of New York city.

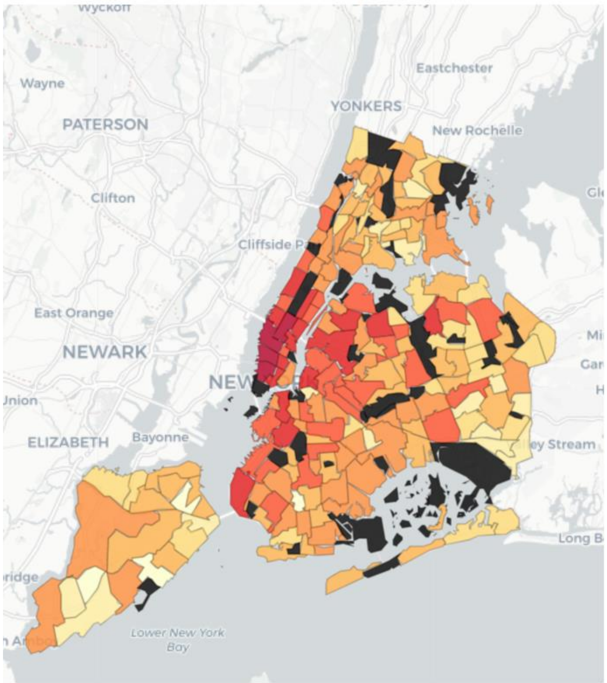
Price / Sqft



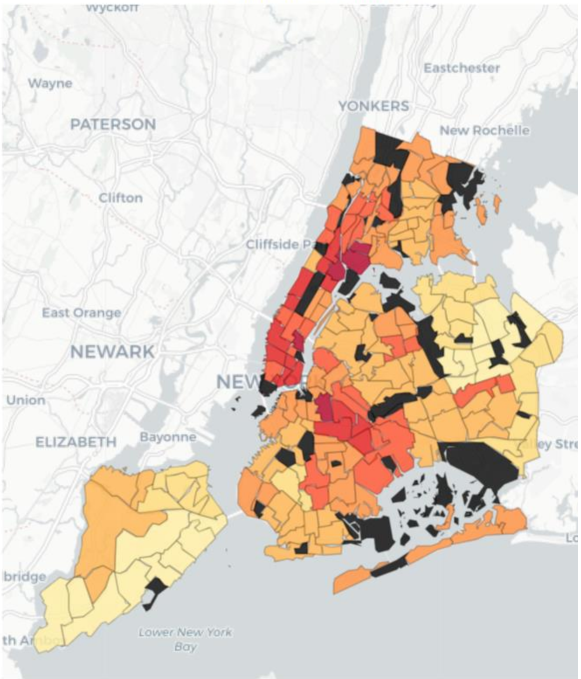
Value Growth in few past years



Restaurant with grade A



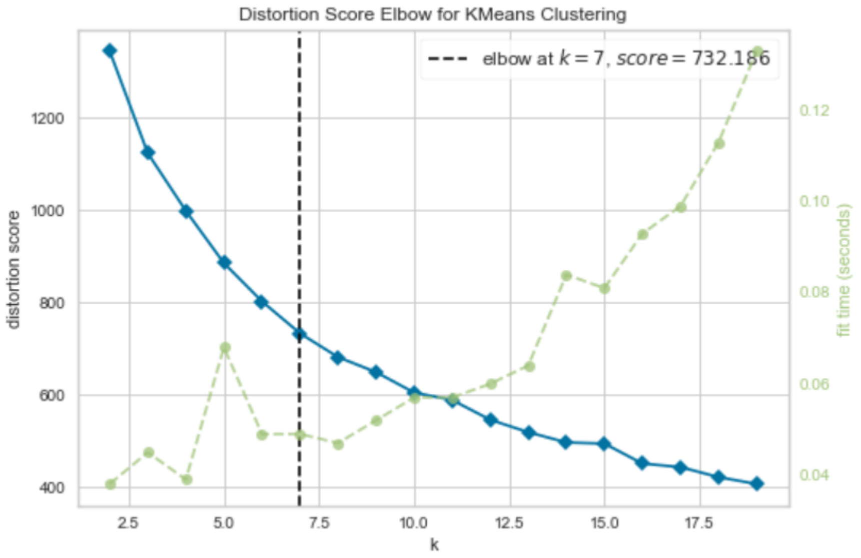
Number of Crimes



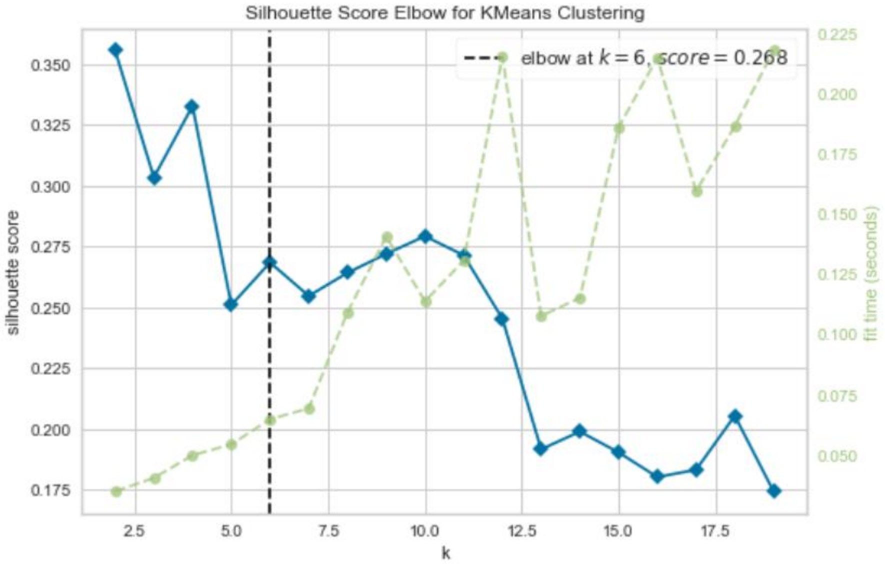
3.5 Clustering of Neighborhoods

By clustering the neighborhoods we can find similar neighborhoods to each other based on desired features. We use the K-means clustering method for clustering. The main question here is the optimized number of final clusters. To find out this value two different methods were used and the average of their suggestions are applied:

Elbow Method for K means



Silhouette Score for K means



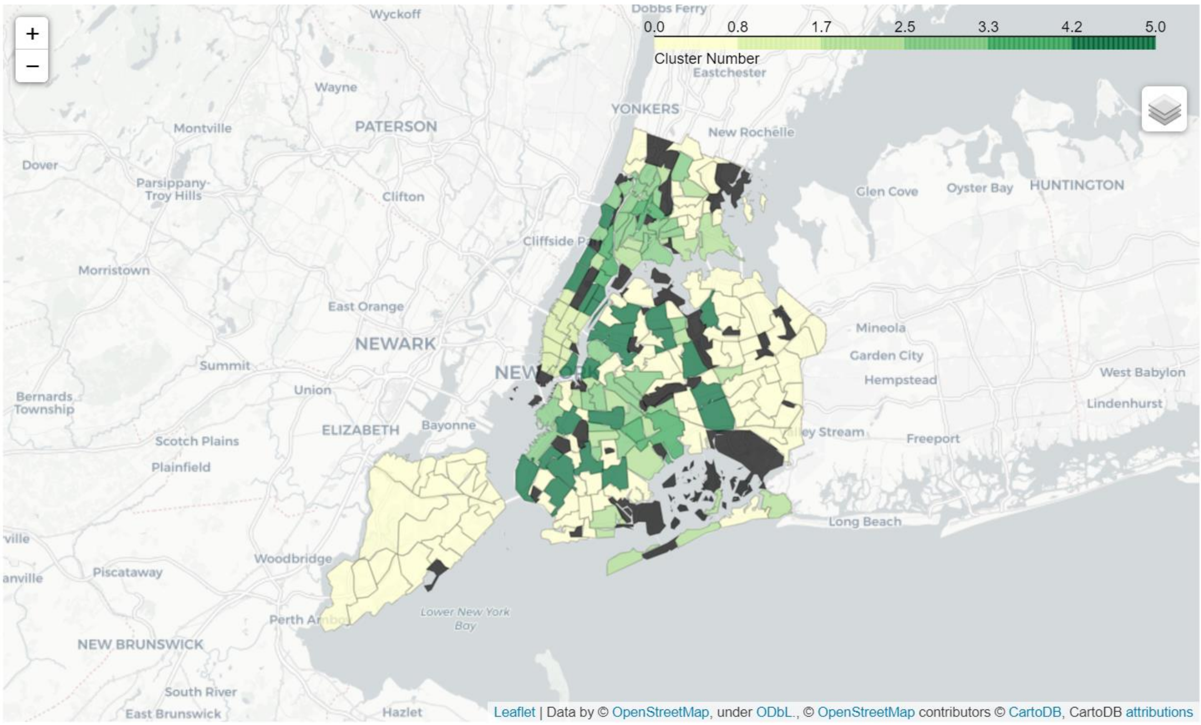
So, we considered 7 final clusters and will continue to use Scikit Learn library in Python for clustering. Clustering is done with 12 different initial cluster centroids selection.

4. Results and Discussions

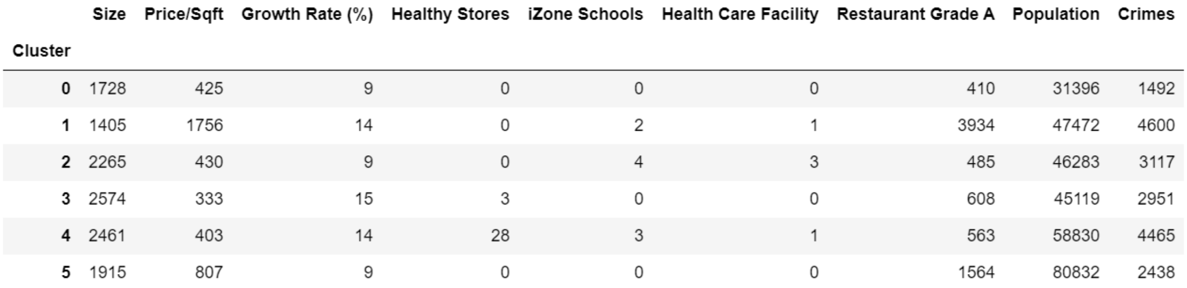
In this section we will superimpose the neighborhoods in different clusters with different colors on the map of New York.

4.1 Neighborhood Clusters on the Map

Following figure shows different cluster of neighborhoods in different colors.



Summary table of features for each cluster group is as follows:



Cluster	Size	Price/Sqft	Growth Rate (%)	Healthy Stores	iZone Schools	Health Care Facility	Restaurant Grade A	Population	Crimes
0	1728	425	9	0	0	0	410	31396	1492
1	1405	1756	14	0	2	1	3934	47472	4600
2	2265	430	9	0	4	3	485	46283	3117
3	2574	333	15	3	0	0	608	45119	2951
4	2461	403	14	28	3	1	563	58830	4465
5	1915	807	9	0	0	0	1564	80832	2438

Based on this table and previous observations:

- Regions 1, 3, and 4 which corresponds to Manhattan, Bronx and Staten Island had the most growth amongst the other. However, one can exactly extract the neighborhoods in these areas using this information.
- Between these regions average price per square feet in Manhattan is higher respect the others.
- Cluster 4 (some regions of Bronx and also Brooklyn) also has good nearby healthy stores, schools and health care facility.
- Based on scatter plots, number of high-grade restaurants has positive correlation with the price of buildings on that region.
- The size of sold buildings in Bronx are higher than other regions, in average.
- The most expensive region is Manhattan area.

5. Conclusion

In conclusion with the help of data science we developed a program to segment different neighborhoods in the city based on desired features. Here in this study we used a wide range of available data of New York city to help people who wants to buy a new house in this region.

Especially, we were interested in sales price of sold buildings during the years 2016 and 2019. We focused and residential buildings. We explore different features such as size of the property, the value growth over the past years, available health care facilities, schools, healthy stores and high grade restaurants near by. We also analyzed different venues in each neighborhood using Foursqure location data. We extracted top five venues in each neighborhood which one can take into account when buying a new house.

Although we focused on a special city and also some features which may differs from one person to another, but it could easily be extended to other locations and tuned with others needs. It would be a great supporting tool to narrow down people's choices when looking for the new house. This tool also give the real state agents lots of insights about different features of a property and similar locations in the city, which in turn help them give better advices to their costumers.