Clustering of New York City Neighborhoods Based on Housing Price, Inventory and Venues nearby

Introduction

Millennials are reaching the ages of getting married and having children, which prompts them to buy their first homes. However, there are so many factors to take into account when it comes to picking a suitable neighborhood, such as housing prices, pre-k to 12 schools, etc.

Therefore, this project will use location data well as real estate data in New York City to recommend suitable neighborhoods for purchasing homes for people, especially couples with the intention to raise their kids in NYC. In addition, these neighborhoods can also be of interest for real estate companies who would like to develop new projects.

Data

Location data from Foursquare API as well as real estate data from StreetEasy will be utilized for the analysis.

Real estate data, namely median sales price as well as inventory of houses in each neighborhood in New York City will be combined with location data that home purchasers are interested in, including categories such as school, Doctor's Office, Outdoors & Recreation, etc.

Location Data

Coordinates of NYC neighborhoods will be obtained from IBM dataset, and up to 100 venues within 1km radius from each neighborhood coordinate in the following categories will be selected:

- Food & Drink Shop
- School
- Medical Center
- Outdoors & Recreation
- Residence

Real Estate Data

Average total inventory and median price of houses for sale in the first quarter of 2020 in each neighborhood of NYC will be obtained from StreeEasy as csv files, which will be employed as features in addition to the data of various venues in each neighborhood.

Data Cleaning

Both coordinates of NYC neighborhoods and real estate data will be imported as csv files. Among all the tables imported, housing price is the only one that has missing values. After computing the mean of Q1 2020 housing prices, I noticed that there are 60 neighborhoods which do not have available data. Therefore, I first used Q4 2019 data to fill the missing values, which leaves 47 neighborhoods still missing values. The remaining missing values for the neighborhoods are filled in with the average Q1 2020 housing price in the borough that they belong to respectively.

Afterwards, real estate data as well as location data are merged into one data frame, which reveals the fact that the real estate data do not include relevant information for neighborhoods in Staten Island. Therefore, for consistency, the current study will exclude neighborhoods in Staten Island for the analysis.

The final data frame includes 7 columns: inventory, housing price, food and drink venues, schools, medical facilities, outdoor and recreation space, residence.

Methodology

To successfully divide neighborhoods in NYC into distinct clusters, I will employ K-Means machine learning algorithm to perform the segmentation. I chose K-Means as it is a partition-based clustering algorithm and divides data into non-overlapping subsets.

Before using the K-Means algorithm, normalization will be conducted as the features have different units.

After normalization, all the features will be fed to K-Means clustering algorithm for neighborhood segmentation, which will divide neighborhoods into k groups to facilitate the process of picking a suitable neighborhood to purchase a home.

The optimal k is selected through comparing the Calinski-Harabasz index, also known as the Variance Ratio Criterion. It is suitable for the situation when the ground truth labels are not known. A higher Calinski-Harabasz score relates to a model with better defined clusters.

The index is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters (where dispersion is defined as the sum of distances squared).

As shown in the figure 1 below, the optimal k is 5 as k=2 will only define 2 clusters, which provides limited information. Therefore, n_clusters=5 will be passed as one of the parameters.

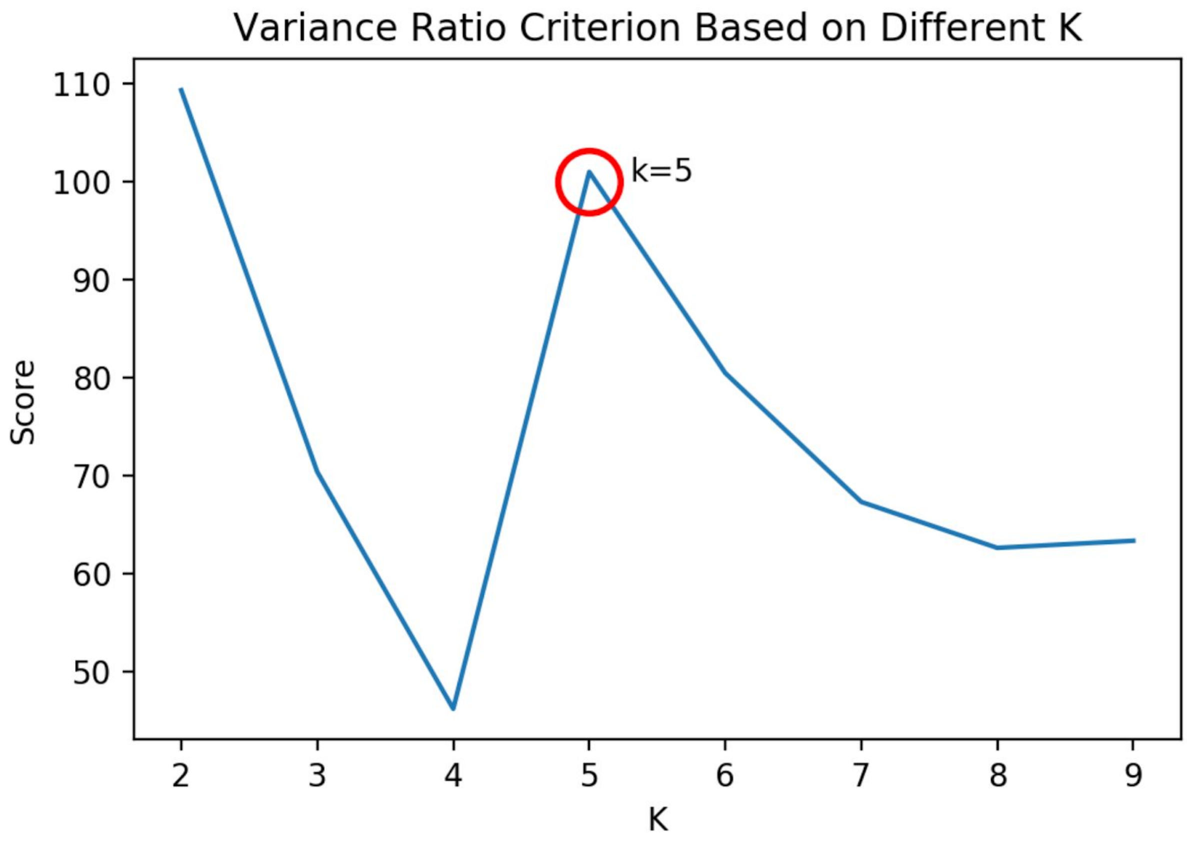

Figure 1

Results

After running the K-Means algorithm, cluster labels ranging from 0 to 4 are generated for each row in the dataframe. Combining the cluster labels with the latitudes and longitudes of NYC neighborhoods, I plot the following map using folium.

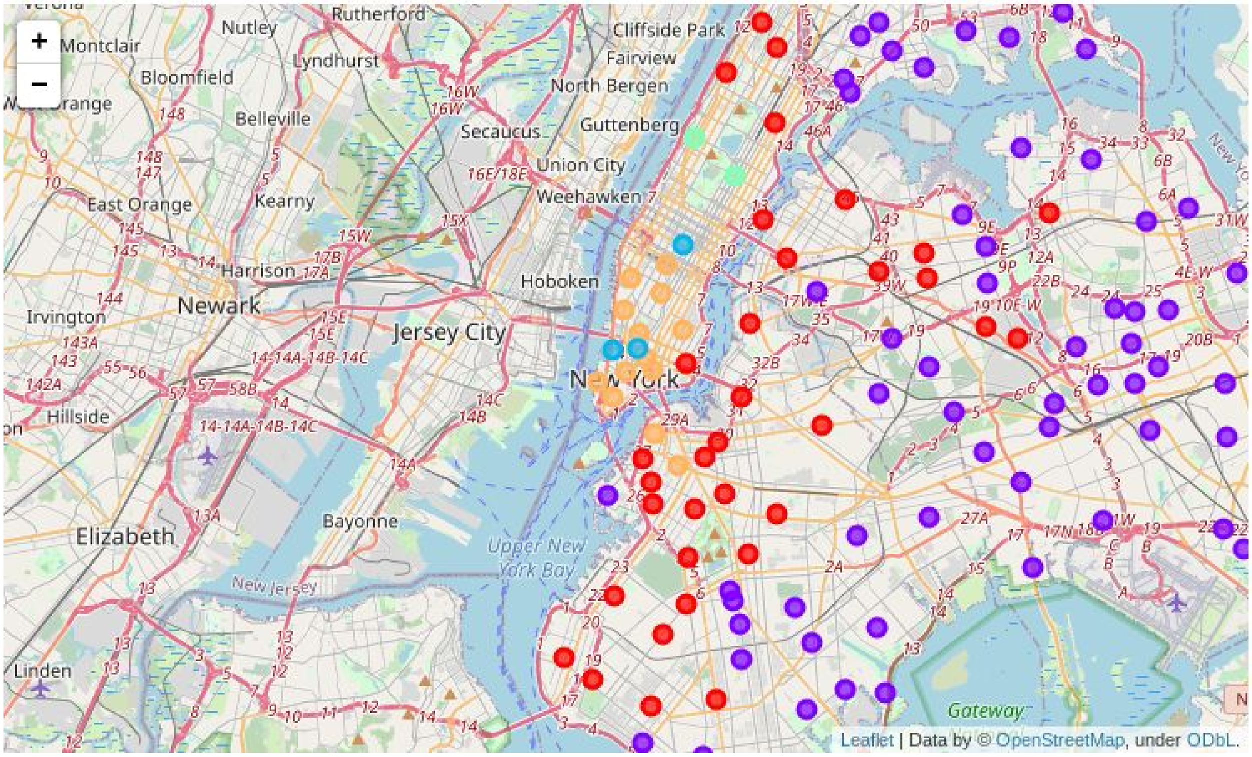

Figure 2

The basic characteristics of the 5 clusters can be summarized as below:

Cluster 1 (Red):

These neighborhoods tend to be near the center of Manhattan, with medium housing prices around $800,000 as well as a variety of facilities available. The commute time is on average 20-30 minutes to midtown. They are suitable for people who are looking for homes with high quality-price-ratio, and enjoy relatively short commutes as well as quietness.

Cluster 2 (Purple):

If you are looking to purchase a home but with a limited budget, look no further. This cluster includes neighborhoods in Bronx, Queens as well as Brooklyn that are comparatively far away from the center of Manhattan. But the average housing price is about $500,000, with limited facilities nearby.

Cluster 3 (Blue):

These neighborhoods sit in the middle of Manhattan, with average sales price of $2-3 millions. They all have an abundance of facilities, such as food and drinks, medical centers as well as recreation space. They are suitable for people with a substantial amount of money and would like to enjoy city life.

Cluster 4 (Green):

It is funny that the algorithm puts the upper west side and upper east side in a cluster. These two neighborhoods are famous for their quietness, closeness to central park, an array of medical services as well as an abundance of housing inventory.

Cluster 5 (Orange):

These neighborhoods are also either in Manhattan or in very close proximity to Manhattan, with the average housing price being around $1 million. They are good choices for people who love city life but are not ultra rich.

Discussion

Based on the K-means clustering, there are 5 distinct clusters that can be used to segment neighborhoods in NYC. In this section, we will delve deeper into the clusters and gain some insights of the clusters and their correlations with the housing price, inventory and venues nearby,

Overall Distribution

Figure 3 below shows the overall distribution of the clusters with regard to medium housing prices, inventory and total number of venues nearby.

As you can see, the five clusters all have their distinct features and are separated nicely. Cluster 3 (label 2) includes neighborhoods that have generally high housing prices and higher number of venues nearby. Cluster 4 (label 3) is characterized by its high inventory number, which cluster 2, 1, 5 are basically incrementing linearly for both housing prices and number of venues.

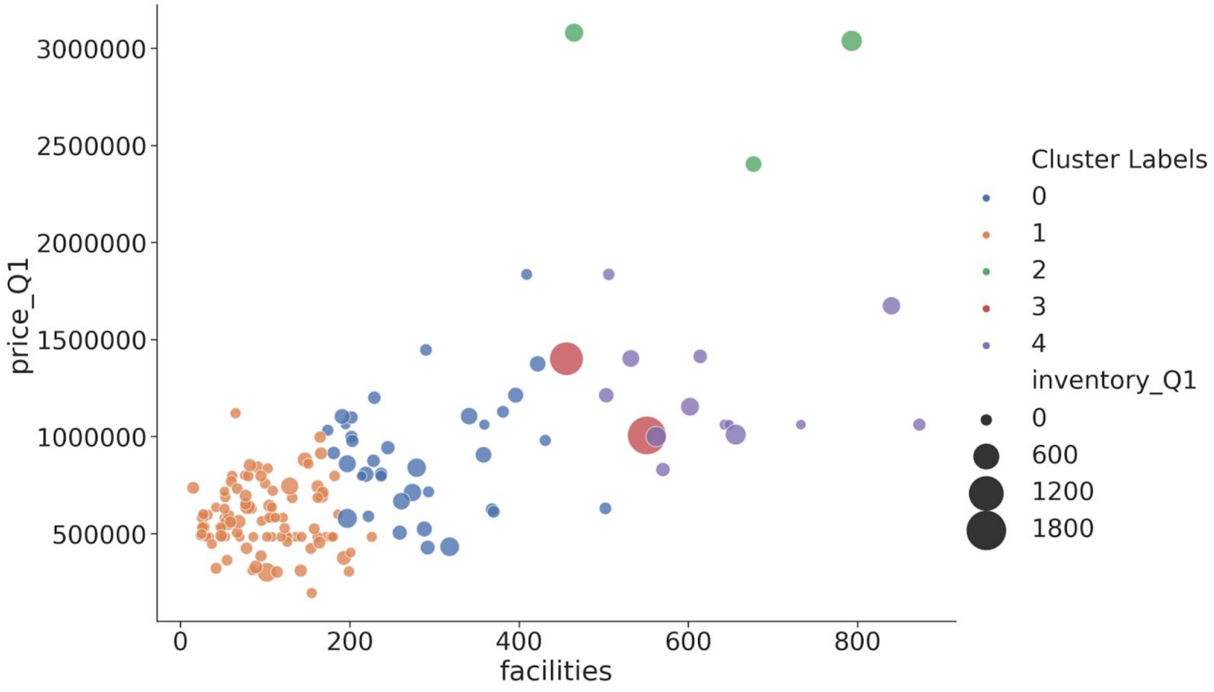

Figure 3

Housing Prices and Venues for the Clusters

Figure 4 shows the average and dispersion of housing prices in each cluster more intuitively. In terms of average housing price, the clusters ranked from low to high are: cluster 2, cluster 1, cluster 5, cluster 4 and cluster 3 (significantly higher than the other clusters).

And the housing prices in cluster 1 and 5 are comparatively more dispersed, while cluster 2 and 4, especially cluster 4, are much less dispersed. It means that the houses in cluster 1 and 5 have both relatively low and high selling prices, and the houses in cluster 2 and 4 tend to have similar selling prices and be closer to the average selling price.

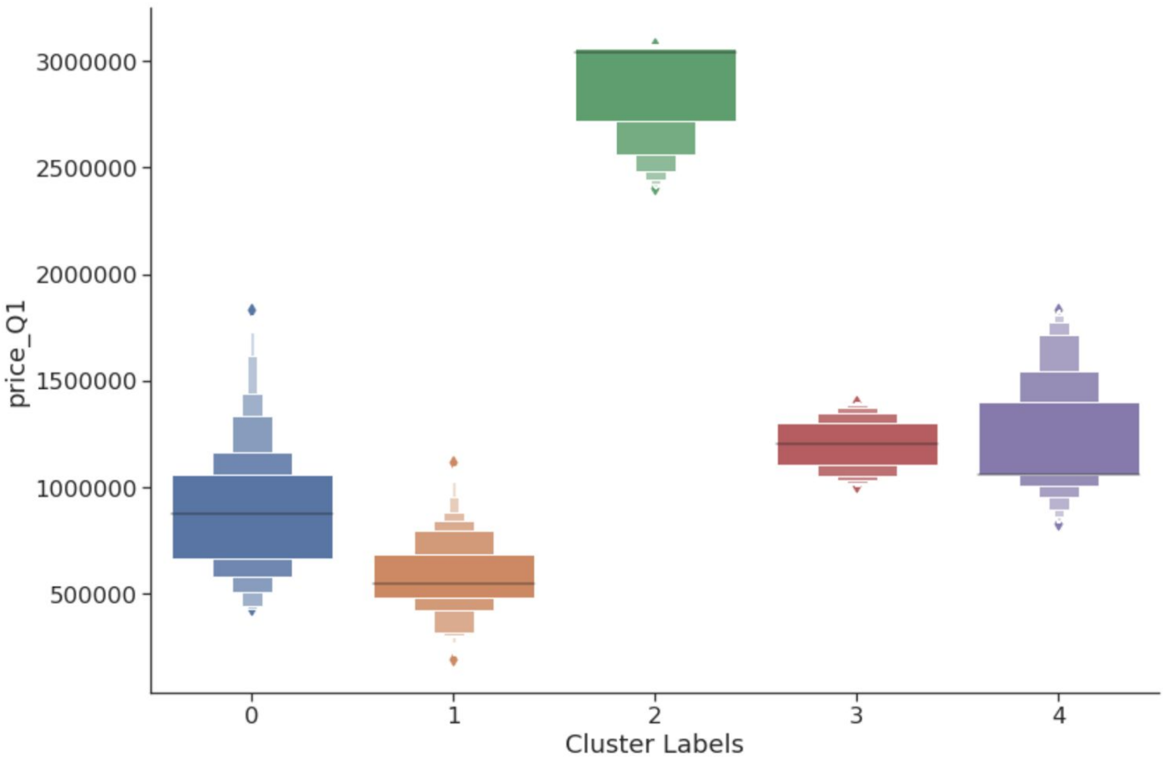

Figure 4

When it comes to the different categories of venues in the clusters, as shown in the figure 5 below, cluster 1 and 2 have relatively lower numbers of venues in all five categories. Cluster 3, 4 and 5 all have plentiful and various venues. Cluster 4 is characterized by its abundant medical facilities. And the distribution of venues for cluster 3 and 5 are quite similar. They both have ample outdoor and recreation facilities, medical facilities, food & drinks venues nearby.

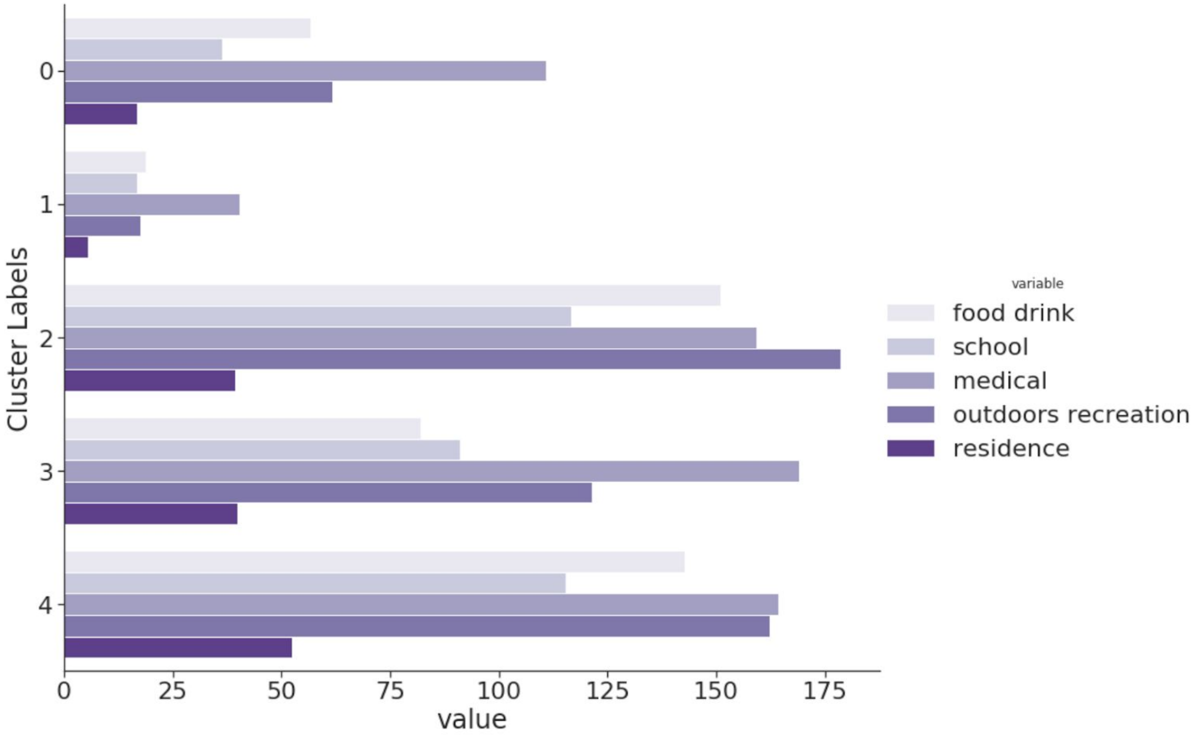

Figure 5

Correlations

In order to dig deeper into the relationships among the features, I used a cluster map to intuitively compare the correlations of each pair of features. As we can see in figure 6, the categories of venues nearby tend to highly correlate with each other. Therefore, we will use the total number of venues to calculate the correlation between venues and housing price.

In addition, the cluster labels are actually more closed correlated to venues nearby than housing price as well as inventory.

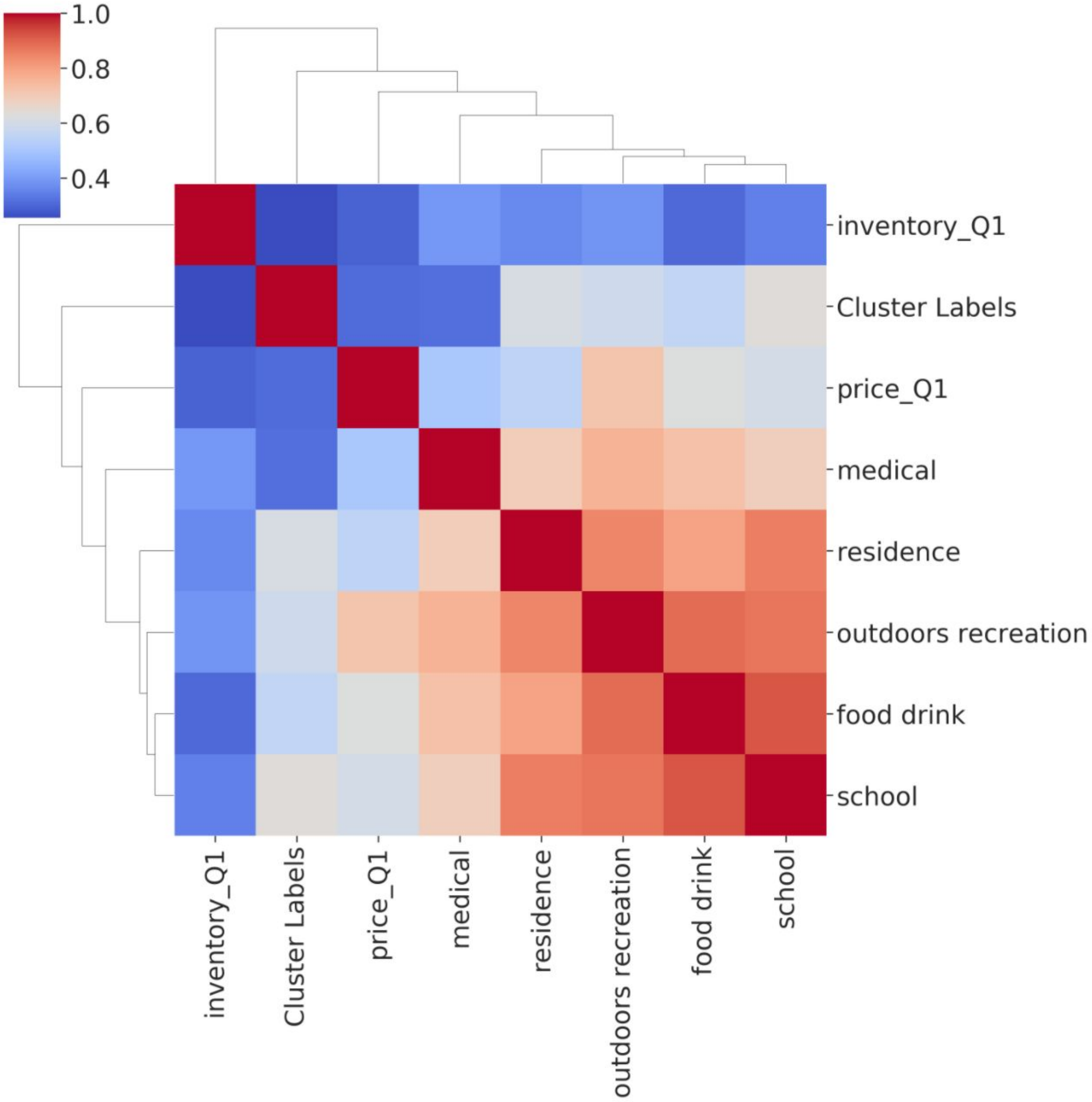

Figure 6

Now, let us delve into the correlation between housing prices and number of venues. As shown in figure 7, when the total number of facilities nearby grows, the housing prices tend to grow as well. Based on the slope, we can conclude that the total number of venues and housing prices have a slight positive correlation.

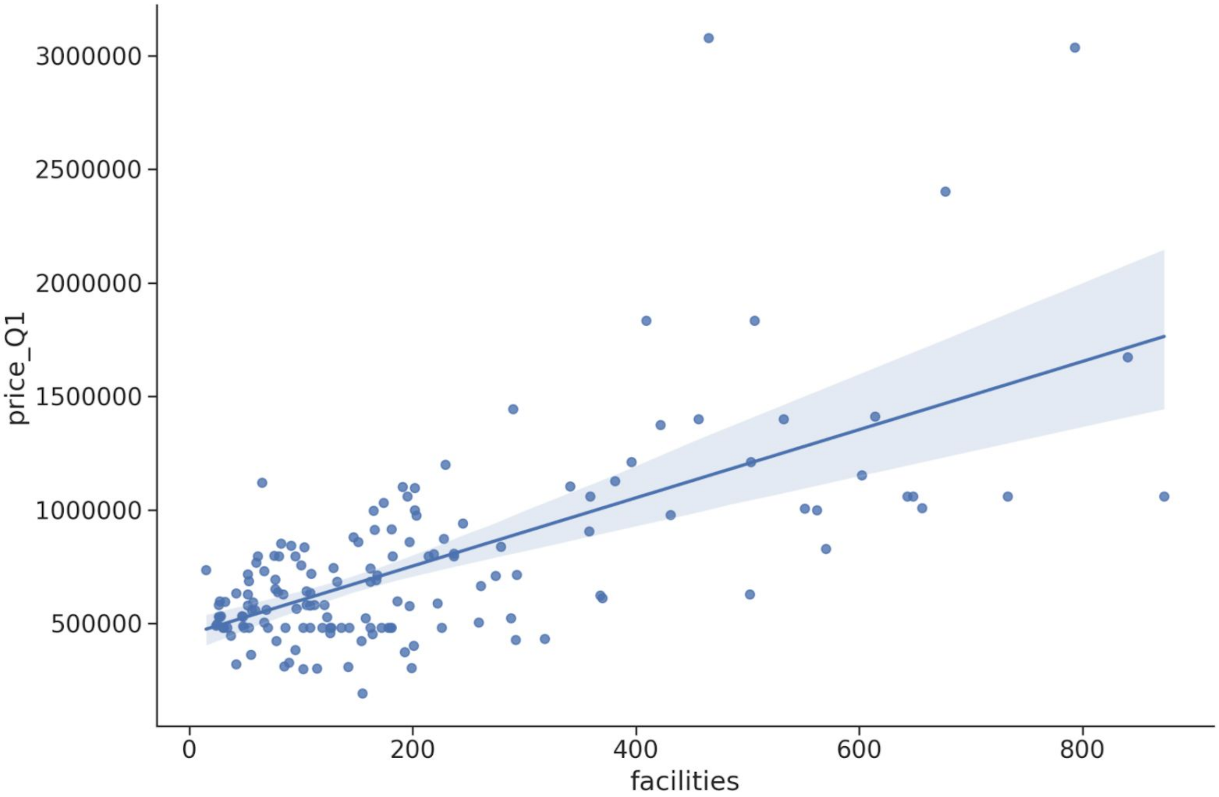

Figure 7

Conclusion

As millennials are reaching the ages of becoming homeowners, picking a place to purchase is a huge task and problem for many people.

This project used K-means clustering algorithm as well as location /real estate data to help segrement neighborhoods in NYC into 5 categories, which suit the needs of people with different budgets, as well as preferences of various facilities such as schools and medical facilities.

However, there are certain limitations of the project: as the real estate data obtained from StreetEasy does not contain neighborhoods in Staten Island, the final clustering does not include the majority of neighborhoods in Staten Island. People who are interested in furthering this project, should try to include these data.

To conclude, customers with a substantial amount of money and enjoying city life could go with cluster 3; customers who enjoys city life but are not willing to pay $2-3 millions for their homes should pick cluster 5; customers who would like to be close to central park and an abundance of medical centers/schools can choose cluster 4; customers who enjoys the proximity of Manhattan, but would like some quietness and have medium amount of budget can select cluster 1; and finally, customers who have limited budget and do not mind long commutes to midtown can pick cluster 2.