Finding a suitable property zone base on the venues of neighborhood and real estate values in New York

Marrugo Nicolas, April 22 of 2019

1. Introduction

1.1 Background

Nowadays, according with the 2018 homebuyer report driven by Nerdwallet, discovered that 75% of North Americans still say that have a house is a top priority, but approximately just 15% of the people reported had purchased a home in the past five years, and 32% intend to do so in the next half-decade.

Despite the fact this percentage seems very low, according with CBRE Econometric Advisors since 2017, the U.S homeownership rate is stable around 63.7%, after some years of recovering from the drop of the global financial crisis of 2007-2008, that caused a decreasing in the rate of at least 10 %, having reports of homeownership below of 60%.

Taking into account, that the number of people buying a house per year is quite stable, lead to suppose that the number of renter-occupied households will be stable, but actually is continuing to grow in most markets, driven by population growth and household formation. Where cities like Chicago and Detroit will be actually stable, while cities like New York City and Boston will be weak growth in the renter-occupied households.

The reasons why the market have this behavior is basically due to the lack of inventory of houses, limited construction of new houses, construction of luxury properties, having zones with rules that restrict a higher density development or simple that investors keep the new acquisitions as rentals. This lead to have a real challenge in order to find a suitable house zone that fulfill the desires of the client in terms of location, price and neighborhood.

1.2 Problem

Taking into account the previous statements, is possible determine that from the point of view of buyers a new client with the intention of buy a house, the first challenge to affront is actually finding a suitable home to move to, in terms of average house price and the location (venues near of the house). This project aims to help the client to find a suitable house zone for buying, analyzing the venues of each neighborhood and the average price of a house for the city of New York, giving as a result a list of neighborhoods with its corresponding average price and promising price behavior which accomplish the client venues or location preferences.

1.3 Interest

In that order of ideas, this project will fulfill the necessities of mid-range and late-range aged persons that are looking for a new or used house to buy in New York city, considering desirable venues around the house, the relation between house price and location, and its corresponding behavior in the real estate market.

2. Data acquisition and cleaning

2.1 Data sources

For developing this project is necessary real state data for all the neighborhoods in New York, this information is provided by the New York Times listings database and the New York City government records, which provide data from the last year, the last 3 months and the current day average house price for 286 neighborhoods. The information gather is given by an analysis of price and sales trends as well as many other metrics to give readers an idea of current conditions for this market and nearby markets, as well as historical market trends.

On the other hand, the venues information is obtained by the foursquare database will give the 70 venues near to the location of each neighborhood in a radius of 500 meters. Finally, the location information is given by the New York (City), Department of City Planning, where has a total of 5 boroughs and 306 neighborhoods, each neighborhood that exist in each borough have the latitude and longitude coordinates.

2.2 Data cleaning

The location data gathered the New York (City), Department of City Planning is in json format, from which is necessary take the parameters of borough, neighborhood, latitude, longitude, and then this data is converted into a pandas dataframe.

The real estate data had to be collected by neighborhood from the New York Times listings database and the New York City government records, due to some of the neighborhoods in the real estate data are separated in the location data, was necessary a rearranged of the data in order to have an average price per house for each neighborhood in the 306 of the location dataset.

Then the real estate data was cleaning from Not a Number values and then was normalize using the min-max function. Finally, the venues data obtained from the foursquare database was gather in Json format, from where the name, latitude, longitude and category of the venue were extracted and converted into a pandas dataframe.

Once this data was extracted, the venues for each neighborhood was grouped in order to find the proportion that each category is present in each neighborhood. In this way is possible determine the 10 most common venues categories per each zone.

2.3 Feature selection

After data cleaning, there were 9039 venues for 306 neighborhoods, were 416 unique categories were identified, but the distribution of these venues in each neighborhoods is represented in the figure 1.

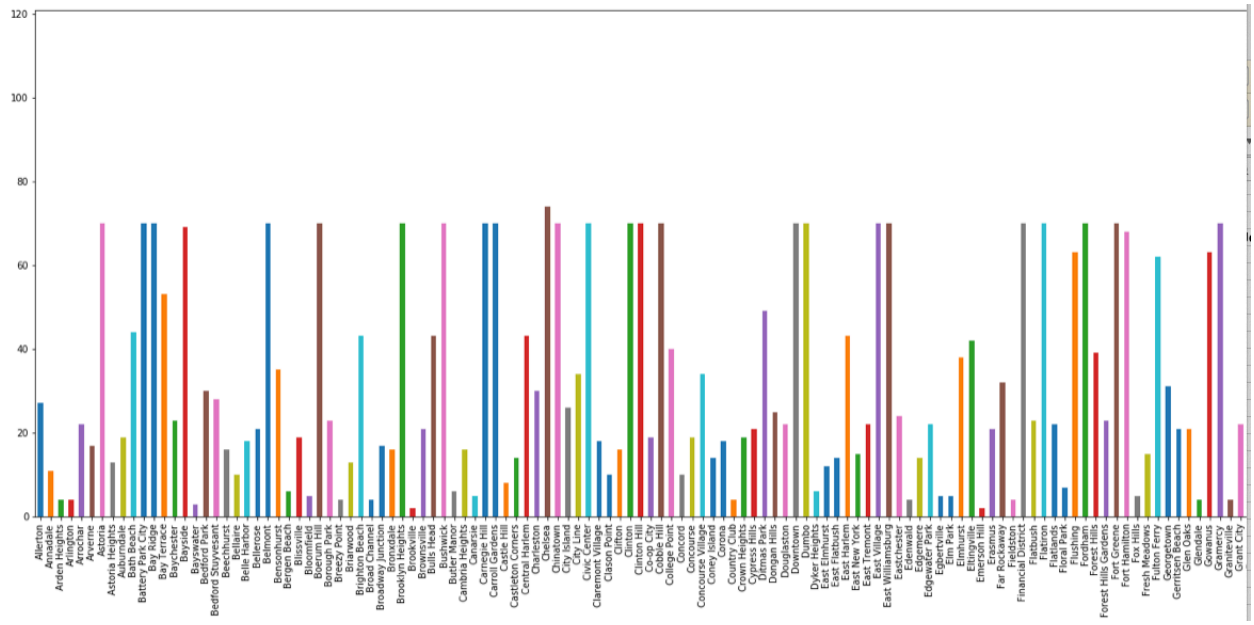

Figure 1. Distribution of number of venues per neighborhood.

Using the distribution of the venues in all the neighborhoods and mean of 30 venues per neighborhood was found, then is possible to analyze the distribution of categories of each neighborhood, as is shown in figure 2.

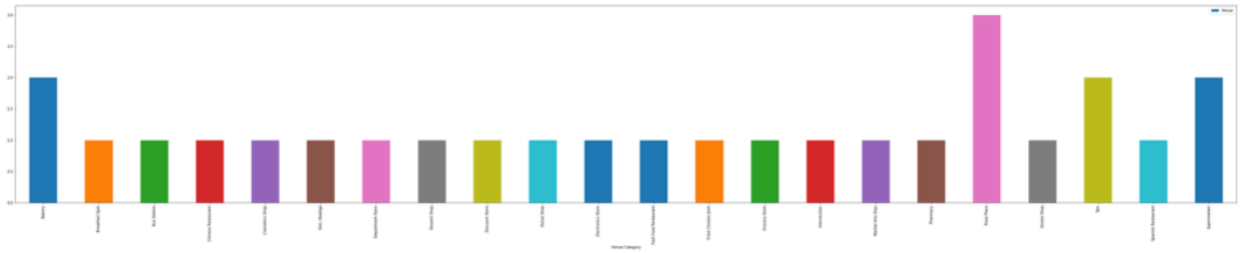

Figure 2. Distribution of unique venues categories for the neighborhood Allerton.

Taking into account that for this neighborhood 27 venues were selected from the foursquare database, 22 unique categories were identified, in this order the ideas is feasible select at least 10 most repetitive categories on each neighborhood. Additional to this data the normalized value of the average price per house today was added for the cluster features. Given the following result:

Borough	Bronx	Bronx	Bronx	Bronx	Bronx
Neighborhood	Allerton	Baychester	Bedford Park	Belmont	Bronxdale
price nowNormal	0.012022	0.042995	0.017294	0.046172	0.0000001

1st Most Common Venue	Pizza Place	Pizza Place	Pizza Place	Italian Restaurant	Italian Restaurant
2nd Most Common Venue	Supermarket	Breakfast Spot	Mexican Restaurant	Pizza Place	Chinese Restaurant
3rd Most Common Venue	Deli / Bodega	Mexican Restaurant	Chinese Restaurant	Deli / Bodega	Performing Arts Venue
4th Most Common Venue	Spa	Mattress Store	Diner	Bakery	Breakfast Spot
5th Most Common Venue	Chinese Restaurant	Fried Chicken Joint	Supermarket	Grocery Store	Eastern European Restaurant
6th Most Common Venue	Intersection	Bank	Pharmacy	Dessert Shop	Gym
7th Most Common Venue	Donut Shop	Pet Store	Sandwich Place	Bar	Mexican Restaurant
8th Most Common Venue	Breakfast Spot	Fast Food Restaurant	Bar	Mediterranean Restaurant	Pizza Place
9th Most Common Venue	Spanish Restaurant	Electronics Store	Fried Chicken Joint	Mexican Restaurant	Spanish Restaurant
10th Most Common Venue	Fast Food Restaurant	Sandwich Place	Spanish Restaurant	Spanish Restaurant	Paper / Office Supplies Store

Table 1. Features Selected for the clustering algorithm.

The features matrix for the clustering algorithm was set as 11 features that are the normalize average price house price of the neighborhood, following for the category of the 10 venues most repetitive in each neighborhood in New York.

3. Methodology

The first exploratory data analysis made was the calculation of the percentage difference between the average price today with respect its price 12 months ago (value 1) and 3 months ago (value 2), then this data was used to define the real estate index using taking into account the following rules:

Value 1	Value 2	Condition	trend	Real estate Index
>0	>0	Value 1 > Value 2	Negative	Appreciation
<0	<0	Value 1 > Value 2	Negative	Depreciation
<0	>0	Value 1 < Value 2	Positive	Appreciation

>0	<0	Value 1 > Value 2	Negative	Depreciation
>0	>0	Value 1 < Value 2	Positive	Appreciation
<0	<0	Value 1 <Value 2	Positive	Depreciation

Table 2. Rules for evaluating the Real estate index.

In this order of ideas, is possible to now if the average price of a house is increasing with a positive trend but still the value is lower than previous registers (depreciation) or the house price is decreasing with a negative trend but still the value is higher than previous registers (appreciation), as is shown in the following table:

Borough	Bronx	Bronx	Bronx	Bronx	Bronx
Neighborhood	Allerton	Baychester	Bedford Park	Belmont	Bronxdale
Price 1 y agoNormal	0.055220591	0.070238245	0.031627338	0.06297011	0.016990212
Price 1 y ago	414917	500833	279940	459252	196201
price 3 month agoNormal	0.016108254	0.051738187	0.023405476	0.05469014	0.000768271
price 3 month ago	260583	484722	306488	503292	164083
price nowNormal	0.012021534	0.042995358	0.017293817	0.04617156	0
price now	267667	549583	315654	578492	158250
Real estate percentage 12month ago	-55.01238479	8.870361711	11.31428716	20.61221244	-23.98167457
Real estate percentage 3month ago	2.646572047	11.80185704	2.903812402	12.99931546	-3.685939968
Real state index	Positive trend (+%) and an appreciation on the average price of +2.647% in the last 3 months.	Positive trend (+%) and an appreciation on the average price of +11.802% in the last 3 months.	Negative trend (-%) and an appreciation on the average price of +2.904% in the last 3 months.	Negative trend (-%) and an appreciation on the average price of +12.999% in the last 3 months.	Positive trend (+%) and an depreciation on the average price of -3.686% in the last 3 months.

Table 3. Exploratory analysis of the average price of house per neighborhood.

Once the real estate index for each neighborhood is defined, is possible to identify the most common venues for each neighborhood using the venues data from foursquare, where 9039 venues were extracted with 416 unique categories, calculating the categorical value for each unique category, giving the following categorical values:

Neighborhood	Allerton
Accessories Store	0
Adult Boutique	0
Afghan Restaurant	0
African Restaurant	0
Piercing Parlor	0
Pilates Studio	0
Pizza Place	1
Platform	0
Playground	0
Arcade	0
Arepa Restaurant	0
Argentinian Restaurant	0

Table 4. categorical values for a part of 12 venues categories from the total of 416 venues categories.

With the categorical values of the venues is possible to evaluate which are the probability of have that type of venue per neighborhood, and finally gather the 10 most common venues per each of the 306 neighborhoods.

Neighborhood	Allerton	Annadale	Arden Heights	Arlington
1st Most Common Venue	Pizza Place	Pizza Place	Pharmacy	Bus Stop
2nd Most Common Venue	Supermarket	American Restaurant	Coffee Shop	Furniture / Home Store
3rd Most Common Venue	Spa	Restaurant	Deli / Bodega	Coffee Shop
4th Most Common Venue	Bakery	Train Station	Pizza Place	Field

5th Most Common Venue	Breakfast Spot	Deli / Bodega	English Restaurant	Event Service
6th Most Common Venue	Fried Chicken Joint	Diner	Ethiopian Restaurant	Event Space
7th Most Common Venue	Spanish Restaurant	Sports Bar	Event Service	Exhibit
8th Most Common Venue	Bus Station	Park	Event Space	Eye Doctor
9th Most Common Venue	Fast Food Restaurant	Dance Studio	Exhibit	Factory
10th Most Common Venue	Smoke Shop	Event Space	Eye Doctor	Falafel Restaurant

Table 5. The 10 most venues per each neighborhood.

Once all the dataset of features was created, the data from location, venues and real estate was gathered together, in order to train the K-means algorithm.

4. Classification Modeling

Taking into account that the purpose of the algorithm is to identify the set of neighborhoods in New Yok that share similar venues around and similar average house price, the most promising machine learning that would accomplish with this requirement is the k-means clustering. This machine learning algorithm is design to have a good performance with large dataset, like the one used in this project; another reason to develop a K-means clustering is that the data does not contain noise due to the massive categorical values, for this reason there is no need to use algorithms robust to noise like DBSCAN (density based spatial clustering application with noise).

According with the location data there is 306 neighborhoods distributing in the following map using the folium library.

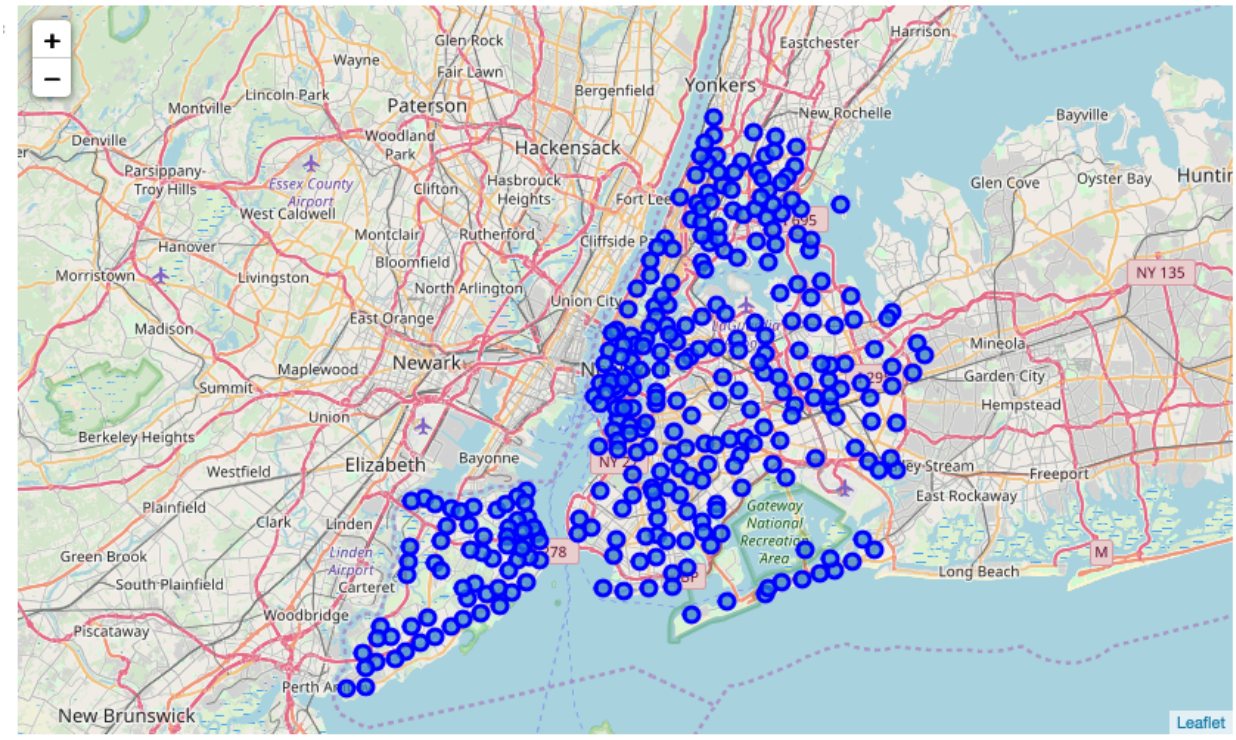

Figure 3. Map of New York with a marker for each neighborhood

The location data is merged with the venues data, that is compressed from more or less 30 venues per neighborhood to the 10 most common categories of those venues for each neighborhood. Finally, the location and venues data are merge with the normalize average price of a house in each neighborhood.

Once all the data is merged the K-means clustering algorithm will analyze the 10 most common venues in a radius of 500 meters of the center of each neighborhood, and the average price of the house in that neighborhood, to separate the total features data (location data, real estate data, venues data) in 8 clusters.

This number of clusters was selected after trying different numbers from 5 to 20 and identify that the best clustering in order to avoid clusters with 1 neighborhood was 8 clusters. The 306 neighborhoods were classified in 8 cluster as the following figure shows.

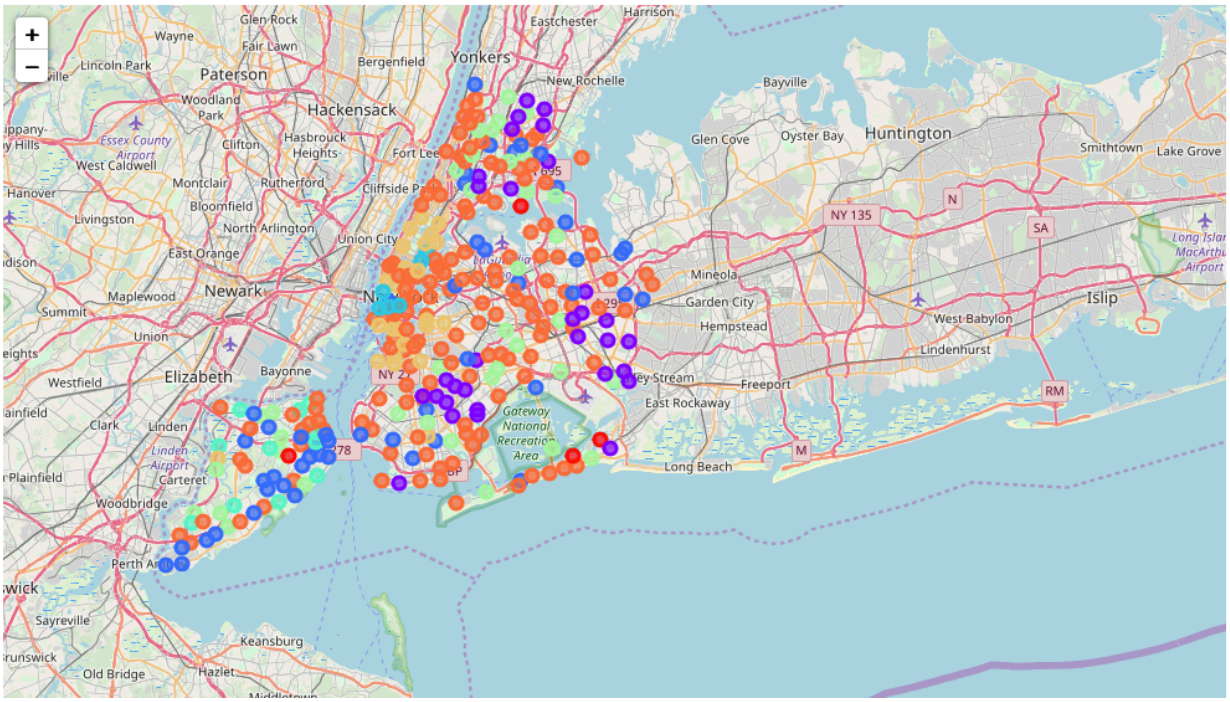

Figure 4. Clusters of the neighborhood of New York

5. Results of classification

In order to examine each cluster classified with the K-means algorithm is necessary add the real estate index corresponding for each neighborhood in the cluster, then is possible make the plot of the map using folium to show the location of each neighborhood with a marker that contain the average price of house today, the real estate index, and the depreciation of appreciation on the average price of the house in the last 3 months.

Once each cluster have the complete information is possible to identify the maximum and minimum price of the cluster, to give an idea of the real estate price of these neighborhoods, and also to give a description of the most common venues of the cluster that among all neighborhoods in the cluster.

For each cluster the information is presented in a dataframe as is shown in table 6.

Borough	Bronx	Bronx	Bronx
Neighborhood	Morrisania	Eastchester	Claremont Village
Latitude	40.823592	40.887556	40.831428
Longitude	-73.901506	-73.827806	-73.901199
Real estate percentage 12month ago	-115.331429	-14.422189	-53.001924

Real estate percentage 3month ago	0	-1.166694	-16.875893
price now	227500	337792	339982
Real state index	Negative trend (-%) and a depreciation on the average price of -4.497% in the last 3 months.	Positive trend (+%) and a depreciation on the average price of -1.167% in the last 3 months.	Positive trend (+%) and a depreciation on the average price of -16.876% in the last 3 months.
price nowNormal	0.007608	0.019726	0.019967
Cluster Labels	0	0	0
1st Most Common Venue	Discount Store	Caribbean Restaurant	Bakery
2nd Most Common Venue	Bus Station	Bus Station	Chinese Restaurant
3rd Most Common Venue	Fast Food Restaurant	Diner	Deli / Bodega
4th Most Common Venue	Metro Station	Deli / Bodega	Supermarket
5th Most Common Venue	Donut Shop	Metro Station	Bus Station
6th Most Common Venue	Grocery Store	Bus Stop	Grocery Store
7th Most Common Venue	Sandwich Place	Platform	Caribbean Restaurant

8th Most Common Venue	American Restaurant	Donut Shop	Liquor Store
9th Most Common Venue	Liquor Store	Pizza Place	Gift Shop
10th Most Common Venue	Bowling Alley	Seafood Restaurant	Pizza Place

Table 6. Part of the dataframe with the information of cluster 0.

The dataframe of each cluster contains 19 columns, that refer the borough, neighborhood, latitude, longitude of each neighborhood in the cluster, also show the percentage of depreciation or appreciation of the average price of a house today with respect the values gathered 12 months ago and 3 months ago. Also, the dataframe show the average price of a house today and its percentage of similarity with the maximum cost of a house in New York which correspond to a 9260000 $, finally the real estate index is added.

5.1 Cluster 0

The first cluster label as 0 contains 31 neighborhoods in which the maximum price is 1’260,000$ and have a percentage of similarity with the maximum price in New York of 12.1%. The minimum price is 227,500$ and have a percentage of similarity of 0.76%.

The venues that are more frequent in a radius of 500 meters of the center of each neighborhood are Caribbean Restaurant, Chinese Restaurant, Gym, Fried chicken Joint, Fast Food restaurant, bus station, supermarket.

The distribution of the neighborhoods of the cluster is shown in figure 5.

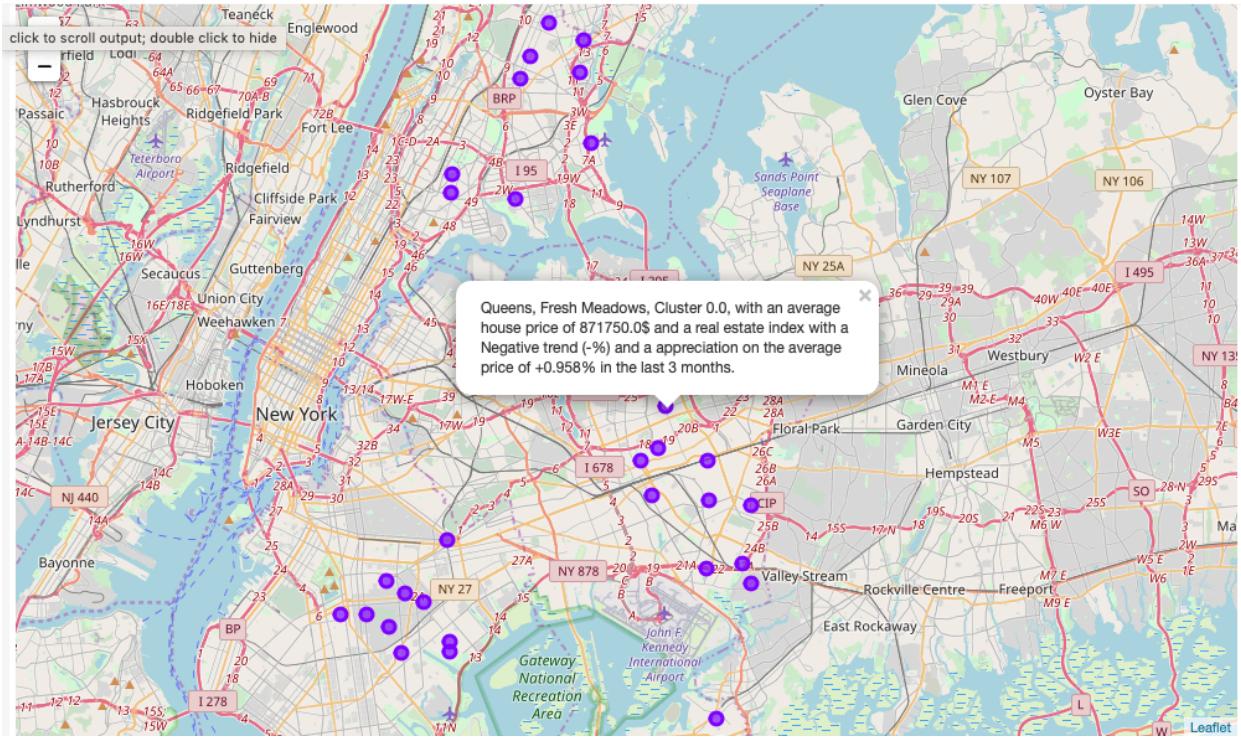

Figure 5. Folium map with the location of the neighborhoods in the cluster 0.

The figure shown the information contain in one market in this case with the neighborhood of Fresh Meadows in the borough Queens, which have an average house price of 871,750$ and have a negative trend in the real estate index this means that the price is decreasing with the time but in the last 3 months the price has an appreciation of + 0.958%

5.2 Cluster 1

The second cluster label as 1 contains 50 neighborhoods in which the maximum price is 1’200,000$ and have a percentage of similarity with the maximum price in New York of 11,45%. The minimum price is 158,250$ and have a percentage of similarity of 0.0%, this mean that actually this price is the cheapest average house price in New Yok.

The venues that are more frequent in a radius of 500 meters of the center of each neighborhood are Italian Restaurant, Pizza place, Bust stop, deli / Bodega, Chinese restaurant, Convenience store, pharmacy.

The distribution of the neighborhoods of the cluster is shown in figure 6.

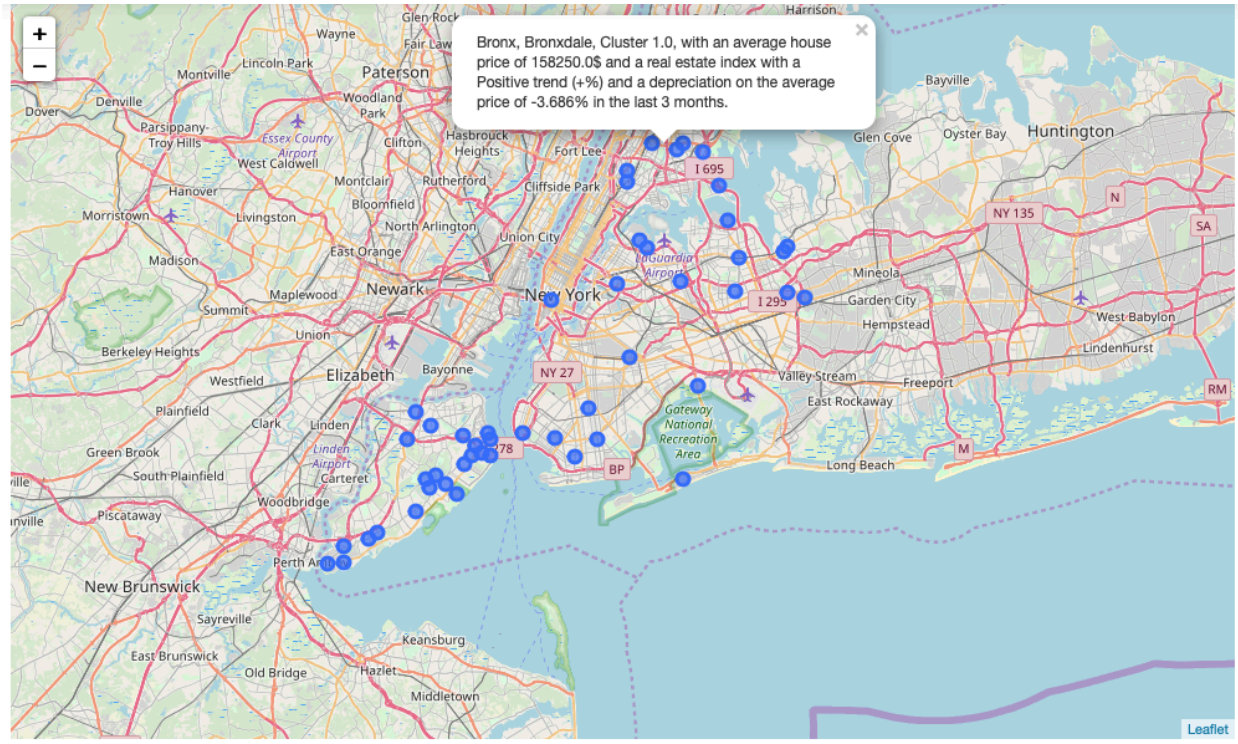

Figure 6. Folium map with the location of the neighborhoods in the cluster 1.

The figure shown the information contain in one market in this case with the neighborhood of Bronxdale in the borough Bronx, which have an average house price of 158,250$ (the cheapest average house price in New York) and have a positive trend in the real estate index this means that the price is increasing with the time but in the last 3 months the price has a depreciation of -3.9686%.

5.3 Cluster 2

The third cluster label as 2 contains 6 neighborhoods in which the maximum price is 9’260,000$ and have a percentage of similarity with the maximum price in New York of 100%. this mean that actually this price is the most expensive average house price in New Yok. The minimum price is 5’300,000$ and have a percentage of similarity of 56,49%.

The venues that are more frequent in a radius of 500 meters of the center of each neighborhood are Gym, Italian restaurant, cocktail bar, Boutique, Coffe shop, Greek restaurant.

The distribution of the neighborhoods of the cluster is shown in figure 7.

Figure 7. Folium map with the location of the neighborhoods in the cluster 2.

The figure shown the information contain in one market in this case with the neighborhood of Sutton Place in the borough Manhattan, which have an average house price of 9'260,000$ (the most expensive average house price in New York) and have a negative trend in the real estate index this means that the price is decreasing with the time but in the last 3 months the price has an appreciation of + 45.248%.

5.4 Cluster 3

The fourth cluster label as 3 contains 11 neighborhoods in which the maximum price is 1'020,000$ and have a percentage of similarity with the maximum price in New York of 9.46%. The minimum price is 228,100$ and have a percentage of similarity of 1.53%.

The venues that are more frequent in a radius of 500 meters of the center of each neighborhood are Bus stop, Bagel shop, deli/ Bodega, Gym, Recreation center, Park, liquor shop.

The distribution of the neighborhoods of the cluster is shown in figure 8.

Figure 8. Folium map with the location of the neighborhoods in the cluster 3.

The figure shown the information contain in one market in this case with the neighborhood of New Brighton in the Staten Island, which have an average house price of 426,750$ and have a negative trend in the real estate index this means that the price is decreasing with the time but in the last 3 months the price has an appreciation of +16.95%.

5.5 Cluster 4

The fifth cluster label as 4 contains 37 neighborhoods in which the maximum price is 1’020,000$ and have a percentage of similarity with the maximum price in New York of 9.46%. The minimum price is 234,333$ and have a percentage of similarity of 0.83%.

The venues that are more frequent in a radius of 500 meters of the center of each neighborhood are deli/ Bodega, Pizza place, American restaurant, Spanish restaurant, Chinese restaurant, supermarket, coffee shop.

The distribution of the neighborhoods of the cluster is shown in figure 9.

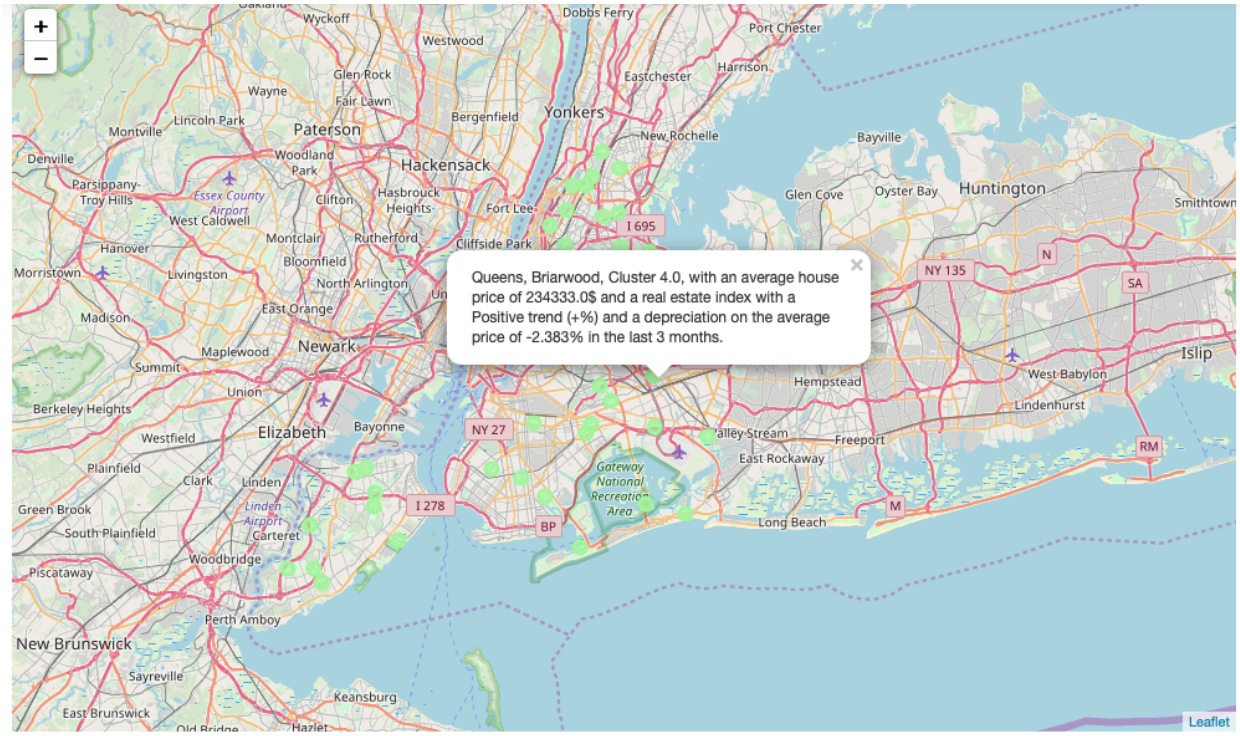

Figure 9. Folium map with the location of the neighborhoods in the cluster 4.

The figure shown the information contain in one market in this case with the neighborhood of Briarwood in the Queens, which have an average house price of 234,333$ (which is the lower average house price of the cluster) and have a positive trend in the real estate index this means that the price is increasing with the time but in the last 3 months the price has a depreciation of -2.383%.

5.6 Cluster 5

The sixth cluster label as 5 contains 24 neighborhoods in which the maximum price is 4’720,000$ and have a percentage of similarity with the maximum price in New York of 50,11% (more than the half of the maximum price in New York. The minimum price is 1’810,000$ and have a percentage of similarity of 18.14%.

The venues that are more frequent in a radius of 500 meters of the center of each neighborhood are Italian Restaurant, Chinese Restaurant, Bar, Night Club, Coffee shop, Wine shop.

The distribution of the neighborhoods of the cluster is shown in figure 10.

Figure 10. Folium map with the location of the neighborhoods in the cluster 5.

The figure shown the information contain in one market in this case with the neighborhood of Turtle Bay in the borough Manhattan, which have an average house price of 3'530,000$ and have a positive trend in the real estate index this means that the price is increasing with the time and in the last 3 months the price has an appreciation of +7.932%.

5.7 Cluster 6

The seventh cluster label as 6 contains 142 neighborhoods in which the maximum price is 1'990,000$ and have a percentage of similarity with the maximum price in New York of 20.12%. The minimum price is 163,532$ and have a percentage of similarity of 0.058%.

The venues that are more frequent in a radius of 500 meters of the center of each neighborhood are Fast Food Restaurant, Pizza Place, Pharmacy, Supermarket, Coffee shop, Italian Restaurant.

The distribution of the neighborhoods of the cluster is shown in figure 11.

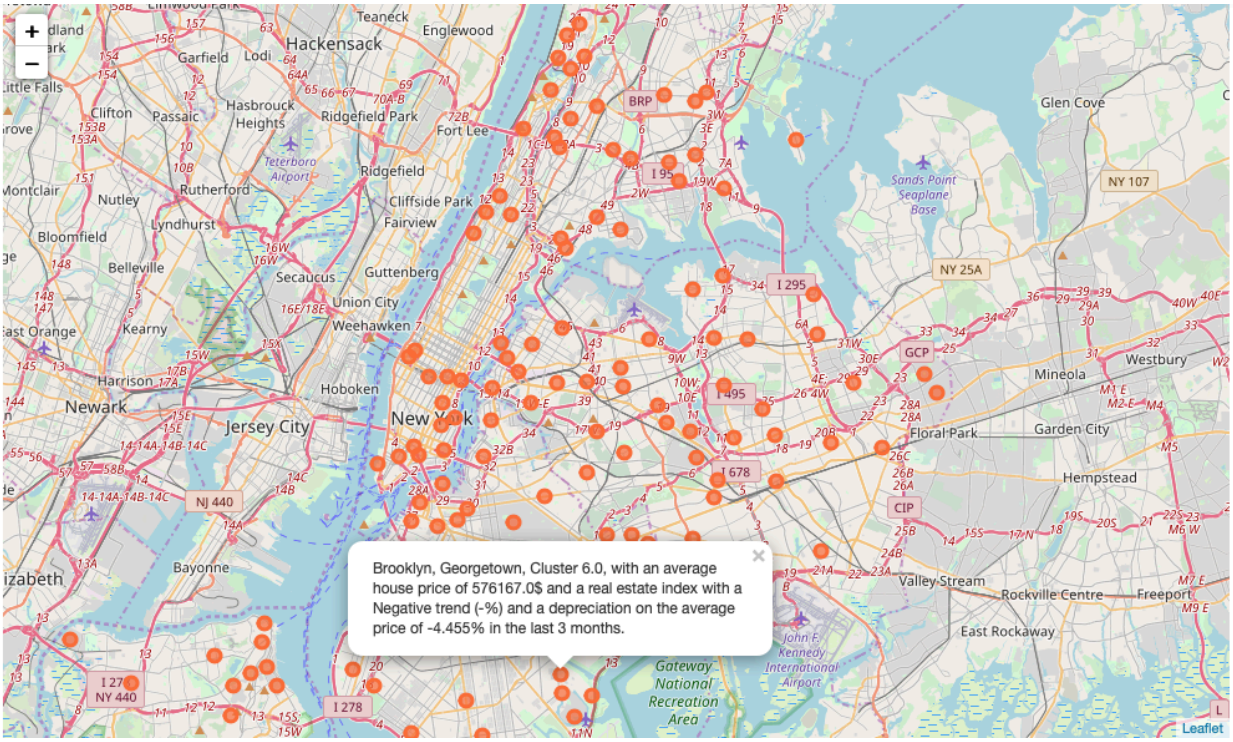

Figure 11. Folium map with the location of the neighborhoods in the cluster 6.

The figure shown the information contain in one market in this case with the neighborhood of Georgetown in the borough Brooklyn, which have an average house price of 576,167$ and have a negative trend in the real estate index this means that the price is decreasing with the time and in the last 3 months the price has a depreciation of -4.455%.

5.8 Cluster 7

The eighth cluster label as 7 contains 4 neighborhoods in which the maximum price is 1’020,000$ and have a percentage of similarity with the maximum price in New York of 9.46%. The minimum price is 439,291$ and have a percentage of similarity of 3.087%.

The venues that are more frequent in a radius of 500 meters of the center of each neighborhood are Tennis court, Park, Yoga studio, Playground.

The distribution of the neighborhoods of the cluster is shown in figure 12.

Figure 12. Folium map with the location of the neighborhoods in the cluster 7.

The figure shown the information contain in one market in this case with the neighborhood of Bayswater in the borough Queens, which have an average house price of 466,300$ and have a negative trend in the real estate index this means that the price is decreasing with the time and in the last 3 months the price has a depreciation of -5.469%.

After the K-means clustering algorithm, 8 clusters were obtained and after analyzing each cluster is possible identify that the algorithm top priority features for classification was the normalize average price of the house per neighborhood, because from the 8 cluster is possible to distinguish the following socio-economy insights:

Cluster	# per cluster	Minimum average price	Percentage of similarity	Maximum average price	Percentage of similarity	Economy	Venues relationship
0	31	227,500$	0.76%	1’260,000$	12.1%	Low price to medium price	Fast food and restaurants venues
1	50	158,250$	0.0%	1’200,000$	11,45%	Low price to medium price	Commercial venues with food restaurants
2	6	5’300,000$	56,49%.	9’260,000$	100%	High price	Boutique and unique restaurants
3	11	228,100$	1.53%.	1’020,000$	9.46%.	Low price	Bust stop, industrial and

							commercial venues
4	37	234,333$	0.83%.	1’020,000$	9.46%.	Low price	Food and commercial venues
5	24	1’810,000$	18.14%	4’720,000$	50,11%	Medium price to high price	Night live venues and restaurant venues
6	142	163,532$	0.058%.	1’990,000$	20.12%	Low price to medium price	Popular venues (supermarket, restaurant, pharmacy, fast food)
7	4	439,291$	3.087%.	1’020,000$	9.46%.	Low price	Open yard venues (tennis court, park, Playgorunds)

Table 7. Socio-economic analysis of each cluster.

6. Discussion

For the Economy insights was necessary divide the average house price in 3 categories (low, medium, high). Where the highest average house price in New York is about 9’260,000$ and because this value corresponds to a limited number of neighborhoods, the category begins from more than 4’000,000, the medium average house price was set from 1’020,000$ up to 4’000,000$, while lower average price is below of 1’020,000$.

For the social insights is possible identify that cluster 0 have neighborhoods that have venues related with food and fast food and usually are located in the center of the boroughs of New York. This type of venues and location can define that this cluster is for neighborhoods with high density of housing with average price that are from low up to the beginning of medium categories.

The cluster 1 have venues related with commercial activities and fast food venues, the neighborhoods related in this cluster are located approximately in the borders of the borough of New York except by Staten Island, this indicates that the average price of house in this neighborhood are between lower and the beginning of medium category, due to the distance to the center of the boroughs.

The cluster 2 have venues related with luxury activities like boutique or expensive restaurant, those types of venues defined the neighborhoods involved in this cluster to be exclusive in New York, and a justification of this is that these neighborhoods have the most expensive average house price in all New York and are located exactly in Manhattan.

The cluster 3 have venues related with commercial activities, green areas and entertainment activities, but the average price of the house in these neighborhoods are in the lower category price

like the neighborhoods of the cluster 1, but with the difference that these neighborhoods are located in the borough of Staten Island.

The cluster 4 have venues related with food like restaurants, which indicates that these neighborhoods are characterized to be near to working places or event places that gather huge amount of people, but although the neighborhoods are distributed in strategic places around the boroughs of New York the average house price is in the lower category.

The cluster 5 have venues related with night live as clubs, bar, cocktail and expensive restaurants, which characterized these neighborhoods to be in the most popular zone for party or hangout in New York, for this reason the average price of the house is between the medium and high categories.

The cluster 6 have venues related with popular zones of New York with high residential density, this means that the venues are mixed in the way that is possible to have restaurants, banks, pharmacy and diverse types of venues, for this reason this cluster have the bigger number of neighborhoods but all of them have an average house price between low and half of the medium categories.

The cluster 7 have venues related with outside activities like parks, gyms and playgrounds, for this characteristic these neighborhoods are not included in the clusters 3 or 4, these neighborhoods have an average house price of low price to the beginning of the medium category and are distributed in different boroughs in New York.

The clustering made in the project allows to separate the 306 neighborhoods in New York by social-economic features have a good dispersion of the neighborhoods avoiding cluster with all the neighborhoods or with just 1, in addition, the result support that the project can be a good resource to identify different zones that share same characteristics according with the current neighborhood that the client likes or even according with the venues and average price that the client prefer.

In addition to these social-economic features the project can be complement with crime information of each neighborhood, to include safety index in the clustering algorithm, or even can include current sales information for each neighborhood from a commercial database like the New York Times listings, in order to improve the value obtained by the client from this project.

7. Conclusion

Taking into account that from the point of view of buyers a new client with the intention of buy a house, the first challenge to affront is actually finding a suitable home to move to, in terms of average house price and the location (venues near of the house). This project reached a first step in a solution of help the client to find a suitable house zone for buying, based on the venues of each neighborhood and the average price of a house for the city of New York.

The project made an unsupervised analysis of the neighborhoods in New York, using K-means clustering to found 8 Clusters that are not just divide the neighborhoods according with location or the average house price, but also based on the unlabeled data of the venues around of 500 meters of the neighborhood.

The machine learning algorithm identified cluster that are compose by neighborhoods with lower average price and are located in zones near to the center of boroughs of New York, identify the neighborhoods with extreme high average house price and luxury venues near in other cluster, also identify medium to high average house price neighborhoods that have near night live venues.

The algorithm was able to discriminate the neighborhoods with a lower and the beginning of medium average house price that correspond more than 70% of the neighborhoods in New York, in 6 different clusters depending in if the neighborhoods are in the border of the borough, are located near to work places, in the center of borough or just with high residential building density.

In that order the ideas the project was able to cluster all neighborhoods in New York based on Social-economic features, which ones help the client to find a suitable zone for house acquisition based on the type of venues that the client preferred to be near of the house and/or the desired location, because with this information was possible identify the cluster that contains neighborhoods which those features.

Once the cluster was selected, the information displayed per cluster was in dataframe format and folium map format, compose by the borough, neighborhood, latitude, longitude of each neighborhood in the cluster, also the percentage of depreciation or appreciation of the average price of a house today with respect the values gathered 12 months ago and 3 months ago and the real estate index in the average price of the house to increase or decrease with the past of the time; allowing the user to choose among the list which zone are more profitable for buying a house based in the zone with a real estate index positive and appreciation with time, or just choose the zone that adjust to the budget of the user, or just the zone with some type of desirable venues.

8. Future directions

Although, the project identified the zone more suitable for buying a house depending in the venues around the neighborhood and the average house price of them, the project does not include information like crime rate, that can include another feature like safety index, that will lead to have a more specific neighborhoods separation in more clusters given a more refined result.

Another type of information that can improve the project is the house sales information, that could be the definite next step in the project to help the user no just to find a suitable zone based in his desired venues and budget, but also find a house offer using the sales options per neighborhood.