Clustering New York Neighborhoods by their Real State Market

Juan Tornero Lucas

December 2, 2020

1. Introduction

The goal of this study is to give insights to a real state investor from Manhattan. Which areas are the best to invest accroding to your needs? Which is the main characteristics of the real state market in every neighbourhood?

Let's say you are a real state investor that wants to focus their next steps in the New York area, but first you want to know which areas fit best your appetite. Maybe you are more focused in suburban areas with low price per square meter, or maybe you are into the luxury segment and want to know which places the high end clients prefer.

With this study, we'll categorize neighbourhoods into different clusters that will describe which kind of investment fits them better.

2. Data

We use the data provided in the Kaggle repository [NYC Property Sales](#)

The data will be prepared with main focus in the numeric columns that feature properties of the sales and the neighborhood they are located.

The first thing we realize is that some of the columns we expect to be numeric have non-numeric values such as "-"

TAX LASS AT SENT	BLOCK	LOT	EASE-MENT	BUILDING CLASS AT PRESENT	ADDRESS	...	RESIDENTIAL UNITS	COMMERCIAL UNITS	TOTAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	YEAR BUILT	TAX CLASS AT TIME OF SALE	BUILDING CLASS AT TIME OF SALE	SALE PRICE
2A	392	6		C2	153 AVENUE B	...	5	0	5	1633	6440	1900	2	C2	6625000
2	399	26		C7	234 EAST 4TH STREET	...	28	3	31	4616	18690	1900	2	C7	-
2	399	39		C7	197 EAST 3RD STREET	...	16	1	17	2212	7803	1900	2	C7	-
2B	402	21		C4	154 EAST 7TH STREET	...	10	0	10	2272	6794	1913	2	C4	3936272
2A	404	55		C2	301 EAST 10TH STREET	...	6	0	6	2369	4615	1900	2	C2	8000000

We drop all rows containing such values and convert the columns into numeric ones.

After this I look for the main statistical characteristics of the dataframe. There are some extreme outlayers, which I consider normal regarding the market we study and can be indicators of particular neighborhood characteristics. But also there is data with values = 0 in rows like price and square feet. So I also drop rows with values = 0 in this columns.

	RESIDENTIAL UNITS	COMMERCIAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	YEAR BUILT	SALE PRICE
count	48244.000000	48244.000000	4.824400e+04	4.824400e+04	48244.000000	4.824400e+04
mean	2.566537	0.249171	3.358117e+03	3.669753e+03	1827.765173	1.153281e+06
std	17.465481	10.988072	3.143590e+04	2.947491e+04	464.361153	1.340131e+07
min	0.000000	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000e+00
25%	1.000000	0.000000	1.413000e+03	8.280000e+02	1920.000000	8.042000e+04
50%	1.000000	0.000000	2.140000e+03	1.620000e+03	1931.000000	4.800000e+05
75%	2.000000	0.000000	3.071000e+03	2.520000e+03	1961.000000	8.300000e+05
max	1844.000000	2261.000000	4.228300e+06	3.750565e+06	2017.000000	2.210000e+09

Finally, for the feature selection, except from the neighborhood we drop all non-business related properties: tax status, transaction ID's, dates...

3. Methodology

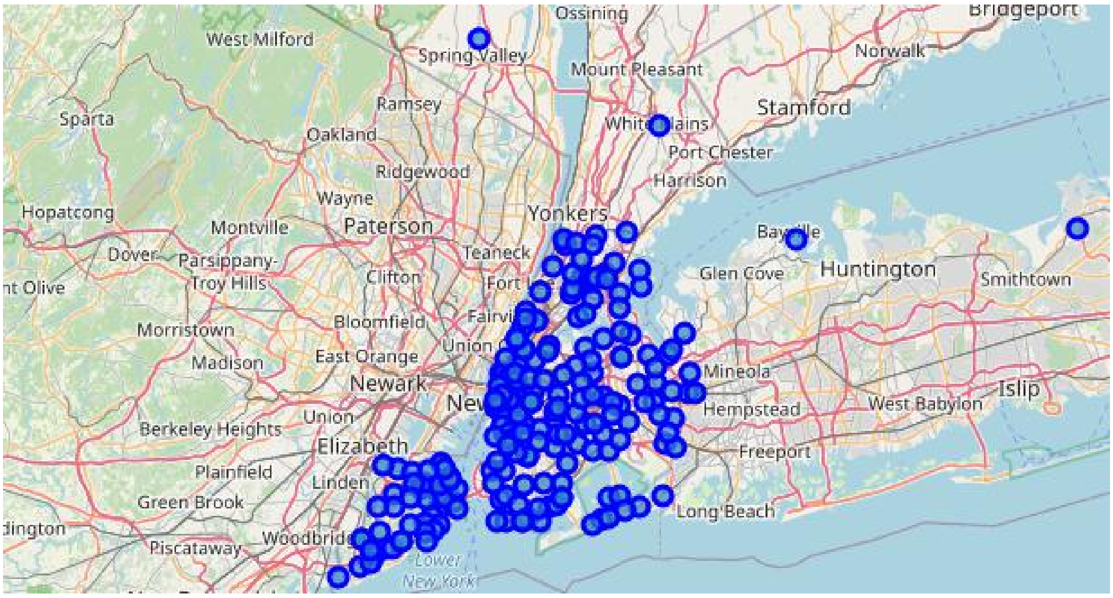
3.1 Exploratory Data Analysis

As stated before, exploring the data we see that some features as price we can see some outlayers, which means we can expect some clusters to have very few (or even individual) neighborhoods that contain particular real state markets.

In order to have a pair of latitude, longitude features so we can locate our clusters on the map, we use Four Square API and look for all neighborhoods that appear in our data.

NEIGHBORHOOD	latitude	longitude
ALPHABET CITY	40.725101	-73.979584
CHELSEA	40.746490	-74.001526
CHINATOWN	40.716492	-73.996254
CIVIC CENTER	40.713680	-74.002403
CLINTON	43.048405	-75.378502
...	...	...
TRAVIS	40.593159	-74.187920
WEST NEW BRIGHTON	40.634548	-74.112083
WESTERLEIGH	40.621216	-74.131813
WILLOWBROOK	40.603161	-74.138474
WOODROW	40.543438	-74.197647

And using Folium we show them on the map.



3.2 Clustering

I prepare the data by grouping it by neighborhood and using the mean value of the columns. After that I perform a z-score test, as this normalization technique allows to see how many sigmas above or below average a point is, and if it above or below the mean looking at its sign.

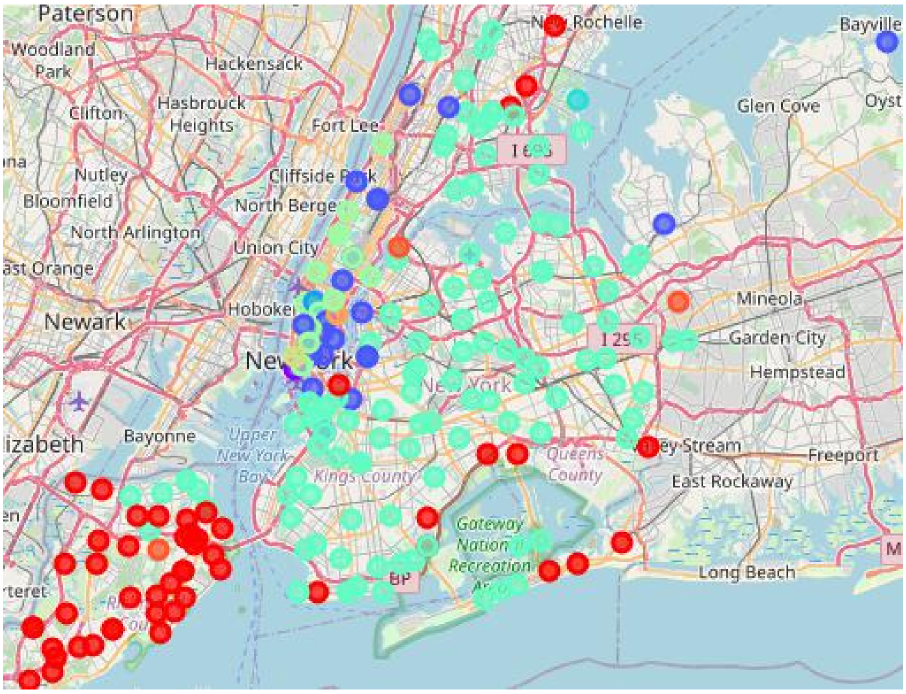
	NEIGHBORHOOD	RESIDENTIAL UNITS	COMMERCIAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	YEAR BUILT	SALE PRICE
0	AIRPORT LA GUARDIA	-0.027833	-0.023051	-0.057744	-0.075383	0.257996	-0.075124
1	ALPHABET CITY	0.297355	0.002033	-0.048576	0.102343	-0.435555	0.510731
2	ANNADALE	-0.081135	-0.021796	0.017222	-0.068912	0.888876	-0.081720
3	ARROCHAR	-0.076282	-0.014428	-0.016864	-0.068813	0.821332	-0.085918
4	ARVERNE	-0.058125	-0.022029	-0.032317	-0.071712	0.572430	-0.101734
...	...	...	...	...	...	...	...
220	WOODHAVEN	-0.059387	-0.019445	-0.048340	-0.074195	-0.431916	-0.083679
221	WOODLAWN	0.028829	-0.019601	-0.012508	-0.003436	-0.299247	-0.051804
222	WOODROW	-0.080563	-0.017180	0.055149	-0.053093	1.106755	-0.076934
223	WOODSIDE	0.031509	-0.003342	-0.039419	-0.014112	0.003825	0.003985
224	WYCKOFF HEIGHTS	0.024723	-0.009254	-0.041871	-0.012157	-0.318107	-0.014530

The machine learning model used was K-means, as it is an efficient unsupervised model which particularly works well with this kind of data.

After trying out different number of K's, the one that seemt to better fit the datawas K = 10, as it allowed to have big clusters of more 'typical' neighborhoods and then have some clusters of particular real state market neighborhoods.

4. Results

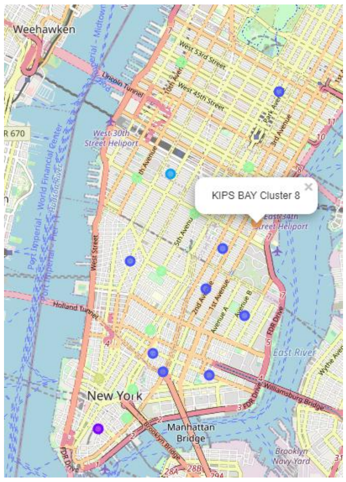
This is the map of clusters we obtain after performing K-means algorthm for K=10 in our dataset grouped by neighborhoods.



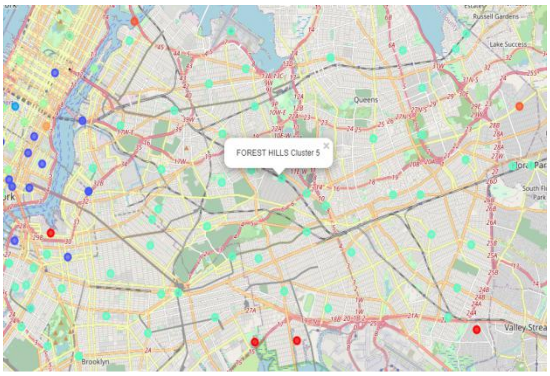
It's particularly interesting how the clustering map fits the district map of New York.

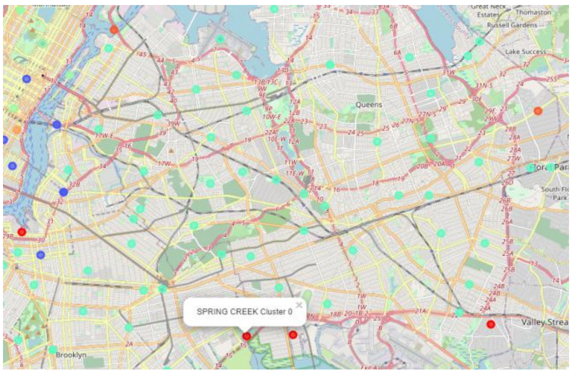
Source: Wikipedia

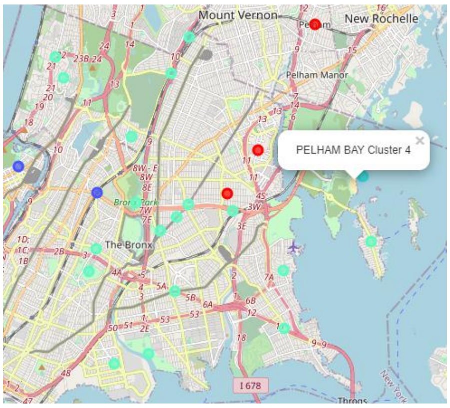
The clustering is particularly diverse in the Manhattan area, which contains six different clusters. This shows a very diverse and particular real state market for different areas of the island.



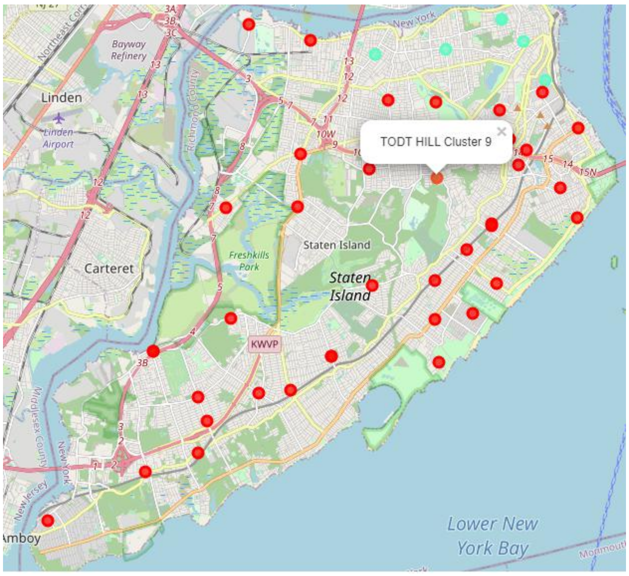
Then we have Brooklyn, Bronx and Queens with a similar and homogenous market, though with several exceptions along the coast.







Finally we have Staten Island, with a particular but also homogenous market.



5. Discussion

This is the average normalized properties of every cluster:

Cluster Labels	RESIDENTIAL UNITS	COMMERCIAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	YEAR BUILT	SALE PRICE
0	-0.076669	-0.014816	0.024778	-0.059899	0.635396	-0.084291
1	3.405901	2.598201	0.390365	13.230329	0.261414	21.382825
2	0.358508	0.041087	-0.006520	0.228732	-0.502612	0.570890
3	1.807157	0.834276	0.114838	4.213276	0.035964	8.926857
4	-0.146083	0.045930	9.147790	-0.003200	0.441828	-0.136968
5	-0.021957	0.003666	-0.003068	-0.030935	-0.119977	-0.034015
6	0.813605	0.092065	0.053578	0.887521	-0.338234	1.297307
7	0.056475	0.264367	0.147588	2.269567	-0.107117	5.683085
8	4.719419	1.529007	0.340317	2.819208	0.097142	7.178048
9	0.609836	0.016310	2.378295	1.055336	0.523045	0.254027

We see for example that the Manhattan clusters are the ones with higher prices, particularly the Cluster 1 (Financial district). Clusters 0 and 5, which are the most frequent are also the ones with more average values.

This table shows the number of neighborhoods for each cluster:

	NEIGHBORHOOD
Cluster Labels	
0	52
1	1
2	23
3	1
4	1
5	124
6	17
7	2
8	1
9	3

6. Conclusion

The study achieved the goal of having insights into the different real state markets you migh find by neighborhood in New York City: it shows the particular Manhattan environment, with a lot of diversity. In contrast wtih more 'regular' properties of Queens and Brooklyn districts.

Probably if we performed the algortihm over individual addresses we could increase the number of clusters, and therefore have more insights into the market.