

Clustering New York Neighborhoods by their Real State Market

Juan Tornero Lucas, December 2020



Introduction


The goal of this study is to give insights to a real state investor from Manhattan.

- Which areas are the best to invest according to your needs?
- Which is the main characteristics of the real state market in every neighbourhood?



Data

I use the data provided in the Kaggle repository [NYC Property Sales](#) . To prepare the data:

1. We clean the numeric columns that contain information about the transactions.
 2. Remove outliers of value 0.
 3. Drop all non-business related columns.
- 

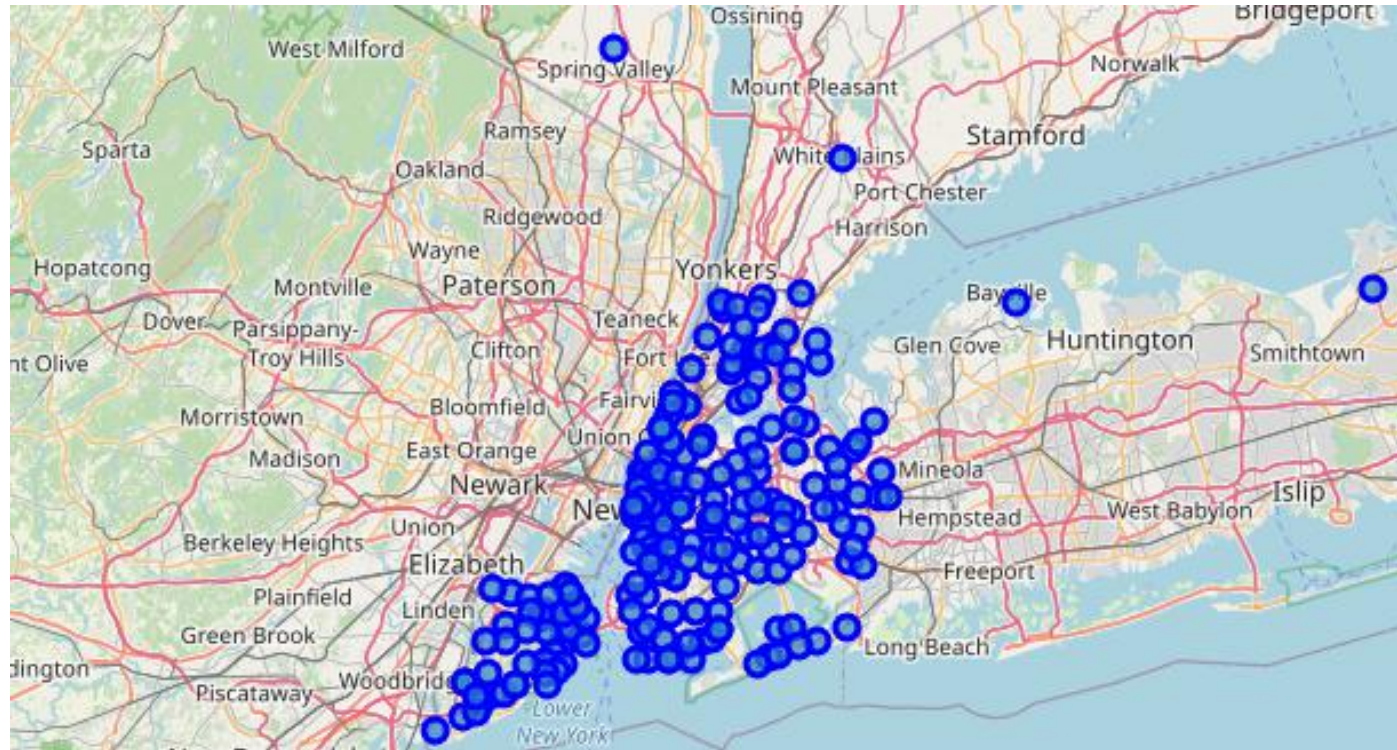
Exploratory Data Analysis

Some columns such as price show the presence of outliers. We expect that some clusters will present extreme values as main characteristics.



Exploratory Data Analysis

Using Four Square API we assign coordinates to all different neighborhoods to be clustered on the map.



Clustering


I perform a Z-Score to the data, 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

Clustering

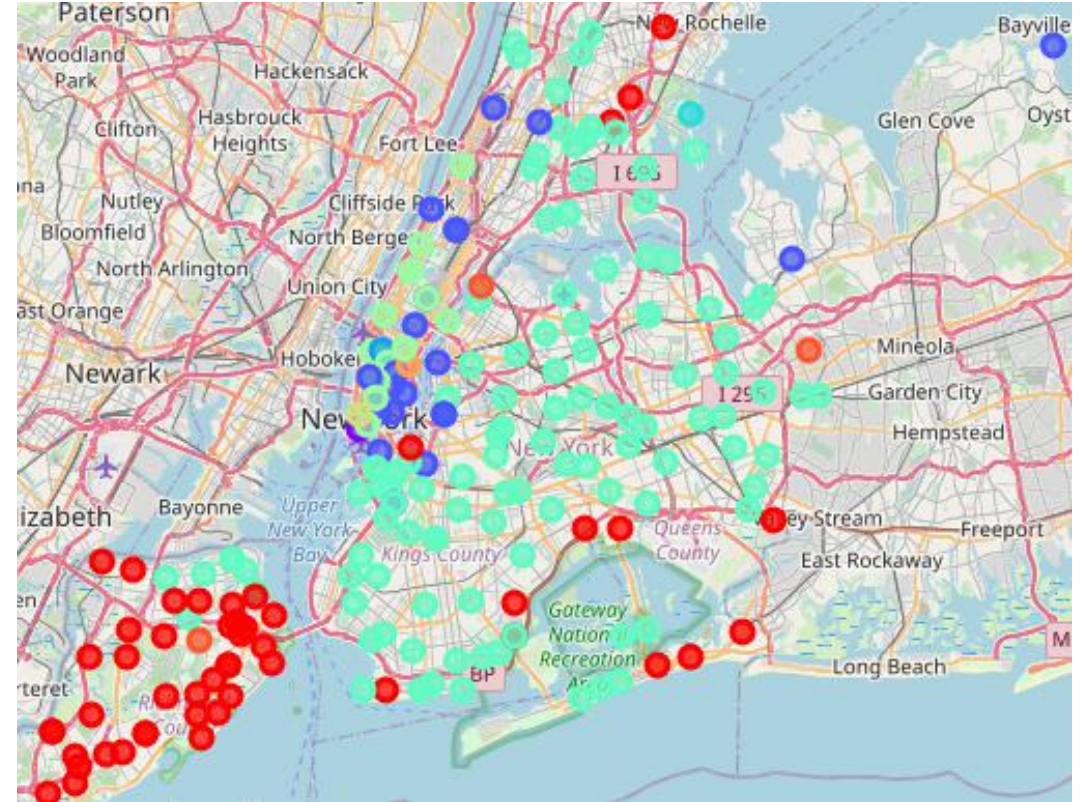
The machine learning model used was K-means, as it is an efficient unsupervised model which particularly works well with this kind of data.

After trying out different number of K's, the one that seemed to better fit the data was $K = 10$

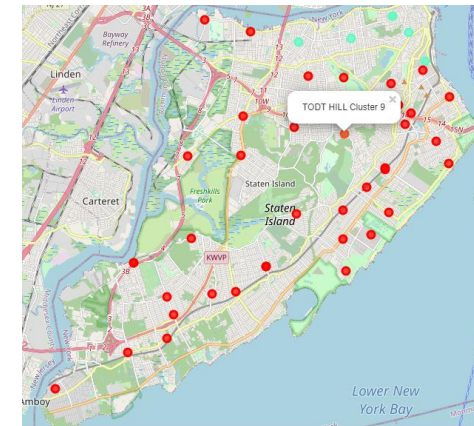
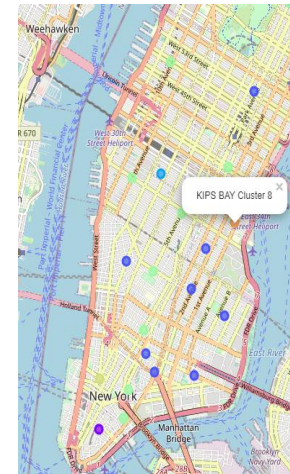
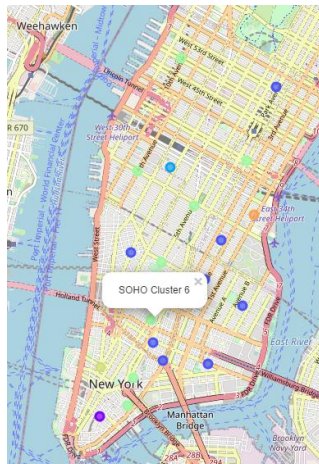
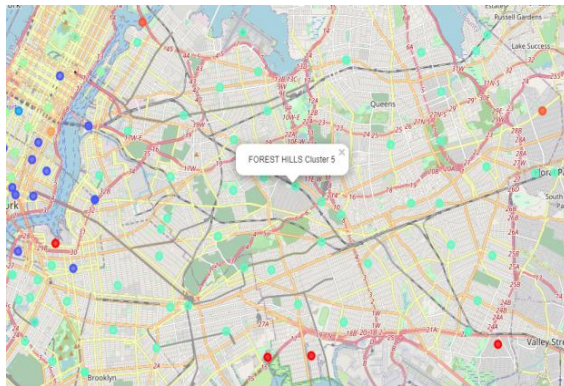
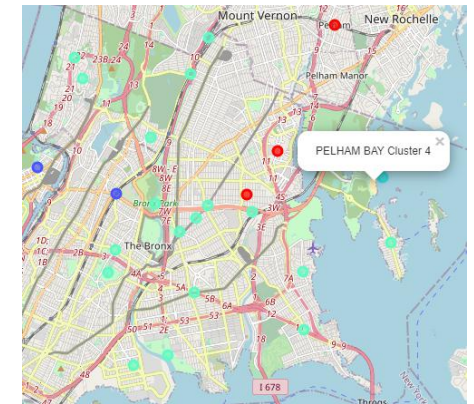
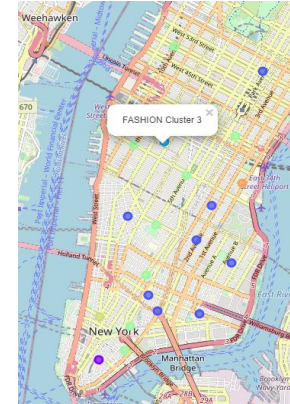
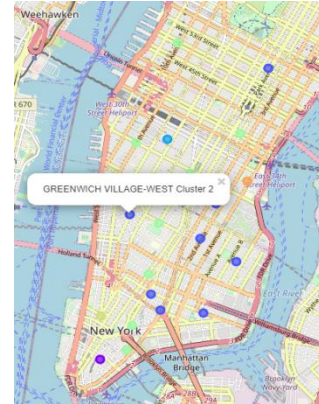
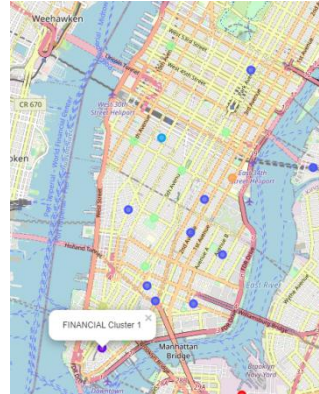
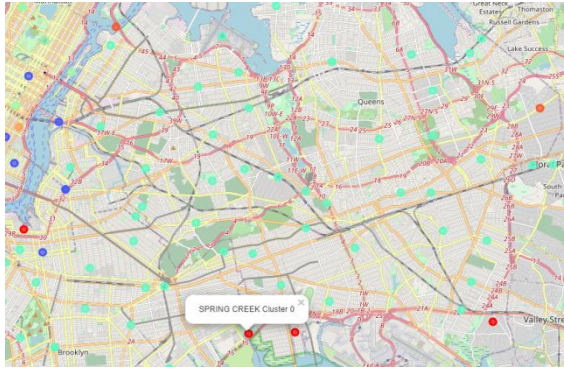


Clustering

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




Clustering



Discussion

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

- The clustering is particularly diverse in the Manhattan area, which contains six different clusters.
 - 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.
- 


Discussion

This is the average normalized properties of every cluster:

	RESIDENTIAL UNITS	COMMERCIAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	YEAR BUILT	SALE PRICE
Cluster Labels						
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

Discussion


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 most average values.



Conclusion

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

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



Thank You!