Data Analytics - Project 2

Christian Altrichter, Jury Andrea D'Onofrio

May 22, 2023

1 Pre-fix

This document provides a clustering analysis for the corpus "17_RealEstate". The document consists of the following features:

- Location
- Address
- Type
- Sale price
- Sale date of building units sold

Each of the above listed categories is subdivided into more features. In total, we have 21 features.

The properties are located in New York across the 5 following districts:

- Manhattan
- Bronx
- Brooklyn
- Queens
- Staten Island

The following columns create a unique key for a property in New York:

- Borough
- Block
- Lot

Thus, we should include these three features into our classifiers. However, as we do not aim to classify by individuality of the buildings, we remain open to the possibility of neglecting either of the features if not needed. Furthermore, as outlined in the ReadMe, there are various sales that represent transfer of deeds. Also, a sale can consist of individual units and / or the entirety of the property. Thus, this needs to be considered when handling the data.

2 Data description

We were given a corpus of real estate sales in New York which recorded the sales between the period of 01 September 2016 and 31 August 2017 (thus, a period of exactly 1 year). We are not aware of any special economic factors (e.g. real estate crash in 2008 or the COVID-crisis in 2020) that impact the real estate sales price in any particular way. Thus, the analysis can be done without any special consideration. The sales were split across 5 districts as follows:

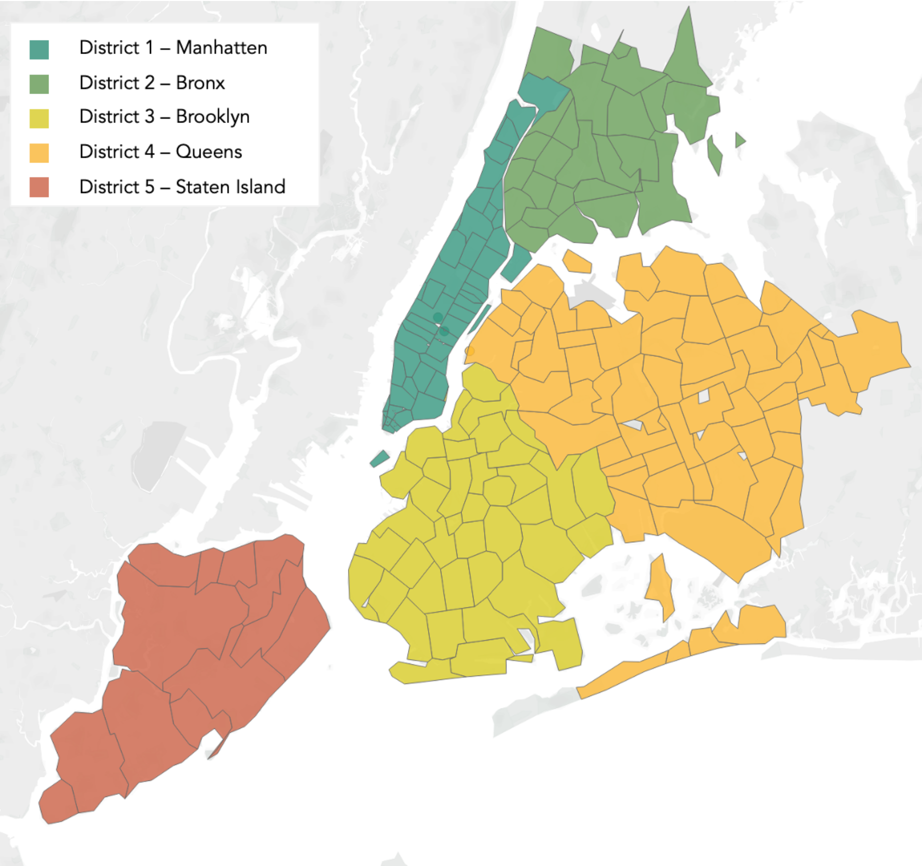


Figure 1: The 5 districts in which sales have taken place

In total, there were 84,548 entries in our data set. For each of the 5 boroughs, we could observe the following statistics:

Borough & ID	**# districts**	**Number of sales**	**% of all sales**	**Avg. sale price**	**StDev**
Manhattan (1)	39	18,306	21.64	$2,632,835	24,097,782
Bronx (2)	39	7,049	8.36	$590’194	2,783,057
Brooklyn (3)	61	24,047	28.44	$834’488	3,935,813
Queens (4)	59	26,736	31.62	$510’348	3,233,883
Staten Island (5)	58	8,410	9.94	$388’444	1,901,766
Total	256	84,458	100.00	$1,056,632	11,405,255

Table 1: Sales description by individual borough

It is evident, that the standard deviation is considerably high due to the various utilization’s of the buildings (e.g. commercial, residential, logistics etc.).

For the following features (that were not previously covered) there are the following amount of unique values present clustered by borough:

	Manhattan	**Bronx**	**Brooklyn**	**Queens**	**Staten Island**
# Building Class Category	44	35	43	41	29
# Tax Class at present	9	9	9	9	9

Table 2: Unique features by borough

Each of the following boroughs has the following mean with regards to the construction year of each sold property. It ought to be noted that we have not considered any entry where there are no values or the value is equivalent to 0.

	Manhattan	**Bronx**	**Brooklyn**	**Queens**	**Staten Island**
Avg. year sold	1952	1944	1943	1949	1969

Table 3: Unique features by borough

Furthermore, it must be mentioned that for the indication of the sale price the year of construction is not highly indicative as we do not know any renovations date and outstanding capital expenditures (abbreviate: CAPEX).

The following table displays cumulative statistics that occurred over the time span of 1 year:

	Manhattan	**Bronx**	**Brooklyn**	**Queens**	**Staten Island**
Residential Units	42,241 [**87%**]	24,225 [**95%**]	48,951 [**92%**]	45,491 [**88%**]	10,324 [**94%**]
Commercial Units	5,486 [**13%**]	1,134 [**5%**]	3,558 [**8%**]	5,517 [**12%**]	670[**6%**]
Total Units	48,504 [**100%**]	25,403 [**100%**]	53,400 [**100%**]	51,841 [**100%**]	11,016 [**100%**]
Land Square Feet	10,496,789	27,823,048	52,092,050	76,578,602	62,793,442
Gross Square Feet	63,904	31,892,145	61,809,317	56,885,502	15,854,601

Table 4: Sales description by individual borough

As each of the types of properties sold have a significance greater or equal than 5.00%, we should consider classifying the properties based on the type of building sold (residential vs. commercial). Furthermore, we see a fluctuation between the boroughs with respect to the share of each units sold based on its usage. Thus, we should also consider performing clustering based on the boroughs themselves.

3 Data Pre-Processing

The data consist of 21 features. For each of the features, we first looked into the official glossary and checked for the value's individual meaning. Based on that, we then decided how to process each vector respectively. In the following section, each of the features will be explored and an explanation will be given with regards to how we proceeded to process it. For each of the features that were not yet an integer, we have transformed them into integer values using the `pandas.factorize` method.

For the columns that have already been an integer value but are represented as an object, we applied the `.astype('int')` function to transform the values from object to integer. Also, as the data set is named as the sale prices in New York we will argue the relevance of each feature with respect to the sale price. Thus, we aim to perform clustering that will ultimately indicate a strong correlation between the features, the sale price and its labels.

For each of the features, we will provide its amount of individual values present in the corpus. This in return lets us consider its usefulness to the clustering. Here, too many and too few are considered to be detrimental to the clustering's performance.

Lastly, we analyze for each feature what pre-processing was necessary to make best use of this feature and, if present, omit any useless data points.

3.1 Borough

The borough describes the name in which the property is located. As aforementioned, it consists of 5 different unique values. We expect the borough to be a strong indicator for the sale price. There were no blank entries. No other pre-processing was necessary here.

3.2 Neighborhood

The neighborhood name is determined by the department of Finance during the course of valuing the properties. Thus, it does not coincide with the actual district. There are a total of 254 distinct neighborhood values. We expect the Neighborhood to be a strong indicator for the sale price. There were no blank entries. No other pre-processing was necessary here.

3.3 Building Class Category

The building class category defines a broader overview of the usages for each property. Also here we expect this to be a strong indicator towards the sale price. There are a total of 47 distinct building class values. There were no blank entries. No other pre-processing was necessary here.

3.4 Tax Class at Present

According to the glossary, there are 4 different tax classes that indicate the use of the property (Class 1, Class 2, Class 3, Class 4). We expect the tax class at present to not be a strong indicator for the sale price because it represents its old usage. Because we have the tax class at time of sale, we omit this column entirely.

3.5 Block

The (tax) block represents a sub-division of the borough. There are a total of 11,566 distinct block values. We expect the block to be weaker indicator for the sale price due to the high variety of distinct values. There were no blank entries. No other pre-processing was necessary here.

3.6 Lot

A (tax) lot is a sub division of a (tax) block, which represents a unique property location. There are a total of 2,627 distinct lot values. There were no blank entries. No other pre-processing was necessary here.

3.7 Easement

The column "easement" was left entirely blank. Thus, we have omitted this column.

3.8 Building Class at Present

The building class at present is used to describe a properties constructive use (thus, the initial purpose it was constructed with). There are a total of 166 distinct building class at present values. We expect the building class at present to not to be a strong indicator for the sale price because it represents its old usage. The value consists of two indicators, the letter which indicates the usage and the number which indicates the construction style. There were are 738 blank entries. Because we have the building class at time of sale, we omit this column entirely.

3.9 Address

There are a total of 67,563 distinct address values. We expect the address to be weak indicator for the sale price due to the high amount of unique values. There were no blank entries. No other pre-processing was necessary here.

3.10 Apartment Number

There are a total of 3,988 distinct apartment numbers. We expect the apartment number to be a weak indicator for the sale price because apartment numbers can repeat themselves if not connected with the exact borough, zip code and street. There were 69,496 blank entries. Due to the high amount of blanks, we have omitted this column entirely.

3.11 Zip Code

There are a total of 186 distinct zip codes. We expect the zip code to be a strong indicator for the sale price. There were no blank entries. No other pre-processing was necessary here.

3.12 Residential Units

The residential units indicates the number of residential units at the listed property. It does not make sense to infer the strength of this feature based on the distinct values as we are dealing with integer numbers. No other pre-processing was necessary here.

3.13 Commercial Units

The commercial units indicates the number of commercial units at the listed property. It does not make sense to infer the strength of this feature based on the distinct values as we are dealing with integer numbers. No other pre-processing was necessary here.

3.14 Total Units

The total units indicates the number of total units at the listed property. It does not make sense to infer the strength of this feature based on the distinct values as we are dealing with integer numbers. No other pre-processing was necessary here. There are 19,762 values with a total unit entry of "0". The ReadMe provided in the zip file indicates that a unit with a value of 0 indicates that the entirety of the building is affected.

However, because we do not know how many units this building consists of, we omit the columns of residential units, commercial units and total units. We can **only** omit these three features as we are relying on the sales price per square feet (which is explained in the following sub-points).

3.15 Land Square Feet

This indicates the land area of the property listed in square feet. It does not make sense to infer the strength of this feature based on the distinct values as we are dealing with integer numbers. We expect the land square feet to be a strong indicator for the sale price. There were are a total of 36,578 entries that either have an entry with value "0" or "-". Since this is a crucial indicator for the sale price that cannot be neglected for a clustering, we will omit all rows with an invalid entry.

3.16 Gross Square Feet

This indicates the total area of all the floors of a building as measured from the exterior surface of the outside walls of the building. It does not make sense to infer the strength of this feature based on the distinct values as we are dealing with integer numbers. We expect the gross square feet to be a strong indicator for the sale price. There were are a total of 39,029 entries that either have an entry with value "0" or "-". Since this is a crucial indicator for the sale price that cannot be neglected for a clustering, we will omit all rows with an invalid entry.

3.17 Year Built

This is the year of construction of the building. As we do not have any information on renovation over the years for older buildings, we expect this to be a weak indicator for the sale price. There were are a total of 6,970 entries with the value of "0".

3.18 Tax Class at Time of Sale

There are a total of 4 distinct tax class at time of sale values. We expect this to be a strong indicator for the sale price. There were no blank entries. No other pre-processing was necessary here.

3.19 Building Class at Time of Sale

The building class at time of sale is used to describe a properties constructive use (thus, the initial purpose it was constructed with). There are a total of 166 distinct building class at present values. We expect the building class at time of sale to be a strong indicator for the sale price. The value consists of two indicators, the letter which indicates the usage and the number which indicates the construction style. As the construction style (e.g. cape cod style for a house) is quite specific for a property comparison on this scale, we omitted the information and kept the usage information. There were no blank entries.

3.20 Sale Price [in USD]

The sale price indicates a sale price of the transaction. As initially mentioned, there are sale prices with a record of $0.00 (total amount: 10,228). There are furthermore various small transaction amounts which ought to be filtered out as we consider this to be noise in our data set. Here, we have have applied an initial filtering and removed all values below USD 50,000. This is a reasonable amount in our opinion in which no real estate can be bought within the New York real estate market. To further help us filtered out noisy data, we have created the below mentioned "Sale Price per Gross Square Feet".

3.21 Sale Date

This indicates the sale date between a period of one year. Here, since there were not specific economic conditions, we do not expect this to be a strong indicator for the sale price. There were no blank entries.

3.22 Sale Price [in USD] per Gross Square Feet - Created by us

Lastly, we have created another column which gives a true indication of the sale price with respect to other features. The following average prices could be observed per borough.

	Manhattan	**Bronx**	**Brooklyn**	**Queens**	**Staten Island**
Avg. SP per GSQF	1189	280	538	464	357

Table 5: Average Sale Price / Gross Square Feet

When looking at the real estate market reports, we have found the following statistics:

	Manhattan	**Bronx**	**Brooklyn**	**Queens**	**Staten Island**
Avg. SP per GSQF	1,548	389	1,114	1,284	475

Table 6: Average Sale Price [USD] / Gross Square Feet based on Elliman Report 2022 and 2023

However, as these reported prices are from the years 2022 and 2023 we have to remove the accumulated inflation from these prices such that a comparison is possible. According to statistics of the USD federal reserve, the US has incurred an inflation of 25.76%. Thus, we discounted average sale prices per borough based on todays market reports are as follows:

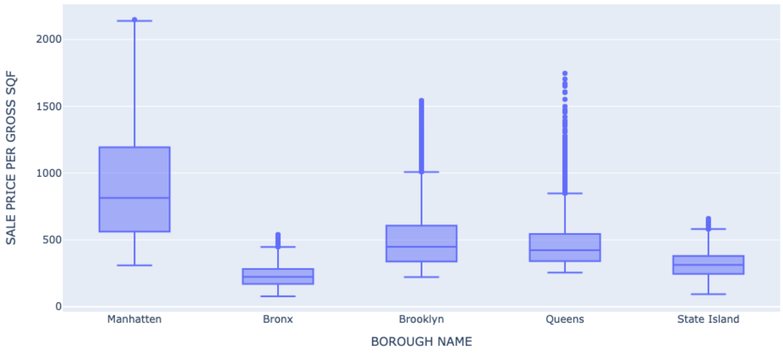


Figure 2: Box plot for Sale Price per Gross Square Feet

	Manhattan	Bronx	Brooklyn	Queens	Staten Island
Avg. SP per GSQF	1,230	309	885	1,020	377

Table 7: Average Sale Price [USD] / Gross Square Feet based on Elliman Report adjusted to inflation and discounted back to 2016

The discrepancy between actual average sale price per SQF in 2016 and the observed prices within our data set in 2016 is considered to be further noise to our data set. Thus, we will further eliminate data sets that are outside of ± 75% of the average sale price per SQF (according to the market reports).

While reading the official real estate reports of New York, we have found the following observation:

1. The market reports cluster their own reports per borough into different subareas to determine the average sale price per SQF. E.g. Brooklyn has been divided into 8 sub-clusters in one of the reports. Thus, going forward we should be cautious with applying the clustering algorithms to the entire data set instead of one specific borough.

4 Clustering

Prior to clustering, we wanted to get a feel of how clear the data is differentiable by its features. Thus, we have first created a heat map. Due to the high dimensional and complexity in our data set, we have used the Spearman correlation to compute the heat map (instead of using Pearson that can only deal with linear data and non-complex data). Also, any heat map that will be used from here on will use the Spearman correlation. The results can be shown as follows:

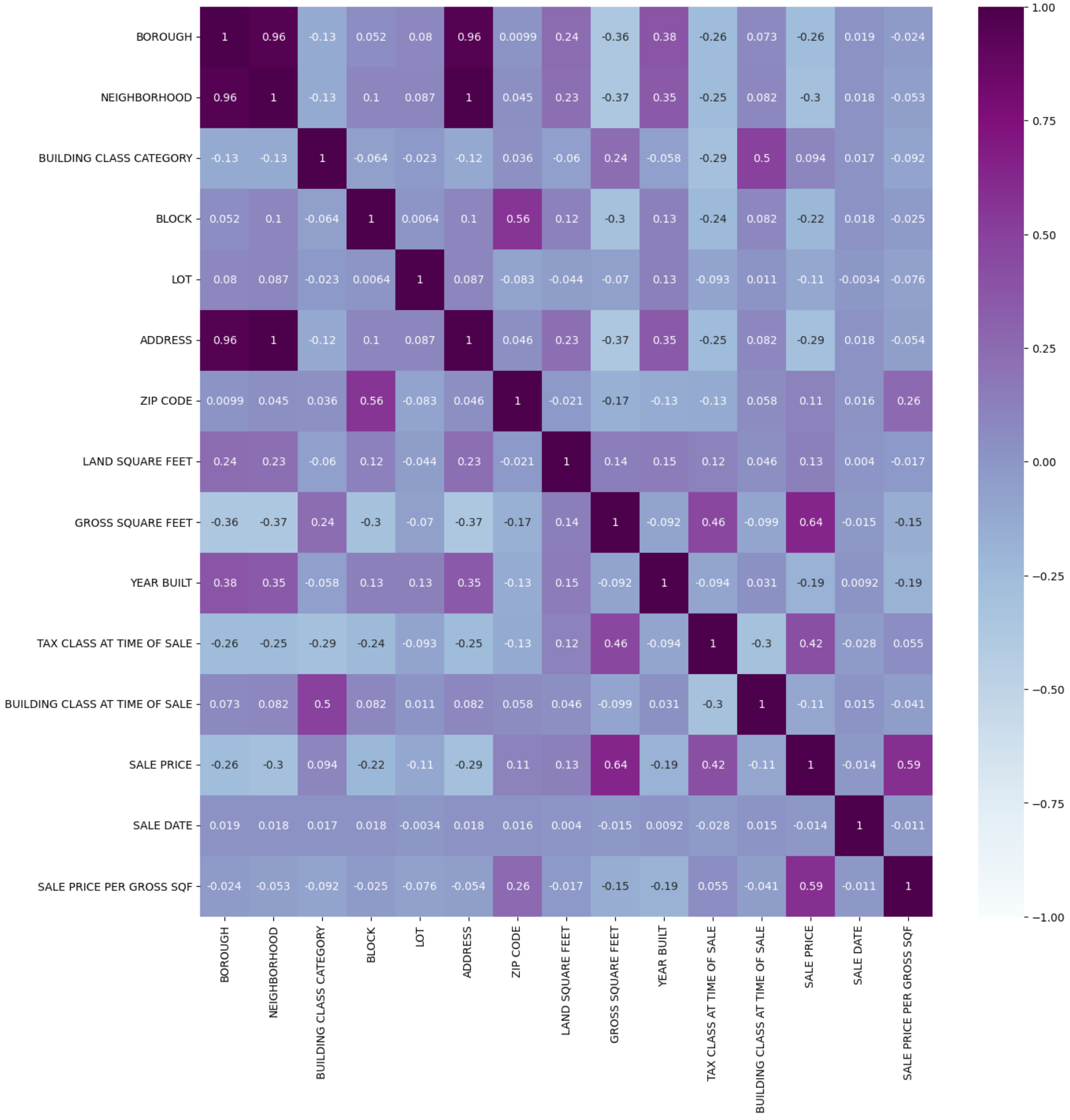


Figure 3: Heat Map of our Features

Quite evidently, there is very little correlation between the data set and the sole indicator we are interested in, the sale price per gross square feet. We assumed this to be due to the high complexity and information present within our data set. Thus, to truly be able to cluster based on prices and the features that are correlated with the prices, we wanted to exclude further data with little significance. According to various sources, the Spearman correlation and its importance can be interpreted as follows:

P-Value	**Interpretation**
p = 1.0	Perfect positive monotonic correlation
1.0 p ≥ 0.8	Strong positive monotonic correlation
0.8 p ≥ 0.4	Moderate positive monotonic correlation
0.4 p 0.0	Weak positive monotonic correlation
p = 0.0	Perfect positive monotonic correlation
0.0 p ≥ -0.4	Weak negative monotonic correlation
-0.4 p ≥ -0.8	Moderate negative monotonic correlation
-0.8 p ≥ -1.0	Strong negative monotonic correlation
p = -1.0	Perfect negative monotonic correlation

Table 8: Spearman correlation interpretation according to Newcastle University of England

Thus, we oriented ourselves along the above definitions and excluded all features where the Spearman correlation is weak. However, as the location for any real estate will have an ultimate impact on the sale price as well as the sale price per SQF, we have lowered our threshold of weak correlations to 0.25. We have therefore dropped all feature vectors that have a small correlation with respect to the sale price and the sale price per SQF. If any of the correlations between the two metrics fall below 0.25, we have dropped the respective feature. This resulted in the following new heat map:

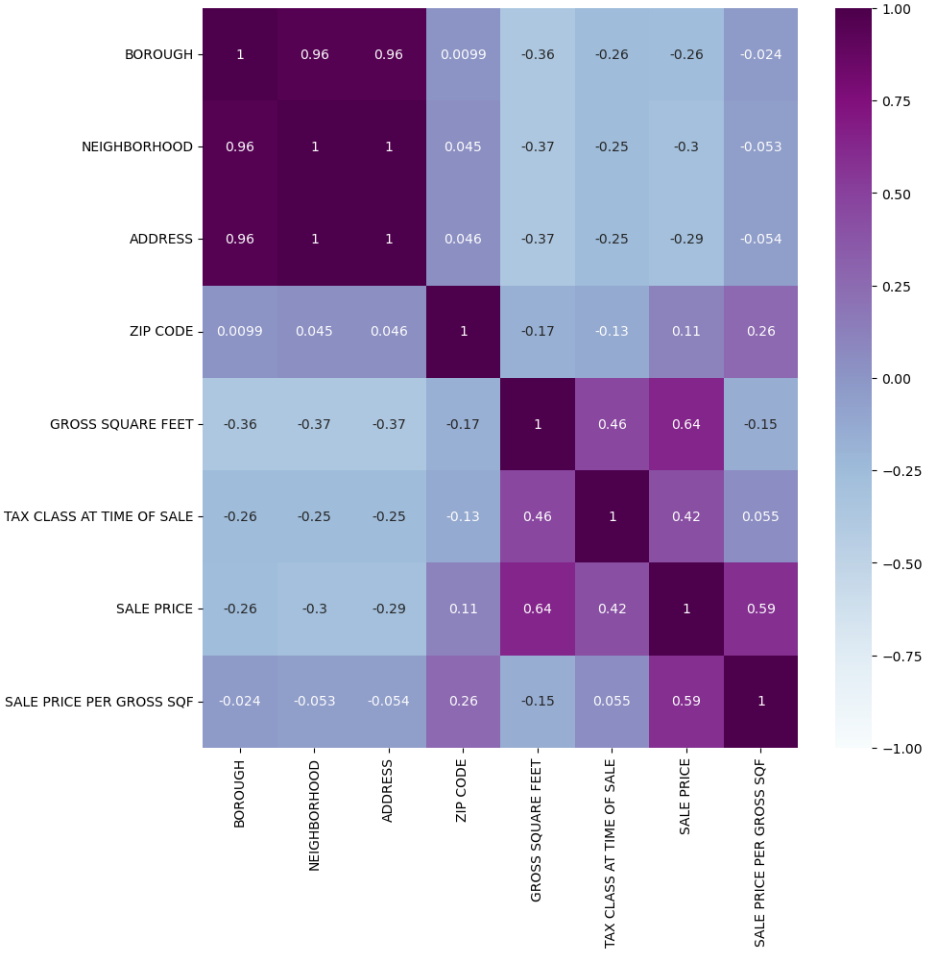


Figure 4: Heat Map of our Features - Post processing with a threshold of 0.25 respective to Sale Price and Sale Price per Gross Square Feet

4.1 PCA analysis & visualization

To understand our data better for the clustering, we performed a PCA analysis with two main components and visualized it in a two-dimensional space. Ideally, we would label the data based on a strong indicator of the sale price per gross square feet. As example, the zip codes would be the best indicator for the sale price per gross square feet. However, due to the high amount of unique values, we opted for the visualization in terms of boroughs. The result is as follows:

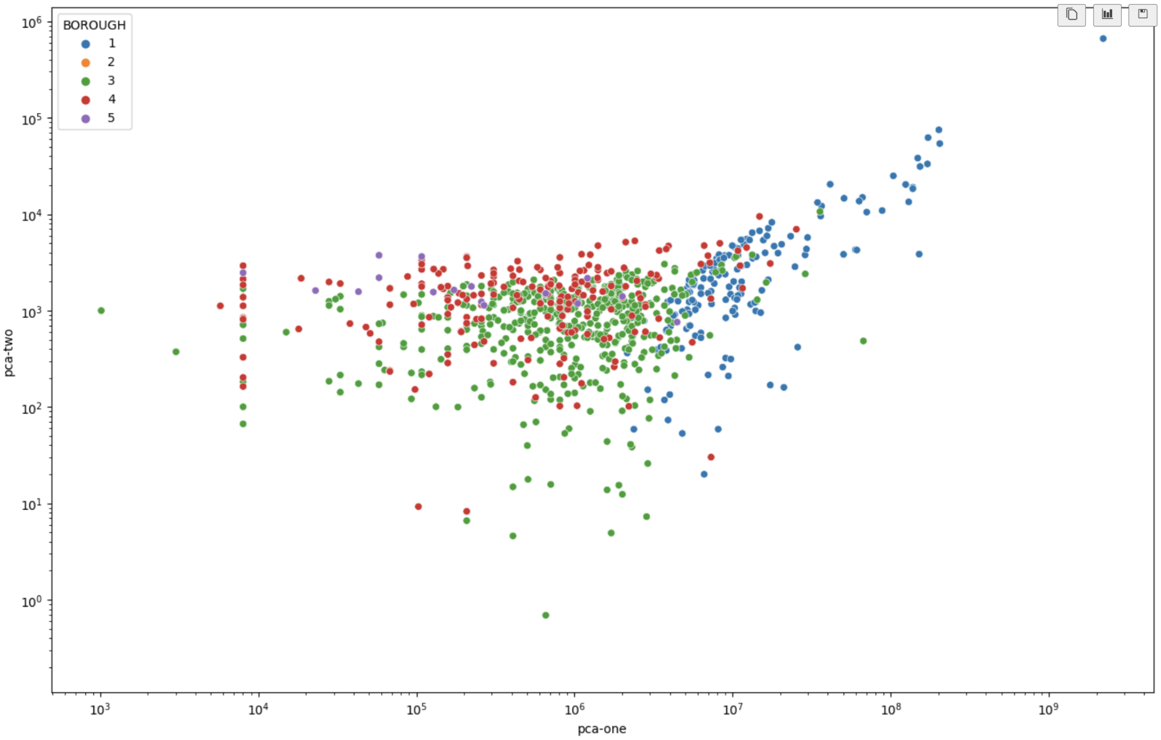


Figure 5: Principal Component Analysis with two dimensions

As initially mentioned in section 3.22 the clustering for a determining sale price per gross square feet based on e.g. borough is not fairly straight forward. This is due to the high dimensionality within our data set but also due to the high complexity. We have visualized the interconnection between the data points on a logarithmic scale. Even though it has been depicted on a logarithmic scale, one can see the high density between cluster points. It is interesting to observe however that Manhattan can be clearly distinguished from the other data sets.

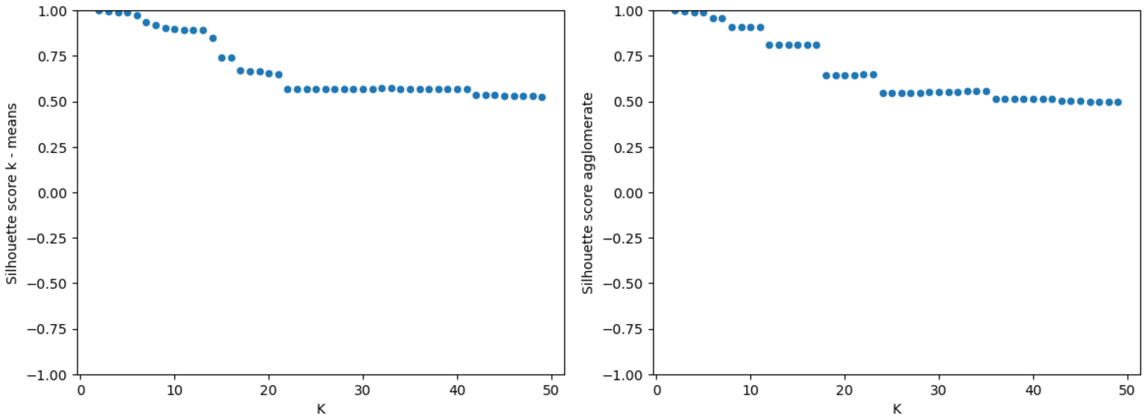
Before clustering, it ought to be mentioned that the clustering is an unsupervised learning problem. That means effectively that we do not possess a ground truth and can therefore not truly validate the clustering with respect to its features. Nonetheless, if possible, we will attempt to do so through the silhouette score, a visualization based on labels and a Spearman correlation analysis between the computed labels and the features.

4.2 K-Means and Agglomerate Clustering

As initial clustering, we have opted to analyze our data based on two very rudimentary clustering algorithms, k-means and agglomerative clustering. For both clustering we measured its performance respectively based on the silhouette score. Subsequently, we have compared the labels to existing features in a heat map to analyze any potential existing correlation between the clustering and the other feature vectors.

Both clustering algorithms have been performed for a range of k = 2 up to but not including k = 50. The silhouette scores are respectively as follows:

4.2.1 Testing Various K and reporting their respective Silhouette Scores



(a) Silhouette score k-means

(b) Silhouette score agglomerate

When taking a look at the cluster size now we can depict the following:

Cluster #	**Cluster Size**
Cluster 0	23973
Cluster 1	1
Cluster 2	1
Cluster 3	48
Cluster 4	3

Table 9: Cluster size based on labels - K-Means

Cluster #	**Cluster Size**
Cluster 0	23976
Cluster 1	45
Cluster 2	1
Cluster 3	3
Cluster 4	1

Table 10: Cluster size based on labels - Agglomerative

Subsequently, the heat map for the following two have been created and depicted. We have chosen k=5 as the silhouette score here is quite high and it would also represent the amount of boroughs (should there exist a correlation). Thus, the heat map is as follows:

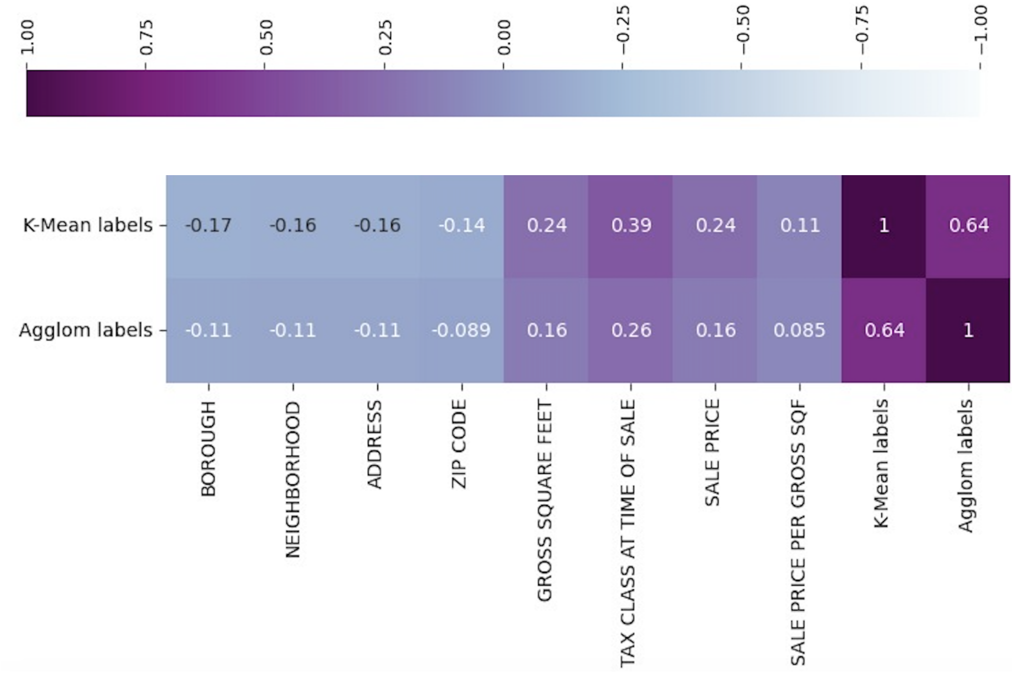
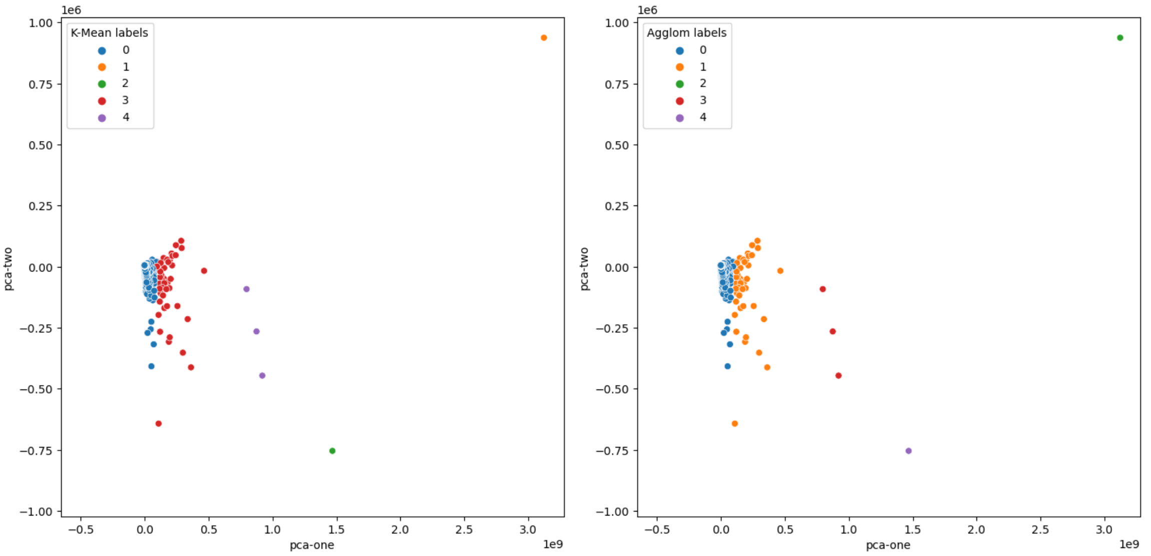


Figure 6: Heat map including the labelling of the two clustering algorithms at $\mathbf{k} = \mathbf{5}$

Here, it can be observed that both clustering algorithms have clustered based on one particular feature as they have the strongest correlation with respect to the tax class at time of sale (based on max k-means Spearman correlation). Further observations are that each feature has received some significance and correlation to the labels. Depending on the application, having importance on the other features in the clustering can be either seen as positive or negative. Both clustering can be visualized with the previously computed PCA analysis as follows:



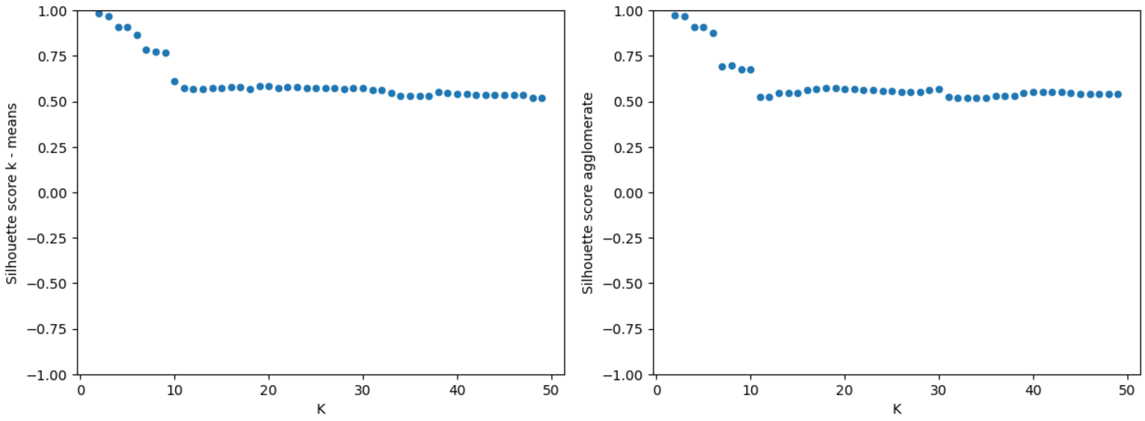
(a) Principal Component Analysis k-means depiction (b) Principal Component Analysis agglomerative depiction

When visualizing the clustering, it can be seen that the clusters are quite dense and the sizes of each cluster is not well distributed.

Thus, to improve on our clustering with respect to sale price and sale price per gross square feet, we have taken Manhattan as individual component and performed another clustering algorithm on this data set alone. We have opted to isolate Manhattan as it could also be differentiated in the previous PCA analysis from the other boroughs. Ideally, going forward each clustering should be done on each individual borough separately. However, due to the limited scope of this assignment, we have opted to only choose one (thus, Manhattan) and perform a more detailed analysis of such.

4.2.2 K-Means and Agglomerative on one borough: Manhattan

When applying the clustering algorithms on only one borough, we cannot observe a distinguishable difference within the silhouette scores.



(c) Silhouette score k-means

(d) Silhouette score agglomerate

When taking a look at the cluster size now we can depict the following:

Cluster #	**Cluster Size**
Cluster 0	647
Cluster 1	1
Cluster 2	1
Cluster 3	30
Cluster 4	3

Table 11: Cluster size based on labels - K-Means

Cluster #	**Cluster Size**
Cluster 0	30
Cluster 1	647
Cluster 2	1
Cluster 3	3
Cluster 4	1

Table 12: Cluster size based on labels - Agglomerative

For both cluster sizes we could not see a clear improvement to the previous clustering based on the entire data set.

However, when taking a look at the heat maps with respect to the labels, we can observe the following patterns. Both algorithms clearly found distinguishable patterns with higher correlations compared to the clustering performed on the entire data set. For k-means, the patterns were clustered in particular by gross square feet, tax class at time of sale and sale price. Similarly, the agglomerative clustering identified also the three classes in terms of absolute correlation as prime indicator for its clustering. Furthermore, the heat map distinguishes itself better from the clustering based on the entire data set, as it drops importance on certain features (e.g. neighborhood and address). Which means that the clustering clearly focuses on only three features rather than having a focus on three but still assigning importance to others.

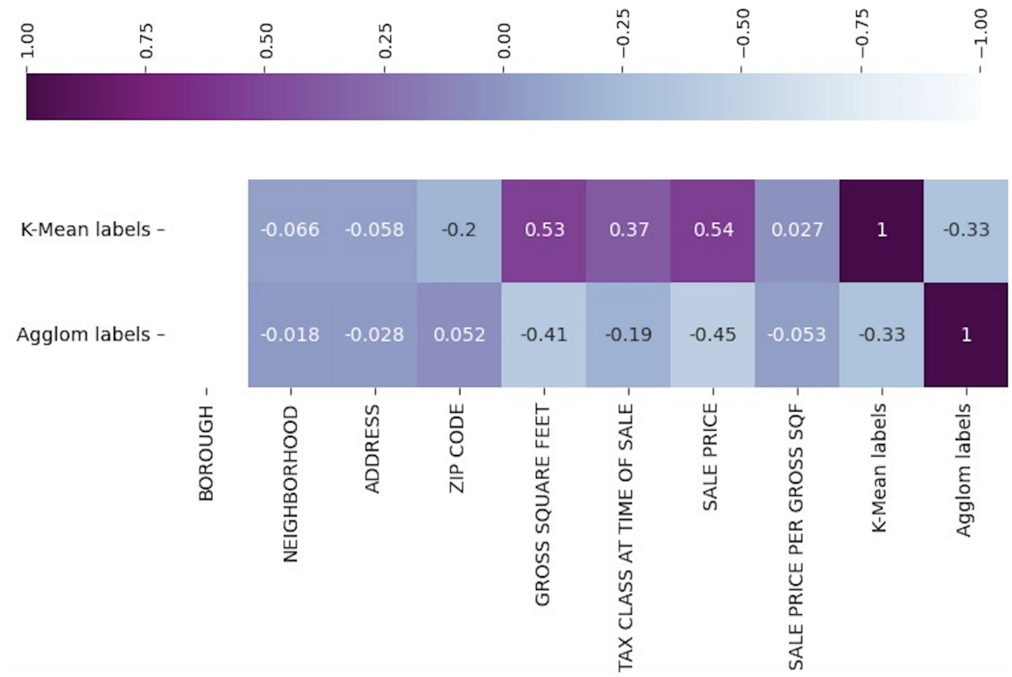
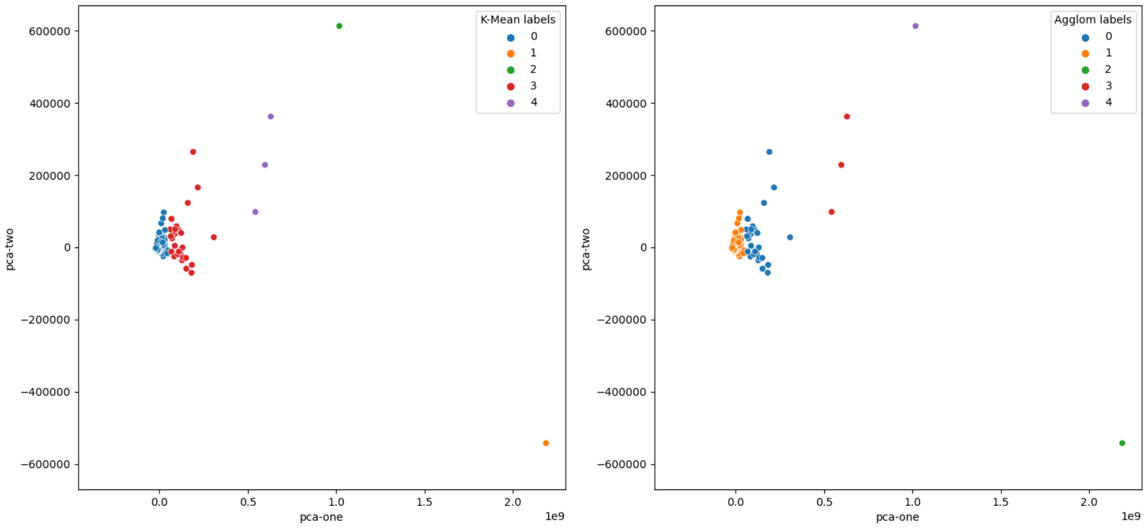


Figure 7: Heat map including the labelling of the two clustering algorithms at $\mathbf{k = 5}$

Thus, the k-means and agglomerative clustering performed much better when a borough is being separated. When taking a look at the PCA analysis in terms of labels, we cannot observe any clear difference to the previous clustering attempt.



(a) PCA k-means depiction on one borough - Manhattan (b) PCA agglomerative depiction on one borough - Manhattan

5 Spectral clustering

As previously seen in k-means and agglomerative clustering, the clustering algorithms are inconclusive with respect to the entire data set. Therefore, we will continue to perform the clustering algorithm based on only one borough. Here, we will use as example borough ”Manhattan”. The results can then actively be compared between the clustering of borough Manhattan through K-Means and agglomerative clustering and Spectral clustering.

Both of the previous clustering suffer from high dimensionality and can therefore not cluster effectively and efficiently. In addition, both algorithms are sensitive to noise, in particular the agglomerative clustering. As researched, there are various clustering algorithms that perform better on high dimensional data. We have opted for the Spectral clustering as it is outlined to be one of the most robust graph-based clustering algorithms.

5.1 Testing Various K and reporting their respective Silhouette Scores

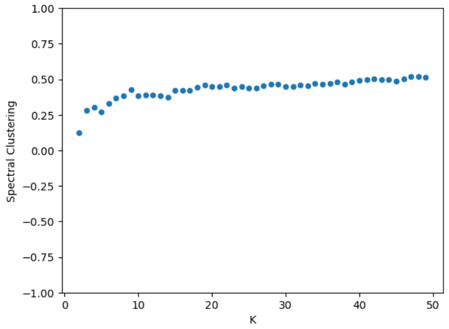


Figure 8: Silhouette score spectral clustering

With respect to the silhouette score, we observed that the spectral clustering converges faster to a silhouette score and performs more stable with multiple iterations. However, comparing the actual silhouette score with k-means and agglomerative clustering, both perform relatively equal. All three clustering algorithms converge around a silhouette score of 0.5. As this indicates of how well each individual data point fits within its cluster, we can now conclude that despite applying a more sophisticated clustering algorithm, the data might still be in a too high dimensional space. However, as already previously seen, the silhouette score is not the sole indicator for the success of a clustering algorithm.

Cluster #	**Cluster Size**
Cluster 0	144
Cluster 1	168
Cluster 2	117
Cluster 3	159
Cluster 4	94

Table 13: Cluster size based on labels - Spectral

As can be seen in the above table, the clustering behaves much more stable with respect to its cluster sizes.

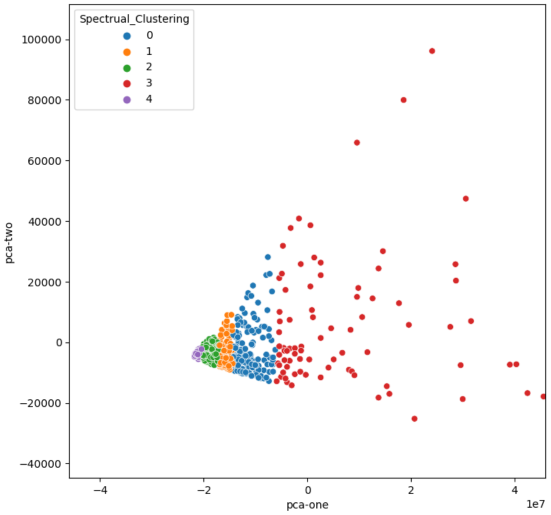


Figure 9: Principal Component Analysis spectral depiction on one borough - Manhattan

When looking at the depiction of the PCA components with respect to the labeling of the clusters, we observed a much clearer clustering pattern with respect to the data points. Also, we could observe that the cluster size is more equally distributed compared to k-means and agglomerative clustering. Thus, in this respect spectral clustering outperformed the trivial clustering algorithms.

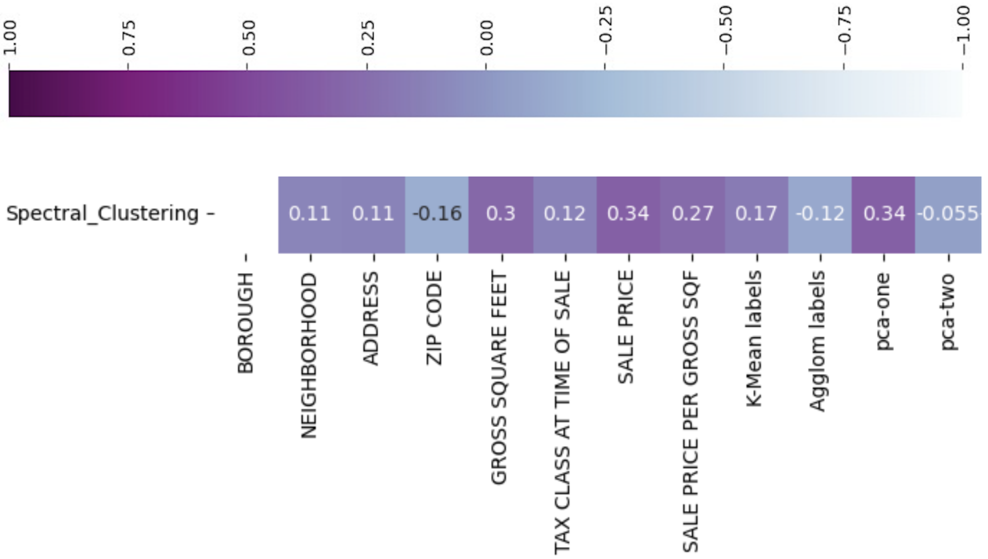


Figure 10: Heat map including the labelling of the spectral clustering algorithms $\mathbf{k} = \mathbf{5}$

Lastly, when inspecting at the Spearman correlation matrix with respect to the individual features and the spectral clustering labels, we observed that the spectral clustering is more stable with respect to individual features. Here, it found significance in all features. Comparing this k-means and agglomerative clustering, we have found that these clustering algorithms have put more emphasis during their clustering on one a few features.

Both approaches possess benefits and disadvantages. Thus, depending on the purpose of the clustering, one should either choose one or the other.

6 Outlook & Recommendations

As the real estate market is highly complex, the clustering could be also split into further categories depending on the desired outcome of the clustering (e.g. split data set by tax class at time of sale (thus, usage).

Second, further data could be gathered to have more information on the variables that influence ultimately the sale price. Here, we were only given for instance the year of construction. This in itself is a weak indicator for the sale price and the sale price per gross square feet as we are missing the year of a full scale and / or partial renovation. Here, we could furthermore include a feature which provides information about the capital expenditures within the property. This would provide further context on the specific sale price.

Going forward, one could now take the data that has previously been omitted from our data set due to missing or irrelevant sale prices and predict the sale price on the feature vector and observe whether the clustering results (due to more amount of data) gets any better. However, as this is beyond the scope of our project, we will not perform a regression with respect to the sale price.