NYC Property Sales

Big Data Fundamentals

Vinina Sunny

MSc Data Analytics

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Most of the current trending technologies have their foundation in the concepts of Big Data, Deep Learning and Machine learning for analysis of complex, critical and large data sets to reveal trends, patterns and give predictions which otherwise are too complex to deal with by using traditional data processing applications.

The real estate sector has become one of the most unpredictable sectors due to the enormous fluctuation in the house prices. The real estate industry being immensely important due to the range of stakeholders from private companies, investors to regulatory bodies it is in high demand to get better understanding of the factor impacting the prices as well as get a prediction on the future trend. This report focuses on applying the idea of machine learning to build a model to predict the future sale value of New York City real estate properties.

This dataset contains the record of every apartment or building sold in New York City market over the period of 12 months. The dataset is a cleaned and concatenated version of New York City's Department of Finance's Rolling Sales. The data mainly provides the following information of the apartment/ building units sold:

- Sale Price
- Location
- Type
- Address
- Sale Date

It also provides references of some complex fields

- BOROUGH: Digit reference of the borough the property is located in; in order these are Manhattan (1), Bronx (2), Brooklyn (3), Queens (4), and Staten Island (5).
- BLOCK; LOT: In New York there is a unique key called BBL elaborated as combination of borough, block, and lot

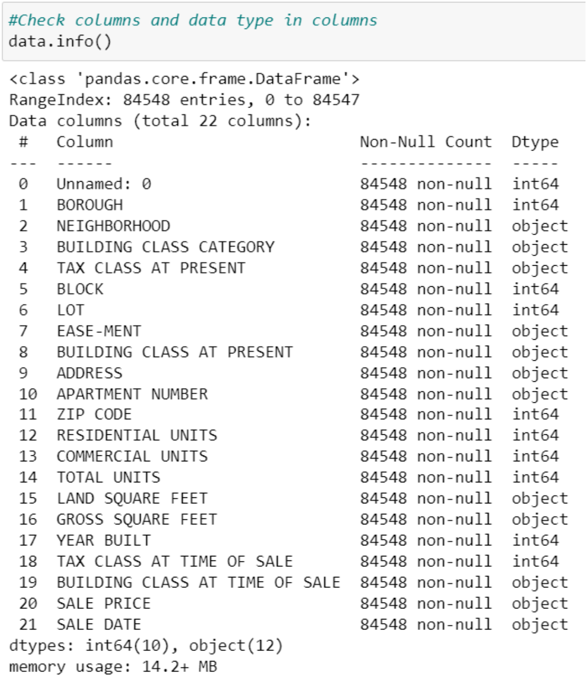
- BUILDING CLASS AT TIME OF SALE & BUILDING CLASS AT PRESENT: Building Type at various points in time.

Analysis of the data should give us a proper overview on the price fluctuations and help in predict the future in New York City Market. To further proceed with the model, it is necessary to deal with few key problems and challenges in the dataset which is explained in the next chapter.

Chapter 2

Key challenges and Problems

One important issue to deal with while analysing the data was doing the missing value analysis and preparing and cleaning up the data to properly give to the model. This was especially necessary while dealing with the Sale Price as it was the target feature to be used to predict the future value of houses. While checking for null values in the dataset, it stated that there were zero null values though on further investigation the few redundant values were identified from the dataset in certain columns. In our case, instead of null values there were '-' or whitespaces which made cleaning a bit harder task. Approximately 17% of the dataset had '-' or equivalent null value in Sale price column and which had to be dropped to make the dataset a bit more precise.



```
#Check columns and data type in columns
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 84548 entries, 0 to 84547
Data columns (total 22 columns):
 #   Column                          Non-Null Count  Dtype 
---  ------                          --------------  ----- 
 0   Unnamed: 0                      84548 non-null  int64 
 1   BOROUGH                         84548 non-null  int64 
 2   NEIGHBORHOOD                    84548 non-null  object
 3   BUILDING CLASS CATEGORY         84548 non-null  object
 4   TAX CLASS AT PRESENT            84548 non-null  object
 5   BLOCK                           84548 non-null  int64 
 6   LOT                             84548 non-null  int64 
 7   EASE-MENT                       84548 non-null  object
 8   BUILDING CLASS AT PRESENT       84548 non-null  object
 9   ADDRESS                         84548 non-null  object
 10  APARTMENT NUMBER                84548 non-null  object
 11  ZIP CODE                        84548 non-null  int64 
 12  RESIDENTIAL UNITS               84548 non-null  int64 
 13  COMMERCIAL UNITS                84548 non-null  int64 
 14  TOTAL UNITS                     84548 non-null  int64 
 15  LAND SQUARE FEET                84548 non-null  object
 16  GROSS SQUARE FEET               84548 non-null  object
 17  YEAR BUILT                      84548 non-null  int64 
 18  TAX CLASS AT TIME OF SALE       84548 non-null  int64 
 19  BUILDING CLASS AT TIME OF SALE  84548 non-null  object
 20  SALE PRICE                      84548 non-null  object
 21  SALE DATE                       84548 non-null  object
dtypes: int64(10), object(12)
memory usage: 14.2+ MB
```

Figure 2.1: Quick information on values in Dataset

Other major problem was dealing with categorical values of the dataset, on analysis it was identified columns namely ‘EASE-MENT' and 'APARTMENT NUMBER' mainly had empty entries which was not stated in the initial analysis. Few numerical columns also had ’-’ and backspaces which needed to be cleaned. Some of the data made no sense such that having value 0 in ZIP CODE, TOTAL UNITS, GROSS SQUARE FEET and YEAR BUILT columns, also there were 0 as the Sale Price value. Depending on the missing values it was not always feasible to just drop those rows as it might have major impact on the prediction, like if majority of the data is missing removing those values or columns from the dataframe may cause severe information loss. If we consider the feature ‘GROSS SQUARE FEET’ which is extremely important to predict the value, dropping all missing values will drastically reduce the relevant information available. Another approach for the same could be filling missing values with the mean of other values which though might not always be appropriate.

The key purpose of the report is to identify the trends in Sale price using 12-month historical data of the New York city market by analysing the data against all the independent parameter and variables and further analysing it using a supervised and unsupervised methods. Beginning with unsupervised learning analysis, k-means clustering analysis was performed on the dataset to give a clear idea on the classification impacting the Sale price of the houses. Although the while performing the analysis it was understood that such analysis cannot be performed on the dataset for which the reason will be explained in the latter part of the report. For supervised training linear regression analysis will be conducted on the dataset to give a predication on the trend of the Sale Price.

Chapter 3

Exploratory Data Analysis and summary of the dataset

The dataset consists of 84548 rows and 22 columns which implies that dataset itself is quite large. The dataset simply has too many things to explore, we will first check to see which categories hold the most observations and then plot relationship target and features.

Firstly, we will start with 'TAX CLASS AT PRESENT' feature and check the total sale of unit for every tax class and plot the relationship. Also, the sum sale price obtained for each tax class apartments.

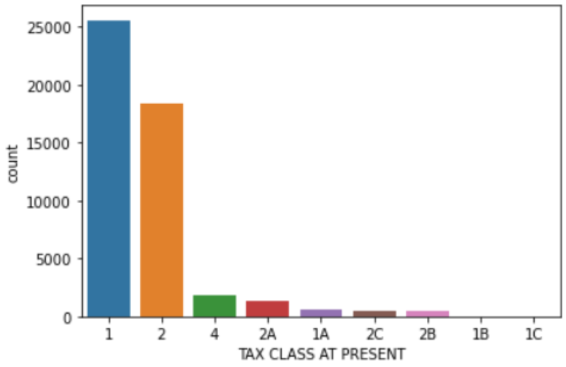


Figure 3.1: Sale Count vs Tax Class

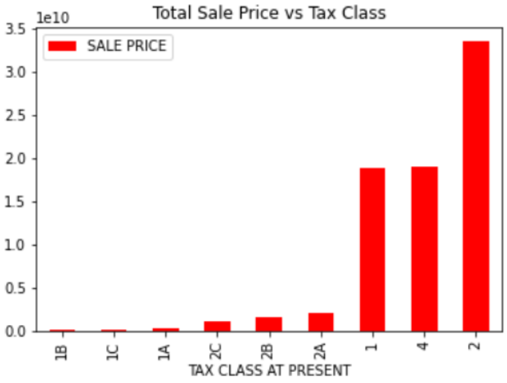


Figure 3.2: Total Sale Price vs Tax Class

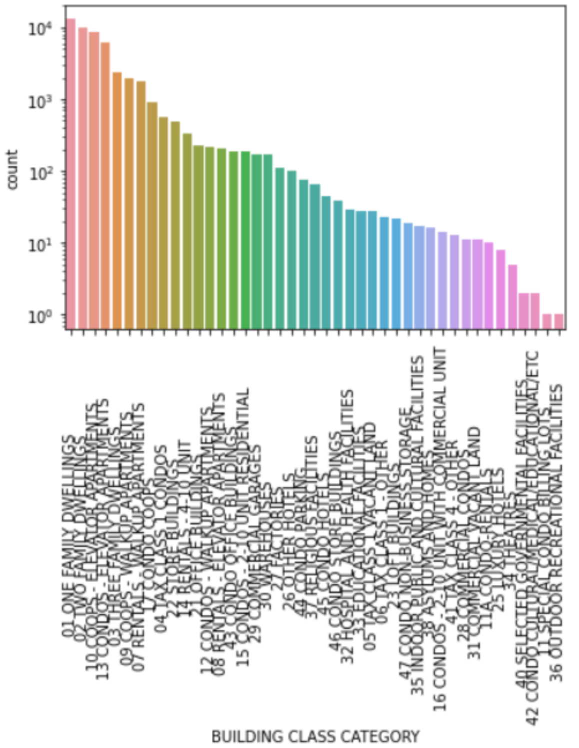
It can be seen that the maximum number of units sold are under Tax class 1 while the cumulative sale price obtained for the Tax class 1 apartment was comparatively less as to units having Tax class as 2 now though the count of sale for 1 was high.

On the contrary if we check the figures for Sale price using the feature TAX CLASS AT TIME OF SALE the highest can be seen for units which were under Tax class 4 at the time of sale. (See Table 3.1)

Table 3.1: Tax Class time of Sale with aggregate Sale price

TAX CLASS AT TIME OF SALE	SALE PRICE(mean)
1	599000.0
2	985000.0
4	1260000.0

Now, let’s consider another independent variable ‘BUILDING CLASS CATEGORY’ which we can compare with Sale price.



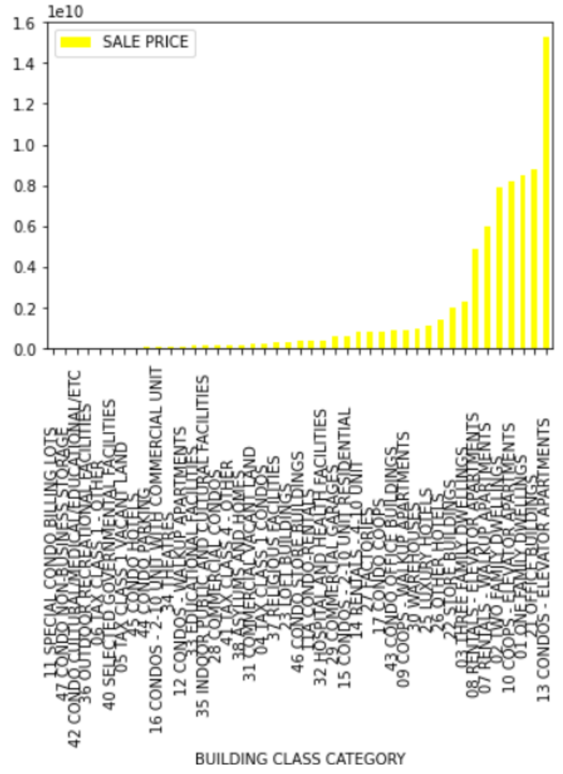


Figure 3.3 : Units sold as per Buidling class category

Figure 3.4 : Sale Price as per Category

One Family Dwelling apartments were the most sold in the span of 12 months (see Fig 3.3) whereas most costly building class category was for Condos – Elevator Apartments (see Fig 3.4). As this feature has many unique values, considering most representative 5 values to reduce the complexity of the model.

Furthermore, the heatmap of the numerical values shows that there is a huge correlation between LAND SQUARE FEET and GROSS SQUARE FEET and between TOTAL UNITS and RESIDENTIAL UNITS (see fig. 3.4) as the sale price increases with the square feet of the apartment and number of units available with respect to that. Only anyone of these correlated features needs to be considered as having both will not further improve the model.

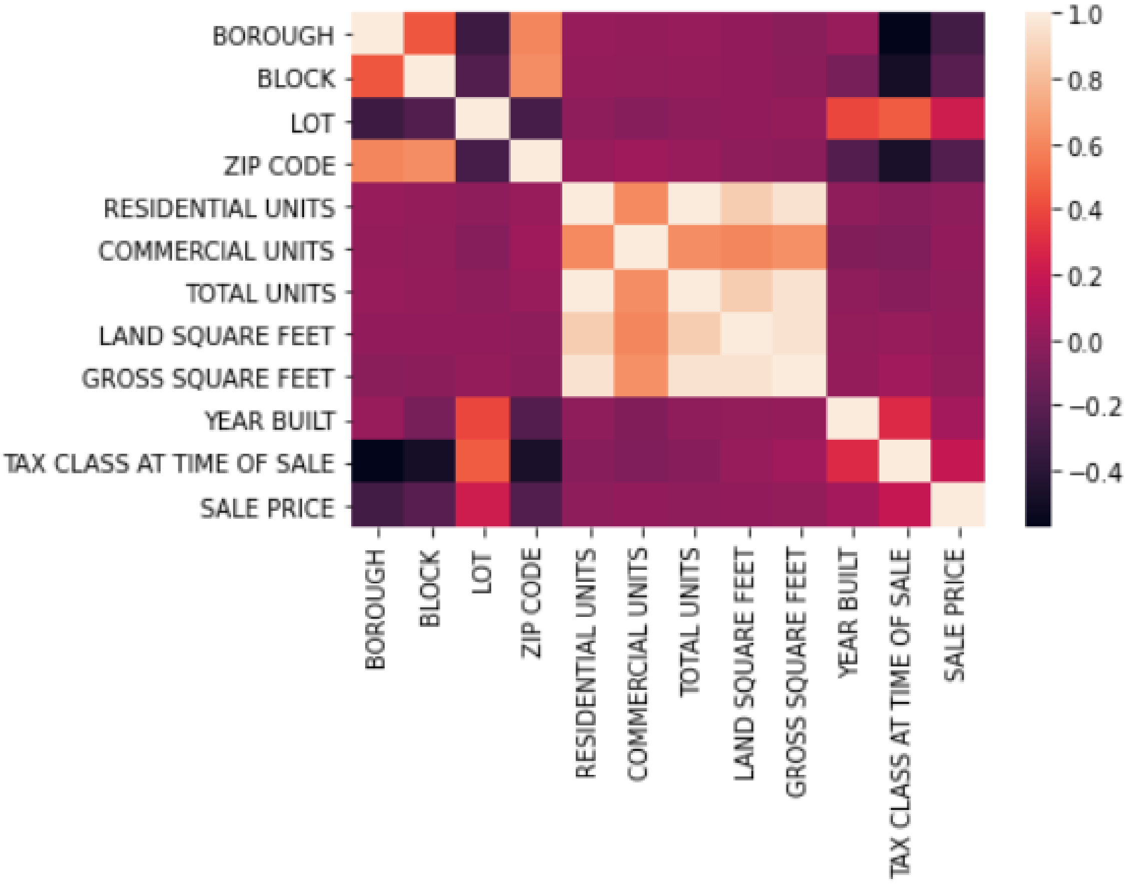


Figure 3.5: Heatmap of Numerical Variables

After cleaning up the data the SALE PRICE when plotted still has some outliers (see fig.3.5).

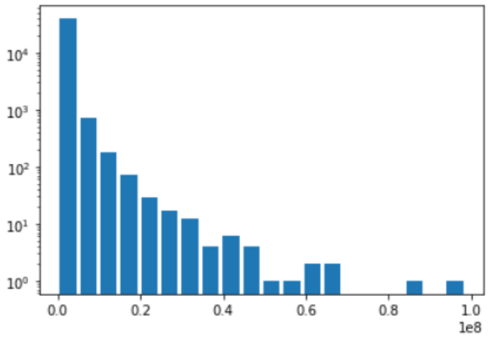


Figure 3.6: Taking log of SALE PRICE

It can also be noted that the SALE PRICE feature is skewed with respect to the numbers. While trying to fit a symmetrical model to the dataset, skewness is removed and the same is done for all numerical features while further exploring. While further exploring the data, the TOTAL UNITS has many unique values, so it's difficult to come to conclusion of total units sold or the price of the total units sold for the time period so only main values are kept to further simplify the model. The BOROUGH feature when considered it can be seen that Manhattan is expensive in 2016 & 2017 in BOROUGH, however sale goes down a notch in 2017 whereas Bronx and Brooklyn show increment in the sale prices whereas the lowest prices is for the units in Staten Island (see fig 3.6) which is not as big commercial city compared to others in BOROUGH.

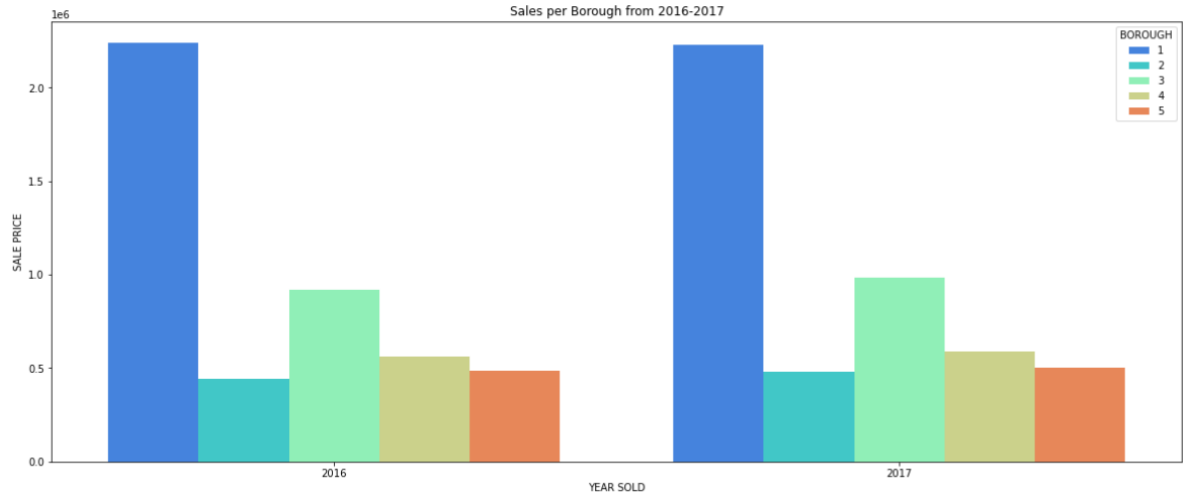


Figure 3.7: Sale Price vs Borough from 2016-2017

Finally exploring the sale date for units sold if we see the trend in the sales for the years 2016 and 2017 we can see that the maximum property sale was from January to August in 2017 whereas it was from September to December in 2016 (see fig 3.6). The same trend can be seen if we separately plot the graph for each Borough as per the property (see fig 3.7)

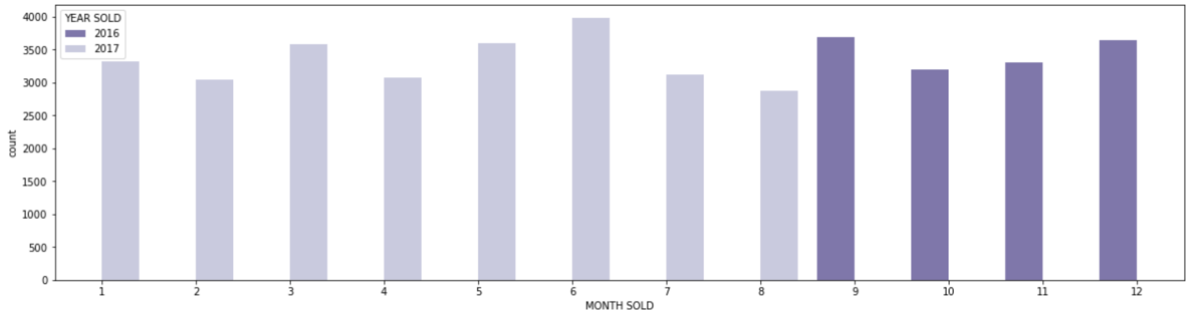


Figure 3.8: Property sales in 2016 and 2017

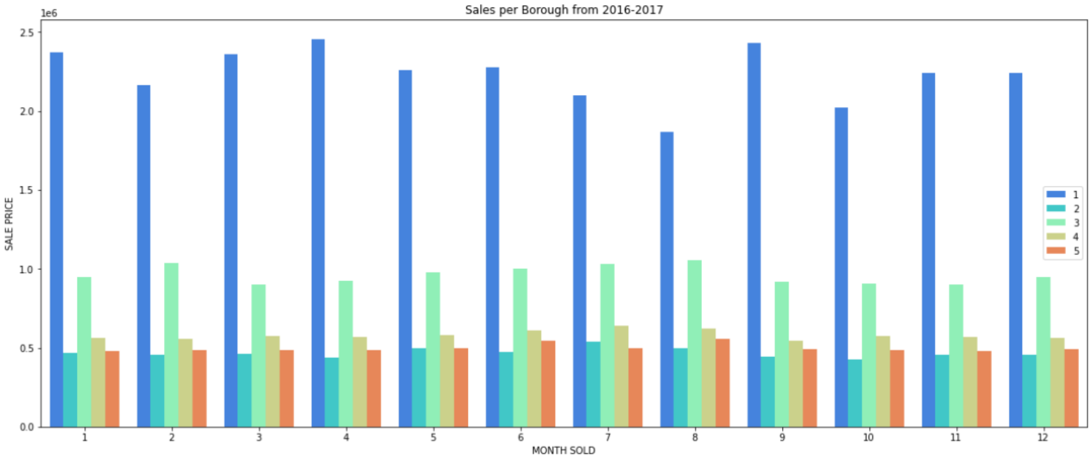


Figure 3.9: Sale per Borough 2016-2017

Now if we consider the feature NEIGHBORHOOD in the dataset, we can understand that Upper West Side was the way too expensive neighbourhood as compared to other (see fig 3.8).

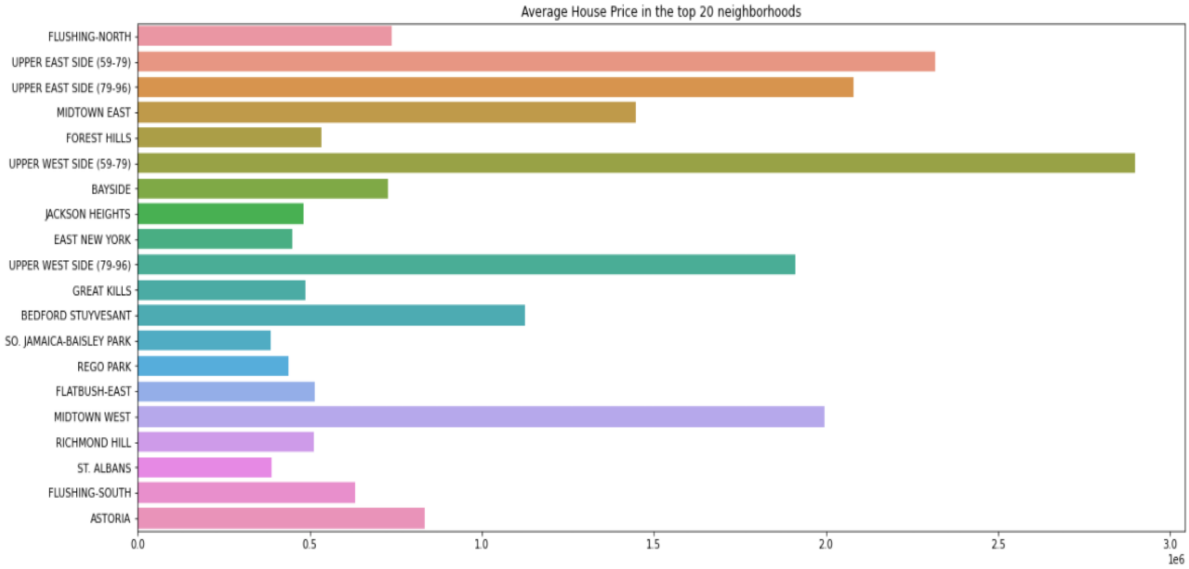


Figure 3.10: Average House Price in the top 20 neighbourhoods

Chapter 4

Unsupervised Approach

Clustering

Cluster analysis is one of the common unsupervised machine learning algorithms for grouping unlabelled datasets. Rather than making predictions as in supervised learning here in clustering we categorize the data. It basically forms sub-groups or clusters of observations of the dataset, same group datapoints are similar while different group data points are not similar. Our objective is to group the dataset to give us similar groupings for properties as per the Sale Price. This dataset is not especially suitable for unsupervised dataset as the data is in a continuous scale. This is further examined by using K-Means clustering.

4.1 K- Means

In K- Means clustering algorithm the number of clusters to split the dataset is pre-defined, we split the data in set of k-groups. To begin we prepare the data for clustering, rows are treated as observations while columns as variables. The data is then standardized using scaling technique so that the variables are comparable and relevant columns are then selected for clustering.

In K-Means, k stands for the number of clusters while Means stands for the centroid of the cluster/group where every cluster has a centroid. To define the optimal number of clusters for the dataset we are using the elbow method which tries to fit the model in a range of values for K. For this range of values for K (in our case 1-11), the elbow method runs k-means on the dataset and an average is computed for every value of K. By default, it calculates the sum of square distance from each point to its pre-defined center, basically within cluster sum of square(wss).

$$\text{minimize}\left({}_{k}\sum_{k=1} W(C_k)\right)$$

For this dataset, the number of clusters defined by the elbow method is 6 as it is the point of variation on the curve (see fig 4.1)

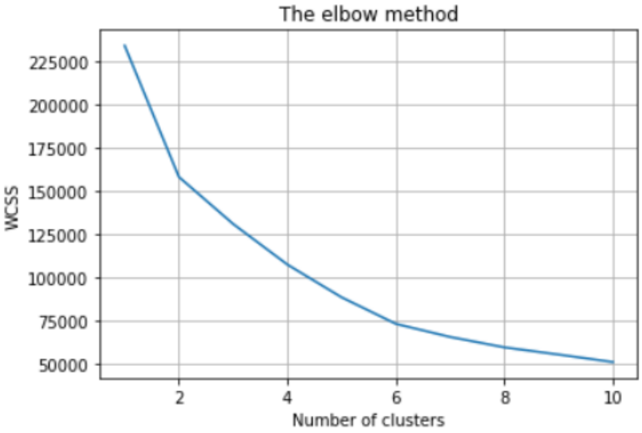


Figure 4.1: Elbow plot for NYC Property Sales

K-Means is then applied with number of clusters defined as 6 and centroid of each cluster is also calculated and then a plot is generated to examine. The Figure 4.2 clearly represents that the dataset is not suitable for k-means clustering learning as the clusters formed are too distorted and the result obtained are poor.

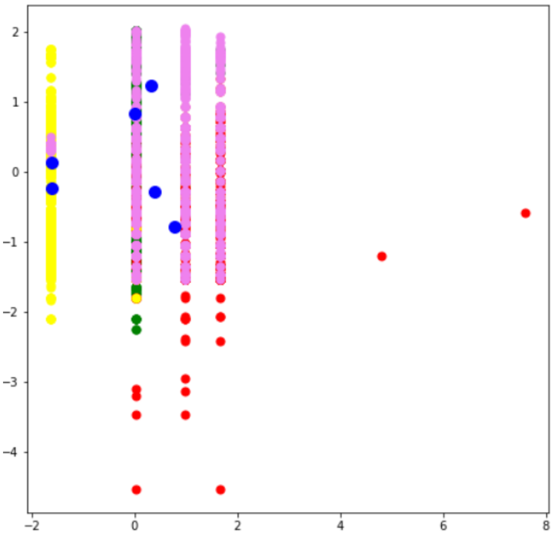


Figure 4.2: K- Means Clustering

Furthermore, the Silhouette Score is also calculated which the measure of how similar an observation is to other observations within the same cluster compared to items in other clusters. It has a range from -1 to 1, 1 representing complete clustering. In our case the score obtained is 0.3327177299790694 which is not at all satisfactory.

Chapter 5
Supervised Approach

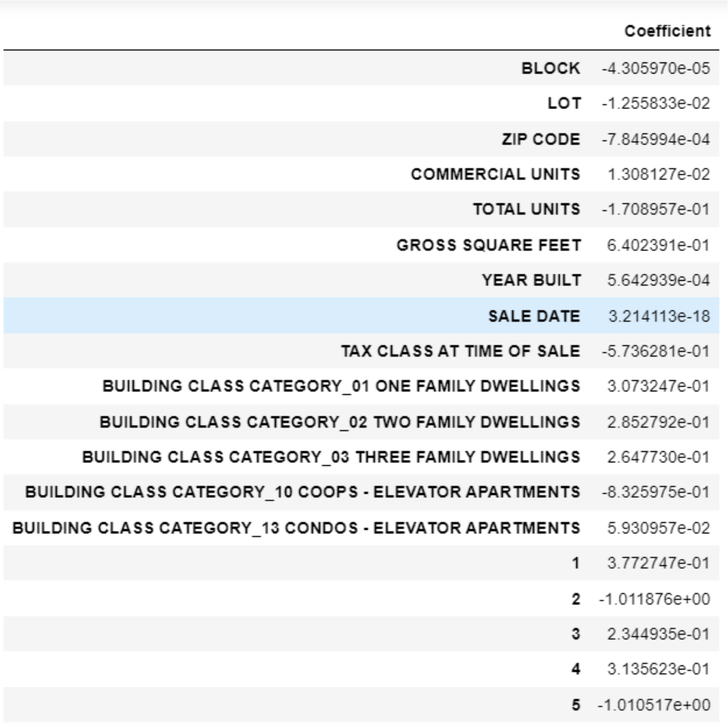
Training a system to do classification of the data by giving it examples of a relatively similar data having the right classification is mainly defined as a supervised approach. The objective of doing this analysis is to find which parameters are influencing the Sale price of New York city market and how. Regression methods mainly involves a series of coefficients which can be further analysed. Also, supervised regression model is the most fit model in our case. The amount of data reserved for regression is 30%

Many factors affect the Sale Price, in order to get further insight into how this rate changes, a regression analysis using linear regression was performed. Though to begin with the data needs to be manipulated to the right form to be used for modelling. To reduce the complexity, we will delete some features that are redundant. BUIDING CLASS AT PRESENT and BUILDING CLASS AT TIME OF SALE are redundant along with BUILDING CLASS CATEGORY; thus, we will drop the initial one. All necessary geological information is available in the ZIP CODE, so we will drop all other related features. We will choose only one of TAX CLASS AT PRESENT and TAX CLASS AT TIME OF SALE. The skewness of the numerical data is removed and all the categorical features are encoded using one hot encoding technique using the below code:

```
one_hot_encoding = ['BUILDING CLASS CATEGORY','TAX CLASS AT TIME OF SALE']

dummies = pd.get_dummies(data[one_hot_encoding])

dummies = pd.concat([dummies, pd.get_dummies(data["BOROUGH"])], axis=1) #BOROUGH are integers, doing seperately

dummies.info(verbose=True, memory_usage=True)
```

Linear regression has been selected here rather than logistic regression as it is somewhat simpler mathematically and can be used for a continuous scale. A linear coefficient is defined for every attribute, multiplying the respective coefficients with the sum of these attributed gives the output.

The data is then split into train and test data 70% and 30% respectively to create a linear regression model using the test data. The model is evaluated by checking the coefficients. The coefficients tells that holding all other features fixed a 1 unit increase in Gross square feet is associated with an increase of price by 6.402391e-01 (see fig 5.1) and so on for all the variables.



	Coefficient
BLOCK	-4.305970e-05
LOT	-1.255833e-02
ZIP CODE	-7.845994e-04
COMMERCIAL UNITS	1.308127e-02
TOTAL UNITS	-1.708957e-01
GROSS SQUARE FEET	6.402391e-01
YEAR BUILT	5.642939e-04
SALE DATE	3.214113e-18
TAX CLASS AT TIME OF SALE	-5.736281e-01
BUILDING CLASS CATEGORY_01 ONE FAMILY DWELLINGS	3.073247e-01
BUILDING CLASS CATEGORY_02 TWO FAMILY DWELLINGS	2.852792e-01
BUILDING CLASS CATEGORY_03 THREE FAMILY DWELLINGS	2.647730e-01
BUILDING CLASS CATEGORY_10 COOPS - ELEVATOR APARTMENTS	-8.325975e-01
BUILDING CLASS CATEGORY_13 CONDOS - ELEVATOR APARTMENTS	5.930957e-02
1	3.772747e-01
2	-1.011876e+00
3	2.344935e-01
4	3.135623e-01
5	-1.010517e+00

Figure 5.1: Linear Coefficient for each attribute

Predictions for linear regression are on a continuous scale so a different metric is required to do without manipulation of the output. The three commonly used evaluation metrics are:

Mean Absolute Error - The average percentage of the incorrect prediction.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error - The total sum of the squares of the average percentage of the incorrect prediction.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE) - The square root of the mean of the squared errors

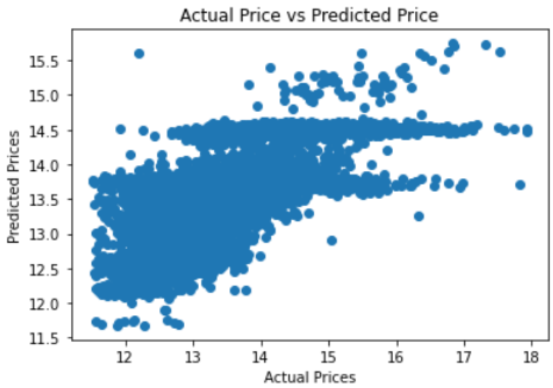
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The results of the linear regression fit are shown in Table 5.1. All the above functions are loss functions and we want them to be minimal.

Table 5.1: Performance of Linear Regression Model

	Linear Regression	Cross Verified
Mean Absolute Error	0.408987437783886	0.4092366572790706
Mean Squared Error	0.5622895086412408	0.3167949852524592
Root Mean Squared Error	0.5622895086412408	0.5628454363788155

We can also create plots to verify our predictions and see how well they perform.



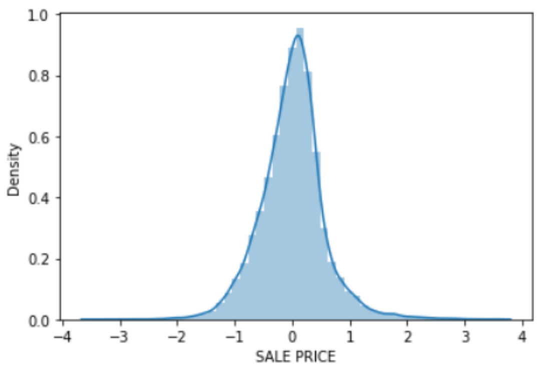


Figure 5.2: Actual Price vs Predicted

Figure 5.3: Distribution of Sale price

The scatter plot shows that the data is almost trying to form a line so our predication is fairly good and also the histogram plot shows the data in bell shape (Normally Distributed) further suggests the model has given good predictions.

Chapter 5

Reflections on the chosen analysis

In case restarting the assignment, I would have chosen a better dataset where the both supervised and unsupervised techniques would have been applicable and both methods give a fairly good result as the results obtained by the performed analysis is still open to various interpretations. As we can see that the k-means clustering analysis gives a very bad result as the dataset itself is not suited for unsupervised learning technique. We have tried to perform clustering by trying to group relevant information like location, age and availability of the apartment/ building to give us an estimate or conclusion on the Sale Price though the clusters obtained are not apt for the model.

If we consider the regression model, the linear regression model developed is quite fit for the Sale Price prediction though other regression methods still needs to be applied and the dataset further could have been segmented to improve the quality of each regression. A simple decision tree method could have been used for further segmentation and few more features like APARTMENT NUMBER could have been explored or a separate analysis for luxury properties, commercial and residential, etc could have been done.

To conclude, this report could only represent how chosen data could be analysed. We could also say that the unsupervised technique is not at all suitable for the model as the connections were very weak and output is vague. Supervised learning definitely fit the dataset better, though the output can be debatable. Overall, no special trend or highlights were found.

Appendix A

Software versions, data and included packages

Python version: Python 6.3.0 IDE: Jupiter Notebook

Dataset derived from:

https://www.kaggle.com/new-york-city/nyc-property-sales

Packages used:

- Pandas
- matplotlib.pyplot
- seaborn
- sklearn
- cluster from sklearn
- metrics from sklearn
- silhouette_score from sklearn.metrics
- scale from sklearn.preprocessing
- LinearRegression from sklearn.linear_model

Bibliography

[1]Housing Price Prediction via Improved Machine Learning Techniques – url: https://doi.org/10.1016/j.procs.2020.06.111 (visited on 09/11/2021)

[2]Scikit Learn (2018). Clustering. url: http://scikit-learn.org/stable/modules/ clustering.html (visited on 15/11/2021)

[4]One-hot-encoding–url:**Error! Hyperlink reference not valid.** (visited on 17/11/2021)

[3]Elbow Method- url:https://www.scikit-yb.org/en/latest/api/cluster/elbow.html(visited on 18/11/2021)