New York House Price | Juan Manuel Gonzalez Rincon

The mobiliary market is one of the most profitable business now days, according to A&P firm, the utility of it is about 3%-12% depending on factors as type of being and the place where it is located, in the case of US, New York, Los Angeles and Miami are the preferred cities for investment due to its diverse economy, development and growth potential influenced by the valorization of the properties.

The analysis of data related to house prices becomes important due to the importance of this market, that's why we as a mobiliary company analyzed the prices of properties in New York, the dataset we have is composed by six numerical variables regarding a specific property as: **Price:** Value in USD of the property, **Beds:** Number of beds in the property, **Bath:** Number of baths in the property, **Property sqft:** Squared footage of the property, **Latitude:** Latitude coordinate of the property and **Longitude:** Longitude coordinate of the property.

First, we start by the exploratory analysis of each variable by taking the mean, standard deviation and quartile ranges, we realize that the data has some outliers as houses with maximum number of beds of 50 or baths and the same happened with all the numeric variables, that's why we delete the outliers of data greater than 1.5 the interquartile range and lower than 1.5 the interquartile range as the limit of the boxplot whisker, getting the following processing in the case of the variable number of baths:

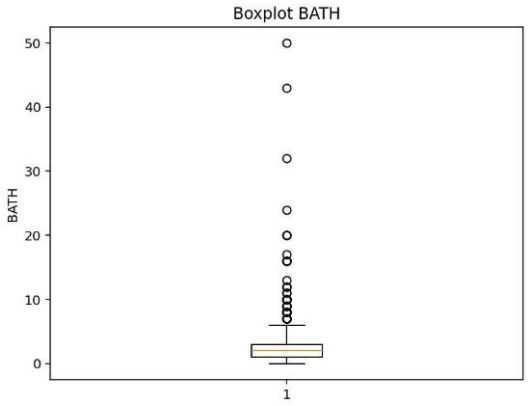

Chart1. Boxplot with outliers.

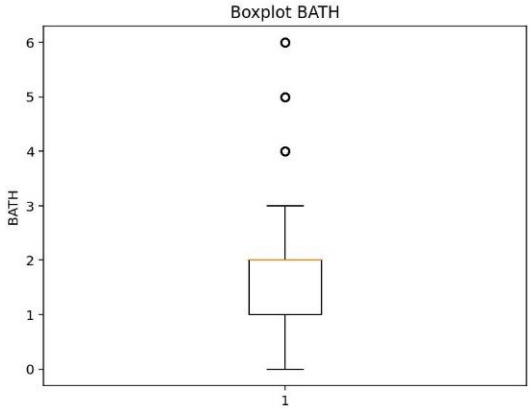

Chart2. Boxplot without outliers.

After the processing, finally we got the key values of the main variables as the average number of real baths (1.98) and beds (1.44) or the average area of a house (1,695 ft2) in New York. The average price of a house is USD$ 845,329 and the distribution of prices had a Skewness of 1.32 and a Kurtosis of 1.53 that means there is a more concentration of prices bellows the mean price but very close to it as we can see in the **Chart3.**

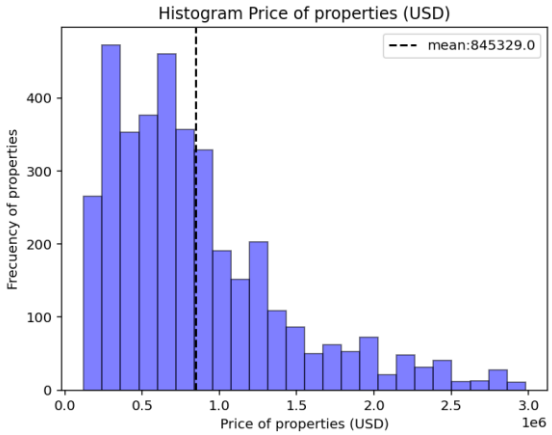

Chart3. Histogram of prices of houses.

As the key value to interpret is the price, we created a scatterplot to see the relationship with the area of the house which can have implicit the number of beds and baths, we got a slight linear relationship that spread as the value of the house increases, for the purpose of this plot we had to normalize the variables due to the difference of ranges of the variables.

Chart4. Scatterplot of Price vs Area

To verify the relationship of the independent variables: number of baths, number of beds and area of the house and the dependent variable: price, we created a correlation heatmap also based on normalized data.

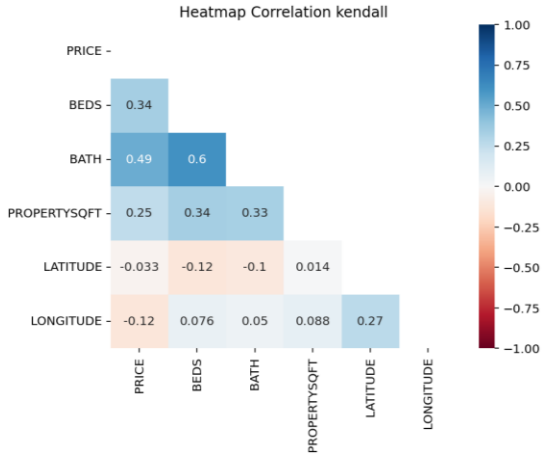

Chart5. Correlation Heatmap.

According to the EDA done, we will create some strategies to boost our mobiliary business, the first project is to divide in cells the agents according to the zone of the houses in sale, for this purpose we use the K-means algorithm based on latitude and longitude measures of the houses, we found that the appropriated number of groups is 5 where the cluster silhouette score is maximized, following this segmentation we got the groups of **Chart6** which are very similar to the boroughs defined in the city

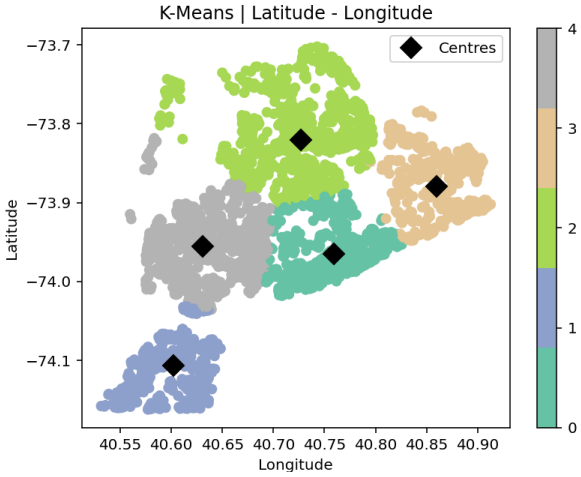

Chart6. K-means scatter plot based on location of houses.

Map1. Official boroughs of New York.

The second strategy is to fit a linear function based on one independent variable to predict the prices of new houses that are going to be sell based on the current data, for that purpose we created a linear approximation of the relation between area of the house and price of the house, the optimized result was:

$$Price\ of\ houses = 292.47 * Area\ of\ house + 349{,}454.38$$

Equation1. Result of linear fitting

The linear model lost accuracy once the value of the house increases, as a first approximation the model will work but it could improve once we add more variables to the linear fitting.

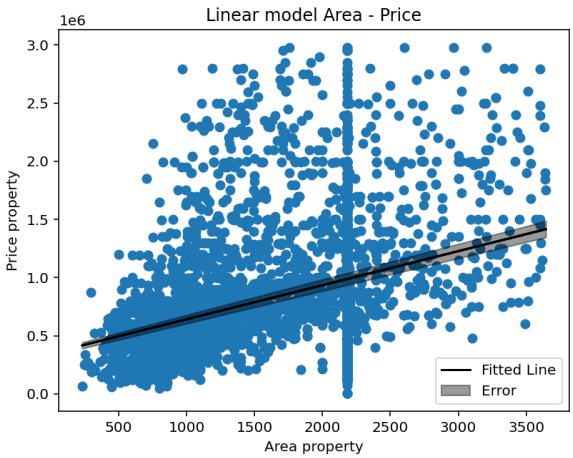

Chart7. Linear fitting of price with area

My Repository: https://github.com/Juan980207/A_Data_Science.git
My Code: **https://github.com/Juan980207/A_Data_Science/blob/main/Clustering.ipynb**
Data: **https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market/data**
Boroughs in New York: https://en.wikipedia.org/wiki/Boroughs_of_New_York_City#/media/File:5_Boroughs_Labels_New_York_City_Map.svg