HOUSE PRICE DATASET

1. Introduction:

In this project, we will analyze the House Price Dataset with the objective of generate important findings for the real estate market, validate the relationship between its variables and create predictions regarding the most important ones.

2. Summary:

The process conducted in this dataset was:

2.1 Data cleaning:

2.1.1 Validated the number of null values throughout the entire dataset.

2.1.2 Verified duplicated values.

2.1.3 Replaced NA values for the columns: Alley, MasVnrType, MasVnrArea, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PoolQC, Fence, MiscFeature.

2.1.4 Calculated the median to replace missing values for the variable LotFrontage (graph 1 – Graphs file).

2.1.5 Calculated the mode to replace missing values for the variables GarageYrBlt and Electrical.

2.1.6 Detected outliers in numerical variables (graphs 2, 3, 4 – Graphs file).

2.2 Data processing:

2.2.1 Plotted the numerical variables to observe the frequency of each value in the dataset.

2.2.2 Calculated the correlation coefficient to verify the numerical variables with the highest linear correlation.

2.2.3 Analyzed the top 4 correlations above 60% between the Sales Price variable and the other respective database variables.

2.2.4 Plotted the categorical variables to observe the frequency of each level in the dataset.

2.2.5 Selected the most relevant categorical variables according to their frequency.

2.2.6 Created boxplot graphs to observe the SalePrice variable behavior vs the most relevant categorical variables selected.

3. Results:

3.1 During the data cleaning process, we could find out that the NA values in the columns mentioned in the point 2.1.3 did not correspond to missing data, the NAs were part of the variables, reason why we modified the NA values by assigning a proper description.

3.2 Most relevant findings of numerical variables frequency plots:

3.2.1 The most frequent lot size in the real estate market is 9.478 square feet, above around 11.000 square feet is not common to find a lot with that size (graph 5 – Graphs file).

3.2.2 It is more common to find buildings with quality between average and good (5 and 7, graph 6 – Graphs file)

3.2.3 Most of the buildings were built before the year 2000 (graph 7 – Graphs file).

3.2.4 The most common size of the basement is 1.000 square feet (graph 8 – Graphs file).

3.2.5 The most frequent living area is between 1400 and 1600 square feet (graph 9 – Graphs file).

3.2.6 Most of the houses have in average 3 bedrooms (graph 10 – Graphs file).

3.2.7 The most popular garage area distribution is around 500 square feet (graph 11 – Graphs file).

3.2.8 The average price of the buildings is between $100.000 and $200.000 (graph 12 – Graphs file).

3.2.9 According to the heatmap (graph 13 – Graphs file), we could say that the next numerical variables are strongly related between them, which means that one depends on the other one:

- GarageArea and GarageCars
- TotRmsAbvGrd and GrLivArea
- 1stFlrSF and TotalBsmtSF

3.2.10 To provide relevant insights, we analyzed the top 4 correlations above 60% between the Sales Price variable and the other database variables:

- SalePrice and OverallQual: These variables are strongly related, we could conclude that when one building does not have any quality score we cannot estimate the price. On the other hand, when the overall quality score increases by one point, the sales price will increase by $45.435. (graph 14 – Graphs file)
- SalePrice and GrLivArea: These variables are strongly related, we could conclude that the average price of each square feet is $18.569. Additionally, when the ground living area increases one square feet, the sales price will increase by $107.13. (graph 15 – Graphs file)
- SalePrice and GarageCars: These variables are related, we could conclude that the average price of each building with 0 garage cars is $60.618. Additionally, when one house increases 0.5 garage cars, the sales price will increase by $68.077. (graph 16 – Graphs file)
- SalePrice and TotalBsmtSF: These 2 variables are related, we could conclude that the average price of each building with 0 basements is $63.430. On the other hand, when the basement area increases 1 square feet, the sales price will increase by $111.10. (graph 17 – Graphs file).

3.2.11 Through the bat plots we determined that the next categorical variables are going to be analyzed because of the frequency of their level, the other ones are pretty plane:

- HouseStyle: The most common houses are the ones that have 1 story. (graph 18 – Graphs file).

- ExterQual: In most of the buildings, the external quality is the average/typical (graph 19 – Graphs file).
- BsmtQual: The distribution of the basement quality is between good and average (graph 20 – Graphs file).
- KitchenQual: The distribution of the kitchen quality is between good and average (graph 21 – Graphs file).
- GarageType: The most common garage types are the attached and the detached (graph 22 – Graphs file).
- Neighborhood: The 3 most popular neighborhood are: Northwest Ames (Names), College Creek (CollgCr), and Edwards (Edwards). (graph 23 – Graphs file).

3.2.12 Observing the SalesPrice variable behavior vs the most relevant categorical variables selected we could conclude:

- SalesPrice vs HouseStyle: The 2 story houses are the one with highest median even over the 2.5 story houses, it has more outliers as well which could mean that the 2 story houses are overrated. (graph 24 – Graphs file).
- SalesPrice vs ExternalQual: The building prices with excellent external quality have a median almost double of the ones that have good external quality, furthermore, their range is superior to the other variable levels, this same behavior happens with SalePrice vs BasementQual and SalePrice vs KitchenQual. (graphs 25, 26 and 27 – Graphs file).
- SalePrice vs GarageType: The BuiltIn garage increases the SalePrice of the buildings with the median around $250.000. (graph 28 – Graphs file).
- SalePrice vs Neighborhood: The neighborhood with the more expensive houses are Northridge Heights, Northridge and Northridge Heights. (graph 29 – Graphs file).

4. Conclusions:

4.1 It is important to try to know the variables of one dataset in order to identify the meaning of each level and avoid losing important information.

4.2 The highest relationships are between the SalePrice variable and Overall Quality variable, SalePrice and GrLivArea, and SalePrice and GarageCars, we could think that the SalePrice variable would be highly related to the Bedrooms or the FullBath variables but, as we saw, those variables have a low positive linear correlation which means.

4.3 The most common houses are the ones with just 1 story but the most expensive are the ones with 2 story, even their prices are above those buildings with more than 2 story.

4.4 The external, basement and kitchen quality are determinants for the price of one house. The prices of the houses with excellent external, basement, and kitchen quality have the median above the other variable levels.