UNIVERSITY OF PENNSYLVANIA

SCHOOL OF SOCIAL POLICY AND PRACTICE

MASTER OF SCIENCE IN SOCIAL POLICY AND DATA ANALYTICS

CASE STUDY REPORT ASSIGNMENT 3

REAL ESTATE MARKET: EXPLORING THE KEY FACTORS THAT SHAPE THE PRICE PREDICTIONS

Authors:

Rose Barragan Barranco

Qiyao Huang

Yuqin Li

Jingyi Zhang

Youlin Jiang

Lechen Zhu

Professor: Richard Hartwell

Course: MSSP 6070 Practical Programming For Data Science

November 12th, 2023

INDEX

1. EXECUTIVE SUMMARY

In this report, we will support a Real Estate Brokerage understanding of the current market dynamics and customer behavior by leveraging the *HousingPrices* Dataset and examining 16 variables and their relationships.

Our primary goal is to gain insights into the most pivotal factors that contribute to setting an attractive price for each property. Thus, by scrutinizing these aspects, understanding their interconnectedness, and garnering the most valuable insights, we will support the Real Estate Brokerage in their strategic decision-making process by leveraging a set of actionable recommendations, and thereby captivating the interest of the brokerage's clientele seeking residential or investment real estate acquisitions.

2. OVERVIEW OF THE VARIABLES IN THE DATASET

Before starting with the analysis, it is necessary to understand what kind of data we have in our Dataset – "Housing Prices" – collecting a sample of 1460 real estate properties and a set of **16 different numerical variables: 1 dependent variable (DV) and 15 independent variables (IV)** (Montoya, 2016).

- **Sale Price (DV)**: This is the most relevant variable in the Dataset, as it is the case study's target field or dependent variable. It represents the sale price of the real estate property in USD.

- **Yr Sold (IV)**: Year Sold.
- **YearBuilt (IV)**: Original construction date.
- **GarageArea(IV)**: Size of garage in square feet.
- **GarageCars (IV)**: Size of garage in car capacity.
- **Total Rooms (IV):** number of rooms in the house.
- **FullBath (IV)**: Full bathrooms above grade.
- **1stFlrSF (IV)**: First Floor square feet.
- **2ndFlrSF (IV)**: Second-floor square feet.
- **CentralAir (IV)**: Central air conditioning.
- **TotalBsmtSF (IV)**: Total square feet of basement area.
- **OverallCond (IV)**: Overall condition rating.

- **OverallQual (IV)**: Overall material and finish quality.
- **HouseStyle (IV)**: Style of dwelling.
- **LotArea (IV)**: Lot size in square feet.
- **LotFrontage (IV)**: Linear feet of street connected to the property.

After reviewing each variable, we will analyze the relationship between each of them, identifying which are the most relevant for our forecasting model of house prices for the years 2011 onwards.

3. ANALYSIS AND METHODS USED

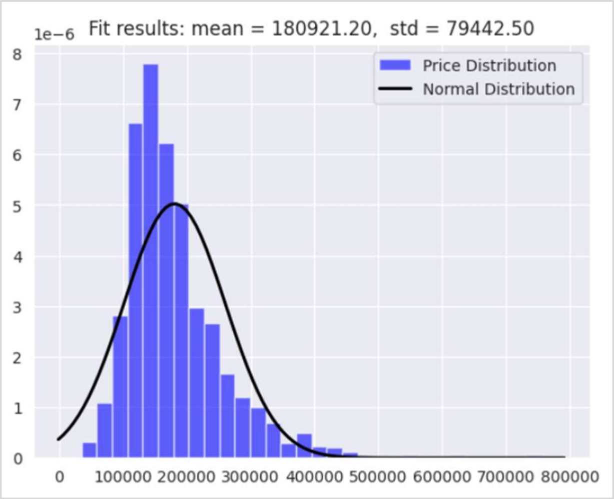
To proceed with the analysis of the Housing Prices Dataset, we will work with Google Colab tools - programming in Python – and Microsoft 365 Toolkit for a thorough analysis of the data, resulting in a comprehensive storytelling that combines a seamless narrative with visualizations that illustrate and facilitate the understanding of the analysis.

4. OVERVIEW OF THE HOUSING PRICES

4.1. Mean and Standard Deviations

The "Housing Prices" dataset includes a total of 1460 real estate properties (houses) whose price range varies between $34,900 to $755,000, with the average price equal to $180,921.20 [Figure #1].

Figure #1 – Housing price distribution and histogram



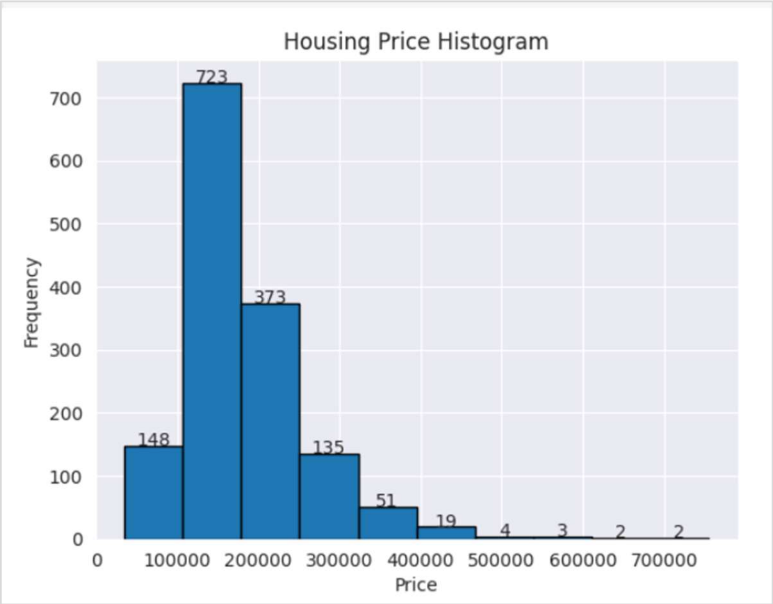
Source: group (own graph).

Looking more closely at the histogram, we observe that the majority of values are concentrated in properties priced between $100,000 and $200,000, amounting to a total of 723 properties in that range of money. If we continue analyzing, we observe that there is another leading group since we find 373 properties between $200,000 and $300,000 [Figure #2].

This may indicate that most buyers in that particular market are young and are buying their first home with their savings or banking financial support; or these are properties for investment purposes, where investors look to buy low, make savvy renovations, and sell high, obtaining a high return of investment and be a profitable income source.

If we now turn our attention to the tail of the histogram, we see that the lowest number of properties fall under the $600,000 - $800,000 range. These homes can be considered luxury, so only a few buyers can get close to a million-dollar value in properties [Figure #2].

Figure #2 – Housing price range

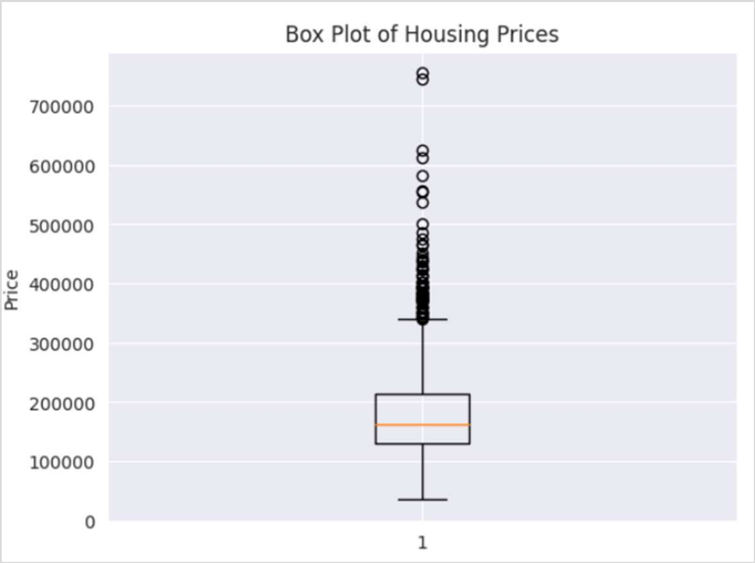


Source: group (own graph).

On the other hand, observing the standard deviation value, which measures how much house prices vary from the mean of housing prices equal to $79,442.50, suggests that house prices tend to be more dispersed around the mean.

Thus, the mean house price is $180,921.20, and the standard deviation is $79,442.50, indicating a wide range (high variability) of house prices in the Dataset, with some houses having prices significantly above or below the mean. This can be seen in Figure #3, which shows the distribution of house prices and how they are dispersed around the mean (orange line).

Figure #3 – Dispersion of the Housing prices around the mean



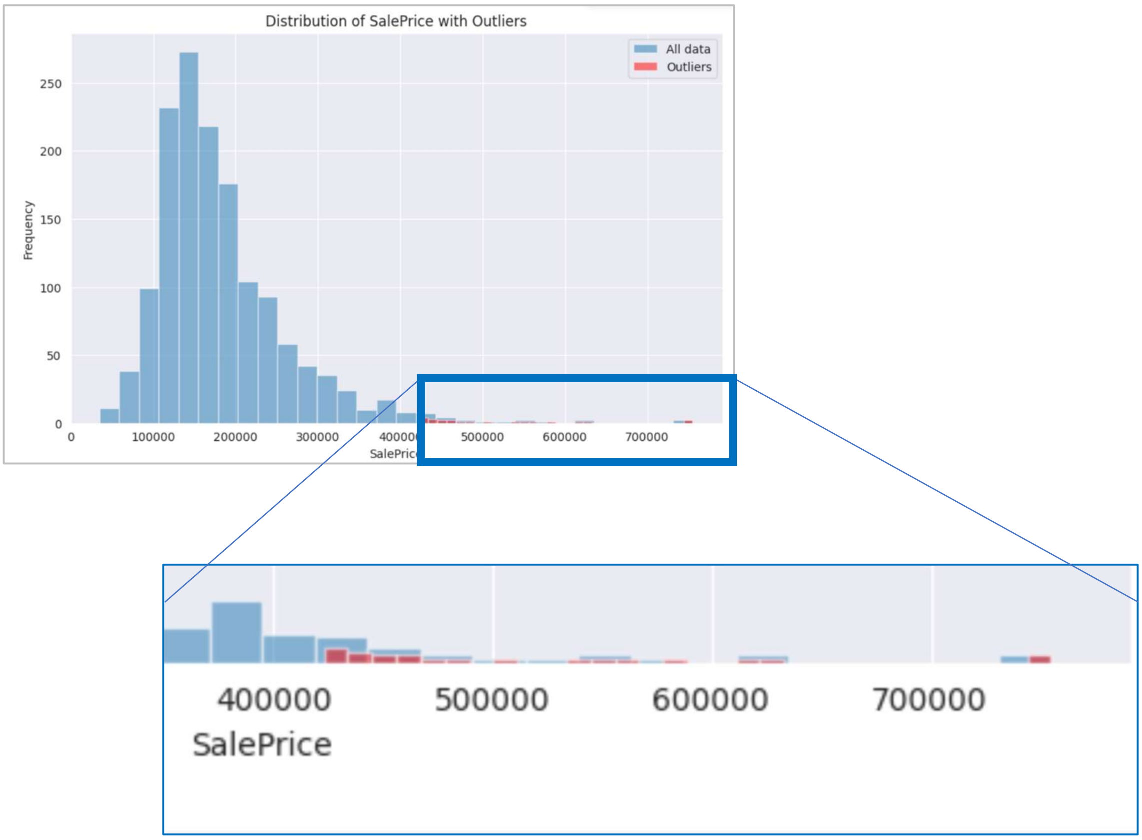
Source: group (own graph).

4.2. Outliers

Following the standard deviation rule, an outlier is identified as any sales price that exceeds more than 3 standard deviations from the mean. Given that the standard deviation of the sales price is $79,442.50 and the mean sales price is $180,921.20, our analysis revealed 22 outliers within the sales price data, starting at approximately $430,000.

These numbers are relatively infrequent and represent prices significantly higher than average. The presence of such outliers suggests that there is a considerable number of homes priced above the average, highlighting the presence of exceptionally high-value properties in the Dataset.

Figure #3 – Outliers' visualization

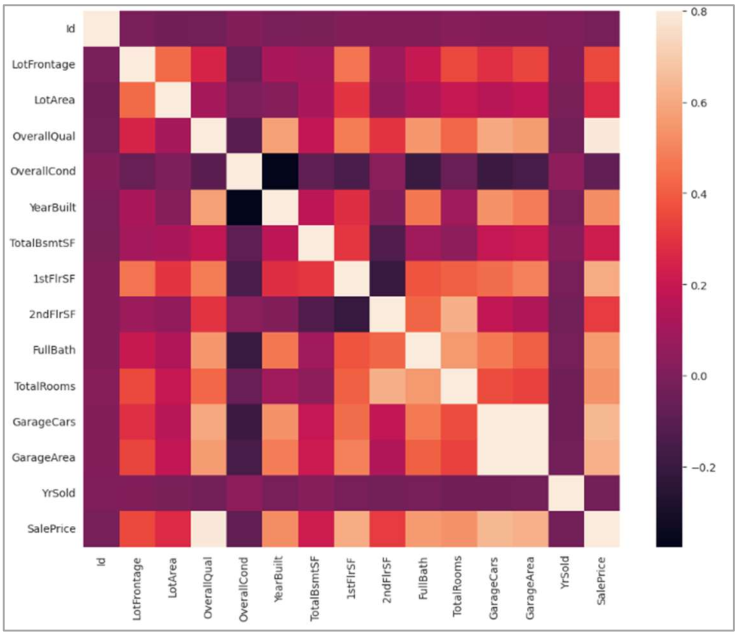


Source: group (own graph).

5. CORRELATION ANALYSIS

As we have explained at the beginning of this report, the market price will depend to a great extent on the relationship between the different variables or characteristics that each real estate property has. That is why it is necessary to analyze the correlation between the independent variables to know which characteristics are the most significant when determining the price of housing [Figure #4 and Figure #5].

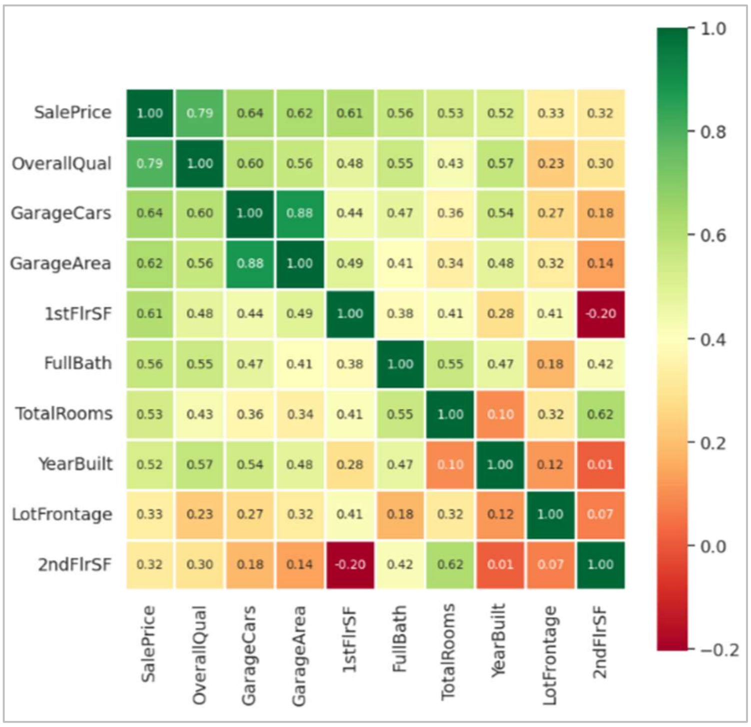
Figure #4 – Correlation Heat map (with outliers)



Source: group (own graph).

Note for the reader: The colors vary from light yellow to dark red. The lighter the shade, the stronger and positive correlation of the coefficients, while dark red shades denote weaker correlations.

Figure #5 – Correlation coefficient Matrix (with outliers)



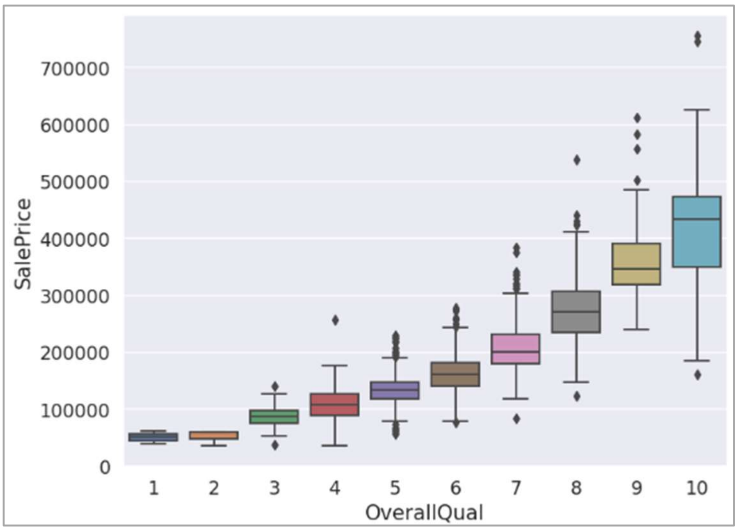
	SalePrice	OverallQual	GarageCars	GarageArea	1stFlrSF	FullBath	TotalRooms	YearBuilt	LotFrontage	2ndFlrSF
SalePrice	1.00	0.79	0.64	0.62	0.61	0.56	0.53	0.52	0.33	0.32
OverallQual	0.79	1.00	0.60	0.56	0.48	0.55	0.43	0.57	0.23	0.30
GarageCars	0.64	0.60	1.00	0.88	0.44	0.47	0.36	0.54	0.27	0.18
GarageArea	0.62	0.56	0.88	1.00	0.49	0.41	0.34	0.48	0.32	0.14
1stFlrSF	0.61	0.48	0.44	0.49	1.00	0.38	0.41	0.28	0.41	-0.20
FullBath	0.56	0.55	0.47	0.41	0.38	1.00	0.55	0.47	0.18	0.42
TotalRooms	0.53	0.43	0.36	0.34	0.41	0.55	1.00	0.10	0.32	0.62
YearBuilt	0.52	0.57	0.54	0.48	0.28	0.47	0.10	1.00	0.12	0.01
LotFrontage	0.33	0.23	0.27	0.32	0.41	0.18	0.32	0.12	1.00	0.07
2ndFlrSF	0.32	0.30	0.18	0.14	-0.20	0.42	0.62	0.01	0.07	1.00

Source: group (own graph).

After performing several correction matrixes (heat map and coefficient matrix), we observe that the factors with the highest relevance in determining the SalePrice are OverallQual, GarageCars, GarageArea, and 1stFlrSF.

- **OverallQual** represents the overall quality of the material and finishes of the house. It indicates a strong positive correlation with 'SalePrice' **(0.79)**, suggesting that the higher the quality of the houses, the higher the selling price [Figure #6].

Figure #6 – Boxplot SalePrice - OverQuall (with outliers)

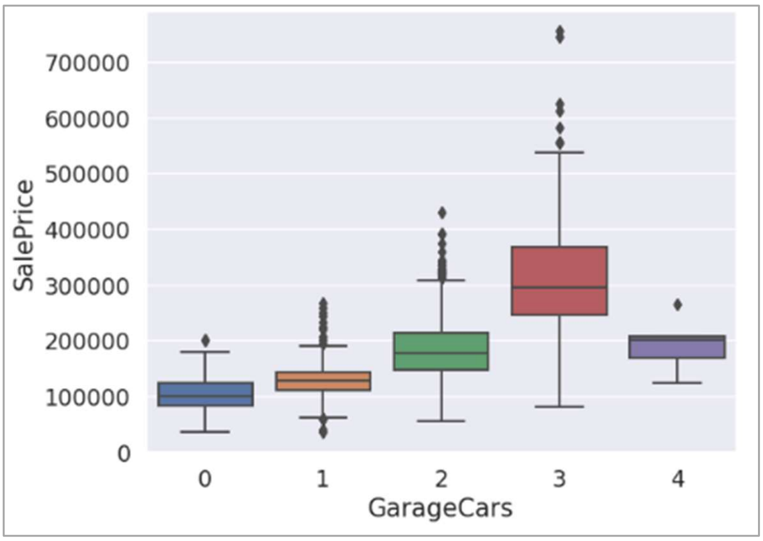


Source: group (own graph).

- **GarageCars**, which represents the garage capacity in terms of car space, also has a significant positive correlation with 'SalePrice' **(0.64).** This indicates that properties with larger garages tend to command higher sale prices, which might reflect buyer preferences for more storage space or an indicator of a more substantial overall property size [Figure #7].

- **GarageArea**, which indicates the size of garage in square feet, has a coefficient of 0.63 with 'SalePrice'. Corresponding to the previous factor 'GarageCars', the high correlation here also illustrates that houses with larger size of garages tend to become more expensive and implies that consumers may favor houses with more available garage space. However, it is also noteworthy that the two IVs, 'GarageCars' and 'GarageArea', are highly correlated with each other, with a coefficient of nearly 0.90. Therefore, when conducting OLS regression, 'GarageArea' is excluded from the model to avoid strong multicollinearity with 'GarageCars'.

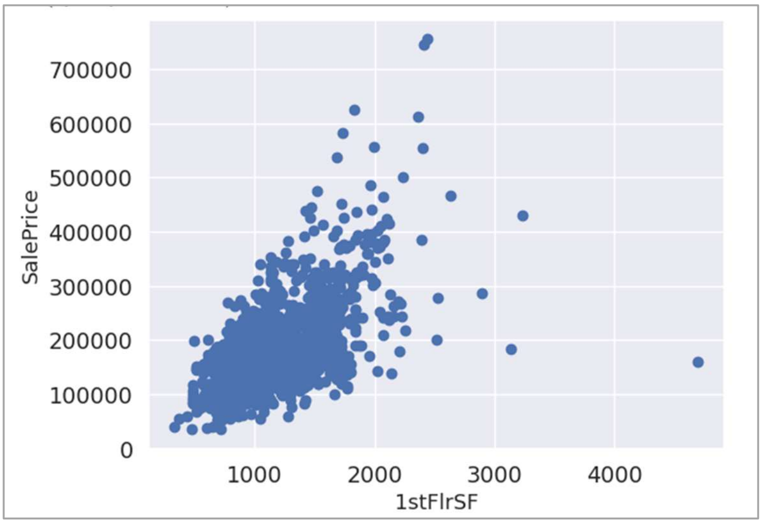
Figure #7 – Boxplot SalePrice – GarageCars (with outliers)



Source: group (own graph).

- **1stFlrSF**, representing the first-floor square footage, is positively correlated with 'SalePrice' **(0.61).** This correlation suggests that houses with more living space on the first floor generally sell at a higher price, likely due to the demand for larger living spaces [Figure #8].

Figure #8 – Scatterplot SalePrice -1stFirSF (with outliers)



Source: group (own graph).

Observing the correlation coefficient matrix, the correlation between OverallQual and GarageCars is moderately low (0.60); the correlation between OverallQual and 1stFlrSF is also moderate (0.48); and the correlation between GarageCars and 1stFlrSF is relatively low (0.44). These values suggest that while there is some degree of correlation, it is not so high as concerning for multicollinearity in a linear regression model.

Therefore, these variables can be considered reasonably independent of each other for the purpose of analyzing their individual effects on ‘SalePrice.’ This makes them good candidates for inclusion in a multiple regression model that aims to understand the impact of several independent variables on a dependent variable without the interference of high inter-correlations.

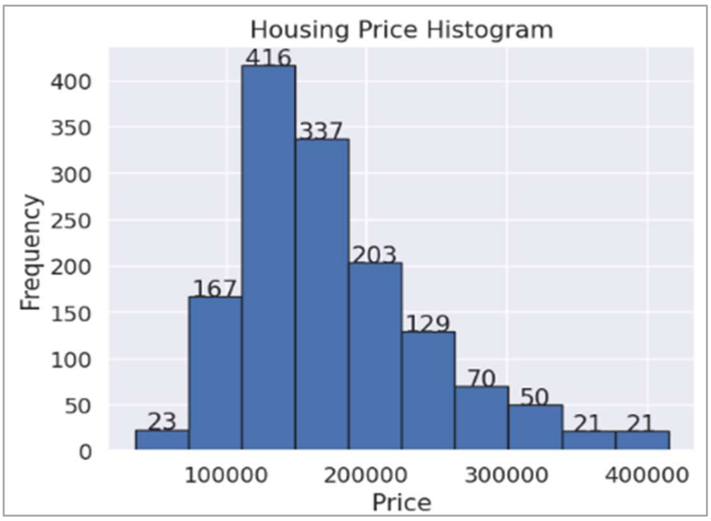
Overall, both (the heat map and the correlation coefficient matrix) support the decision to focus on OverallQual, GarageCars, and 1stFlrSF as variables for analyzing ‘SalePrice’ due to their relatively strong and positive correlations. These factors are likely to be influential in predicting the sale price of houses in the Dataset, making them valuable for regression analysis or other predictive modeling techniques.

6. DATA CLEANING – BEFORE PERFORMING LINEAR REGRESSION ANALYSIS

6.1. Remove Outliers

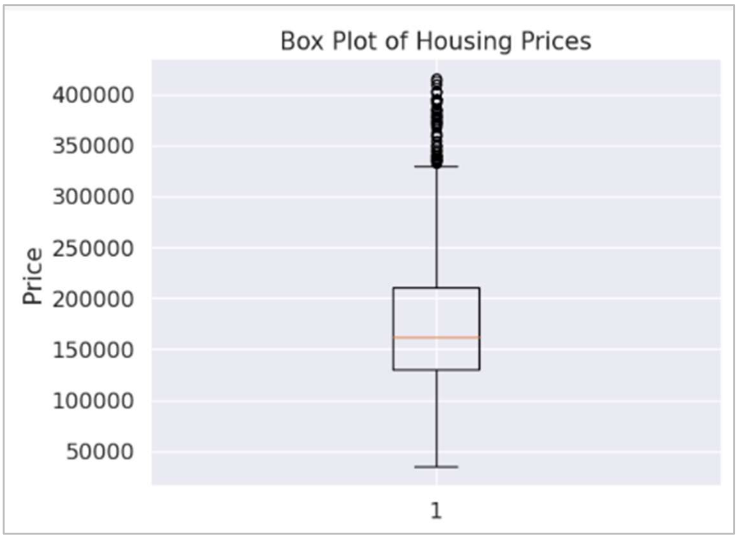
After identifying the outliers in SalesPrice in our Dataset, first, we procced to eliminate the outliers values in Sales Price [Figure #9 and Figure #10].

Figure #9 – New Histogram of SalePrice after removing the outliers values



Source: group (own graph).

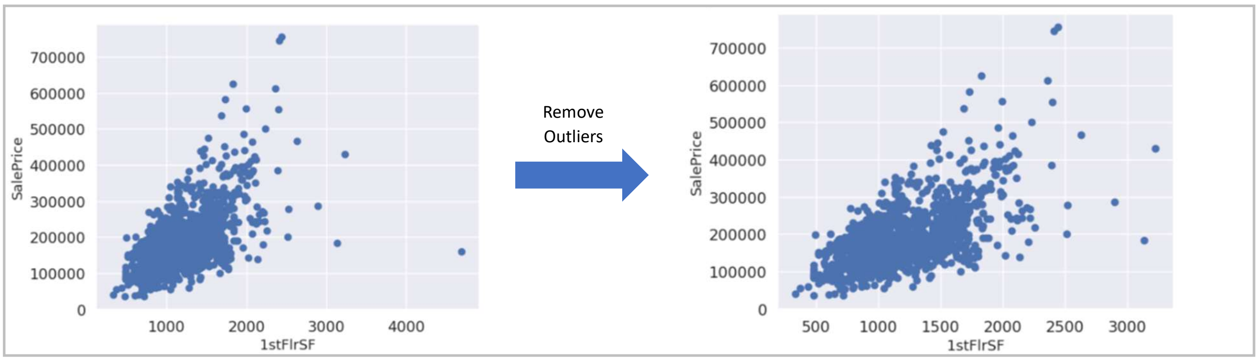
Figure #10 – New Boxplot of SalePrice after removing the outliers values



Source: group (own graph).

And second, we have also proceed to remove the outliers values from the independent variable, 1stFirSF, in order to improve the model's performance [Figure #11].

Figure #11 – Removing the outliers from 1stFirSF



Source: group (own graph).

6.2. Missing Values Assessment

Furthermore, other data-cleaning techniques must be implemented in other variables of our Dataset before performing a multi-linear regression model. Thus, we observe that LotFrontage (Linear feet of street connected to property) and TotalBsmtSF (Total square feet of basement area) have approximately 17.74% and 37.47% missing values, respectively.

Due to the strong correlation observed in previous analyses between LotFrontage and TotalBsmtSF with Sales Price (r > 0.6) [Figure #5], the missing values for these variables might significantly impact subsequent studies. Therefore, imputation was performed:

- For **TotalBsmtSF**, it was assumed that the missing values might be due to houses lacking a basement. For this continuous variable, **zero** was used to fill the missing value.

- For **LotFrontage**, considering that all houses in this Dataset are in the same neighborhood and adjacent houses on the same street often share similar street frontage attributes, the **mode** was chosen for assigning missing values to maintain data consistency and reasonability.

7. TRANSFORMATIONS – BEFORE PERFORMING LINEAR REGRESSION ANALYSIS

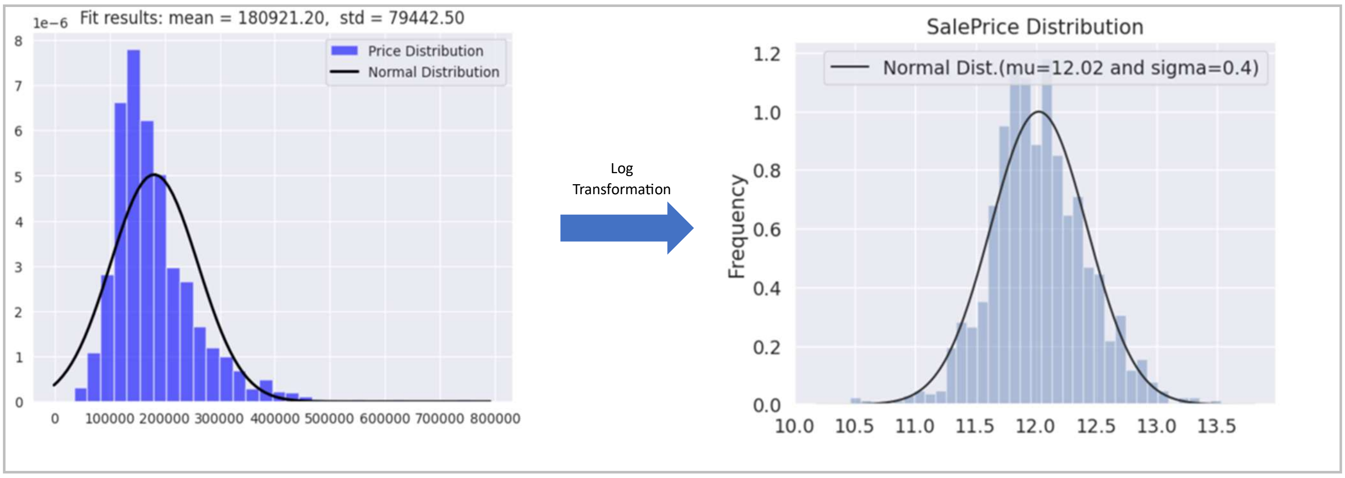
7.1. Requirements for Normality And Homoscedasticity

Before performing the linear regression analysis, further modifications are necessary to make the dependent and independent variables better fit within the model.

As seen in section "*4.1 Mean and Standard Deviation*", the distribution of 'SalePrice' is slightly skewed-right. To address this problem and meet the requirements of normality and homoscedasticity of the residuals for multiple linear regression, a logarithmic transformation is implemented to ensure the normal distribution of the dependent variable.

After transformation, the new variable 'SalePricelog' is more normally distributed and will be used as the new dependent variable in the multiple linear regression [Figure #12].

Figure #12 - Comparison between the Sale Price Distribution vs Normal Distribution



Source: group (own graph).

7.2. Requirements for centering

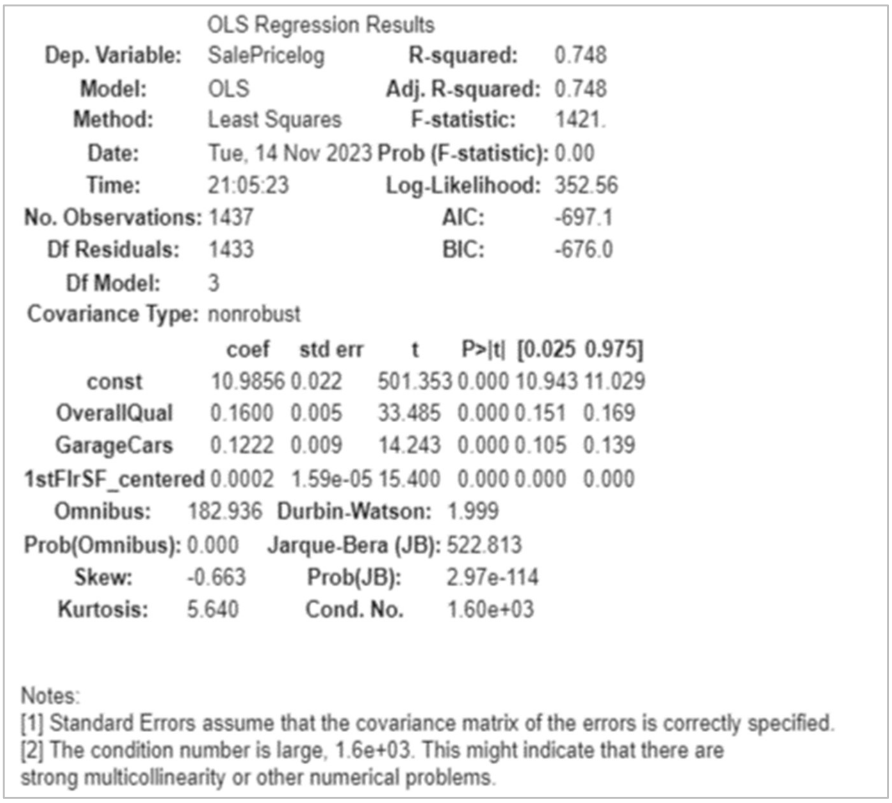
In addition, for the sake of interpretability and the stability of the model, a centering transformation is implemented on the continuous variable '1stFlrSF'.

8. LINEAR REGRESSION ANALYSIS – AFTER PERFORMING DATA CLEANING AND TRANSFORMATIONS

Therefore, after the logarithmic transformation of ‘SalePrice’ and the centering transformation of ‘1stFlrSF’, the final multiple regression model is as follows [Figure #13]:

SalePricelog = β0 + β1*OverallQual + β2*GarageCars + β3*1stFlrSF_centered

Figure #13 - Multiple Linear Regression Analysis



```
                    OLS Regression Results
   Dep. Variable:   SalePricelog          R-squared:        0.748
          Model:    OLS               Adj. R-squared:       0.748
         Method:    Least Squares        F-statistic:       1421.
           Date:    Tue, 14 Nov 2023 Prob (F-statistic):    0.00
           Time:    21:05:23          Log-Likelihood:       352.56
No. Observations:   1437                    AIC:            -697.1
   Df Residuals:    1433                    BIC:            -676.0
       Df Model:    3
 Covariance Type:   nonrobust
                      coef   std err     t     P>|t|  [0.025  0.975]
         const      10.9856  0.022    501.353  0.000  10.943  11.029
    OverallQual      0.1600  0.005     33.485  0.000   0.151   0.169
     GarageCars      0.1222  0.009     14.243  0.000   0.105   0.139
1stFlrSF_centered   0.0002  1.59e-05  15.400  0.000   0.000   0.000
      Omnibus:     182.936  Durbin-Watson:      1.999
Prob(Omnibus):     0.000    Jarque-Bera (JB):   522.813
         Skew:     -0.663       Prob(JB):       2.97e-114
     Kurtosis:     5.640        Cond. No.       1.60e+03


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.6e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Source: group (own graph).

The performance of the model is generally good, as evidenced by the R-squared and Adjusted R-squared of 0.748, which indicates that approximately 74.8% of the variance in ‘SalePricelog’ can be explained by the three independent variables, and by the AIC of -697 that suggests a good fit of the model.

Moreover, the p-values of the three variables are all extremely low, smaller than 0.05 (the threshold used to estimate the statistical significance), which indicates that all independent variables are statistically significant in predicting ‘SalePricelog.’

Among them, the variable OverallQual has the highest coefficient of 0.16, indicating that it has the most substantial impact on 'SalePricelog' among the three variables. 'GarageCars' has the second significant positive impact on the dependent variable, with a coefficient of nearly 0.12.

This analysis concludes that although these three variables are highly correlated to the SalePrice, it must be highlighted that the overall material and finish quality of the houses and the size of the garage in car capacity tend to play a more significant role in the determinants of the housing sale prices than the first-floor square feet.

9. CONCLUSIONS AND KEY TAKEAWAYS

After a thorough analysis, we identified three main factors that significantly impact the selling price of homes: overall quality of materials and finishes (OverallQual), garage capacity (GarageCars), and second-floor square footage (1stFlrSF).

A first analysis concluded that these three factors are correlated but do not exhibit multicollinearity, making them ideal candidates for a multiple regression model. However, before performing such regression modeling, we had to remove outliers in the sales price and fill in missing values for variables like LotFrontage and TotalBsmtSF, strategically imputing values to maintain the completeness of the analysis. In addition, transformations such as logarithmic for 'SalePrice' and centering for 1stFlrSF have helped to fit the data to the assumptions of normality and homogeneity needed for linear regression.

Finally, the linear regression model confirms the relevance of the three highlighted factors. However, it is worth noting that OverallQual (OverallQual) and GarageCapacity (GarageCars) show more significant leverage on sales price, while 1st-floor space (1stFlrSF) has a more minor impact.

So, this analysis highlights the importance of material quality and garage capacity in determining the selling price of real estate properties, suggesting a strategy focused on improving these aspects to maximize property values.

10. RECOMMENDATIONS

Based on the analysis, **recommendations for the real estate brokerage** would be to focus improvement efforts on specific areas that significantly impact the selling price.

- **Focus on property quality:** Since the overall quality of materials and finishes (OverallQual) has a strong correlation to the SalePrice, it is crucial to invest in improvements to the quality of construction and finishes of the homes. This can include upgrades in materials, interior design, and finish details to increase buyers' appeal and perceived value.

  In addition, this high-quality approach to materials and finishes should be **showcased to buyers during marketing campaigns and property tours**.

- **Garage space optimization**: The garage's capacity (GarageCars) significantly influences the selling price. So, it could be considered expanding or upgrading existing garages to provide more space for vehicles, tools, or warehouse storage. This may attract buyers who value additional garage space.

In summary, prioritizing construction quality, upgrading existing garages, and optimizing living space can be effective strategies to increase the perceived and actual value of real estate properties, **attracting a more significant number of buyers or investors into the market.**

11. REFERENCES

[1] Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle. https://kaggle.com/competitions/house-prices-advanced-regression-techniques