Data Analysis and Regression Modeling of Housing Data

Debayan Biswas
Student ID: 22242821
Master of Science in Data Analytics
National College of Ireland
x22242821@student.ncirl.ie

Abstract—**This report outlines the process of data analysis and regression modeling applied on to the housing data set. The objective is to predict the target variable based on a set of features, combining exploratory data analysis, correlation analysis, and multiple linear regression modeling. Python programming language is used for implementation.**

I. Introduction

A. Background

The project outlines the process of data analysis on a housing data set provided in comma-separated value format to investigate factors influencing housing prices using regression. The data set consists of 18 features and 2413 rows. The goal of the analysis is to accurately predict the Sale_Price.

B. Objective

The primary objective of the project is to explore and analyze the factors influencing housing prices using multiple linear regression. The steps to achieve this include Exploratory Data Analysis which is used to understand the pattern, feature distribution, and relationships between different variables, examine the correlation within the data set to identify relationships between features and the target variable, and handle missing values, outliers, and categorical variables appropriately. The other steps include using a multiple linear regression model, to predict the target variable and analyze the coefficients of the features to determine their importance in influencing the target variable and reviewing the effectiveness and performance of the regression model using appropriate evaluation metrics like Mean Squared Error value and R-squared value. By achieving the above steps, the project aims to provide valuable insight into the housing data set, facilitate informed decision-making, and establish a foundation for future work in predictive modeling for housing-related outcomes.

Debayan Biswas

II. Exploratory Data Analysis

A. data set Summary

The data set provided has 18 features in total having 4 object type (categorical data), 12 integer type, and 2 float type datatype. On further investigation, it can be inferred that some numeric columns do have categorical values as well. The 17 features present are independent features and one feature that is Sale_Price is the dependent feature. The value of Sale_Price is influenced by the other 17 features.

B. Unique Values

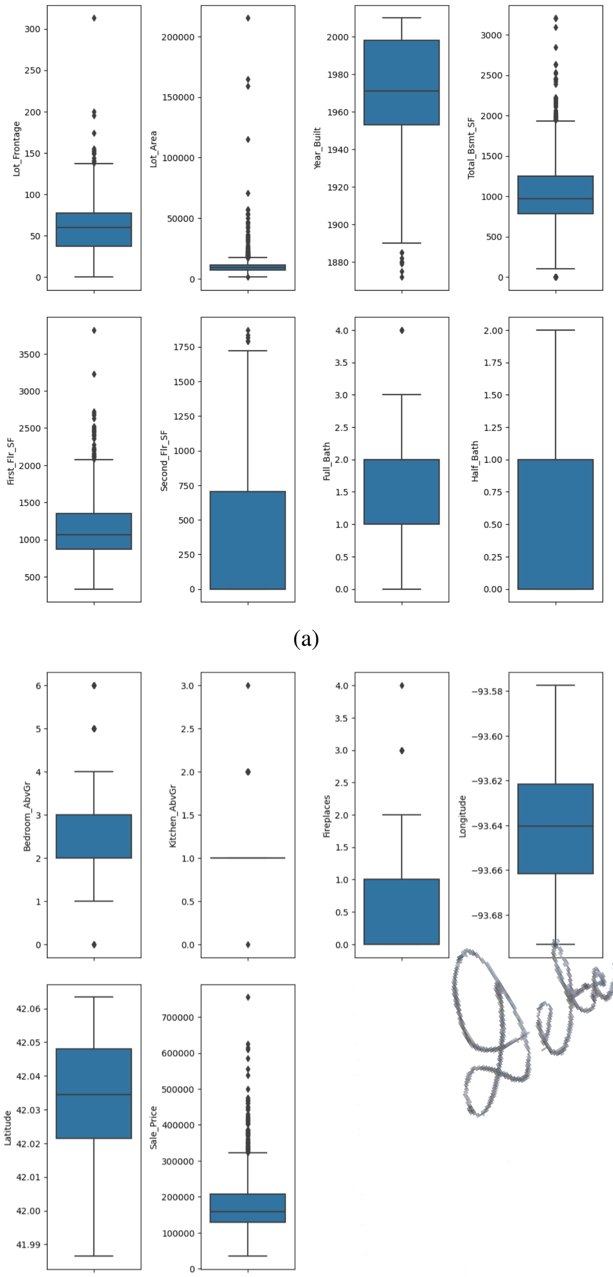
The data set has unique values present in all the columns. However, some columns have very low uniqueness values, which might indicate that some features might be categorical.

C. Missing Data

The data set does not have any missing data or Null value present. A heat map is generated to double-check for the missing values.

D. Visual Exploration

1. Box plot is applied on the data set (except for the 4 object type categorical columns) for identifying outliers and the spread of the data. Outliers are data points that deviate from the other data points present in a data set. Proper handling of outliers is a significant step for the pre-processing of data, as Outliers can have a consequential impact on the statistical analysis.
From the boxplot in Fig.1, it is evident that outliers are present in the data set and it should be removed during data pre-processing. The Outlier percentage of all the numeric data columns is observed and it can be inferred that the features such as Lot_Area, Total_Bsmt_SF, and Bedroom_AbvGr have some outliers present while Kitchen_AbvGr has 100 percent outliers. The outlier percentage will be again checked after outlier removal during the pre-processing of data. 2. Hist-plot is applied on the 4 object type categorical column with the dependent feature Sales_Price as a reference to understand how the Sale_Price is affected by the categorical features.
Hist-plot is implemented to check how the Sales_Price is influenced by the individual categories of each categorical feature which can be seen in Fig.2.
3. Bar plot is used as a visualization technique to show the distribution of all the categorical features, both object type and the rest 5 numeric categorical data. The Bar plot in Fig.3 shows the frequency or count of observations falling into different categories. The count of each object of the individual features is visible.
4. Distribution plot is used to visualize the distribution of each numeric variable individually to select relevant features



(a)

(b)

Fig. 1: Box plot showing outliers

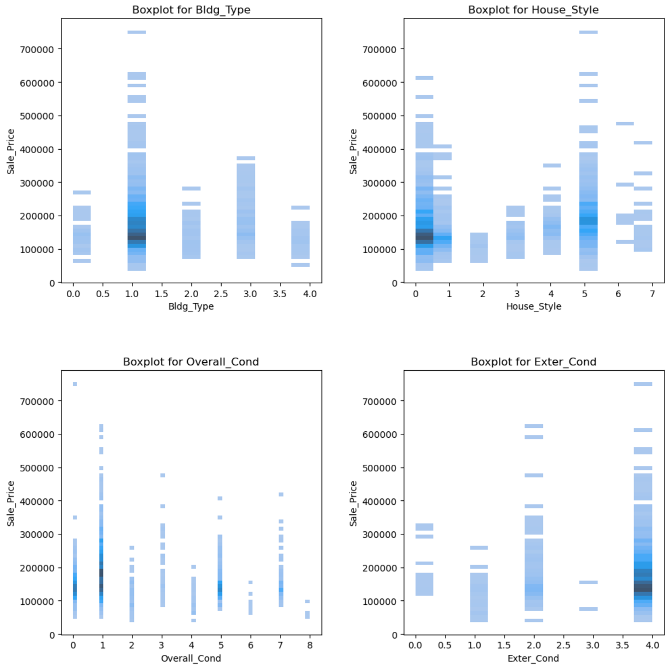


Fig. 2: Histplot of Object type Categorical features

for data analysis. This plot is used to check whether the distribution of the feature is normal or not.
From the observation of Fig.4, it is visible that some of the features consist of skewness. Skewness measures the asymmetry of the distribution of variables. The skewness needs to be removed during pre-processing.

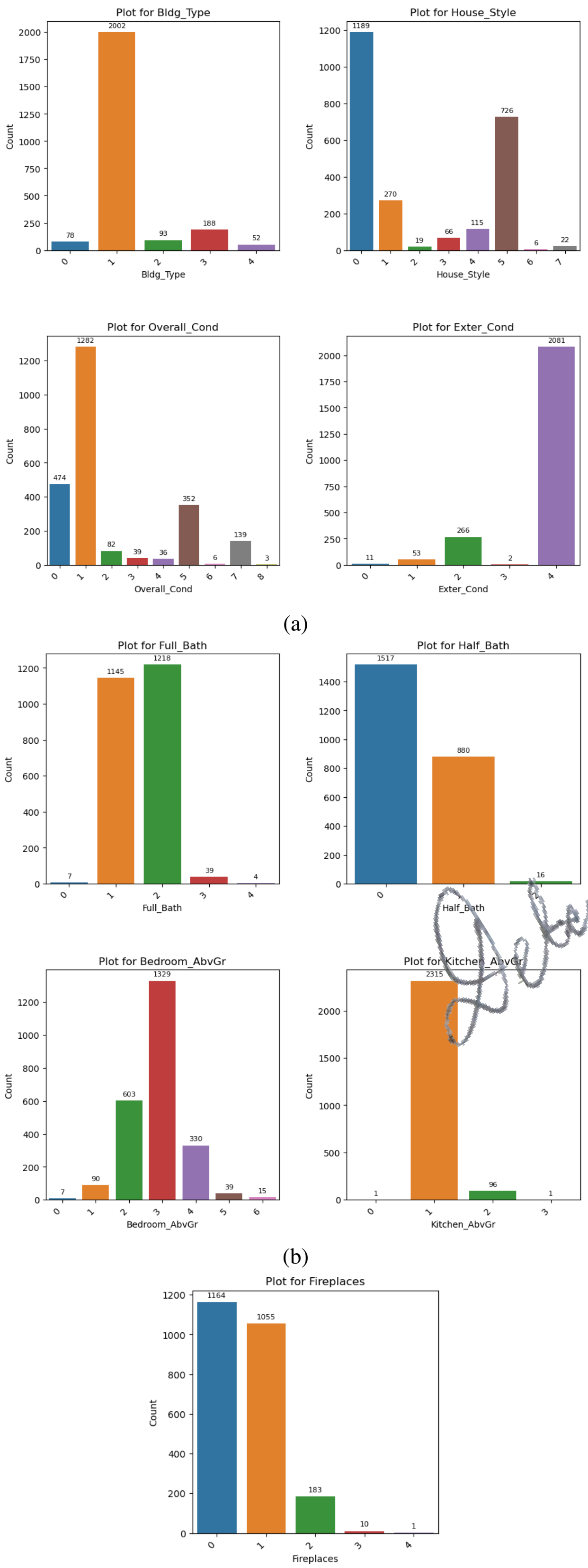
E. Correlation Analysis with further visualization

The correlation matrix is used to understand the linear relationship between numeric variables (except for the object type categorical features) using their correlation coefficients. The heat map in Fig.5 shows the correlation distribution in different cells where the value can be seen ranging from -1 to 1. From the correlation heat map generated, it can be inferred that the columns Year_Built, Total_Bsmt_SF, First_Flr_SF, Full_Bath have positive correlation among them and have tall correlation coefficients close to 1 between the pairs. This suggests a strong linear relationship between them and may indicate multicollinearity. Multicollinearity needs to be handled as it may cause issues in regression analysis.
On checking the skewness of the above 4 features it is observed that the First_Flr_SF has a skewness of more than 1. This indicates a substantial deviation from normality. So to handle the distribution properly, Log Transformation is applied. After the Log Transformation, the skewness of First_Flr_SF is 0.0246 which is now normally distributed.
Quantile-Quantile plot is used to visualize data distribution against the expected normal distribution. Q-Q plot compares the observed data quantiles against the quantiles of expected normal distribution. From the observation, it can be inferred that most of the features follow the expected distribution with less deviation.

III. Data Preparation

A. Encoding Categorical feature

A label encoder is used to encode the object-type Categorical features. Here Label encoder is used to preserve the ordinal relationships between the categories where objects per category are high in number. A label encoder is the most suitable option in this case. Here the object type categorical columns are 'Bldg_Type', 'House_Style', 'Overall_Cond', and

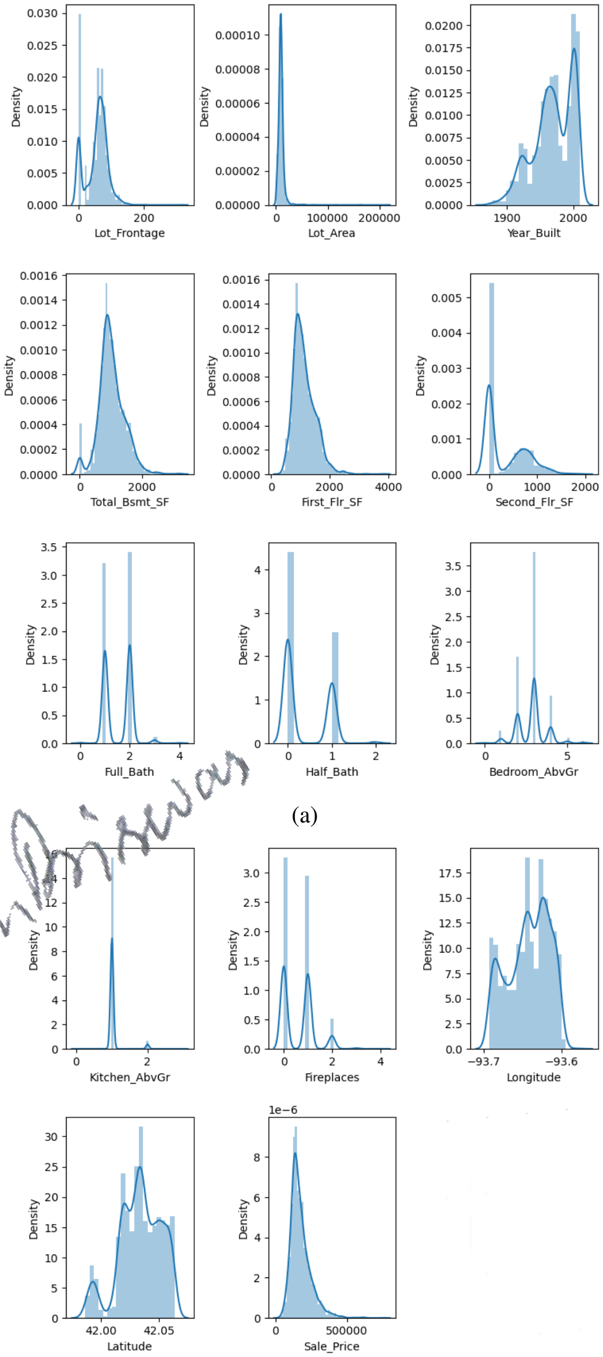


(a)

(b)

(c)

Fig. 3: Bar plot of All Categorical Columns



(a)

(b)

Fig. 4: Distribution plot of all numerical features

'Exter_Cond'.

The encoder encodes the textual value of the categorical columns in numeric values ranging from $0 - 8$ as, per our data set. On rechecking the skewness of the new encoded data set, it is observed that the skew value of some non-collinear features is high. This high value can be further adjusted upon the removal of outliers.

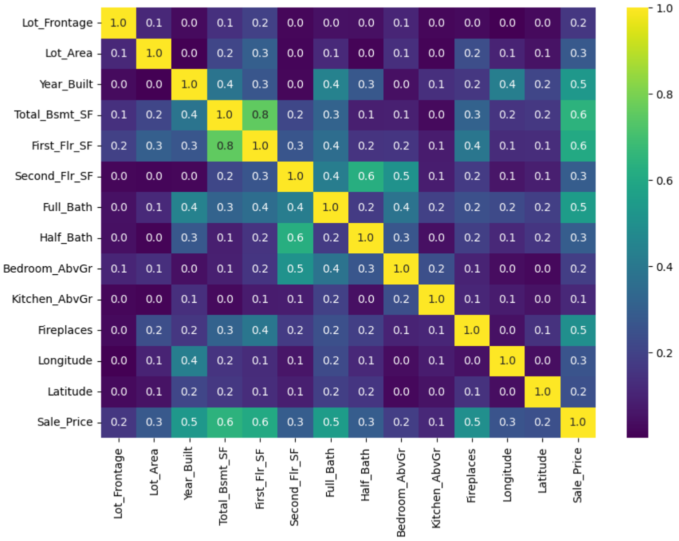


Fig. 5: Correlation Heat map

B. Outliers removal

Box plot is applied on the encoded data set for identifying outliers and the spread of the data. Outliers are present in the data set as can be seen in Fig.1. Outliers can influence the outcome of the prediction and it needs to be removed.

Z Score of all independent features are calculated except for the Sales_Price column. Sales_Price column is not used as it is a dependent feature and is influenced by the other independent features. The z score threshold is considered as 2 for this data set. The rows having a z score above the mentioned threshold are to be removed to achieve clean data. After removing the outliers and resetting the index, the row count falls to 1234. On further checking the skewness, it can be observed that a better skewness is prevalent than the previous.

C. Observing the cleaned data with visualization

Further visualization techniques are used to confirm the data is cleaned and distributed properly. From the observation of the new Correlation Heat map in Fig.6, it can be inferred that the data set is free of multicollinearity. The features Exter_Cond and Kitchen_AbvGr have 0 skewness. That means these features are normally distributed without any deviation.

The Box plot visualization of Fig.7 now confirms that the majority of outliers are successfully removed, which may influence the prediction.

IV. Modelling

A. Data Splitting

The data set is split into 2 subsets, one having all the independent features and the other with only the dependent feature ie Sales_Price.

B. Adding Seed into Training and Test Set

As per the outline of the project, Data needs to be seeded as per the student ID. The Student here is X22242821, so the seed applied is 22242821. By applying a fixed seed, it will generate the exact same result every time the program runs, thus maintaining rigidity and consistency. The clean data set is split into Train and Test sets, where the length of Train and Test data are 925 and 309 respectively.

C. Model Selection

In this data set, two machine learning algorithms, OlS and Ridge Regression model, have been applied to check and compare the performance and effectiveness of both models. Before modelling the machine learning models, the data set is scaled using a standard scaler to standardize the features ensuring every feature has a similar scale.

Ridge Regression is a linear regression or regularization method that adds a penalty term to the traditional least squares objective function which is proportional to the magnitude square of the coefficients. Ridge regression ignores large coefficient values which helps to prevent data over-fitting. The regularisation strength (alpha) is a parameter that can be adjusted as per the regularization. Here the alpha value is taken in log space of range -3 to 4 as the 17 inputs features have their coefficients ranging from e-03 to e$+04$. Both the intermediate and final models have been implemented to compare their effectiveness and efficiency.

Ordinary Least Squares (OLS) regression is used here as a machine learning algorithm to estimate the unknown parameters. Here OLS is used as it is straightforward and intuitive and provides estimates for the coefficients along with standard errors. OLS provides estimates that minimize the sum of the squared residuals and this makes the OLS estimate the best linear unbiased estimates. Both the intermediate model and final model have been used here to compare their effectiveness on the data.

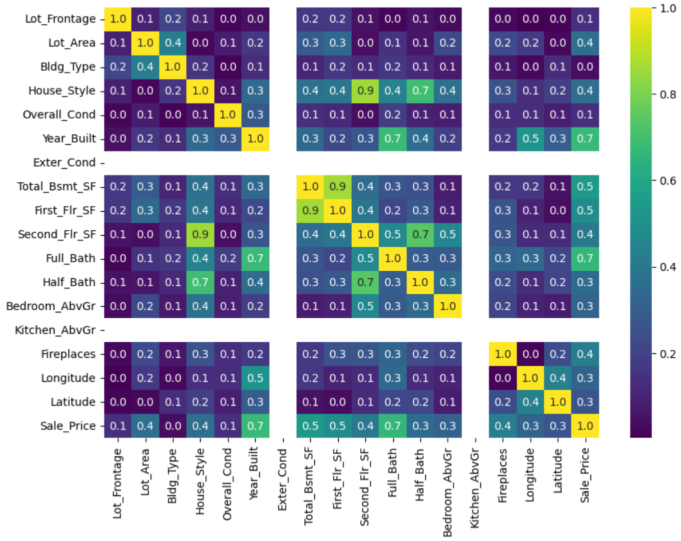


Fig. 6: Heat map after Data Preparation

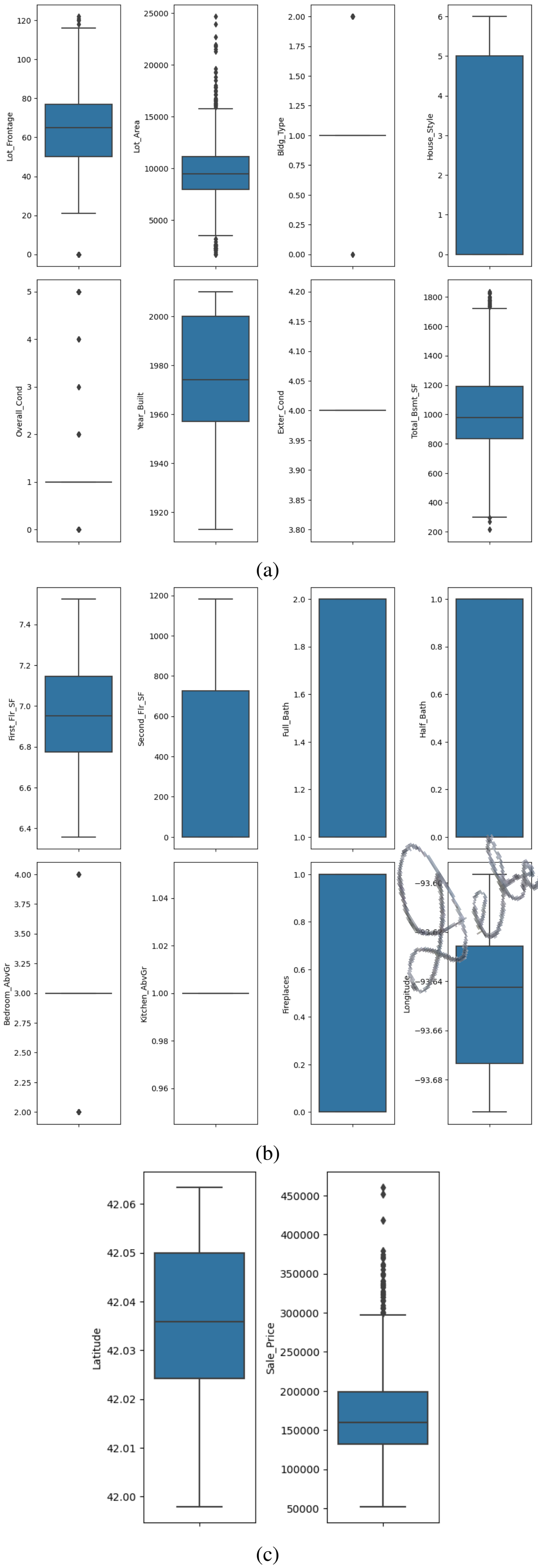


Fig. 7: Box plot after Data Preparation (Outlier removal)

V. Interpretation

Ridge Regression Model:
From both the intermediate and final Ridge Regression Model, it can be observed that the coefficients and the Intercepts fetches same values respectively. This value convergence suggests model stability indicating a consistent solution. The columns Exter_Cond and Kitchen_AbvGr show 0 value which indicates that these two features have no impact on the predicted values. This happened due to multicollinearity. The plot of the coefficients of the Ridge Regression Model can be seen in Fig.8.

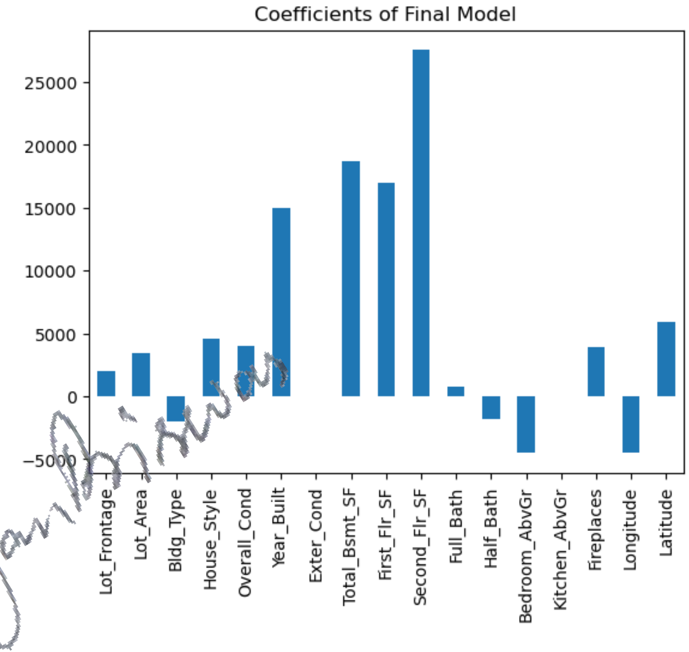


Fig. 8: Coefficient Plot of Ridge Regression Model

OLS Model:
From the observation of both the summary of the intermediate and final OLS model, it can be inferred that both the model's output fetches a similar value of the coefficients. This suggests that the model optimization has successfully converged to stability. The F-statistic value of 295.7 indicates that the model is statistically significant. It is also observed that almost all the P-values obtained are 0 or near 0 value. This combined with a high value of F-Statistics suggests strong evidence against the null hypothesis, depicting at least one of the independent feature has a significant effect on the dependent feature in the model.

VI. Diagnostics

Ols Model:
Shapiro-Wilk Test for Normality and Breusch-Pagan Test for Homoscedasticity have been applied to the OLS model to check the p-value of both tests. The p-value obtained is small suggesting the distribution deviates slightly from normal distribution and strongly suggests against the null hypothesis. Heteroscedasticity may be present in the model.

The presence of Linearity, normality, and Homoscedasticity is checked on the OLS model. The residuals are spread equally around the horizontal line without any pattern which can be seen in Fig.9. This depicts that the model is linear. The Q-Q plot in Fig.10, shows that the majority of the residuals follow the red line, thus confirming the normality of residuals. The homoscedasticity plot in Fig.11 shows that the residuals are randomly spread over the horizontal red line with random spread points. This suggests a minute case of Heteroscedasticity in the model. The histogram of the residuals in Fig.12, suggests a normal distribution trend.

Linearity, normality, and Homoscedasticity are checked on the Ridge Regression model. The residuals are spread equally around the horizontal line without any pattern with less scatter. This depicts that there is linearity in the model. The Q-Q plot shows that the majority of the residuals follow the green line, thus confirming the normality of residuals. The homoscedasticity plot shows that the residuals are randomly scattered over the horizontal green line with random spread points. This suggests heteroscedasticity in the model. The histogram of the residuals suggests a normal distribution trend.

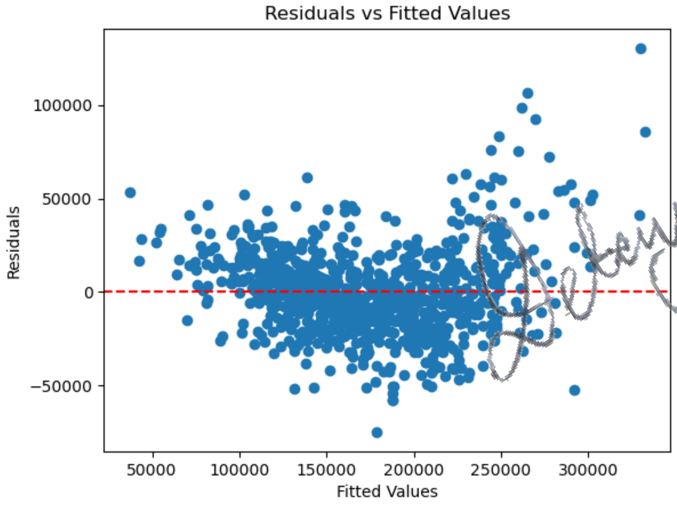


Fig. 9: Residuals vs Fitted Plot for Linearity

The final model meets most of the Gauss-Markov assumptions like Linearity, Independence, and Normality but shows Heteroscedasticity.

VII. Evaluation

The Ridge Regression model gives two slightly different results for the Intermediate and Final Models. The Mean Square Error and R-square for the Intermediate model are 574253862.0523953 and 0.830457239866615. The Mean Square Error and R-square for the Final model are 574253862.0523627 and 0.8304572398666247. The mean squared difference is the average squared difference between the actual values and the model's prediction. The value of MSE is influenced by the scale of the dependent variable.

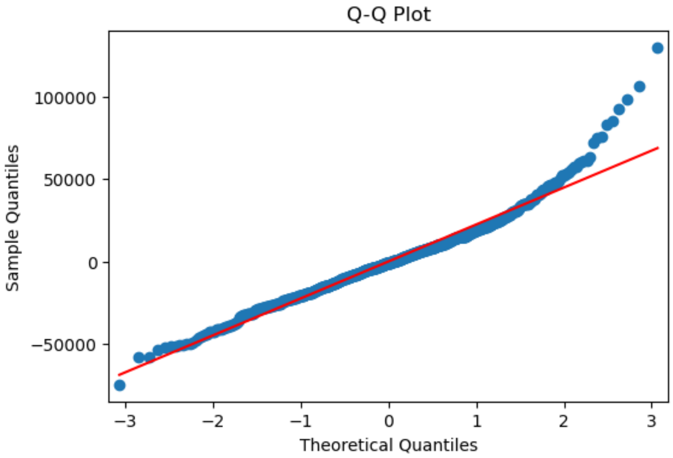


Fig. 10: Q-Q Plot for Normality of Residuals

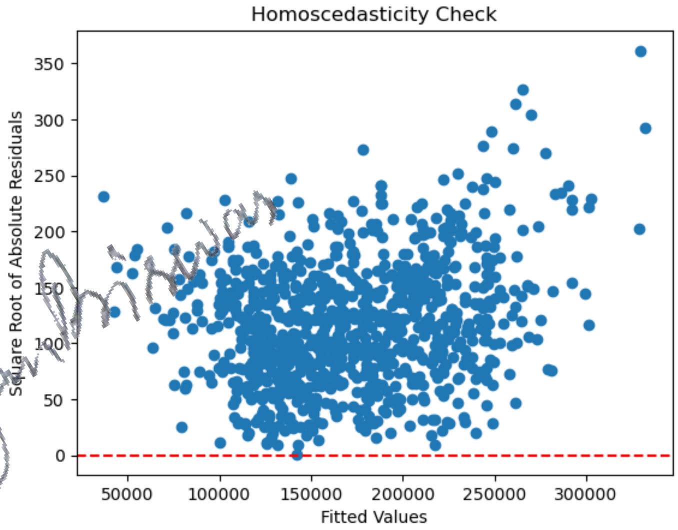


Fig. 11: Homoscedasticity

The R-squared value depicts the effectiveness in explaining the variability of the dependent variable.

The Ols model when evaluated gives the same result for both the Intermediate and Final Models. The Mean Square Error is 572659942.0445561 and the R-square value is 0.8309278289482366. The R-square value is very close to 1 and this signifies that the model perfectly explains the variability in the target variable. The mean squared difference is the average squared difference between the actual values and the model's prediction. The scale of the dependent variable influences the value of MSE.

The Test Final model result for the Ols Model gives a better result for both the cases of R-squared and Mean Square Error (MSE). The MSE is however on the higher side. The MSE can be better worked on by adjusting the skewness of some of the right-skewed variables. The MSE can be further minimized if the skewness and outliers from the dependent variable can be removed. This will also remove heteroscedasticity ensuring

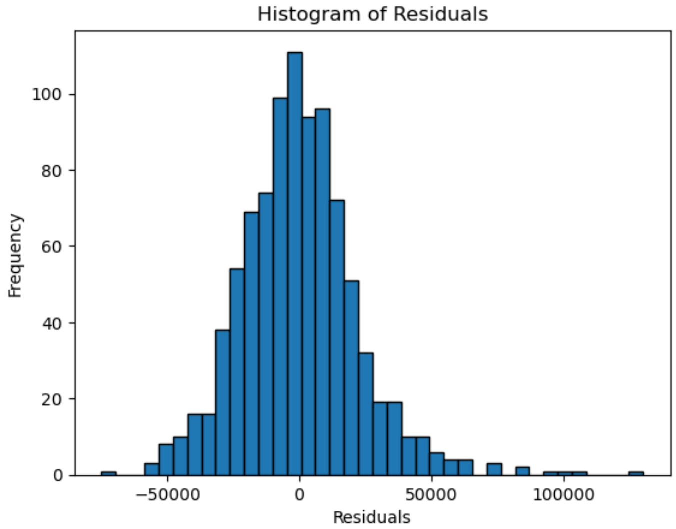


Fig. 12: Histogram of Residuals

more reliability of statistical inferences and predictions made by the regression model. However, removing the outliers and fixing the skewness for the Dependent Column Sale_Price is not a correct approach as the value of Sale_Price is solely dependent on the Independent features and not vice-versa.

VIII. Interpretation of the parameters of the final model

Interpreting the parameters of the final regression model analysis involves understanding the meaning of the coefficients associated with each independent feature. Each coefficient represents the change in the predicted dependent variable per unit of change in the corresponding independent variable taking the rest of the variables as constant. The coefficients depict a wider range of values. This can be further reduced by scaling the dependent variable. However, the dependent variable can be anything in the real world and will only be influenced by the independent variables. Thus outlier removal and fixing skewness was not performed on the dependent variable.

IX. Conclusion

A multiple linear regression strategy was used on the Housing data set to assess the effect of variables such as Lot_Frontage, Lot_Area, Bldg_Type, House_Style, Overall_Cond, Year_Built, Exter_Cond, Total_Bsmt_SF, First_Flr_SF, Second_Flr_SF, Full_Bath, Half_Bath, Bedroom_AvgGr, Kitchen_AvgGr, Fireplaces, Longitude, Latitude on the Sale_Price.

The model was built using the Ordinary Least Square and Ridge Regression method. The Ols Model provided a better result than the Ridge model with a higher R-squared value and lower MSE value. Linear relationship (between dependent and independent variable), independence of residuals, homoscedasticity, and normality were quantified in the Jupyter notebook using visualization and regression.

The R2 value which is 0.8309278289482366, indicate model correctness.

This project report demonstrates that all the independent factors had a significant effect on the dependent feature, Sale_Price.

Acknowledgments

The completion of this project has been made possible through the support of Professor John Kelly, for his guidance, advice, and critical feedback.

References

[1] *Statistics Unplugged Edition:3e*. by Sally Caldwell

[2] Introduction To Computer Science Using Python. *A Computational Problem Solving Focus*. by Charles Dierbach

[3] Statology: lists all of the basic statistics tutorials https://www.statology.org/tutorials/

[4] Understanding Ordinary Least Squares (OLS): The Foundation of Linear Regression by Vitor C Sampaio https://tinyurl.com/VictorCSampaio

[5] Datacamp *: Handling Machine Learning Categorical Data with Python Tutorial*. https://www.datacamp.com/tutorial/categorical-data

[6] Python for Data Analysis, by W. McKinney, edition 2, 2017, O'Reilly Media publishers

[7] Github, Python Data Science Handbook, by J. VanderPlas, https://jakevdp.github.io/PythonDataScienceHandbook/

[8] Econometrics, by K. Nirmal Ravi Kumar, edition 1, 2020, CRC Press Publishers