**House Price Prediction Model Using CRISP-DM Methodology**

**Programming In R**

**Table of Contents**

Introduction............................................................................................. 3

Overview of the CRISP-DM Framework................................................... 3

1. Business Understanding….................................................................... 4

2. Data Understanding............................................................................. 4

3. Data Preprocessing ............................................................................ 14

4. Modelling ........................................................................................... 16

5. Evaluation............................................................................................18

6. Deployment ....................................................................................... 21

Conclusion.............................................................................................. 21

**Implementing a House Price Prediction Model**

**A CRISP-DM Framework**

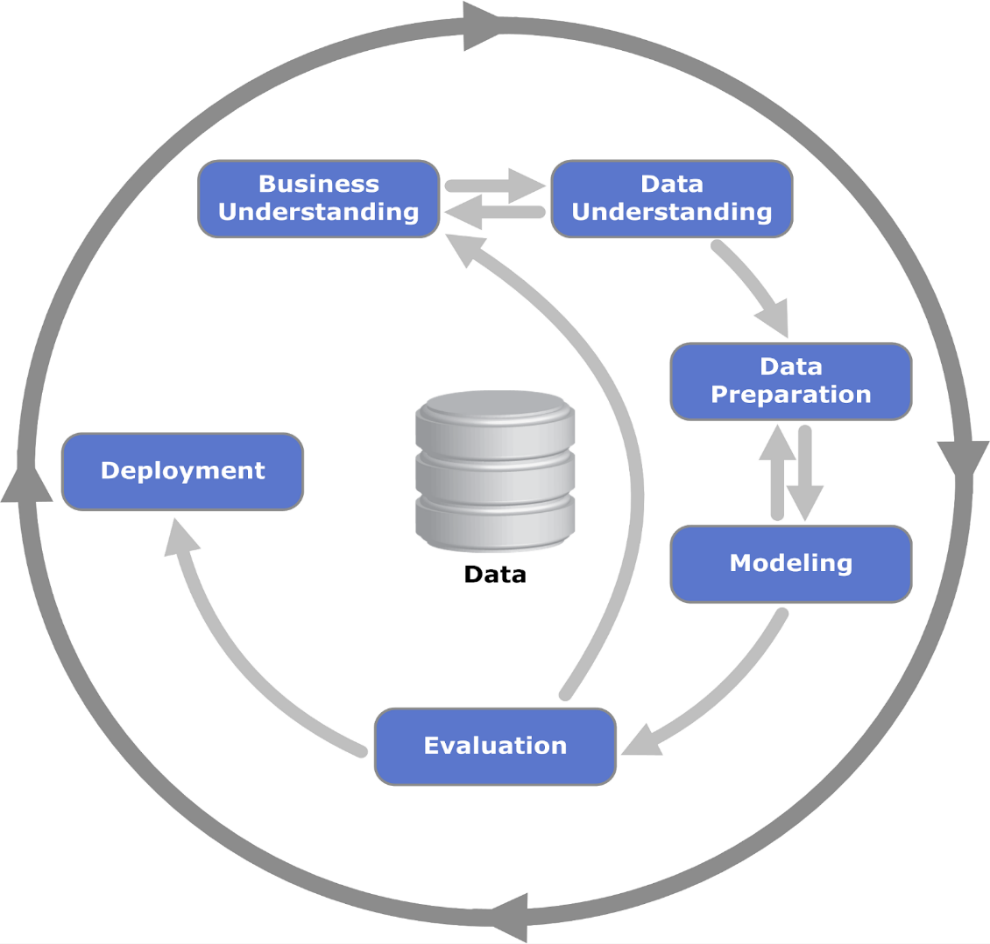
**Introduction**

The real estate market is never the same, and there are lots of variables that decide the worth of a property, be it where, how big, and state—maybe some attributes. The real estate industry is skyrocketing with house prices one of the most important tasks within the real estate and finance fields; this helps stakeholders make informed moves in buying, selling, and investing in properties. The report describes the development of house price prediction models using the Random Forest and Multiple Linear Regression machine learning techniques under the CRISP-DM framework.

**Overview of the CRISP-DM Framework**

CRISP-DM is considered one of the most effective process models for data mining and predictive analytics projects. It offers a systematic framework within which to organize, implement, and evaluate data-driven projects to guarantee that resultant models are effective and generalizable.

**CRISP-DM consists of six phases:**



1. **Business Understanding**

The initial phase of the CRISP-DM model concentrates on business understanding and detecting business objectives that lead to data mining goals. In the case of forecasting house prices, the major goal is to create a model that can make a precise forecast for the selling price of a house based on given features like locality, utilities, and the number of rooms.

**Research Questions:**

* *How does house prices vary across different neighborhoods?*
* *What are the relationships between various features of the houses?*

**Purpose of This Model:**

1. **Data Understanding**

In the data understanding phase, our main objective is to bring together and investigate the data so that we can make sure we know its structure, quality, and its hidden problems. The house-price predicting model usually consists of a number of significant characteristics among which are:

**2.1) About the Dataset:**

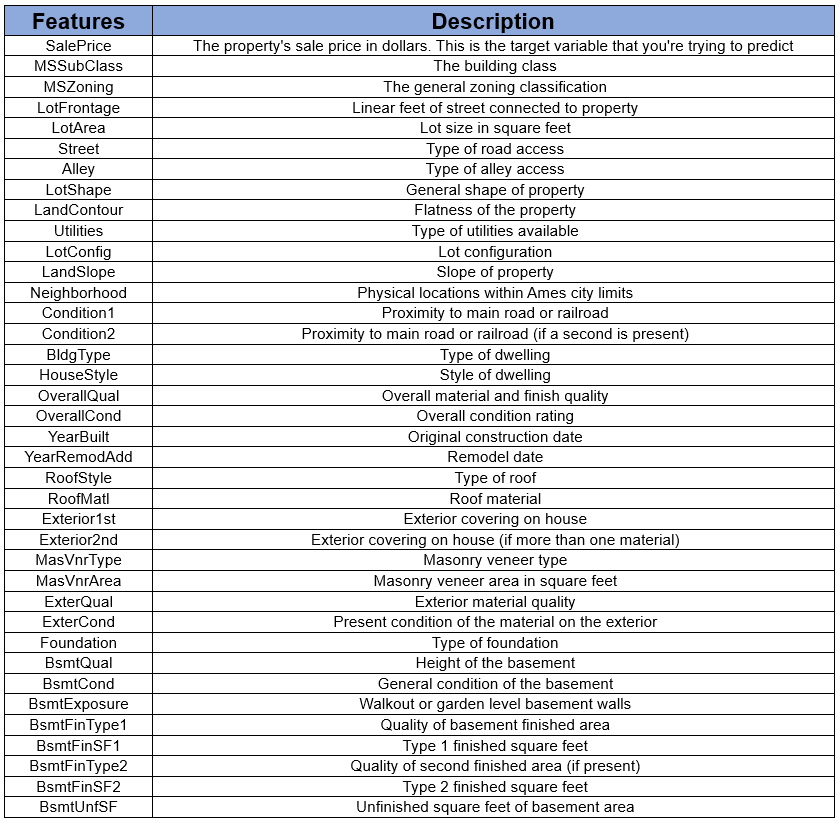
The data has the information related to American housing and was obtained from:

<https://www.kaggle.com/code/chanakyavivekkapoor/house-price-prediction/input>

* **Data Information:**

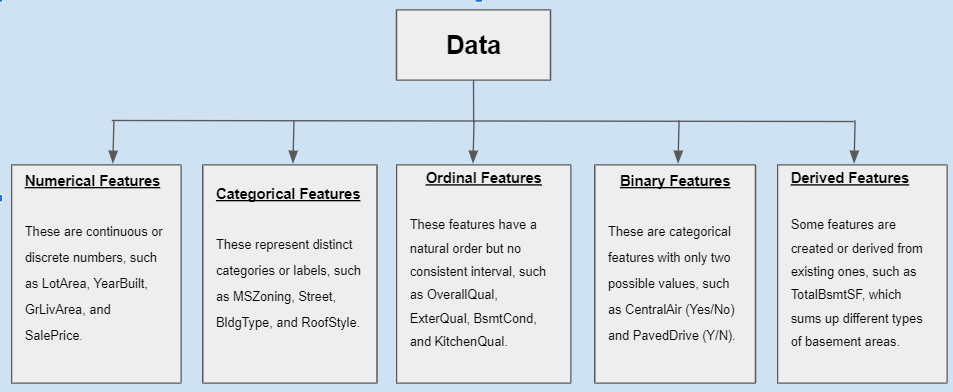
The dataset comprises of **81 headers** and **2919 rows**, containing all types of property attributes, which can be used in order to predict house prices, such as physical characteristics, location, and conditions of sale.

The below table shows information of each feature in detail:





The features can be broadly categorized into:



**2.2) Exploratory Data Analysis (EDA):**

1. **Visualization**

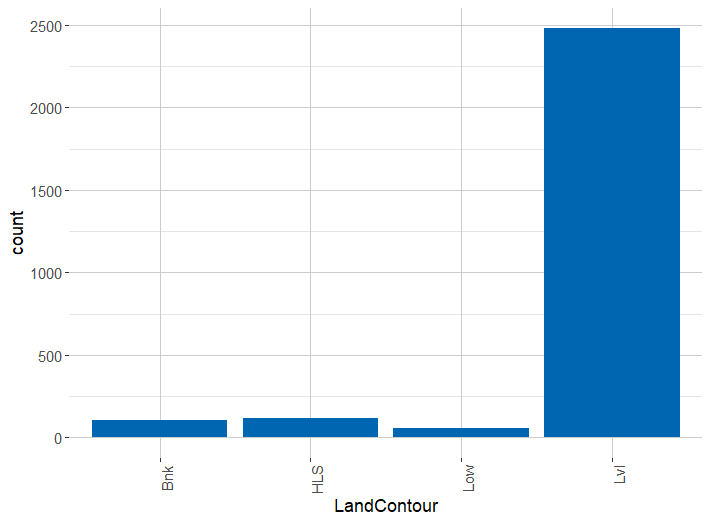
Visualization in predictive modeling is important since identify patterns, trends, and outliers in the data; to drive feature selection and preprocessing; and to understand model performance and the different roles of features in the predictions.

**BarPlot representation of categorical features**

**Description of features with Bar charts:**

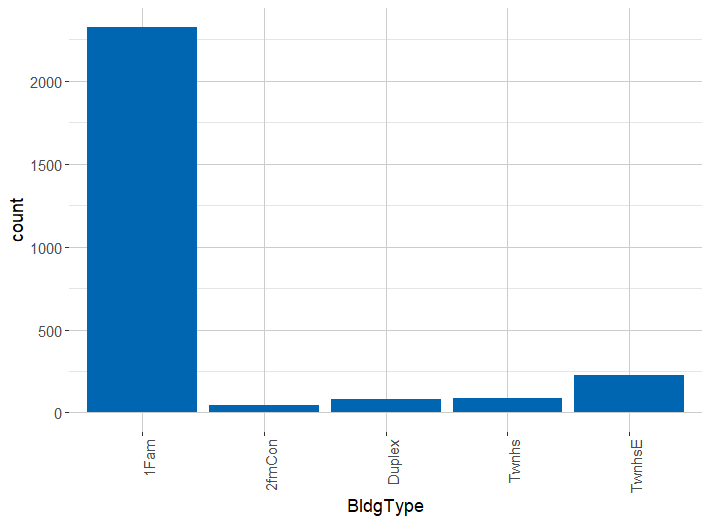
* **LandContour**

The majority of the properties are on a "Level" (Lvl) land contour, indicating flat terrain, with only a small percentage on other contours like "**Bnk**", "**HLS**", and "**Low**".



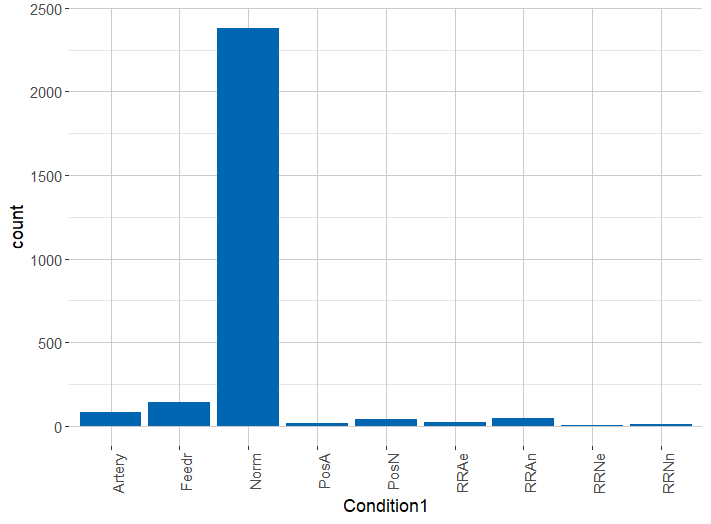
* **BldgType:**

The most common building type in the dataset is "1Fam"—single-family homes. It is followed by other varieties, well behind in number, like "TwnhsE" and "Duplex.".



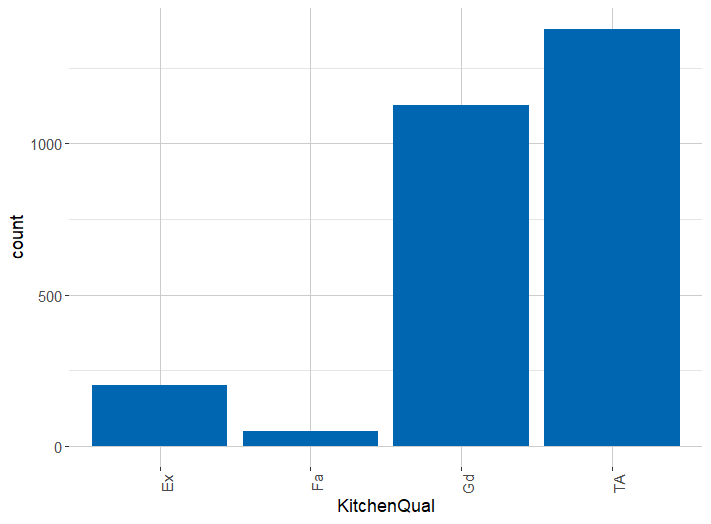
* **Condition1**

Most of the properties are in "Normal" condition with respect to proximity to various conditions—like railroads or main roads—and very few are in "Feedr" and "Artery" conditions.



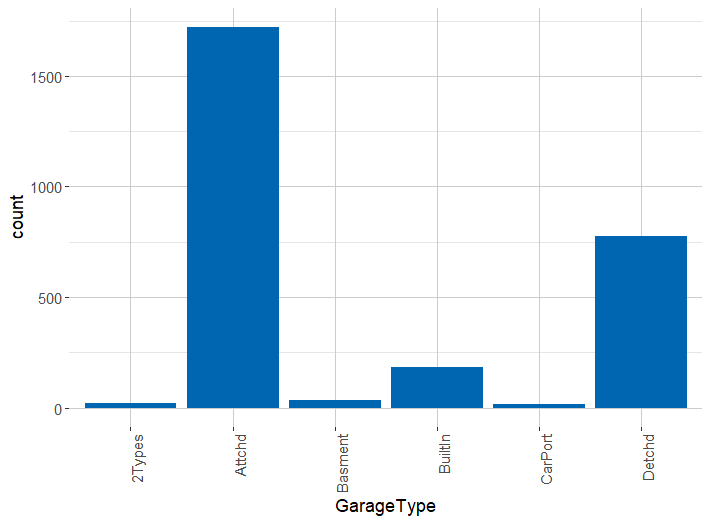
* **KitchenQual**

A majority of the properties have "TA" (Typical/Average) kitchen quality, followed by "Gd" (Good), with very few having a "Fa" (Fair) or "Ex" (Excellent) kitchen quality.



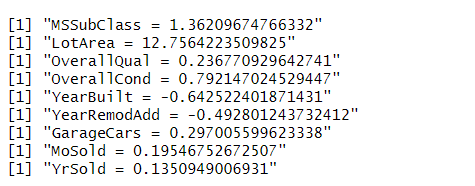
* **GarageType**

Attached Garage is on top of the list with a count of more than 1500 houses, followed by detached and Builtin in that order. 2 Types, Basement, CarPort are almost the same.



1. **Statistical Use:**

This dataset contains a lot of outliers, which distort the data. The skewness measures were included to obtain the asymmetry of the distribution of a variable around its mean. We check this to see that data is symmetrically distributed without/or less outliers.



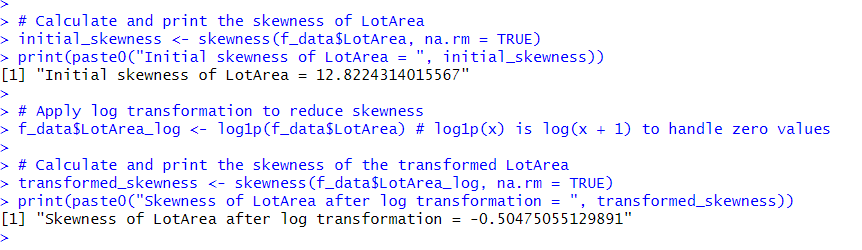
**Analyse Skewness Results:**

Based on the observation,

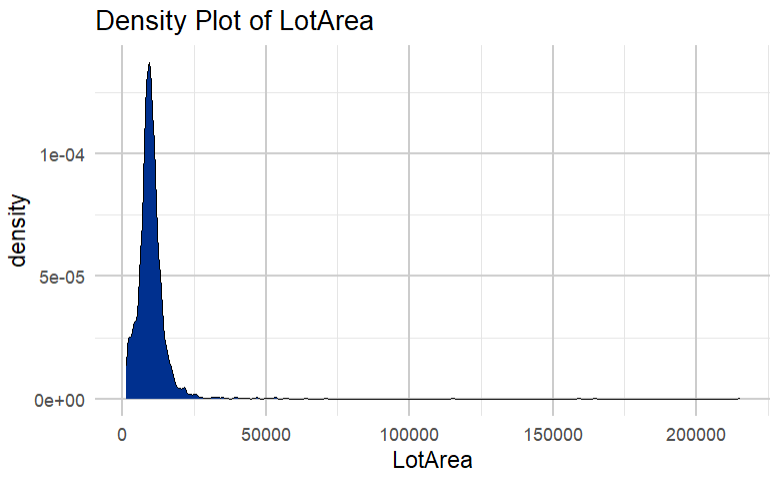
* **LotArea** has High Positive Skewness.
* **OverallCond** has moderate positive Skewness.
* **YearBuilt** has moderately negative Skewness.
* **MoSold** and **YrSold** has Low Positive Skewness(near-symmetry).

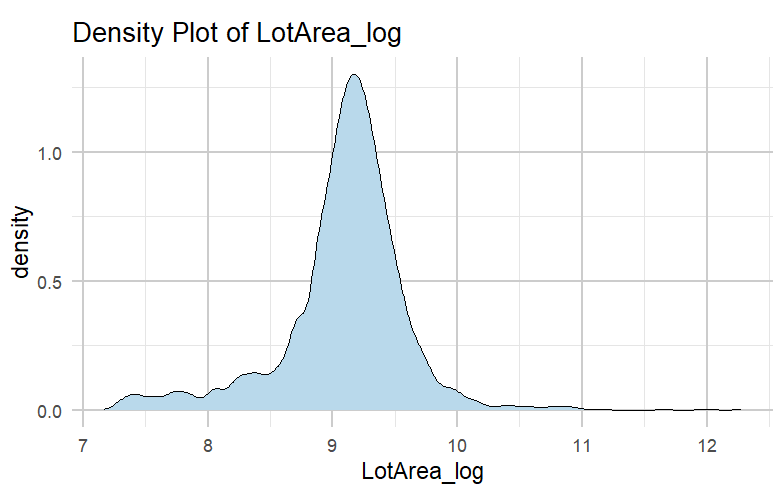
**Skewness Correction:**

Applied log transformation to reduce skewness

****

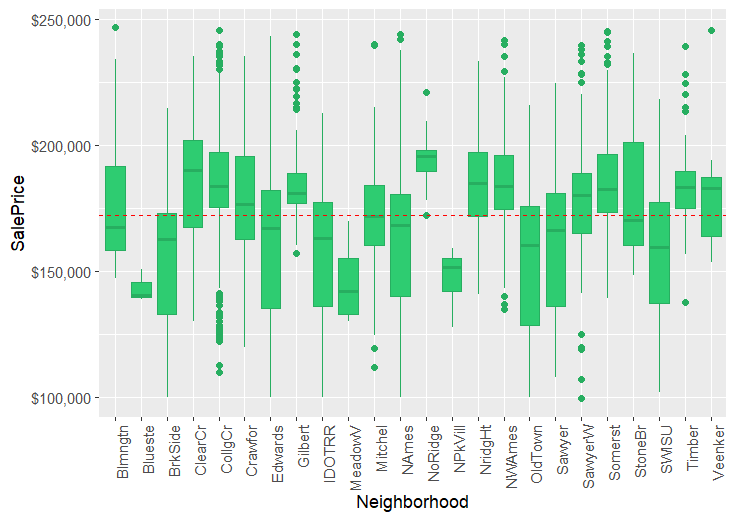
**Density graph to check the distribution of data and skewness:**

****

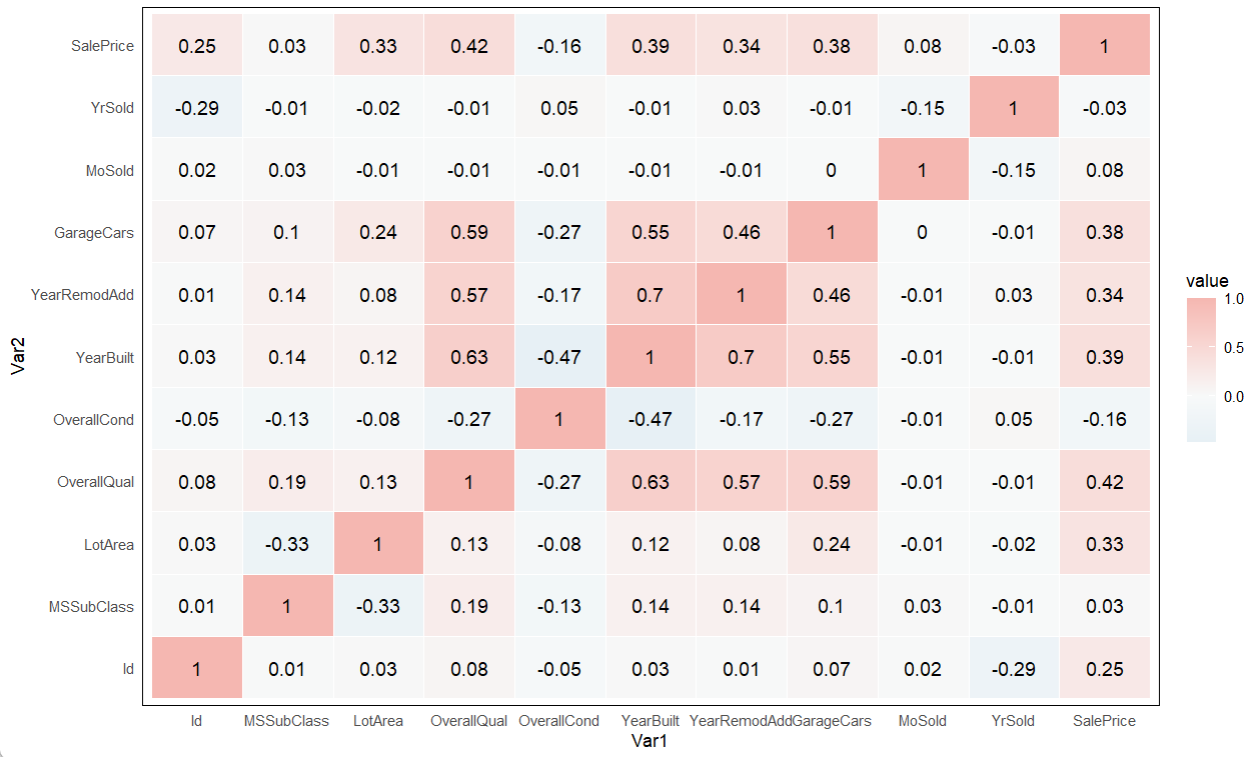
****

**Boxplot to check the prices in different area**

* Observed range, median, and variability of house price in each neighborhood.
* The red dashed line shows the overall mean sale price across all neighborhoods, which offers some sort of yardstick by which the performance in individual neighborhoods may be measured.

****

****

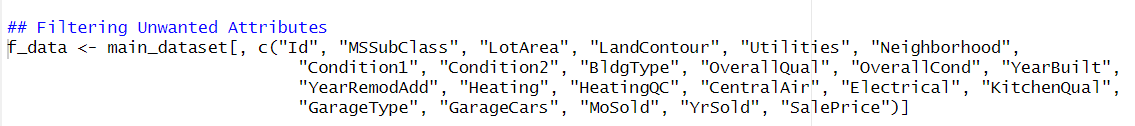


1. **Data Preparation**

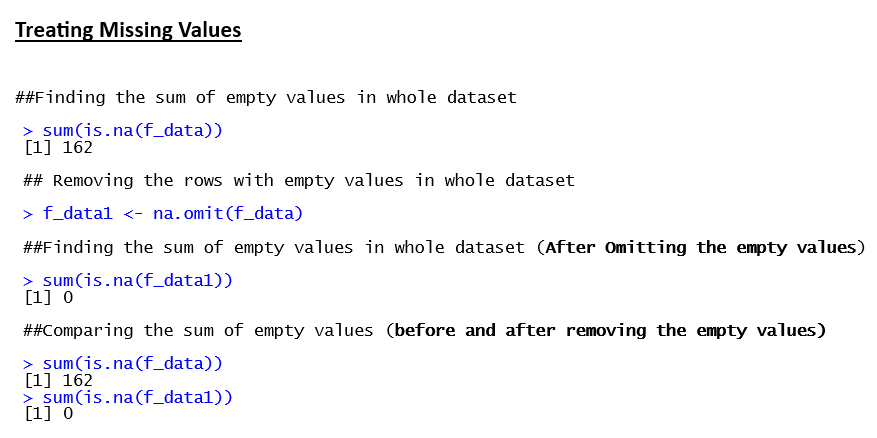
Data Preparation is the most important phase in methodology, ensuring that clean and well-structured data is prepared for modeling by getting rid of any column or data that may hamper prediction attributes through designated steps.

**Data Cleaning:**

* **Removing Unwanted Attributes:** There were 81 columns in our dataset andall features were not relevant for the predictive modeling task and were removed. Below is the code snippet where we filter unwanted attributes, and only 23 attributes will remain:

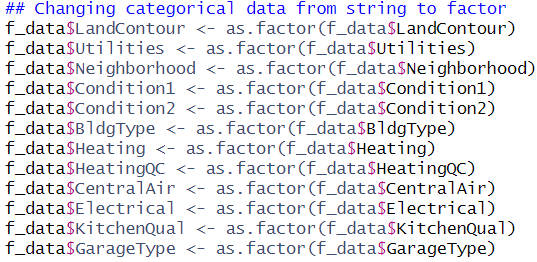


* **Handle Missing Values:** The presence of missing values inside the dataset can effectively influence the performance of machine learning models and can tamper the accuracy of modeling. According to the House Pricing Dataset, there were 162 empty values in the dataset that were detected by ***is.na*** and removed by ***is.omit***.



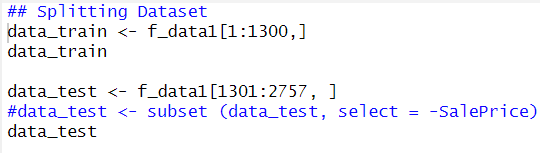
**Data Transformation:**

* **Converting Categorical Variables:** In the f\_data dataframe, which contain categorical variables, are being converted from their string (or character) format to factors**.** In R, factors are used to deal with categorical data in a better way so that it can be used in statistical modeling where the distinct categories are treated as levels.



This step ensures that the machine learning algorithms interpret and use these categorical features in a correct manner while training the model.

* **Splitting the Data into Training and Testing Sets:** The dataset was divided manually in two subsets where 1300 rows were used for training and the rest for testing. In this way, there will be a clear distinction between the data used in the training of the model and the data used in its evaluation.



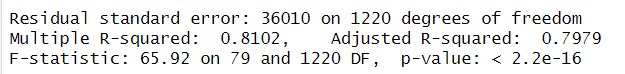
This was chosen for exact control of the rows using this for training and testing, in the interest of consistency and reproducibility.

1. **Modeling**

The modeling phase involves the development, training, and testing of predictive models to solve a house price-related business problem. For the sake of current research, two machine learning models are implemented with the intention of assessing their performance to determine which model gives superior predictive accuracy: **Random Forest** and **Multiple Linear Regression.**

**4.1. Multiple Linear Regression**

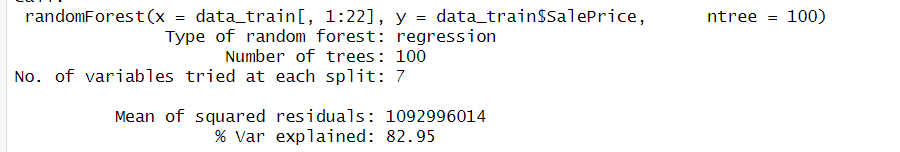
**Multiple Linear Regression** is a parametric approach based on the assumption of linearity in the relationship that exists between independent variables (predictors) and the dependent variable, which is the target. Logically, this method fits a linear equation to the observed data, thus being very easy and interpretable to apply.

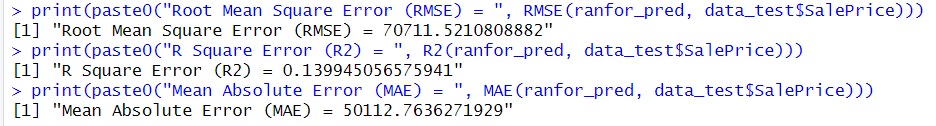


**Performance Summary**: The amount of variance in house prices explained by the multiple linear regression model is 81.02%, quite comparable to that of the random forest model. The Adjusted R-squared value considers the number of predictors—in other words, the complexity of the model—is well worth it at 79.79%. The lower residual standard error shows that predictions from the model are more consistent, with smaller average errors. The large value of the F-statistic confirms that the predictors as a whole are significant in explaining the variation of the house prices.

**4.2. Random Forest**

**Random Forest** is a form of ensemble learning where many different decision trees are created during training, and it outputs the mean of the predictions in order to perform regression tasks. It becomes very effective at capturing complex nonlinear relationships within the data itself.





**Performance Summary**: The Random Forest model accounts for 82.95% of the variance, so it does capture most of the underlying patterns in the data. However, the **RMSE** and **MAE** values are quite high, so the model's predictions have considerable average errors. On the other hand, the low R² value of 0.1399 might tell that while the model is complex and does capture variance, it might not generalize well on unseen data, likely due to overfitting.

**Model Comparison:**

On this dataset, the performance of the Multiple Linear Regression model in general will be better than that of the Random Forest model. While Random Forest could capture most of the variance, its predictive accuracy was less reliable and pointed to possible overfitting by returning bigger errors. In contrast, Multiple Linear Regression performed quite consistently, so it should be better at predicting house prices.

1. **Evaluation**

The following metrics are used to evaluate the models:

**R-squared (R²)**: Proportion of the variance explained by the model; takes values from 0 to 1, and the closer to 1, the better the fit.

**MSE**: This is the average of the squares of the differences between the real and predicted values.

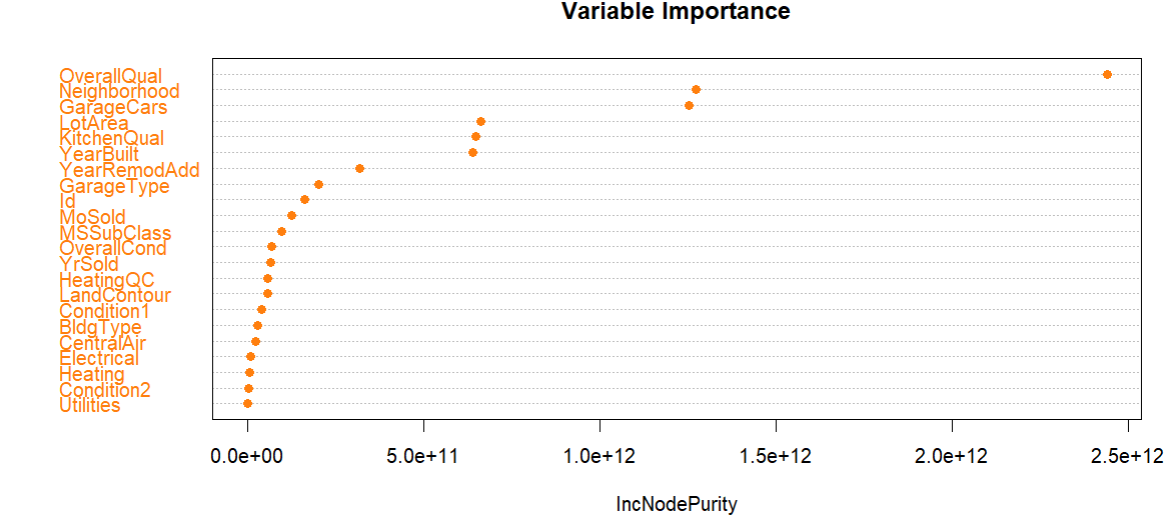
**RMSE:** It is the root mean squared error, which is the square root of MSE returning the error magnitude in the same units as the output.

One of the metrics is **Mean Absolute Error**: it provides the average absolute difference between the actual and predicted values.

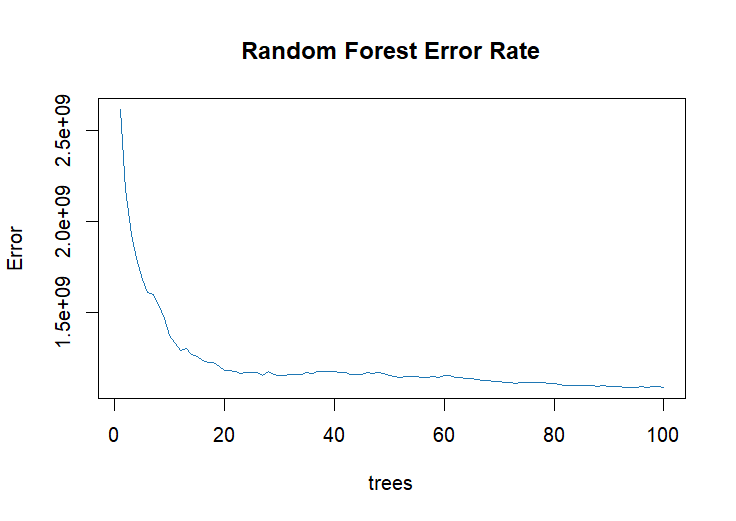
**Evaluation of the Models**

1. **Random Forest Evaluation**

* **R-squared**: For the Random Forest, the R² was 0.1399, meaning it told the variance of house pricing less well.



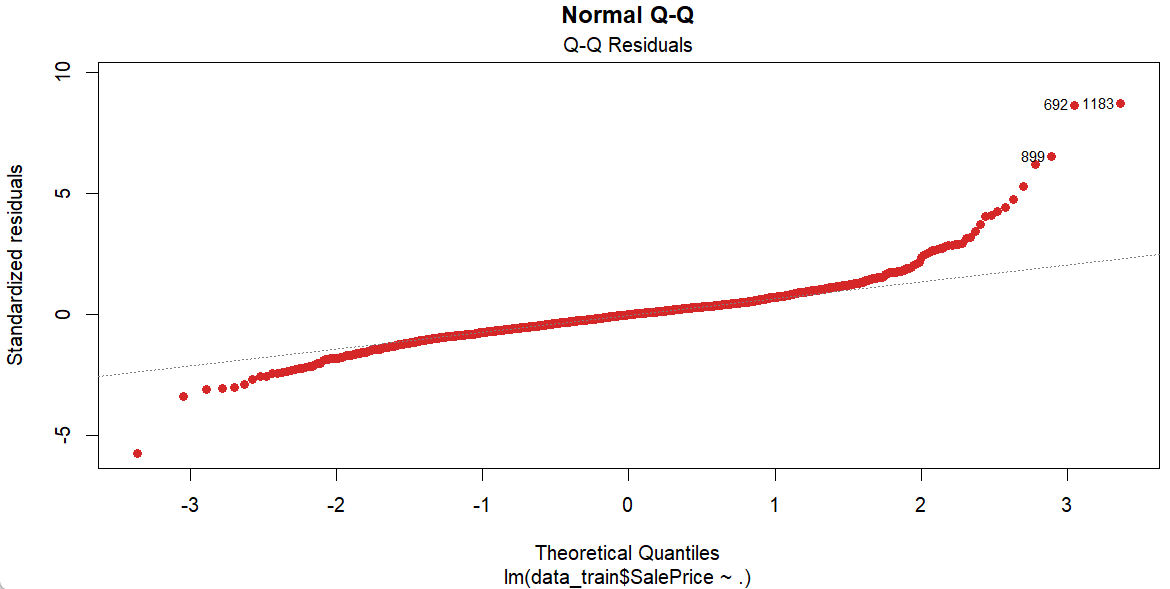
* **RMSE and MAE**: The model returned an RMSE of 70,711.5270,711.5270,711.52 and an MAE of 50,112.7650,112.7650,112.76. This would present an average very large error of predictions. This may be indicative of such problems as overfitting or high sensitivity of the model to the complexity of the data.



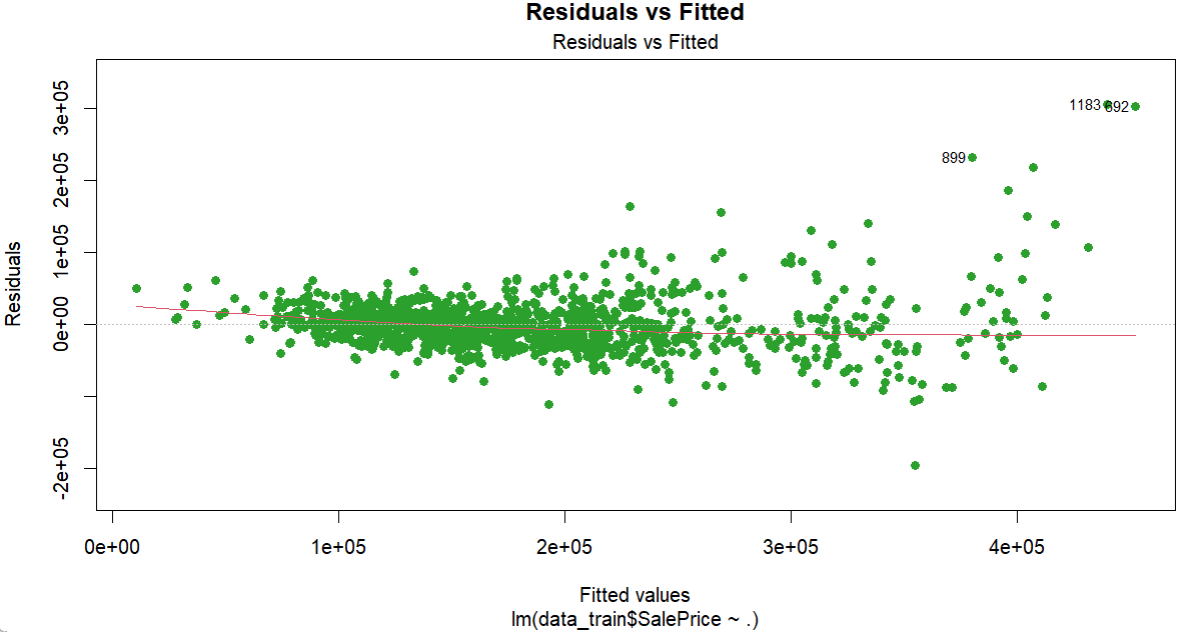
* **Conclusion**: Whereas this random forest model was very strong in capturing complex relationships, this was a case of poor performance, probably because of the nature of the dataset or the configuration of the model itself.

1. **Multiple Linear Regression Evaluation**

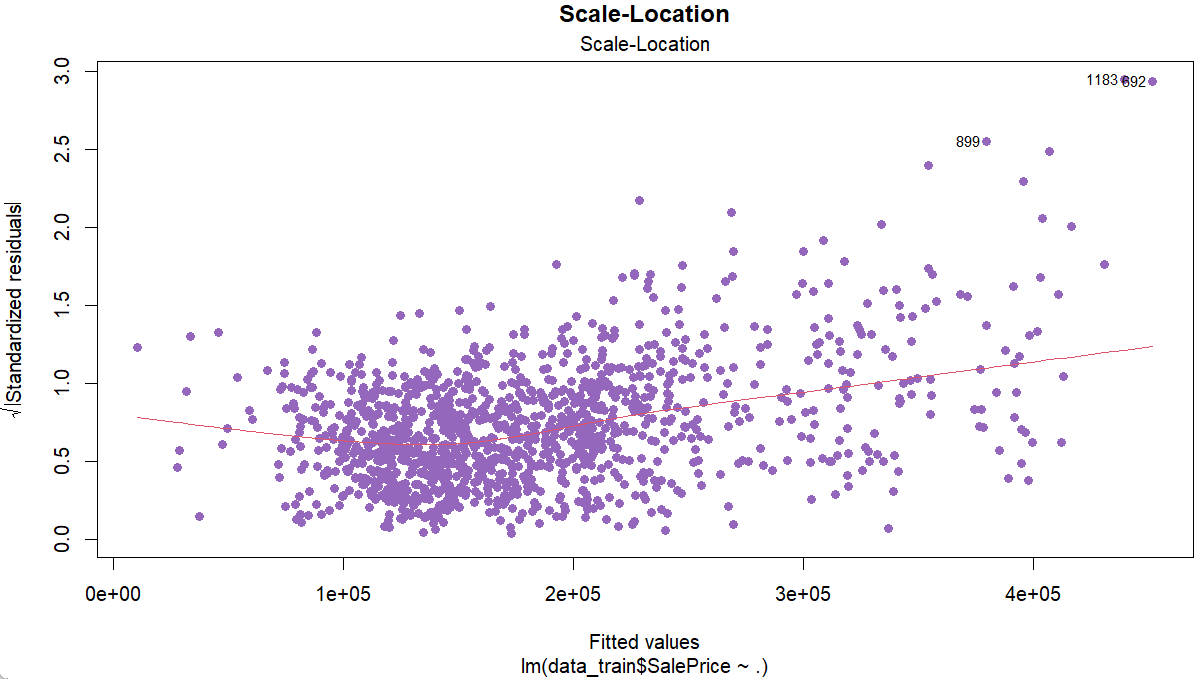
* **R-squared and Adjusted R-squared**: Multiple Linear Regression came out well, keeping R-squared at 81.02% and Adjusted R-squared at 79.79%, which means it perfectly captured those linear relationships of the predictors with the house price.



* **Residual Standard Error (RSE)**: The RSE was 36,01036,01036,010, indicating that the predictions are very near to the actual values.



* **F-statistic**: The resulting F-statistic was 65.9265.9265.92 on 79 and 1220 degrees of freedom, with a p-value of <2.2×10−16< 2.2 \times 10^{-16}<2.2×10−16. This proved that the model was statistically significant; that is, its predictors were relevant and helped in the accuracy of the model.



**Conclusion**: The model of Multiple Linear Regression demonstrates good and stable fit; hence, it is more trustworthy for this dataset than the model of Random Forest.

1. **Deployment**

This deployment plan ensures that the multiple linear regression model is not only accurate and efficient in house price prediction but also user-friendly and sustainable across time.

**Conclusion**

This report focuses on the application of machine-learning models to predict house prices using the CRISP-DM methodology. Correspondingly, enough data cleaning, missing values treatment, and outlier removal were done to prepare the dataset for modeling. We have made use of a couple of regression models, namely, the **Random Forest** and a simple **Multiple Linear Regression**, for the purpose of comparison. While more focused emphasis is given on RMSE, R², and MAE as the principal evaluation metric of every one of them. Random Forest did the job quite well, but the potentially higher error rates and more signal noise indicate that it might have picked up the complex data excessively. In comparison, the simpler setup of **Multiple Linear Regression** produced an even and realistic relapse prognosis that was equally reliable at both ends. MLR should be, thus, used for predictions about house prices within this scope. This work highlights the need for a fusion between domain knowledge and statistical machine learning in providing pricing outputs which should be credible and clear to the real estate stakeholders.