Kent State University

Ambassador Crawford College of Business and Entrepreneurship

Course: Business Analytics

Section: BA-64036-005

Batch: Fall 2023

Instructor: Prof. Mostafa K. Ardakani, Ph.D.

Final Project: Finding the Optimal Model for Accurate Predictions

Group – II

Group Members	Contribution
Spandana Sodadasi	R Code, Report
Keerthi Tiyyagura	Report
Anusha Banda	Report
Samyuktha Ananthan	Power point presentation & Voice Recording

➢ Project Goal:-

In the realm of data analysis and prediction, the choice of an appropriate model is pivotal, as it directly impacts the quality of insights and decisions derived from the data. A well-suited model ensures optimal performance, aligning with the specific characteristics inherent in the dataset. Hence, the objective of our project is to meticulously evaluate and compare various models such as Regression, Decision Tree, and Logistic Regression, with the aim of identifying the model that best aligns with the House Prices dataset. Through this process, we seek to unlock meaningful insights and enable well-informed decision-making based on the data at hand.

➢ Overview of the Data:-

1. Dataset Used for Modeling:

We use House Prices.csv dataset for training the model and perform prediction on the given testing dataset. The dataset consists of 13 variables which will be described as follows,

- **LotArea:** Lot size in square feet
- **OverallQual:** Rates the overall material and finish of the house. 10 Very Excellent; 9 Excellent; 8 Very Good; 7 Good; 6 Above Average; 5 Average; 4 Below Average; 3 Fair; 2 Poor; and 1 Very Poor.
- **YearBuilt:** Original construction date
- **YearRemodAdd:** Remodel date (same as construction date if no remodeling or additions)
- **BsmtFinSF1:** Finished square feet
- **FullBath:** Full bathrooms
- **HalfBath:** Half baths
- **BedroomAbvGr:** Number of Bedrooms above the ground
- **TotRmsAbvGrd:** Number of rooms above the ground
- **Fireplaces:** Number of fireplaces
- **GarageArea:** Size of garage in square feet
- **YrSold:** Year sold
- **SalePrice:** The sale price of the property

2. Descriptive Analysis:

Descriptive statistics are crucial in data analysis as they offer a comprehensive overview of the main features of a dataset. By summarizing and presenting data in a clear and concise manner, descriptive statistics provide valuable insights into the distribution, central tendency, and variability of the data. This summary aids in understanding the overall structure of the dataset, identifying patterns, trends, and outliers. Through measures such as mean, median, mode, range, and standard deviation, we can gain a deeper understanding of the data. In essence, the summary of a dataset serves as the foundation for subsequent analysis, which allows us to draw meaningful conclusions and make data-driven decisions.

```
summary(HP_train)
```

```
    LotArea       OverallQual      YearBuilt     YearRemodAdd
 Min.   :  1491   Min.   : 1.000   Min.   :1880   Min.   :1950
 1st Qu.:  7585   1st Qu.: 5.000   1st Qu.:1954   1st Qu.:1968
 Median :  9442   Median : 6.000   Median :1973   Median :1994
 Mean   : 10795   Mean   : 6.136   Mean   :1971   Mean   :1985
 3rd Qu.: 11618   3rd Qu.: 7.000   3rd Qu.:2000   3rd Qu.:2004
 Max.   :215245   Max.   :10.000   Max.   :2010   Max.   :2010
   BsmtFinSF1       FullBath        HalfBath       BedroomAbvGr
 Min.   :   0.0   Min.   :0.000   Min.   :0.0000   Min.   :0.000
 1st Qu.:   0.0   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
 Median : 384.0   Median :2.000   Median :0.0000   Median :3.000
 Mean   : 446.5   Mean   :1.564   Mean   :0.3856   Mean   :2.843
 3rd Qu.: 728.8   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :2260.0   Max.   :3.000   Max.   :2.0000   Max.   :8.000
  TotRmsAbvGrd      Fireplaces       GarageArea         YrSold
 Min.   : 2.000   Min.   :0.0000   Min.   :   0.0   Min.   :2006
 1st Qu.: 5.000   1st Qu.:0.0000   1st Qu.: 336.0   1st Qu.:2007
 Median : 6.000   Median :1.0000   Median : 480.0   Median :2008
 Mean   : 6.482   Mean   :0.6278   Mean   : 472.6   Mean   :2008
 3rd Qu.: 7.000   3rd Qu.:1.0000   3rd Qu.: 576.0   3rd Qu.:2009
 Max.   :14.000   Max.   :3.0000   Max.   :1390.0   Max.   :2010
   SalePrice
 Min.   : 34900
 1st Qu.:130000
 Median :163000
 Mean   :183108
 3rd Qu.:216878
 Max.   :755000
```

```
summary(HP_test)
```

```
    LotArea      OverallQual    YearBuilt     YearRemodAdd    BsmtFinSF1
 Min.   : 1300   Min.   :2    Min.   :1890   Min.   :1950   Min.   :   0.0
 1st Qu.: 7493   1st Qu.:5    1st Qu.:1958   1st Qu.:1966   1st Qu.:   0.0
 Median : 9380   Median :6    Median :1976   Median :1994   Median : 407.5
 Mean   : 9713   Mean   :6    Mean   :1974   Mean   :1985   Mean   : 426.1
 3rd Qu.:11629   3rd Qu.:7    3rd Qu.:2002   3rd Qu.:2004   3rd Qu.: 687.0
 Max.   :27650   Max.   :9    Max.   :2009   Max.   :2010   Max.   :1646.0
    FullBath        HalfBath       BedroomAbvGr    TotRmsAbvGrd
 Min.   :0.000   Min.   :0.0000   Min.   :1.000   Min.   : 4.000
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.250   1st Qu.: 5.250
 Median :2.000   Median :0.0000   Median :3.000   Median : 6.000
 Mean   :1.578   Mean   :0.3778   Mean   :2.967   Mean   : 6.633
 3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.: 8.000
 Max.   :2.000   Max.   :2.0000   Max.   :5.000   Max.   :12.000
   Fireplaces       GarageArea        YrSold       SalePrice
 Min.   :0.0000   Min.   :  0.0   Min.   :2006   Min.   : 35311
 1st Qu.:0.0000   1st Qu.:388.5   1st Qu.:2007   1st Qu.:132475
 Median :0.0000   Median :491.0   Median :2008   Median :166250
 Mean   :0.4333   Mean   :475.4   Mean   :2008   Mean   :172587
 3rd Qu.:1.0000   3rd Qu.:604.8   3rd Qu.:2009   3rd Qu.:200725
 Max.   :2.0000   Max.   :871.0   Max.   :2010   Max.   :395192
```

3. Data Preparation:

Data preparation is an important phase in the data analysis process, involving the cleaning and organization of raw data to ensure its quality and suitability for further analysis. This step addresses issues such as missing values, outliers, and inconsistencies, enhancing the overall reliability of the dataset. In our case, we do not have any missing values.

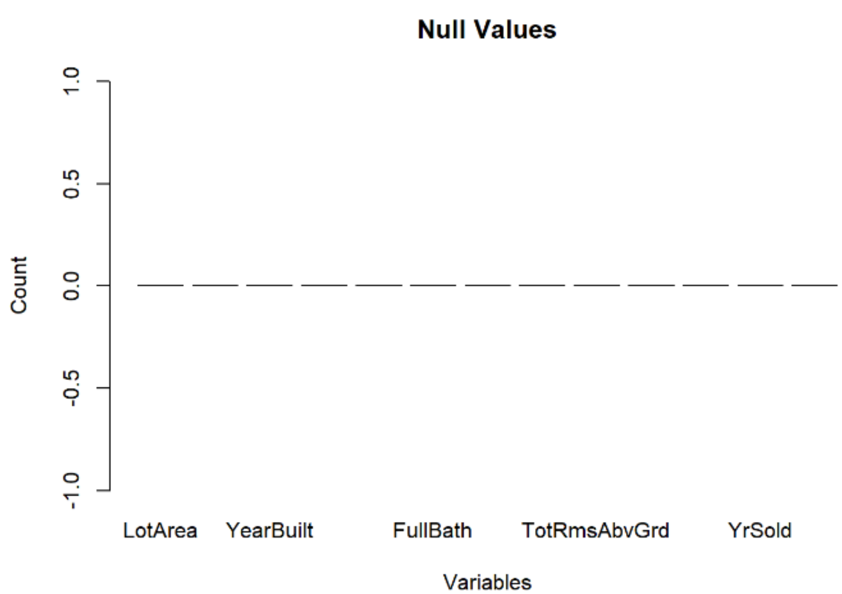
```
Missing_HP_train = colSums(is.na(HP_train))
Missing_HP_test = colSums(is.na(HP_test))
print(Missing_HP_train)
```

```
     LotArea  OverallQual    YearBuilt YearRemodAdd   BsmtFinSF1     FullBath 
           0            0            0            0            0            0 
    HalfBath BedroomAbvGr TotRmsAbvGrd   Fireplaces   GarageArea       YrSold 
           0            0            0            0            0            0 
   SalePrice 
           0 
```

```
print(Missing_HP_test)
```

```
     LotArea  OverallQual    YearBuilt YearRemodAdd   BsmtFinSF1     FullBath 
           0            0            0            0            0            0 
    HalfBath BedroomAbvGr TotRmsAbvGrd   Fireplaces   GarageArea       YrSold 
           0            0            0            0            0            0 
   SalePrice 
           0 
```

```
Plot_HP_train <- barplot(Missing_HP_train, main = "Null Values", xlab = "Variables", ylab = "Count")
Plot_HP_test <- barplot(Missing_HP_test, main = "Null Values", xlab = "Variables", ylab = "Count")
```

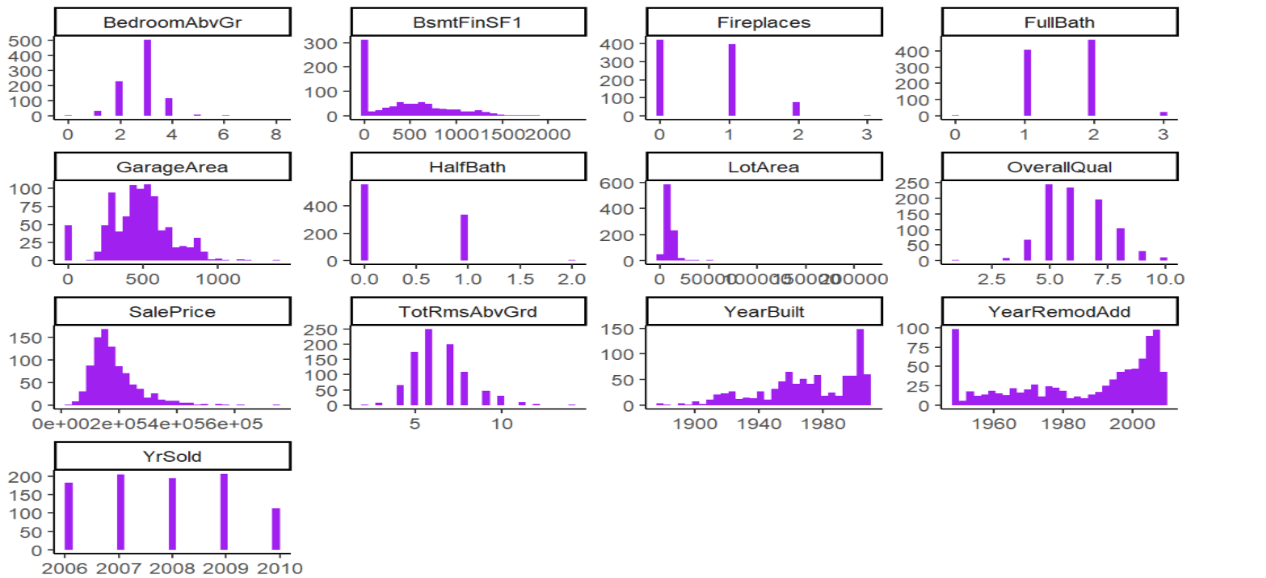


4. Data Exploration:

Data Exploration, encompassing both descriptive analysis and visualization, is another very important step in the data analysis process that aims to gain deeper understanding of the dataset. While descriptive analysis provides a statistical summary of the data, data visualization employs graphical representations to offer a more intuitive and comprehensive understanding. As we have already covered the descriptive analysis we will now move forward with the visualization.

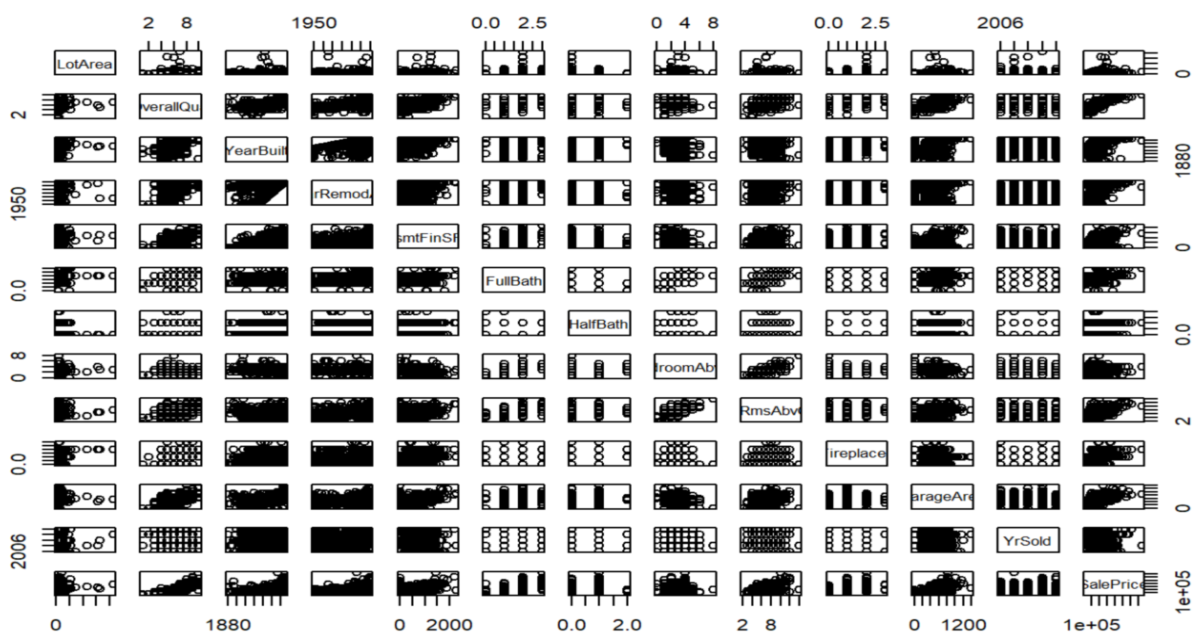
Data visualization is the process of presenting information graphically, utilizing charts, graphs, and visual elements to convey complex data patterns and trends in a concise and understandable manner.

(a) **Histogram:**



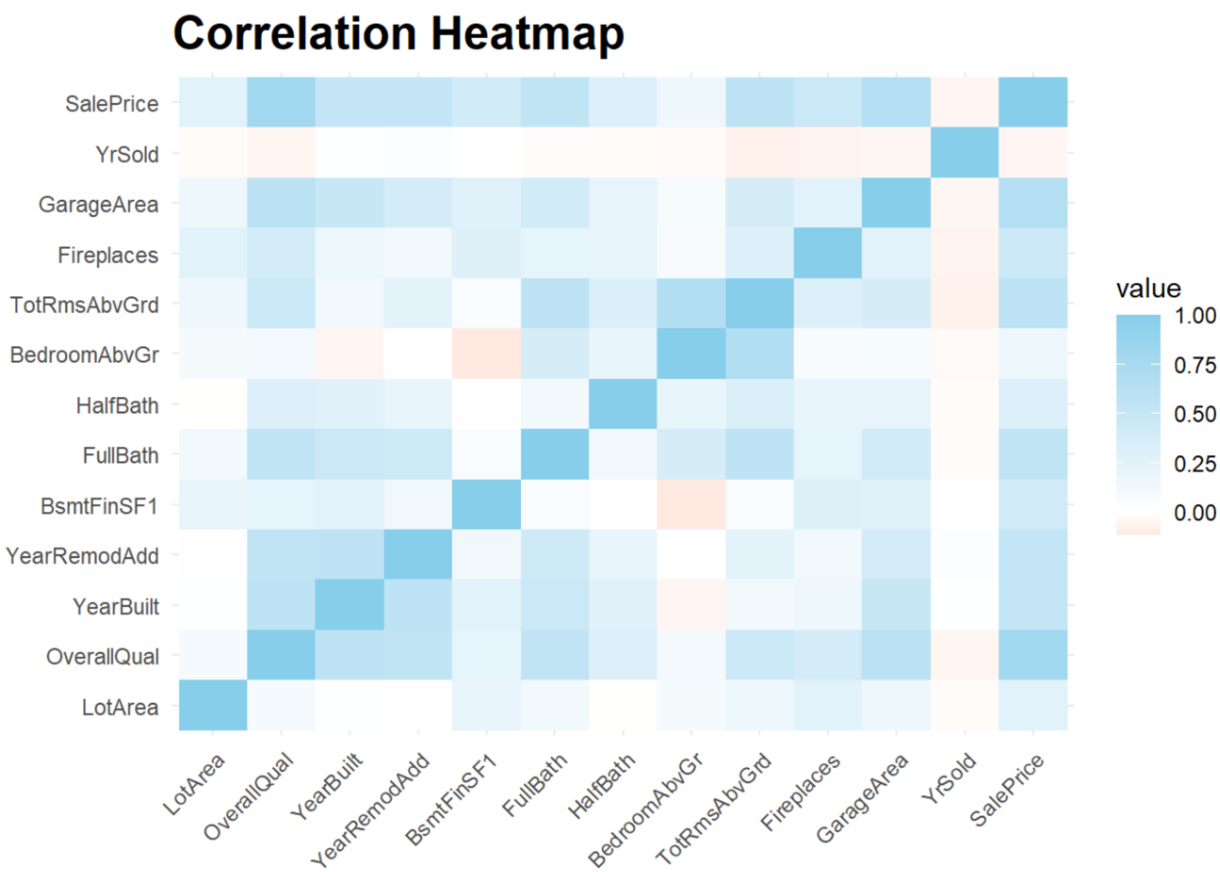
A histogram helps us to visualize the data distribution of each variable. For example, when examining SalePrice, a left-sided skewness can be observed which is evident from the longer tail on the left side.

(b) Pairs:



Pairs visually display the scatter plots for each pair of variables in a dataset, facilitating the examination of relationships and patterns. In the context of SalePrice as the response variable, commonalities among variables like LotArea, OverallQual, YearRemodAdd, BsmtFinSF1, BedroomAbvGr, TotRmsAbvGrd, Fireplaces, GarageArea can be observed.

(c) Correlation Heatmap:

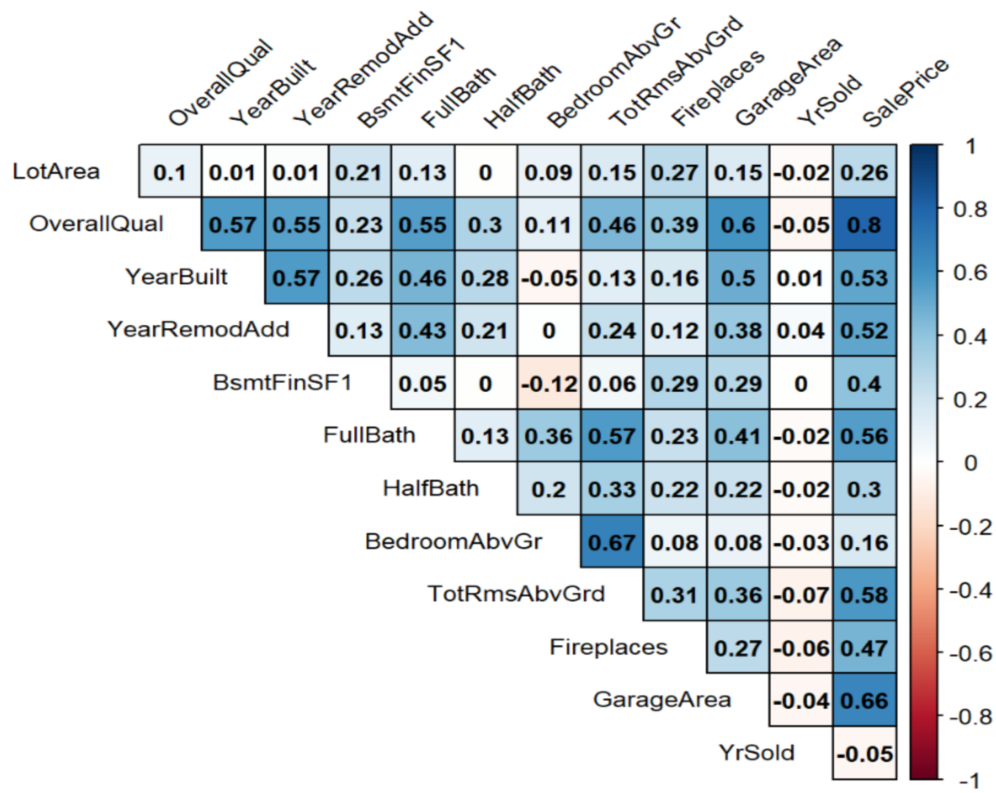


The above correlation plot is a statistical measure that quantifies the degree to which the output variable changes based on the independent variable. It indicates the direction and strength of the linear relationship between them. The correlation coefficient, typically ranging from -1 to 1, helps assess whether an increase in one variable corresponds to an increase, decrease, or no change in the response which is the SalePrice.

➢ Feature Selection:-

Feature selection is a critical aspect of the data preprocessing phase as it involves choosing the most relevant and impactful variables for the predictive model. The importance of feature selection is underscored by its role in identifying the subset of variables that significantly influence the target variable, SalePrice. Stepwise Regression is a feature selection technique that systematically adds or removes predictors from a model based on a chosen criterion, such as AIC or BIC. This model consists of a subset of predictors considered most relevant for explaining variation in the dependent variable. The application of ANOVA and linear regression models, along with the examination of the correlation matrix and heatmap also serve as a robust methodology for feature selection. By scrutinizing the P-values from these analyses, we can pinpoint the features— LotArea, OverallQual, YearRemodAdd, BsmtFinSF1, BedroomAbvGr, TotRmsAbvGrd, Fireplaces, GarageArea that exhibit a meaningful impact on predicting the SalePrice, ensuring a more focused and effective model.

(a) Correlation Plot showing the influence of different features on the output variable.



(b) Anova Analysis of the selected features.

```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq  F value    Pr(>F)
LotArea        1 4.2155e+11 4.2155e+11  317.273 < 2.2e-16 ***
OverallQual    1 3.6167e+12 3.6167e+12 2722.061 < 2.2e-16 ***
YearRemodAdd   1 7.6161e+10 7.6161e+10   57.322 9.240e-14 ***
BsmtFinSF1     1 2.2966e+11 2.2966e+11  172.850 < 2.2e-16 ***
BedroomAbvGr   1 6.2001e+10 6.2001e+10   46.664 1.560e-11 ***
TotRmsAbvGrd   1 3.1449e+11 3.1449e+11  236.697 < 2.2e-16 ***
Fireplaces     1 2.2765e+10 2.2765e+10   17.134 3.814e-05 ***
GarageArea     1 1.0419e+11 1.0419e+11   78.418 < 2.2e-16 ***
Residuals    891 1.1838e+12 1.3287e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

➢ Predictive Analysis:-

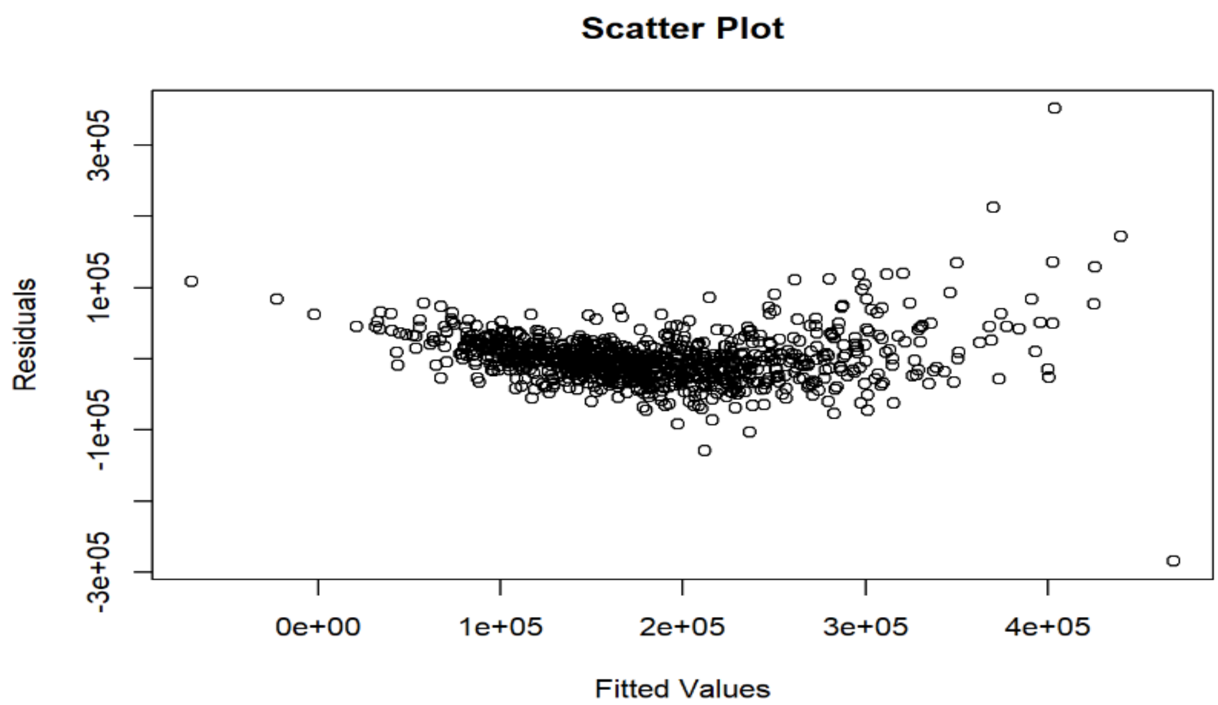
Predictive analysis employs statistical algorithms and machine learning to uncover patterns in historical data for making informed predictions about future outcomes. It is crucial for strategic decision-making and planning, providing valuable insights. We are applying various models, including Regression, Decision Tree, and Logistic Regression, to the House Prices dataset to enhance our understanding and anticipate trends.

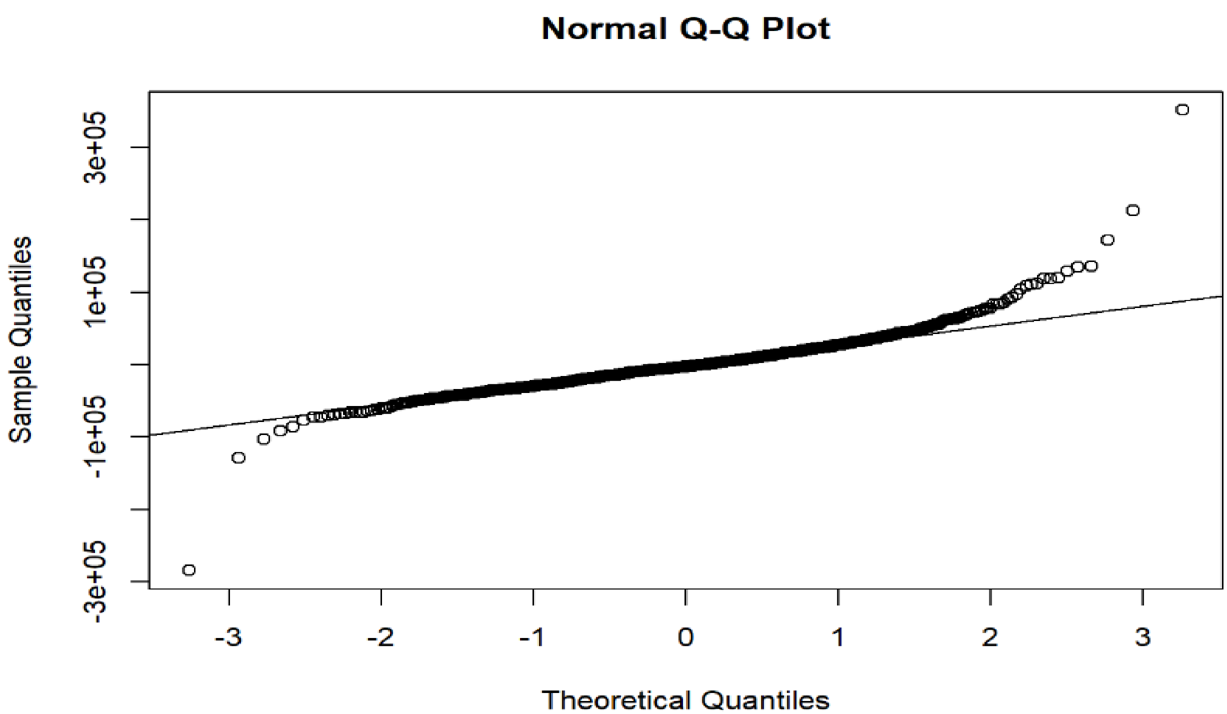
In the first step of Predictive Analysis, we build both Regression and Decision Tree Model to accurately predict the price of a house on various features.

1. **Regression Model:** Regression is a statistical method that analyzes the relationship between dependent and independent variables, allowing the prediction of future outcomes on the testing set based on training set.

```
        1         2         3         4         5         6         7         8 
 94075.22 171627.24 218101.94 228209.70 119889.17 107930.68 281644.41 173420.57 
        9        10        11        12        13        14        15        16 
127760.91 197511.00 203689.64 101766.04 118887.99 162365.27 147712.39  74898.47 
       17        18        19        20        21        22        23        24 
 12443.90 111624.78 236313.95 193429.05 194400.16 173068.94 165569.61 165876.39 
       25        26        27        28        29        30        31        32 
201928.85 153464.02 285380.26 231465.28 237098.55 224920.92 234725.31 116573.49 
       33        34        35        36        37        38        39        40 
297917.98 189673.43 260277.80  90211.01 209122.67 250771.08 240699.50 231851.38 
       41        42        43        44        45        46        47        48 
204809.92 240719.76 129990.40 155203.48 172136.02 188985.99 165583.95 331877.43 
       49        50        51        52        53        54        55        56 
211235.11 197374.80 144243.26 122160.86 141752.08 154565.51  99759.76 168596.96 
       57        58        59        60        61        62        63        64 
153341.00 131835.84 207807.49 195629.51 102836.49 273440.83 180786.26 277568.19 
       65        66        67        68        69        70        71        72 
237402.85 197048.72 152279.87 122471.51  34524.47 149231.83  41705.75 206487.05 
       73        74        75        76        77        78        79        80 
135318.72 187202.95 215295.16 197733.58  30961.02 210883.45  81079.38 118151.72 
       81        82        83        84        85        86        87        88 
221396.89 279005.21 194762.80 276020.46 147245.11 109376.67 135025.85 338196.61 
       89        90 
188099.52 192711.98 
```

- **Assumptions Check:** After testing the assumptions of the linear regression model, we can conclude that the scatter plot reveals that the relationship between the fitted values and residuals is not entirely random; there appears to be some pattern, indicating potential issues with the model. Additionally, the quantile-quantile plot shows deviations from the expected straight line, suggesting that the residuals might not follow a normal distribution. These observations indicate that the linear regression model may not fully meet the assumptions. As a result, we would further explore another model called decision tree for the same dataset.





2. **Decision Tree Model:** A decision tree model is a predictive algorithm that maps out potential outcomes based on a series of decision rules derived from the data. It simplifies complex decision-making processes, making it valuable for classification and regression tasks.

```
       1        2        3        4        5        6        7        8 
125471.0 125471.0 202038.8 202038.8 125471.0 125471.0 263933.7 202038.8 
       9       10       11       12       13       14       15       16 
125471.0 170146.5 125471.0 125471.0 125471.0 125471.0 170146.5 125471.0 
      17       18       19       20       21       22       23       24 
125471.0 125471.0 202038.8 170146.5 170146.5 125471.0 170146.5 170146.5 
      25       26       27       28       29       30       31       32 
170146.5 125471.0 263933.7 202038.8 263933.7 202038.8 263933.7 125471.0 
      33       34       35       36       37       38       39       40 
388831.3 202038.8 202038.8 125471.0 202038.8 202038.8 202038.8 202038.8 
      41       42       43       44       45       46       47       48 
170146.5 202038.8 125471.0 125471.0 170146.5 125471.0 125471.0 342542.1 
      49       50       51       52       53       54       55       56 
202038.8 170146.5 125471.0 125471.0 125471.0 125471.0 130751.8 170146.5 
      57       58       59       60       61       62       63       64 
170146.5 125471.0 202038.8 202038.8 125471.0 170146.5 170146.5 388831.3 
      65       66       67       68       69       70       71       72 
202038.8 170146.5 125471.0 125471.0 125471.0 130751.8 125471.0 202038.8 
      73       74       75       76       77       78       79       80 
130751.8 202038.8 202038.8 202038.8 125471.0 268469.5 125471.0 125471.0 
      81       82       83       84       85       86       87       88 
202038.8 263933.7 202038.8 263933.7 125471.0 125471.0 130751.8 388831.3 
      89       90 
170146.5 202038.8 
```

Comparing both Regression and Decision Tree Model:

Model	R-Squared Value	Adjusted R-Squared Value	RMSE Value
Linear Regression	0.8037	0.802	29382
Decision Tree	-0.633	-0.631	37430

Interpretation:- To determine the most appropriate model for the provided dataset, we assessed two specific models such as linear regression and decision tree. Our evaluation relied on key metrics like R-squared value, adjusted R-squared value, and RMSE value. A preferred model should exhibit a high adjusted R-squared value and a low RMSE value. Our analysis revealed that the decision tree model had a completely negative adjusted R-squared value and a higher RMSE compared to the linear regression model. Consequently, we can conclude that the decision tree model is not suitable for this dataset.

[**Note:** Adjusted R-squared is a modified version of R-squared that accounts for the number of predictors, R-squared measures the proportion of variance explained by the model, and RMSE (Root Mean Squared Error) quantifies the average prediction error in the model.]

In the second step, we use a Logistic Regression Model to divide the 'OverallQual' variable into two levels of classes "0" and "1" and make prediction using the categorical output.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 49  8
         1  6 27

               Accuracy : 0.8444
                 95% CI : (0.7528, 0.9123)
    No Information Rate : 0.6111
    P-Value [Acc > NIR] : 1.258e-06

                  Kappa : 0.6693

 Mcnemar's Test P-Value : 0.7893

            Sensitivity : 0.7714
            Specificity : 0.8909
         Pos Pred Value : 0.8182
         Neg Pred Value : 0.8596
             Prevalence : 0.3889
         Detection Rate : 0.3000
   Detection Prevalence : 0.3667
      Balanced Accuracy : 0.8312

       'Positive' Class : 1
```

Interpretation:- The logistic regression model, applied to the categorical variable (OverallQual), demonstrates its effectiveness, as evident from the confusion matrix. The accuracy of 84.44%, specificity of 89.09%, and high precision of 81.82% showcase the model's ability to distinguish between classes. Moreover, the model generalizes well, as indicated by its strong performance on the test set. Logistic regression proves to be a robust choice for handling categorical variables and making reliable predictions.

- **Insight/Conclusion:-** The primary objective of our project is to determine the most fitting model for the dataset. Initially, when the output variable was numerical (SalePrice), we applied two models: linear regression and decision tree. Despite the decision tree being unsuitable, the linear model, while not meeting all assumptions, remains comparatively favorable. In the subsequent step, as the output became categorical, logistic regression was performed, and it exhibited satisfactory performance metrics in the confusion matrix. This insight highlights the selection of models based on the nature of the output variable, with linear regression being preferable for numerical outputs and logistic regression for categorical ones.

*****