

**PAPER OUTLINE:**

[**PROJECT GOAL** 3](#_Toc153530078)

[**OVERVIEW OF DATA** 4](#_Toc153530079)

[**I ) DATA EXPLORATION ANALYSIS (DEA)** 4](#_Toc153530080)

[**II) DETAILS OF MODELING STRATEGY** 9](#_Toc153530081)

[**III) ESTIMATION OF MODEL PERFORMANCE** 13](#_Toc153530082)

[**Establishing Models** 13](#_Toc153530083)

[**IV) PREDICTION PERFORMANCE** 16](#_Toc153530084)

[**V) MODELS COMPARISON:** 21](#_Toc153530085)

[**VI) INSIGHTS AND CONCLUSION** 23](#_Toc153530086)

# **PROJECT GOAL**

The goal of this project is to create a model that accurately predicts the price of a house based upon the Zillow Zestimate home valuation system. This project will achieve the goal through formulating models through the methods of regression analysis, a decision tree, and classification. Through the analysis of the training data in R, this project is intended to provide accurate predictions of the test dataset. The results will help to solve the problem of Zillow’s Zestimate home valuations inaccurately predicting the price of houses, which is generally the largest purchase that an individual makes in their lifetime, and improve upon the margin of error through machine learning. This has implications for various stakeholders within the company of Zillow, including executives, as well as customers.

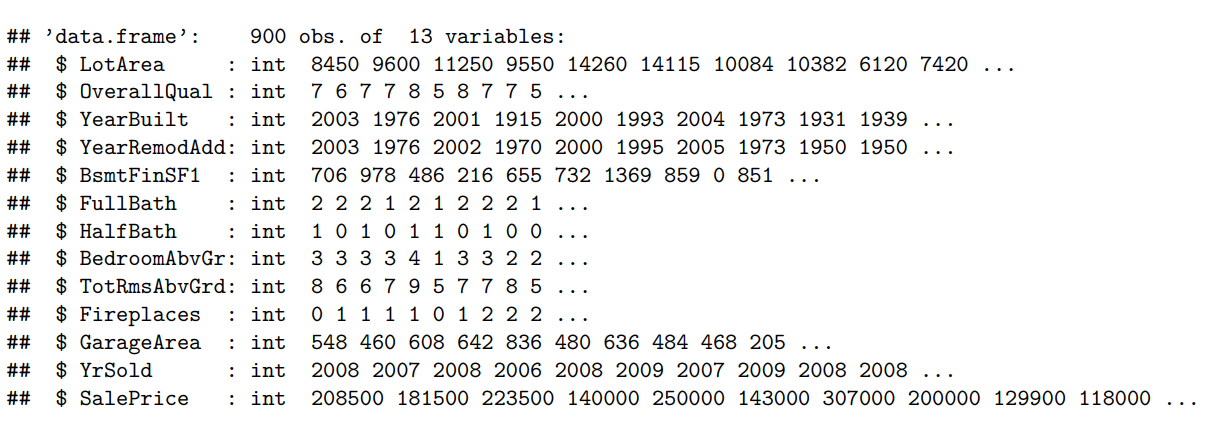
# **OVERVIEW OF DATA**

The dataset that is used in this project, titled House\_Prices, represents home values based upon Zillow’s Zestimate system. It represents a simplified version of the data that is used in the Zillow Prize competition, which is a competition that is used to help improve the accuracy of the Zestimate system, with the winning algorithm being awarded a prize of one-million dollars. According to the programming language software R, the dataset consists of 900 observations within 13 variables. These variables consist of lot size (LotArea), overall quality, (OverallQual), year built (YearBuilt), year remodeled (YearRemodAdd), finished basement square feet (BsmtFinSF1), full bathrooms (FullBath), half baths (HalfBath), bedrooms above ground (BedroomAbvGr), total rooms above ground (TotRmsAbvGrd), fireplaces (Fireplaces), size of garage (GarageArea), year sold (YrSold), and sale price (SalePrice). Each of these independent variables will be used to create a model that accurately predicts the dependent variable, sale price (SalePrice).

# **I ) DATA EXPLORATION ANALYSIS (DEA)**

The Data exploration analysis is divided into 2 parts, which are Descriptive Analysis and Visualization.

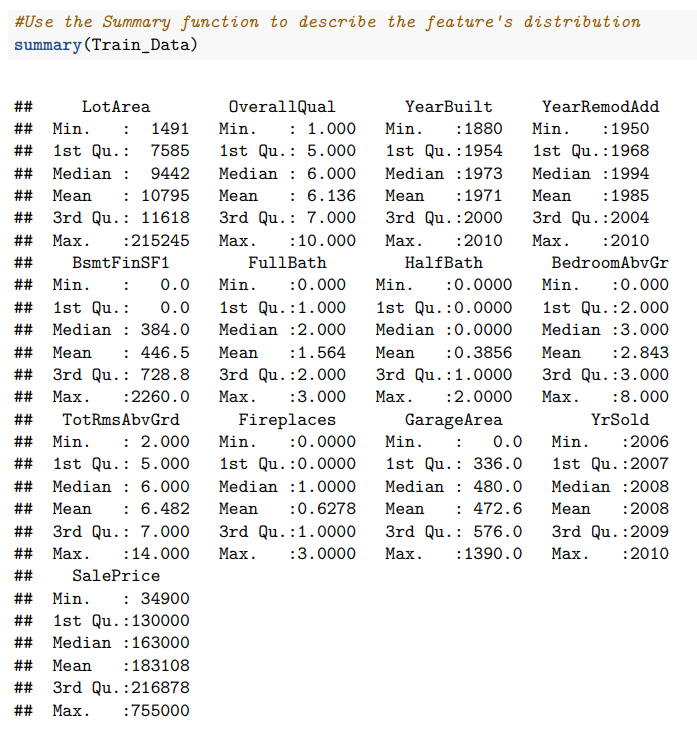
**Picture 1.1: Data Structure**



Firstly, in describing the dataset, we conducted a thorough analysis to ensure data completeness, revealing there are no missing values throughout the entire dataset. Among the 900 observations, we identified 13 variables characterized as integers, rendering them well-suited for various analytical methodologies such as Regression, Decision tree, and Classification approaches.

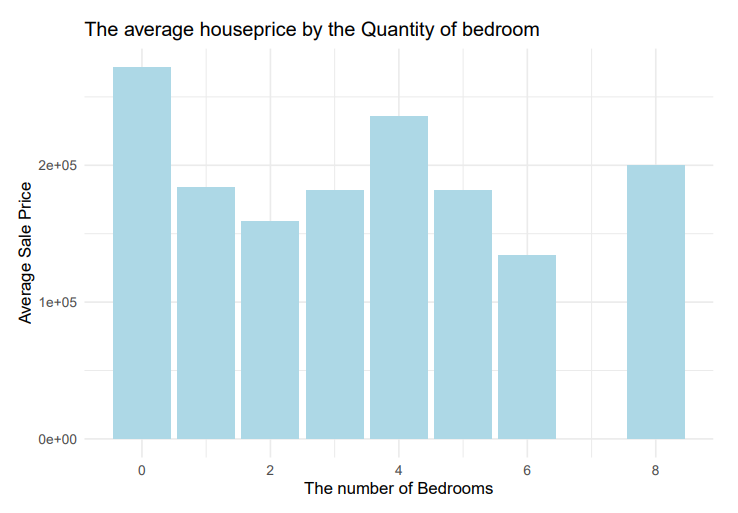
Recognizing the potential bias introduced by attributes with disparate ranges, we took a preventive measure. Leveraging the `preprocess()` function from the “Caret” package, with the 'Range' method, we crafted a model for Min-Max normalization of the House\_Price dataset variables. This normalization model was then applied to the SalesPrice data frame using the predict function, and the resultant values were stored in a new dataframe named Predict\_normalization. An examination of the summary statistics demonstrated the successful normalization of both variables, placing them within a standardized range of 0 to 1 for equitable comparisons.

**Picture 1.2: Data Structure**



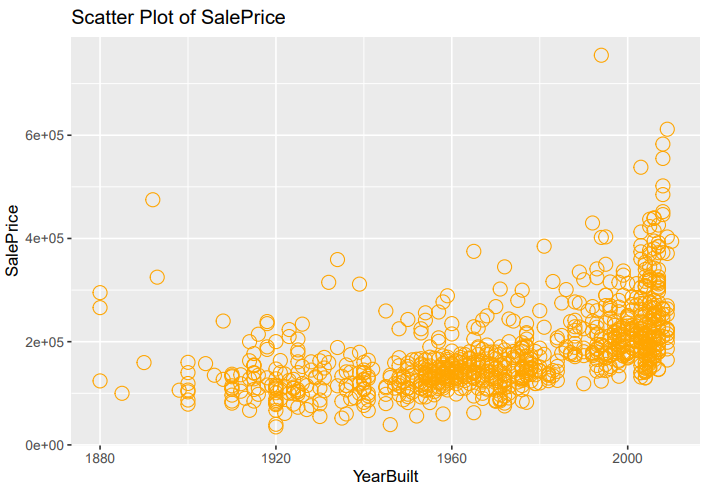
Given that all features are numerical, the summary function played a crucial role in providing key statistical metrics such as minimums, medians, means, maximums, and quartile values. This meticulous data exploration and normalization process lays the foundation for subsequent analyses, ensuring a nuanced understanding of the numerical attributes within the dataset.

**Picture 1.3: Graph 1**



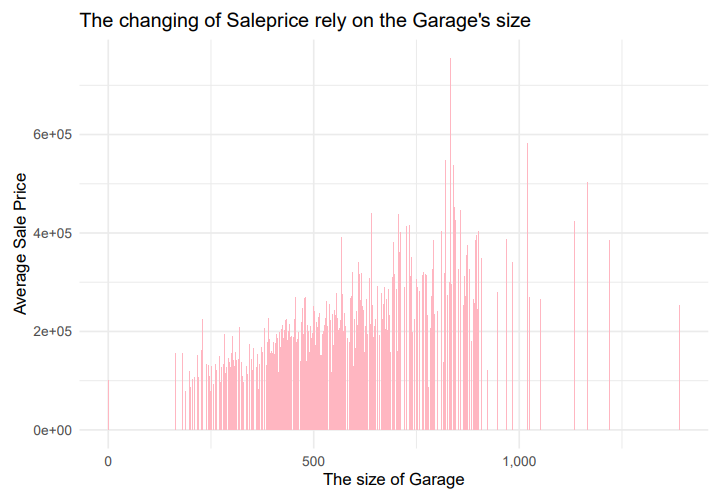
Secondly, in the visual analysis presented, Graph 1 illuminates the relationship between the number of bedrooms and the corresponding sale prices of houses. The graph categorizes bedrooms into six types above the ground, ranging from no bedroom to eight bedrooms. Notably, the average SalePrice analysis indicates that houses without bedrooms above the ground tend to command higher prices than their counterparts.

**Picture 1.4: Graph 2**



Moving on to Graph 2, it offers insights into the Distribution of SalePrice based on the year of house construction. A discernible trend reveals a higher tendency in the number of houses built since the 1960s. Although this trend persists, there is a significant increase in house prices starting from the 2000s.

**Picture 1.5: Graph 3**



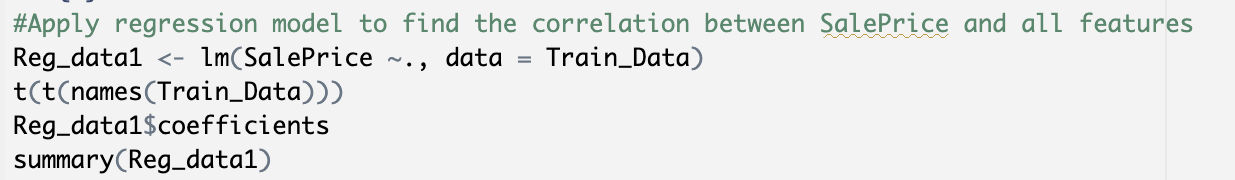
Graph 3 delves into the impact of Garage size on the average sale prices, unequivocally demonstrating that a larger garage size correlates with higher house prices. Specifically, when the garage falls within the 600 to 700 square ft range, the sale price peaks. These graphical representations contribute valuable insights into the relationships between key variables and average house prices within our analytical context.

# **II) DETAILS OF MODELING STRATEGY**

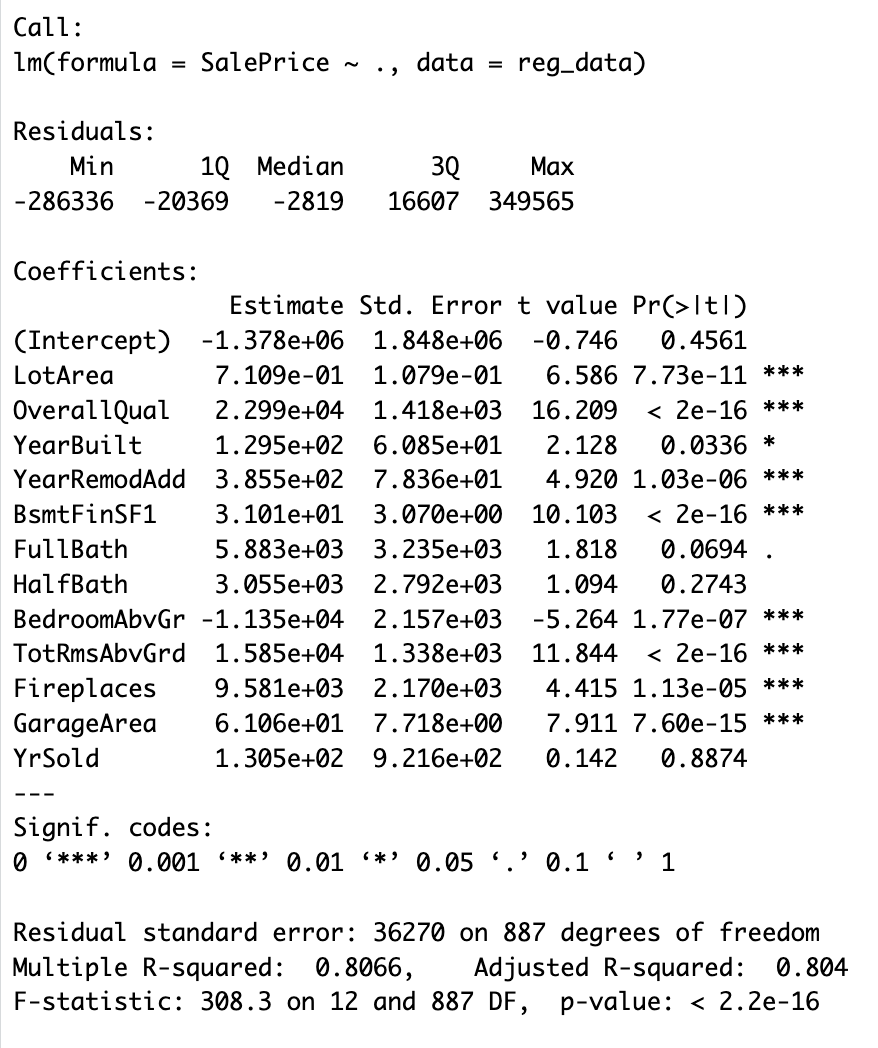
***Regression Model***

The regression model (Reg\_data1) was formulated with SalePrice as the dependent variable, and the rest of the features as independent variables, utilizing the ‘lm’ function and the ‘Train\_Data’ dataset to complete this. Prior to viewing a summary of the data, the names of each of the 13 columns of the ‘Train\_Data’ was analyzed, as well as their coefficients based upon the ‘Reg\_data1’. Once this regression model was summarized, it became clear that not all of the features were considered significant, which was shown by some variables having larger p-values than the rest. The features of FullBath, HalfBath, YrSold, and YearBuilt showed very little significance to the model. The rest of the variables, including LotArea, OverallQual, YearRemodAdd, BsmtFinSF1, BedroomAbvGr, TotRmsAbvGrd, Fireplaces, and GarageArea all showed very small p-values, thus indicating significance of the variable to the model. A summary of the model can be seen below in Picture 2.2.

**Picture 2.1. Regression model code**



**Picture 2.2: Summary of regression model**



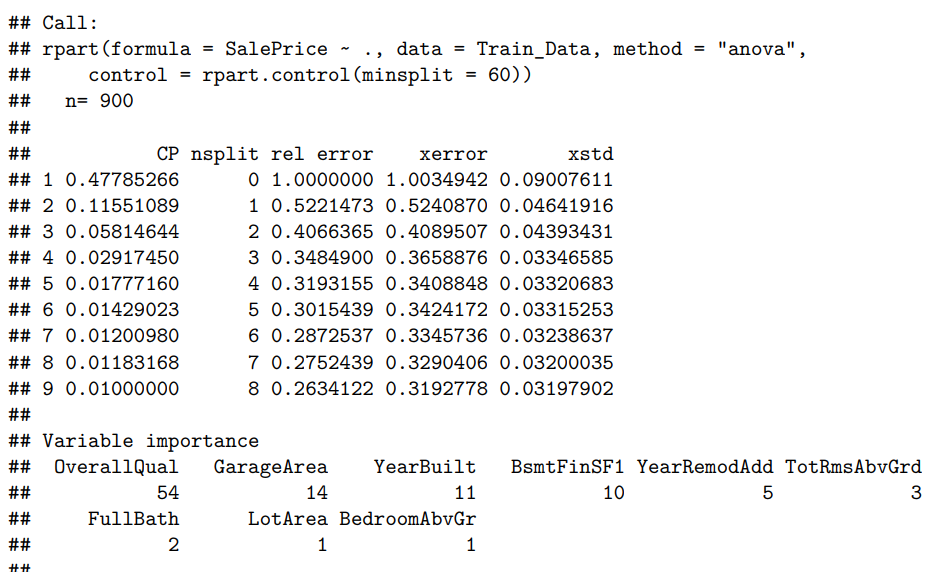
***Decision Tree***

**Picture 2.3: Using “Rpart” for predicting the SalePrice**



The modeling strategy for the Decision Tree is initiated with the construction of the model using the `rpart` function in R. The objective of the model is to predict the 'SalePrice' based on all available features in the dataset (`Train\_Data`). The method employed is 'anova', suggesting that the model is designed for predicting the Regression variable where our outcome is the House\_price, which is the continuous feature. The `minsplit` parameter is set to 60 through the `control` argument, indicating that a node in the tree must have a minimum of 60 observations before it is further split. This choice aims to control the complexity of the tree and prevent overfitting.

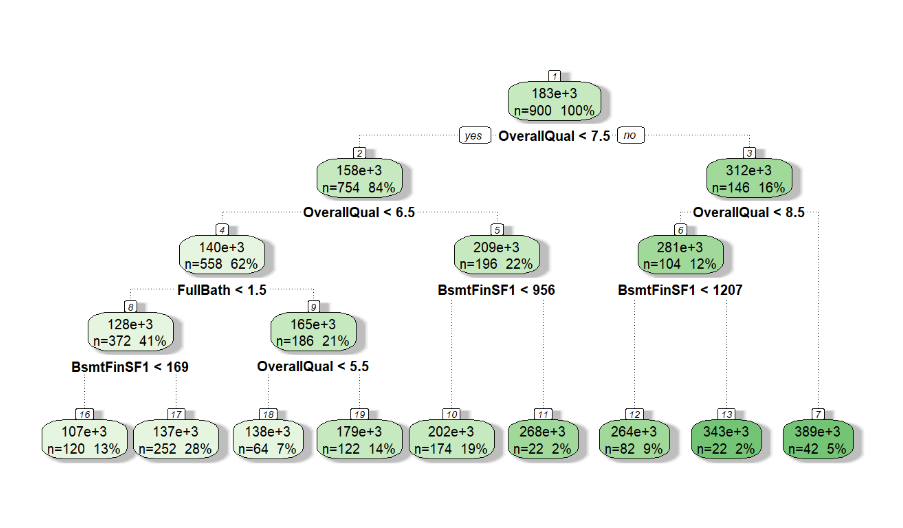
**Picture 2.4: Summary of Decision tree**



Following the model construction, a summary of the Decision Tree is generated using the `summary` function. This summary provides insights into the structure of the tree, including details on the splits, the significance of each variable, and the overall performance.

To enhance interpretability, the Decision Tree’s visualization is now plotted, and it shows different splits and branches based on our Decision nodes. From the top of the tree, OverallQual is the most significant feature, which becomes the root node. Recall that OverallQual is the numerical values, then the Decision Tree strategy is to divide the feature into 2 subsets. For instance, The OverallQual was divided into 2 subsets. If OverallQual < 7.5, we will have the first scenario in the left-hand-side, if OverallQual > 7.5, means the answer is No, then we will go to the branch in the right-hand-side.

**Picture 2.5: Decision tree visualization**



# **III) ESTIMATION OF MODEL PERFORMANCE**

## **Establishing Models**

***Regression***

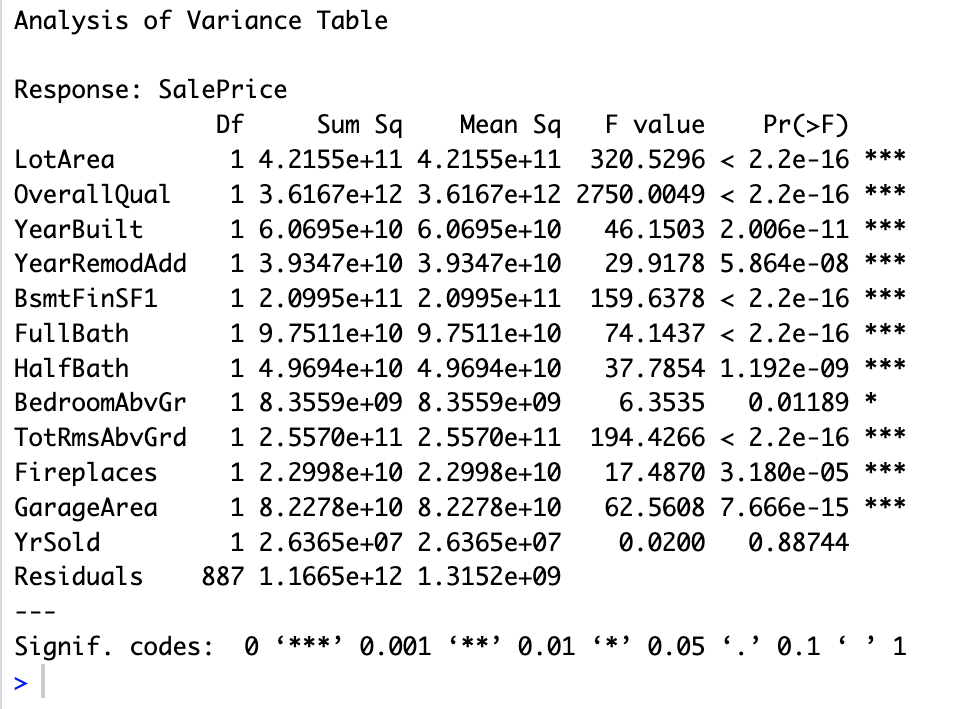
The regression model is expected to perform well and yield somewhat accurate predictions. This is due to the ‘test’ data consisting of an r-squared value of .8066, which shows that the features of the model account for explaining approximately 81% of the variability is accounted for by the features. Given this information, it is expected that the model is not completely accurate in explaining the house prices, however around 81% accurate.

According to the summary of the regression model, all p-values are very small, suggesting their significance. Despite this all coefficient values are positive except for the intercept and the bedrooms above ground (BedrromAbvGr) feature, which appear to have negative coefficient values. What this indicates is that the line of best fit for this regression model crosses the y-axis below the point of 0. In terms of the feature being negative, this suggests that as the sale price of a house increases, the number of bedrooms above ground decreases. This agrees with the findings in the data exploration analysis. Furthermore, it is clear that the r-squared and multiple r-squared values are very similar with values of .8066 and .804, respectively. This result suggests the goodness of fit of this regression model for the dataset.

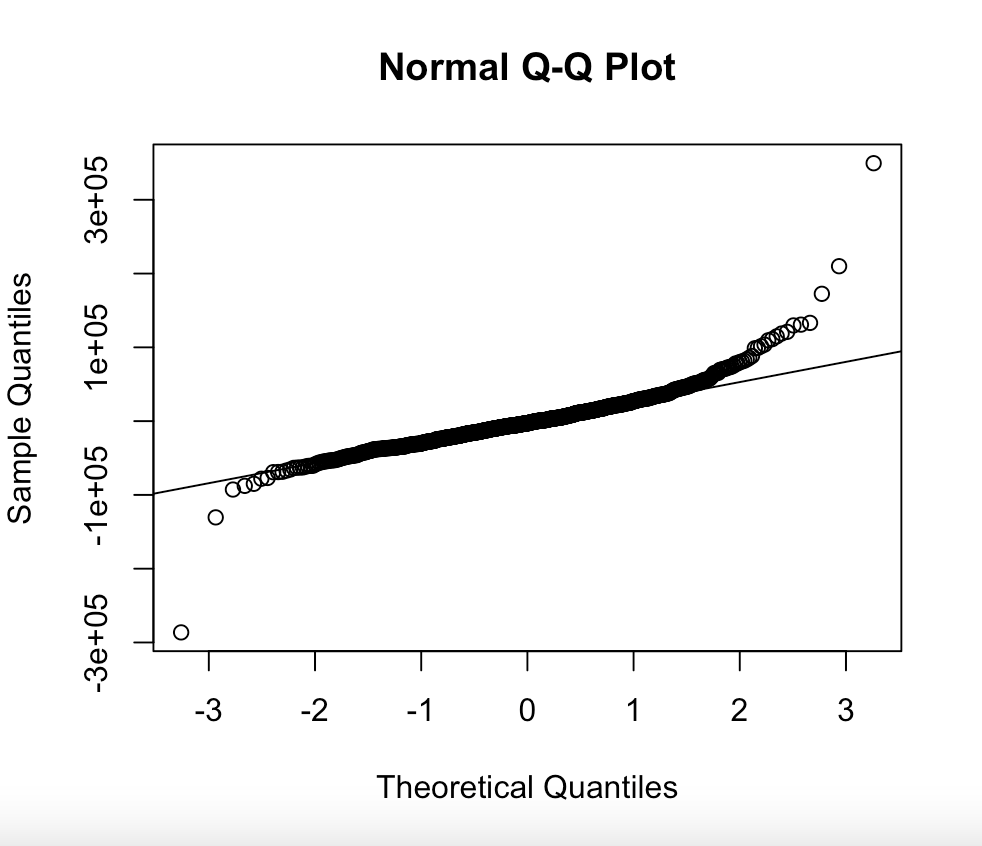
To continue, an ANOVA was conducted in order to review the variability of each feature and analyze each feature’s significance to the model, which can be seen below in Picture 3.1. Based upon the ANOVA, it is clear that the feature OverallQual is the most significant to the regression model based upon the Sum Square value as well as the F value. Other highly significant features in this model include LotArea, TotRmsAbvGrd, and BsmtFinSF1. Therefore, these variables are the driving factors that affect the sale price of a house in this regression model.

Lastly, the residuals were analyzed using a Quantile-Quantile, or qq, plot. This plot assesses the distribution of residuals and shows whether or not they are normally distributed, or symmetric. According to the qq plot of this regression model, it is clear that the residuals follow a symmetric distribution based upon their placement of the qq line, as can be seen below in picture 3.2. Given all of this information, it is clear that the regression model will likely be adequate in predicting the sale price of a house.

**Picture 3.1: ANOVA Table**



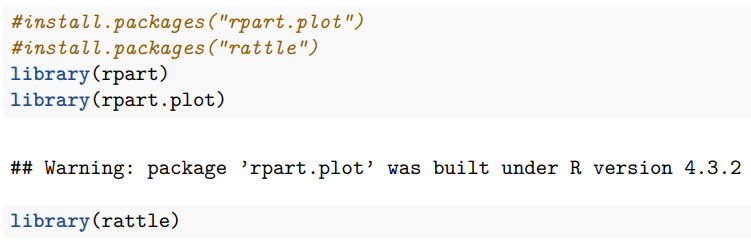
**Picture 3.2: Regression Q-Q Plot**



***Decision Tree***

At the beginning, picture 3.3 shows there are three libraries required to be installed and loaded.

**Picture 3.3: Required tools and libraries**



The Decision tree model started with using the “Rpart” function, which is also one of the most vital libraries because “Rpart” will be used for predicting the SalePrice (Dependent variable) with all other 13 variables (Independent variables). The “control” parameter in the “rpart” function can also be used as the hyperparameter, which specifies how many minimum observations in a node before it considers splitting (control=rpart.control(minsplit = 60)) that node further involves. Picture 3.4

**Picture 3.4: Using “Rpart” function**



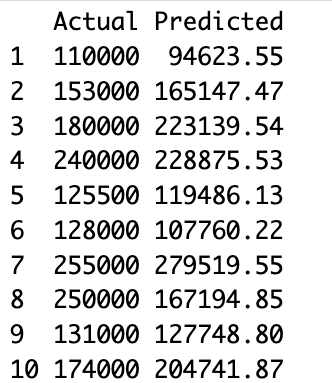
# **IV) PREDICTION PERFORMANCE**

***Regression Model***

To assess the accuracy of the regression model various methods were utilized including r-squared, root square mean error (RMSE), and the mean absolute error (MAE) utilizing the predict function on the ‘test’ dataset and comparing those values to the ‘Actual’ values. First, these values were compared side by side in a data frame (predict\_data). The regression model of the ‘predict’ data had an adjusted r-squared value of .8039, which is very similar to the initial adjusted r-squared model of the ‘test’ data, which was .804.

Additionally, the model showed a prediction of an RMSE of 28,930.40, and a MAE of 22,040.83. What this indicates is that the difference between the average actual and predicted values, measured by RMSE, is $28,930.40. As for MAE, or the average absolute error of actual and predicted values, is $22,040.83. While these numbers do not show complete accuracy in the regression model’s performance, these numbers are not off by a considerable amount. For example, the average house can cost hundreds of thousands of dollars. Thus, an RMSE and MAE of $28,930.40 and $22,040.83, respectively, shows considerable accuracy in predicting the dependent variable SalePrice. The data frame consisting of the actual and predicted data is shown below.

**Picture 4.1: Actual vs. predicted values**

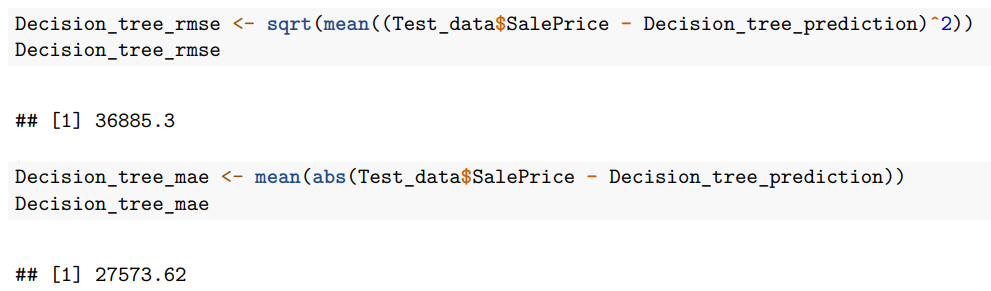
******

***Decision Tree Model***

Next step, we apply the “summary” function, giving the idea to identify important variables, and getting a sense of the model's overall performance. To further assess the predictive accuracy of the decision tree model, several metrics were employed (Picture 4.1). The root mean squared error (Decision\_tree\_rmse) was calculated, measuring the square root of the mean squared differences between the actual and predicted sale prices. The sum\_square error value shows as 36885.3, which means the predicted SalePrice is lower than the Actual one 36885.3 unit. Additionally, the mean absolute error (Decision\_tree\_mae) is another measure of the average magnitude of errors, but it considers the absolute values of the errors without squaring them.

These metrics offer diverse perspectives on model performance, considering aspects such as precision, accuracy, and overall predictive power. The analysis of these metrics will provide a comprehensive understanding of how well the decision tree model performs in predicting sale prices on the 'test' dataset.

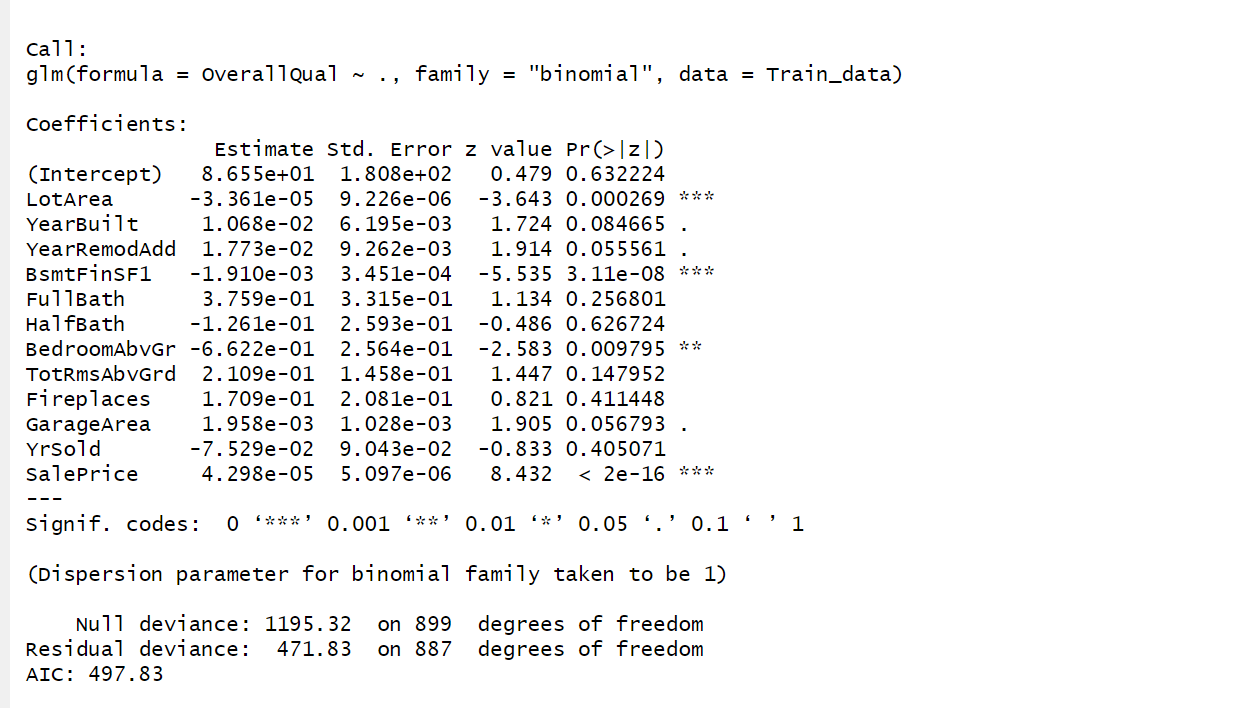
**Picture 4.2**: Summary function in Decision Tree



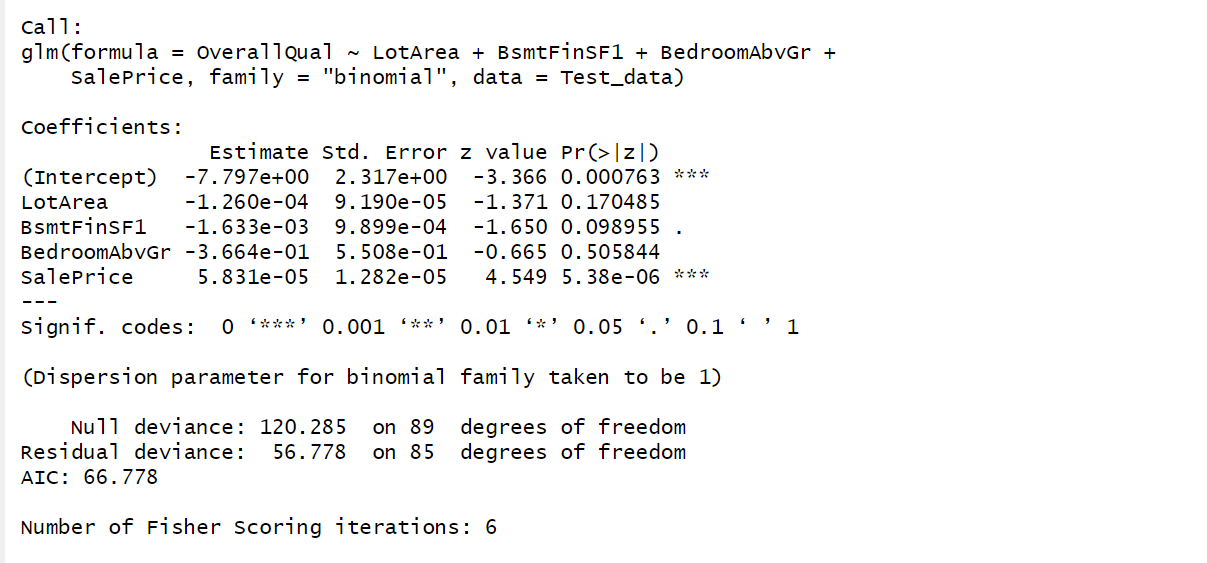
***Classification Model***

For feature selection, the model(Model\_5) is created and analyzed with logistic regression. By analyzing the logistic regression model\_5, I determined that LotArea, BsmtFinSF1, BedroomAbvGr and SalePrice are the features that have influence on the OverallQual Variable. Hence, new model (Model\_6) with features - LotArea, BsmtFinSF1, BedroomAbvGr and SalePrice is created and trained a logistic regression model (Model\_6). The prediction for categorical variable OverallQual is performed and at last, a confusion matrix is generated to evaluate the model's performance.

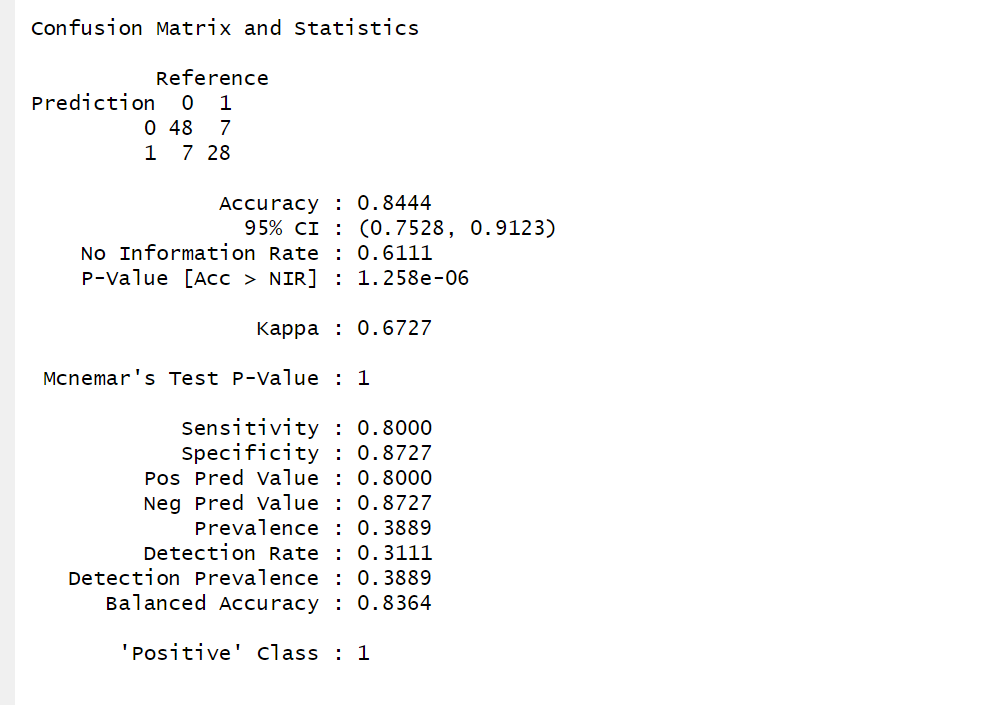
**Picture 4.3:** Create the Logistic model

******

**Picture 4.4**

******

**Picture 4.5**

******

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.

# **V) MODELS COMPARISON:**

In order to make well-informed forecasts about future events, predictive analysis uses statistical algorithms and machine learning to find patterns in previous data. It offers vital insights and is essential for strategic planning and decision-making. We are using a variety of models, adding Logistic Regression, Decision Tree, and Regression to the House Prices dataset to improve our comprehension and identify patterns.

To accurately estimate the price of a house based on multiple factors, we create both Regression and Decision Tree Models in the initial step of predictive analysis.

**1. Regression Model:** By analyzing the relationship between dependent and independent variables, regression is a statistical technique that makes it possible to predict future results on the testing set based on the training set.

**Assumptions Check:** We can conclude that the scatter plot shows that the relationship between the fitted values and residuals is not totally random after testing the assumptions of the linear regression model; there appears to be some pattern, suggesting potential problems with the model. Furthermore, deviations from the anticipated straight line are visible in the quantile-quantile plot, indicating that the residuals may not have a normal distribution.

These findings suggest that the assumptions made by the linear regression model might not be entirely met. Consequently, we would investigate a different model for the same dataset called a decision tree.

**2. Decision Tree Model:**

A decision tree model is a type of predictive algorithm that plots possible results according to a set of data-driven decision rules. It is useful for tasks involving regression and classification because it makes complicated decision-making processes simpler.

**Comparing between 3 Models:**

**Interpretation:-** We evaluated two distinct models, namely linear regression and decision tree, in order to ascertain which one would be best for the dataset that was provided. Important metrics such as RMSE value, adjusted R-squared value, and R-squared value were used in our analysis. A favored design ought to have a low RMSE value and a high adjusted R-squared value. In comparison to the linear regression model, the decision tree model had a higher RMSE and an adjusted R-squared value that was entirely negative, according to our analysis. As a result, we can say that this dataset is not a good fit for the decision tree model.

[**Note:** R-squared quantifies the percentage of variance explained by the model, adjusted R-squared takes the number of predictors into account, and RMSE (Root Mean Squared Error) measures the average prediction error in the model.]

In the second step, we split the "OverallQual" variable into two levels of classes, "0" and "1," using a logistic regression model, and we use the categorical output to make a prediction.

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.

# **VI) INSIGHTS AND CONCLUSION**

***Regression***

The regression model was modeled using the ‘lm’ that is utilized in linear regression models. ‘SalePrice’ was modeled as the dependent variable, with all other variables being used as predictor, or independent, variables of the ‘Train\_Data’ dataset. The column names of the dataset, as well as the coefficients, were analyzed.

According to the summary of the regression model (Reg\_data1), YearBuilt, FullBath, HalfBath, and YrSold have the largest p-values and thus show the least significance in describing the variance of the data. On the other hand, the features of LotArea, OverallQual, YearRemodAdd, BsmtFinSF1, BedroomAbcGr, TotRmsAbvGr, Fireplaces, and GarageArea, show the most significance thus showing the smallest p-values. This model consists of a multiple r-squared value of .8066 and an adjusted r-squared value of .804. The similarities in these values show the fitness of the model for this dataset. Additionally, this summary shows that the intercept and the feature BedroomAbvGr are both negative. This suggests that the feature has an inverse relationship with the dependent variable.

Furtherformore, an ANOVA was conducted to assess the variable and significance of various features in the regression model. The OverallQual feature was identified as the most significant feature, followed by LotArea, TotRmsAbvGr, and BsmtFinSF1. These variables are considered key drivers affecting the SalePrice. Additionally, a QQ plot was used to analyze residuals, confirming their symmetric distribution and suggesting that the regression model is likely to be effective in predicting house prices.

The accuracy of a regression model was assessed using various metrics such as adjusted r-squared, the root mean square error (RMSE), and mean absolute error (MAE). The model’s adjusted r-squared for the predicted data closely resembled that of the initial ‘test’ data. The model’s RMSE was found to be $28,930.40, and MAE was $22, 040.83, indicating a relatively small difference between the average actual and predicted values. While these values don’t demonstrate perfect accuracy, they are considered satisfactory given the substantial range of hour prices.

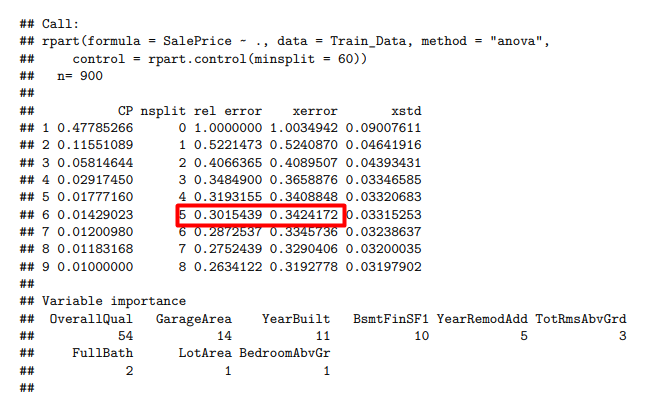
***Decision Tree***

The decision tree model was constructed using the “Rpart" function with the following parameters: SalePrice was modeled as a function of all other variables in the Train\_Data dataset, using the ANOVA method, and employing control parameters with a minimum split requirement of 60. The dataset comprised 900 observations.

The complexity parameter (CP) was used to control the size of the tree. The tree was initially built with a CP of 0.47785266, resulting in a fully grown tree with 0.4778527 as the complexity parameter for the root node. Subsequently, the tree was pruned by selecting optimal CP values. The pruning process involved the creation of nodes, each representing a split based on a specific predictor variable. The CP values at each split indicated the improvement in model fit, and smaller values were preferred.

Generally speaking, in the pruning process, the higher the complexity parameter (CP), results in a smaller number of splits. Then, when we get the CP value = 0.01, there are 17 root nodes that are splitted. The process’s evaluation is also based on relative error and the xerror. So, the rel error represents the relative error on the training data set. Whereas the x error is the relative error on the cross validation data set. Moreover, the pruning’s objective is to find what is the optimal tree size or basically we will find where is the point that the relative error starts to decrease, but the xerror or the cross-validation error reaches the error on the validation data set start to increase, which means that from this point on, the model has started to overfit at the point nsplit = 5 (Picture 6.1).

**Picture 6.1: Pruning process**



The variable importance section indicates the contribution of each predictor variable to the overall model performance. The higher the value, the more important the variable. Notably, OverallQual, GarageArea, YearBuilt, and BsmtFinSF1 were identified as the most influential predictors.

The tree structure consists of nodes, each associated with a specific subset of observations. For instance, Node 1 represents the entire dataset, and subsequent nodes (e.g., Nodes 2 and 3) represent subsets resulting from binary splits. The splits are based on predictor variables such as OverallQual, YearBuilt, GarageArea, FullBath, and YearRemodAdd.

The mean and mean squared error (MSE) for each node provide insight into the central tendency and dispersion of the observations within that node. Pruned subtrees, represented by nodes with smaller complexity parameters, are crucial for achieving a balance between model complexity and predictive accuracy.

In summary, the decision tree model captures the relationships between predictor variables and the target variable (SalePrice) by recursively partitioning the data into homogenous subsets. The pruning process ensures that the model generalizes well to new, unseen data, and the variable importance highlights the key features influencing the model's predictive performance.