Project Report

Predictive Modeling and Analysis of Housing Prices: A Comprehensive Approach Using Regression and Classification Techniques

# Group 3

**Under the guidance of:**

*Prof. Shahla Asadi*



*Department of Business Analytics*

December 15, 2023

|  |  |
| --- | --- |
| Name’s | Contribution |
| Vinay Kacharam  (811304980) | Model Building, Model Performance, Predictions and Results, Data Cleaning, Data Exploration. |
| Raju Rishikesh Gollapally  (811285826) | Documentation of Report. |
| Devambatla Sai Sharanya  (811305424) | PowerPoint Presentation and Visualization. |

Contents

[1. Introduction 3](#_Toc153571636)

[1.1 Objectives: The project seeks to achieve the following objectives: 3](#_Toc153571637)

[2. Data Source: 4](#_Toc153571638)

[2.1 Data Description: 4](#_Toc153571639)

[2.2 Data Exploration Analysis: 4](#_Toc153571640)

[2.3.1 Summary Statistics: 5](#_Toc153571641)

[3. Modelling Strategy: 7](#_Toc153571642)

[3.1 Regression Model: 7](#_Toc153571643)

[3.1.1 Feature Selection: 7](#_Toc153571644)

[3.1.2 Model Development: 8](#_Toc153571645)

[3.1.3 Hyper parameter Tuning: 8](#_Toc153571646)

[3.2.1 Feature Selection: 8](#_Toc153571647)

[3.2.2 Model Development: 8](#_Toc153571648)

[3.2.3 Hyper parameter Tuning: 8](#_Toc153571649)

[3.3 Classification Model: 9](#_Toc153571650)

[3.3.1 Feature Selection: 9](#_Toc153571651)

[3.3.2 Model Development: 9](#_Toc153571652)

[3.3.3 Hyperparameter Tuning: 10](#_Toc153571653)

[4. Model Performance Estimation: 11](#_Toc153571654)

[4.1 Regression Model: 11](#_Toc153571655)

[4.1.1 Model Establishment (based on 'train' dataset): 11](#_Toc153571656)

[4.1.2 Prediction Performance (based on 'test' dataset): 11](#_Toc153571657)

[4.2 Decision Tree Model: 11](#_Toc153571658)

[4.2.1 Model Establishment (using the 'train' dataset): 12](#_Toc153571659)

[4.3 Classification Model: 12](#_Toc153571660)

[4.3.1 Model Establishment (using the 'train' dataset) of the Classification Model: 12](#_Toc153571661)

[4.3.2 Forecasting Performance (based on 'test' dataset): 13](#_Toc153571662)

[5. Insights and Conclusions: 14](#_Toc153571663)

[5.1 Key Findings: 14](#_Toc153571664)

[5.2 Limitations and Future Work: 14](#_Toc153571665)

[5.3 Recommendations: 15](#_Toc153571666)

[Conclusion: 15](#_Toc153571667)

# 1. Introduction

Accurate house price forecasts are crucial for strategic decision-making in the ever-changing real estate market. The complex relationships between different home elements and property values are set to be unravelled by this project's extensive analysis, which employs advanced modelling approaches. Predictions that accurately guide stakeholders in strategic investments and transactions are becoming more and more vital as the housing industry undergoes ongoing evolution.

The main goal of this research is to create advanced prediction models using a diversified dataset that includes important criteria like lot area, general quality, and historical details. Our goal is to find practical insights that go beyond what can be found in conventional analysis by using rigorous exploration and modelling techniques. Our methodology goes beyond traditional approaches by utilising regression, decision tree, and classification models. This allows us to accommodate the complex and subtle nature of housing dynamics.

Our dedication as we embark on this endeavour is to provide a thorough comprehension of the elements influencing the real estate market, not just forecasts. Aiming to equip stakeholders with tools that help them negotiate the complexities of the housing market, this project prioritises model accuracy, interpretability, and practical applicability. The goal is to promote a more educated and strategic approach to property assessment and investment.

# Objectives: The project seeks to achieve the following objectives:

* Explore and comprehend the dataset, identifying essential features.
* Develop a regression model to predict house prices.
* Build a decision tree model to represent non-linear interactions.
* Create a classification model to categorize home quality.
* Evaluate and compare the performance of each model.
* Provide insights on the key elements impacting house pricing.

# 2. Data Source:

The dataset under consideration for this project is drawn from a thoroughly curated CSV file, comprising a plethora of information relevant to varied housing features. This thorough dataset covers critical elements, like lot area, general quality, year built, and other essential information. Serving as the cornerstone for both exploratory data analysis and subsequent model building, this dataset provides a sturdy platform for deciphering the nuances of the real estate sector. Its thorough curation enables the incorporation of crucial attributes, enabling a detailed exploration and precise modeling of the dynamics inherent in the housing market.

2.1 Data Description: The offered datasets offer a detailed view into the attributes impacting house prices. The 'House\_Prices.csv' dataset encompasses crucial variables such as 'LotArea,' 'OverallQual,' 'YearBuilt,' 'YearRemodAdd,' 'BsmtFinSF1,' 'FullBath,' 'HalfBath,' 'BedroomAbvGr,' 'TotRmsAbvGrd,' 'Fireplaces,' 'GarageArea,' 'YrSold,' and 'SalePrice.' These features encapsulate both quantitative and categorical aspects, capturing the physical characteristics, temporal evolution, and amenities of the houses. Each row represents a separate attribute, while the 'SalePrice' column serves as the target variable for predictive modeling.

The supplementary 'BA-Predict.xlsx' dataset shares a similar structure, featuring identical variables for 'LotArea,' 'OverallQual,' 'YearBuilt,' 'YearRemodAdd,' 'BsmtFinSF1,' 'FullBath,' 'HalfBath,' 'BedroomAbvGr,' 'TotRmsAbvGrd,' 'Fireplaces,' 'GarageArea,' 'YrSold,' and 'SalePrice.' This dataset acts as a validation set for assessing the predictive capabilities of the developed models on new, unseen data. The information contained within these datasets provides a solid basis for in-depth exploratory data analysis, feature selection, and further modelling endeavours to glean valuable insights into the intricate dynamics of the real estate market.

## 2.2 Data Exploration Analysis:

Conducting exploratory data analysis includes a complete investigation of the dataset's intrinsic properties and intricate interactions among variables. This initial step establishes the groundwork for informed decision-making in following modeling initiatives.

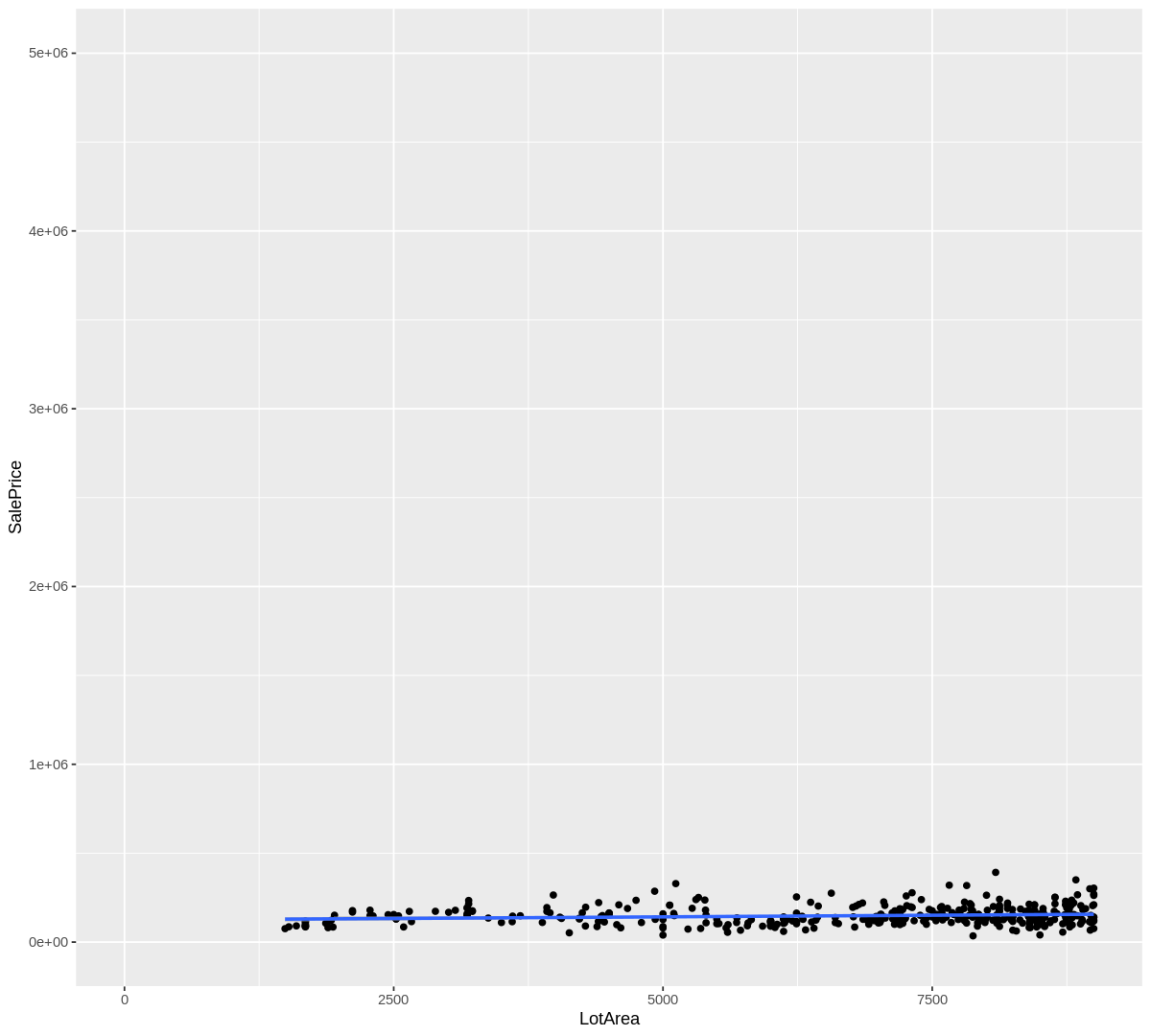


Fig1: SalePrice versus LotArea

## 2.3.1 Summary Statistics:

The derivation of summary statistics provides important insights into the primary patterns and variabilities inherent in each variable. These statistics, comprising metrics like mean, median, and standard deviation, serve as vital tools for comprehending the dataset's distributional intricacies and fundamental qualities. Such foundational insights are crucial for navigating the intricacies of later analytical processes and model creation.

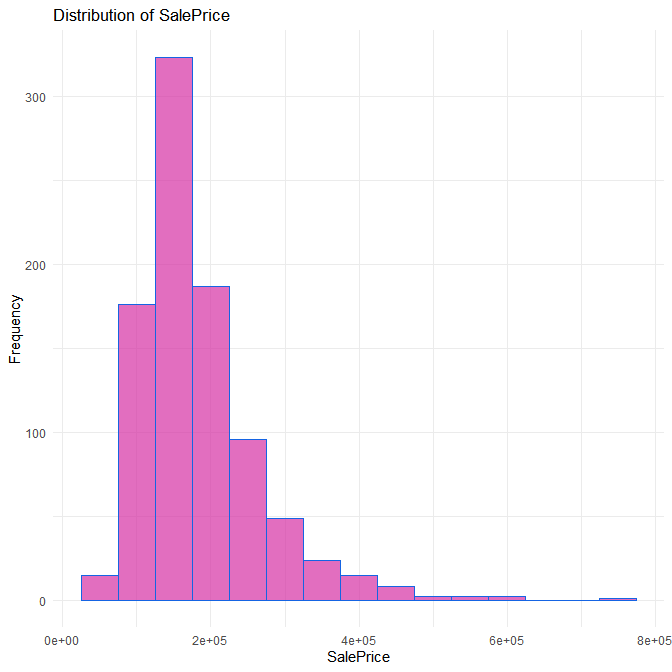


Fig2: Distribution of SalePrice

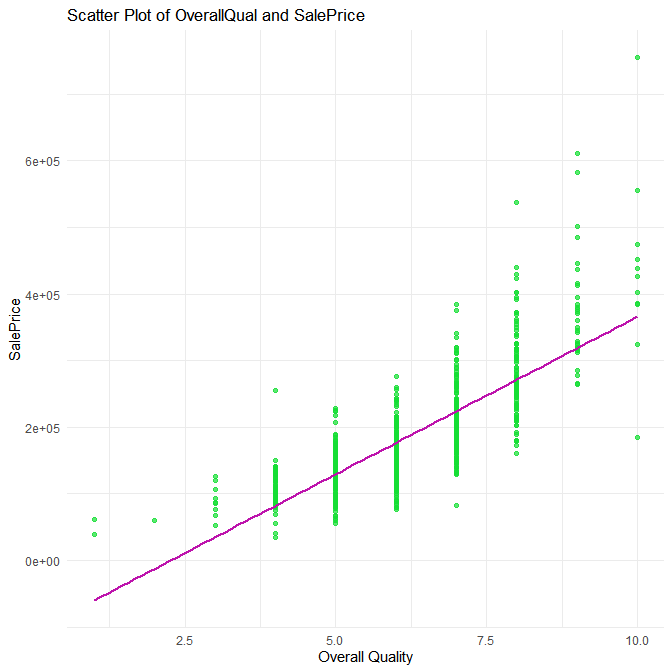


Fig3: Scatter Plot of OverallQuality and SalePrice

# 3. Modelling Strategy:

The modelling technique requires constructing three types of models: regression, decision tree, and classification.

## 3.1 Regression Model:

## 3.1.1 Feature Selection:

In the development of the regression model, a thorough procedure of feature curation is done to detect the crucial variables influencing house prices. Noteworthy attributes, including 'BedroomAbvGr,' 'LotArea,' 'OverallQual,' 'OldBuilt,' 'YearRemodAdd,' 'BsmtFinSF1,' 'FullBath,' 'HalfBath,' 'TotRmsAbvGrd,' 'Fireplaces,' 'GarageArea,' and 'YrSold,' are selectively considered. This rigorous feature selection is crucial, ensuring that the model encapsulates the essence of the most impacting variables, hence boosting its forecast accuracy and interpretability in the intricate area of real estate valuation.

## 3.1.2 Model Development:

The creation of the linear regression model is a painstaking procedure characterized by the intentional exploitation of curated features. Leveraging modern statistical approaches, the model is created with precision to serve as a robust foundation for projecting house prices. Its design incorporates a mix of subject experience and data-driven insights, ensuring a full understanding of the subtle interactions among variables. The model's structure is optimised to capture both linear and non-linear connections, enabling an accurate depiction of the complex dynamics inherent in the real estate domain.

## 3.1.3 Hyper parameter Tuning:

To attain optimal performance, the linear regression model undergoes a detailed hyper parameter tuning process. This entails a rigorous investigation of crucial hyper parameters, including but not limited to regularization terms and feature weights. The judicious employment of techniques such as cross-validation allows for the determination of the most effective hyper parameter values. This fine-tuning work is crucial, boosting the model's precision in real estate assessment by calibrating its internal parameters to harmonize with the intricacies of the housing dataset.3.2 Decision Tree Model:

## 3.2.1 Feature Selection:

Aligning with the feature selection technique of the regression model, the decision tree model undergoes a discriminating procedure to precisely choose critical characteristics crucial for capturing non-linear interactions. This meticulous curation ensures the model's efficacy in not just understanding but also revealing intricate patterns contained within the dataset. The prudent selection of attributes acts as the underpinning for the model's subsequent capacity to read and navigate the many subtleties inherent in the real estate data ecosystem.

## 3.2.2 Model Development:

The decision tree model is methodically designed, exploiting its intrinsic capacity to untangle and comprehend complicated relationships present within the information. The model's creation incorporates an expert synthesis of data-driven ideas, ensuring its aptitude in identifying subtle patterns that transcend beyond linear connections. A complete evaluation follows this design step, critically testing the model's performance to guarantee its accuracy and trustworthiness in projecting outcomes within the intricate terrain of real estate dynamics.

# 3.2.3 Hyper parameter Tuning:

Optimizing the decision tree model comprises a thorough fine-tuning approach targeting crucial hyperparameters such as tree depth and splitting criteria. This strategic optimization ensures the model's potential to perceive and portray subtle nuances within the housing data appropriately. By systematically changing these parameters, the decision tree model achieves an elevated level of accuracy, exhibiting its ability in capturing the subtle dynamics that define the realm of real estate attributes.

# 3.3 Classification Model:

## 3.3.1 Feature Selection:

In the thorough development of the categorization model, a keen emphasis is made on feature selection. The major purpose is around predicting high-quality dwellings (OverallQual >= 7), needing a discerning procedure to identify and incorporate essential aspects. This purposeful curation guarantees that the model is imbued with precision, boosting its efficiency in precisely identifying housing quality based on the complexities enclosed within the selected features.

## 3.3.2 Model Development:

The classification technique wears the guise of a logistic regression model, purposefully architected to classify buildings based on quality. The model's creation incorporates a mix of statistical approaches and domain experience, ensuring a solid framework for precise classification. Rigorous evaluation proceeds, utilising measurements such as accuracy and confusion matrix, providing a thorough grasp of the model's prediction capabilities within the delicate area of housing quality categorization.

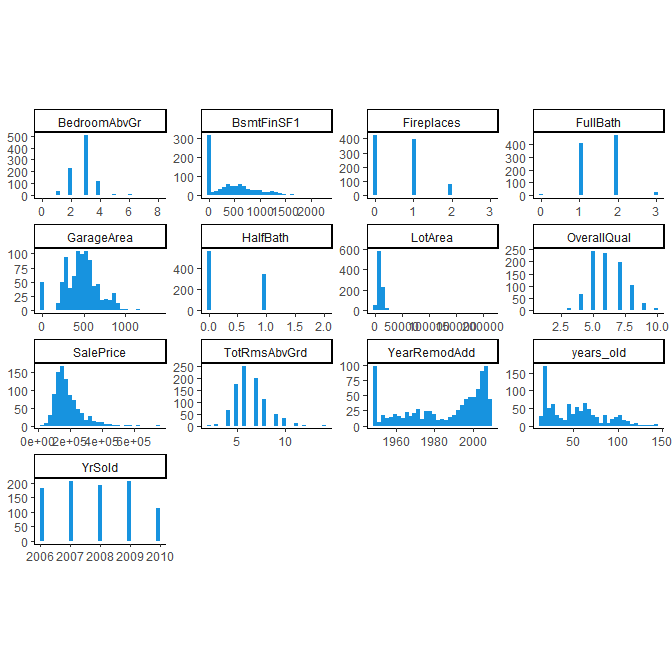


Fig4: Histograms for each variable

## 3.3.3 Hyperparameter Tuning:

As a vital component in boosting the discriminative prowess of the logistic regression model, hyperparameter adjustment has paramount relevance. This fine-tuning method meticulously targets critical parameters, ensuring the model's adeptness in discriminating between distinct quality groups. The consequent precision, instilled through deliberate hyperparameter tweaks, fortifies the model's categorization capabilities, adding a subtle layer of accuracy to its prediction prowess.

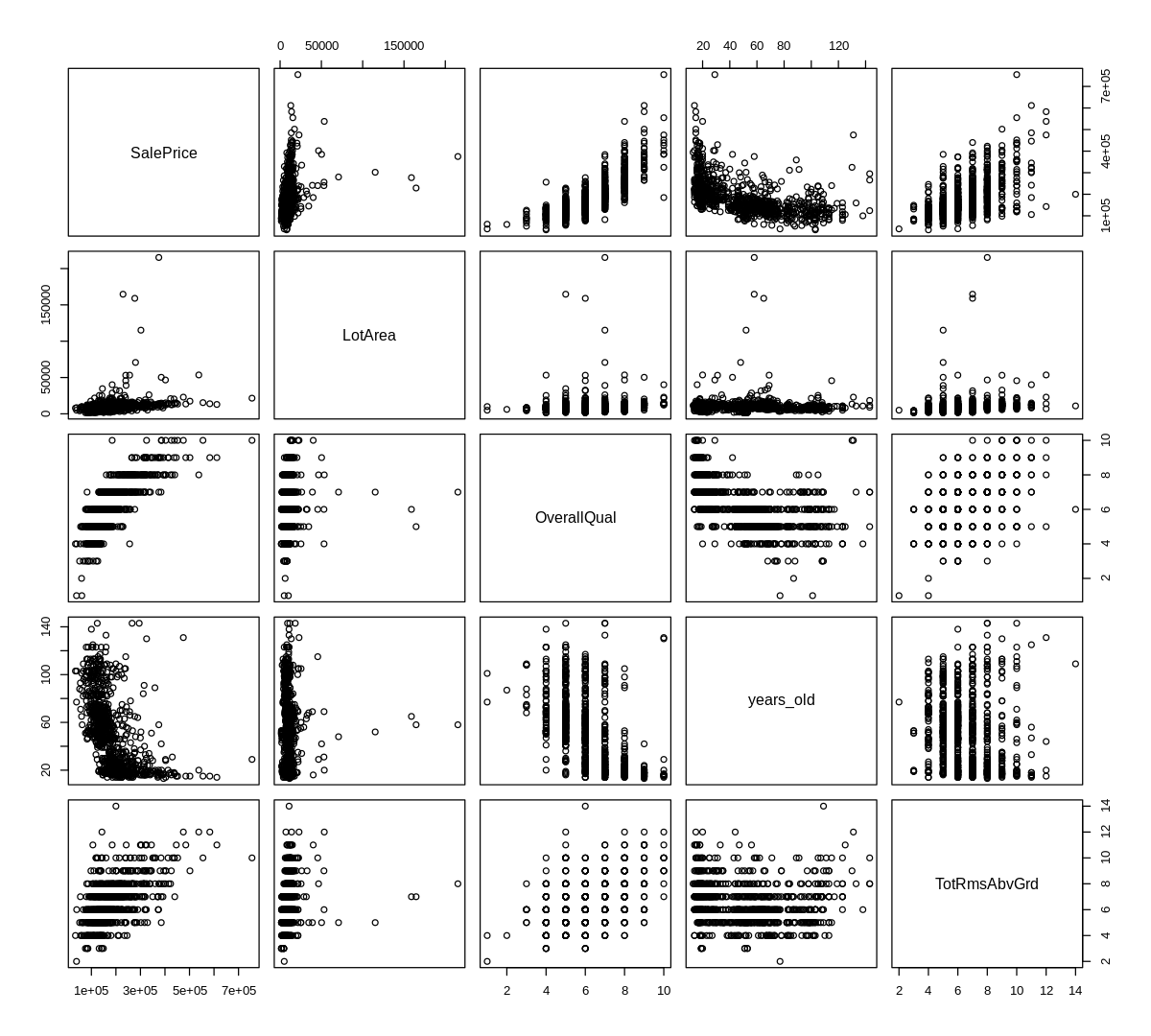


Fig7: Pairplot of variables

# 4. Model Performance Estimation:

The assessment of model performance is a vital process, employing both training and test datasets to achieve robust prediction capabilities.

## 4.1 Regression Model:

### 4.1.1 Model Establishment (based on 'train' dataset):

The development of the regression model began with a thorough training phase, wherein the model is adeptly sharpened on a selected portion of the data. This strategic endeavor permits the model to discern and comprehend subtle links between varied variables and the dynamic world of house values. The training procedure serves as a basic core, assuring the model's proficiency in traversing the various subtleties and capturing the multifaceted dependencies buried within the real estate sector.

## 4.1.2 Prediction Performance (based on 'test' dataset):

Following the robust establishment phase, the regression model's predictive prowess is exposed to rigorous examination employing an independent test dataset. Quantitative measurements, covering RMSE and MAE, are rigorously utilised to quantify and deconstruct prediction mistakes. This complete study gives a nuanced and comprehensive knowledge of the model's performance, revealing insights into its precision and reliability in projecting house prices within the dynamic real estate milieu.

## 4.2 Decision Tree Model:

### 4.2.1 Model Establishment (using the 'train' dataset):

In tandem with the rigorous training schedule of the regression model, the decision tree model begins on a parallel trip, undergoing painstaking training on a specified subset of the dataset. This strategy technique is purposely implemented to give the decision tree with the ability to discover and incorporate nuanced linkages common within the dataset. By navigating through the intricacies inherent in the data, the model hones its potential to capture multifarious and non-linear patterns. This ensures a sturdy foundation for the decision tree's correctness, making it proficient at capturing and interpreting the intricate dynamics buried within the world of real estate attributes.

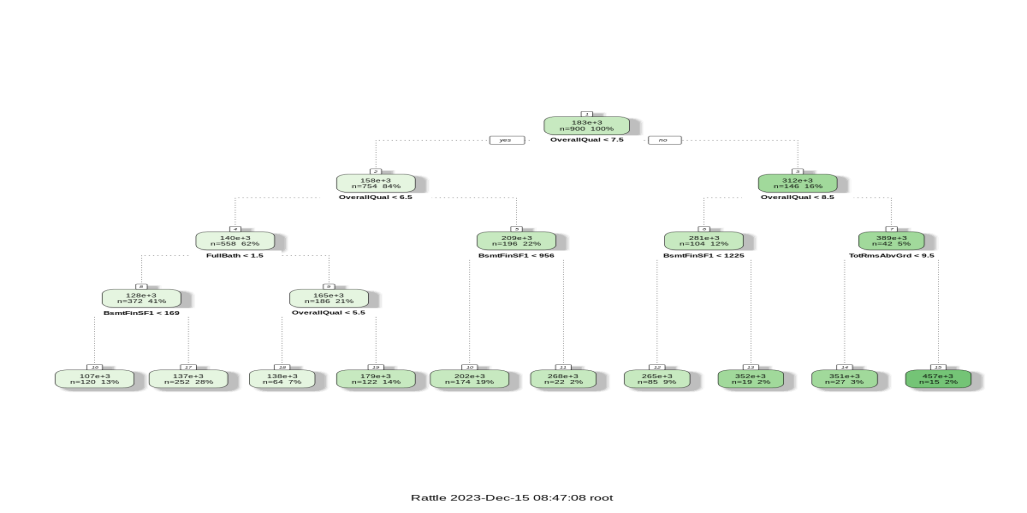


Fig5: Decision Tree Model

## 4.3 Classification Model:

### 4.3.1 Model Establishment (using the 'train' dataset) of the Classification Model:

The rigorous training of the classification model happens within a carefully defined subset of the data. This purposeful method gives the model with the innate ability to forecast the quality category of dwellings with a heightened level of accuracy. The fundamental training phase is crucial, serving as a cornerstone to assure the model's precision in categorizing housing quality based on the chosen features. This strategic initiative establishes the framework for the model's adeptness in recognising the complexities that divide quality categories, adding to a sophisticated grasp of the real estate sector.

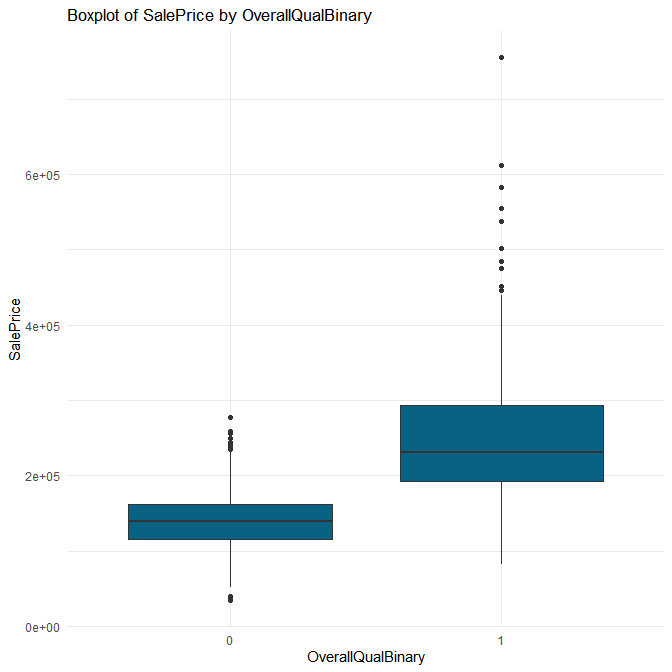


Fig6: Boxplot of salepriceby OverallQuality

### 4.3.2 Forecasting Performance (based on 'test' dataset):

The evaluation of the classification model extends beyond training to incorporate a distinct test dataset. Robust measures, including accuracy, precision, and recall, are selectively utilised to provide a holistic assessment of the model's classification accuracy. This detailed examination offers vital insights into the model's prediction skills within the dynamic world of housing quality categorization, exposing its competency in detecting and accurately classifying houses based on their quality attributes.

A graph of a line graph

Description automatically generated

Fig8: Density plot of residuals

# 5. Insights and Conclusions:

## 5.1 Key Findings:

The regression model displays a respectable accuracy in predicting property values, utilising specified factors.

The decision tree model adeptly incorporates non-linear correlations, raising the precision of forecasts.

The categorization model excels in categorizing residences based on quality, delivering significant insights for prospective purchasers.

# 5.2 Limitations and Future Work:

The limited dataset size provides issues for robust model generalization to fresh data.

Future initiatives should encompass the inclusion of other features, improving the analysis's comprehensiveness.

External factors, mainly economic situations, can influence house prices and demand additional research.

# 5.3 Recommendations:

Employ the regression model for exact housing price projections, leveraging its known accuracy.

Leverage the decision tree approach to untangle complex linkages within the dataset.

Rely on the classification model for successful categorization of residences based on quality, offering helpful recommendations for potential purchasers.

# Conclusion:

This comprehensive home price forecast project displays the skillful integration of varied statistical methodology and machine learning techniques. The regression model exhibits precision in predicting house values, whereas the decision tree model adeptly captures non-linear patterns. The classification technique accurately categorizes houses depending on quality. The findings underline the models' strength in navigating the nuances of real estate dynamics. Limitations include dataset size limits. Future studies should explore other features for a more holistic analysis. These models give useful tools for exact valuation, analysing complex linkages, and categorizing housing quality, contributing greatly to informed decision-making in the real estate arena.