Predicting Perth Housing Prices

STAT5009

Decision Methods & Predictive Analytics Project

Group Name: Creator

31/05/2023

**Predicting Perth Housing Prices**

**Problem Statement and Background**

The objective of this project is to develop a predictive model for housing prices in Perth, Western Australia. By predicting housing prices, potential buyers and sellers can make more informed decisions about the housing market, this will enable them to navigate the market more effectively and optimize their transactions. In addition, government agencies can also use this model to modify government policies to make house ownership more accessible to people.

The cost of buying a house is steadily increasing in many parts of the world, making it more difficult to afford homes. This increase can be for various reasons such as economic growth, population density, and geographic location. The macroeconomic environment had a significant impact on the real estate market, such as the U.S. subprime mortgage crisis turned into a global financial crisis that lead to significant depression in housing market. Housing prices are also influenced by macro policies such as land use and leasing regulations, mortgage market, residential liquidity, volatility, and property taxes [1].

Hence, there can be several parameters in determining the values of house prices. Still, this analysis is limited to the size of the house, the actual area of the house, population data for each suburb, the distance of the house from the CBD, and several highly ranked schools. By leveraging these factors and using machine learning algorithms, models can predict specific area housing prices with varying degrees of accuracy.

The main dataset used in this report is House Pricing data set, The dataset containing information about the prices of housing in Western Australia, is part of this dataset. It contains 33,656 observations of 19 variables recorded between 1990 and 2020.

To make a prediction, this model takes input like size, location, number of bedrooms, number of bathrooms, land area, floor area, distance from CBD, age, nearest school, and its rank. It then combines these parameters with the variables such as the population of that suburb, Australian Citizens, Non-Australian Citizens, and age in the 10 years age group collected from Australian Bureau of Statistics [1].

The data used in this study comes from the Kaggle dataset [2]. The dataset contains historical housing prices in Perth, and it has been used for various analyses by other researchers and data scientists.

**Methods**

The models used in this report mainly include Simple Linear Regression, Multiple Linear Regression, Decision Tree ,Random Forest, house price prediction model can be built by summing up the results from a training dataset, as well as the results from a test dataset, in this proposal, we will outline the steps for building such a model. Our model will be trained using the training dataset, while its performance will be evaluated by using the test dataset. We will experiment with the following model to train on the training dataset:

Simple Linear Regression: Fitting a linear relationship between the house price and each predictor variable separately.

Multiple Linear Regression: Fitting a linear relationship between the house price and all predictor variables simultaneously.

Decision Tree and Random Forest: The performance of each model will be evaluated using cross-validation, so we will be able to select the one that performs the best.

Before constructing the model, the variables are analyzed, which mainly includes exploratory data analysis, analyzing the correlation between variables and exploring which variables have an impact on house prices.

After constructing Simple Linear Regression and the Multiple Linear Regression model, it is necessary to optimize the model, screen out the variables that have significant effects on house prices, eliminate the outliers that affect the model, test whether the model satisfies the assumptions of OLS, and evaluate whether the model satisfies the assumptions of linearity, normality, homoscedasticity, and independence.

The decision tree model is constructed with house price as the outcome variable and other variables as the predicted predictor variables, and the unimportant branches are cut off based on the complexity parameter cp on the basis of the complete tree. Thus, the size of the tree is controlled within the ideal range and an ideal size tree is obtained.

Random forest model, construct a random forest model with house price as the outcome variable and other variables as predictor variables, and give the importance of the variables.

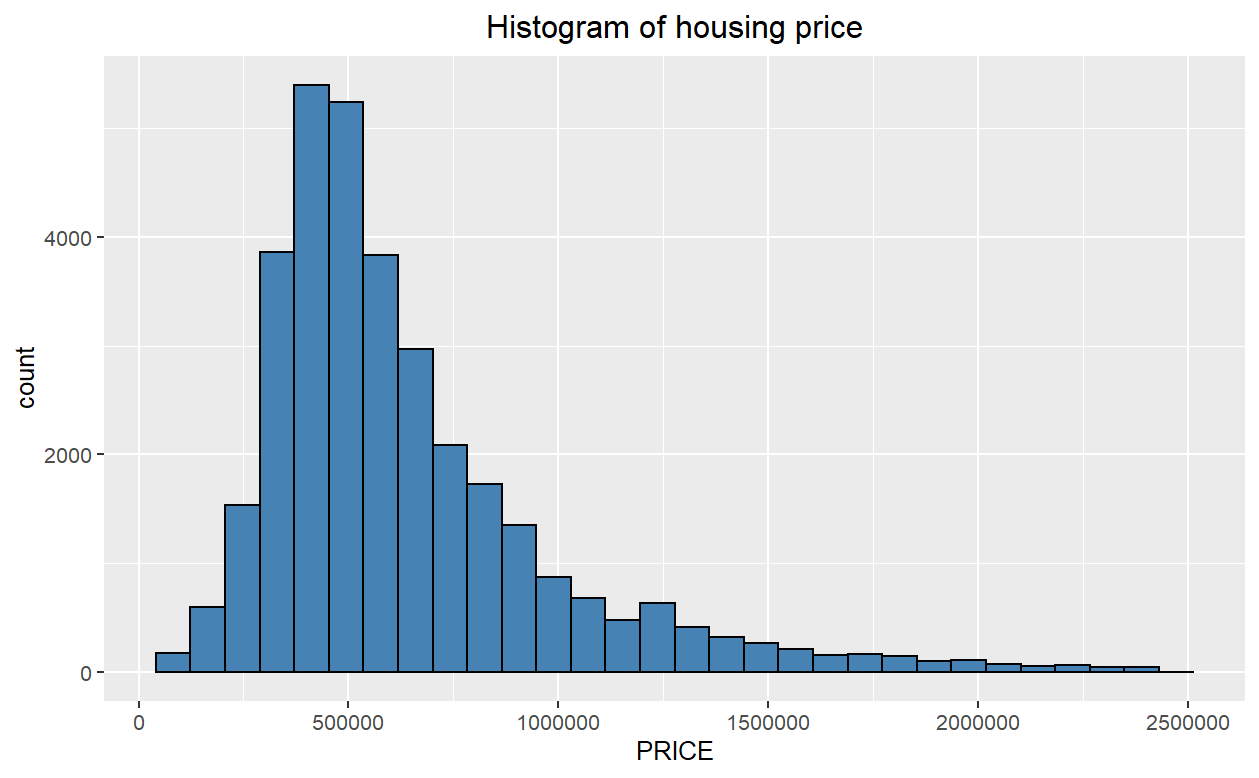
**Model prediction evaluation**

Before constructing the models, 80% of the data set is randomly selected as the training set for training the models, and the remaining 20% of the data is used as the test set for evaluating the accuracy of the models.

In this report, Simple Linear Regression, Multiple Linear Regression, Decision Tree and Random Forest models are constructed with the training set data, and the models trained in the test set are used to predict the house price of the test set respectively, and the predicted The predicted house price is compared with the actual house price in the test set, and the accuracy of the models is evaluated by calculating the mean squared error, mean absolute error and other indicators, and the model with the highest accuracy is selected.

**Results**

The histogram below shows that housing prices in Perth are mainly concentrated in the 20,000-200,000 range, with housing prices over 200,000 and less than 20,000.



Overall, the number of bedrooms in a house is positively correlated with the price of the house, and as the number of bedrooms in a house increases, the price of the house also tends to increase.

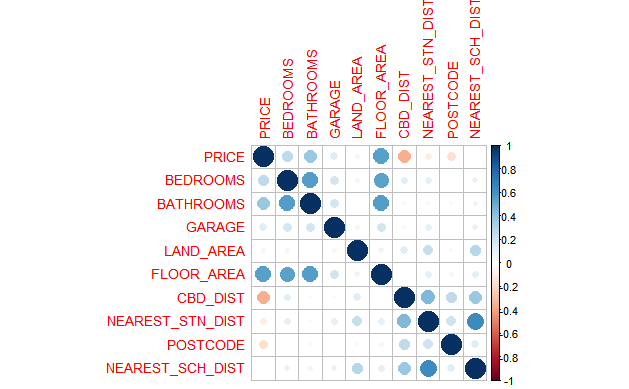


**Data cleaning**

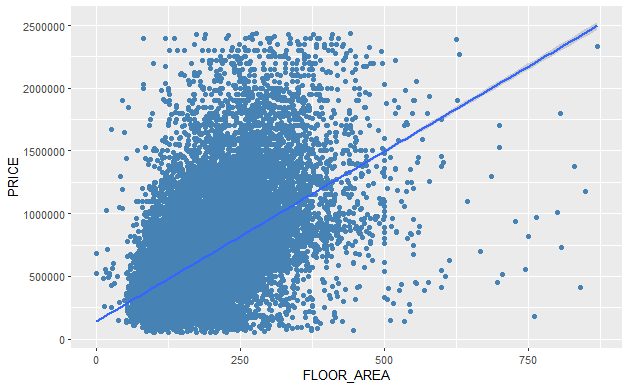
Before building the model, the data is cleaned, which mainly includes converting the variable types of the data, transforming the character variables into numeric variables, deleting some meaningless variables, interpolating the missing values present in the data, and using the mean values of the neighboring data to interpolate the missing values, and finally getting clean data.

**Correlation analysis of variables**

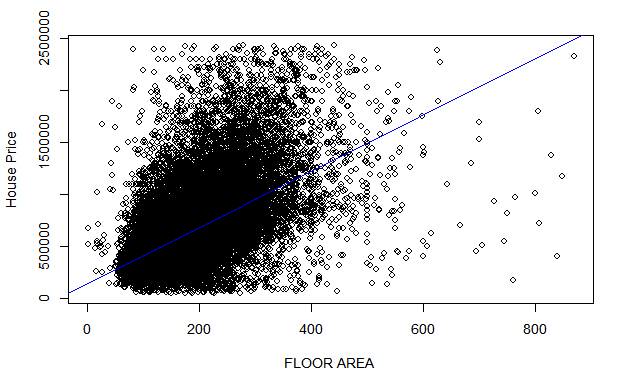
From the correlation coefficient graph below, it can be seen that price has a high positive correlation coefficient with FLOOR AREA and Price has a large negative correlation coefficient with CBD DIST.



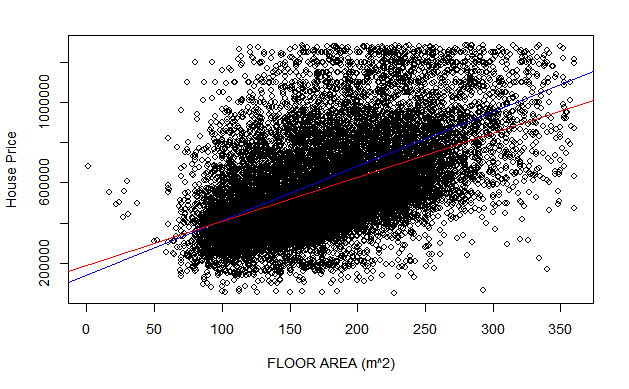
A scatter plot of price versus FLOOR AREA is plotted and fitted lines are generated. The graph shows that there is a significant linear positive correlation between FLOOR AREA and price.



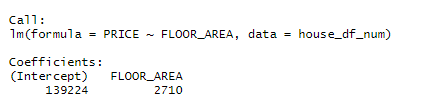
Plotting the scatter plot of price versus FLOOR AREA and fitting a straight line, the figure below shows that there is a linear relationship between FLOOR AREA and price.



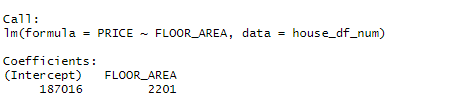
Before constructing Simple Linear Regression, the outliers of PRICE and FLOOR AREA are removed first, and the outliers have some influence on the fitted straight line.



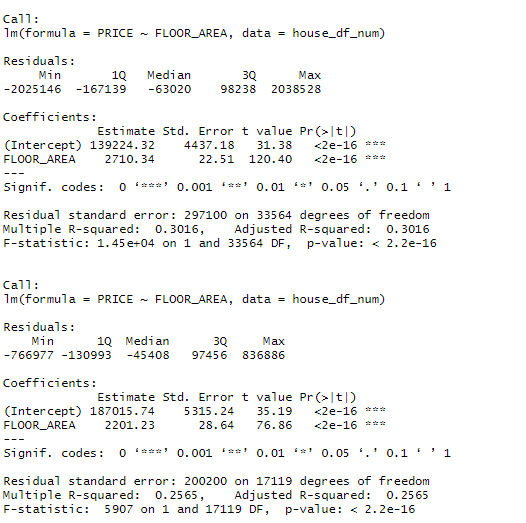
The output of the regression model shows that there is a positive linear relationship between FLOOR AREA and PRICE, and on average, for every 1 unit increase in FLOOR AREA, Price will increase by 2710 units.



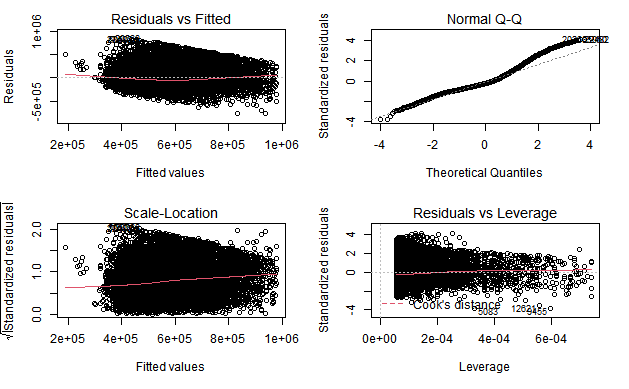
After excluding the extreme values, the regression coefficient of FLOOR AREA is 2201, which means that for each unit increase in FLOOR AREA, Price will increase by 2201 units.



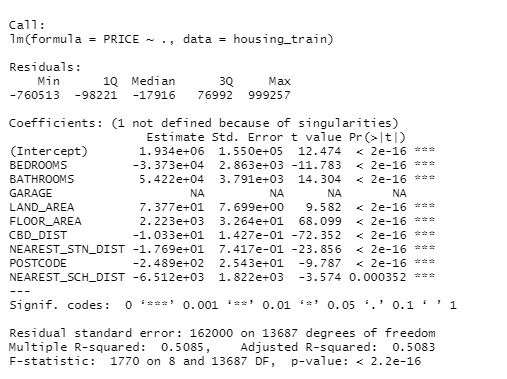
Excluding the outliers, the decrease of the regression coefficient of FLOOR AREA and the slope of the fitted straight line are smoother.



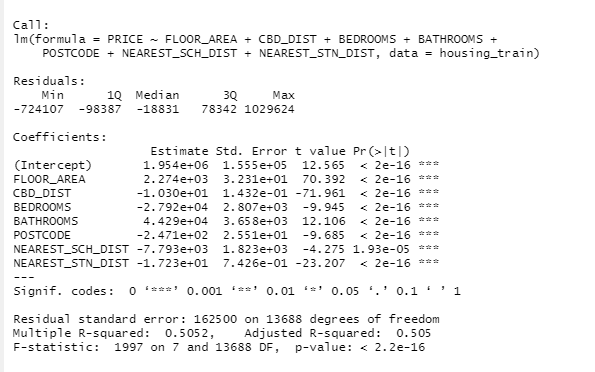
The model was evaluated, and the assumption of normality was not well satisfied by the model and there were still some strong influence points.



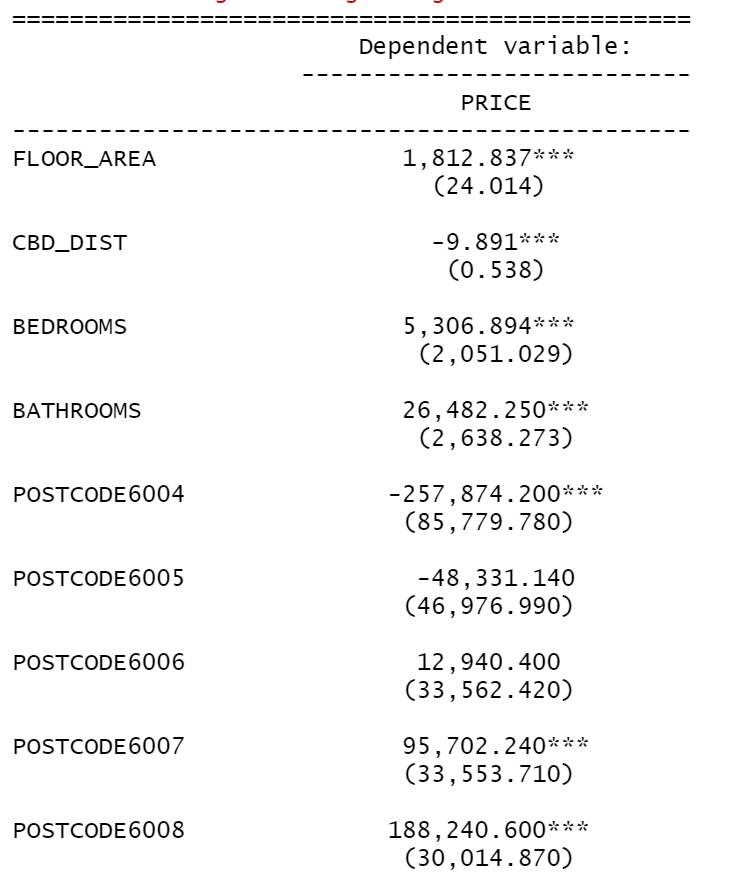
A multiple linear regression model was constructed with PRICE as the dependent variable and other variables as independent variables, and the model output showed that the R-squared of the model was 0.5083, indicating that the model was able to explain 50.83% of the variance of PRICE, but the model still had some variables that did not pass the significance test.

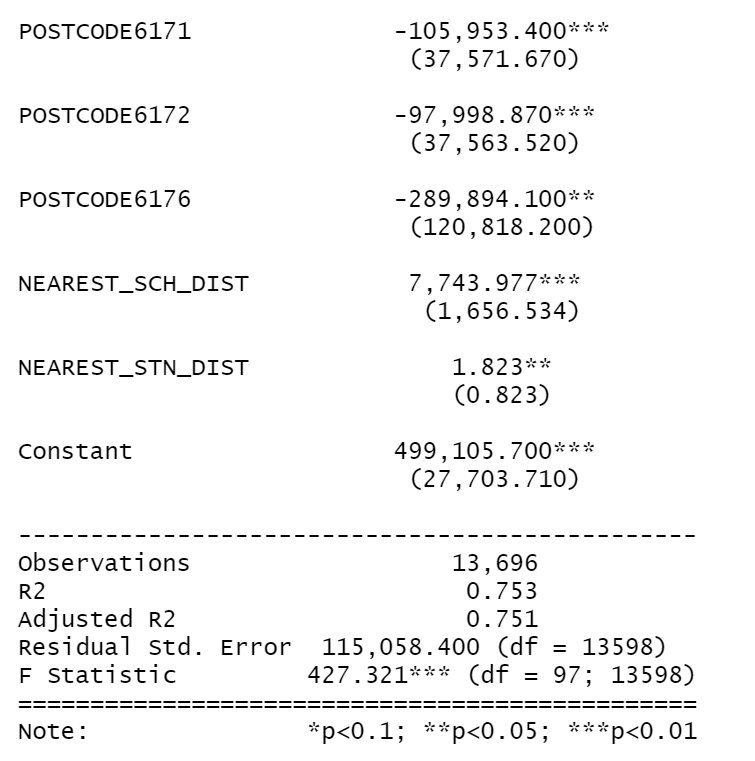


While checking upon the GARAGE column, it doesn't have NA values, but NA values seen in coefficient output are due to the multicollinearity between the variables. So, GARAGE is removed, and regression analysis is performed again.



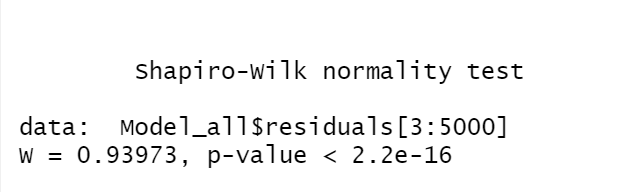
Converted POSTCODE variable from int to categorical using as. factor function and visualizing the coefficient. There is a significant change in the R-squared value compared to the model in which the POSTCODE column changed to a categorical variable. For that, we accept this new model.



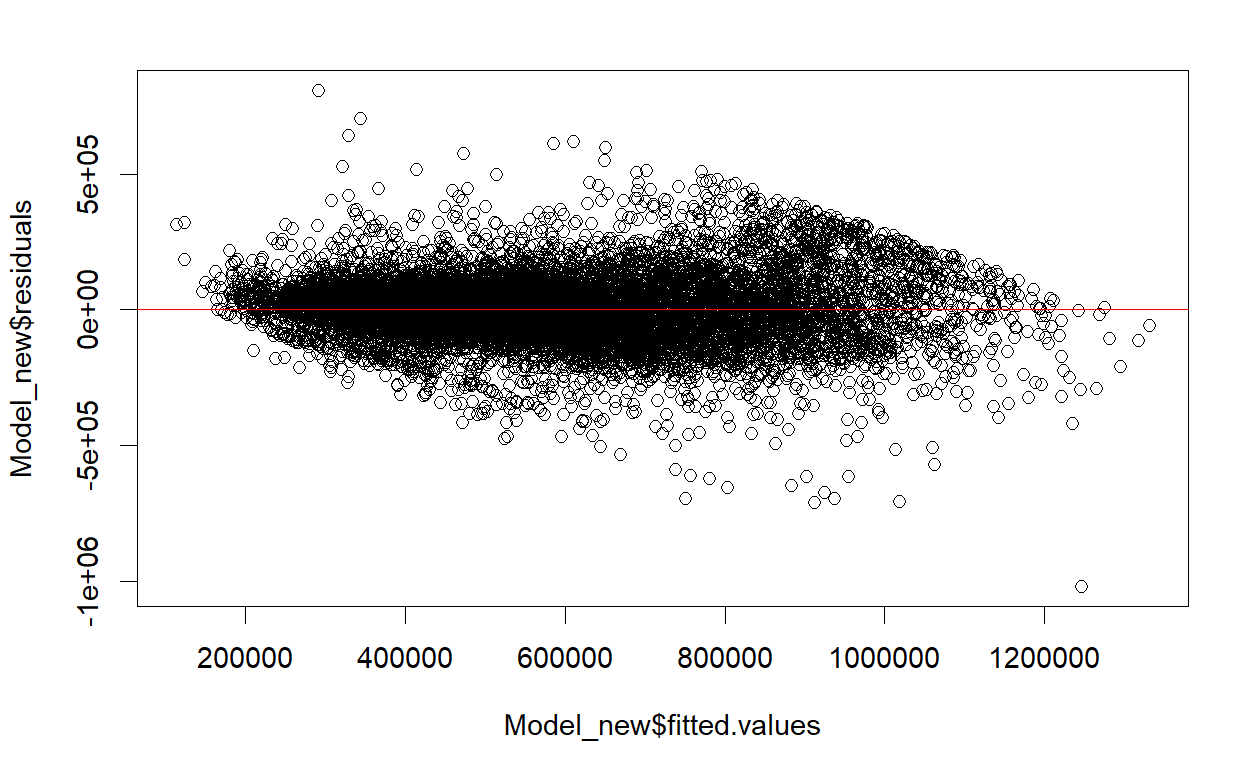


the RMSE result is 115896.4, which means the model has good enough to predict.

The Shapiro-Wilk normality test on the subset of residuals (Model\_all$residuals[3:5000]) resulted in a test statistic of W = 1 and an extremely small p-value (< 2e-16), indicating strong evidence against the null hypothesis of normality. Therefore, we can conclude that the subset of residuals does not follow a normal distribution.

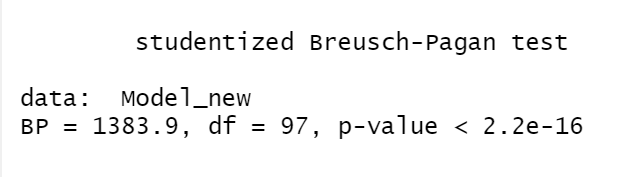


Model has normal distribution and p-value < 0.05



p-value less then 0.05 conclude that heteroscedasticity is present in the regression model.

The studentized Breusch-Pagan test on Model\_all resulted in a test statistic of BP = 2997, with degrees of freedom (df) equal to 9, and an extremely small p-value (< 2e-16). This indicates strong evidence against the null hypothesis of homoscedasticity, suggesting the presence of heteroscedasticity in the residuals of the model.

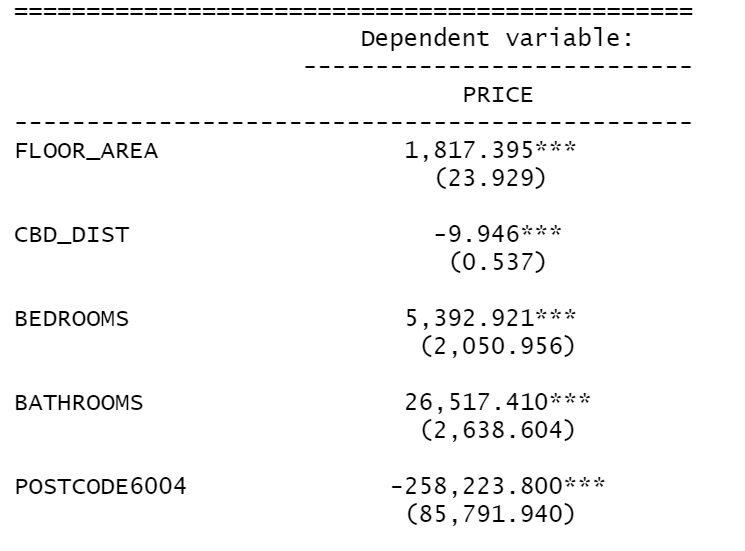


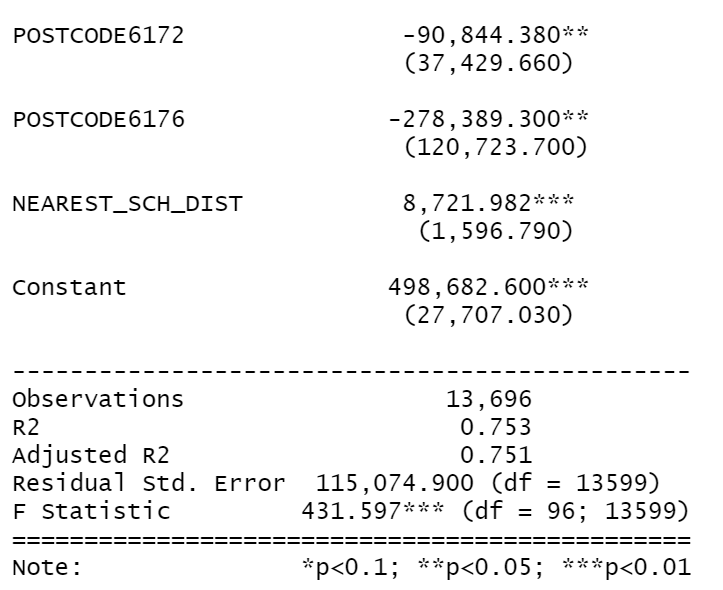
The null hypothesis of homoscedasticity in regression analysis is that the error terms (residuals) have constant variance across all levels of the predictor variables. In other words, it assumes that the spread of the residuals is the same for all values of the predictors.

The error message "there are aliased coefficients in the model" suggests that there is perfect multicollinearity or linear dependence among the predictors in our model Model\_all.

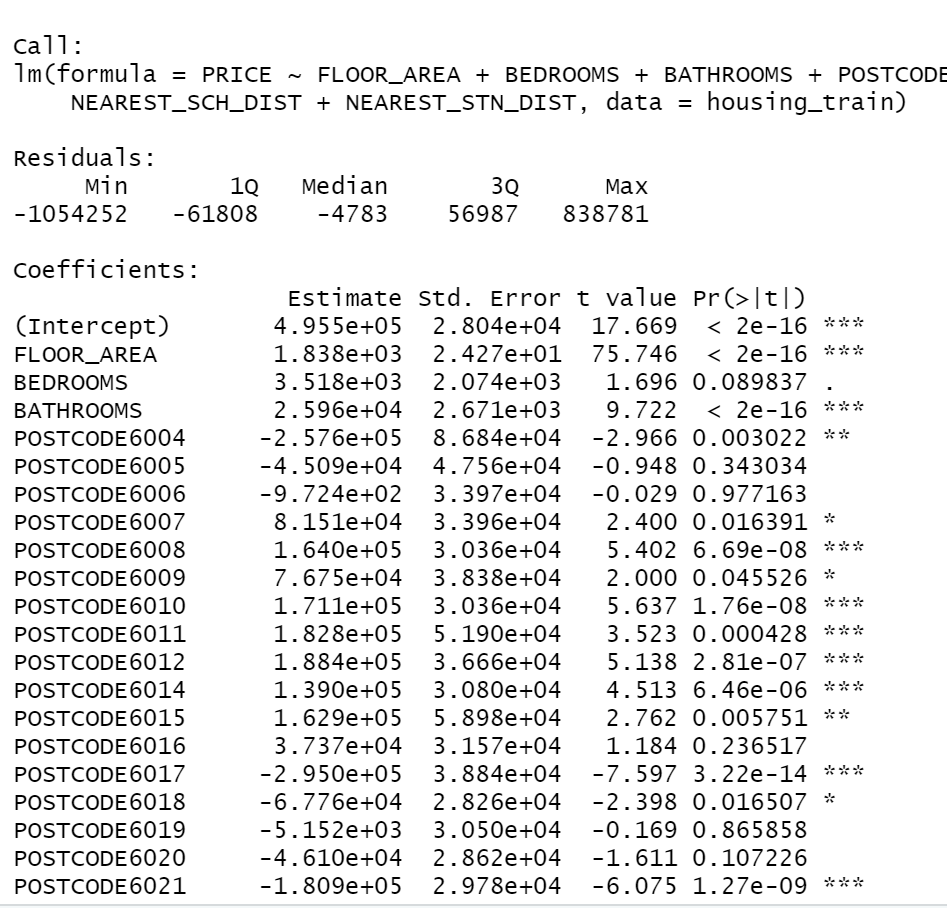
To resolve this issue, identifying and removing one or more predictors that are causing the multicollinearity is required.

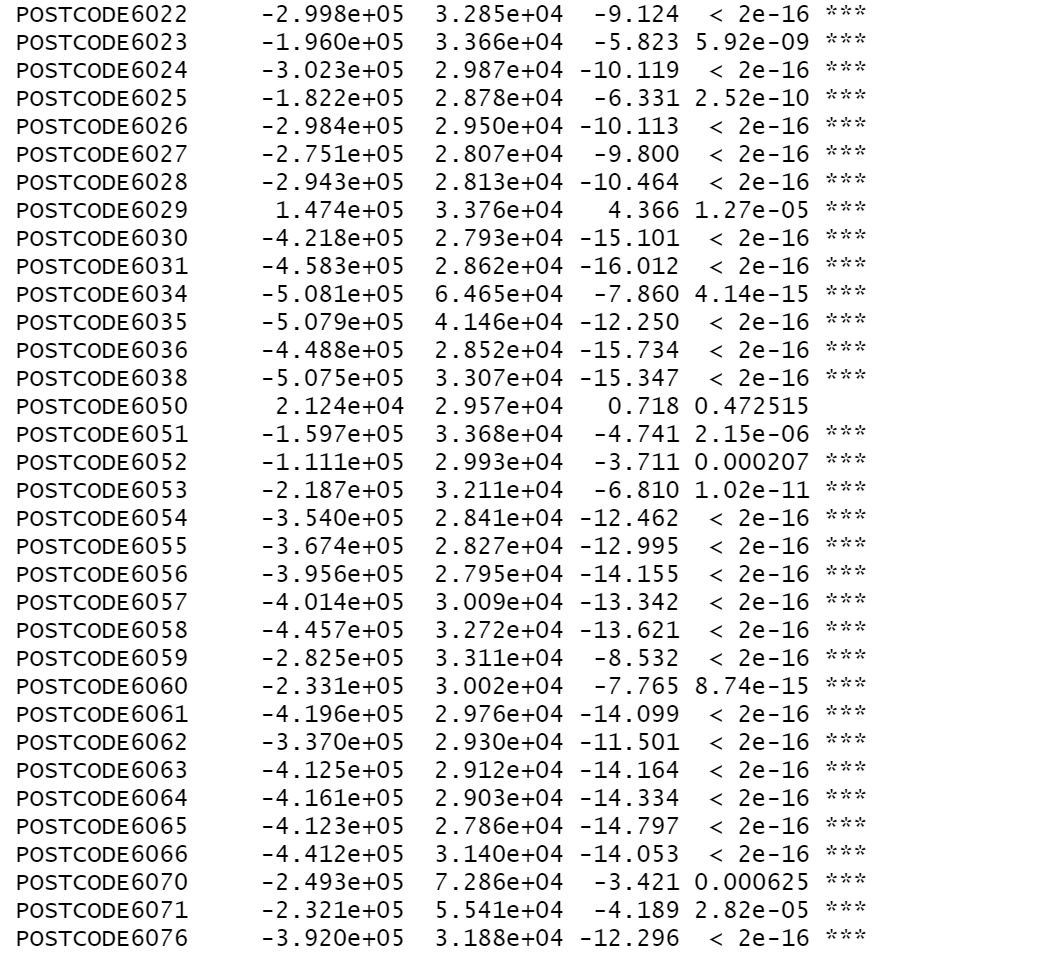
Removed NEAREST\_STN\_DIST and analyzed the coefficient.

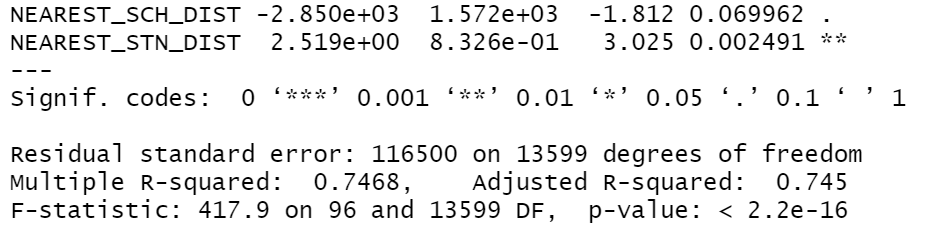




We can conclude that multicollinearity exists on POSTCODE and CBD DIST ,by removing either of the columns. Removing CBD\_DIST and analysis the regression coefficient. The final multiple linear regression model is obtained as follows:

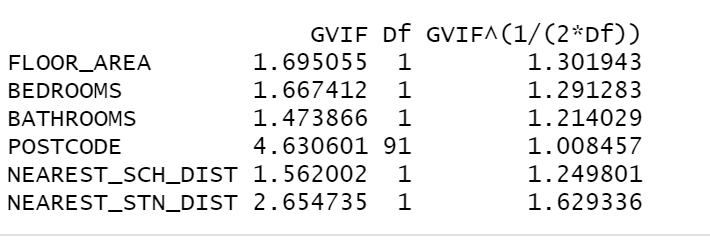




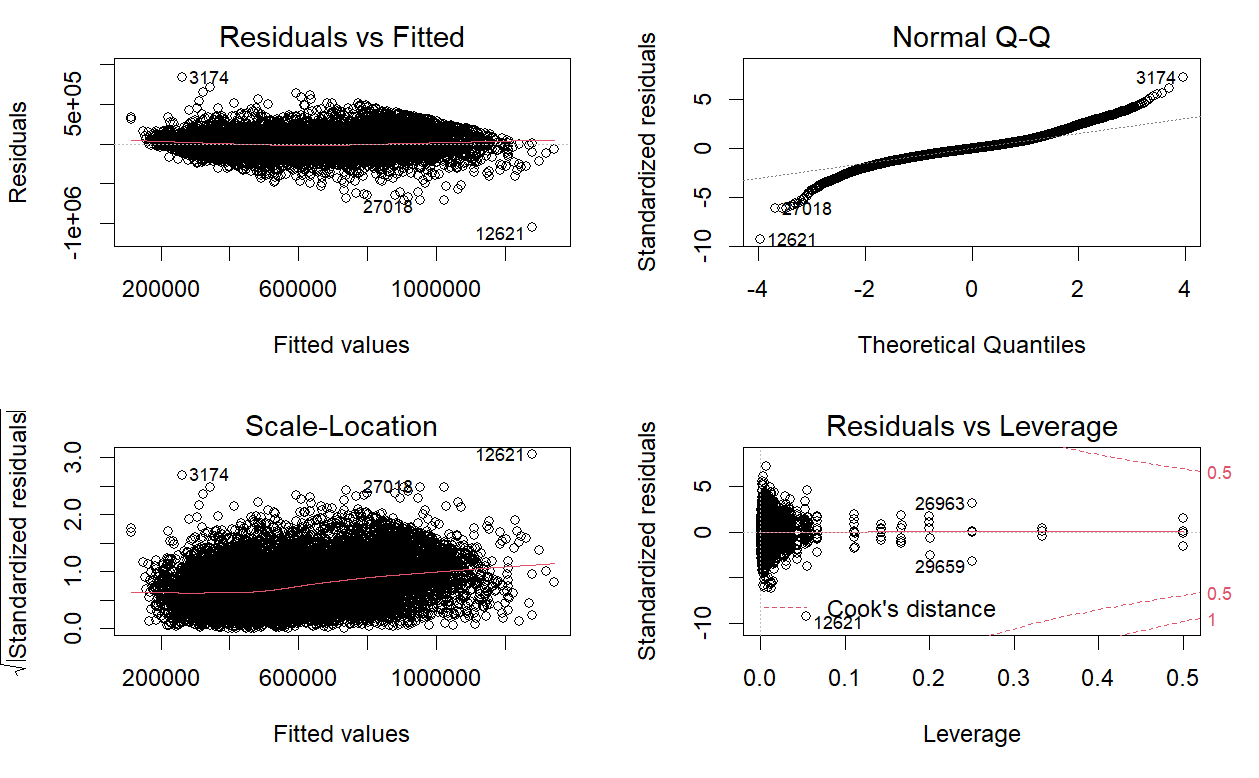


The R-square of the model is equal to 0.745, indicating that the model can explain 74.5% of the variance in house prices. The fit of the model has been substantially improved compared to the previous one.

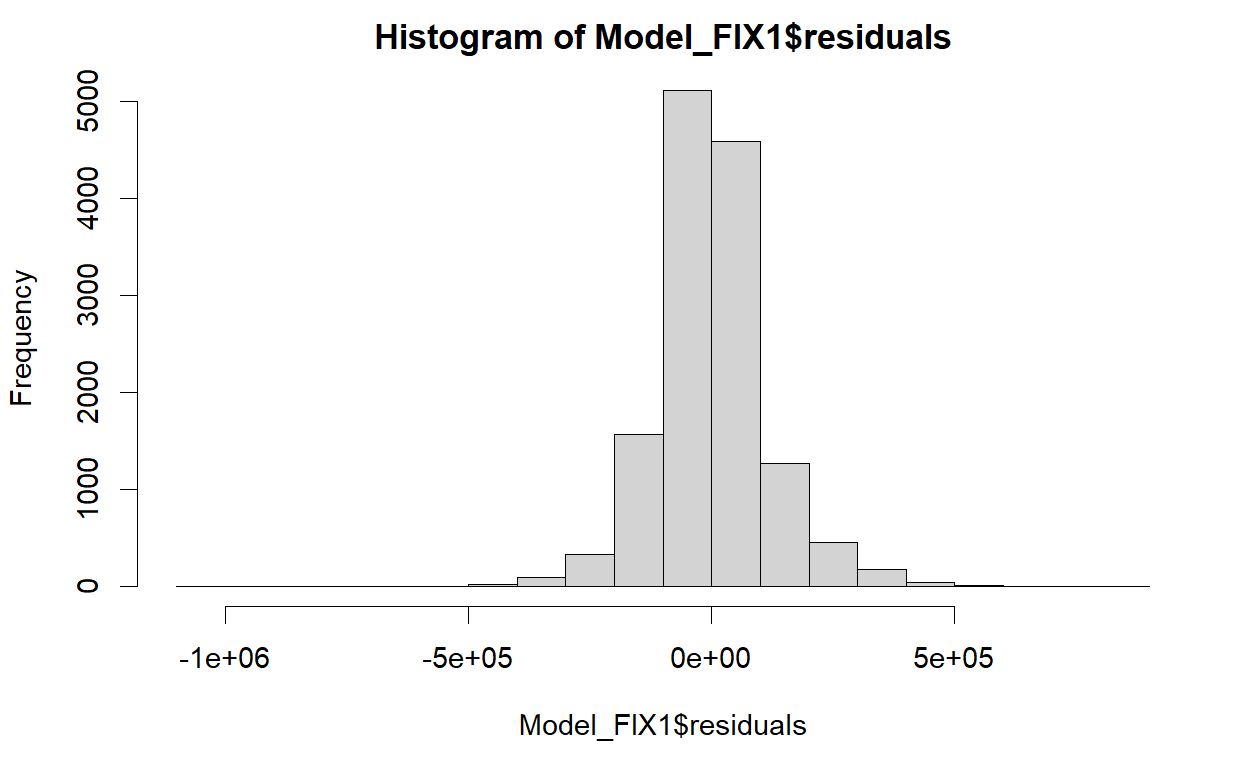
Generally, VIF values below 5 indicate low multicollinearity, so based on these values, there is no significant multicollinearity issue among the predictor variables in the model.



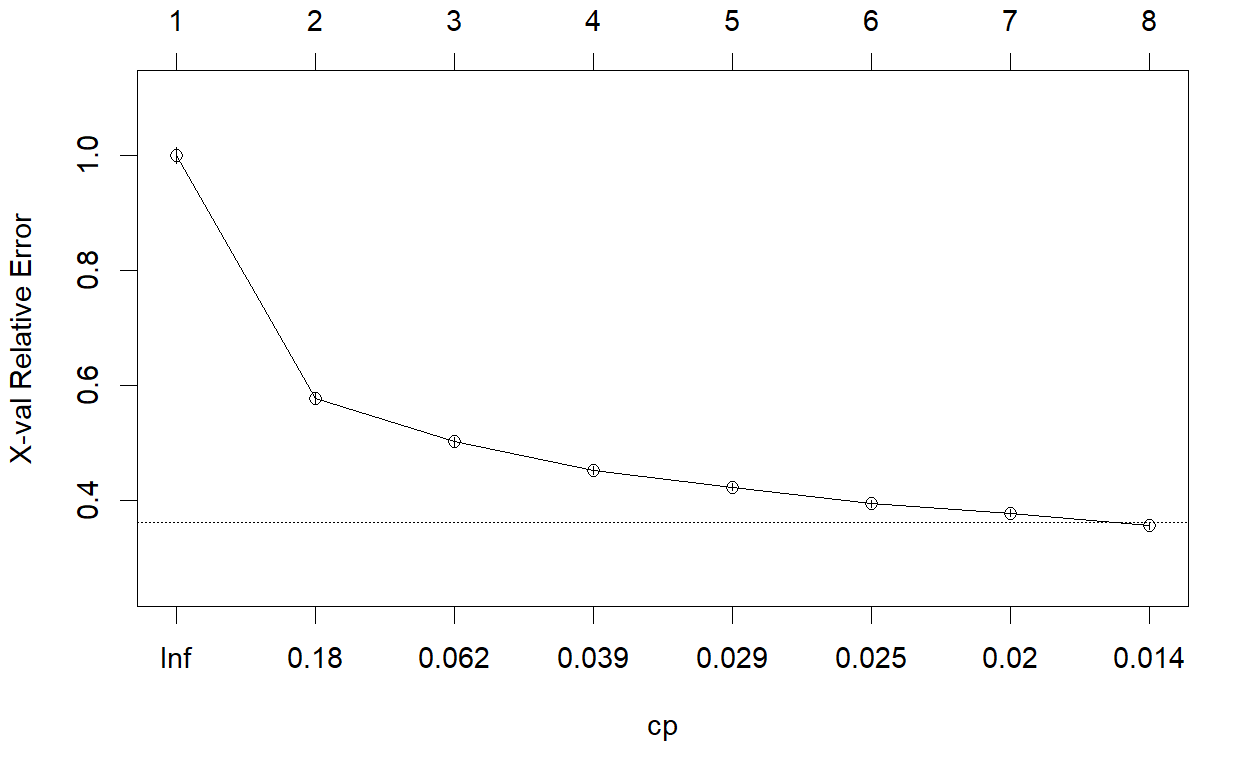
Based on the four graphs below, the model satisfies the assumption of linearity, normality, and homoscedasticity with only a few extreme values.

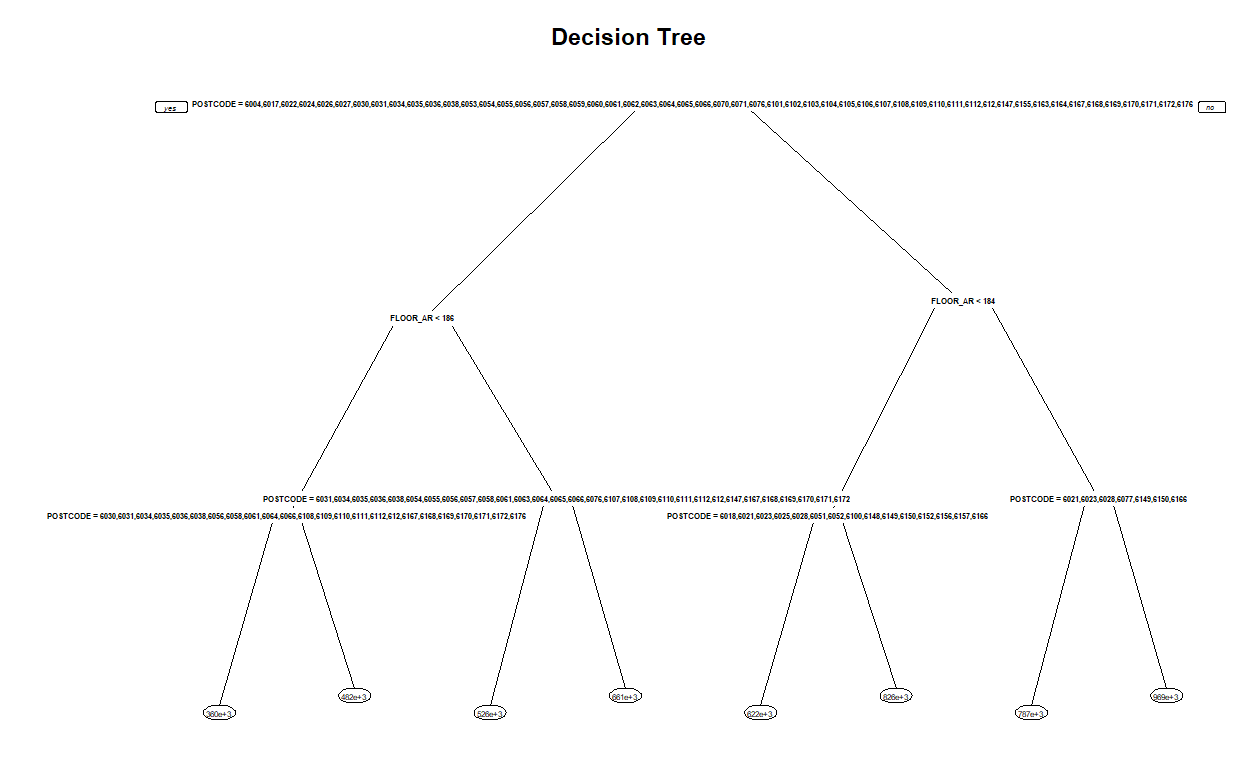


The residuals of the multiple linear regression model approximately follow a normal distribution.

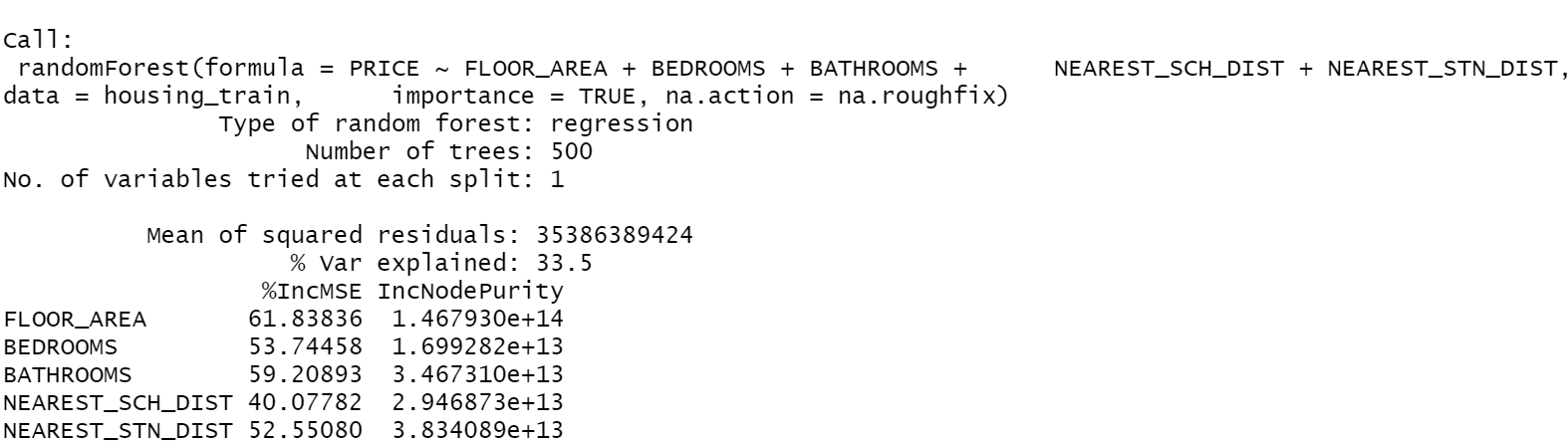


By constructing the decision tree model and optimizing the decision tree model with the cp value, the final decision tree model was obtained as shown in the figure below, the decision tree model was trained using the training set and the accuracy of the model was calculated in the test set, the RMSE of the decision tree model was 138885.6, compared with the previous multiple linear regression model, the accuracy of the multiple linear regression model was higher.





Random forest model, by constructing a random forest model and outputting the importance of variables. Based on the random forest model, the house prices were predicted in the test set, and the RMSE was obtained by comparing with the actual house prices, and the value of RMSE was 262031.8, which was higher than the previous decision tree model and multiple linear regression model.



**Conclusions and Lessons Learned**

By constructing simple linear regression, multiple linear regression, decision tree and random forest models respectively, we constructed models for predicting house prices based on these four types of arithmetic put, trained the models in the training set, and evaluated the accuracy of the models based on the test set. The multiple linear regression is easier to be understood by non-specialists than other models, and it can be used to evaluate the price of a house faster for home sellers, home buyers and agents to give a reasonable price for a house quickly.

Through this course, I have mastered the complete process of carrying out a project, mastering the methods of data collection, data cleaning, data analysis, model construction, model optimization, model evaluation, model testing, model comparison, etc., and can apply the knowledge learned to solve practical problems.

***References***

[1] Dan, A. , AC Sánchez, & Åsa Johansson. (2011). Housing markets and structural policies in oecd countries. Oecd Economics Department Working Papers.

[2] Australian Bureau of Statistics, “Australian Bureau of Statistics,” Australian Bureau of Statistics, 20 April 2023. [Online]. Available: https://www.abs.gov.au. [Accessed 24 April 2023].

[3] M. S. ZAINAL, “Kaggle,” 2021. [Online]. Available: https://www.kaggle.com/datasets/syuzai/perth-house-prices/discussion?select=all\_perth\_310121.csv. [Accessed 24 April 2023].