Melbourne Housing Price Prediction

Fahmi Harum

6/2/2020

1.Introduction

Data has been essential for nowadays growth, data can be manipulated and elaborated by using statistical and analytics knowledge. Through this process data can be one of the most precious things in the world. This project concerned on the data of Melbourne Housing Price.

Since housing price are fluactutating in trends for nowadays, the prediction of its price for future would be a much a help. To proceed with that, this project would develops few models and compare them for it best accuracy for this problem by calculating its R-squareds,Sum Squared Error(SSE)s and Sum Squared Total(SST)s.

The data is acquired from Kaggle website with a total size of 12 megabytes with 2 datasets. These datasets contains 34 columns in total and 34857 rows distinctly.However, not both datasets are used, only full housing data will be used that consists of 21 columns and 34857 rows.

This project will conduct analysis, visualisation and analytics to the dataset and conclude best model to be used for the price prediction purposes.

Datasets

Kaggle : https://www.kaggle.com/anthonypino/melbourne-housing-market/download (https://www.kaggle.com/anthonypino/melbourne-housing-market/download)

- Melbourne_housing_FULL.csv
- Melbourne_HOUSE_PRICE_LESS.csv

Aim & Objectives

The aim of this project is to come up with machine learning algorithms that are able to make use of the dataset stated above in Dataset Section, which will forecast the price of the houses in Melbourne in the future.

Few external libraries are used in data exploration and visualiation, model development and predictions. Factors from the dataset will be make used in a optimized manner to produce best and most accurate model. A total of 5 models would be created and measured for error using SSE and SST methods. Then, most optimized model would be concluded for prediction purpose.

2.Methodology & Analysis

R-Squared

R-Squared(R2) is a statistical measure of fits thah shows the variation of a dependent variable is explained by the independent variables in a regression models. This method will be used as all of the models are based on regression.

R-squared(R2) of a 100% proportion means that all changes in other independent variables are related to changes in the variable that is currently measured.

R-Squared Formula:

```
R^2= 1-(SSE/SST)
```

Dataset Overview

Before proceeding further, a sneak peak into the dataset would help with understanding the data even better. Below, missing values are checked by columns, and 10 columns are infersted with missing values which will need data preprocessing and cleaning.

```
##
## Total Number of [rows vs columns] in the dataset-
##  34857 21
## --------------##-----------------
##
##  Datatypes of all the columns in the dataset-
##  character character integer character integer character character character charac
ter character integer integer integer integer numeric integer character numeric numeri
c character character
## --------------##-----------------
##
##  Names of all the columns in the dataset-
##  Suburb Address Rooms Type Price Method SellerG Date Distance Postcode Bedroom2 Bat
hroom Car Landsize BuildingArea YearBuilt CouncilArea Lattitude Longtitude Regionname
Propertycount
```

```
## Total NA in the dataset in all in the columns-
##
##  100964
## --------------##-----------------
##
##  Names of NA columns in the dataset-
##
##  Price Bedroom2 Bathroom Car Landsize BuildingArea YearBuilt Lattitude Longtitude
##
##  Total NA by column in the dataset-
##
##  0 0 0 0 7610 0 0 0 0 0 8217 8226 8728 11810 21115 19306 0 7976 7976 0 0
## --------------##-----------------
```

Columns Type

Before the data cleaning, the need of classifying columns to it suitable type is crucial, as fews columns are unuseful and fews can be considered as factors. After column recategorization, we would like to recheck the columns' types.

```
##        Suburb       Address         Rooms          Type         Price
##      "factor"   "character"      "factor"      "factor"     "integer"
##        Method       SellerG          Date      Distance      Postcode
##      "factor"      "factor"   "character"     "integer"      "factor"
##      Bedroom2      Bathroom           Car      Landsize  BuildingArea
##     "integer"      "factor"     "numeric"     "integer"     "numeric"
##     YearBuilt   CouncilArea     Lattitude    Longtitude    Regionname
##     "integer"      "factor"     "numeric"     "numeric"      "factor"
## Propertycount
##      "factor"
```

Data Pre-processing

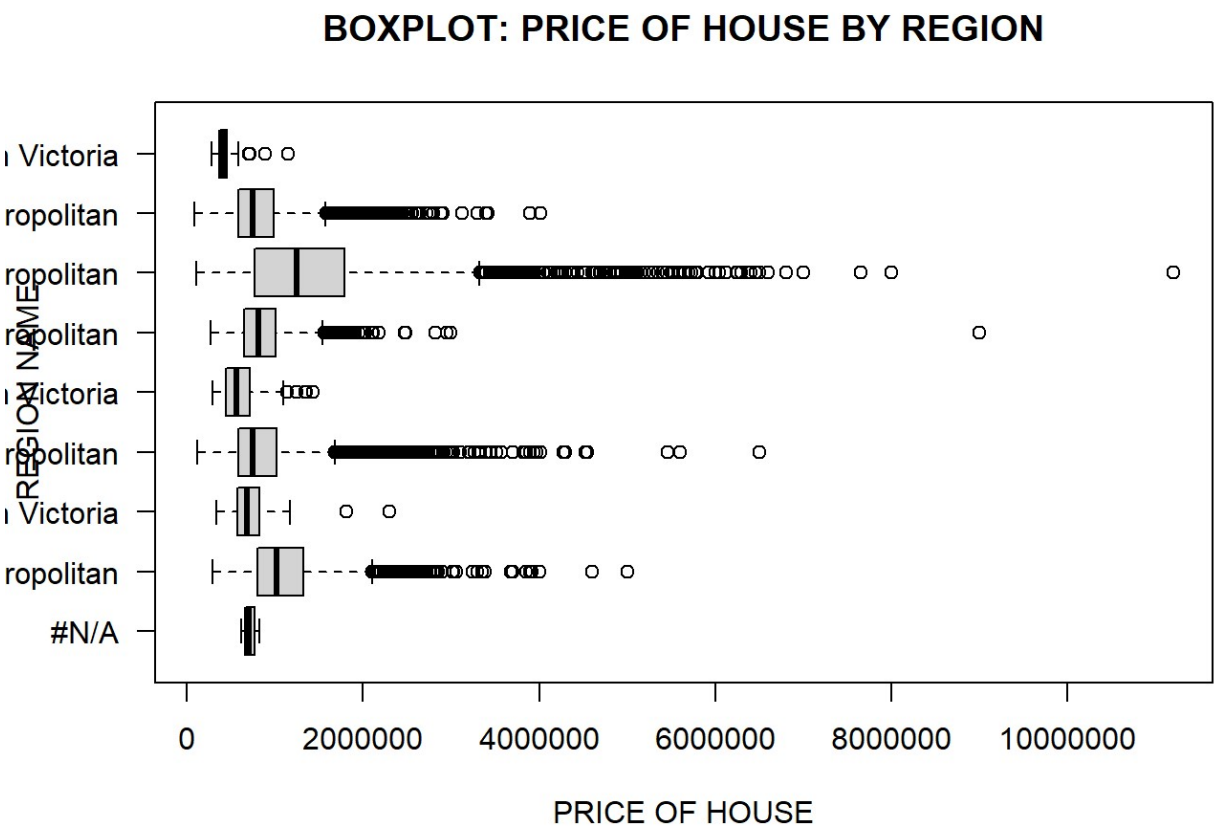
Missing Values

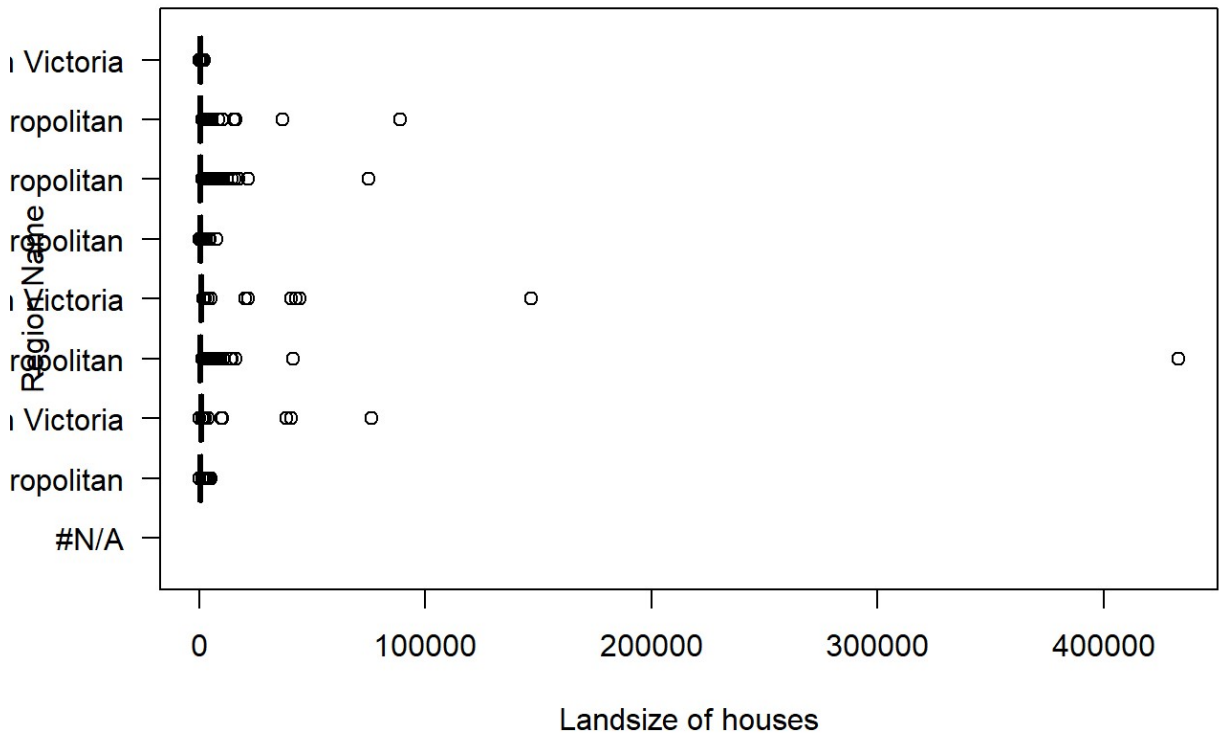
After reviewing the data types, the missing values are rechecked. Later the NA is fixed by; The price column would eliminate the NA values as it acts as a dependent variable, BuildingArea column is removed because is has low factor to the prediction and consist of 60% of missing values, Rooms column is eliminated and the values are substituted to bedrooms2 columns as it has same values with Rooms.

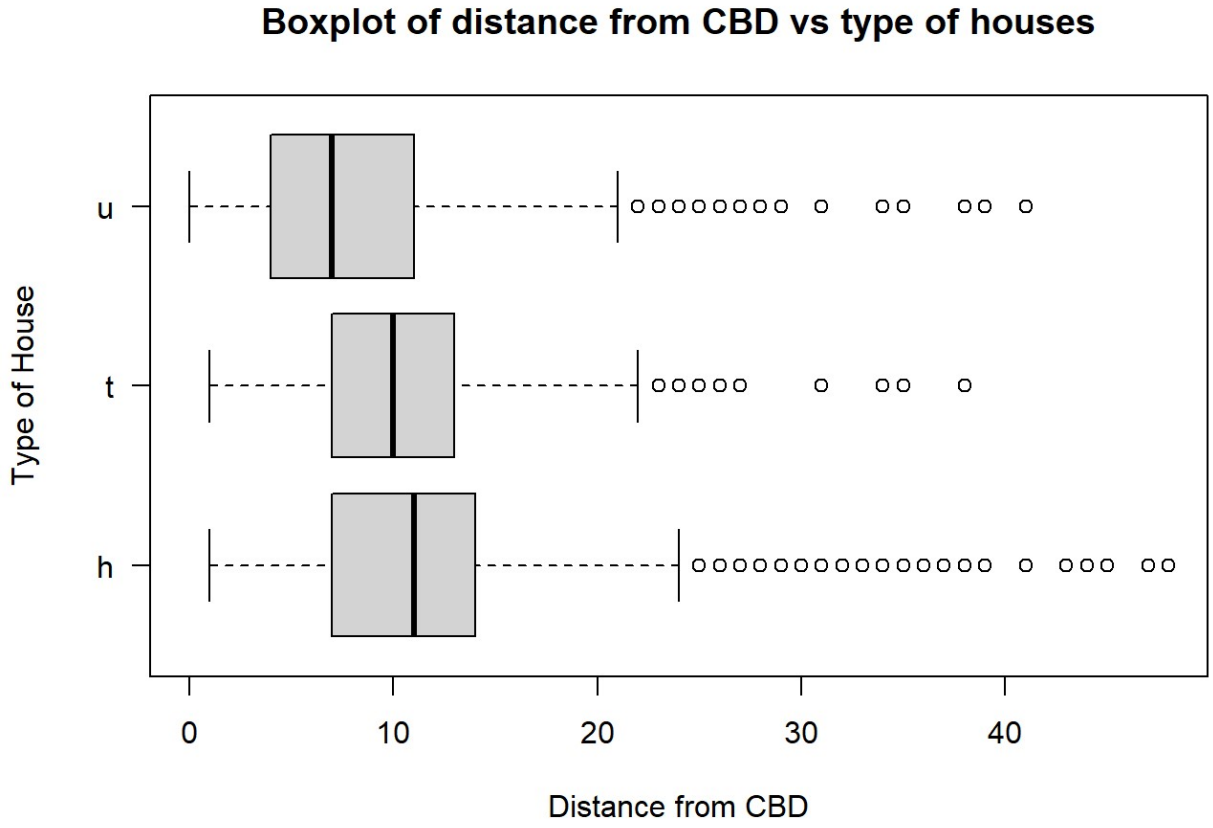
```
##        Suburb       Address         Rooms          Type         Price
##             0             0             0             0          7610
##        Method       SellerG          Date      Distance      Postcode
##             0             0             0             1             0
##      Bedroom2      Bathroom           Car      Landsize  BuildingArea
##          8217          8226          8728         11810         21115
##     YearBuilt   CouncilArea     Lattitude    Longtitude    Regionname
##         19306             0          7976          7976             0
## Propertycount
##             0
```

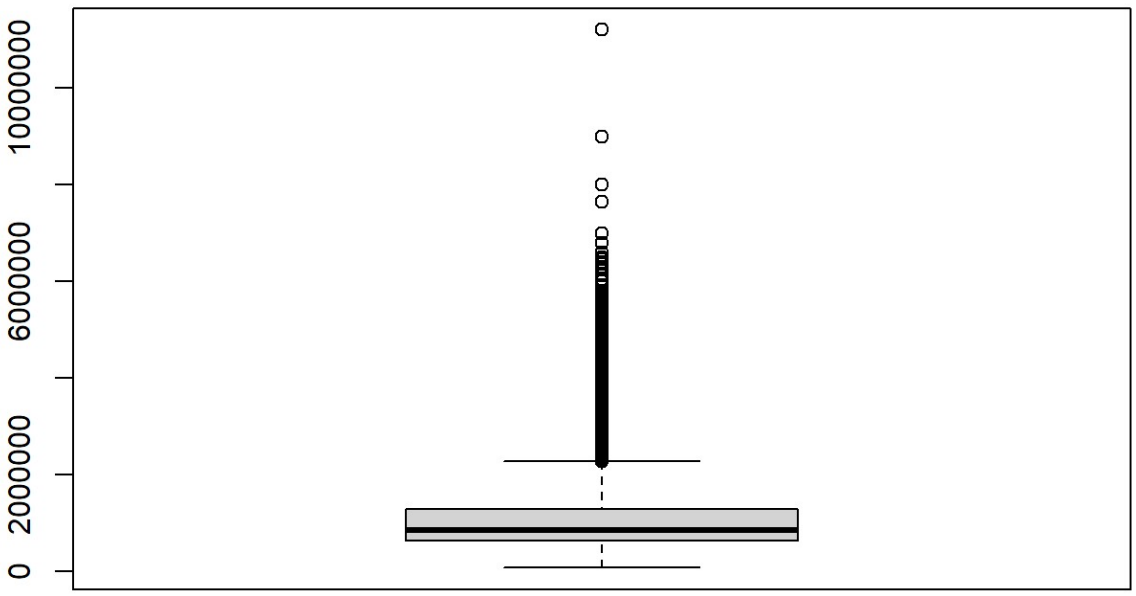
Outliers

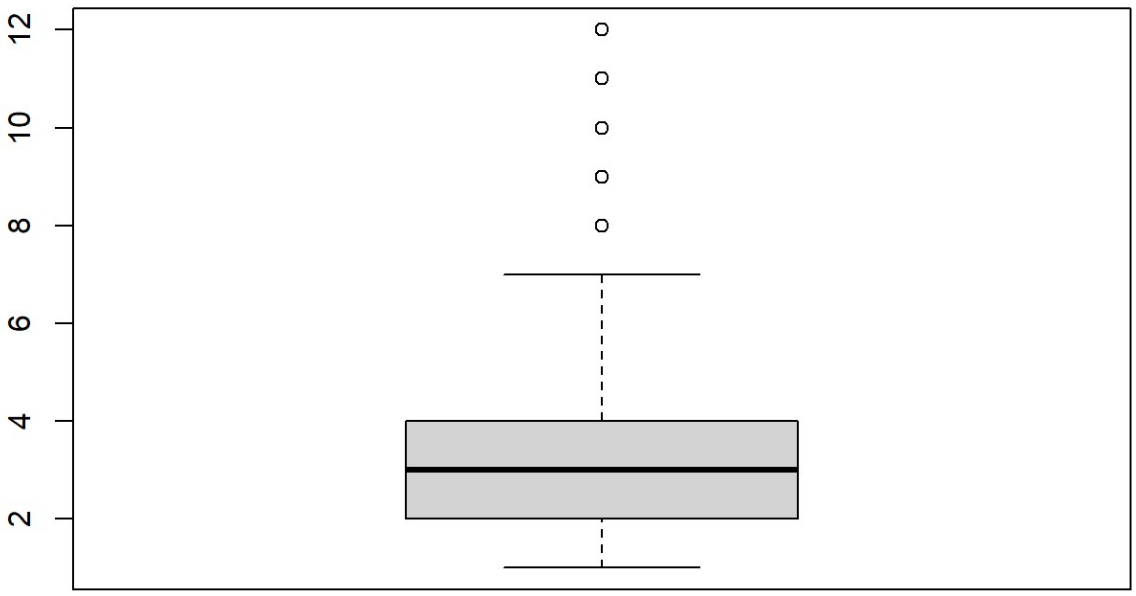
From the plot below, it can be seen many outliers are ingested in few columns. However, the outliers are eliminated by eliminating the rows, and the result of outliers elimination are shown below.



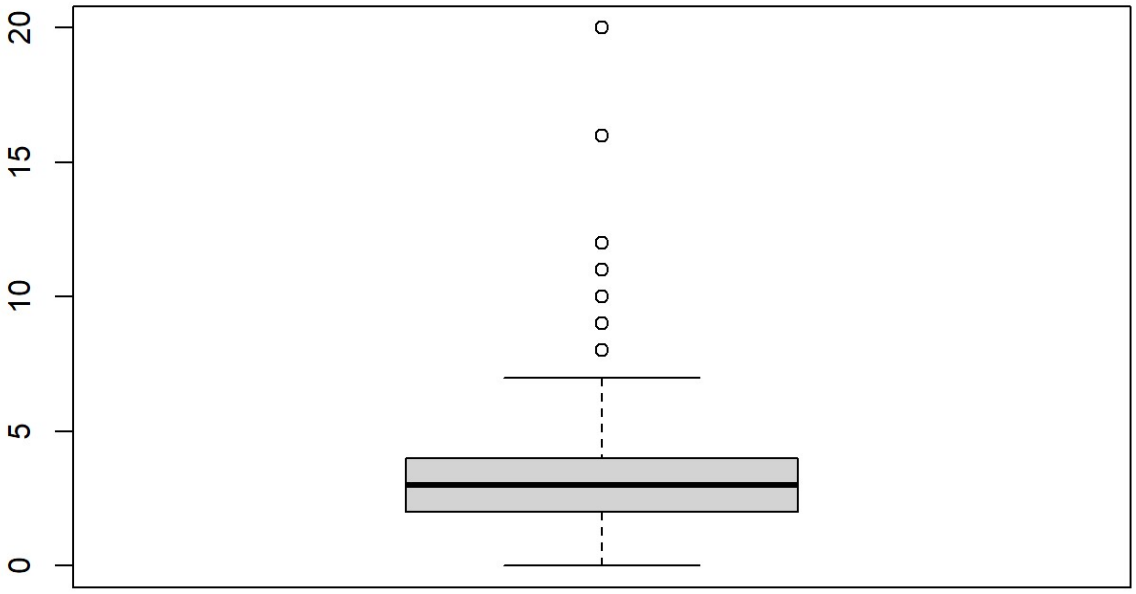

BOXPLOT OF LANDSIZE OF HOUSES BY REGION
Region_Name
Victoria
ropolitan
ropolitan
ropolitan
Victoria
ropolitan
Victoria
ropolitan
#N/A
0
100000
200000
300000
400000
Landsize of houses

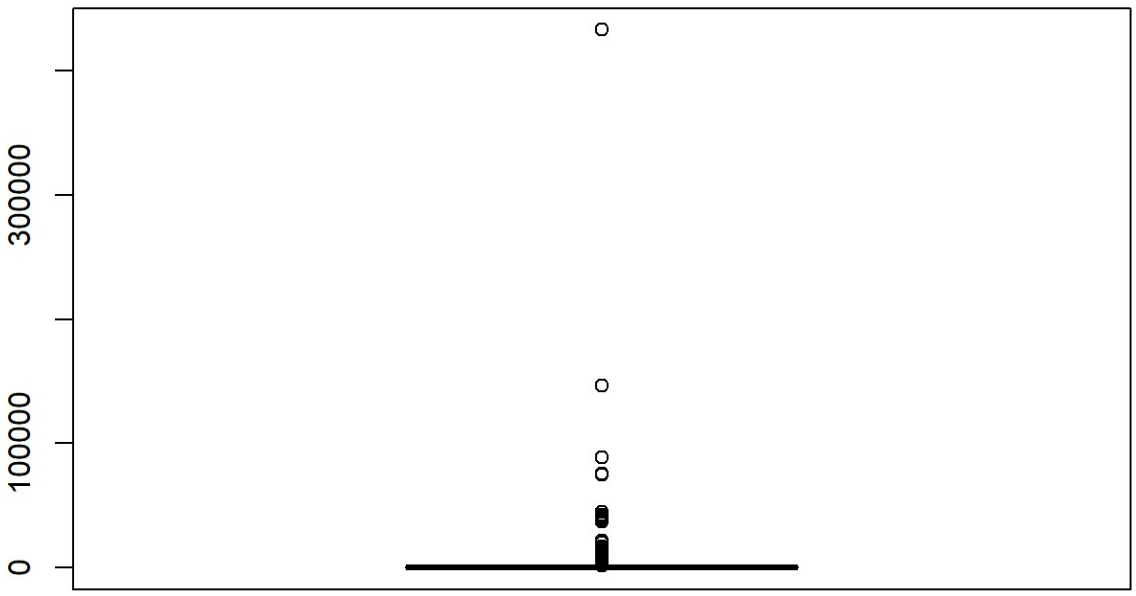

Boxplot of distance from CBD vs type of houses
Type of House
u
t
h
0
10
20
30
40
Distance from CBD

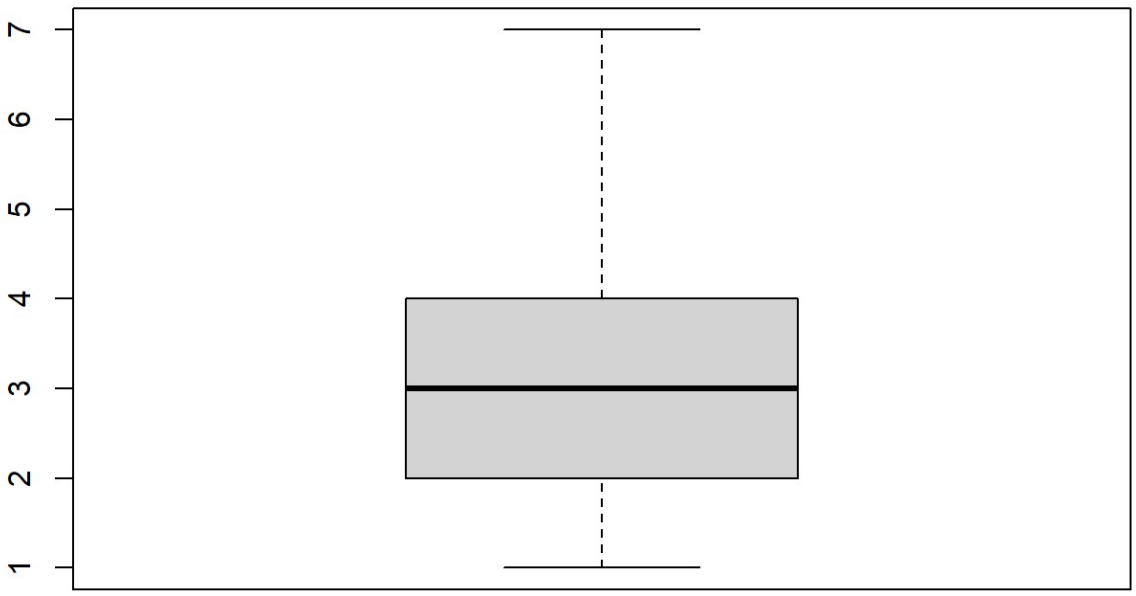

PRICE OF HOUSES.
0
2000000
6000000
10000000

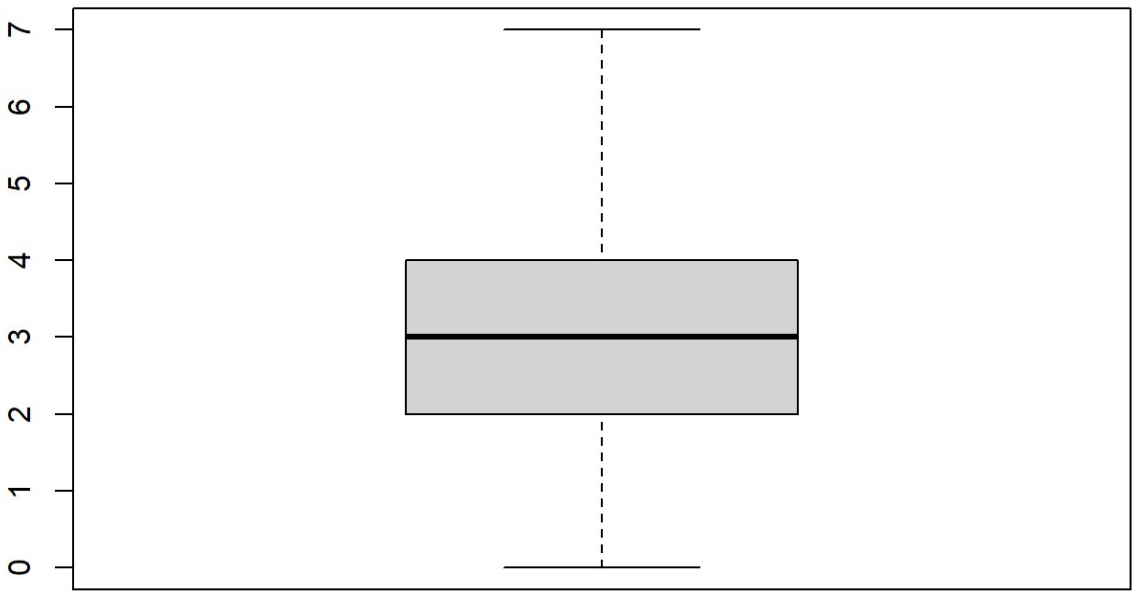

NUMBER OF ROOMS.
2
4
6
8
10
12

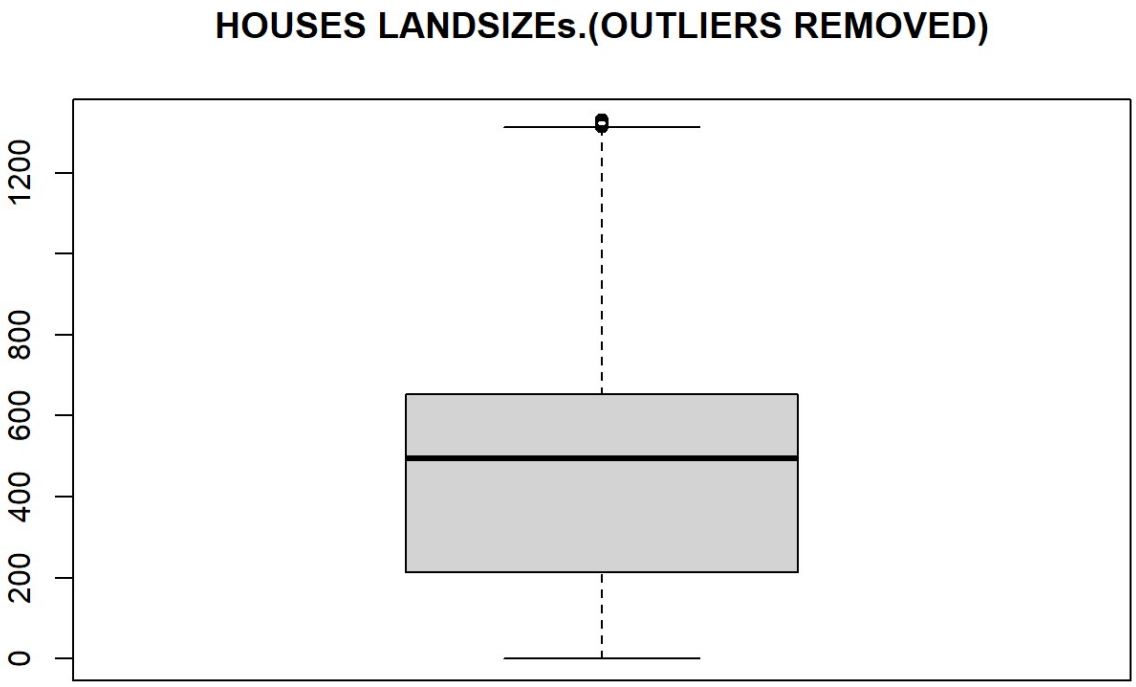
NUMBER OF BEDROOMS.




LANDSIZE OF HOUSES.
0
100000
300000


NUMBER OF ROOMS.(OUTLIERS REMOVED)
7
6
5
4
3
2
1


NUMBER OF BEDROOMS.(OUTLIERS REMOVED)
0
1
2
3
4
5
6
7



Columns Fixation

Landsize

Since the number of rooms and the landsize are dependent, the missing values of landsize can be assumed from the numbers of rooms in the house. The average room size taken is 120 metre square to fill into the NA columns.

Car

Median will bes used to replace NA in car column.

Year

Since, the year built is unknown, 0 will replaced NA as it will be converted to numeric.

Council Area

The rows that has NA values would be eliminated from the datasets.

```
##        Suburb       Address         Rooms          Type         Price
##             0             0             0             0             0
##        Method       SellerG          Date      Distance      Postcode
##             0             0             0             0             0
##      Bedroom2      Bathroom           Car      Landsize     YearBuilt
##             0             0             0             0             0
##   CouncilArea     Lattitude    Longtitude    Regionname Propertycount
##             0          6247          6247             0             0
```

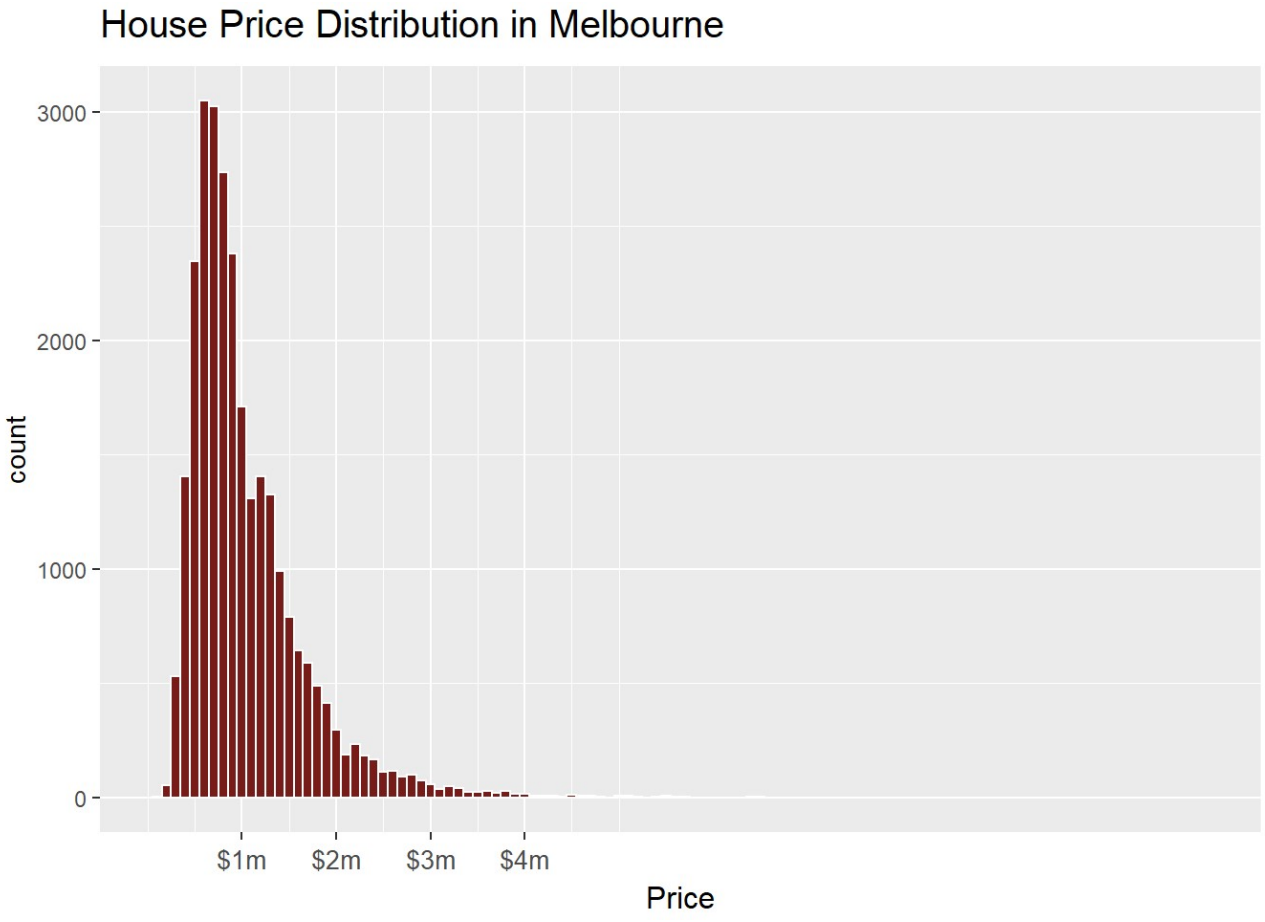
```
## [1] 26763    20
```

Date

Melbourne is located in Australia, which is considered as a four season country. It is needed to take into account for this prediction. The months will be classified according to the season.

Data Analysis

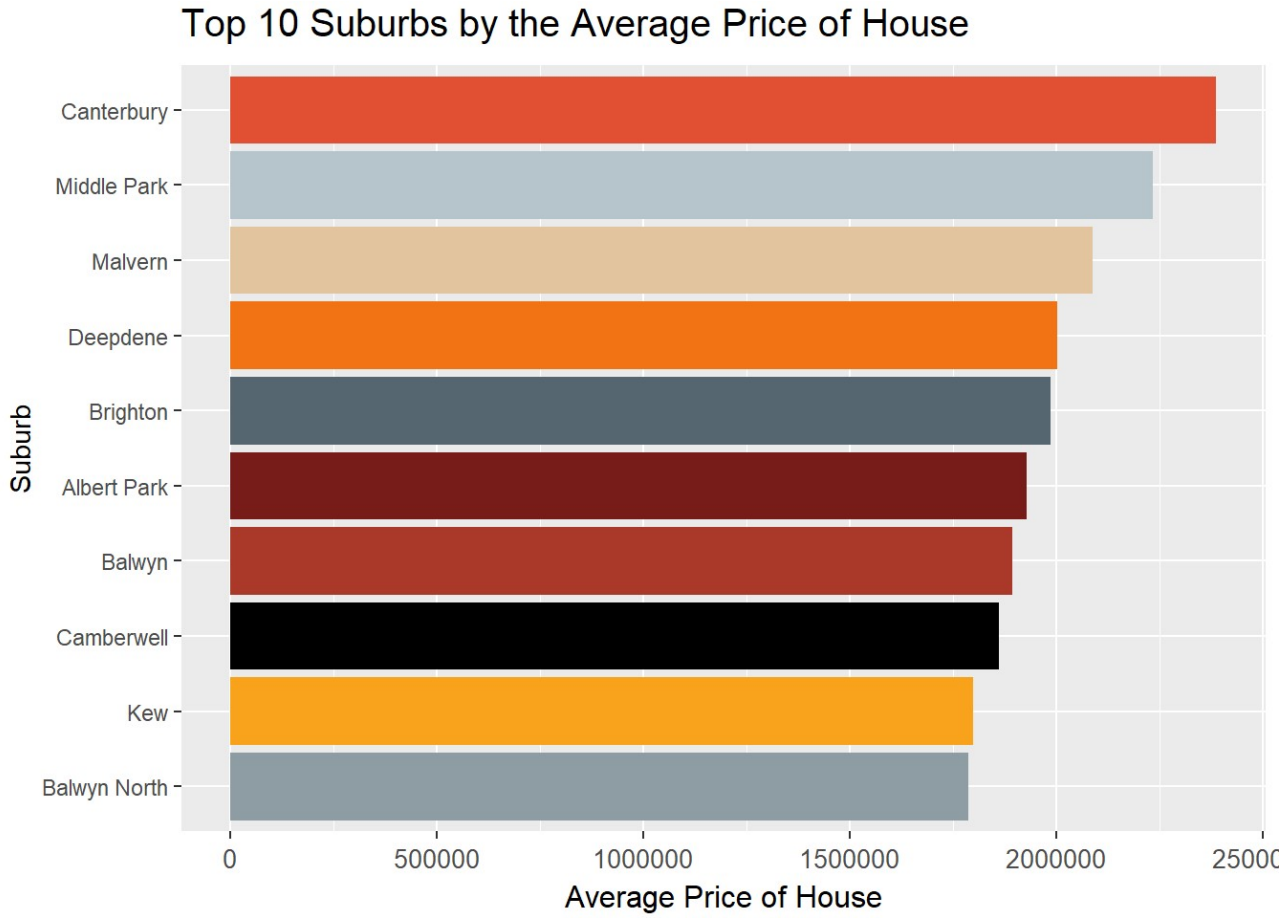
House Price Distribution



Based on the graph above, it is shown that on average, most price ranges below 1 million and very few

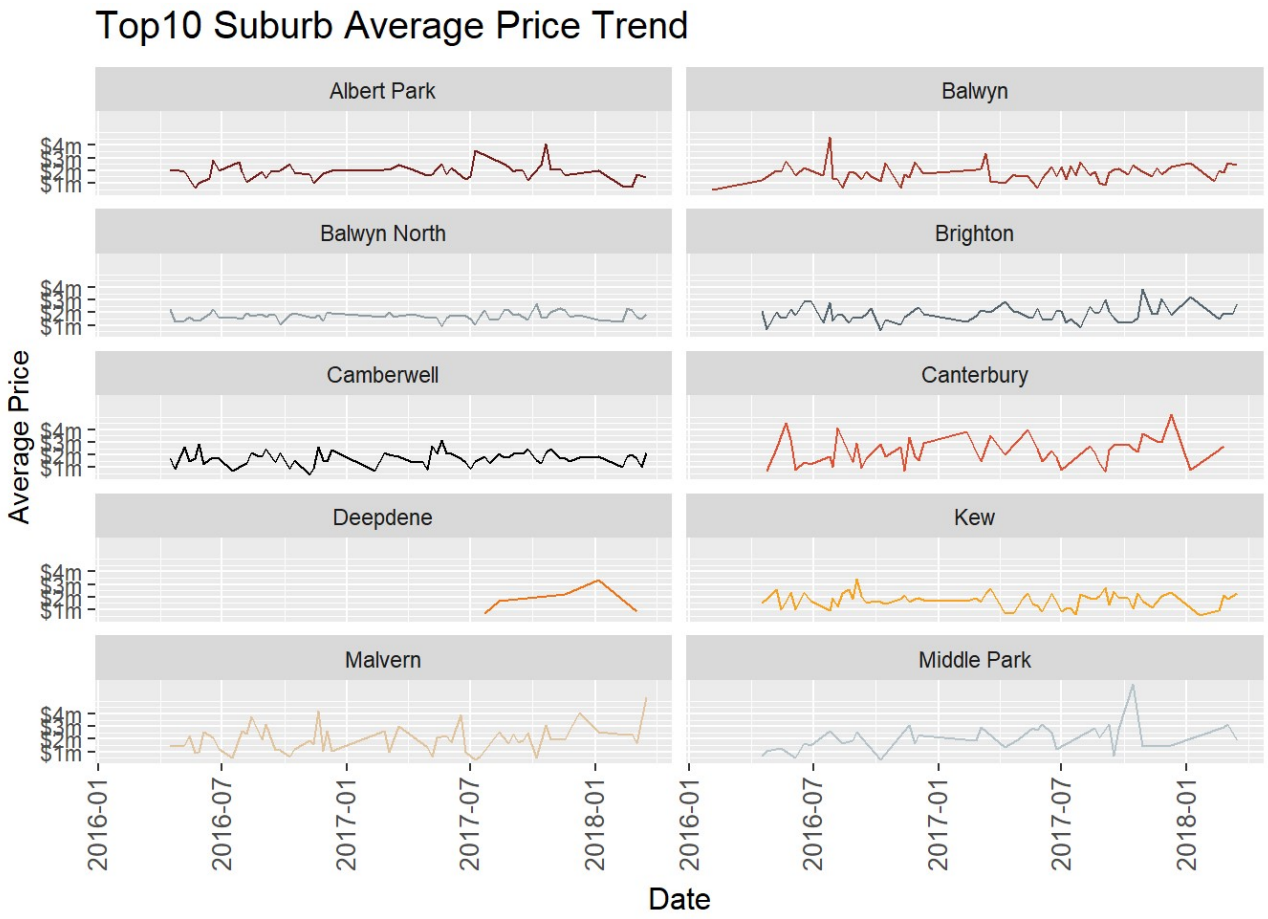
are priced more than 3 millions.

Top 10 Suburbs by the Average House Price



Above is the top 10 suburbs that has the most houses and the rate of average price in the respective areas.

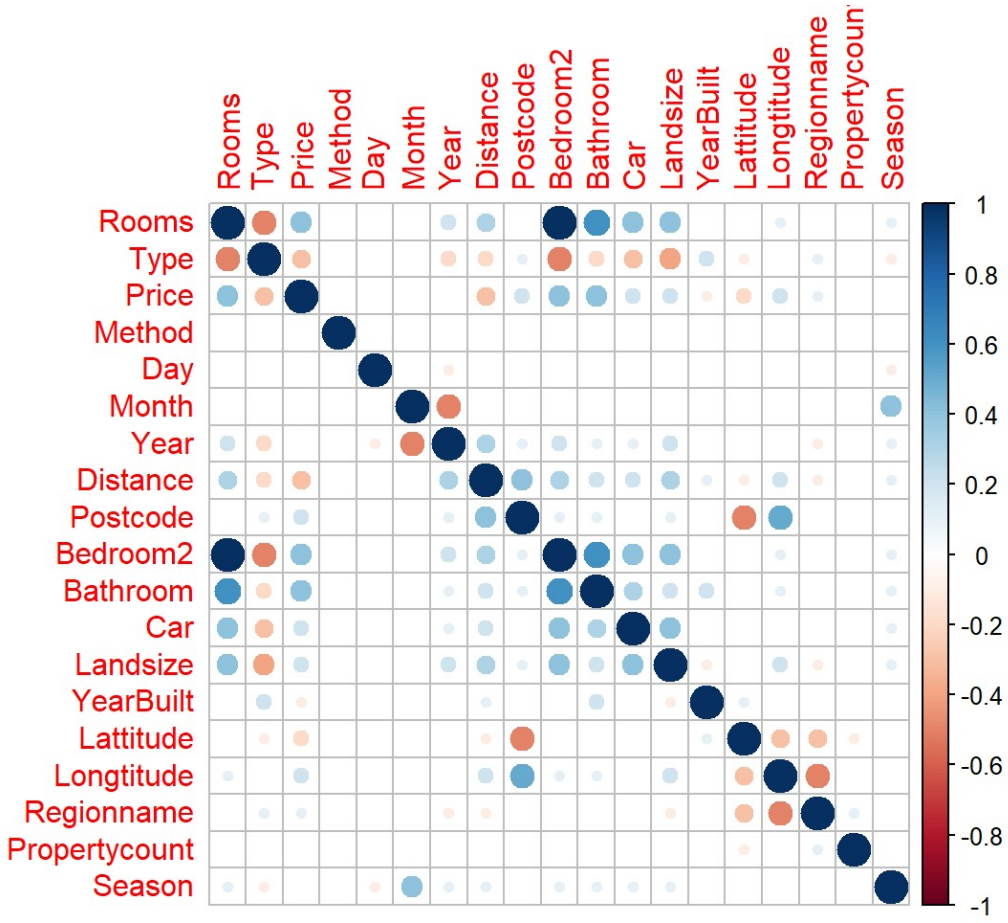
Top 10 Suburbs Price Trends



The graphs depict the house pricing trends of top suburbs with houses. Most suburbs has a fluctuating price trends.However, most of the suburbs maintain the prices range of around 1 to 2 millions and few fluctuate can be very aggresive maybe cause by internal and external factors. The most stable trends is the Balwyn North area and the most fluactuating trend goes to Malvern area.

Data corelation

The plot below depicts the corelation between the attributes.



It can be infered that bedroom, bathroom, car parking lots, landsize, type and price play most of the crucial part in this pricing markets. They will be considered as the major factors to the prediction of this project.

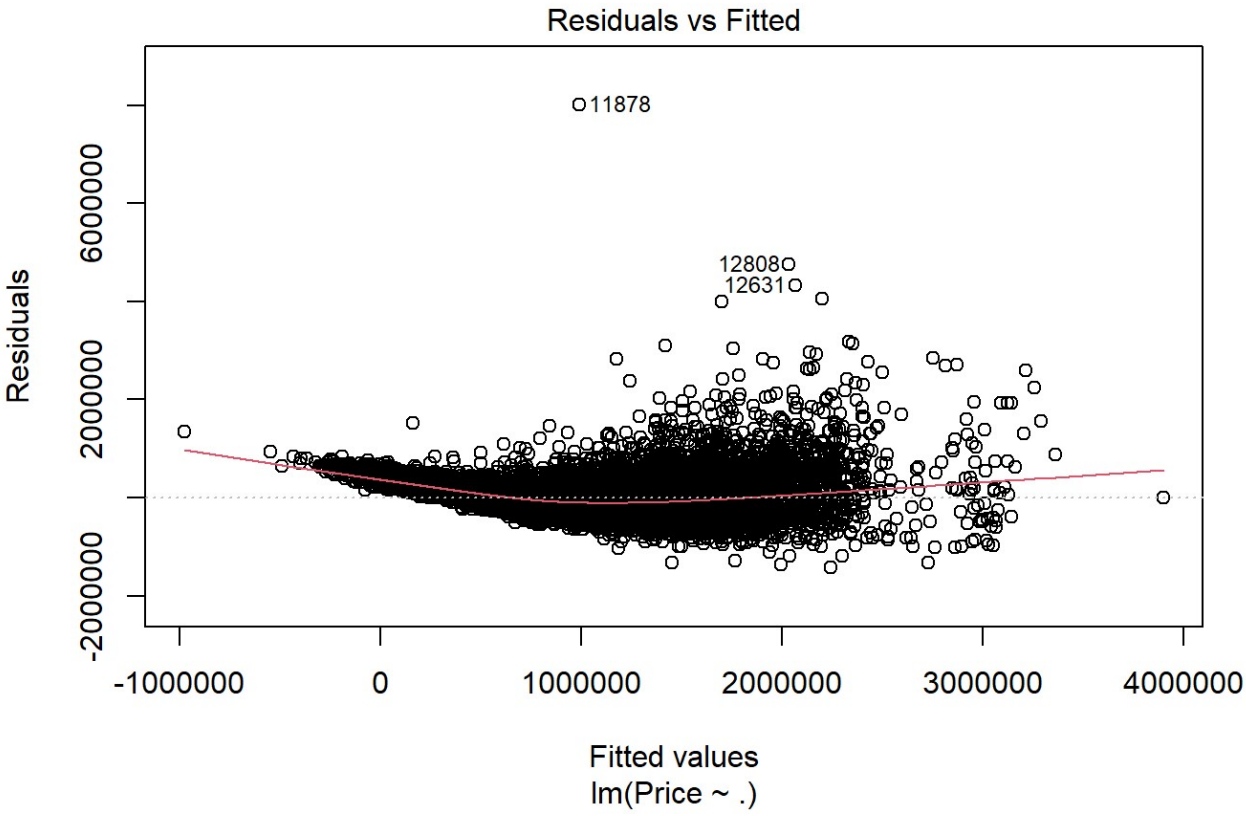
Model Development

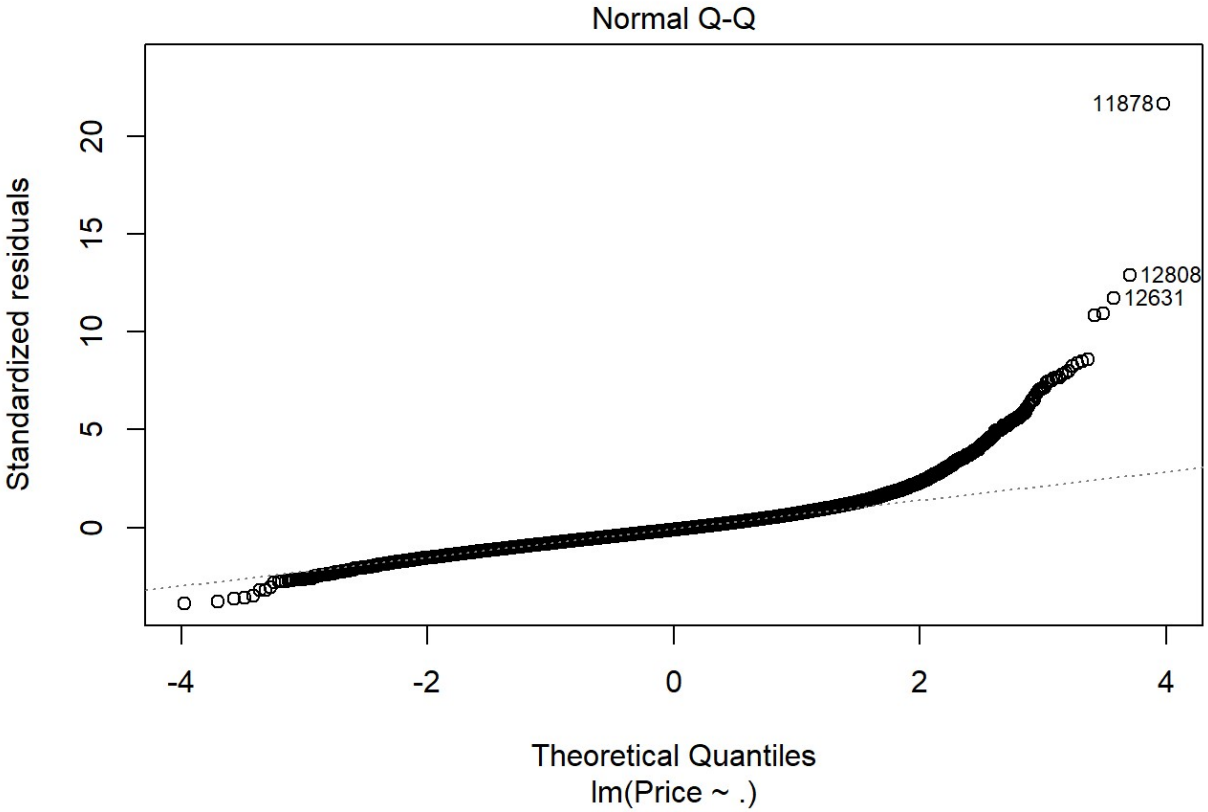
Training & Testing Dataset

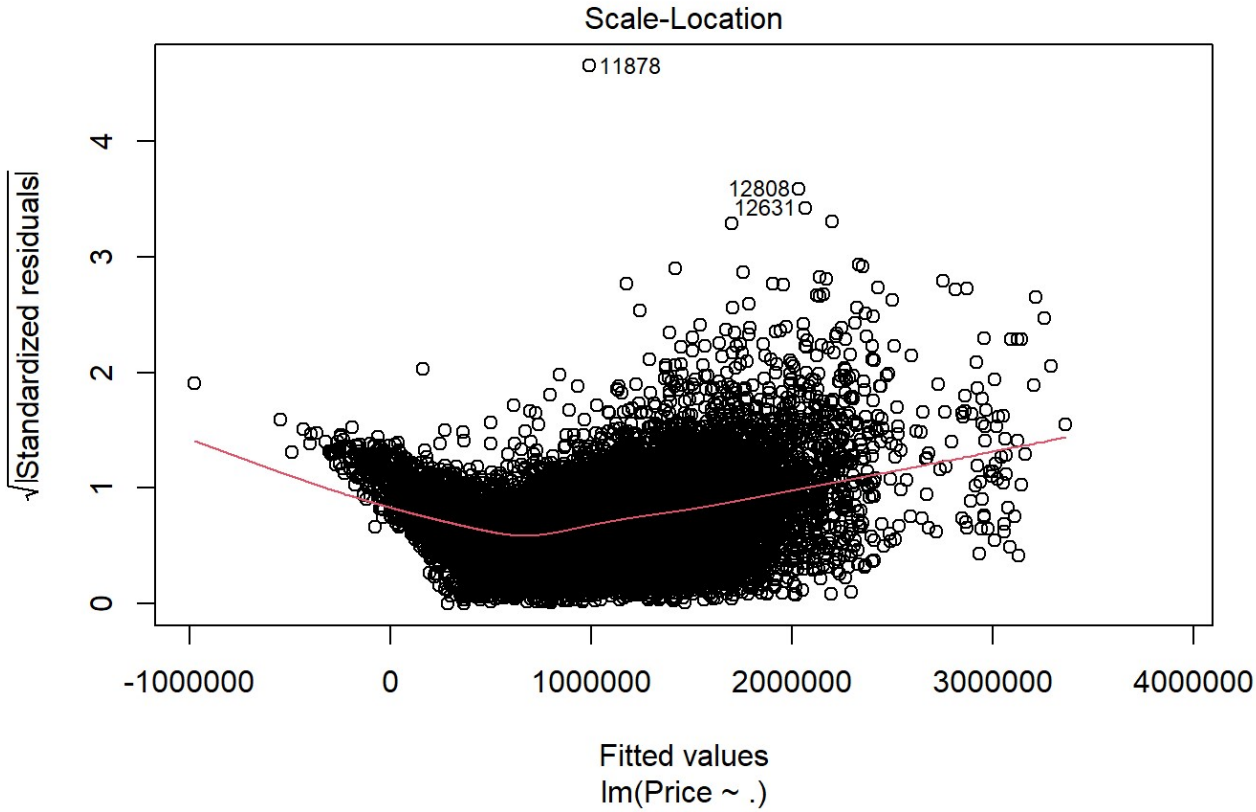
For training and testing data, the rooms column will be removed as it shows a lot of collinearity to bedroom2. Then, the cleaned dataset would be segregated to training set and testing set in the ratio of 70%:30% respectively.

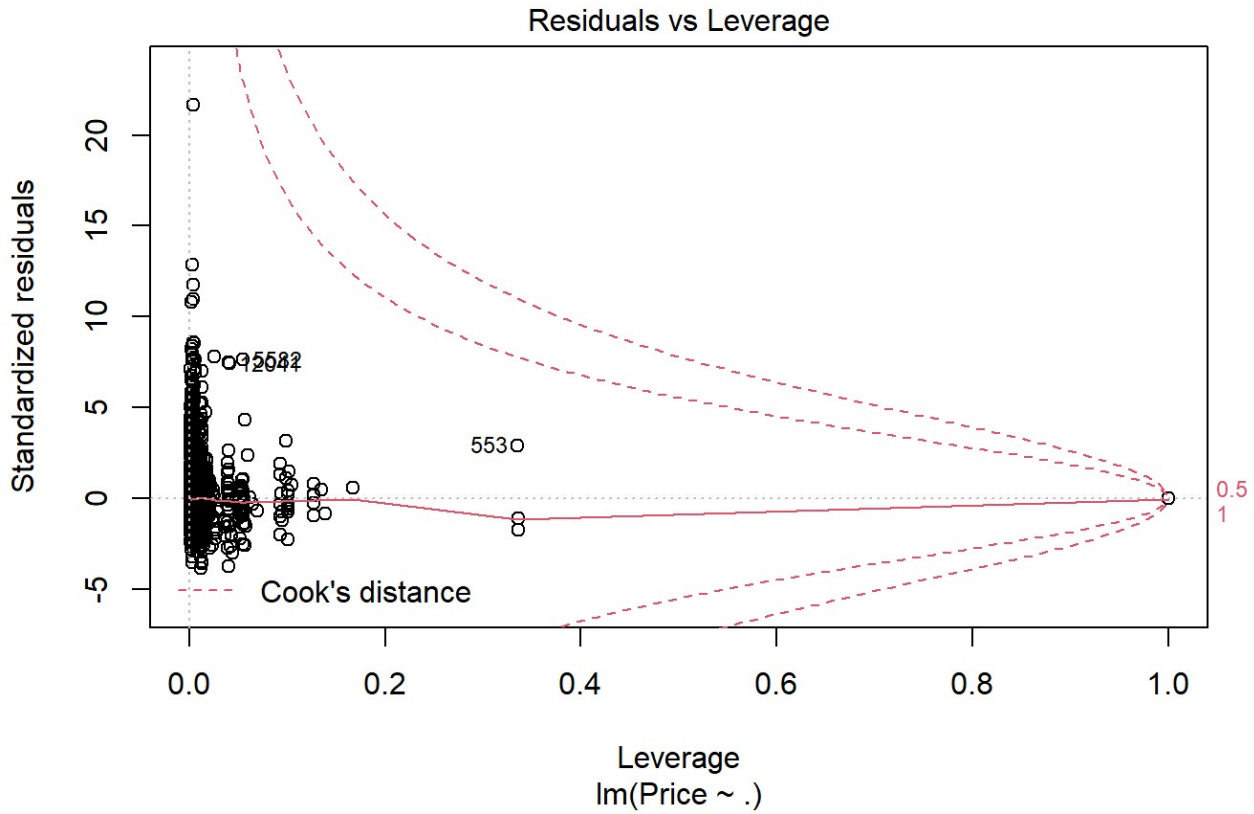
Model Training

Linear Regression




Normal Q-Q
Standardized residuals
0
5
10
15
20
11878
12808
12631
-4
-2
0
2
4
Theoretical Quantiles
lm(Price ~ .)


Scale-Location
11878
12808
12631
√|Standardized residuals|
0
1
2
3
4
-1000000
0
1000000
2000000
3000000
4000000
Fitted values
lm(Price ~ .)



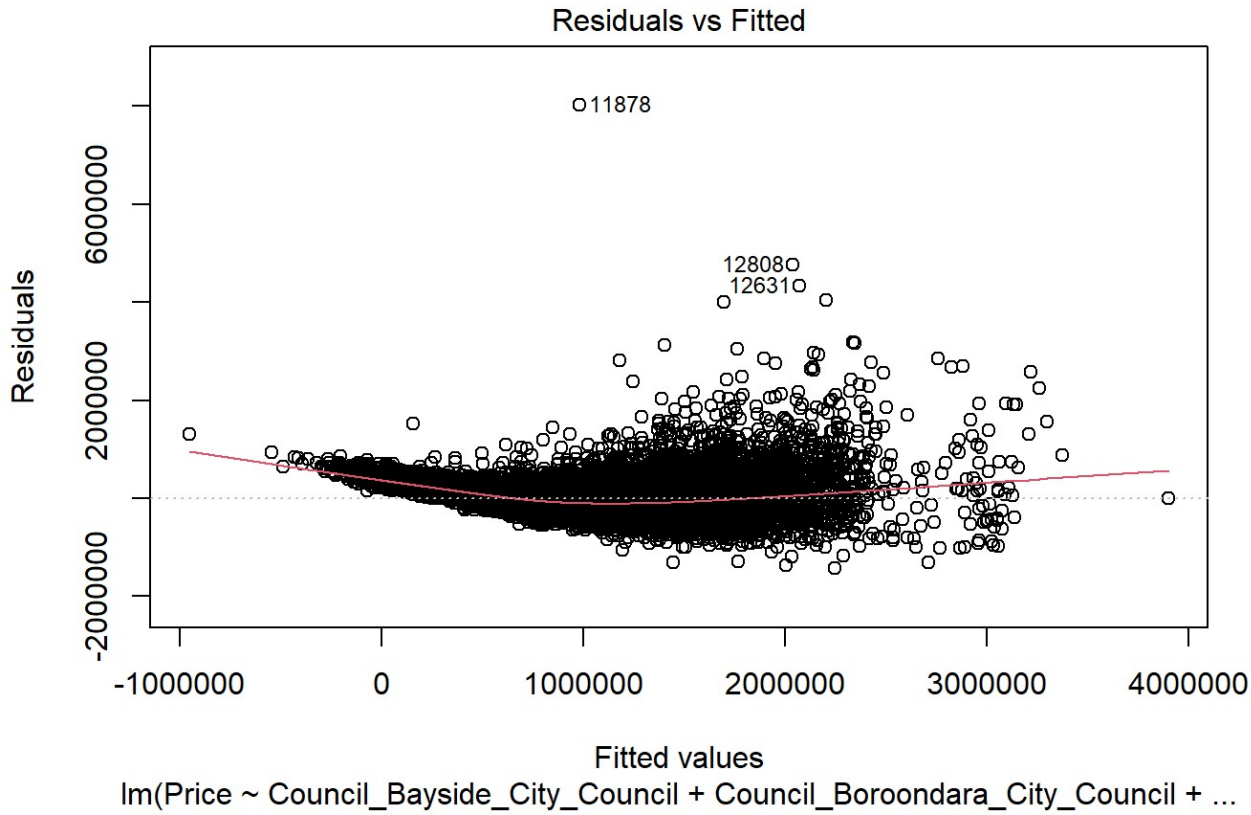
R-Squared

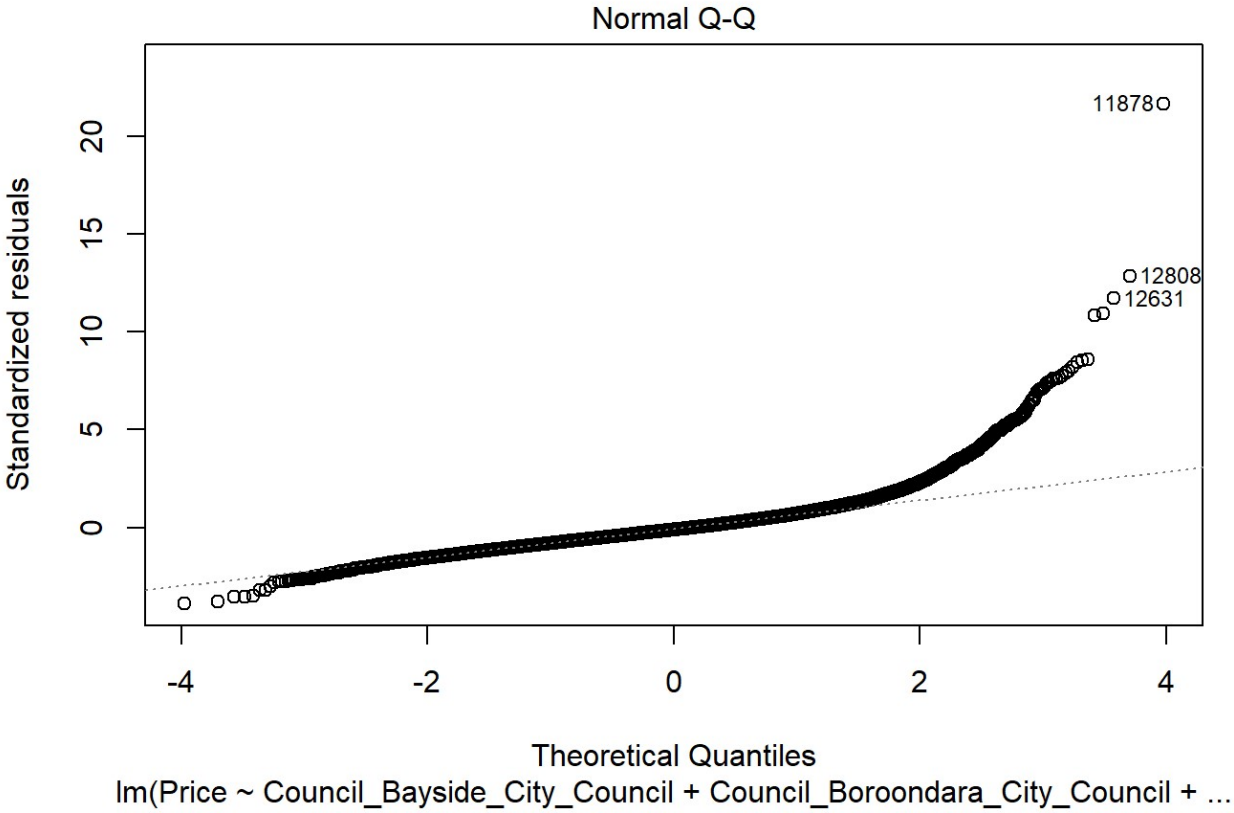
```
## [1] 0.6517716
```

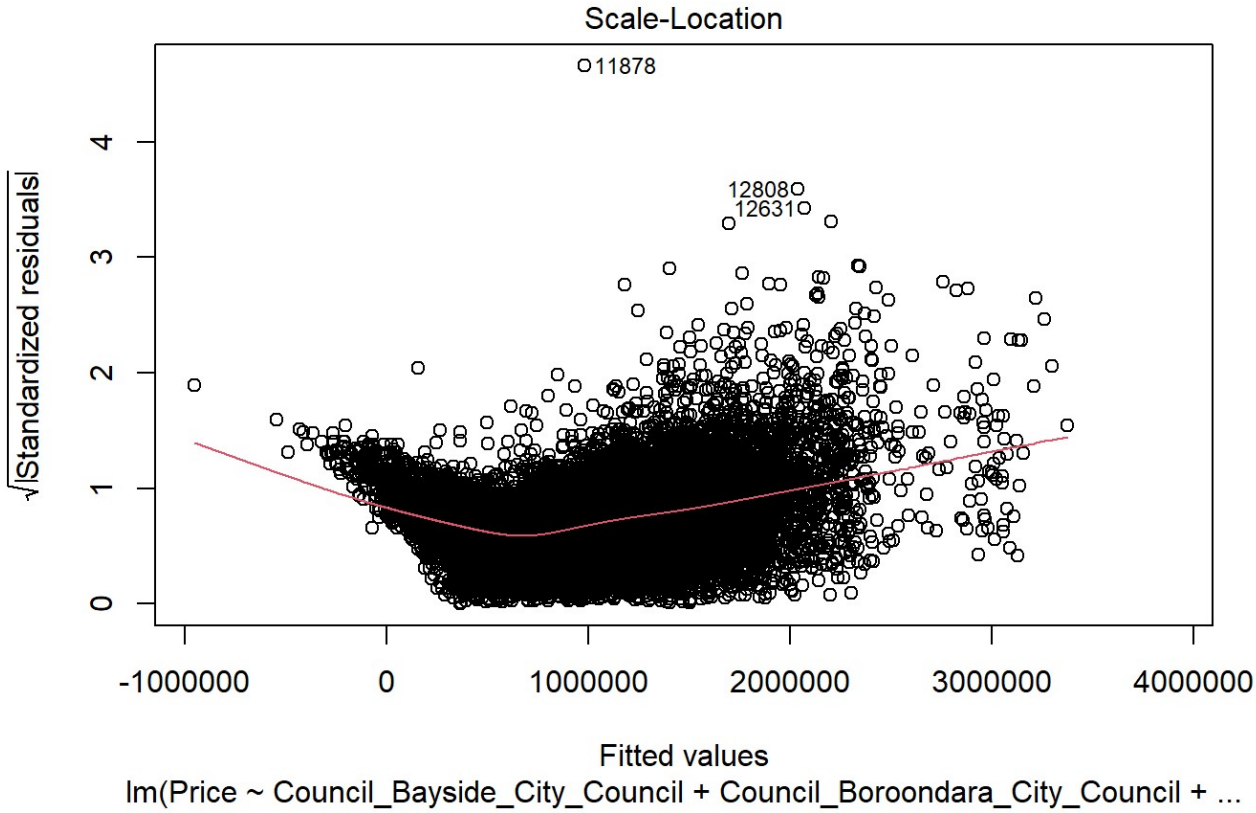
Stepwise Model

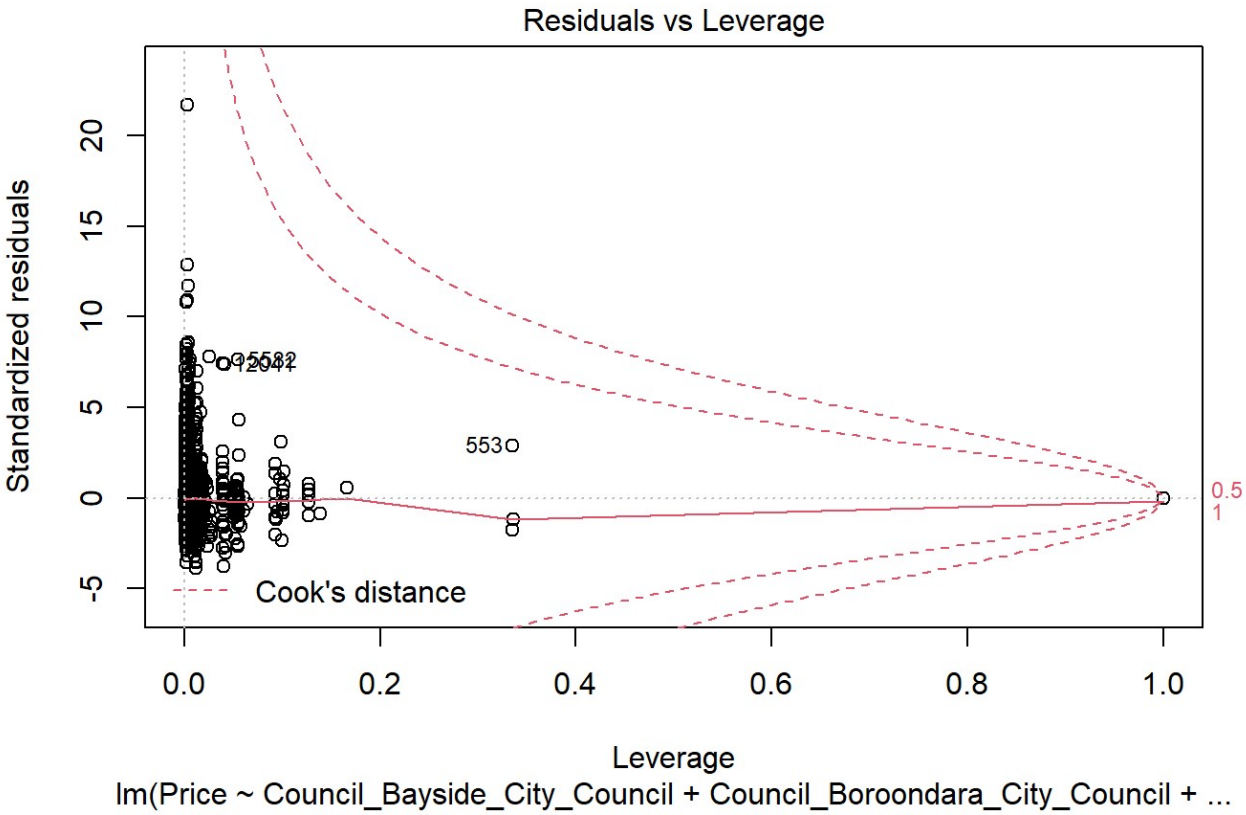
R-Squared

```
## [1] 0.6522785
```

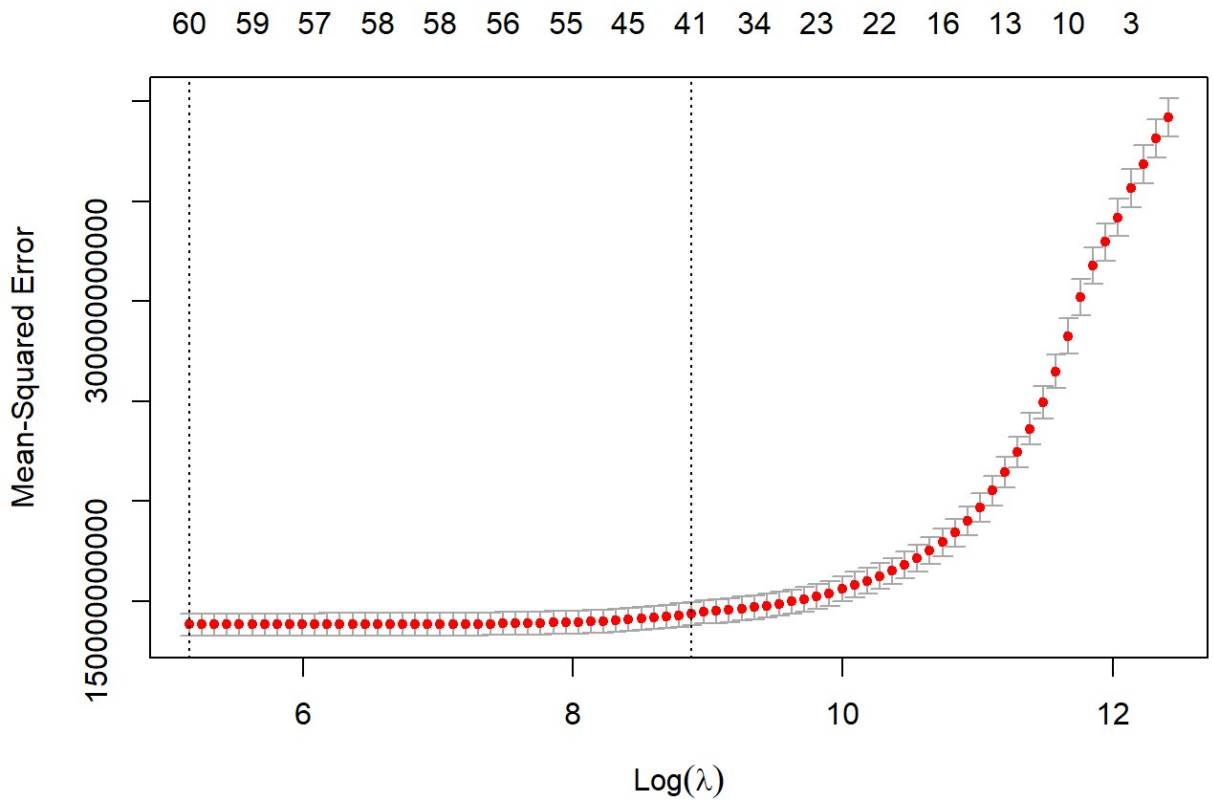

Residuals vs Fitted
11878
12808
12631
Residuals
6000000
2000000
-2000000
-1000000
0
1000000
2000000
3000000
4000000
Fitted values
lm(Price ~ Council_Bayside_City_Council + Council_Boroondara_City_Council + ...


Normal Q-Q
Standardized residuals
0
5
10
15
20
11878
12808
12631
-4
-2
0
2
4
Theoretical Quantiles
lm(Price ~ Council_Bayside_City_Council + Council_Boroondara_City_Council + ...


Scale-Location
11878
12808
12631
√|Standardized residuals|
0
1
2
3
4
-1000000
0
1000000
2000000
3000000
4000000
Fitted values
lm(Price ~ Council_Bayside_City_Council + Council_Boroondara_City_Council + ...


Residuals vs Leverage
Standardized residuals
20
15
10
5
0
-5
12041
553
0.5
1
Cook's distance
0.0
0.2
0.4
0.6
0.8
1.0
Leverage
lm(Price ~ Council_Bayside_City_Council + Council_Boroondara_City_Council + ...

Lasso Regression

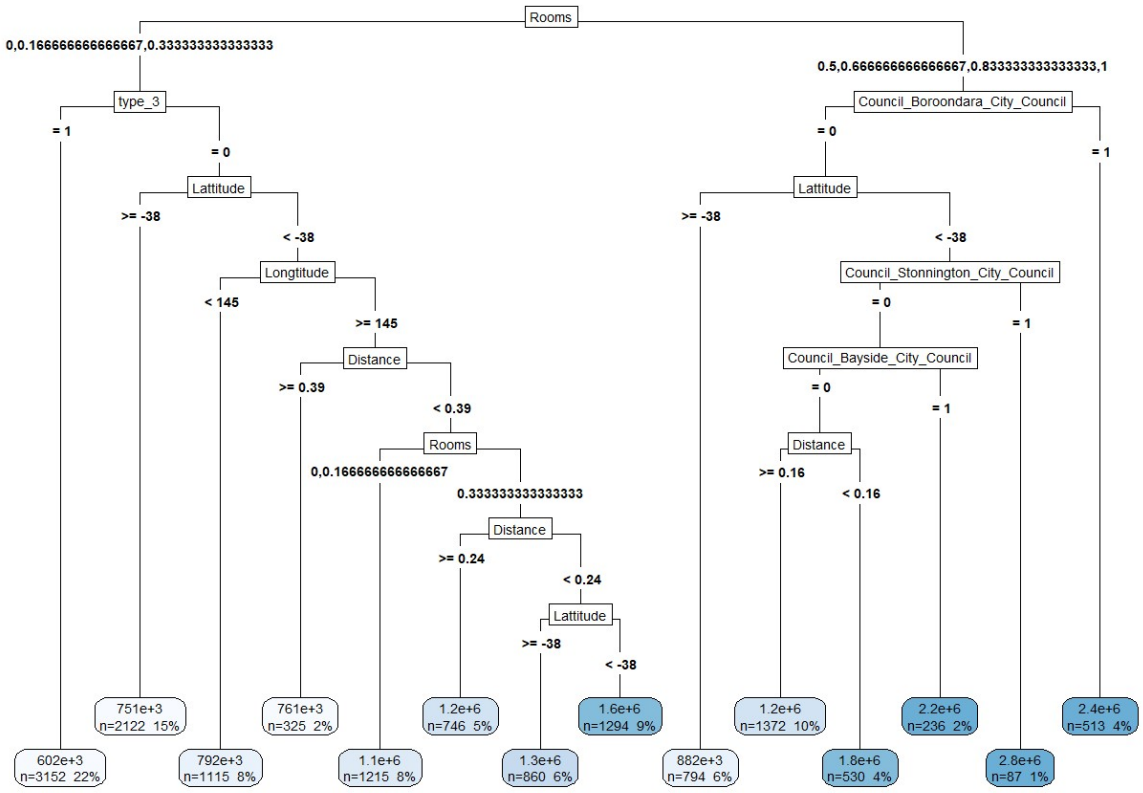



```
## [1] 173.6668
```

R-Squared

```
## [1] 0.6512648
```

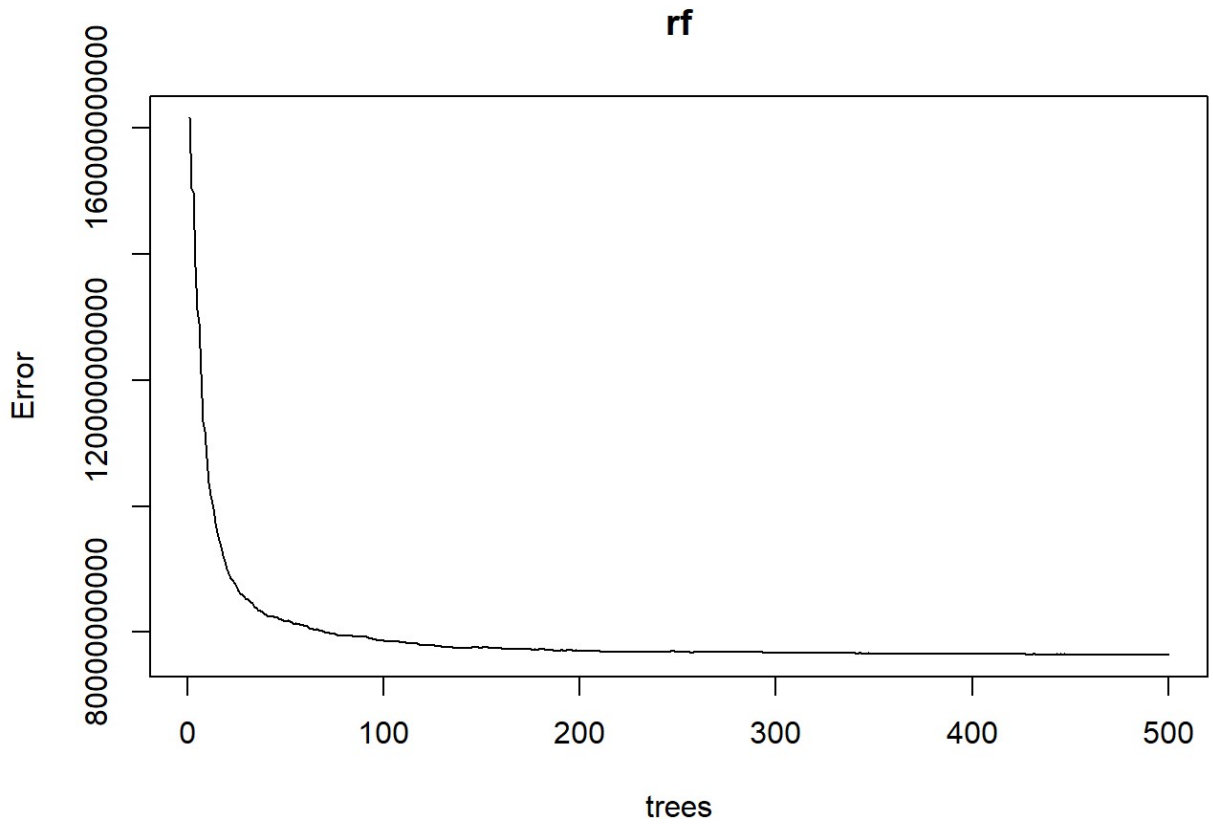
Decision Tree



R-Squared

```
## [1] 0.5631654
```

Random Forest



R-Squared

```
## [1] 0.8096577
```

3.Result

R-Squared

Evaluating models with R-Squared requires the observation of the highest value to 1. The higher the value, the more accurate the model is.

Random Forest

From the project, it can be seen that Random Forest has the highest value among other models, so it can be assumed and taken as the most accurate among other models. The comparison of the values is shown below.

method	r2	sm.r2	lm.r2	dt.r2	r2.rf

method	r2	sm.r2	lm.r2	dt.r2	r2.rf
Linear Regression	0.6517716	NA	NA	NA	NA
Stepwise Model	NA	0.6522785	NA	NA	NA
Lasso Regression	NA	NA	0.6512648	NA	NA
Decision Tree	NA	NA	NA	0.5631654	NA
Random Forest	NA	NA	NA	NA	0.8096577

Below is the summary of the predicted price from the random forest models.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  327179  691861  955172 1089792 1337421 4518275
```

4.Conclusion

Melbourne Housing Price dataset is a huge dataset that needed extra works needed to put into cleaning, preprocessing, analysing and prediction modelling. However, this project is managed to be done despite the constraints that has occured in it.

The main attribute or variable that is taken into account of the prediction is the Price attribute, few others attributes that may affect the prediction such as the landsize, room, car lot, bathroom and type are used as relative variables for the model building. The most optimum model for this price prediction is the Random Forest model, which has the highest R2, 0.810, compared to other models. As for this projects, a constraint has occured, the dataset has a lot of missing values which results less data for less accuracy.For future works, a more comprehensive dataset would do a better job and the uses of the common algorithm may has lower prediction accuracy, by ensembling differents algorithm a more accurate result can be acquired.