A house that has a sign on the side of a building

Description generated with very high confidence

**House Price Prediction**

***“Multiple Linear Regression, Time Series Forecasting, Hypothesis Testing and ANOVA”***

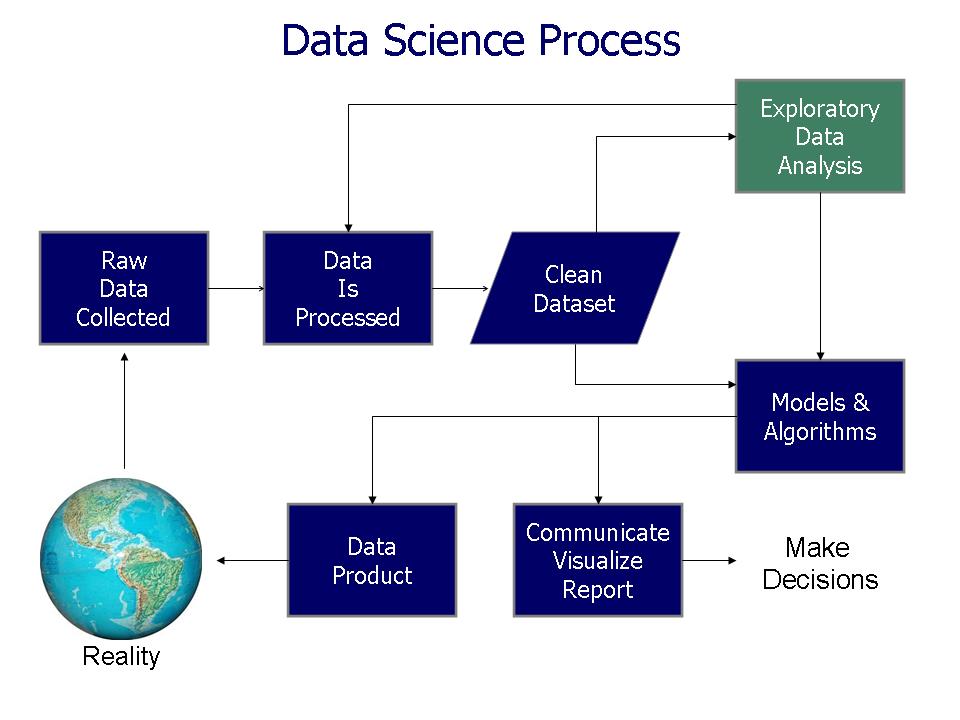
**Shephalika Shekhar**

**Table of Contents:**

|  |  |  |
| --- | --- | --- |
| ***Sr No*** | ***Topic*** | ***Page No*** |
| 1 | Introduction | 2 |
| 2 | Data | 2-3 |
| 3 | Problems to be solved | 4 |
| 4 | Data Preprocessing | 4-8 |
| 5 | Methods and Process | 8-31 |
| 6 | Evaluations and Predictions | 32-33 |
| 7 | Conclusion & Future Work | 33 |

1. **Introduction:**

If we ask a person going to buy a house to describe their dream house, then there would be many factors they would tell such as the number the bedrooms, condition and grade of the house, number of bathrooms etc. But there are many more factors involved for predicting the price of houses. A person won't probably begin with the height of the basement ceiling or the proximity to the roads. Here the dataset consists of many factors that can be useful for predicting house prices. There are many steps involved to make accurate predictions on the dataset which includes data preprocessing, exploratory data analysis then model building, evaluating the model based on their accuracy and then make predictions.



1. **Data**

The data set consists of many numerical as well as categorical variables that might have an affect on house price prediction which can be determined by model building using Multiple Linear Regression. Also, there is “date” variable through which we can also use time series analysis and forecasting methods to see if we can predict price in the future based on the past data.

Data has been taken from: <https://www.kaggle.com/harlfoxem/housesalesprediction>

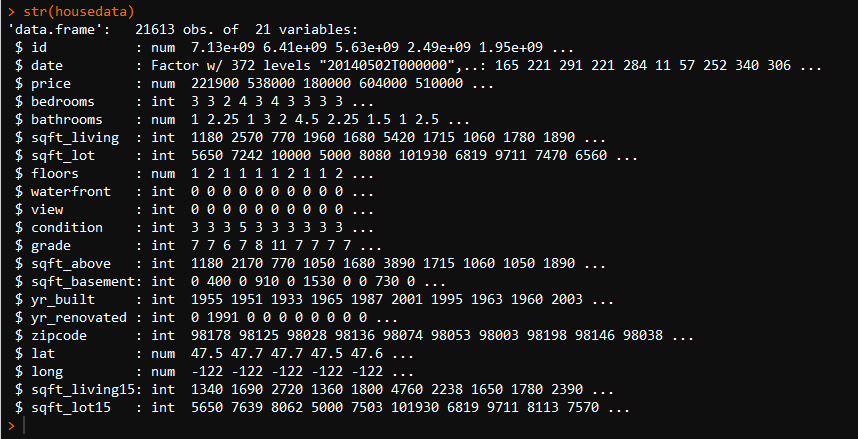
The dataset consists of house sale prices (21000 rows) with many factors associated to it.

Raw data was collected, and it consists on many variables – Dependent variable: Price

Independent Variables that affect price:

1. Id: numeric
2. Date: converted to numeric
3. Bedrooms: categorical
4. Bathrooms: categorical
5. Sqft\_living: numeric
6. Sqft\_lot: numeric
7. Floors: categorical
8. Waterfront: categorical
9. View: categorical
10. Condition: categorical
11. Grade: categorical
12. Sqft\_above: numeric
13. Sqft\_basement: numeric
14. Yr\_built: numeric
15. Yr\_renovated: numeric
16. Zipcode: categorical
17. Lat: numeric
18. Long: numeric
19. Sqft\_living15: numeric
20. Sqft\_lot15: numeric

Below screenshot shows the data types of the variables of the data set.

****

1. **Problems to be Solved**

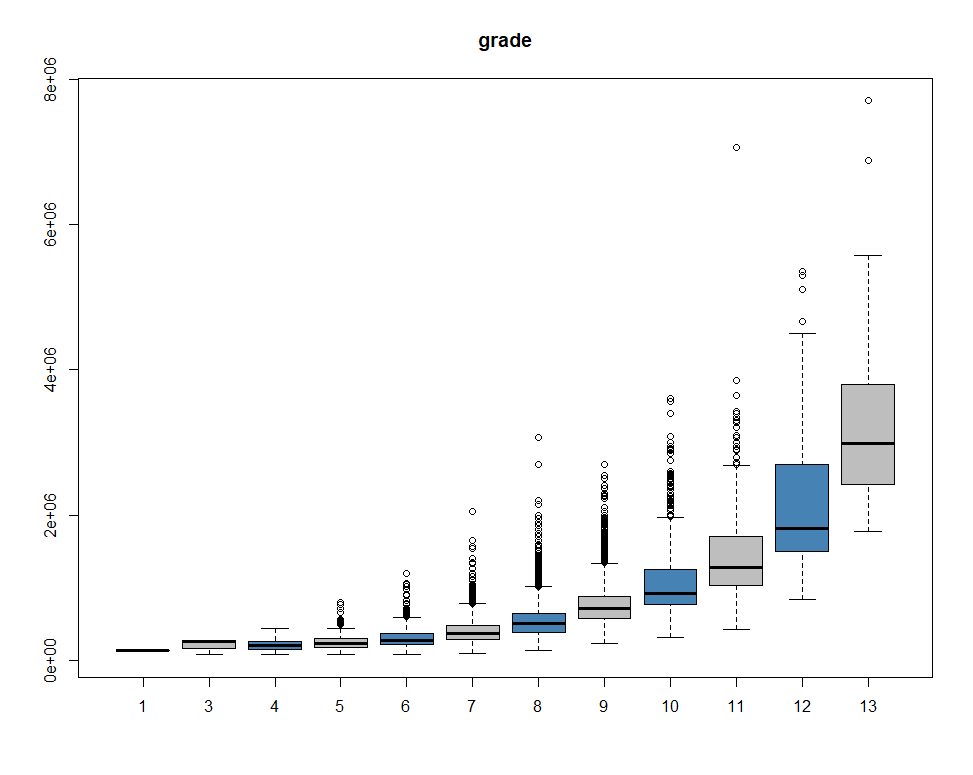
The goal is to determine which factors that influence the house price the most and may be useful to predict the house sales. Also, to see if the house sales depends on the past data based on the date variable that is present in the dataset. These problems are to be solved by building various models and applying algorithms and analysis to determine the best factors and future values of the price.

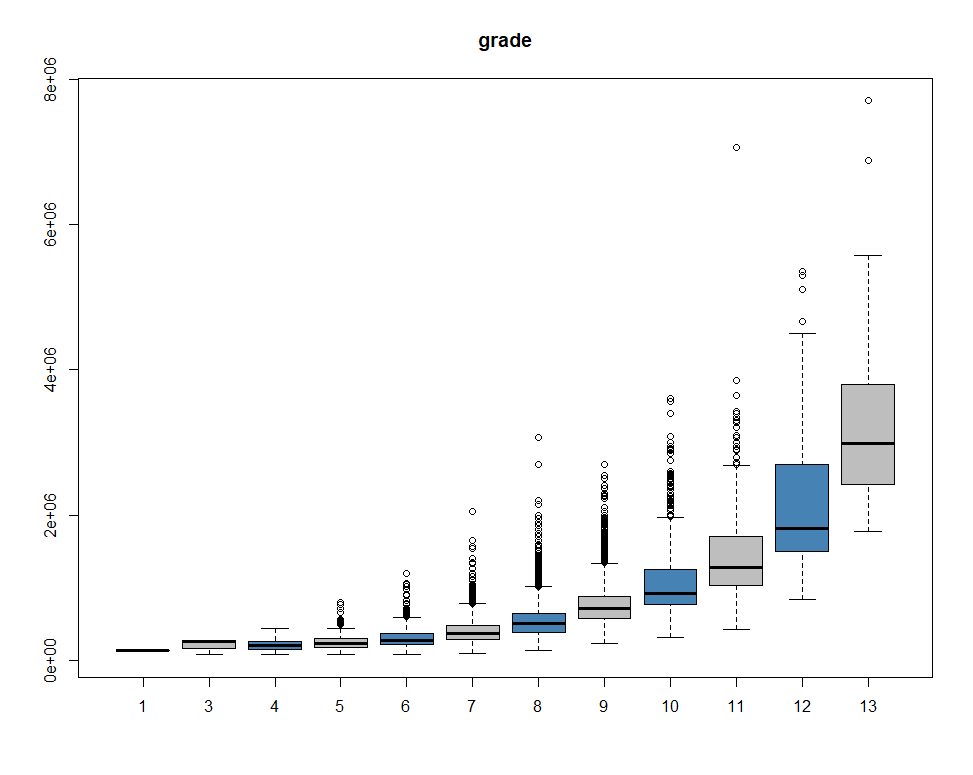
1. **Data Processing**

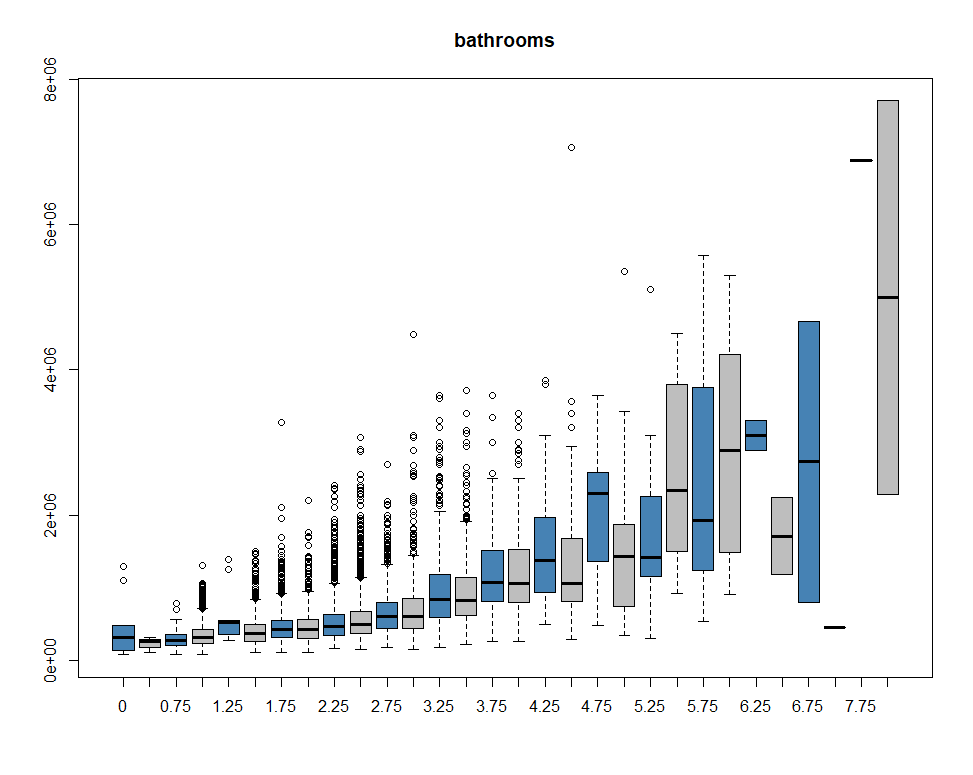
There we no missing values in the data. Bedrooms variable had an outlier which has been rectified before building the model. The dataset has been divided into training and testing data set. Date was converted to numeric variable for multiple linear regression while Date was converted to date format as required for time series analysis. Below are the box plots for independent variables vs Price which show that these are categorical variables and how they vary with price.

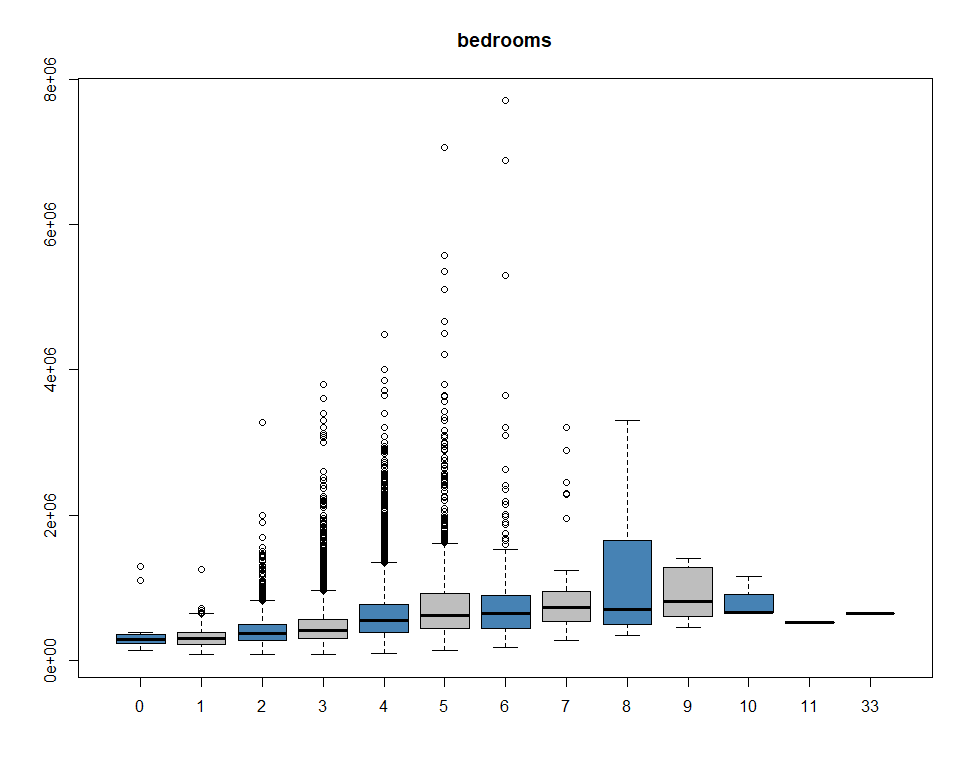
A close up of a map

Description generated with high confidence









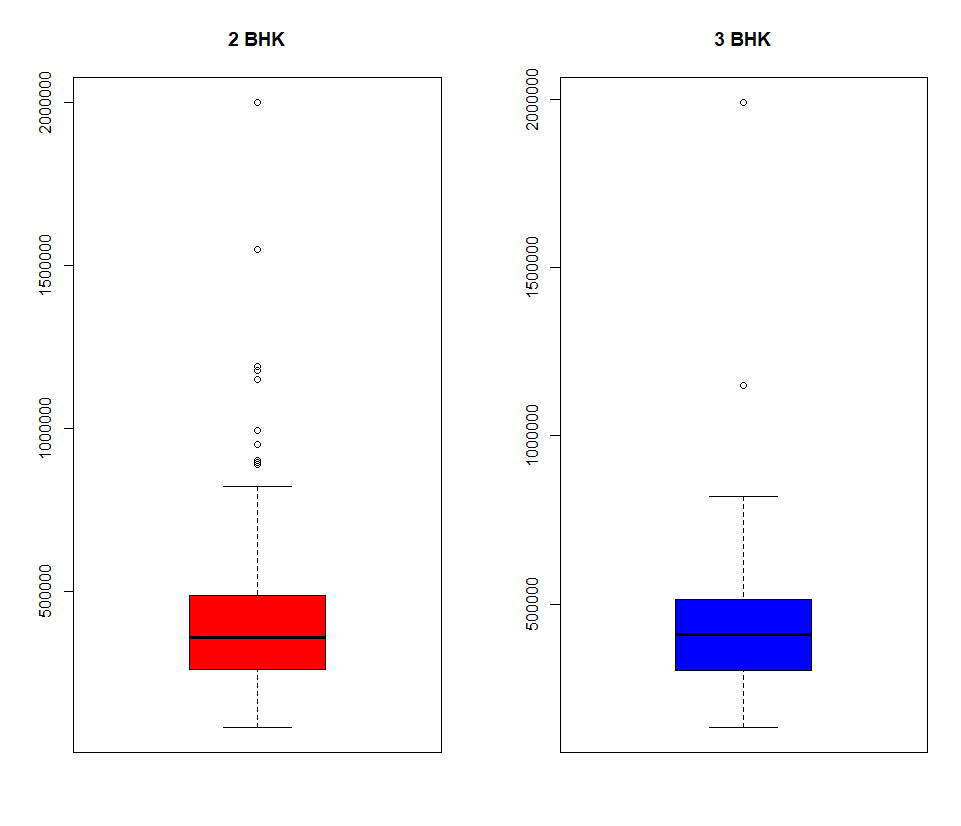
**Two sample One-tailed Hypothesis Testing:**

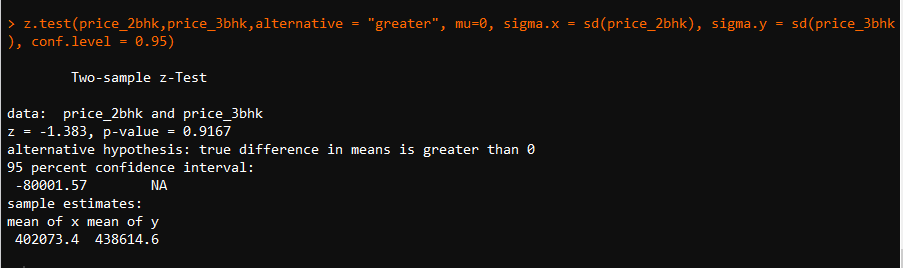
Null Hypothesis: Mean price of 2 BHK houses is less than 3 BHK houses

Alternative Hypothesis: Mean price of 2 BHK houses is greater than 3 BHK houses.

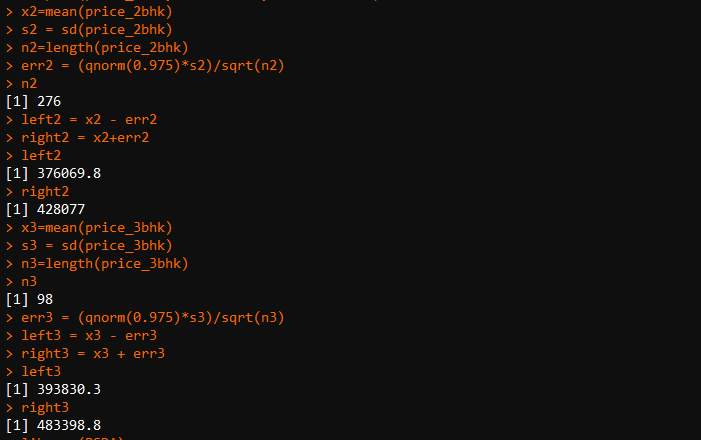
To check hypothesis, z test has been done where p value is greater than 0.05. Thus, at 95% confidence interval, there is enough evidence to accept null hypothesis.

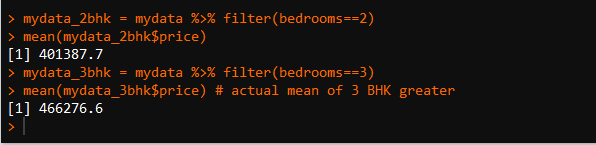
**This test can be modified by 2BHK houses in a particular downtown area to be compared with 3BHK houses in another suburb area: limitations due to no such variable present in data set.**

****

****

**Confidence Intervals for 2BHK and 3BHK Houses –  
Taken random samples from main dataset and this test data produces the confidence interval.**





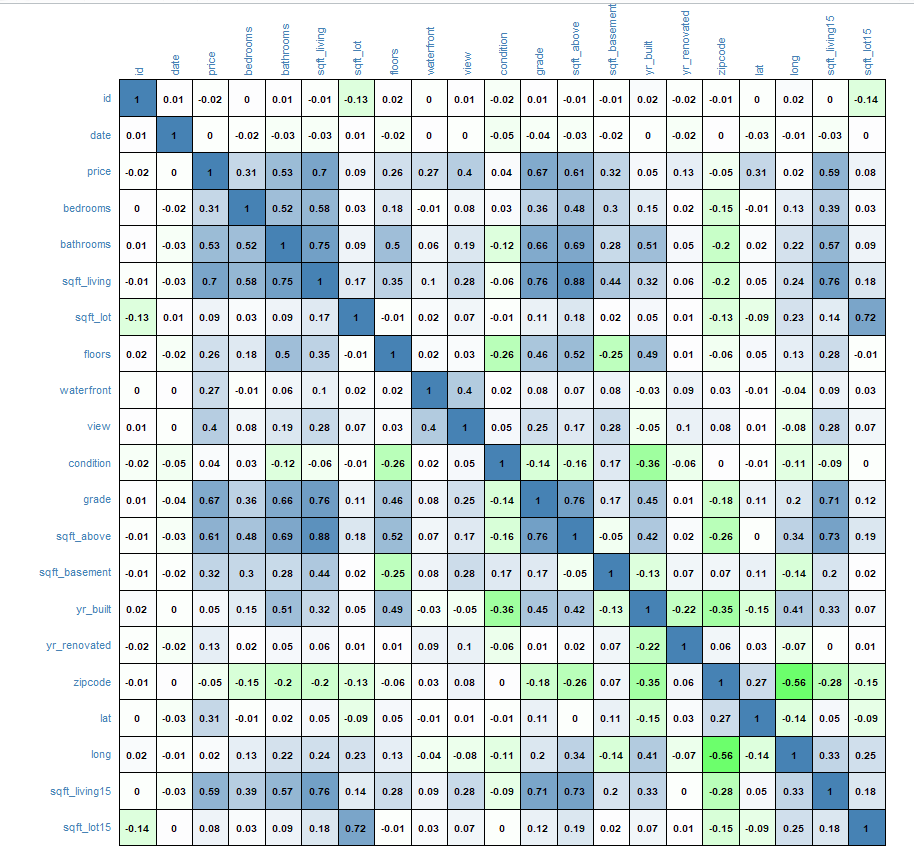
1. **Methods and Processes:**

**Multiple Linear Regression:**

***Steps Involved -***

1. Load the dataset and libraries in R
2. Divide the data set into training and testing
3. Checking the linear association between variables
4. Multicollinearity exists between variables, but they will taken care later after model building and calculating VIF
5. Convert the categorical variables converted to factors and then to be used in model.
6. Build different regression models using training data set and find the best model based on RMSE value.
7. Selection of different models based on different metrics like AIC, BIC and p value
8. Residual analysis
9. Make predictions on test dataset and find the accuracy

**Checking correlation values among the variables - Below is the correlation plot of data:**



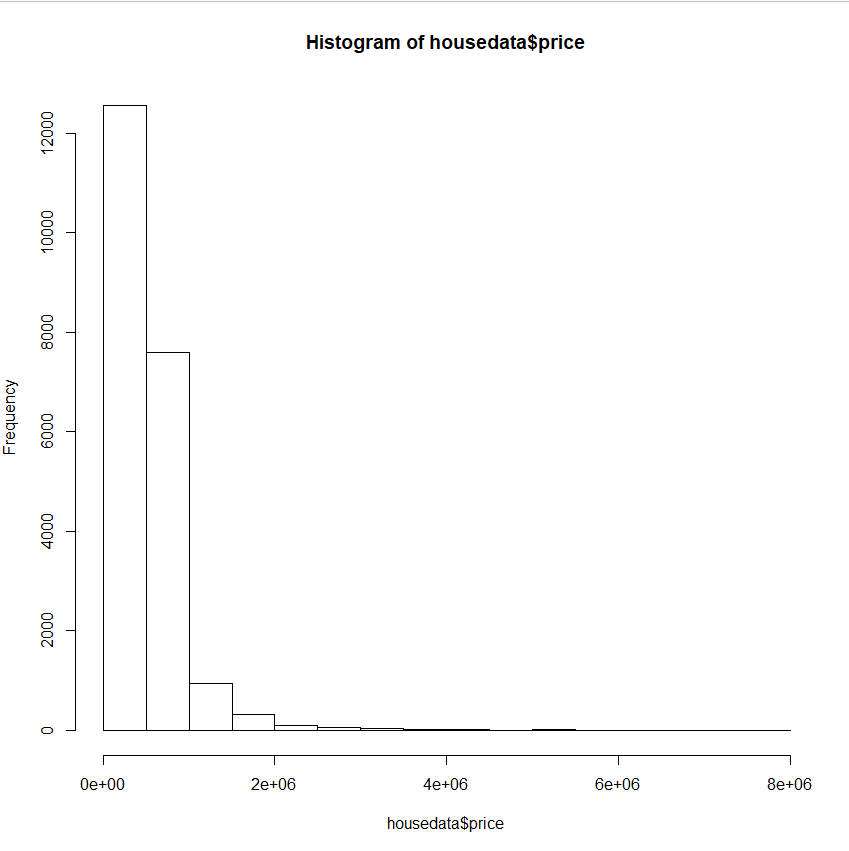
A close up of a map

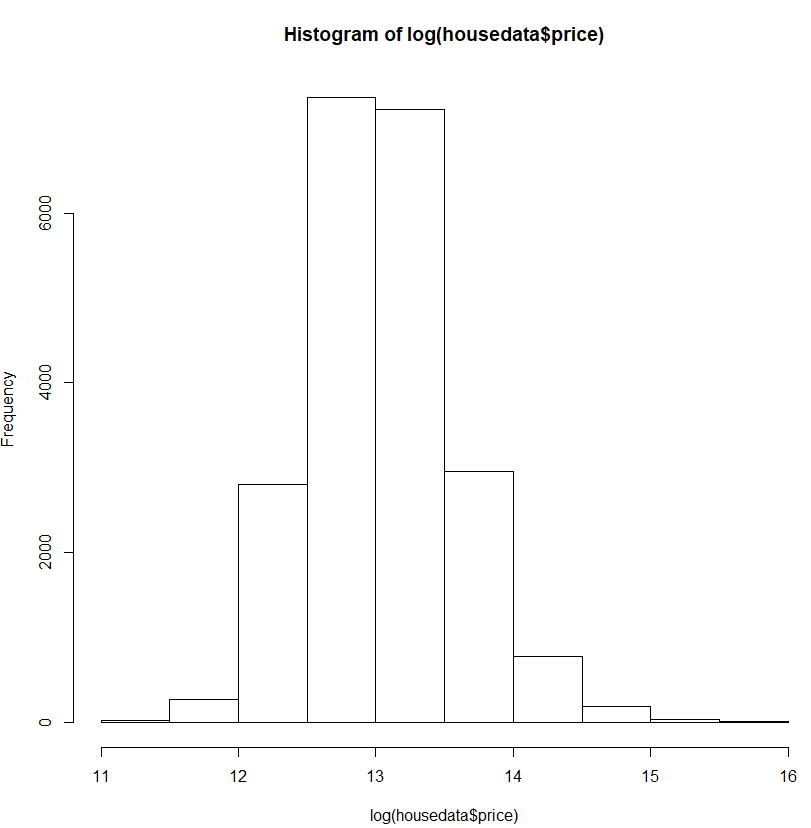
Description generated with high confidence

* Linear associations exist between price and some independent variables like sqft-living, bathrooms, sqft\_above. So linear model can be built using them.
* Multicollinearity exist between independent variables:
  + 1. Correlation between sqft\_living and sqft\_above is (0.88)
    2. Correlation between sqft\_living and sqft\_living15 is (0.76)
    3. Correlation between sqft\_living and grade is (0.76)
    4. Correlation between sqft living and bathrooms is (0.75)

But multicollinearity can be taken care later by checking VIF value of the model.

**Building the models and Residual Analysis of each model:**

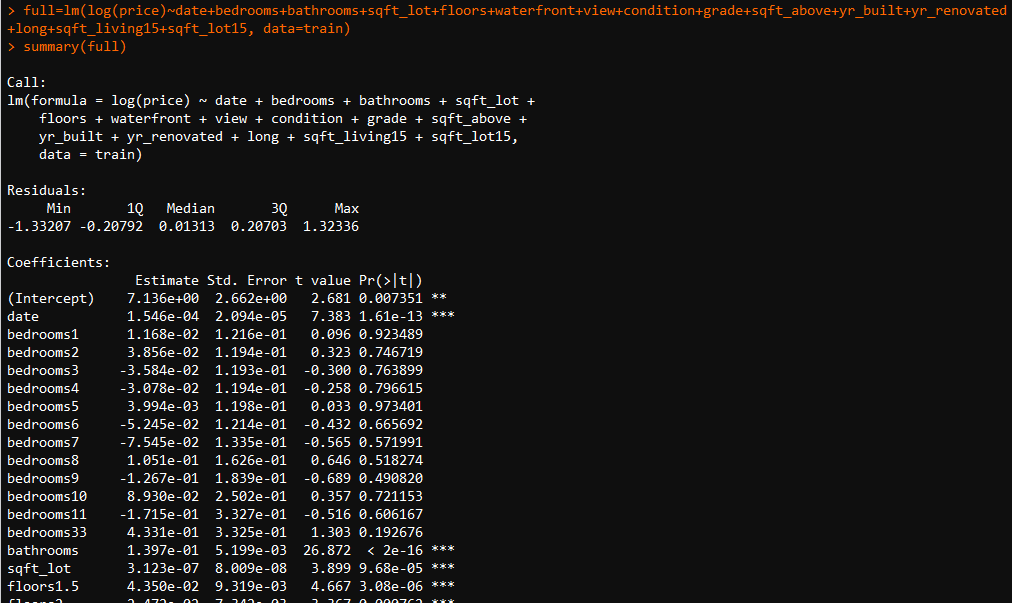
****

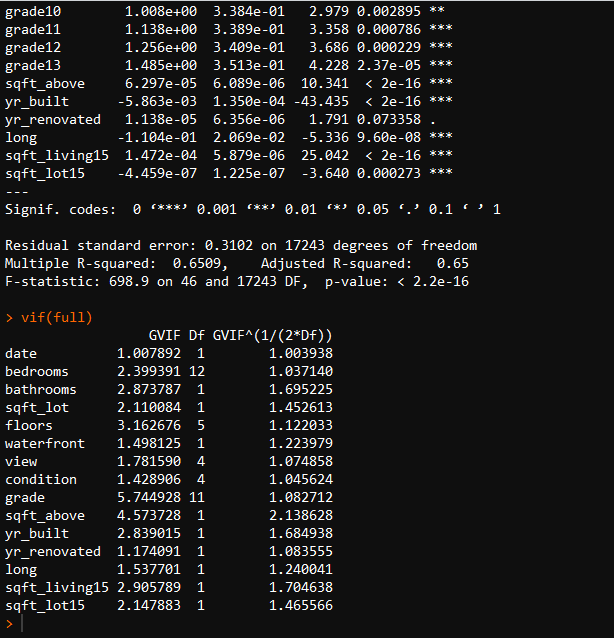


Taking log of price normalizes the data.

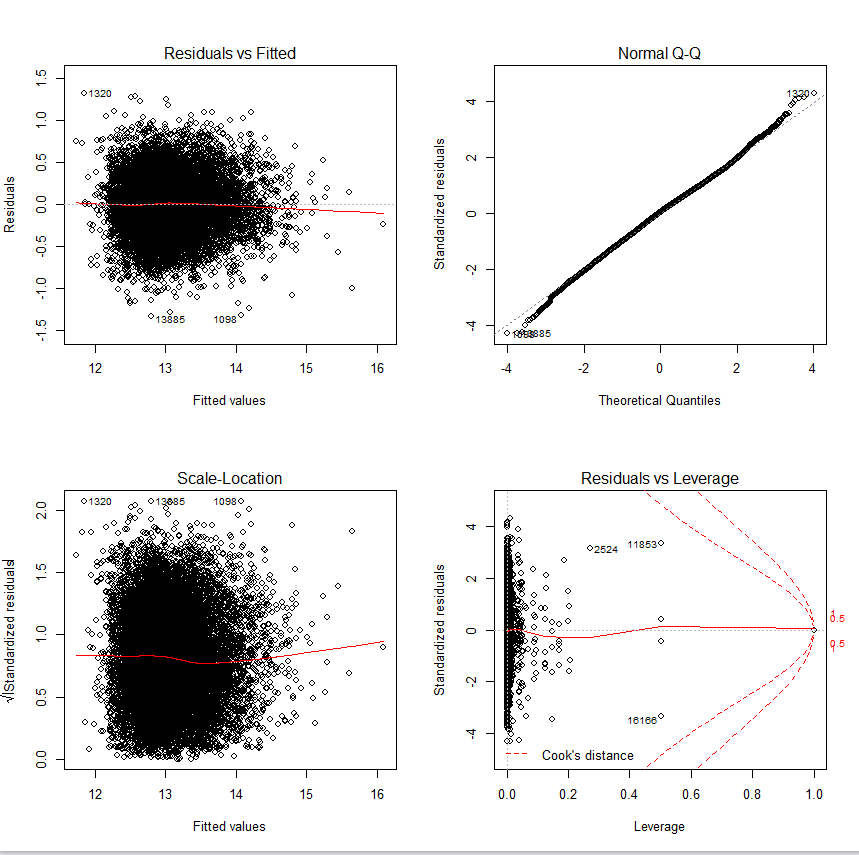
**Model 1 – Elimination by P value:**

**Taken all variables into account in the first model and then eliminating the variables one by one depending on the p value.**



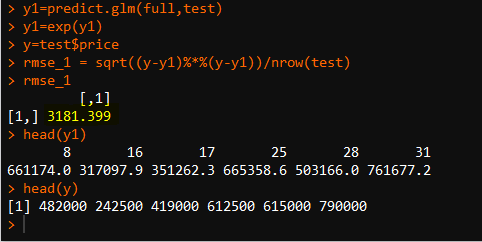


Residual Analysis of the above model: The residuals are normally distributed from QQ Plot.

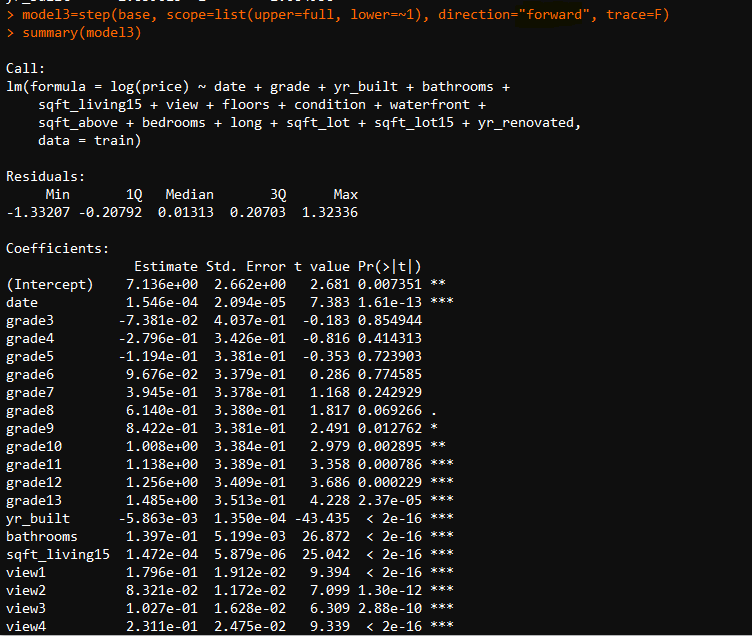


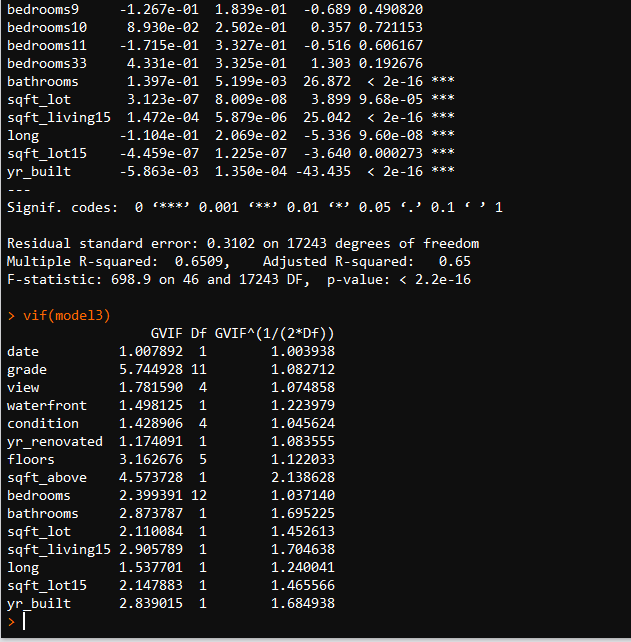
Below is RMSE and prediction for model1:

The predictions are close and the **RMSE value is 3181.399**. This RMSE will be compared with other models to determine the best model with lowest RMSE.



**Model2: Model by using step function “forward” selection and applying log on price.**

****

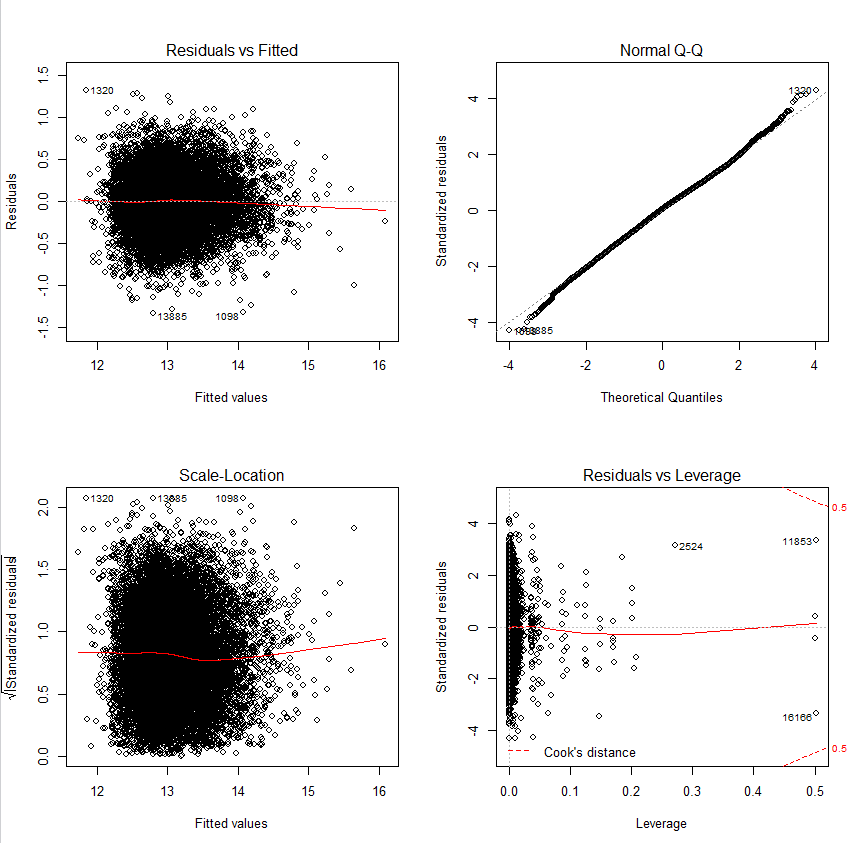
****

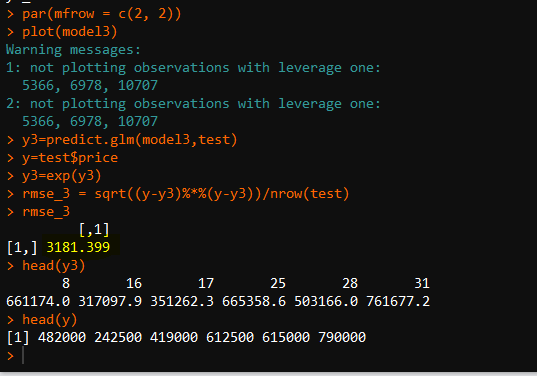
Residual Analysis of the above model: The residuals are normally distributed from QQ Plot.

Below is RMSE and prediction for model1:

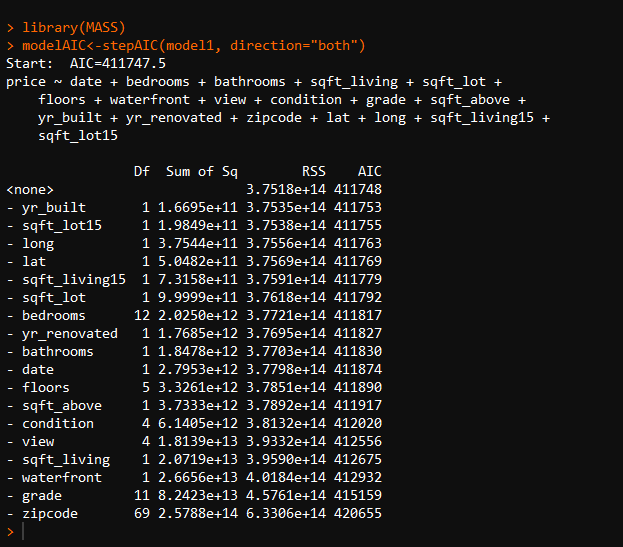
The predictions are close and the **RMSE value is 3181.399**. This RMSE will be compared with other models to determine the best model with lowest RMSE.

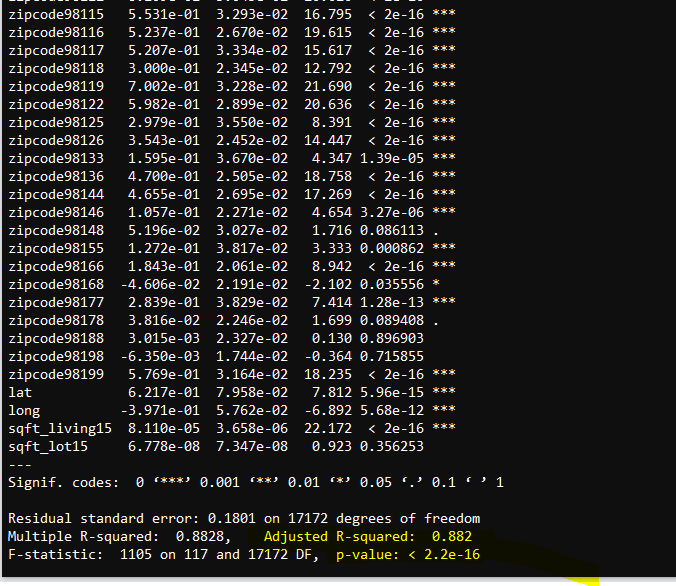
Forward selection with step function yields the same model as model1.

****

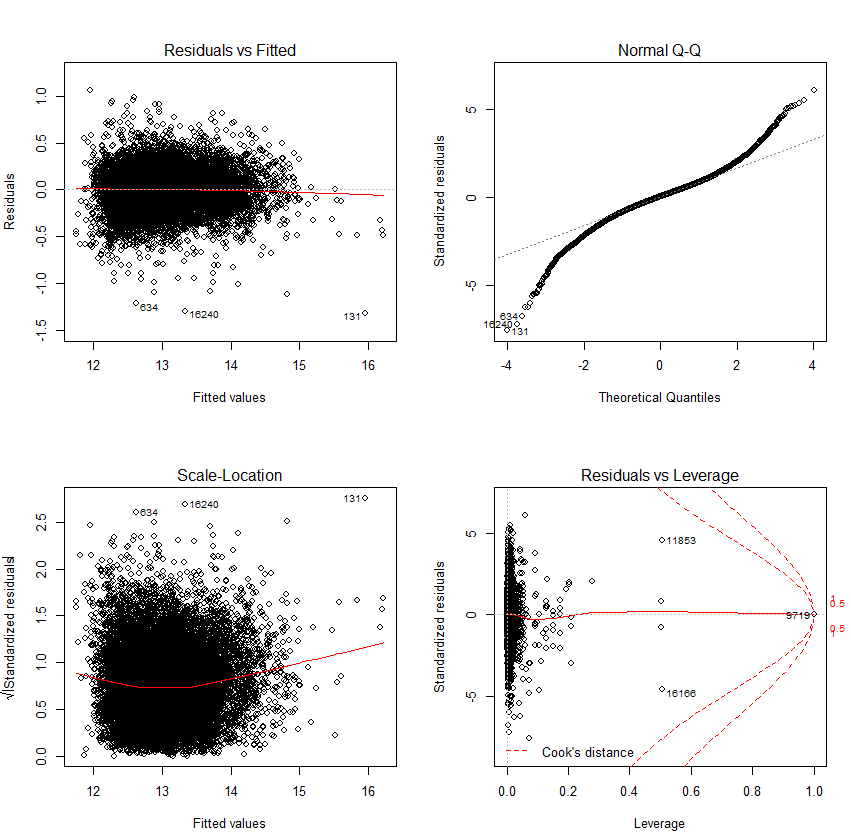
****

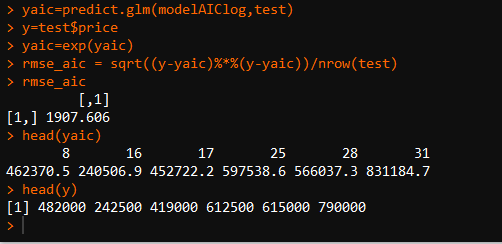
**Model3: Using step function “both” with AIC metric for selection- Best model with lowest RMSE.**

****

****

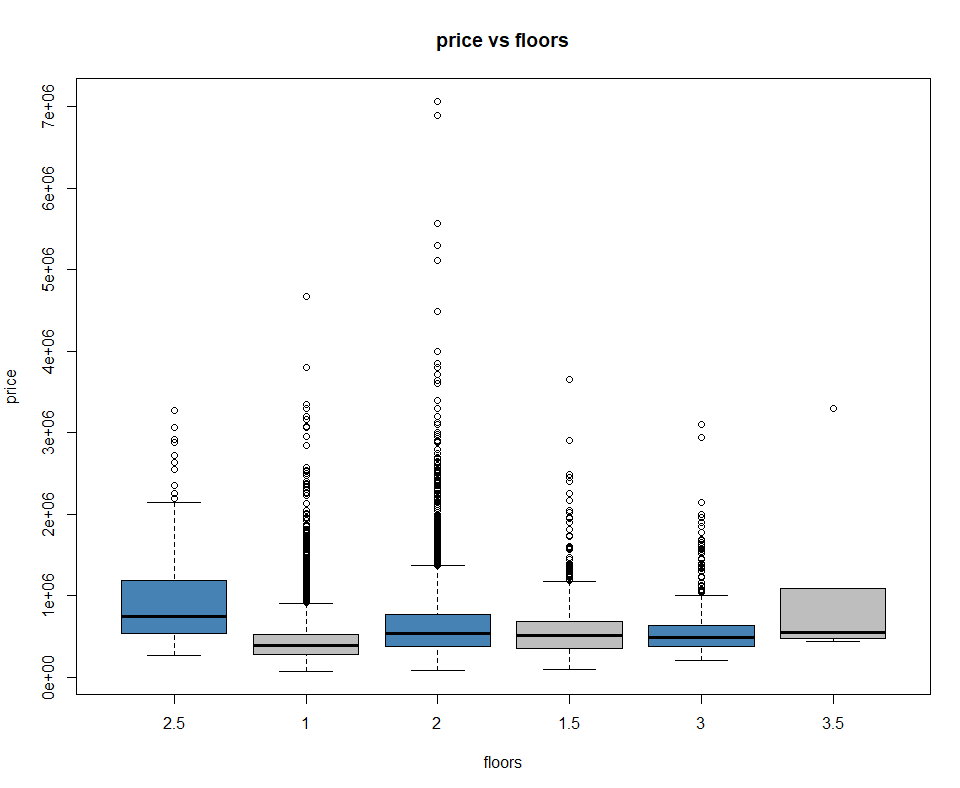
**Residual Analysis and Predictions of Model3: Predictions are very close, and this has been tested on test dataset. This model has lowest RMSE.**

****

****

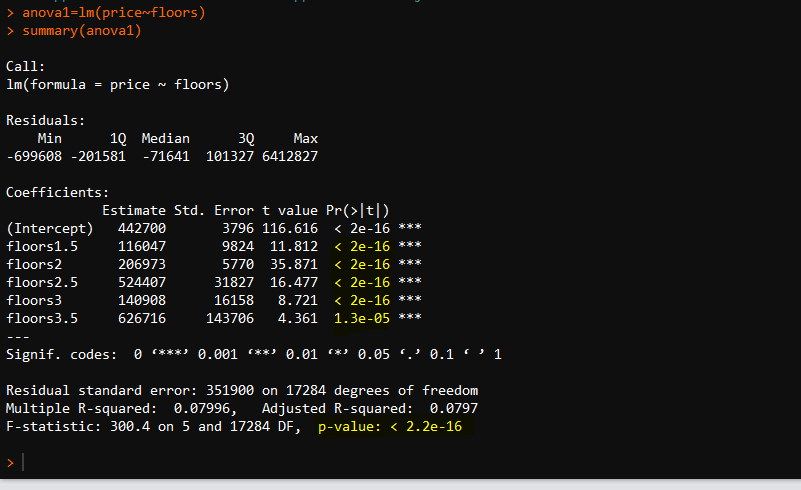
***Model3 has the lowest RMSE value 1907.606 among the three models as stated above. So model3 is the best model by multiple linear regression and can be used for prediction of house prices.***

**ANOVA: Extension to Linear Regression**

****

**Null Hypothesis: Group means of price of all houses with different floors are equal**

**Alternative Hypothesis: Group means of price of all houses with different floors are not equal**

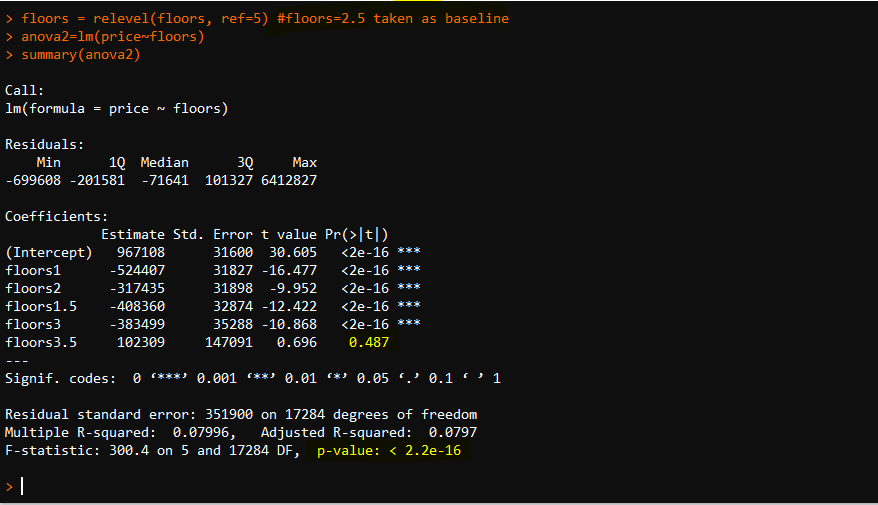
****

**After building the model, we find that F test is satisfied as p value for model is less than 0.05 at 95% confidence which means that at least one of the group means in price for different types of floors are different.**

Here, the baseline has been taken as floor1 in the model and group means of all houses differ for different kinds of floors from floor1.

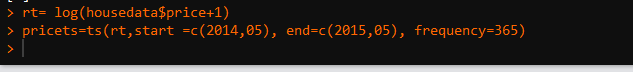
Releveling the baseline of floor to 2.5 . Before floor=1 is taken as baseline and other floor’s group price means are compared with floor1 group mean.

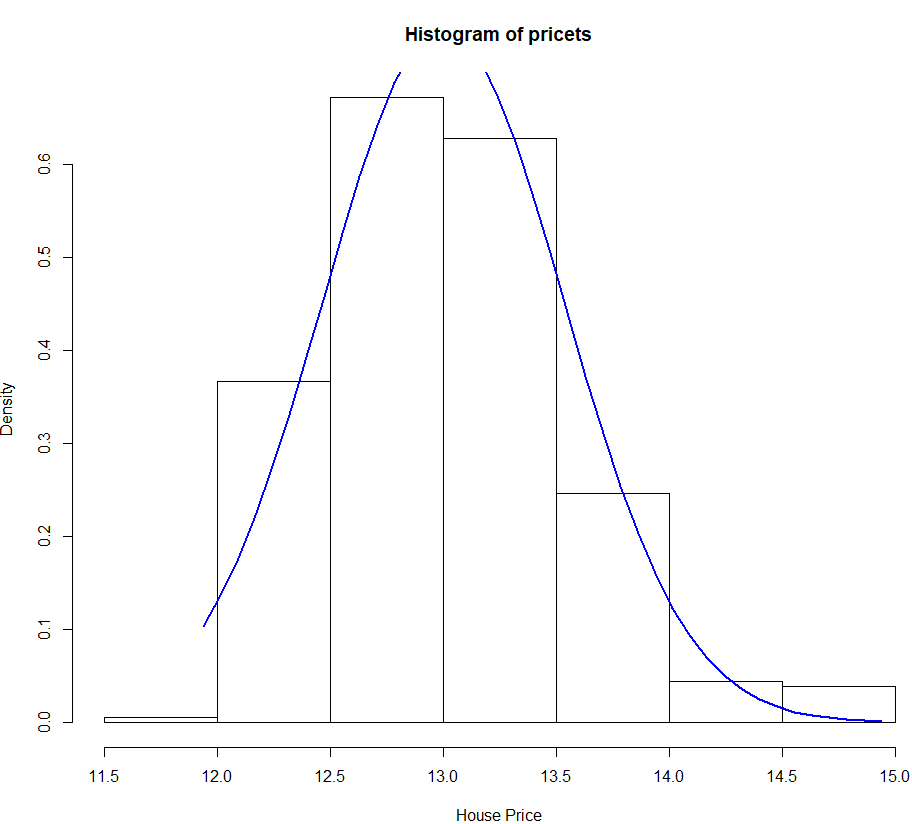
Rebuilding the anova model and we can find that other types of floors have different group price means from floor2.5 except floor3.5 which has same group mean for price as floor2.5(because pvalue>0.05).

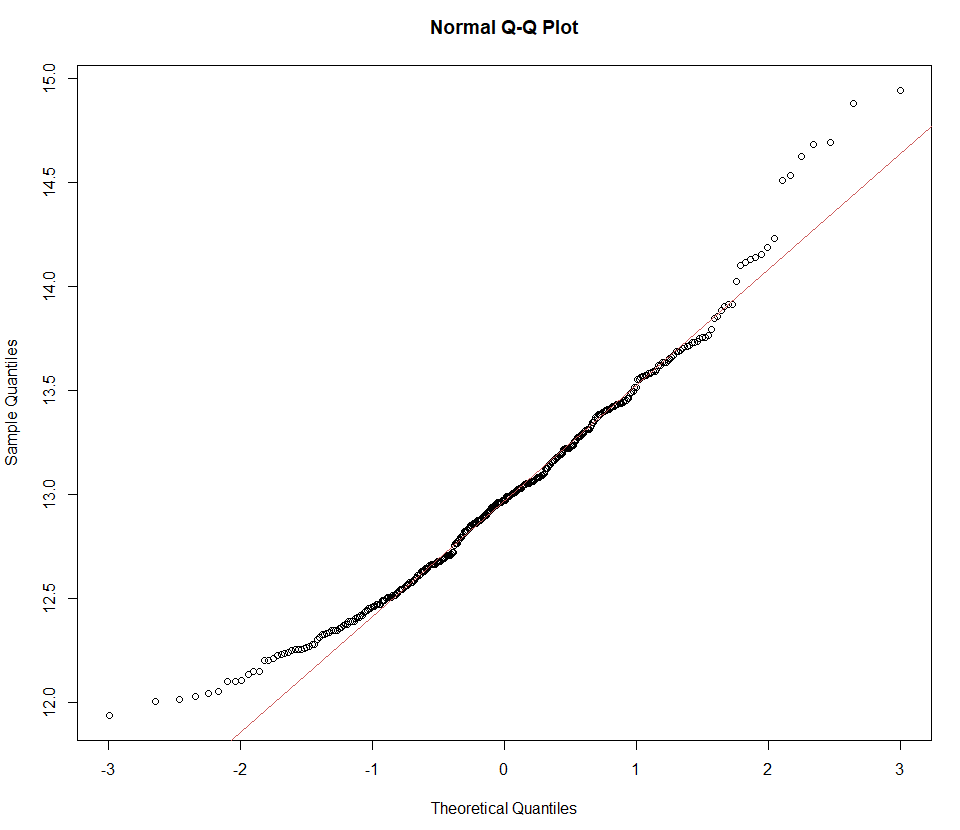


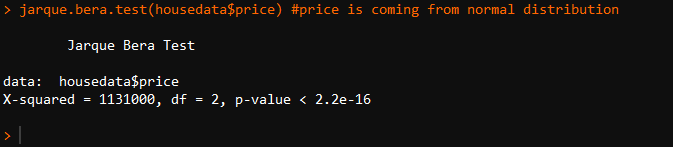
**Time Series Analysis and Forecasting:**

**Time Series Object Normality Tests:** Histogram, QQ Plot and Jarque Bera test show that price is coming from almost normal distribution and is slightly right skewed.



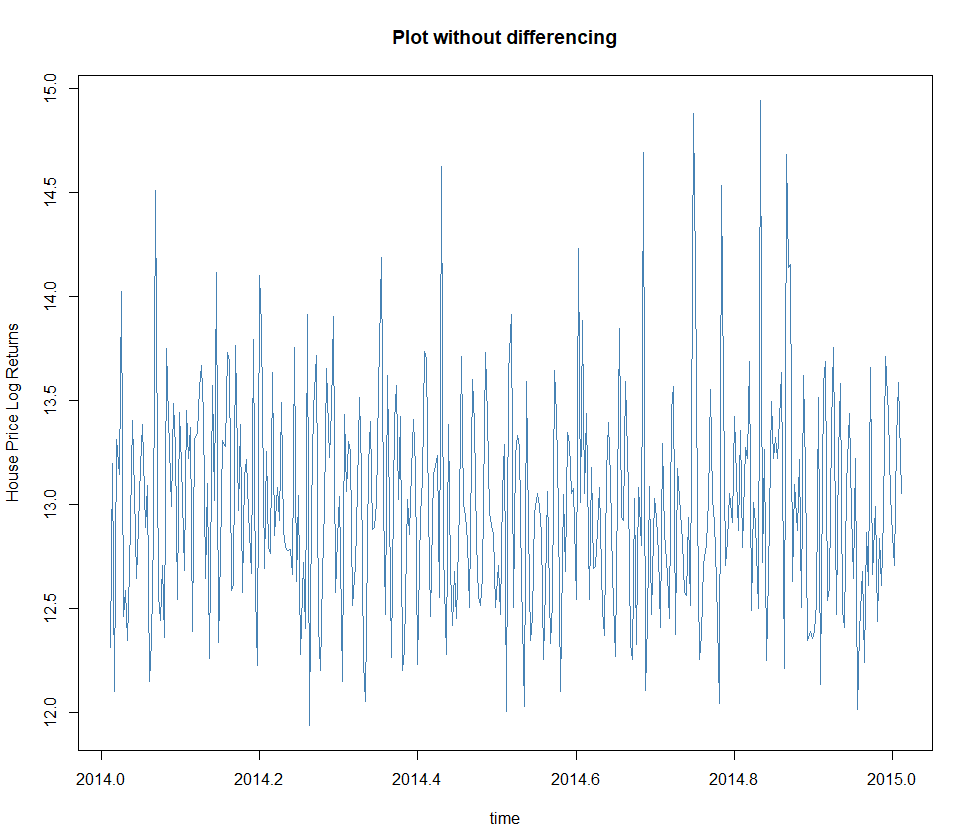


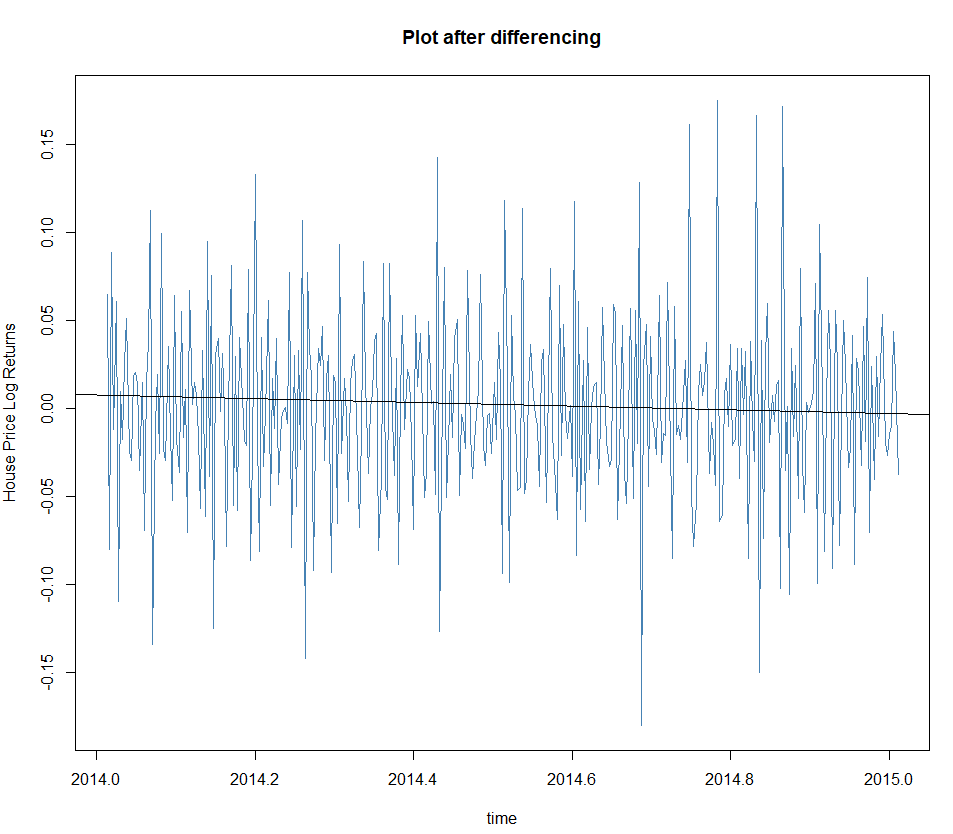




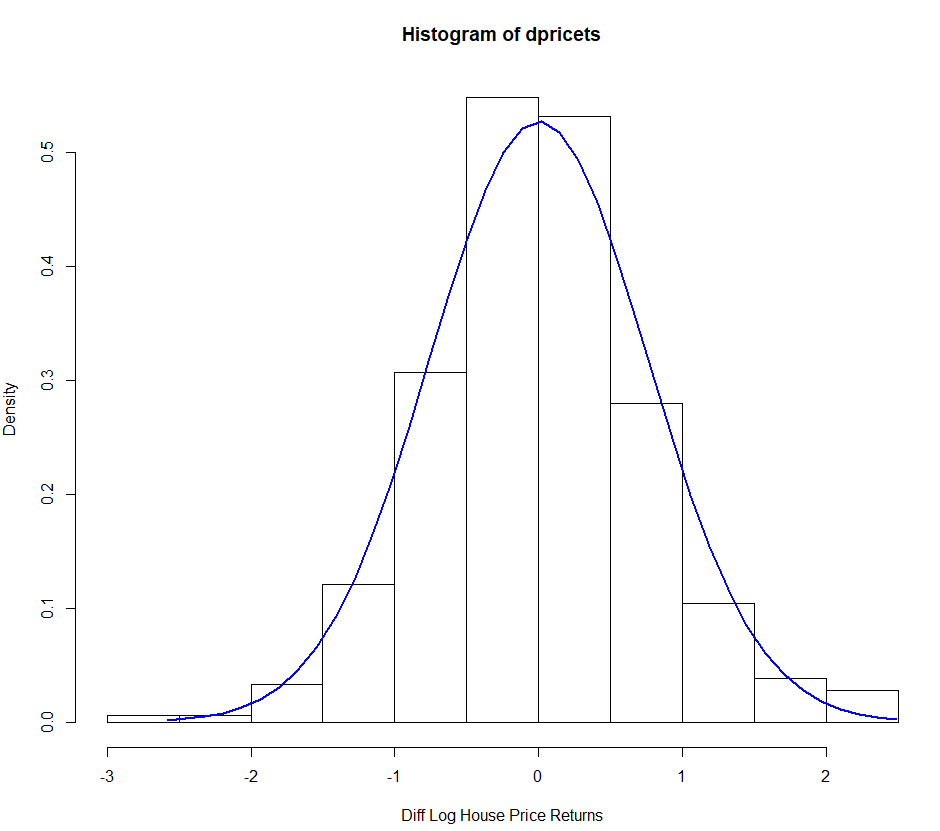
**Time Series Plots:**

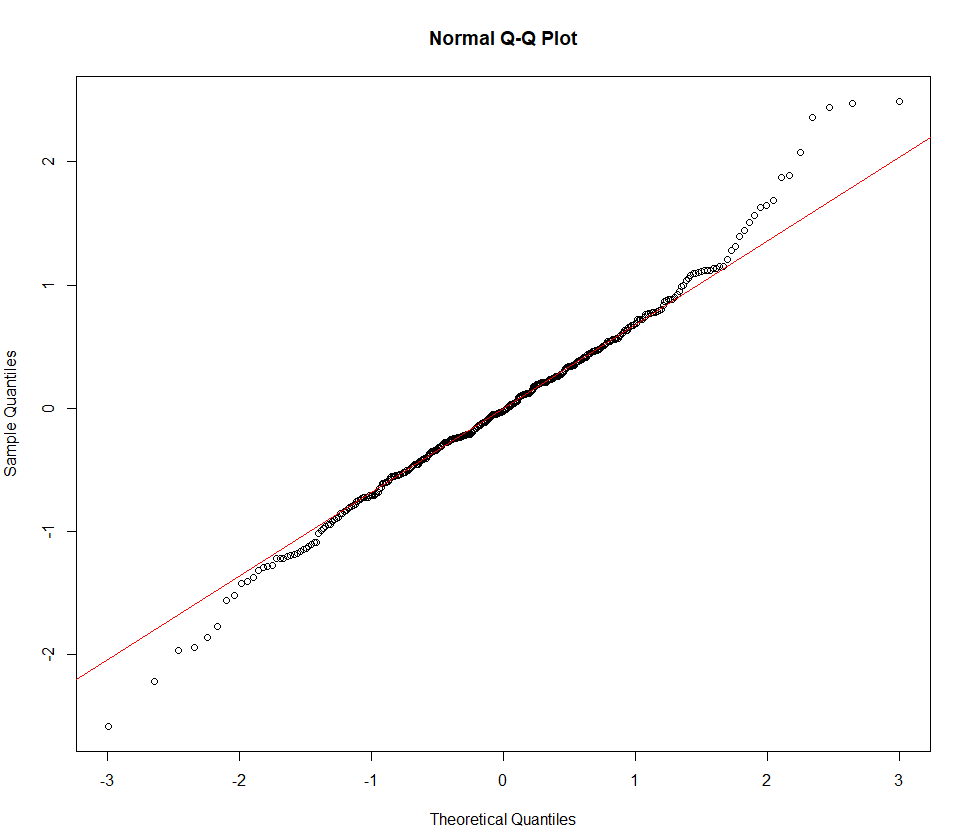
Mean and variance looks almost constant with time: Stationary series

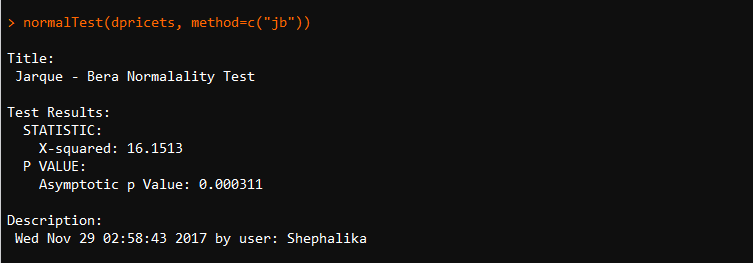




**Differenced Time Series Object Normality Tests:** Histogram, QQ Plot and Jarque Bera test show that price is coming from almost normal distribution.





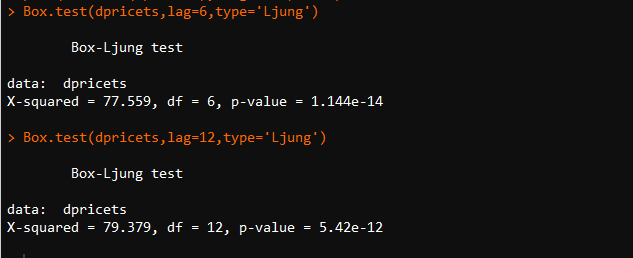


**Plots and test show that serial correlations exist**

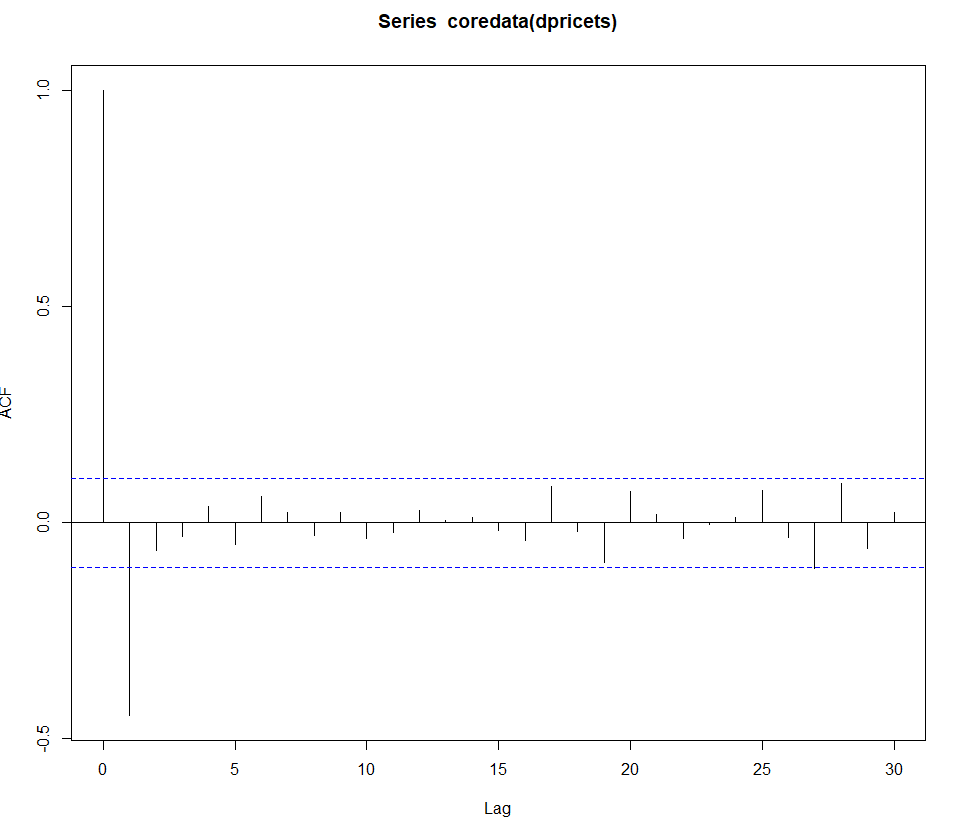
**Series is not a white noise: Confirmed by Ljung Box test as p value is less than 0.05 – Performed on diff time series object**

**#H0: Series is not correlated, and autocorrelations of time series object is zero**

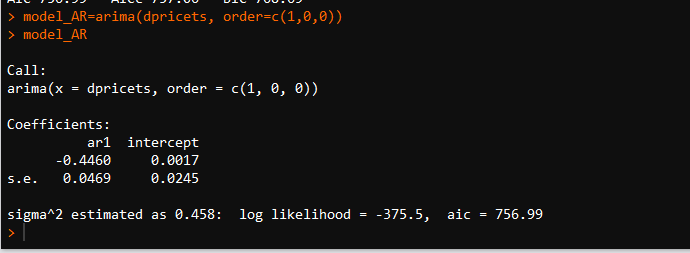
**#H1: Series is correlated**

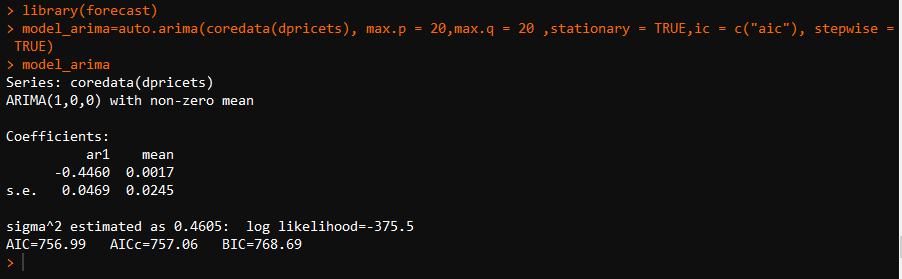
****

ACF plot decays quickly which shows that series is serially correlated.

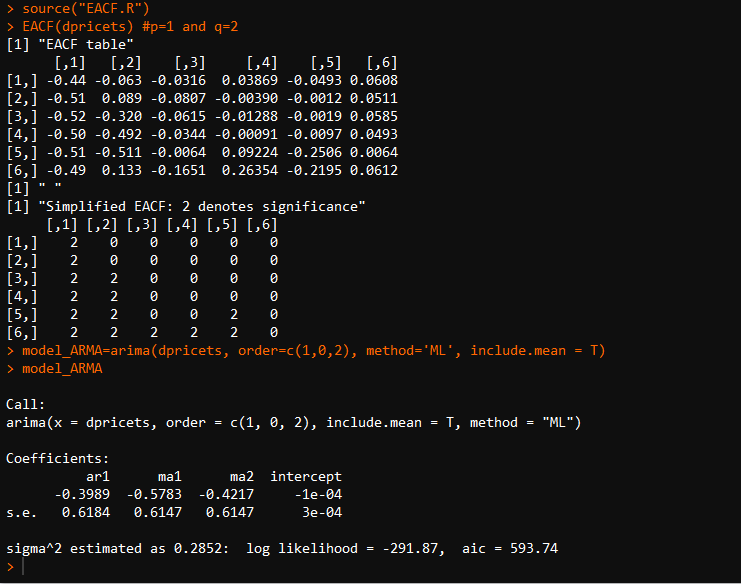


**Model1: Model Using AR (1,0,0) and auto.arima model:  
Both are same models with same AIC value.**

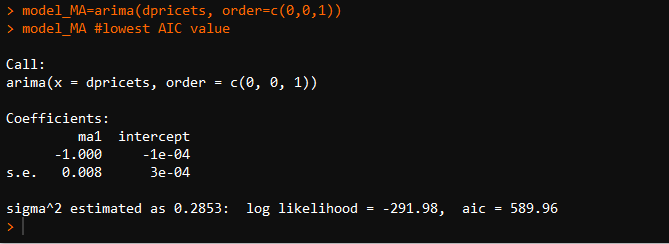




**Model 2: Model using EACF**



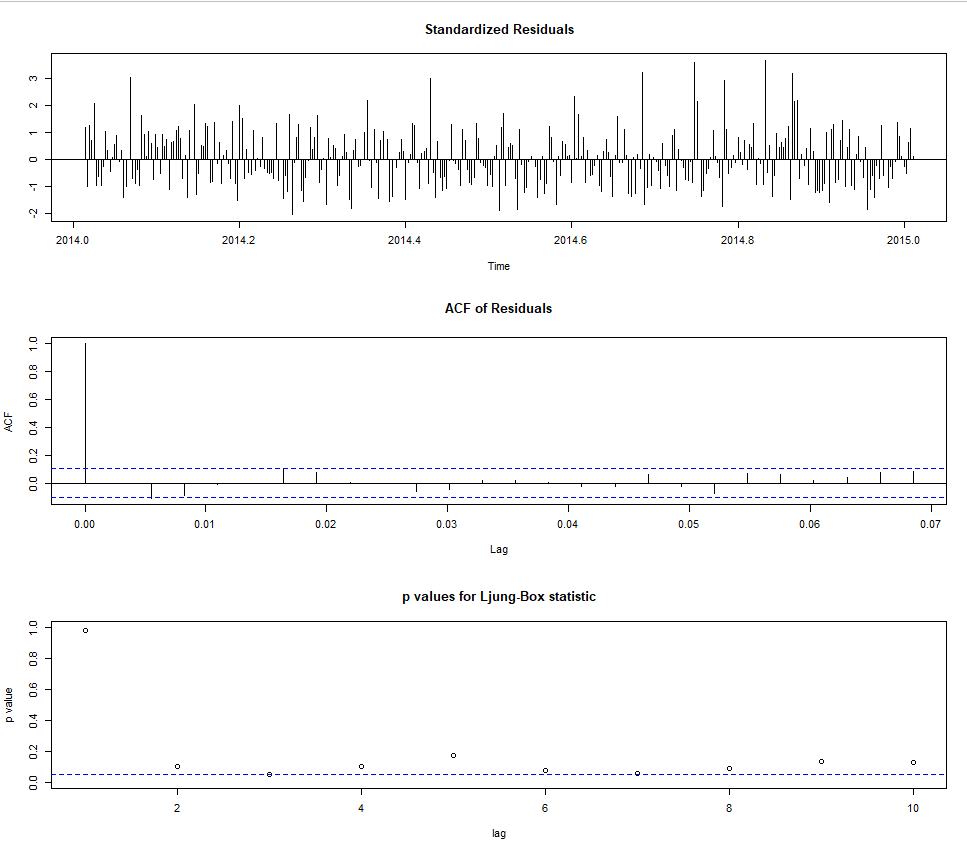
**Model 3: Model Using MA – Model with lowest AIC value**



Residual Analysis for MA model:

A close up of a map

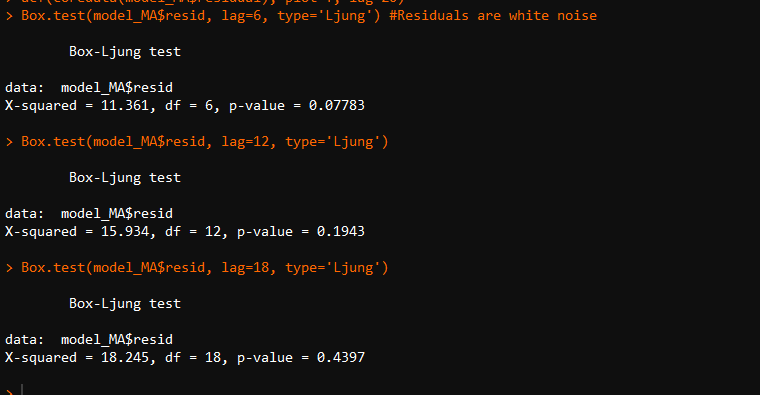
Description generated with high confidence



A close up of a mans face

Description generated with very high confidence

Residual Analysis for MA model: Ljung test result says that the residuals is white noise(independent). Assumption met.

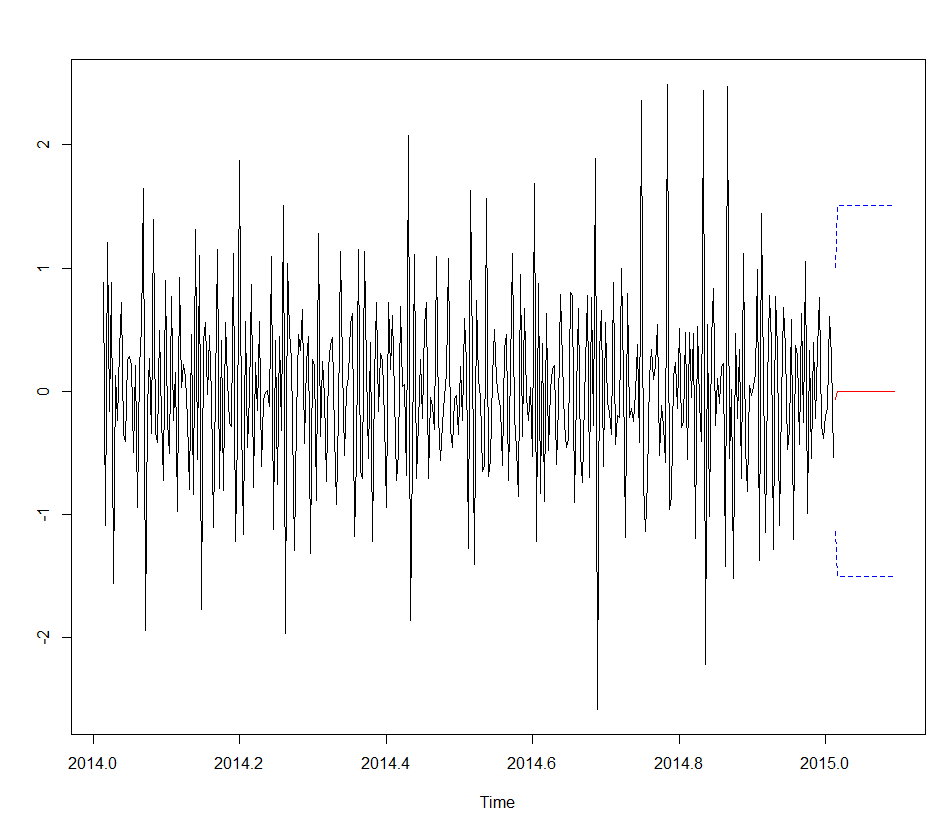


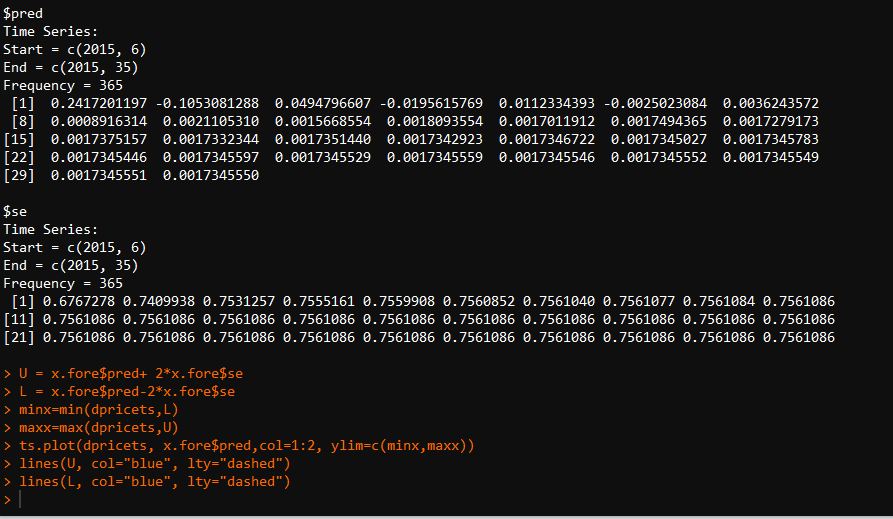
1. **Results and Findings:**

Evaluations for best model for linear regression has been made based on lowest RMSE value.

Prediction from Linear Regression:

Best model is model3 and predictions from model3 are almost close to the actual values. The factors used to build model 3 are useful variables for price prediction.

Predictions for future with time series:



1. **Conclusion and Future Work:**

Significant relationship has been established between the factors and the price. Models has been developed and evaluated to closely predict the house prices. Time series analysis show that the future price can be predicted based on the past values.

## Limitations:

More attributes can be added to understand and predict house prices in a improved manner.

## Potential Improvements or Future Work:

Time series analysis can be done in a better way. Other algorithms can be used on the dataset and more models can be built and then these models can be compared with other ones to find the best model for predicting the price of a house.