Modelling and Predicting the House Price of Different Areas of Bangalore Using the Concept of Multiple Linear Regression Model

Ishika Naskar

ishika.naskar@stat.christuniversity.in

February 2023

1 Abstract:

House prices differ every year for different locations, for different neighbourhood facilities and for different amenities, so there is a need for a process so as to predict house prices on the basis of the available data for the future purposes. Predicting the price of the houses helps the seller to determine the rational selling price of a house and also helps the buyer to select which house to buy and when to buy. There are three factors that influence the price of a house which include infrastructure, amenities and neighbourhood. In this article we divide the whole dataset according to the three above mentioned factors and apply the multiple linear regression to fit a model on the dataset which can be used for further price prediction. We also find whether the fitted model satisfies the assumptions of linear regression equation or not. Lastly we identify the factors that mostly influence the price of a particular house.

2 Key terms:

House price, Modelling, Prediction, Multiple linear regression, Autocorrelation, Residual Analysis.

3 Introduction:

Houses are one of the most fundamental needs of human being so as to live. Houses are long term properties where people live. For buying a house, the buyer need to invest a large amount of money. So it is of immense important to study all the relevant factors that are affecting the house price and also affect the quality of the house. The most important task of the seller is to determine the price of house with maximum accuracy.

The price of houses can be determine in many ways. One of the very popular is to fit a model with the help of the available data set and make future prediction based on the model. For this the accuracy of the model should be quite high so as to make accurate predictions.

In present days of artificial intelligence, these kind of predictions can be done very easily with large accuracy. Generally several machine learning models like linear regression analysis, support vector regression, k nearest neighbours, random forest regressor, decision trees, CatBoostregression, XGB regression etc.

In this article we have considered the housing dataset for Bangalore. We use the multiple linear regression model to the available housing dataset.

4 Literature review:

[1] P. Durganjali, et al., proposed a house resale price prediction using classification algorithms. In this paper, the resale price prediction of the house is done using different classification algorithms like Logistic regression, Decision tree, Naive Bayes and Random forest is used and we use AdaBoost algorithm for boosting up the weak learners to strong learners. Several factors that are affecting the house resale price includes the physical attributes, location as well as several economic factors persuading at that time. Here we consider accuracy as the performance metrics for different datasets and these algorithms are applied and compared to discover the most appropriate method that can be used the reference for determining the resale price by the sellers.

[2] Ayush Varma , et al., proposed house price prediction using machine learning and neural networks. Housing prices keep changing day in and day out and sometimes are hyped rather than

being based on valuation. Predicting housing prices with real factors is the main crux of our research project. Here we aim to make our evaluations based on every basic parameter that is considered while determining the price. We use various regression techniques in this pathway, and our results are not sole determination of one technique rather it is the weighted mean of various techniques to give most accurate results. The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied. We also propose to use real-time neighborhood details using Google maps to get exact real-world valuations.

[3] Sifei Lu, et al., proposed a hybrid regression technique for house prices prediction. With limited dataset and data features, a practical and composite data pre-processing, creative feature engineering method is examined in this paper. The paper also proposes a hybrid Lasso and Gradient boosting regression model to predict individual house price. The proposed approach has recently been deployed as the key kernel for Kaggle Challenge “House Prices: Advanced Regression Techniques”. The performance is promising as our latest score was ranked top 1% out of all competition teams and individuals.

5 Objective:

In this article our objectives are as follows:

- To find the relationship of the dependent variable with the other independent variables.
- To fit a multiple linear regression model on the given data set.
- To check the accuracy of the fitted model.
- To check the accuracy of the fitted model.

6 Methodology:

6.1 Data description:–

Here we consider the housing data set for the Bangalore metropolitan city. In this data set we have 32 columns and 1951 rows. The columns are as follows:

Price, Area, No. of bedrooms, Resale, Maintenance, Gymnasium, Swimming Pool, Landscape

Garden, Jogging track, Rain water harvesting, Indoor Game, Shopping mall, Intercom, Sports facility, ATM, Club house, School, 24*7 security, Power backup, Car parking, Staff quarter, Cafeteria, Multipurpose rooms, Hospital, Washing machine, Gas connection, AC, Children's play area, Lift available, BED, Vaastu compliant, Golf course.

For all the columns except the Price, Area and No. of bedrooms, 0 indicates absence and 1 indicates presence.

6.2 Theoretical Concepts:–

The theoretical concepts used in this article are explained below:

Multiple linear regression is a linear model that explains the linear relationship that exists between one dependent variable and more than one independent variables.

The multiple linear model is expressed using the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where Y is the dependent variable.

$X_1, X_2, \cdots, X_k$ are the k independent variables.

β_0 is the intercept term of the linear regression equation.

$\beta_1, \beta_2, \cdots, \beta_n$ is the regression coefficient of Y on $X_1, X_2, \cdots, X_K$ respectively.

ϵ is the random error in predicting Y with the help of the independent variables. It is called the residual error.

Assumptions: In case of linear regression we consider the following assumptions:

1. The dependent and independent variables show a linear relationship between the slope and the intercept.

2. The independent variables are non-stochastic.

3. The independent variables are non- correlated.

4. The expected value of the residual (error) is zero.

5. The variance of the residual (error) is constant across all observations.

6. The value of the residual (error) is not correlated across all observations.

7. The residual (error) values follow the normal distribution with zero mean and constant variance.

We use the method of least squares to estimate the regression coefficients.
In this method, we try to minimize the sum of square of the distance between a data point and the regression line.
After that we also test for the significance of the predictors in predicting the response variable.
Here the null hypothesis is that all the regression coefficient are equal to zero against the alternative hypothesis is that at least one regression coefficient is not equal to zero.
Decision Rule: If the p value corresponding to a predictor is less than 0.05, then we reject the null hypothesis and conclude that the predictor is significant.
Then we verify whether the fitted model satisfies the assumption of the multiple linear regression model:

- **Check for linearity** For this we use the matrix scatter plot and the correlation coefficient values.
  If the value of correlation coefficient is close to 1, then the corresponding two variables are highly positively correlated.
  If the value of correlation coefficient is close to -1, then the corresponding two variables are highly negatively correlated.

- **Check for homoscedasticity of the residuals:** If the scatter plot of residual vs fitted value shows a random nature about the zero line, then the residuals are homoscedastic in nature. We also use studentized Breusch-Pagan test in this case where the null hypothesis is that the residuals are homoscedastic.

- **Check for normality of the residuals:** We use the normal QQ plot in this case and also use the Shapiro-Wilk test where the null hypothesis is that the residuals are normality distributed.

- **Check for multicollinearity of the predictors:** For that purpose, we use VIF (Variation Inflation Factor). VIF is the measure of the amount of multicollinearity present in the multiple linear regression analysis. When there is a correlation between multiple regressors in the regression model, the multicollinearity exists. VIF is given by:

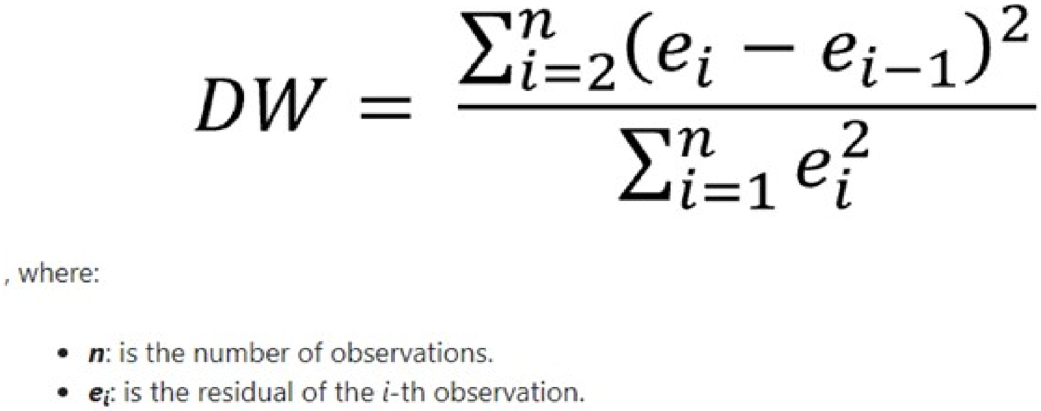
$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the adjusted r square when the i^{th} independent variable is regressed on the remaining ones. If VIF > 5, then variables are highly correlated.

- **Check for the presence of the outliers:** We use the cook's distance measure to detect the presence of outliers.

- **Check for the presence of autocorrelation:** When the residuals of a regression model are not independent of each other, then we say that autocorrelation is present. In other words, if the value of residual e_i depends on the value of residual e_{i-1}.

  We use the ACF plot and the Darwin Watson test. For DW test, the null hypothesis is that 1st order autocorrelation is not present against the alternative hypothesis that 1st order autocorrelation exists.



$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

, where:

- ***n***: is the number of observations.
- $\boldsymbol{e_i}$: is the residual of the *i*-th observation.

Finally we use the R square value and the adjusted R square value to find the goodness of the fitted model.

6.3 Data Analysis:–

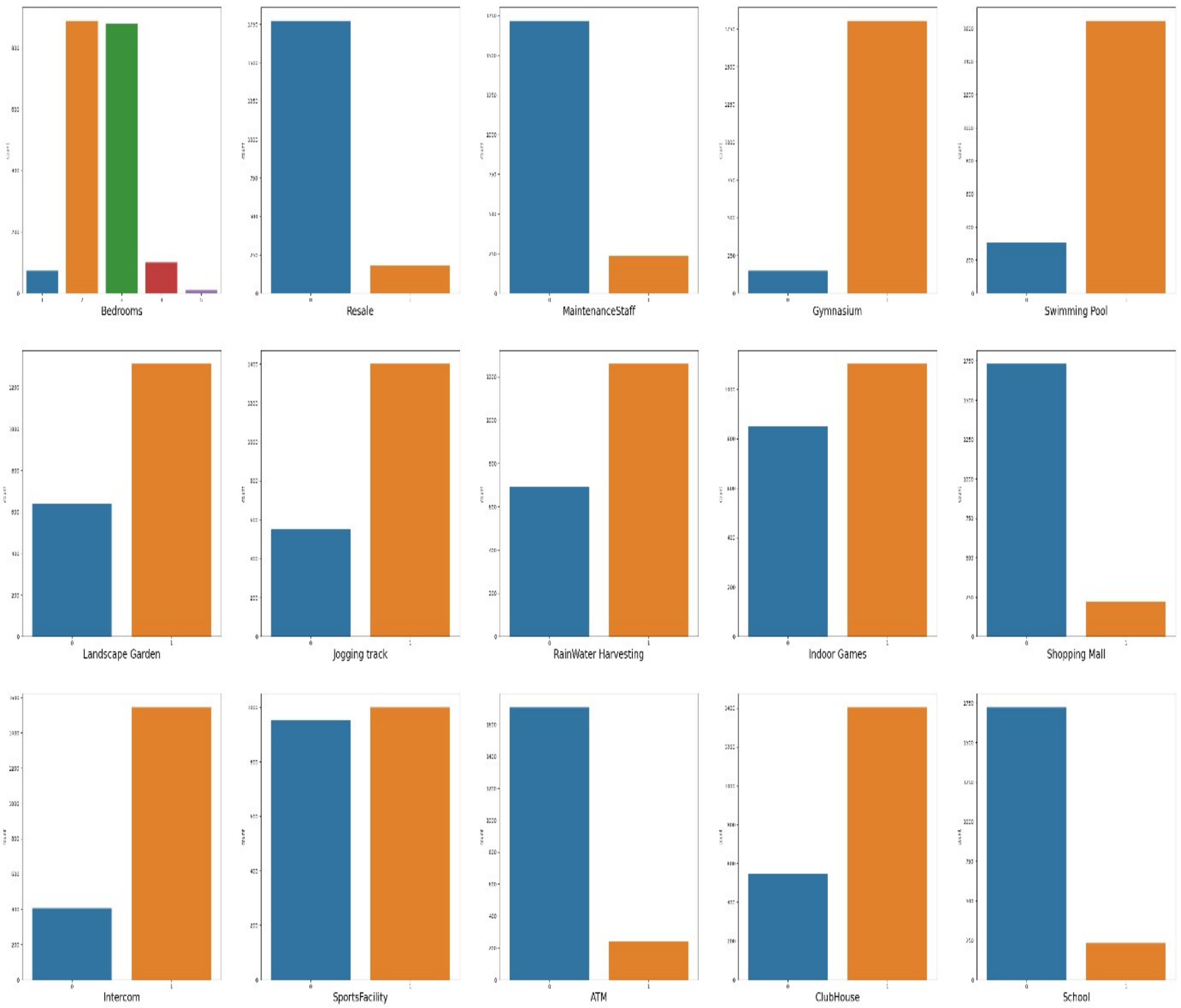
The different methods for analysing the data set are described below:

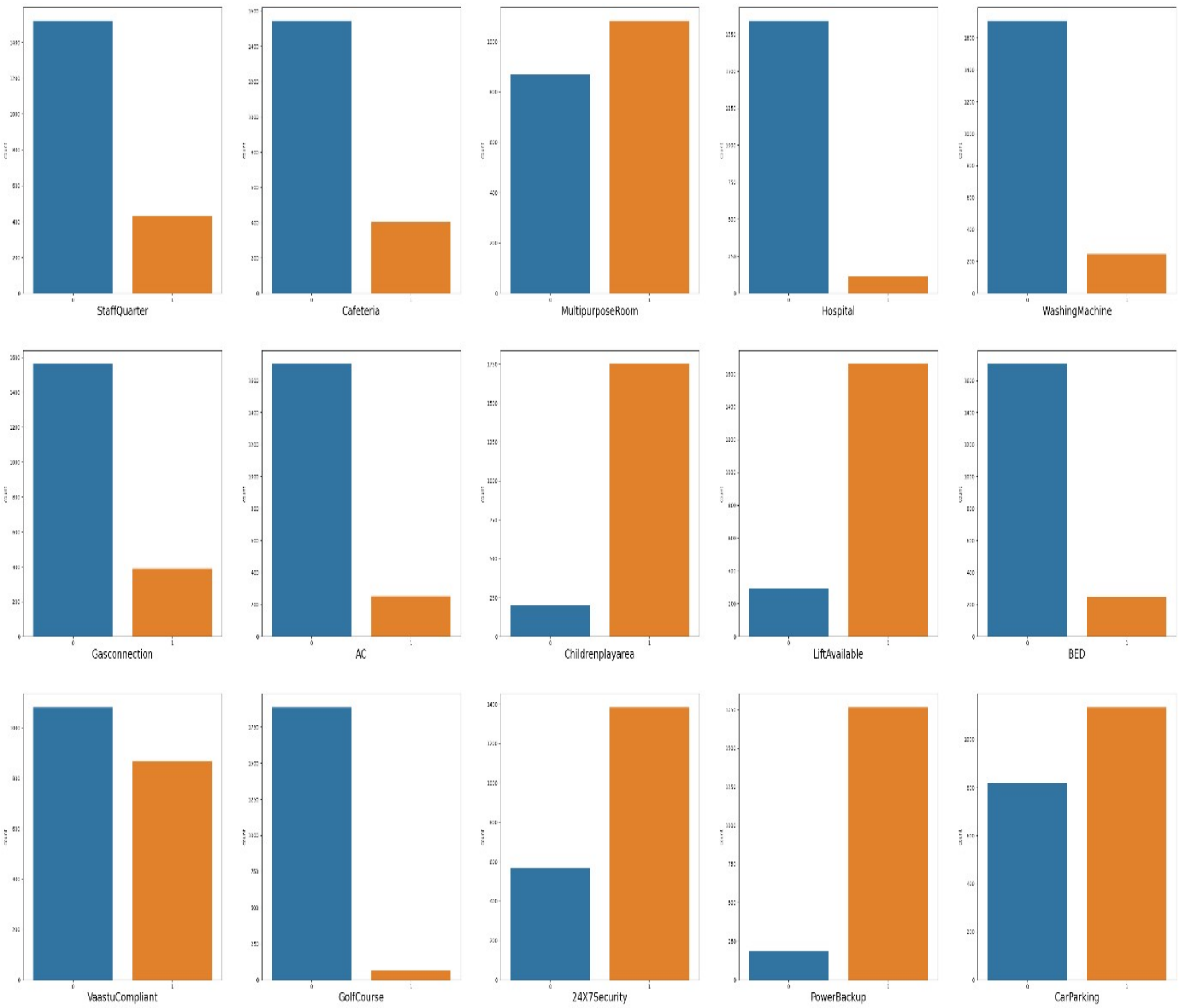
Exploratory Data Analysis(EDA) : We found that the dataset does not contain any missing value.

The following table gives the summary of each of the columns:

Column Names	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Price	2096000	4973500	6950000	9953003	10000000	202700000
Area	525	1147	1330	1516	1610	9900
No. of Bedrooms	1	2	3	2.532	3	5
Resale	0	0	0	0.093	0	1
Maintenance Staff	0	0	0	0.1199	0	1
Gymnasium	0	1	1	0.9241	1	1
Swimming Pool	0	1	1	0.8432	1	1
Landscaped Gardens	0	0	1	0.6725	1	1
Jogging Track	0	0	1	0.7176	1	1
Rain water harvesting	0	0	1	0.6453	1	1
Indoor Games	0	0	1	0.5648	1	1
Shopping Mall	0	0	0	0.1143	0	1
Intercom	0	1	1	0.7919	1	1

Barplot of the categorical variables :


Bedrooms
Resale
MaintenanceStaff
Gymnasium
Swimming Pool
Landscape Garden
Jogging track
RainWater Harvesting
Indoor Games
Shopping Mall
Intercom
SportsFacility
ATM
ClubHouse
School

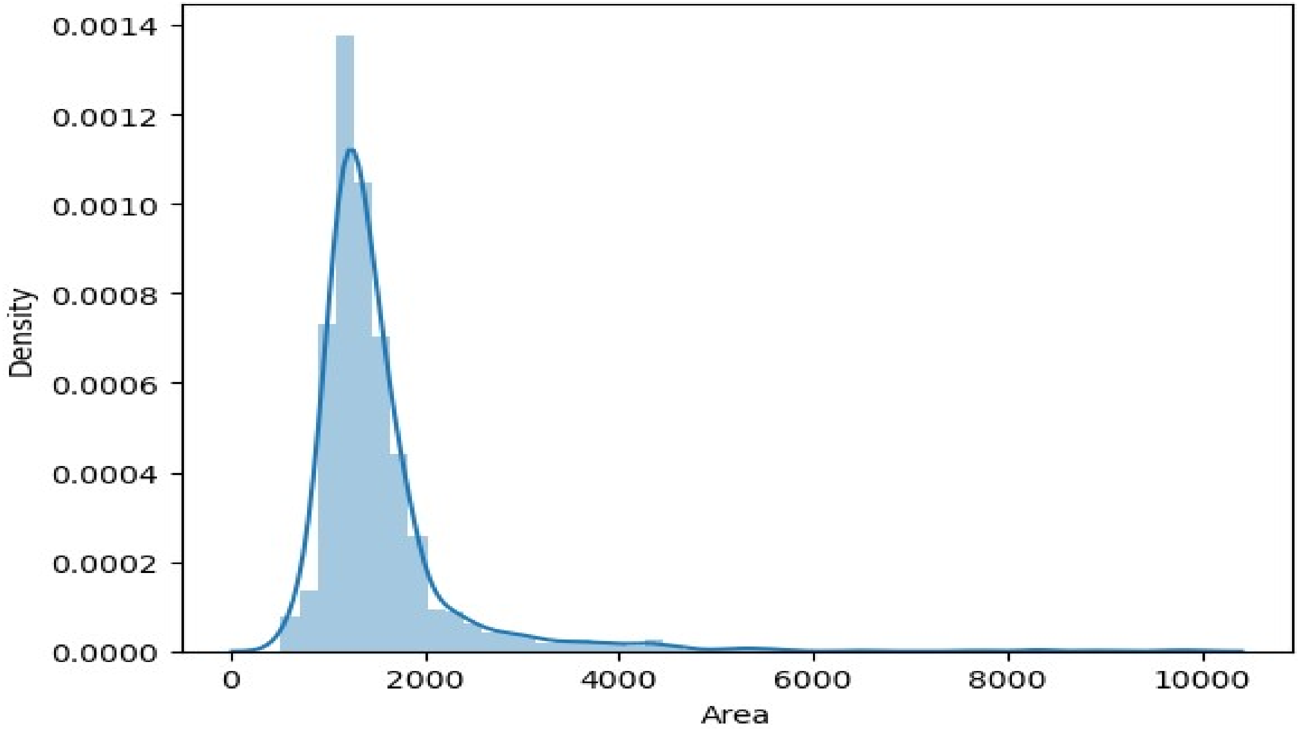


From the barplots we observe that :—

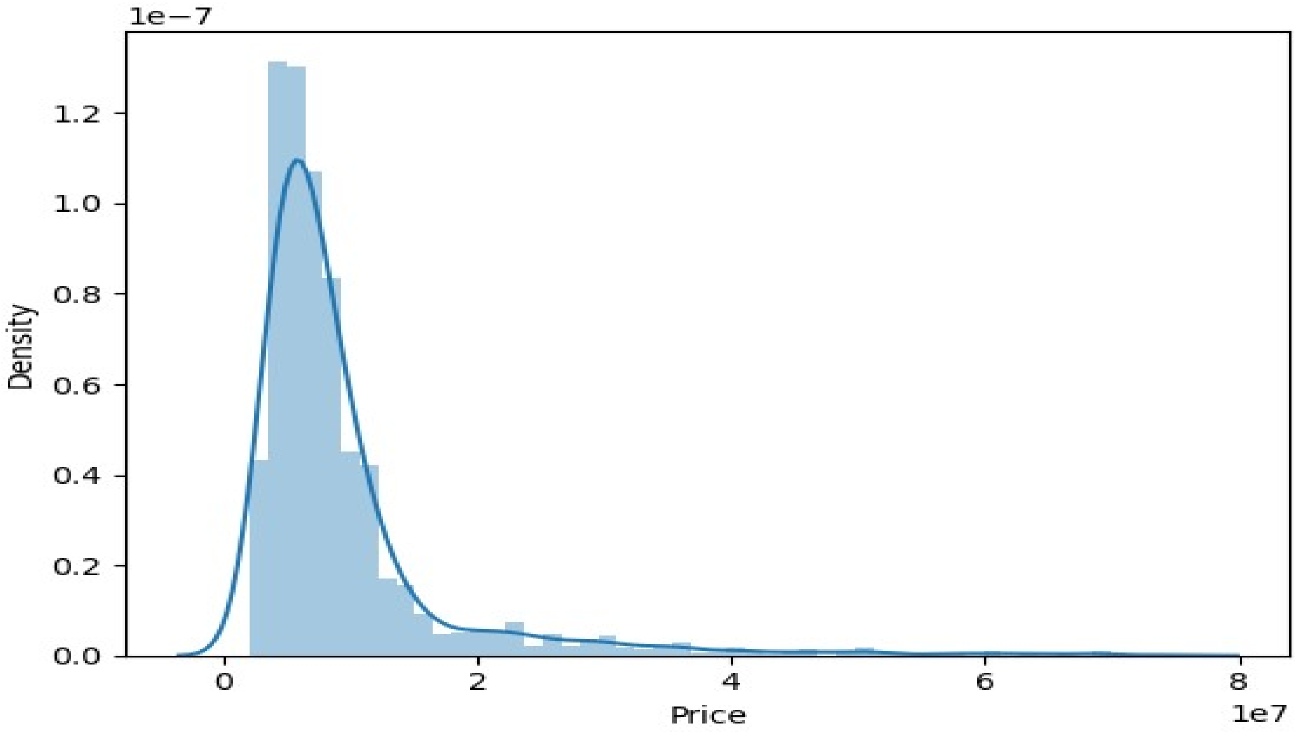
1. Most of the houses has 2 bedrooms.
2. Maintenance staff are not present for most of the houses.
3. Gymnasiums are present in maximum houses.
4. Houses with landscape gardens and jogging tracks are more in number.
5. Most of the houses has rain water harvesting capacity.

6. The difference between the no. of houses with indoor games and houses without indoor games are not very high.
7. Large no. of houses has intercom facility.
8. The difference between the no. of houses with sports facility and houses without this facility are very less.
9. Maximum no. of houses do not have ATM, staff quarter, hospital, cafeteria, ac, gas connection and golf course.
10. All most all the houses have children play area, lift and power back up.

Histogram of the Continuous variables:



From the above graph, we observe that the distribution of the area of the houses is highly positively skewed with a long right tail.



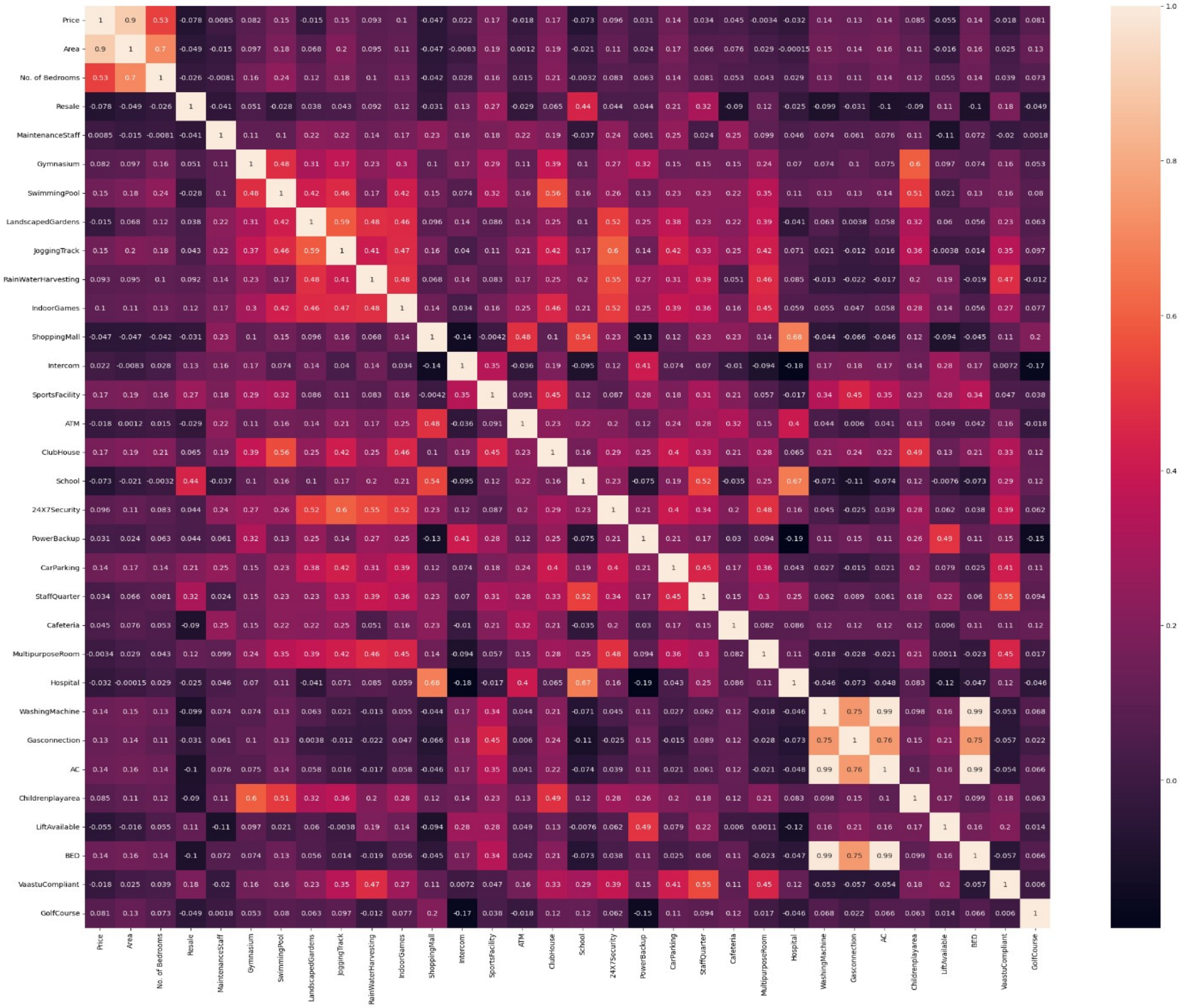
From the above graph we observe that the distribution of the price of the houses is highly positively skewed with a long right tail.

Here the price of the houses are considered as dependent variable and the rest of the variables are considered as independent variables.

Association between the variables: The following table gives the value of the correlation coefficient of the price of the houses with the other independent variables:

Coulmn Names	**Correlation Coefficients**
Area	0.9
No. of Bedrooms	0.53
Resale	-0.08
MaintenanceStaff	0.01
Gymnasium	0.08
SwimmingPool	0.15
LandscapedGardens	-0.01
JoggingTrack	0.15
RainwaterHarvesting	0.09
IndoorGames	0.1
ShoppingMall	-0.05
Intercom	0.02
SportsFacility	0.17
ATM	-0.02
ClubHouse	0.17
School	-0.07
24*7Security	0.1
PowerBackup	0.03
CarParking	0.14
StaffQuarter	0.03
Cafeteria	0.05
MultipurposeRoom	0
Hospital	-0.03
WashingMachine	0.14
Gasconnection	0.13
AC	0.14
Childrenplayarea	0.09
Liftavailable	-0.05
BED	0.14
VaastuCompliant	-0.02
GolfCourse	0.08

Observe that the price of the houses is highly positively correlated with area and no. of bedrooms in the house.



Now we create three different data frames using the variables which are of same type. The data frames are as follows:

The infrastructure data frame contains the variables price, area, no. of bedrooms, resale, rain water harvesting, washing machine, gas connection, ac, bed, vaastu compliant.

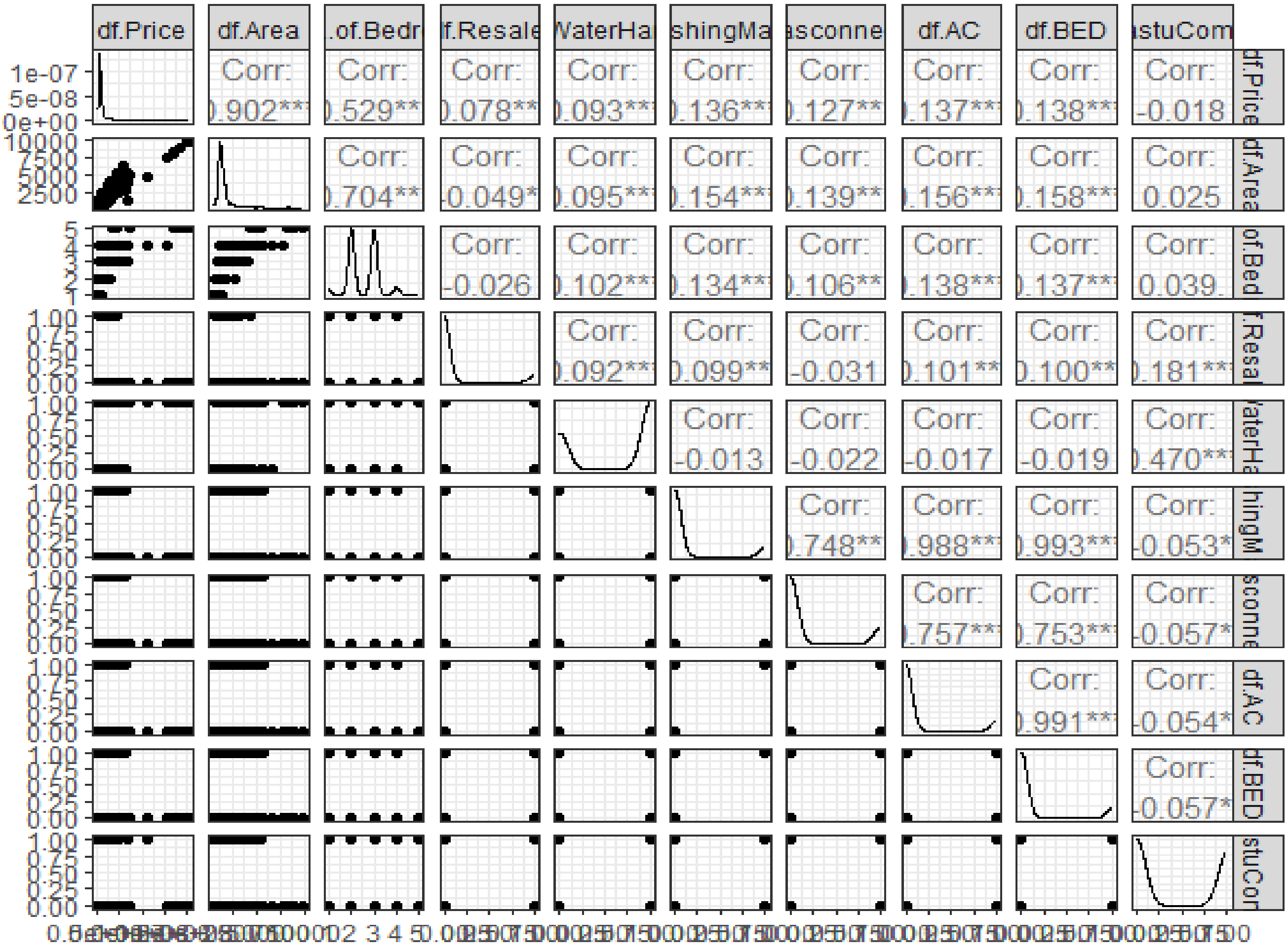
The amenity data frame contains the variables price, maintenance staff, gymnasium, swimming pool, landscaped garden, jogging track, indoor games, intercom, sports facility, 24*7 security, power backup, car parking, staff quarter, multipurpose room, children play area, lift available and golf course.

The **neighbourhood data frame** contains the variables price, shopping mall, atm, club house, school, cafeteria, hospital.

Now we fit the linear regression model in the three different data frame separately and verify whether the assumptions of linear regression model are satisfied or not:

Consider the Infrastructure Data frame: ———

To verify whether the dependent variable is linearly related with the independent variables:



From the matrix scatter plot we observe that price is highly positively correlated with area and no. of bedrooms.

Now we fit the multiple linear regression model:

The following table indicates the estimated value of the regression coefficients, standard error, value of the t statistic, the corresponding p value of the predictors and lastly whether the predictors are significant or not:

Predictors	**Estimated value of coefficient**	**Standard error**	**t-statistic**	**p value**	**Significant or Not Significant**
Intercept	-6716898.7	485487.4	-13.835	<2e-16	Significant
Area	17731.8	217.8	81.425	<2e-16	Significant
No. of Bedrooms	-4035284.9	244841.6	-16.481	<2e-16	Significant
Resale	-1194500.9	415662.1	-2.874	0. 00410	Significant
Rainwater Harvesting	1084676.8	279665.2	3.878	0.000109	Significant
Washing machine	4074799.3	3149536.3	1.294	0.195896	Not significant
Gas connection	265740.1	451658.1	0.588	0.556355	Not significant
AC	409613	2732040.5	0.150	0.880836	Not significant
BED	-4792301.7	3505674.8	-1.367	0.171780	Not significant
Vaastu Compliant	-1312940.4	271258.9	-4.840	1.4e-06	Significant

Therefore, the predictors Area, No. of bedrooms, Resale, Rainwater Harvesting and Vaastu Compliant are significant in predicting the house price.

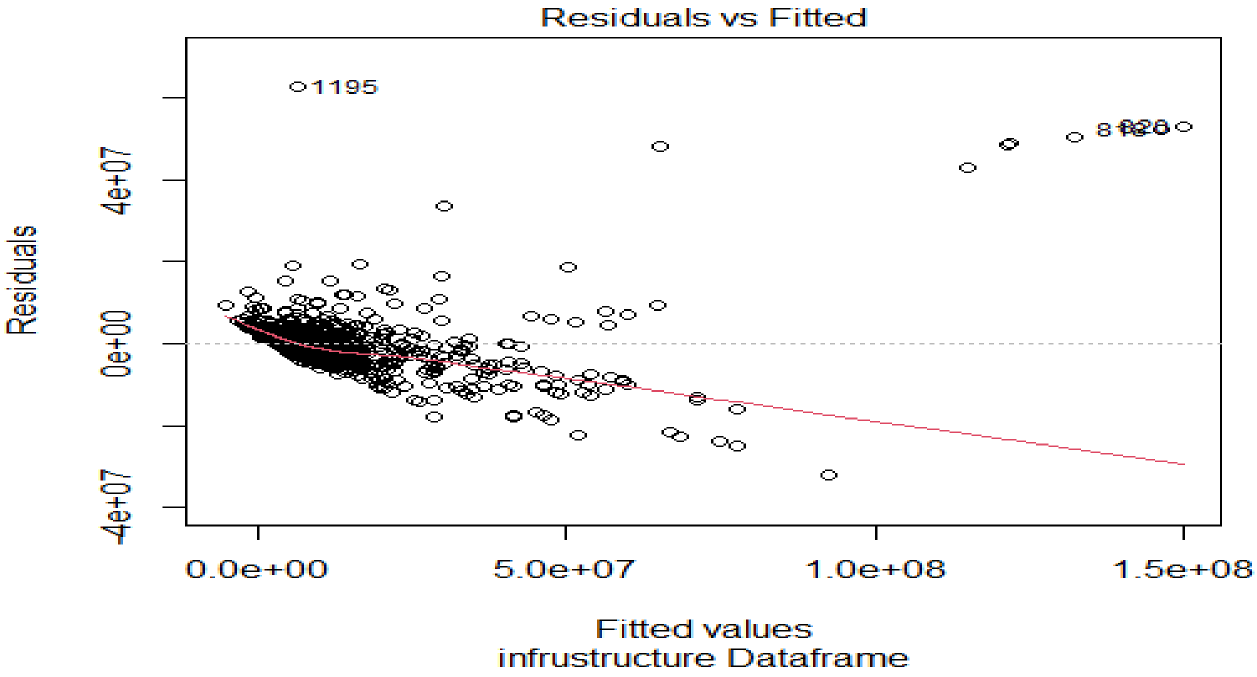
Here the value of the multiple R-squared is 0.8401. This implies that almost 84.01% of the variation in the response variable is explained by the predictor variables.

Here the value of the adjusted R-squared is 0.8394. This implies that the given data fits the multiple regression model quite well.

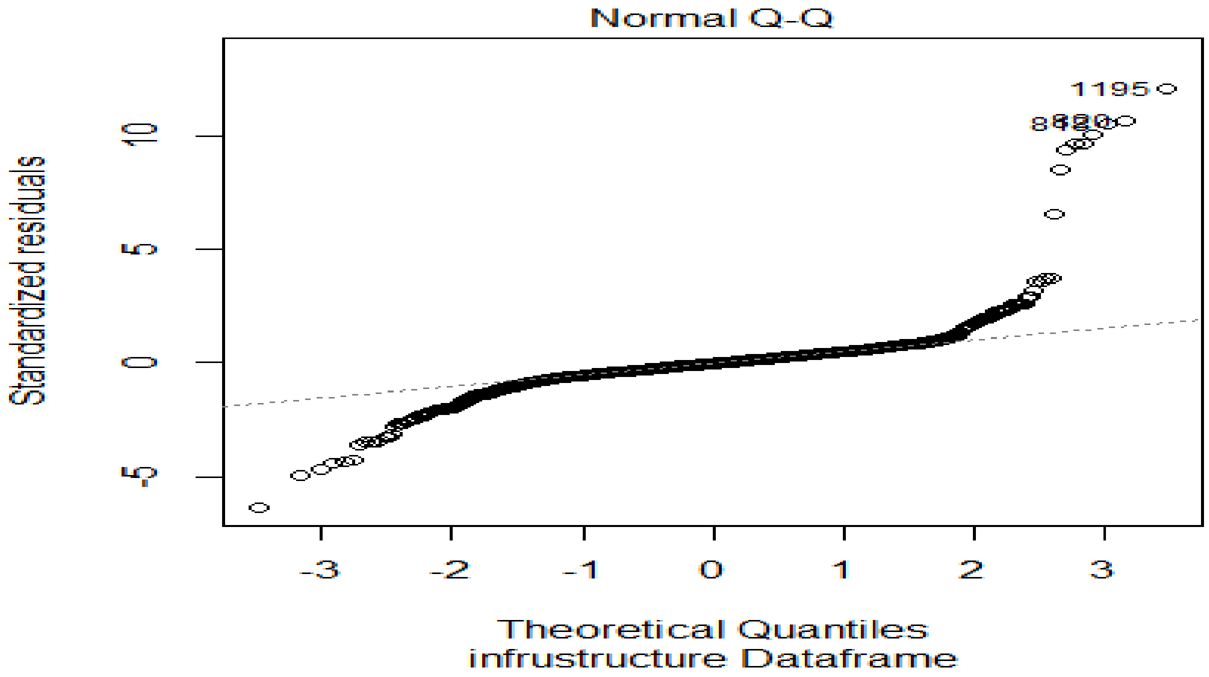
- **Check for multicollinearity:** Here the VIF values for the predictors Washing machine, AC and BED are greater than 5 so, these predictors are correlated with each other.

- **Check for homoscdasticity:** Here the points are not randomly scattered around the zero line.

Also the p value for the BP test is less than 0.05, so we reject the null hypothesis. We conclude that the residuals are not homoscedastic.

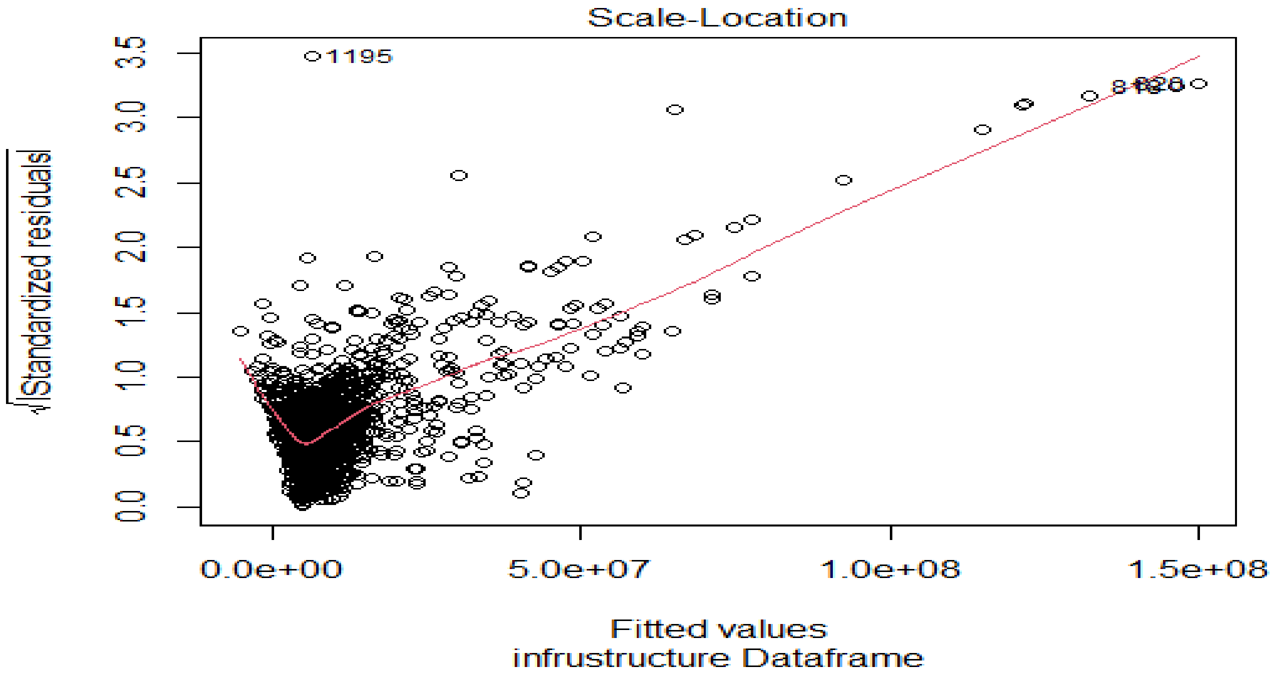


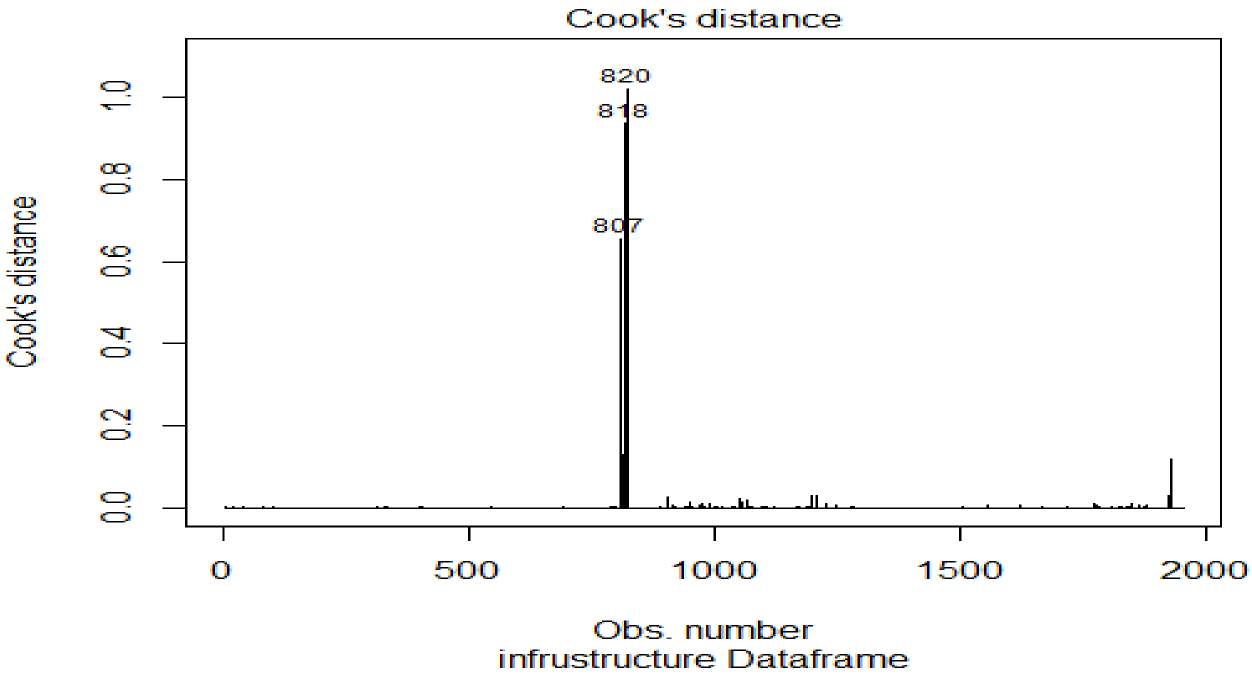
- **Check for normality:**



Here all the points are not close to the line. The p value for the Shapiro wilk test is also less than 0.05, so we reject the null hypothesis. We conclude that the residuals are not normally distributed.

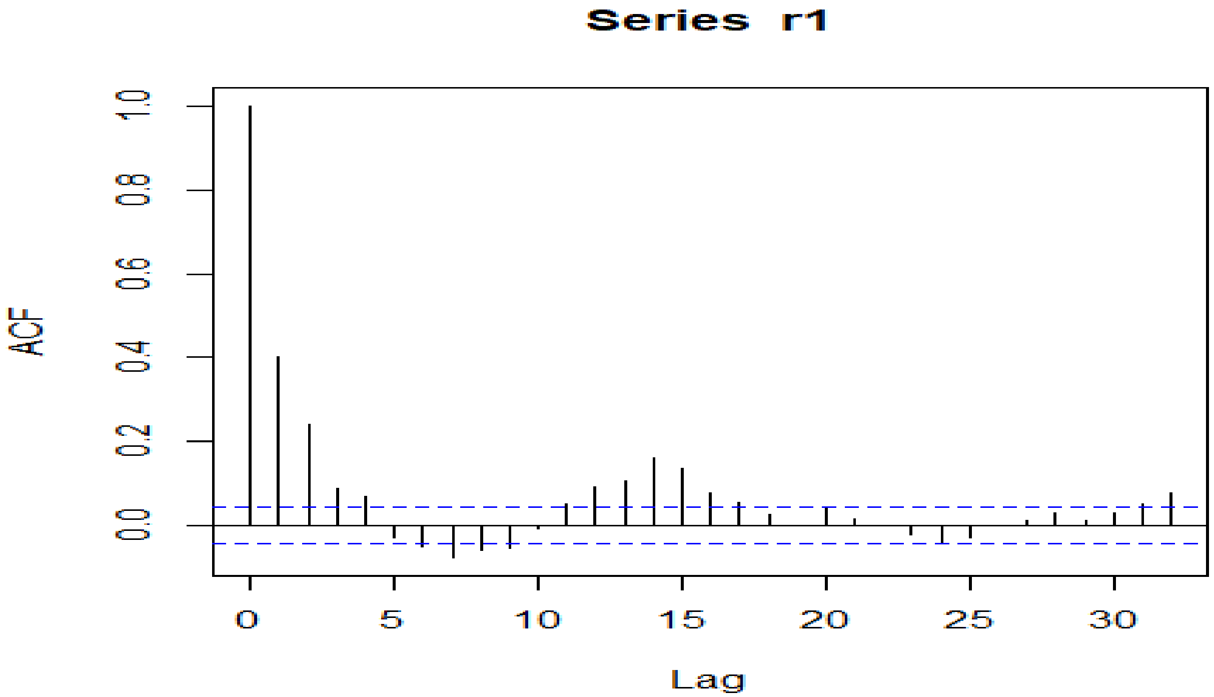
- **Check for outliers:**





From the above diagrams we get that outliers are present in the data.

- **Check for autocorrelation:**



The p value for the Darwin watson test is also less than 0.05, so we reject the null hypothesis. Therefore autocorrelation is present in the residuals.

Now we remove the insignificant variables and fit the model again.

For the new model the value of the multiple R-squared is 0.8399. This implies that almost 83.99% of the variation in the response variable is explained by the predictor variables.

Here the value of the adjusted R-squared is 0.8395. This implies that as we remove the insignificant variables the model accuracy increases.

Consider the Amenity Data Frame: ————

To verify whether the dependent variable is linearly related with the independent variables:

Observe that the prices are moderately positively skewed with gymnasium, sports facility, car parking.

Now we fit the multiple linear regression model:

The following table indicates the estimated value of the regression coefficients, standard error, value of the t statistic, the corresponding p value of the predictors and lastly whether the predictors are significant or not:

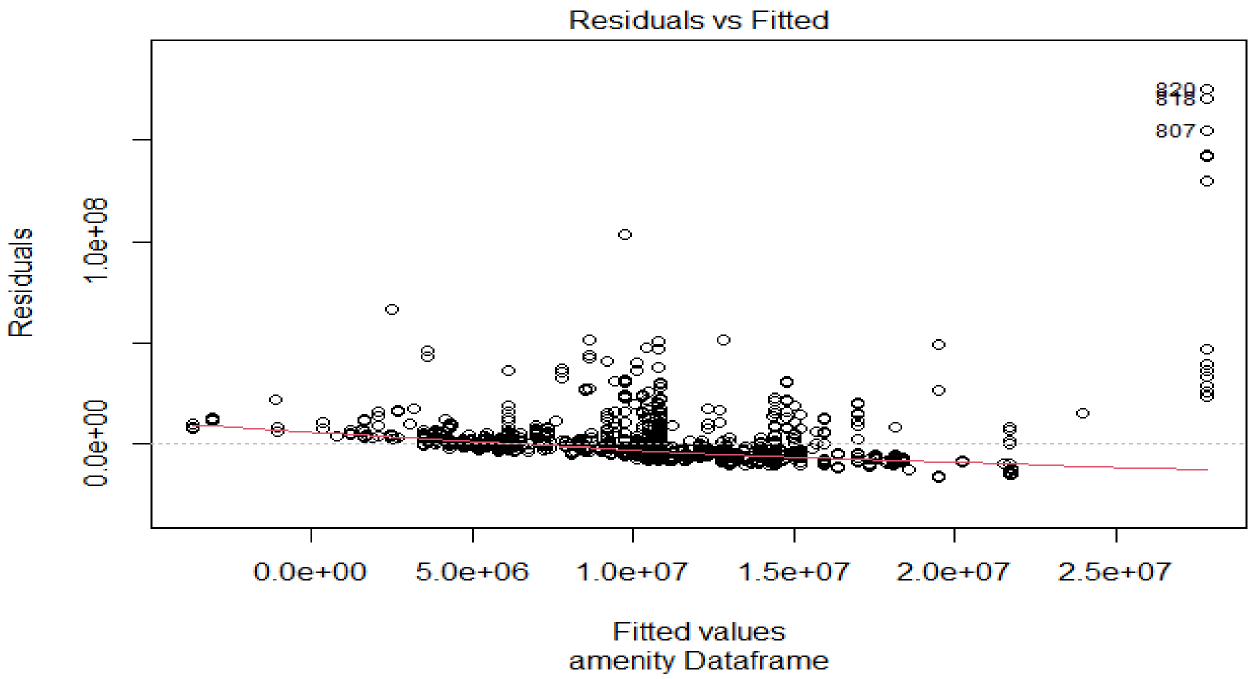
Predictors	Estimated value of coefficient	Standard error	t-statistic	p value	Significant or Not Significant
Intercept	6133806	1310287	4.681	3.05e-06	Significant
MaintenanceStaff	-3626514	939628	-3.860	0.000117	Significant
Gymnasium	-1795367	1424175	-1.261	0.207592	Not significant
SwimmingPool	3718526	1058924	3.512	0.000456	Significant
LandscapedGardens	-5780620	804902	-7.182	9.77e-13	Significant
JoggingTrack	4713145	916799	5.141	3.01e-07	Significant
IndoorGames	1624397	740984	2.192	0.028483	Significant
Intercom	-369377	814532	-0.453	0.650252	Not significant
SportsFacility	4685465	680336	6.887	7.68e-12	Significant
24X7Security	2665898	884099	3.015	0.002600	Significant
PowerBackup	2192110	1251964	1.751	0.080115	Not significant
CarParking	3839513	706866	5.432	6.29e-08	Significant
StaffQuarter	-3585899	823184	-4.356	1.39e-05	Significant
MultipurposeRoom	-3121700	702050	-4.447	9.22e-06	Significant
Childrenplayarea	398154	1222202	0.326	0.744635	Not significant
LiftAvailable	-4485245	959279	-4.676	3.13e-06	Significant
GolfCourse	4149388	1622322	2.558	0.010613	Significant

Therefore, the predictors MaintenanceStaff, SwimmingPool, LandscapedGardens, JoggingTrack, IndoorGames, SportsFacility, 24X7Security, CarParking, StaffQuarter, MultipurposeRoom, LiftAvailable and GolfCourse are significant in predicting the house price. Here the value of the multiple R-squared is 0.1166. This implies that almost 11.66% of the variation in the response variable is explained by the predictor variables.

Here the value of the adjusted R-squared is 0.1093. This implies that the given data does not fit the multiple regression model.

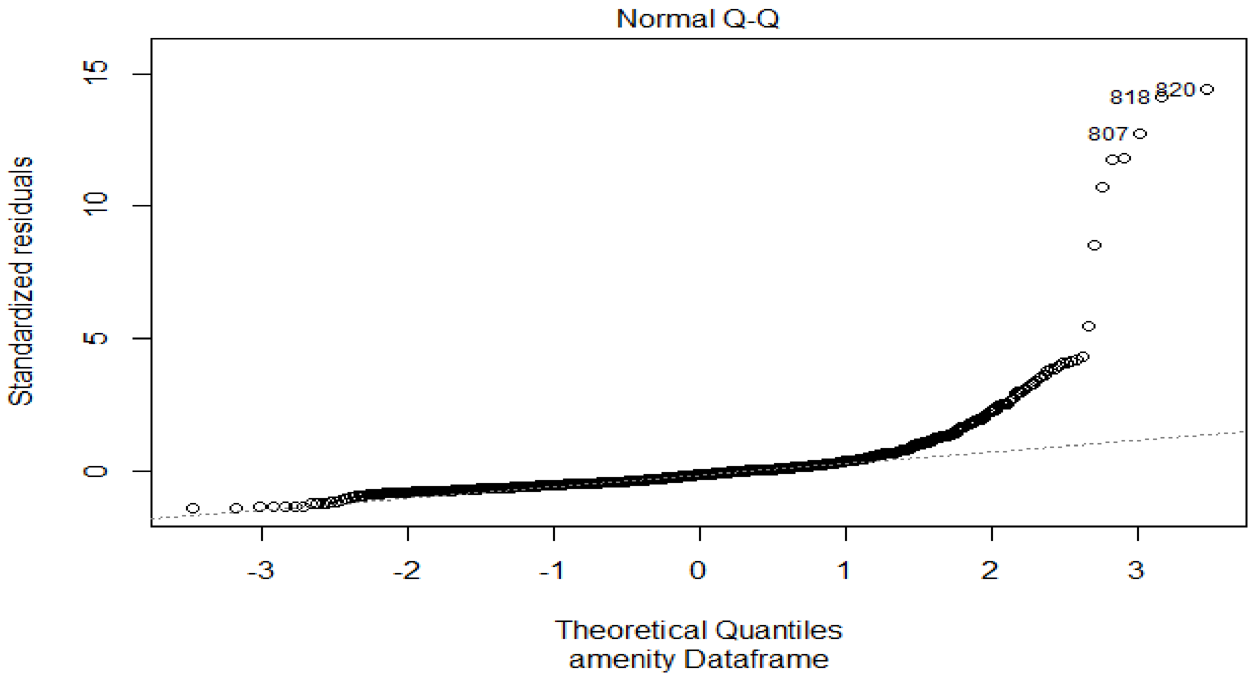
- **Check for multicollinearity:** Here the VIF values for all the predictors are smaller than 5 so, multicollinearity does not exist.

- **Check for homoscedasticity:**



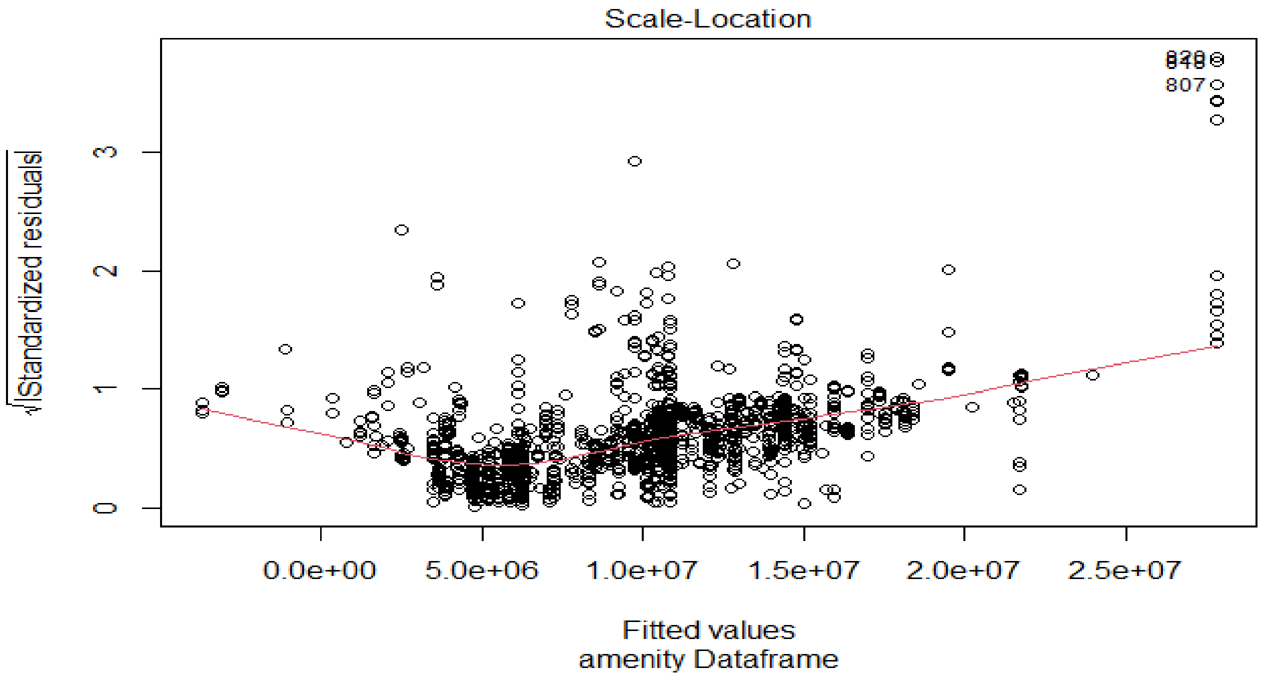
Here the points are not randomly scattered around the zero line. Also the p value for the BP test is less than 0.05, so we reject the null hypothesis. We conclude that the residuals are not homoscedastic.

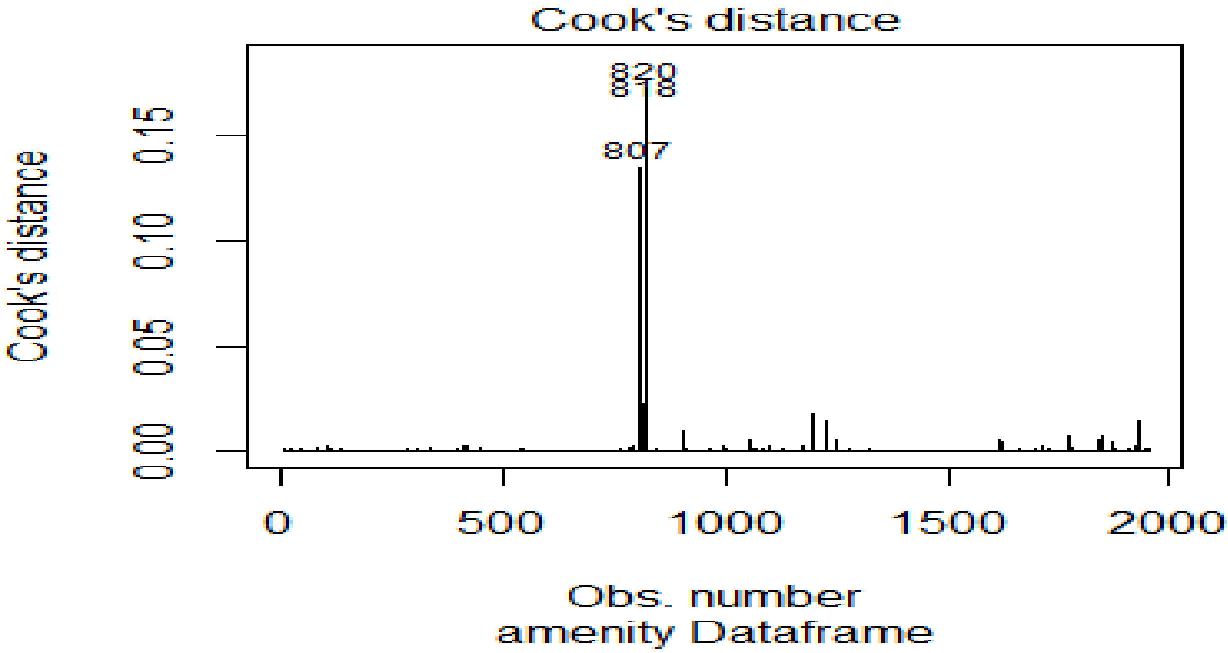
- **Check for normality:**



Here all the points are not close to the line. The p value for the Shapiro wilk test is also less than 0.05, so we reject the null hypothesis. We conclude that the residuals are not normally distributed.

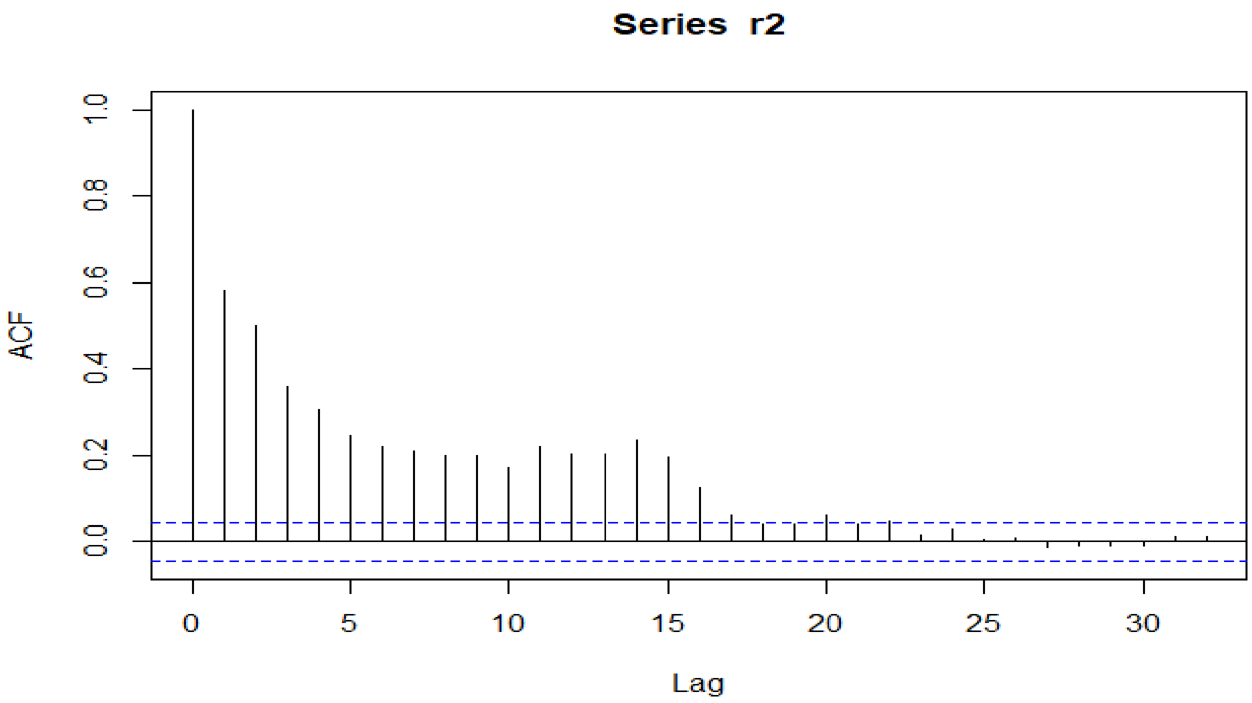
- **Check for outliers:**





From the above diagrams we get that outliers are present in the data.

- **Check for autocorrelation:**



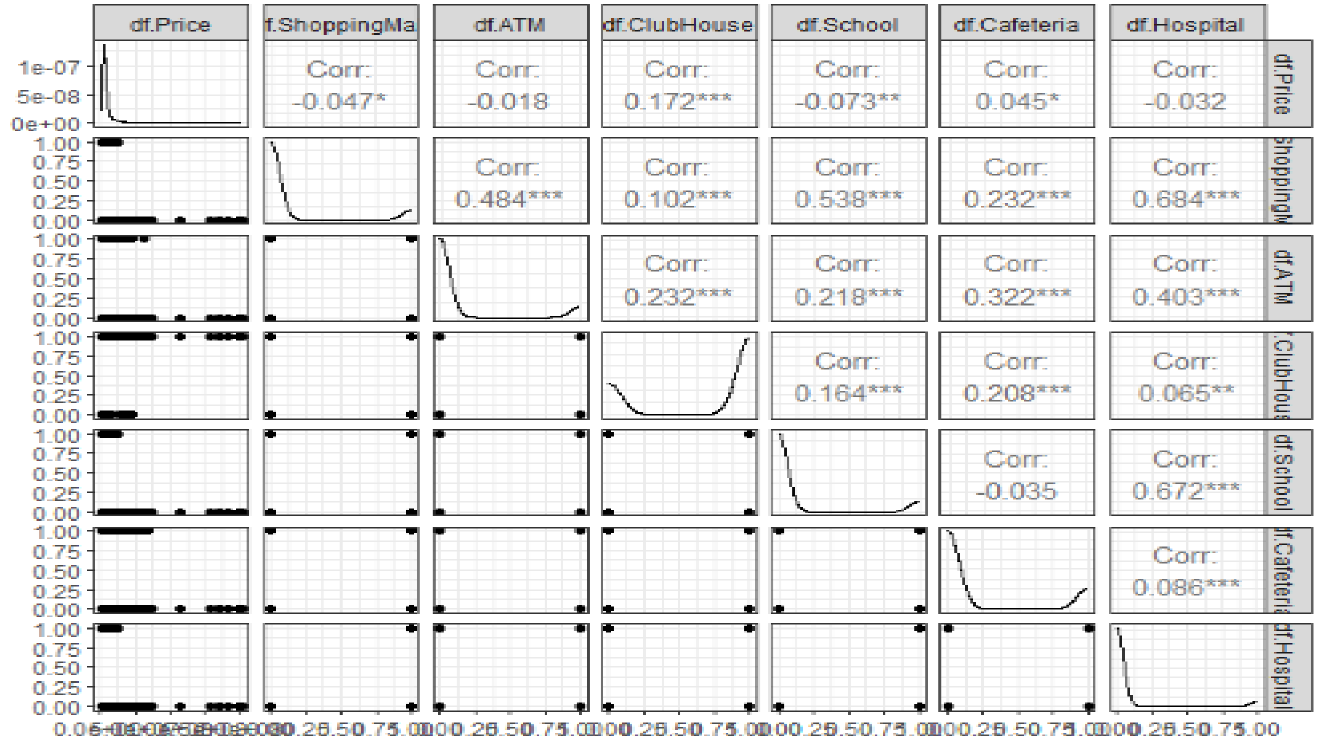
The p value for the Darwin watson test is also less than 0.05, so we reject the null hypothesis. Therefore autocorrelation is present in the residuals.

Now we remove the insignificant variables and fit the model again.

For the new model the value of the multiple R-squared is 0.1157. This implies that almost 11.57% of the variation in the response variable is explained by the predictor variables. Here the value of the adjusted R-squared is 0.1098. This implies that as we remove the insignificant variables the model accuracy is increased to some extent.

Consider the Neighbourhood Data frame: ———

To verify whether the dependent variable is linearly related with the independent variables:



Observe that the prices are moderately positively skewed with club house, cafeteria and moderately negatively correlated with shopping mall, atm, school, hospital.

Now we fit the multiple linear regression model:

The following table indicates the estimated value of the regression coefficients, standard error, value of the t statistic, the corresponding p value of the predictors and lastly whether the predictors are significant or not:

Predictors	Estimated value of coefficient	Standard error	t-statistic	p value	Significant or Not Significant
Intercept	6433529	546999	11.762	<2e-16	Significant
ShoppingMall	-1191771	1343375	-0.887	0.3751	Not significant
ATM	-2488660	1057896	-2.352	0.0187	Significant
ClubHouse	5846780	678298	8.620	<2e-16	Significant
School	-5500566	1263500	-4.353	1.41e-05	Significant
Cafeteria	559803	774001	0.723	0.4696	Not significant
Hospital	5093618	1991596	2.558	0.0106	Significant

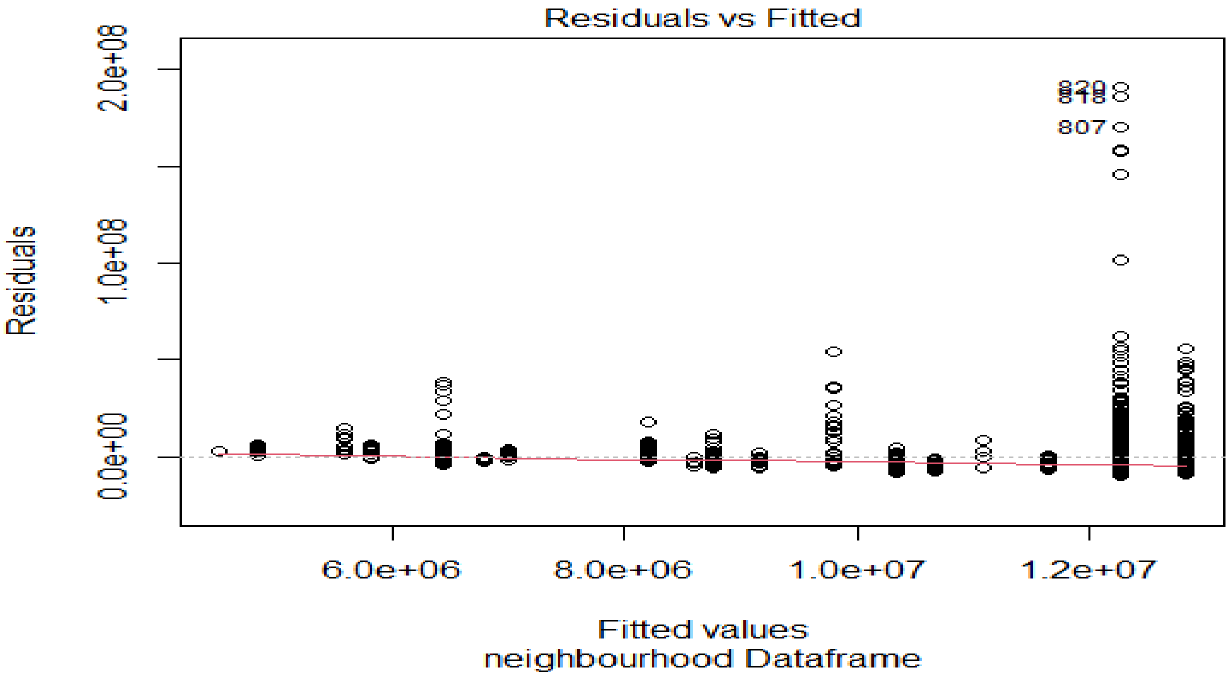
Therefore, the predictors ATM, ClubHouse, School and Hospital are significant in predicting the house price.

Here the value of the multiple R-squared is 0.04541. This implies that almost 4.54% of the variation in the response variable is explained by the predictor variables.

Here the value of the adjusted R-squared is 0.04247. This implies that the given data does not fit the multiple regression model.

- **Check for multicollinearity:** Here the VIF values for all the predictors are smaller than 5 so, these predictors are independent with respect to each other.

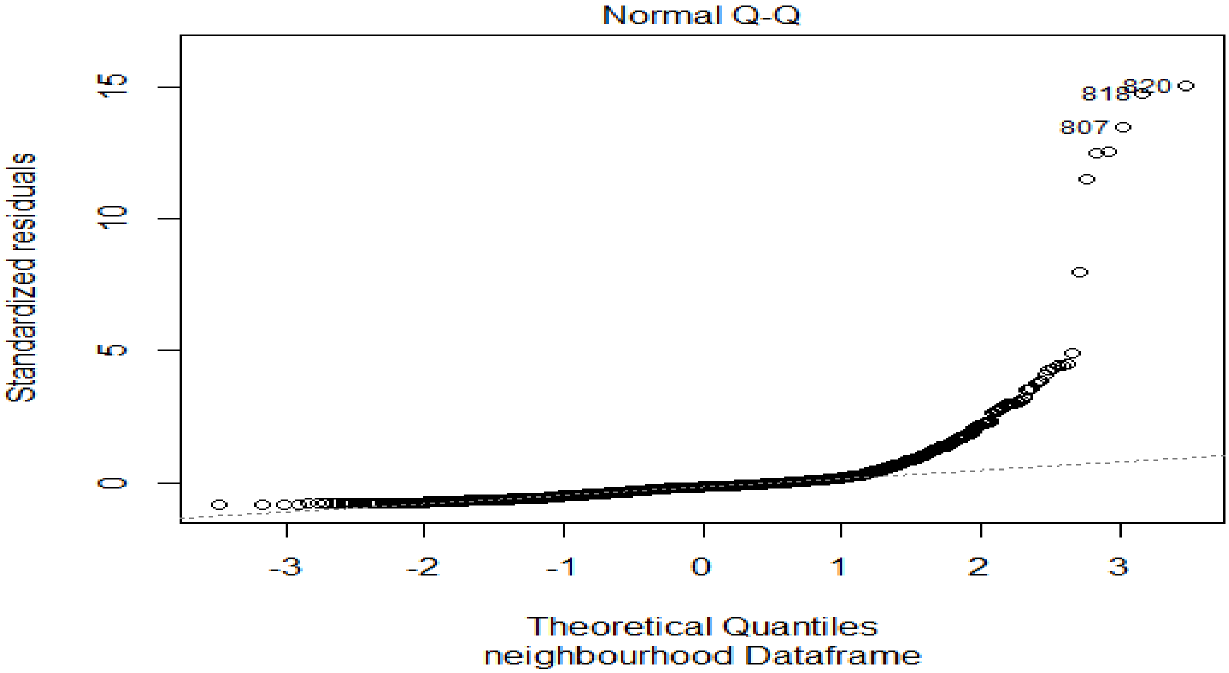
- **Check for multicollinearity:**



Here the points are not randomly scattered around the zero line.

Also the p value for the BP test is less than 0.05, so we reject the null hypothesis. We conclude that the residuals are not homoscedastic.

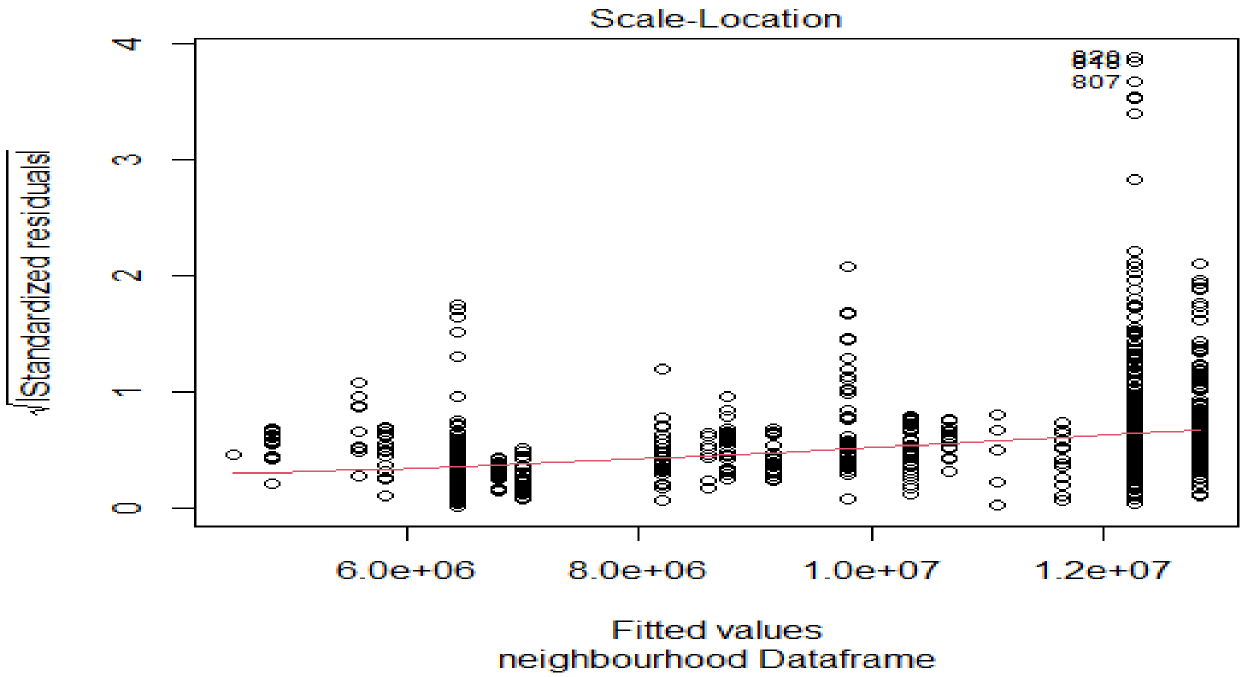
- **Check for normality:**

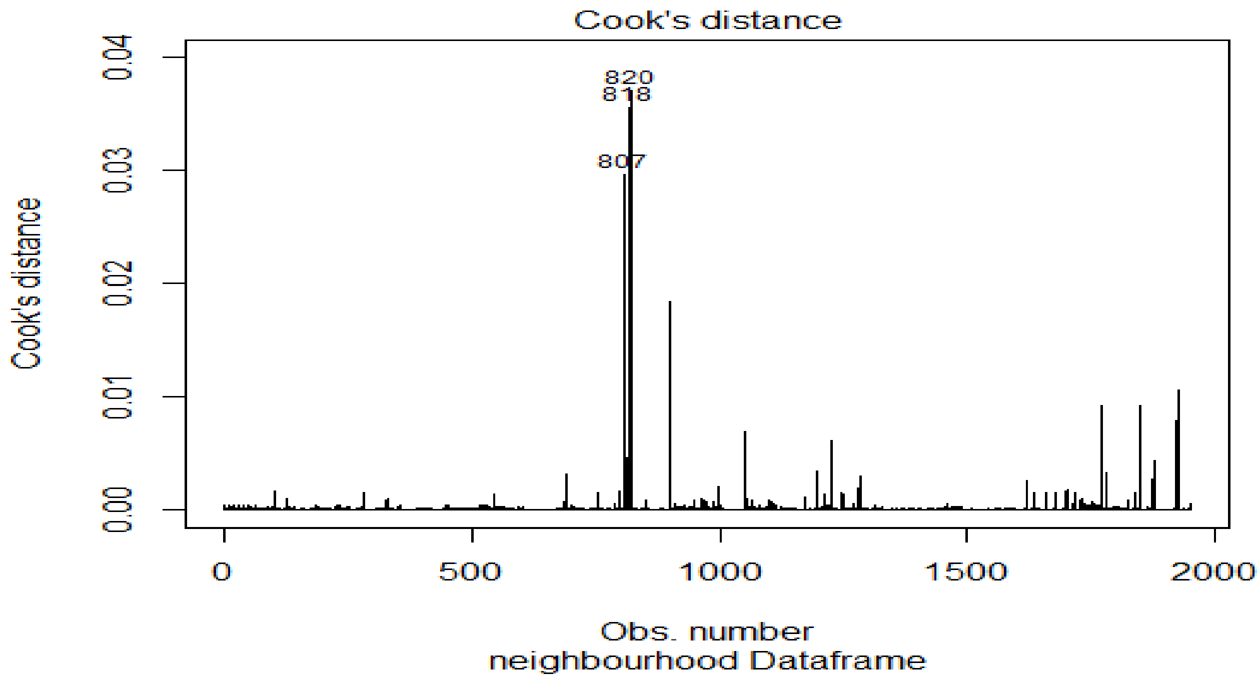


Here all the points are not close to the line.

The p value for the Shapiro wilk test is also less than 0.05, so we reject the null hypothesis. We conclude that the residuals are not normally distributed.

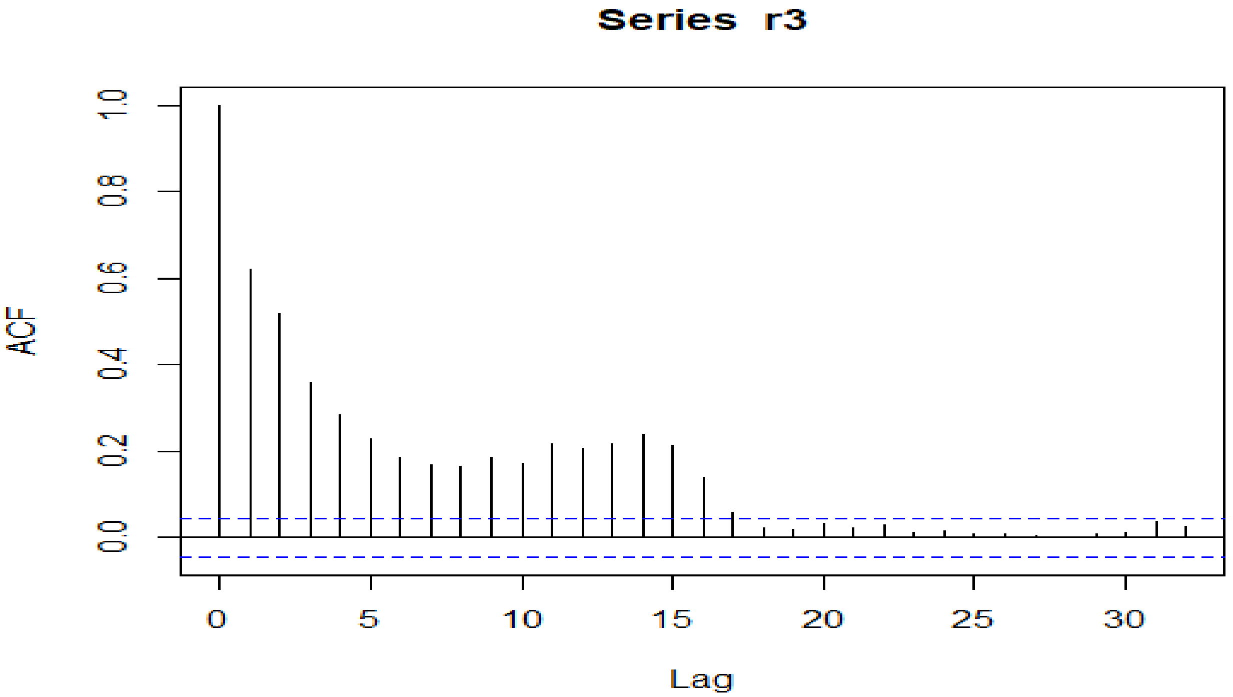
- **Check for outliers:**





From the above diagrams we get that outliers are present in the data.

- **Check for autocorrelation :**



The p value for the Darwin watson test is also less than 0.05, so we reject the null hypothesis.

Therefore autocorrelation is present in the residuals.

Now we remove the insignificant variables and fit the model again.

For the new model the value of the multiple R-squared is 0.04487. This implies that almost 4.5% of the variation in the response variable is explained by the predictor variables.

Here the value of the adjusted R-squared is 0.0429. **This implies that as we remove the insignificant variables the model accuracy is increased to some extent.**

7 Conclusion:

Here we had a large dataset with large number of columns. So as to interpret properly, we create three data frames and fit the multiple linear regression separately.

We get the predictors Area, No. of bedrooms, Resale, Rainwater Harvesting, Vaastu Compliant, MaintenanceStaff, SwimmingPool, LandscapedGardens, JoggingTrack, IndoorGames, SportsFacility, 24X7Security, CarParking, StaffQuarter, MultipurposeRoom, LiftAvailable, GolfCourse, ATM, ClubHouse, School and Hospital are significant in predicting the price of the houses.

And on the basis of the fitted regression model we can predict the price of a house in far future.

8 Future Scope :

Prediction of the house prices are expected to help people who are interested in buying a house so that they can know the price range in the future, then they can plan their finance well. In addition, house price predictions are also helpful for real estate investors to know the future trend of house prices for a particular location. For these purposes this type of study has large scope in the future.

9 References:

[1] P. Durganjali, M. Vani Pujitha, ”House Resale Price Prediction Using Classification Algorithms”, 2019 International Conference on Smart Structure and Systems(ICSSS), Chennai, India, 2019, pp.1-4, doi:10.1109/ICSSS.2019.8882842.

[2] Ayush Varma, Abhijit Sarma, Rohini Nair and Sagar Doshi, ”House Price

Prediction Using Machine Learning And Neural Networks", @2018 IEEE, 2018 Second International Conference on Inventive Communication and Computational Technologies(ICICCT), Coimbatore, India, DOI:10.1109/ICICCT.2018.8473231.

[3] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, "A Hybrid Regression Technique for House Prices Prediction", @2017 IEEE, 2017 IEEE International Conference on Industrial Engineering and Engineering Management(IEEM), Singapore, DOI:10.1109/IEEM.2017.8289904.