Name: Sanika Hande

PRN: 22060641035

Multiple Linear Regression

Introduction:

The dataset is of Real Estate price prediction. Real estate is the property consisting of land and buildings on it, along with its natural resources. Purchasing price of a house depends on various factors such as location, latitude, longitude, convenience facilities, etc. This project focuses on various factors affecting the price of the house. Here the analysis is done in R as well as excel.

Aim:

To fit a multiple linear regression model and study the effect of various factors affecting the house price per unit area.

Data Characteristics:

Data from Kaggle was used for this analysis. There are 414 observations and 8 variables in total. Out of these 8 variables, 7 variables are independent variables and one variable is the dependent variable. There are no missing observations.

The input variables include latitude, longitude, number of convenience stores, house age, date of house purchase, and distance to the nearest railway station.

The output variable is the house price of the unit area

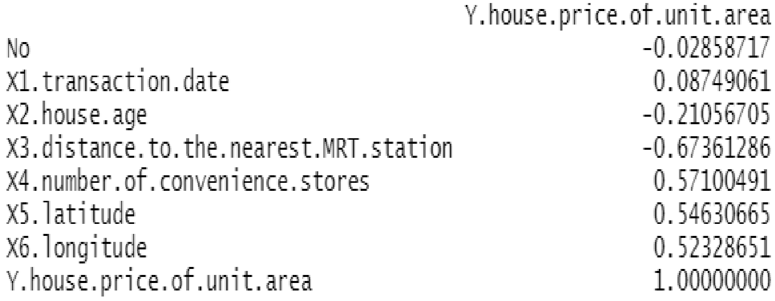
Methodology:

- The first step involves importing the data and checking for missing values.
- Structure of dataset, correlation, and descriptive statistics is calculated.
- Multiple linear regression model is fitted, taking into account only the variables having a positive correlation with the dependent variable.
- The next model is fitted, considering all the independent variables having positive as well as negative correlations with the dependent variable.
- . The adjusted R squared value of all 3 models is found, and the model with the highest value of the adjusted R square is chosen.
- The next step involves plotting the model and checking for the assumptions of linear regression.

- A few tests are also done to check for those conditions.
- We finally conclude whether the model is a good fit or not.

Interpretation:

Correlation between all variables is found. Multicollinearity is absent since no value of the correlation matrix is greater than 0.8.



```
                                        Y.house.price.of.unit.area
No                                                     -0.02858717
X1.transaction.date                                     0.08749061
X2.house.age                                           -0.21056705
X3.distance.to.the.nearest.MRT.station                 -0.67361286
X4.number.of.convenience.stores                         0.57100491
X5.latitude                                             0.54630665
X6.longitude                                            0.52328651
Y.house.price.of.unit.area                              1.00000000
```

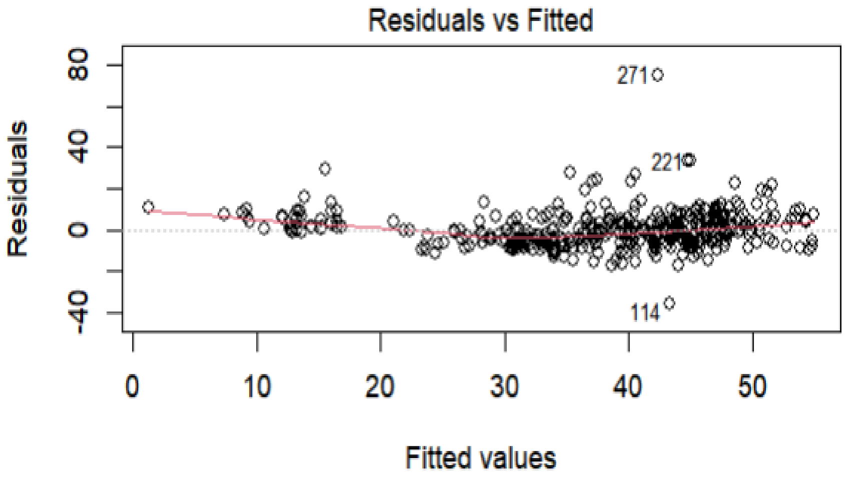
2 models of multiple linear regression are fitted. The first model uses the independent variables having a positive correlation with the dependent variable. The R squared value is 57%. The second model uses all variables having a positive correlation. This model gives an Adjusted R-squared value of 48%.Since the value of the Adjusted R squared value of model 1 is greater, it is a better model than the second model. The regression equation obtained is as follows:

Y= -1.444 + 5.146X1 -2.697X2 – 4.488X3+1.13X4 + 2.25X5 -1.242X6

The assumptions of residual analysis are checked by using the plot command

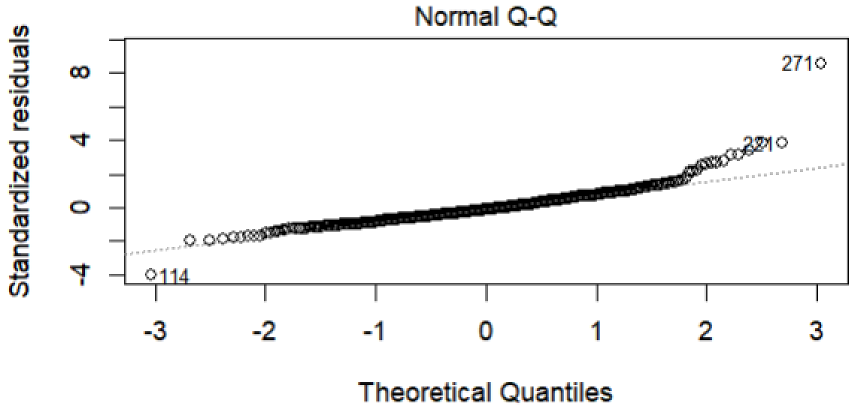
1. Linearity

The residual vs fitted plot is used to check linearity. From the Residual VS Fitted Plot, we can observe an almost straight horizontal line with equally spaced residuals. Hence it is safe to assume that the linearity assumption is satisfied.



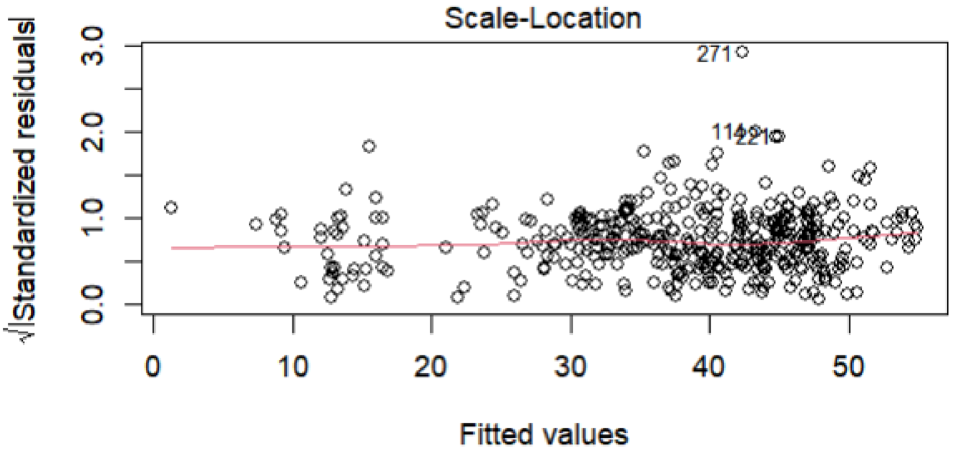
2) Normality

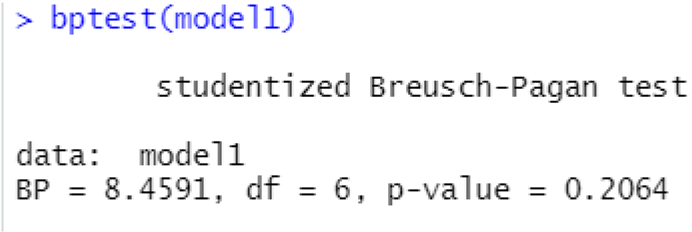
A normal QQ plot is used to check the condition of normality. From Normal QQ. In the plot, we can observe that most observations lie on-line. Hence normality assumption is satisfied**.** Shapiro Wilk test is also performed to check for normality since the p-value is 5.41 which is greater than 0.5 we can say that normality is present.



3) Homoscedasticity

A scale location plot is used to determine homoscedasticity. From the graph, it can be observed that the residuals fall along the horizontal line and hence it can be inferred that homoscedasticity is present. Also from the BP test, we are safe to say that **homoscedasticity is present** since the p-value is >0.05.





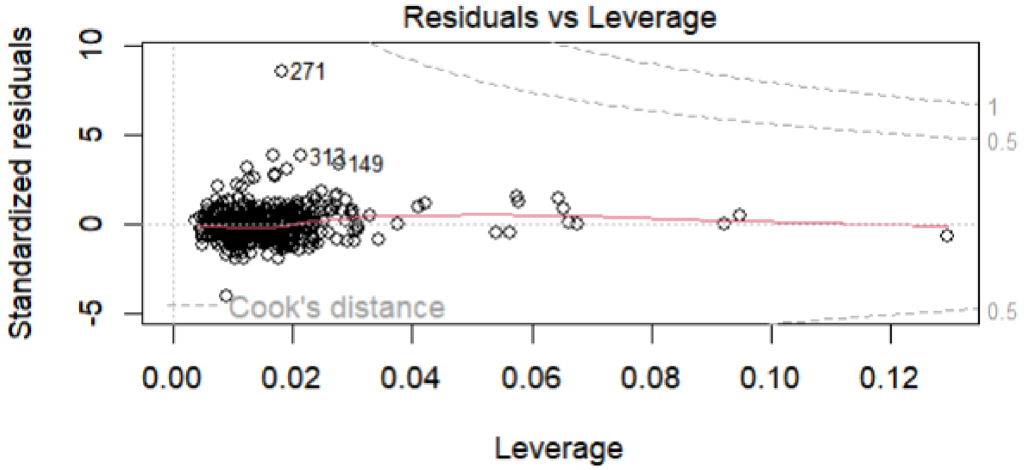
```
> bptest(model1)

        studentized Breusch-Pagan test

data:  model1
BP = 8.4591, df = 6, p-value = 0.2064
```

4) Outlier detection

Residual vs leverage plot is plotted to check for outliers. From the below graph, we can clearly see that there are no outliers present since none of the values lie outside the cook distance. Thus, there are **no outliers** in the data.



Conclusion

From the above interpretation, we can see that the assumptions of linearity, homoscedasticity, outliers, and normality is satisfied. There is also autocorrelation present in the data. Since, the presence of autocorrelation reduces the accuracy of the model, autocorrelation can be removed to increase the accuracy of the model.