Real Estate Price Prediction

REGRESSION MODEL OF REAL ESTATE DATASET

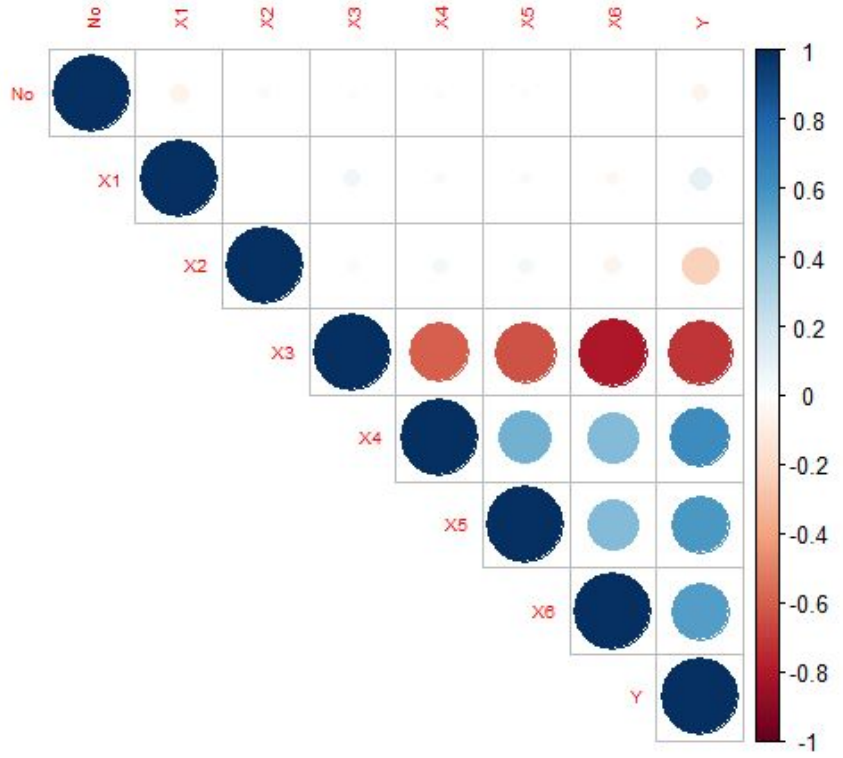
All About Dataset

- Name of Dataset :- Real Estate Data
- Variables:

1. X1 - Transaction Date (Date at which home is bought)
2. X2 - House Age (Age of house from when it was built)
3. X3 - Distance To The Nearest MRT Station
4. X4 - Number Of Convenience Stores
5. X5 - Latitude (Represents the geographical position of property)
6. X6 - Longitude (Represents the geographical position of property)
7. Y - House Price Of Unit Area

- **Source: https://www.kaggle.com/quantbruce/real-estate-price-prediction**

Correlation Plot Between Variables



	X1	X2	X3	X4	X5	X6
X1	1	-0.009308	0.054006	0.024994	0.024389	-0.034151
X2	-0.009308	1	0.025998	0.046235	0.046439	-0.053359
X3	0.054006	0.025998	1	**-0.594963**	**-0.630921**	**-0.714931**
X4	0.024994	0.046235	-0.594963	1	0.472311	0.437196
X5	0.024389	0.046439	-0.630921	0.472311	1	0.435171
X6	-0.034151	-0.053359	-0.714931	0.437196	0.435171	1

Correlation Matrix

- From the correlation plot and correlation matrix we can clearly see that there is a correlation between X3 and X4 , X3 and X5 , X3 and X6 variables.

Multiple Linear Regression Model

Hypothesis:-

H_0: All β's are insignificant.

H_1: At least one β is significant.

Significant variables:- (At 5% level of significance)

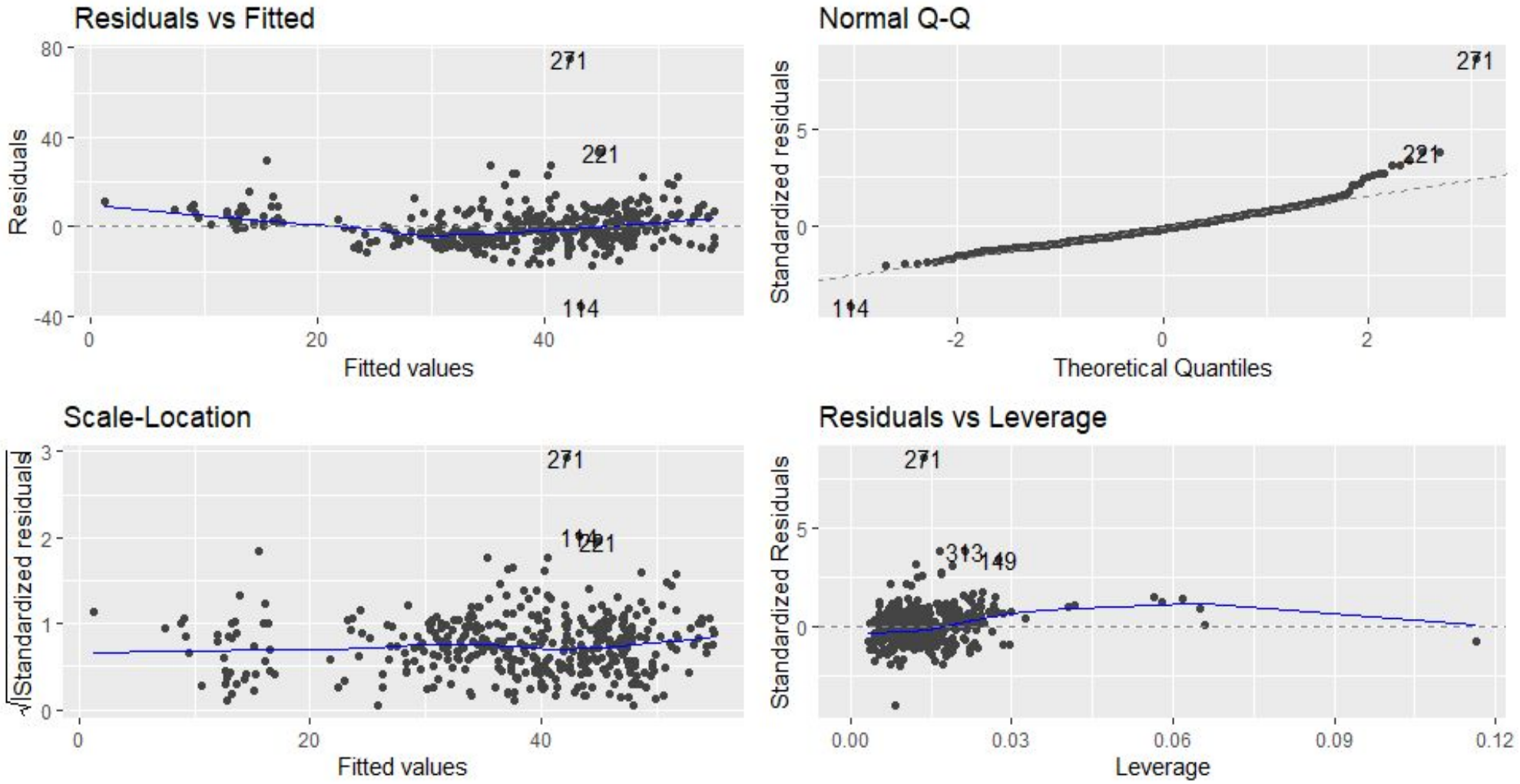
Model:

$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_k X_k$

Y= -15960 + (5.135)***Transaction Date** + (-0.2694)***House Age** + (-0.00435)***Distance To The Nearest MRT Station** + (1.136)***Number Of Convenience Stores** + (226.9)* **Latitude**

Significant Variables	P - Value
Transaction Date	0.00104
House Age	1.04e-11
Distance To The Nearest MRT Station	2e-16
Number Of Convenience Stores	3.17e-09
Latitude	4.36e-07

Assumptions Of Multiple Linear Regression Model



• Multicollinearity

$H_{0:}$ There is no Multicollinearity between the variables.

$H_{1:}$ There is Multicollinearity between the variables.

X1	X2	X3	X4	X5
1.013834	1.013243	2.016855	1.611299	1.585635

Since all the values are lies between 0 < VIF < 5,

Hence there is no multicollinearity among the regressor variables.

• Autocorrelation

$H_{0:}$ Autocorrelation is absent.

$H_{1:}$ Autocorrelation is present.

D-W Statistics	2.1541
P - Value	0.128

Since P-Value is not less than 0.05,

So at 5% l.o.s we do not reject our null hypothesis and conclude that Autocorrelation is absent.

• Heteroscedasticity

$H_{0:}$ Data is homoscedastic (i.e Variance are equal)

$H_{1:}$ Data is heteroscedastic (i.e Variance are not equal)

BP Test	5.7624
P - Value	0.33

Since P-Value is not less than 0.05,

So at 5% l.o.s we do not reject our null hypothesis and conclude that the data is homoscedastic.

• Multivariate Normality

$H_{0:}$ Errors are normally distributed.

$H_{1:}$ Errors are not normally distributed.

K-S Normality Test	0.078433
P - Value	2.105e-06

Since P-Value is less than 0.05,

So at 5% l.o.s we reject our null hypothesis and conclude that the errors are normally distributed.

Summary Of Regression Model

R Squared Value	0.5823
Adjusted R Squared value	0.5772
F - Statistics	113.8
P - Value	2.2e-16

- From the R Squared value we can conclude the 58.23% variation is explained by our model. And also the P - Value is less than 0.05.

- Note that:

If we add interaction term (i.e interaction between X3 and X4 , X3 and X5) and also remove outliers which is present in graphs there is a high chance of getting a good fit model as compared to this model.

Multiple Linear Regression Model

Hypothesis:-

H_0: All β's are insignificant.

H_1: At least one β is significant.

Significant variables:- (At 5% level of significance)

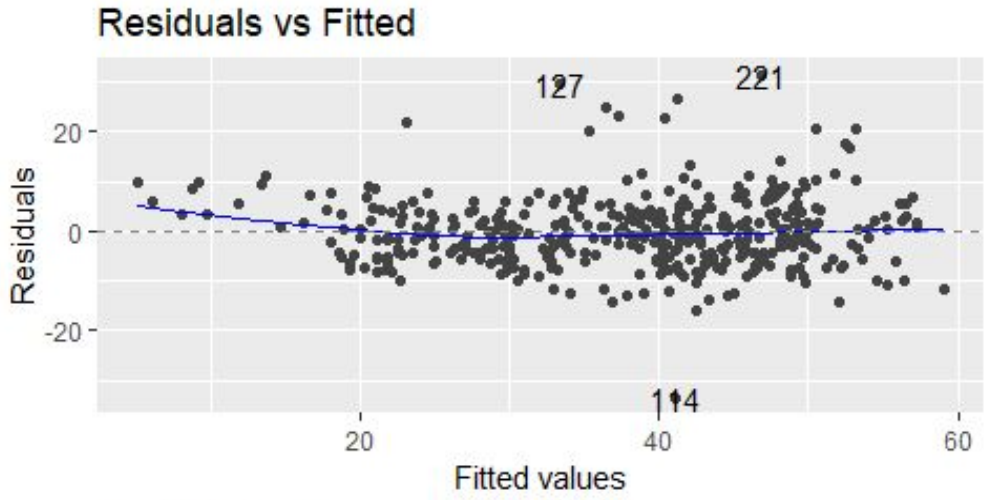
Model:

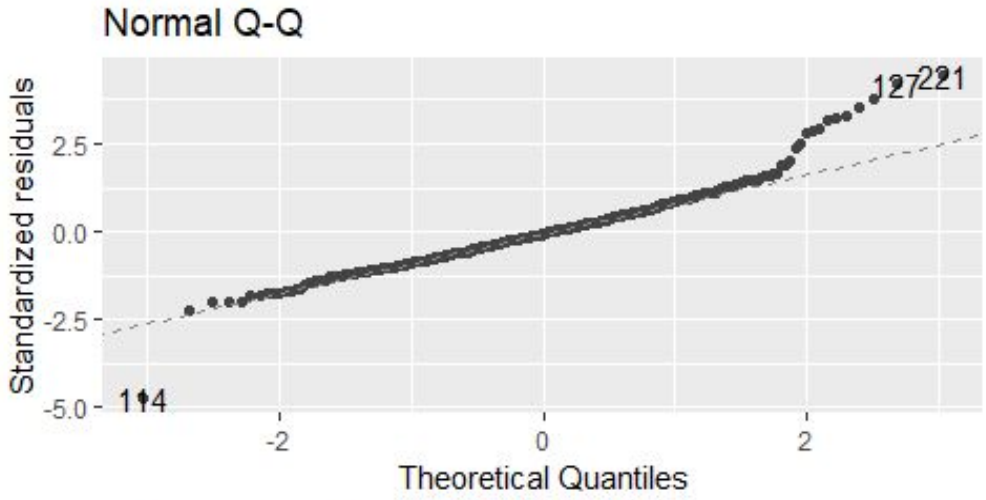
$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_k X_k$

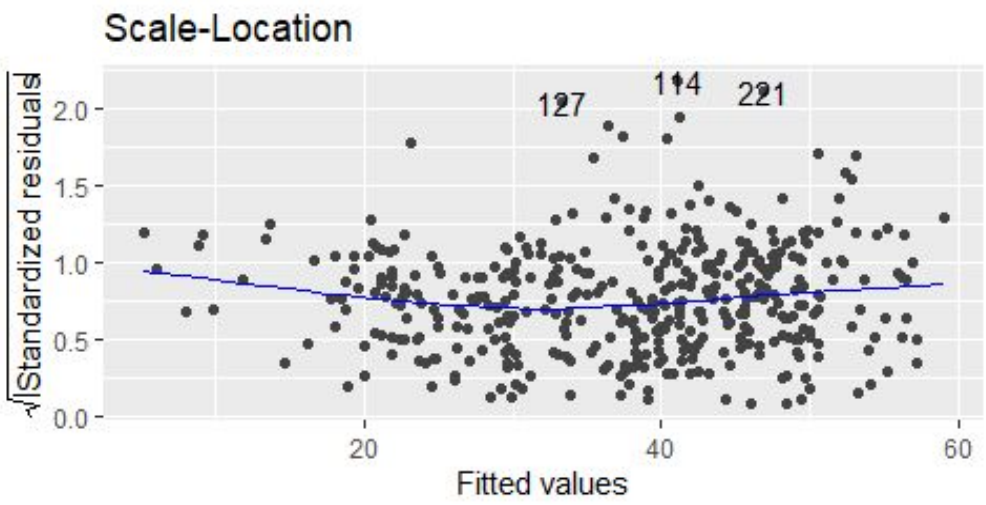
Y= -2.0100 + (4.587)*Transaction Date + (-0.2888)*House Age + (2.920)*Distance To The Nearest MRT Station + (1.430)*Number Of Convenience Stores + (436.8)*Latitude + (-0.00120)*[Distance To The Nearest MRT Station: Number Of Convenience Stores] + (-0.1171)*[Distance To The Nearest MRT Station: Latitude]

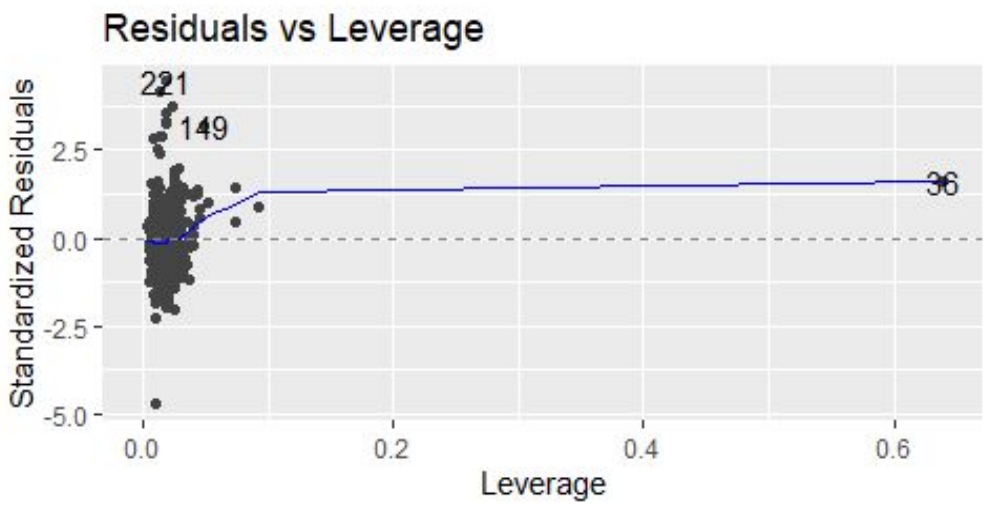
Significant Variables	P - Value
X1	0.000299
X2	2e-16
X3	1.34e-05
X4	1.39e-14
X5	1.18e-15
X3:X4	2.52e-07
X3:X5	1.32e-05

Graphical Representation









Summary Of The Model

R Squared Value	0.7029
Adjusted R Squared value	0.6978
F - Statistics	135.9
P - Value	2.2e-16

- From the R Squared value we can conclude that **70.29%** variation is explained by our model. Now our Multiple Linear Regression Model is good fit for the data.

Conclusion Of The Final Model

1) R-squared Value Of Our Final Model Is 70.29% .
2) From The Residual Vs Fitted Graph We Can See That The Estimated Error Curve Of Our Final Model Is Almost Converge To 0.
3) From The QQ-plot We Can See That The Our Model Behaves Like Normal Except For The Tail Parts.
4) Data Is Homoscedastic.

Thank You !!