Real Estate Price Prediction

Swayam Mehta
BS21DMU003

Date

—

Statistical Data Analysis

—

Dr. Suchismita Das

CONTENTS

1. Introduction

About the dataset:

This is Real Estate pricing data. These estates are spread throughout the city, and the prices have been set based on the convenience and services accessible nearby.

Linear regression makes various data assumptions, including:

- The data's linearity. The predictor (x) and the result (y) are believed to have a linear relationship.
- The residuals' normality. It is assumed that the residual errors are regularly distributed.
- The variance of residuals is homogeneous. The variance of the residuals is considered to be constant
- Error terms in residuals are independent. We should investigate whether this assessment is correct. Potential issues include:
- Relationships between result and predictor are nonlinear.

The data consists of the 7 attributes:

1. Transaction date
2. House age
3. Distance to the nearest MRT station
4. Number of convenience stores
5. Latitude
6. Longitude
7. Price Unit

The dependent variable (X variables) –

- Transaction date, House age, Distance to the nearest MRT station, Number of convenience stores, Latitude, Longitude

The independent variable (Y variable) –

- Price Unit

First 10 Rows of the Dataset:

	Transaction_Date	House_Age	Dist_MRT_station	number_of_conv_stores	Latitude	Longitude	Price_Unit
0	2012.917	32.0	84.87882	10	24.98298	121.54024	37.9
1	2012.917	19.5	306.59470	9	24.98034	121.53951	42.2
2	2013.583	13.3	561.98450	5	24.98746	121.54391	47.3
3	2013.500	13.3	561.98450	5	24.98746	121.54391	54.8
4	2012.833	5.0	390.56840	5	24.97937	121.54245	43.1
5	2012.667	7.1	2175.03000	3	24.96305	121.51254	32.1
6	2012.667	34.5	623.47310	7	24.97933	121.53642	40.3
7	2013.417	20.3	287.60250	6	24.98042	121.54228	46.7
8	2013.500	31.7	5512.03800	1	24.95095	121.48458	18.8
9	2013.417	17.9	1783.18000	3	24.96731	121.51486	22.1

Table 1. 1 First 10 rows

2. Descriptive Statistics

Describing the dataset:

```
> print(data_descriptive)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Transaction_Date	1	200	2013.16	0.29	2013.17	2013.17	0.37	2012.67	2013.58	0.92	-0.16	-1.27	0.02
House_Age	2	200	17.91	11.47	15.75	17.52	12.68	0.00	43.80	43.80	0.35	-0.95	0.81
Dist_MRT_station	3	200	1118.21	1333.37	492.23	832.05	448.42	23.38	6396.28	6372.90	1.74	2.09	94.28
number_of_conv_stores	4	200	4.14	2.86	5.00	4.08	2.97	0.00	10.00	10.00	0.02	-1.13	0.20
Latitude	5	200	24.97	0.01	24.97	24.97	0.01	24.93	25.01	0.08	-0.41	0.18	0.00
Longitude	6	200	121.53	0.02	121.54	121.54	0.01	121.48	121.57	0.09	-1.24	0.78	0.00
Price_Unit	7	200	38.40	13.51	39.40	38.44	14.68	7.60	73.60	66.00	-0.02	-0.60	0.96

Table 2. 1 – Descriptive statistics

The following major observations may be drawn from the above table output:

- The Price Unit (Y variable) has a mean of 38.4, which is lower than its median of 39.4, showing a negative skew of - 0.02.

- The distance to the nearest MRT station and the longitude have a high kurtosis (4th derivative of the moment generating function) of 2.09 and 0.78, respectively.

Graphs and fitted lines:

Let us create some key plots to better understand the distribution of our dependent variable:

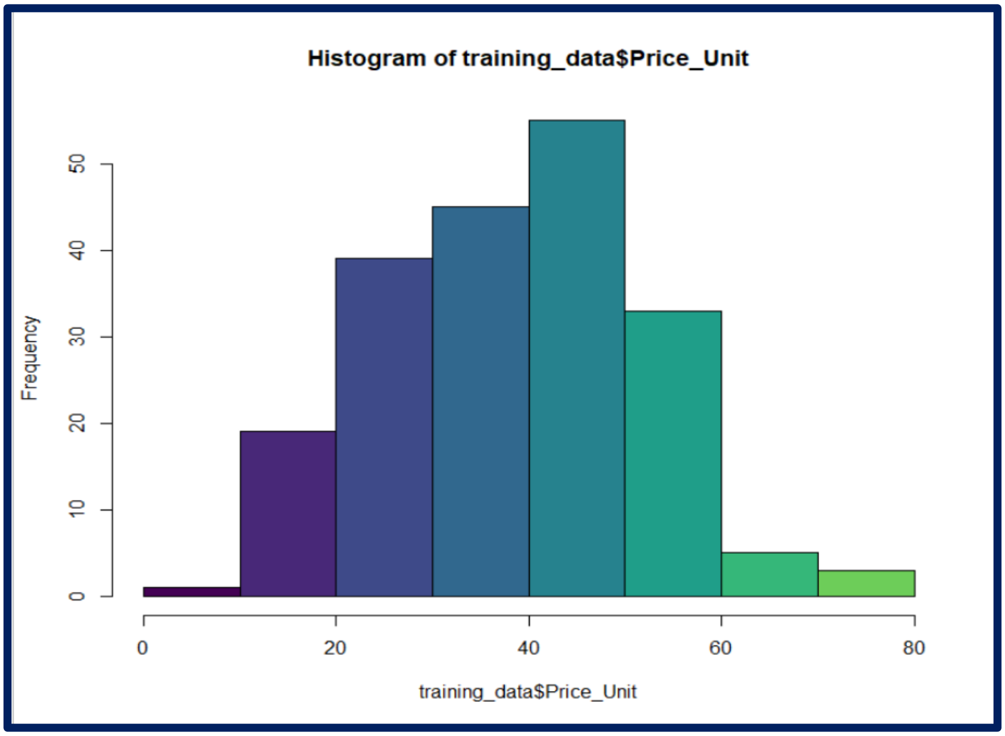
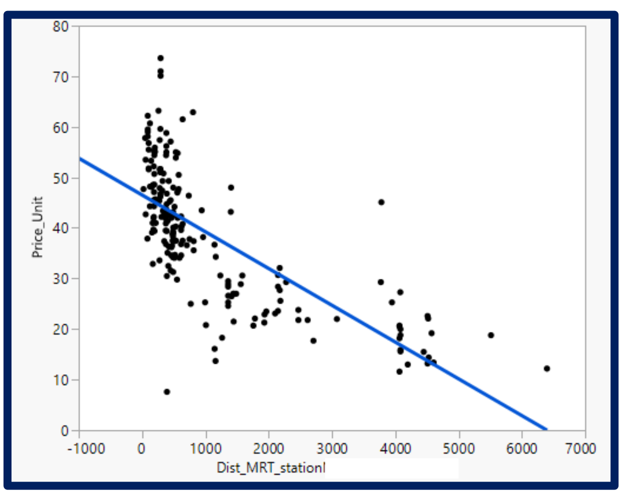


Figure 2.1 – Histogram of Price Unit

As the above figure is a normal approximation which have a normal bell curve.

Scatter Plots:



Price Unit = 46.537567 - 0.0072733 * Dist_MRT_station

Price Unit v/s Dist_MRT_station
Price Unit and Dist_MRT_station appear to be strongly negatively correlated as the points seem to fall on a line. There is a less possibility of a linear relationship.

Figure 2.1 – Scatter plot of Price Unit against Dist_MRT_station

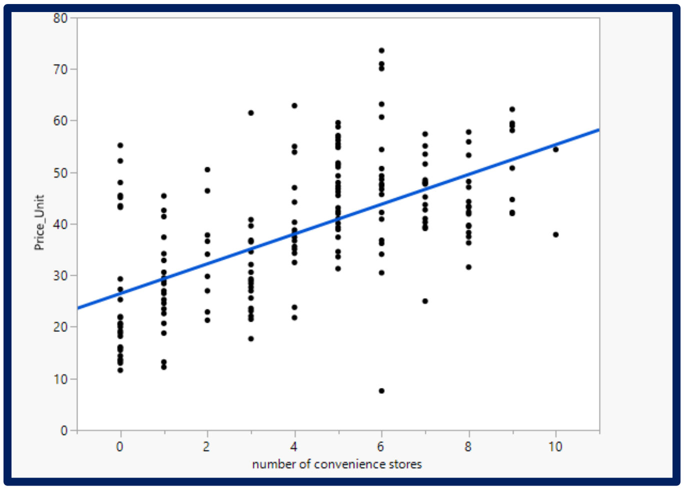


Figure 2.3 – Scatter plot of Price Unit against Number of conv. stores

Price Unit = 26.463564 + 2.8877716 * number of convenience stores

Price Unit v/s Number of conv. stores
Price Unit and number of convenience stores appear to have a positive linear relationship with few points as many of them are away from the line which tells there are residuals also present

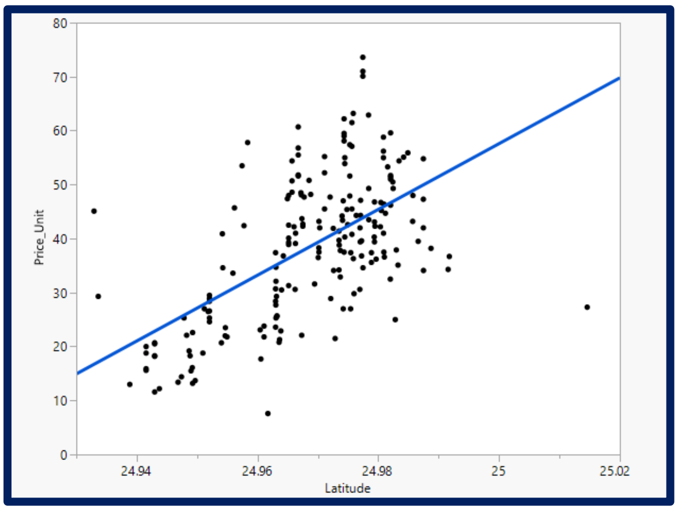


Figure 2.4 – Scatter plot of Price Unit against Latitude

Price Unit = -15180.83 + 609.53739 * Latitude

Price Unit v/s Latitude
Price Unit and Latitude appear to be strongly positively correlated as the points seem to fall on a line. There is a strong possibility of a linear relationship.

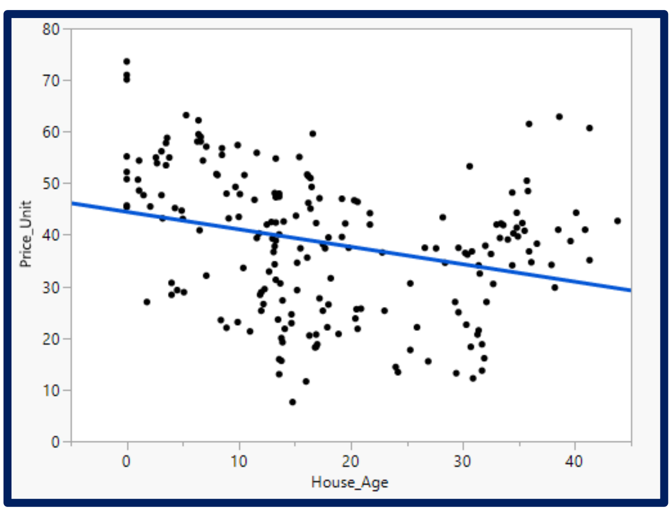


Figure 2.5 – Scatter plot of Price Unit against House age

Price Unit = 44.467135 - 0.3384111 * House Age

Price Unit v/s House Age
Price Unit against House age appear to have less relationship between them as points are scattered on both the sides of the line.

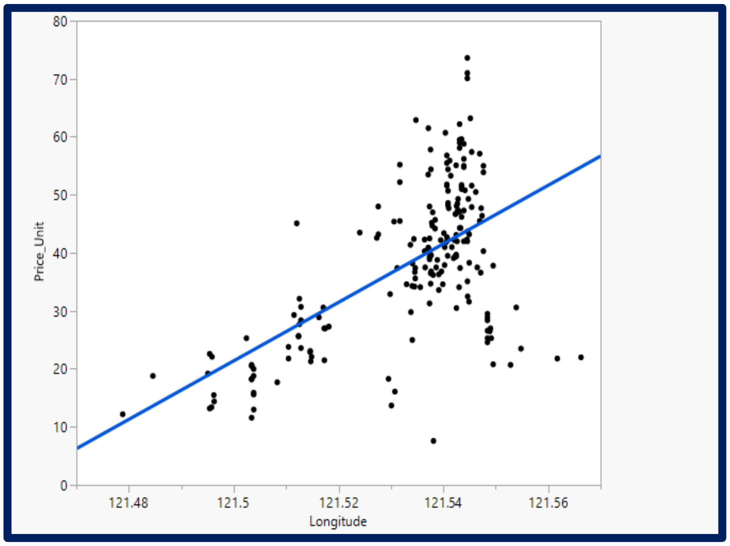


Figure 2.4 – Scatter plot of Price Unit against Longitude

Price Unit = -61230.64 + 504.13195 * Longitude

Price Unit v/s Longitude
Price Unit and Longitude appear to be positively correlated as the points seem to fall on a line. There is a strong possibility of a linear relationship.

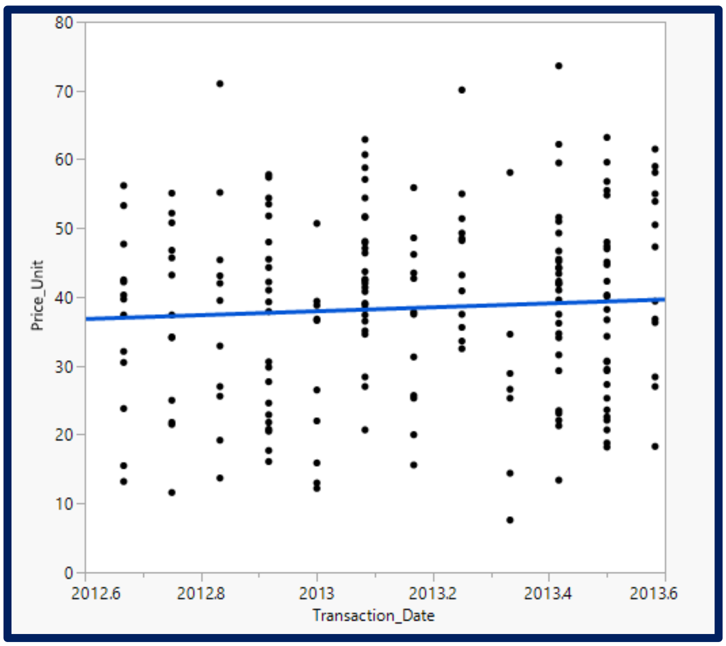


Figure 2.4 – Scatter plot of Price Unit against Transaction date

Price Unit = -5672.611 + 2.8368365 * Transaction Date

Price Unit v/s Dist_MRT_station
Price Unit and Transaction appear to be not correlated as the points are apart from the line on both the sides. There is a zero correlation.

3. Correlation Chart

The correlation coefficient between two random variables X and Y, represented by r(X, Y) or r_{XY}, is a numerical measure of their linear connection and is defined as:

Correlation coefficient (r) matrix of numeric variables:

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

OR

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- r_{XY} provided a measure of linear relationship between X and Y.
- It is a measure of degree of relationship.

	Transaction_Date	House_Age	Dist_MRT_station	number_of_conv_stores	Latitude	Longitude	Price_Unit
Transaction_Date	1.000000	0.000244	0.070241	-0.006477	0.012240	-0.050373	0.060498
House_Age	0.000244	1.000000	0.058864	0.014621	0.023488	-0.103493	-0.287212
Dist_MRT_station	0.070241	0.058864	1.000000	-0.664152	-0.648924	-0.831703	-0.717778
number_of_conv_stores	-0.006477	0.014621	-0.664152	1.000000	0.499150	0.497570	0.610455
Latitude	0.012240	0.023488	-0.648924	0.499150	1.000000	0.508376	0.588713
Longitude	-0.050373	-0.103493	-0.831703	0.497570	0.508376	1.000000	0.587680
Price_Unit	0.060498	-0.287212	-0.717778	0.610455	0.588713	0.587680	1.000000

Figure 3.1 – Correlation Matrix

4. Multiple Linear Regression Prediction Model

We'll use the training data to train our regression model.

A multiple linear regression model looks like this:

$$Y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta k{-}1xk{-}1 + \varepsilon \ldots \ldots \ldots \ldots (1)$$

For Hypothesis testing and the setting of confidence limits, we also assume that ε is normally distributed.

The linearity of the model (1) is defined with respect to the regression coefficients X variables $\beta 1$, $\beta 2$ $etc.$ … in the test are as follows:

1. Transaction date
2. House age
3. Distance to the nearest MRT station
4. Number of convenience stores
5. Latitude
6. Longitude

Y variable for the model is:

1. Price Unit

Regression Output Coefficients and p-value:

```
Residuals:
    Min      1Q  Median      3Q     Max
-36.269  -5.056  -0.733   4.464  28.297

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -1.173e+04  9.147e+03  -1.283 0.201155
training_data$Transaction_Date    4.188e+00  2.016e+00   2.077 0.039107 *
training_data$House_Age          -3.200e-01  5.087e-02  -6.292 2.06e-09 ***
training_data$Dist_MRT_station   -4.426e-03  9.883e-04  -4.479 1.28e-05 ***
training_data$number_of_conv_stores  1.099e+00  2.750e-01   3.995 9.19e-05 ***
training_data$Latitude            2.112e+02  5.893e+01   3.584 0.000428 ***
training_data$Longitude          -1.588e+01  6.695e+01  -0.237 0.812794
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.14 on 193 degrees of freedom
Multiple R-squared:  0.648,	Adjusted R-squared:  0.637
F-statistic: 59.21 on 6 and 193 DF,  p-value: < 2.2e-16
```

Figure 3.1 – Model 1 output

Statistic	Value
Residual standard error	8.14
Multiple R-squared	0.648
Adjusted R-squared	0.637

Table 3.1 – Model output

Model	df	F	p value
Regression	6	59.21	< 2.2e-16
Residual	193		
Total	199		

Table 3.2 – Model output

Explanation of the terms in the table –

1. Multiple R - square root of R2
2. R square – Coefficient of determination given be the formula:

$$R^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

Where, $\text{SSResid} = \Sigma(Y - \hat{Y})^2$

$\text{SST}_o = \Sigma(Y - \bar{Y})^2$

The given results suggest that the P-Value for attribute Longitude is not significant. This clearly indicates that these options have little or no effect on the outcome. Furthermore, the intercept for this model is insignificant.

Model 2 : Removing all the insignificant values

We will now run the model again with these variables:

1. Transaction date
2. House age
3. Distance to the nearest MRT station
4. Number of convenience stores
5. Latitude
6. ~~Longitude~~

```
Residuals:
    Min      1Q  Median      3Q     Max 
-36.218  -4.992  -0.661   4.465  28.250 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     -1.366e+04  4.206e+03  -3.247 0.001375 ** 
training_data$Transaction_Date   4.176e+00  2.011e+00   2.077 0.039123 *  
training_data$House_Age         -3.190e-01  5.056e-02  -6.310 1.85e-09 ***
training_data$Dist_MRT_station  -4.254e-03  6.711e-04  -6.339 1.58e-09 ***
training_data$number_of_conv_stores 1.107e+00  2.724e-01   4.063 7.03e-05 ***
training_data$Latitude           2.120e+02  5.870e+01   3.611 0.000388 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.12 on 194 degrees of freedom
Multiple R-squared:  0.6479,	Adjusted R-squared:  0.6388 
F-statistic: 71.39 on 5 and 194 DF,  p-value: < 2.2e-16
```

Figure 3.2 – Model 2 output

Statistic	Value
Residual standard error	8.14
Multiple R-squared	0.648
Adjusted R-squared	0.639

Table 3.3 – Model 2 output

Model	df	F	p value
Regression	6	71.39	< 2.2e-16
Residual	194		
Total	200		

Table 3.4– Model 2 output

All the factors in this model are important, and the model accurately predicts price in 64.8% of the situations. Adjusted-R^2 is necessary for comparing models with changing numbers of variables, because R^2 will always be higher or equal for a model with a bigger number of independent variables.

5. Model Validation

To assess the model, we will use the VIC test (Variance Inflation factor) and step AIC to see whether the model is optimal:

The **variance inflation factor (VIF),** which evaluates the correlation and intensity of correlation between the predictor variables in a regression model, is the most often used method for detecting multicollinearity.

VIFs are calculated by regressing one predictor against each other predictor in the model. This yields the R-squared values, which can then be input into the VIF calculation.

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

Where Ri is the coefficient for regressing xi on other x's

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

Criteria: VIF > 5 can be an indication of multi collinearity.

Reducing Multicollinearity: Eliminate one or more strongly correlated independent variables.
Output :

```
> vif(model)
    training_data$Transaction_Date            training_data$House_Age     training_data$Dist_MRT_station
                          1.013066                               1.014325                           2.416780
training_data$number_of_conv_stores             training_data$Latitude
                          1.826355                               1.771015
```

Figure 5.1 – VIF Test

We will not use the Step AIC approach because our VIF values are in the criterion.

6. Residuals and QQ plot

Plot – Residuals against model :

The correlation between the residuals and the predicted results, the more the residual value is around 0, the better the predicted results. This graph shows that the residual concentrated around y = 0 so the hypotheses: The error term has a population mean of zero is acceptable.

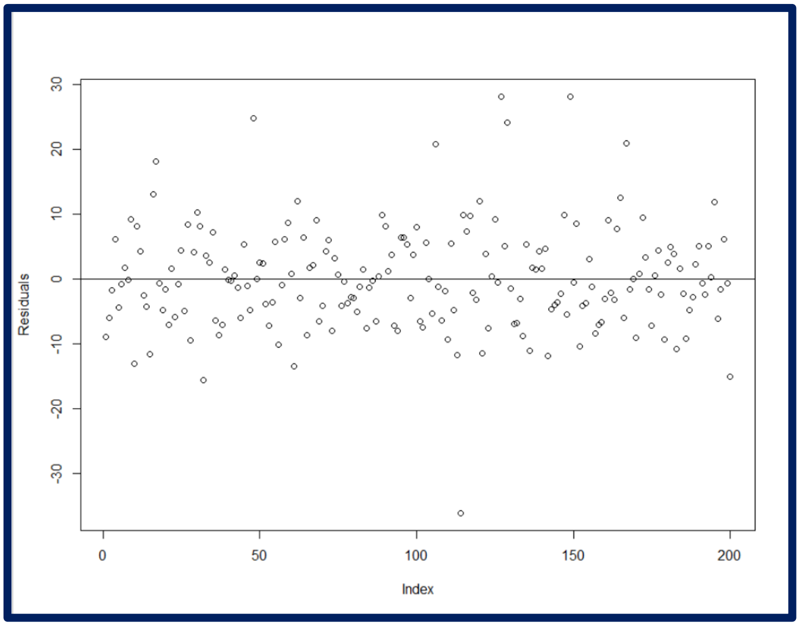


Figure 6.1 – Residuals plot against model

Normal QQ Plot :

Check if the residual has a N(0,0) normal distribution.

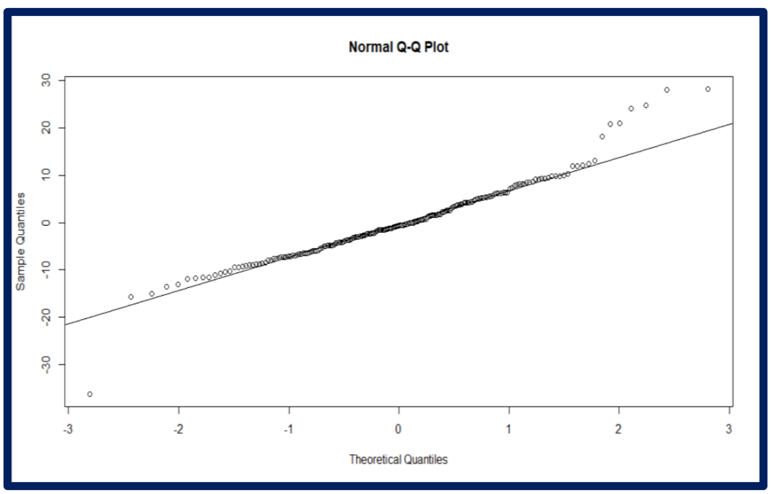


Figure 6.2 – Normal QQ Plot

Analogical reasoning: We can see that the assumption that the data is linearly distributed is valid from both the residual vs fitted plot and the Normal QQ plot. However, several outliers are visible in both of the graphs above.

Influence Index Plot:
Identifying and eliminating outliers in data to enhance model performance.
We'll put the model 2 regressor into the plot and see what happens:

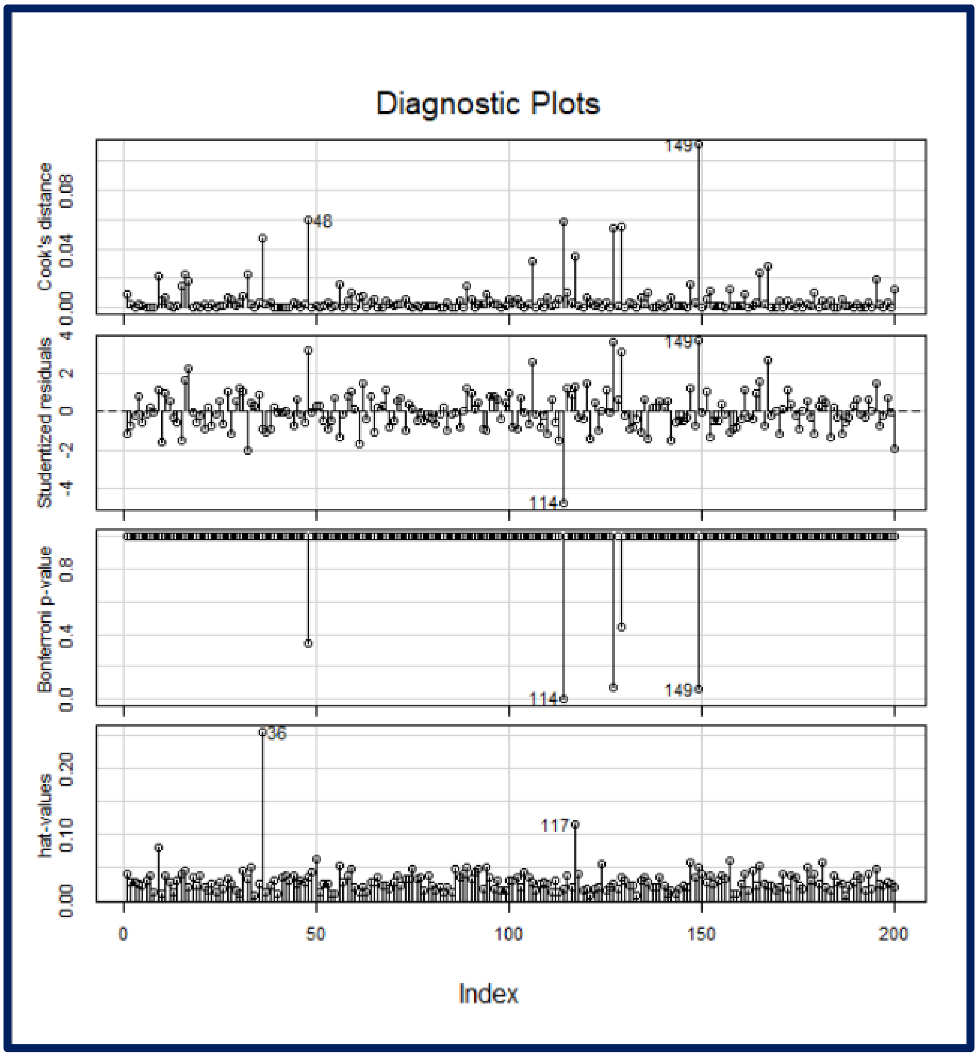


Figure 6.3 – Influence Index Plot

7. Hypothesis testing

To check model utility.

Hypothesis -

$$H_0 : \quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

The dependent and independent variables have no linear relationship.

$$H_A : \quad \beta_j \neq 0 \text{ where } j = 1,2,3,4,5$$

There is at least one independent variable which has a linear relationship with dependent variable.

Anova Table

Formulas -

Source	Sum of squares	Degree of Freedom	Mean squares	F
Treatment	SS_T	k-1	$MS_T = \frac{SS_T}{k-1}$	$F = \frac{MS_T}{MS_E}$
Error	SS_E	N-k	$MS_E = \frac{SS_E}{N-k}$	
Total	TotalSS	N-1		

Figure 7.1 – ANOVA Table Formula

```
> anova(model2)
Analysis of Variance Table

Response: training_data$Price_Unit
                                  Df  Sum Sq Mean Sq  F value    Pr(>F)    
training_data$Transaction_Date     1   133.0   133.0   2.0164 0.1572165    
training_data$House_Age            1  2997.0  2997.0  45.4509 1.746e-10 ***
training_data$Dist_MRT_station     1 18214.9 18214.9 276.2402 < 2.2e-16 ***
training_data$number_of_conv_stores  1  1330.3  1330.3  20.1754 1.212e-05 ***
training_data$Latitude             1   860.0   860.0  13.0424 0.0003877 ***
Residuals                        194 12792.1    65.9                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> |
```

Figure 7.2 – ANOVA Table for the model 2

With significance α = 0.05

F – critical (df1 = 4, df2 = 194) = 2.418 ,

F – statistic = 71.38506

Since F – statistic > F – critical

We reject the null hypothesis. Hence, there is at least one independent variable which has a linear relationship with the dependent variable.

Regression coefficient table and final model

The following table gives the value, standard error (SE), t statistic, p-value and confidence interval of regression coefficients –

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-13656.73963	4206.067543	-3.246914008	0.001374527	-21952.23008	-5361.24917	-21952.23008	-5361.24917
Transaction_Date	4.176344145	2.010800076	2.076956429	0.039122837	0.210508441	8.142179849	0.210508441	8.142179849
House_Age	-0.319013969	0.050557357	-6.309941557	1.85E-09	-0.418726603	-0.219301335	-0.418726603	-0.219301335
Dist_MRT_station	-0.004254374	0.000671138	-6.339040299	1.58E-09	-0.005578038	-0.002930709	-0.005578038	-0.002930709
number_of_conv_stores	1.106658015	0.272367518	4.063105701	7.03E-05	0.569476402	1.643839627	0.569476402	1.643839627
Latitude	212.002239	58.70313526	3.611429578	0.000387732	96.22395128	327.7805267	96.22395128	327.7805267

Figure 7.3 - Regression coefficient table and final model

Analysis of the coefficient table:

- $\beta_{1,}$ $\beta_{4,}$ $\beta 5$ are positive and β_2, β_3 is negative.

A simple summary of the above output is that the fitted line is:

$$\hat{Y} = 4.176X_1 - 0.319X_2 - 0.004X_3 + 1.107X_4 + 212X_5 - 13656.74$$

8. Prediction

Actual by Predicted Plot:

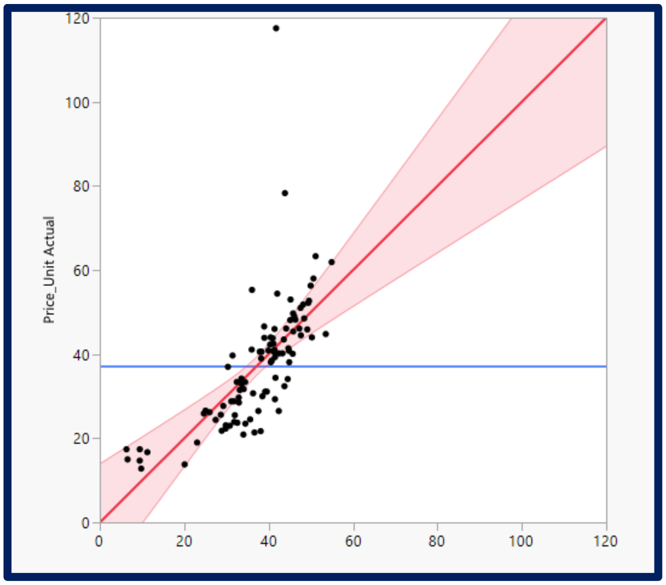


Figure 8.1 – Actual by Predicted Plot

Since the points are close to the fitted line and the confidence bands are narrow, our model fits well. The points on the left and right of the plot that are the furthest from the mean have the most clout and can successfully pull the fitted line toward the point. Points that are vertically away from the line are possible outliers. Both types of points could be detrimental to the fit.

Residual plot :

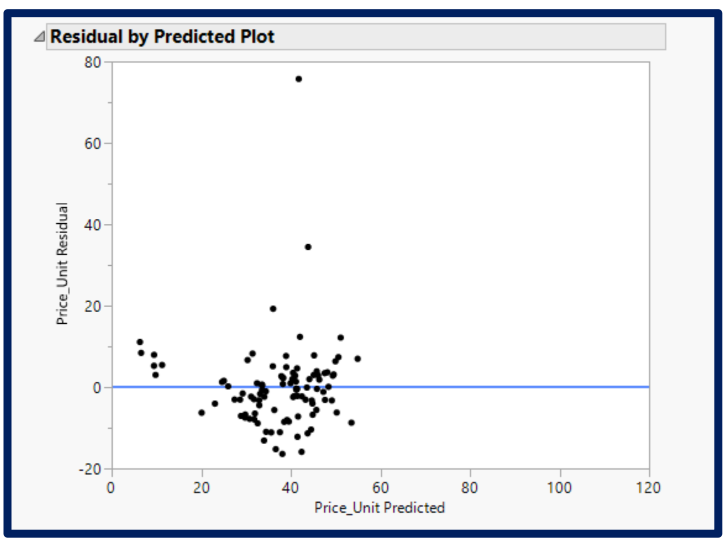


Figure 8.2 – Residual Plot

9. Summary and conclusion

The original pricing values per unit area were quite varied. It would be impossible to develop a model that properly forecasted extreme values; thus, it was critical to eliminate outliers, especially the extremely high ones, in this scenario.

The data utilized in the study most likely came from a single city. The technique used may not be applicable to projecting prices per square meter in another city. Then we should use data from the area around the location we want to research.

10. References

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