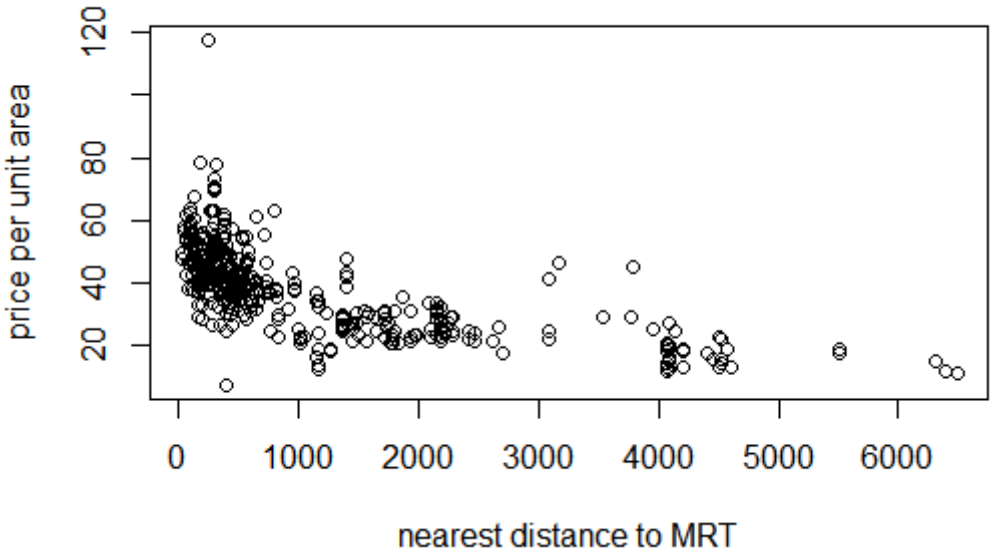
1. Start with a simple regression model giving justification for the selection of the dependent & independent variables.

The data taken for the regression is the market historical data set of real estate valuation collected from Sindian Dist., New Taipei City, Taiwan.
New Taipei City is a city located in Northern Taiwan, completely surrounding the city of Taipei. It is considered a part of the Taipei Metropolitan Area.
The attributes that are collected are

- the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
- the house age (unit: year)
- the distance to the nearest MRT station (unit: meter)
- the number of convenience stores in the living circle on foot (integer)
- the geographic coordinate, latitude. (unit: degree)
- the geographic coordinate, longitude. (unit: degree)

An MRT station is Mass Rapid Transit (MRT), branded as Metro Taipei, is a metro system serving Taipei and New Taipei (known as Taipei County until 2010), Taiwan, operated by government owned Taipei Rapid Transit Corporation.
Taipei MRT is convenient, comfortable, on schedule, significantly reduces commuting time, and shortens distances between spots in the city, it has expanded the living domains of urban residents, transformed the lifestyles of Taipei residents, and significantly improved the living quality of people
The system has been effective in relieving traffic congestion in Taipei, with over two million trips made daily.
Living near an MRT is beneficial due to the easy and efficient commute to any part of the city, there might be more demand for real estate near the MRT stations.



Plotting a scatter plot between nearest MRT station (unit: meter) and the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)
It can be seen that the price is relatively higher when the distance to a MRT station is lower

Taking this into consideration the distance to the nearest MRT station (unit: meter) is taken to be the independent variable and the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) is taken to be the dependent variable for the simple regression model.

The general equation for standard linear regression equation is given by

$$Y = \beta_0 + \beta_1 X + \epsilon$$
$$\beta_0 \; is \; intercept$$
$$\beta_1 \; is \; slope$$

The equation for estimating β_0 and β_1 using Ordinary least squares is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$
$$\hat{\beta}_0 \; is \; estimate \; of \; \beta_0$$
$$\hat{\beta}_1 \; is \; estimate \; of \; \beta_1$$

```
Call:
lm(formula = price_of_unit_area ~ distance_mrt, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-35.396  -6.007  -1.195   4.831  73.483

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.8514271  0.6526105   70.26   <2e-16 ***
distance_mrt -0.0072621  0.0003925  -18.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.07 on 412 degrees of freedom
Multiple R-squared:  0.4538,    Adjusted R-squared:  0.4524
F-statistic: 342.2 on 1 and 412 DF,  p-value: < 2.2e-16
```

Summary of the Regression

The statistics of the standard Linear Regression between nearest MRT station (unit: meter) and the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) are:

$\hat{\beta}_0$ = 45.8514271 with standard error of 0.6526105

$\hat{\beta}_1$ = -0.0072621 with standard error of 0.0004

With 95% confidence the intercept estimate is between 44.6 and 47.1
With 95% confidence the slope estimate is between -0.00805 and -0.00648

The estimated regression line using Ordinary least squares is given by

$$\hat{Y} = 45.8514271 - 0.0072621X$$

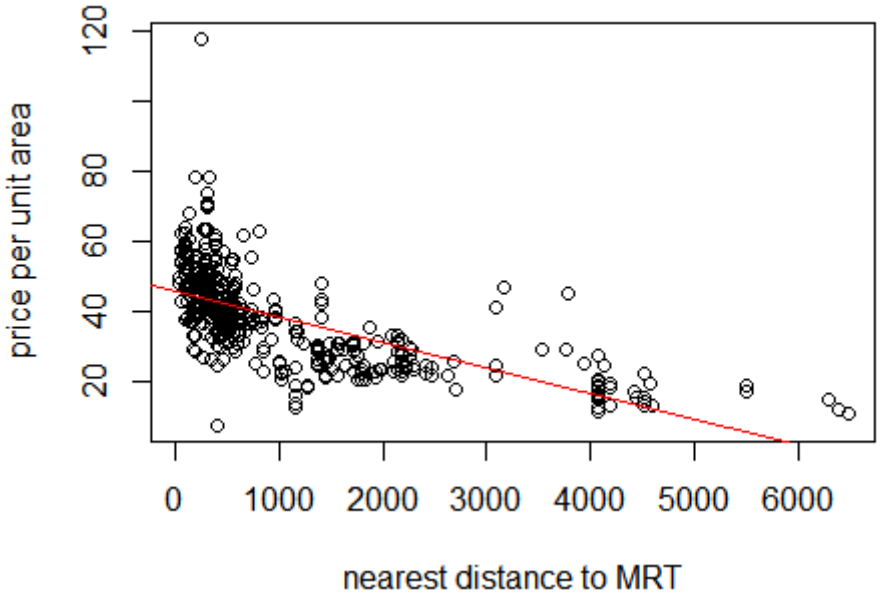
The p-value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable, as the P-value for both the intercept estimate and

slope estimate is 2e-16 which rejects the null hypothesis for significance level of 0.05, which means there is non zero correlation, they are significant.

R Square measures the proportion of variance explained by the independent variable, As this is a linear regression there is not much difference between Multiple R-Squared and Adjusted R-Squared is equal to 0.45, so nearest MRT station (unit: meter) explains 45% of the variance in the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared).

The maximum residual is 73.483 and the minimum residual is -35.396, in general the magnitude of both should be similar otherwise there might be an outlier data point which is causing a difference in the magnitudes.

The elasticity at the points of the mean (1083.89,37.98) is 0.2, so 1% increase in nearest distance to MRT is estimated to reduce house price per unit area by 0.2%



Scatter plot between nearest MRT station (unit: meter) and the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) along with the linear regression line.

2a. Calculate some summary statistics of the variables & comment on the same.

The data has 8 Attributes and 414 rows of data
The Attributes are:

- the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
- the house age (unit: year)
- the distance to the nearest MRT station (unit: meter)
- the number of convenience stores in the living circle on foot (integer)
- the geographic coordinate, latitude. (unit: degree)
- the geographic coordinate, longitude. (unit: degree)

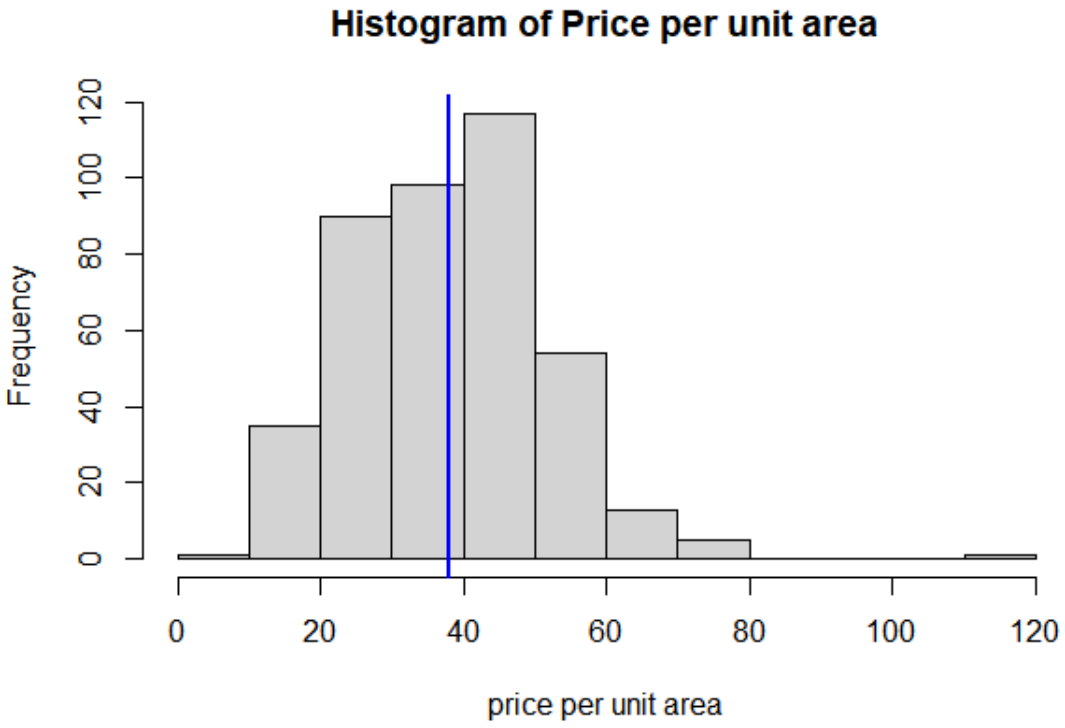
```
> dim(df)
[1] 414   8
> str(df)
tibble[,8] [414 x 8] (S3: tbl_df/tbl/data.frame)
 $ No                : num [1:414] 1 2 3 4 5 6 7 8 9 10 ...
 $ transaction_date  : num [1:414] 2013 2013 2014 2014 2013 ...
 $ house_age         : num [1:414] 32 19.5 13.3 13.3 5 7.1 34.5 20.3 31.7 17.9 ...
 $ distance_mrt      : num [1:414] 84.9 306.6 562 562 390.6 ...
 $ no_of_stores      : num [1:414] 10 9 5 5 5 3 7 6 1 3 ...
 $ latitude          : num [1:414] 25 25 25 25 25 ...
 $ longitude         : num [1:414] 122 122 122 122 122 ...
 $ price_of_unit_area: num [1:414] 37.9 42.2 47.3 54.8 43.1 32.1 40.3 46.7 18.8 22.1
```

The dependent variable for the linear regression is the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)
The Statistics are:

- Mean is 379,800 New Taiwan Dollar/Ping
- Minimum and maximum are 76,000 New Taiwan Dollar/Ping and 1,175,000 New Taiwan Dollar/Ping respectively
- There are no NULL or NA values in the dataset
- The variance and standard deviation are 185.14 and 13.61 respectively

```
> summary(df$price_of_unit_area)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.60   27.70   38.45   37.98   46.60  117.50
> res <- stat.desc(df$price_of_unit_area)
> round(res, 2)
     nbr.val     nbr.null       nbr.na          min
      414.00         0.00         0.00         7.60
         max        range          sum       median
      117.50       109.90     15723.80        38.45
        mean      SE.mean CI.mean.0.95          var
       37.98         0.67         1.31       185.14
     std.dev     coef.var
       13.61         0.36
> describe(df$price_of_unit_area,type=1)
   vars   n  mean    sd median trimmed   mad min   max
X1    1 414 37.98 13.61  38.45   37.63 13.86 7.6 117.5
   range skew kurtosis   se
X1 109.9  0.6     2.14 0.67
```

The skew is 0.6 which is positive, means that the distribution has a longer right tail (right skewed) as mean>median>mode due to most of the distribution towards the left
The kurtosis is 2.11 which is positive, tails are fatter than the normal distribution also called Leptokurtic



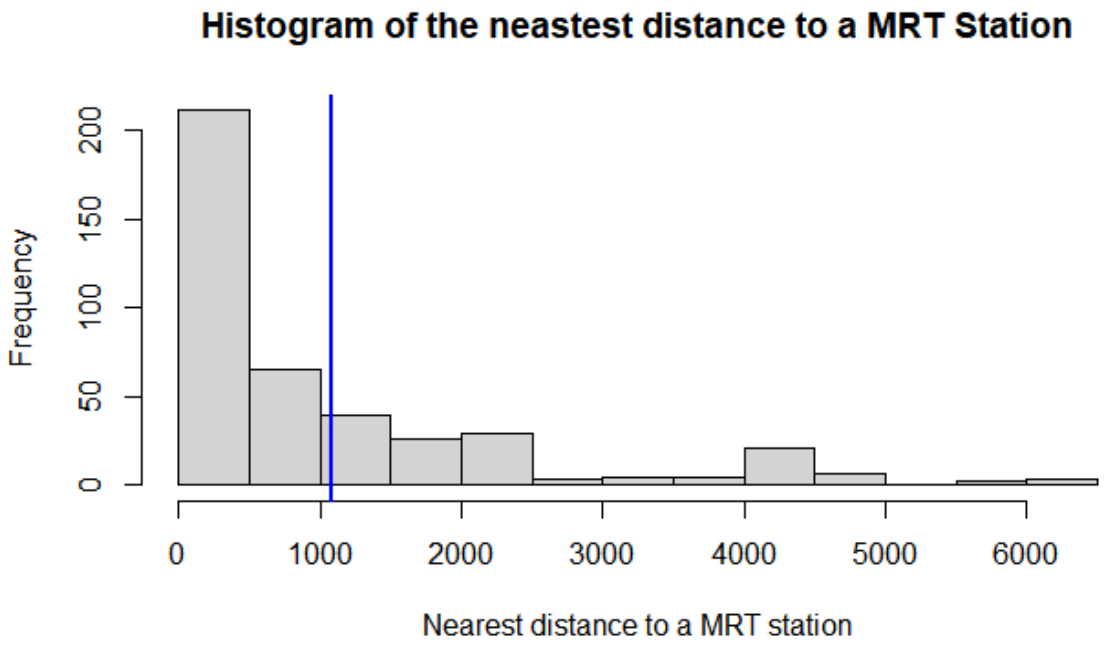
If $(m_r = [\sum(X - mx)^r]/n)$ then the kurtosis formula is given by $(m_4/(m_2)^2 - 3)$ and skewness is given by $(m_3/(m_2)^{3/2})$

The Independent variable for the linear regression is the nearest MRT station (unit: meter)
The Statistics are:

- Mean is 1083.89m
- Minimum and maximum are 23.38m and 6488.02m respectively
- There are no NULL or NA values in the dataset
- The variance and standard deviation are 1592920.63m and 1262.11 respectively

```
> summary(df$distance_mrt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.38  289.32  492.23 1083.89 1454.28 6488.02
> res <- stat.desc(df$distance_mrt)
> round(res, 2)
     nbr.val     nbr.null       nbr.na          min
      414.00         0.00         0.00        23.38
         max        range          sum       median
     6488.02      6464.64    448728.68       492.23
        mean      SE.mean CI.mean.0.95          var
     1083.89        62.03       121.93   1592920.63
     std.dev     coef.var
     1262.11         1.16
> describe(df$distance_mrt,type=1)
   vars   n    mean      sd median trimmed    mad   min
X1    1 414 1083.89 1262.11 492.23  809.27 455.68 23.38
       max   range skew kurtosis    se
X1 6488.02 6464.64 1.88     3.15 62.03
```

The skew is 1.88 which is positive, means that the distribution has a longer right tail (skewed towards right) as mean>median>mode due to most of the data is bunched towards the left
The kurtosis is 3.15 which is positive, tails are fatter than the normal distribution also called Leptokurtic.

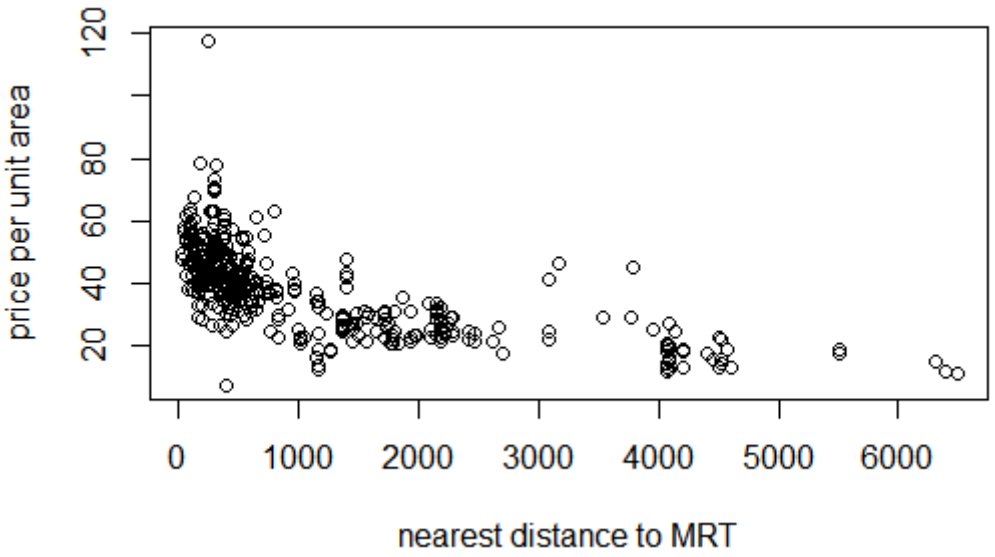


If $(m_r = [\sum(X - mx)^r]/n)$ then the kurtosis formula is given by $(m_4/(m_2)^2 - 3)$ and skewness is given by $(m_3/(m_2)^{3/2})$

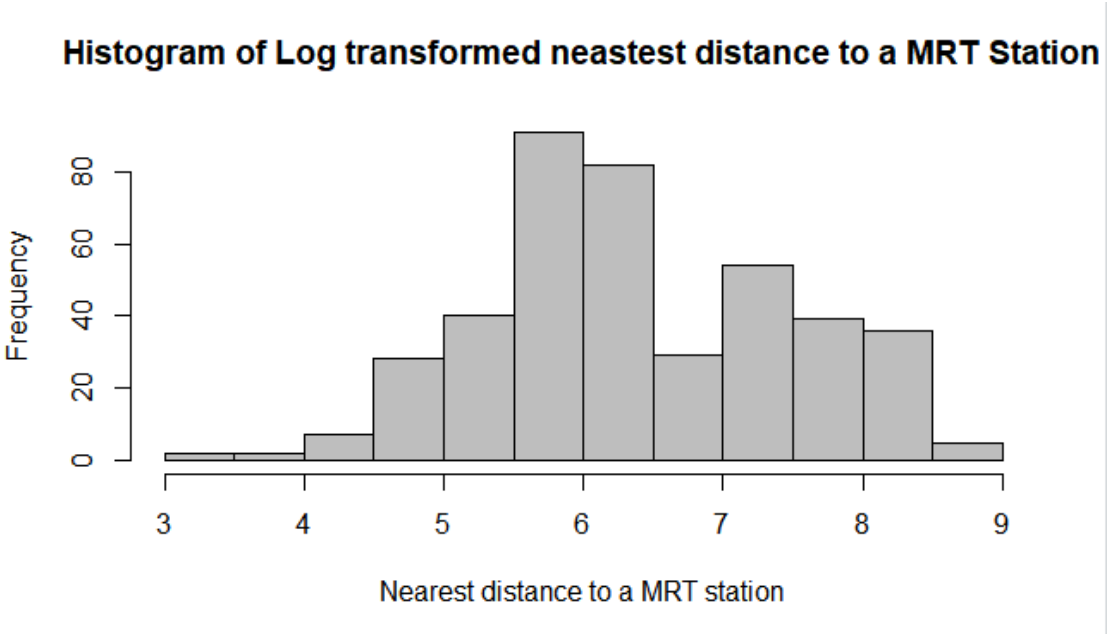
2b. Make a scatter diagram & comment on the relationship between the variables. Based on the plot, discuss whether it suggests a linear or non-linear relationship and then decide whether you should go with a linear or a non-linear model.

The scatter plot shows that the data is more skewed towards lower values of the independent variable(nearest MRT station (unit: meter)) and it has long tail towards the right, the data is skewed towards the left as shown in the figure above, the dependent variable (house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) almost takes a bell curve but has a an outlier data point which causes it to have a long tail towards the right as shown in the figure above
The economic theory assumption is that the closer the distance to a MRT station the more expensive the house price of the unit area which is also explained in 1 we have to work based on this assumption.
The expected curve between the independent and the dependent variable is a decreasing curve with a decreasing rate as the assumption is that after going a certain distance away from the MRT Station the marginal change due to the change in distance will have a decreasing effect on the change of house price per unit area. There will be less impact on the house price due to the distance from nearest MRT station after a certain distance

Due to the skewness in both the independent and the dependent variables maybe a linear model would not be able to fit the data so a logarithmic transformed data is better.

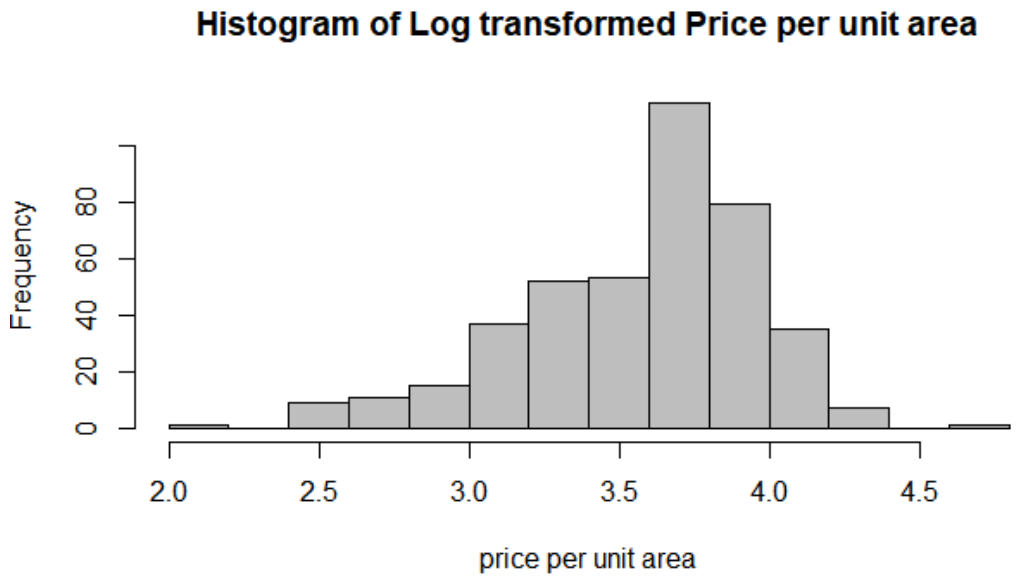


The histogram of logarithmic transferred nearest distance to MRT(in metre)



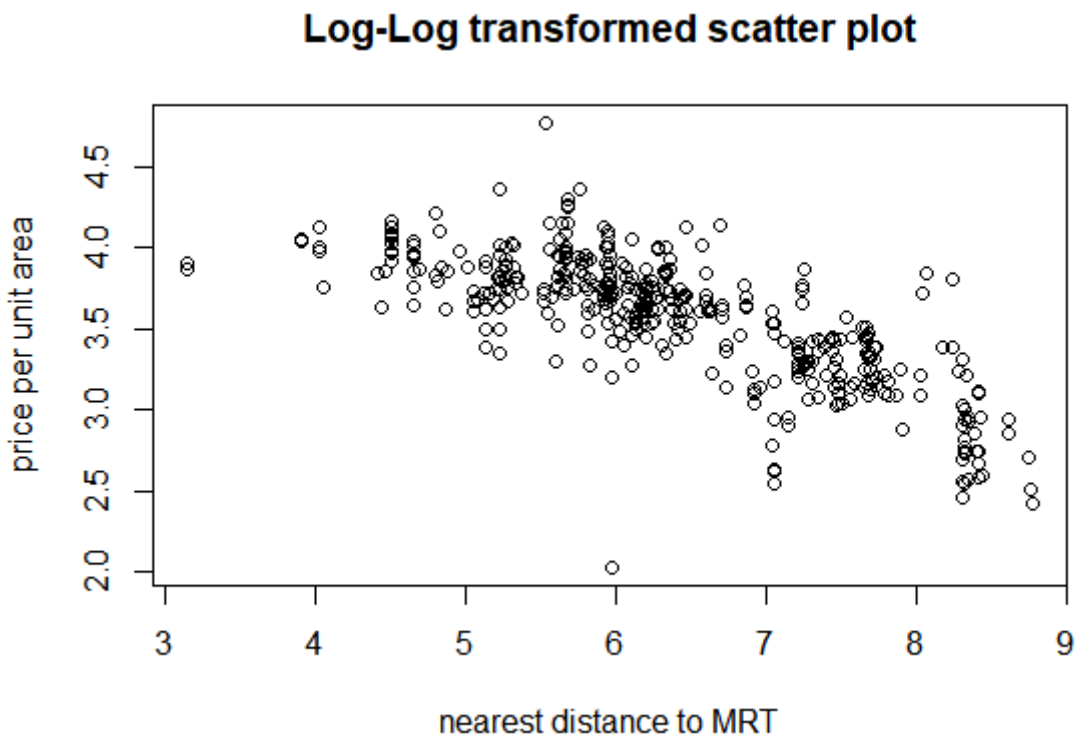
The logarithmic transformed nearest distance to a MRT Station is approximately normal, better than the data which is not transformed.

The histogram of logarithmic transferred house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)



The logarithmic transformed house price of unit area is approximately normal, better than the data which is not transformed.

So a log-log model is better to fit the data, the scatter plot of a log-log transformed data, it can be seen a linear line will be able to fit the data.



The general equation for standard log-log regression equation is given by

$$lnY = \beta_0 + \beta_1 lnX + \epsilon$$
$$\beta_0 \; is \; Coefficient \; of \; the \; log \; transformed \; Independent \; variable$$
$$\beta_1 \; is \; slope$$

When β_1 is less than 0 then the model would be a decreasing function at a decreasing rate So, the estimation is done using a log -log model after considering all the factors of the above.

2c. Estimate the model & interpret the result.

The general equation for standard log-linear regression equation is given by

$$lnY = \beta_0 + \beta_1 lnX + \epsilon$$

$$\beta_0 \; is \; Coefficient \; of \; the \; log \; transformed \; Independent \; variable$$

$$\beta_1 \; is \; slope$$

The equation for estimating β_0 and β_1 using Ordinary least squares is given by

$$\ln \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 lnX$$

$$\hat{\beta}_0 \; is \; estimate \; of \; \beta_0$$

$$\hat{\beta}_1 \; is \; estimate \; of \; \beta_1$$

```
Call:
lm(formula = log(df$price_of_unit_area) ~ log(df$distance_mrt),
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.64982 -0.12285  0.01203  0.14539  0.97039

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.27132    0.07271    72.5   <2e-16 ***
log(df$distance_mrt) -0.26669    0.01121   -23.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.255 on 412 degrees of freedom
Multiple R-squared:  0.5789,    Adjusted R-squared:  0.5779
F-statistic: 566.4 on 1 and 412 DF,  p-value: < 2.2e-16
```

Summary of the Regression

The statistics of the standard log-log regression between nearest MRT station (unit: meter) and the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) are:

$\hat{\beta}_0$ = 5.27132 with standard error of 0.07271

$\hat{\beta}_1$ = -0.26669 with standard error of 0.01121

With 95% confidence the intercept estimate is between 5.13 and 5.41
With 95% confidence the independent variable coefficient estimate is between -0.289 and -0.245

The estimated regression line using Ordinary least squares is given by

$$\ln \hat{Y} = \; 5.27132 - 0.26669 \; lnX$$

The p-value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable, as the P-value for both the intercept estimate and

independent variable coefficient estimate is 2e-16 which rejects the null hypothesis for significance level of 0.05, which means there is non zero correlation, they are significant.

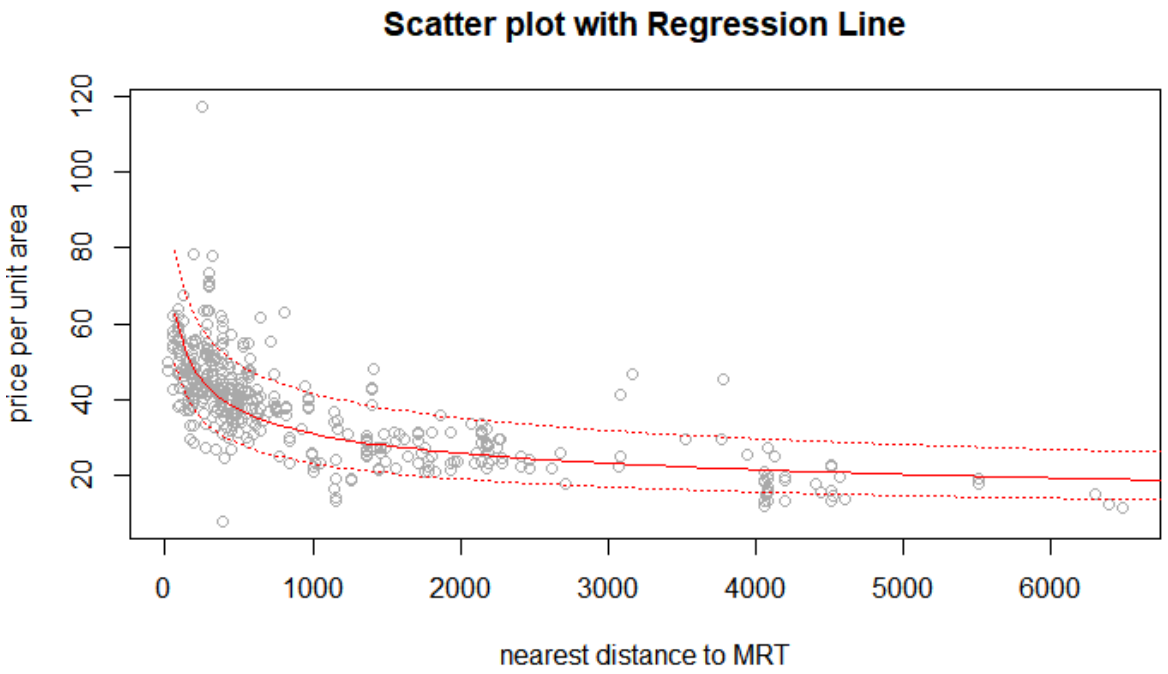
R Square measures the proportion of variance explained by the independent variable, As this is a log-linear regression,R-squared is 0.57, then 57% of the variability in the log-transformed values of the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) is accounted for by the predictor variables that is log transformed nearest MRT station (unit: meter) in the model.
An R-square comparison is meaningful only if the dependent variable is the same for both models. So the R-square from the linear model cannot be compared with the R-square from the log-linear.
The maximum residual is 0.97039 and the minimum residual is -1.64982 the magnitude of both are approximately the same, median is 0.01203 which is close to 0, the non linear regression did better than the linear regression.

For a log-log model the elasticity is constant which is equal to $\hat{\beta}_1$ = -0.26669 so a 1% increase in the distance to the nearest MRT station(meter) there will be a 0.267% decrease in the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

2d. Fit the regression line on the scatter plot & comment on the same.

Scatter plot between nearest MRT station (unit: meter) and the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) along with the log-log regression line.
The log-log curve fits the data better than the linear regression line.

The scatter plot along with the 95% confidence interval regression curves, the dotted lines are the 95% confidence intervals



2e. Make predictions from your model based on the median value of the explanatory variable.

The median value of the explanatory variable that is the nearest MRT station (unit: meter) is 492.2313m. In general there may or may not be a data point near the median value but in this case there are multiple values of the dependent variable(response variable) which are given below.

	No	transaction_date	house_age	distance_mrt	no_of_stores	latitude	longitude	price_of_unit_area
1	13	2012.917	13.0	492.2313	5	24.96515	121.5374	39.3
2	54	2013.083	13.3	492.2313	5	24.96515	121.5374	38.9
3	121	2013.167	13.3	492.2313	5	24.96515	121.5374	31.3
4	122	2013.500	13.6	492.2313	5	24.96515	121.5374	48.0
5	140	2012.667	12.9	492.2313	5	24.96515	121.5374	42.5
6	144	2013.500	13.6	492.2313	5	24.96515	121.5374	40.1
7	219	2013.417	13.6	492.2313	5	24.96515	121.5374	43.8
8	273	2012.750	13.0	492.2313	5	24.96515	121.5374	40.5
9	351	2013.000	13.2	492.2313	5	24.96515	121.5374	42.3

The mean of the dependent variable values (the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)) when the nearest MRT station value is 492.2313m is 407,444.4 Dollar/Ping from the data

```
> mxn <- mean(newdf$price_of_unit_area)
> mxn
[1] 40.74444
```

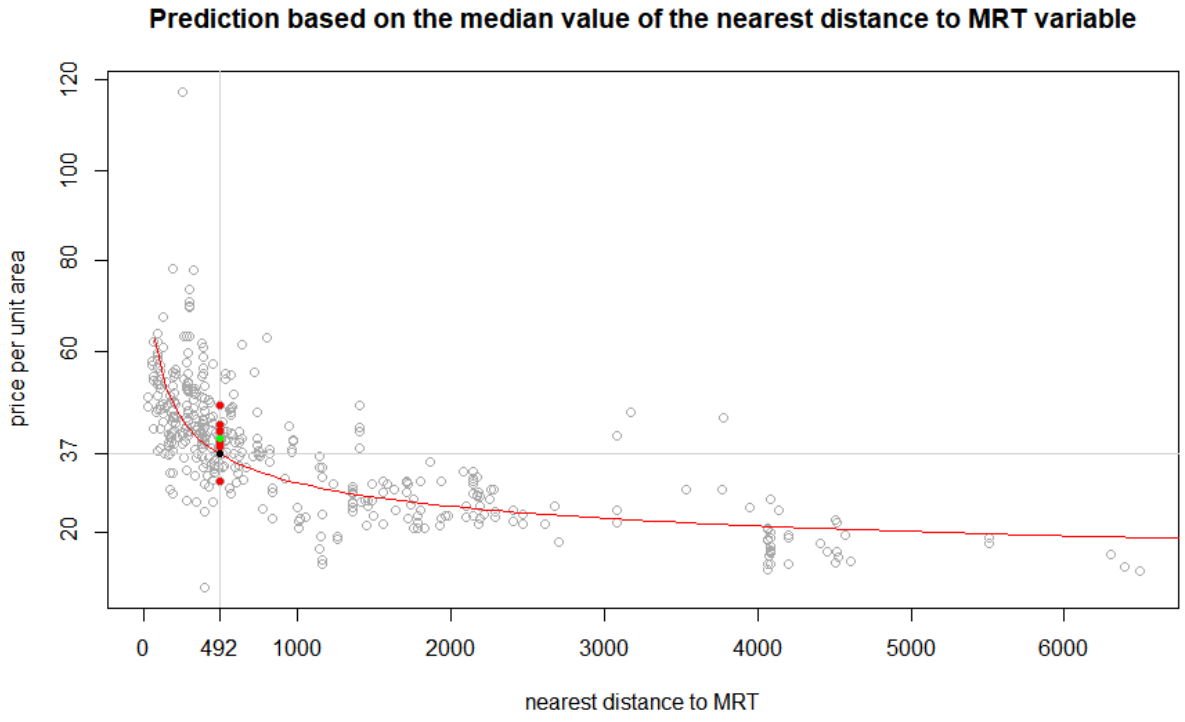
The log-log model can also be transformed into

$$\hat{Y} = e^{5.27132} * X^{-0.26669}$$

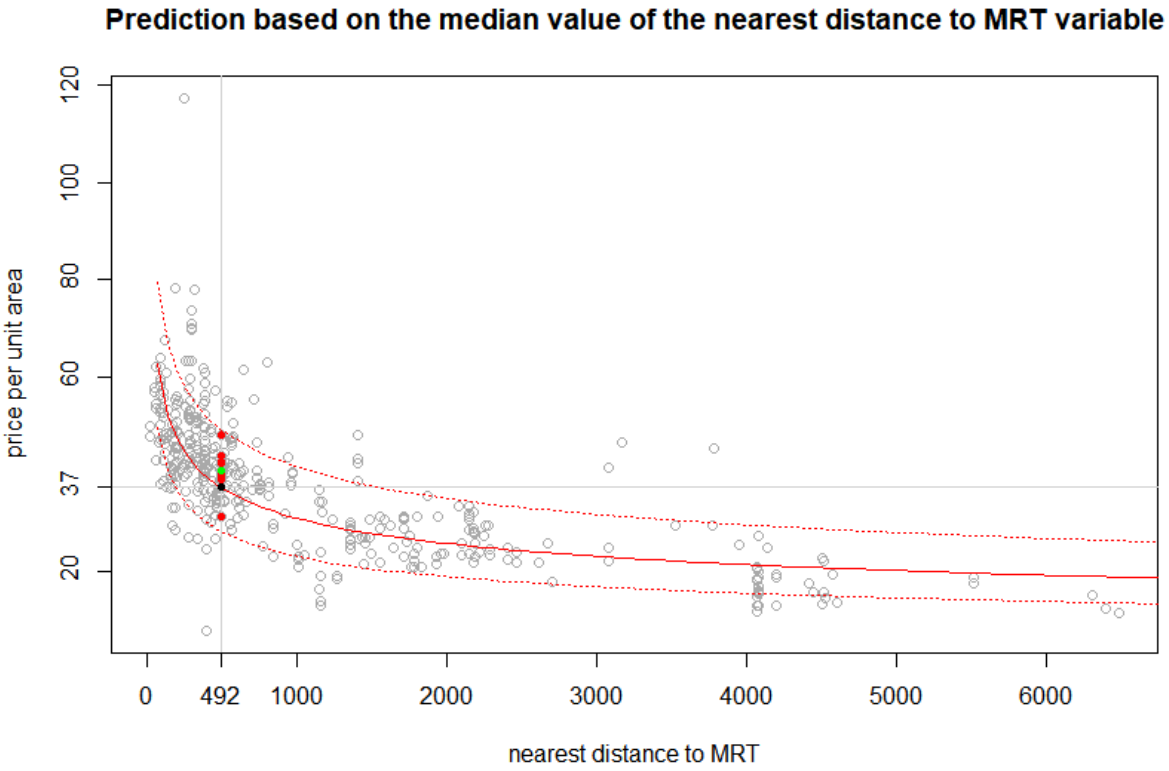
The estimated value of the dependent variable (the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)) when the nearest MRT station value is 492.2313m is **372,668.2** Dollar/Ping

```
> yest <- exp(b1)*med^(b2)
> yest
[1] 37.26682
```

In the graph given below the estimate is the black dot which is 37.26682 (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)), the values of the dependent variable at the median of the independent variable are the red dots, the mean of the values of the dependent variable at the median of the independent variable is represented by the green dot.



The figure below shows that the mean (green dot) lies in the between the 95% confidence interval curves (the red dotted curve)



So the difference between the mean and the prediction from the regression is 407,444.4 - 372,668.2 = 34,776.2.

2f. Run a test of significance on your selected model & discuss the result.

The number of samples in the given the data are 414 and there are 2 variables considered for the model, so the model has 412 degrees of freedom
The Hypothesis is:

$$H_0 : \hat{\beta}_1 = 0$$
$$H_1 : \hat{\beta}_1 \neq 0$$

Where H_0 is the null hypothesis which states that the coefficient of the logarithmic transformed independent variable is 0, which means logarithmic transformed independent variable is no use in explaining the logarithmic transformed dependent variable
The alternate hypothesis H_1 states that logarithmic transformed independent variable can explain the logarithmic transformed dependent variable.

This would be a two tail test let α be 0.05 then the critical values for the two tail test are 2.5 percentile $t_{(0.025,412)}$ = 2.249625 and the 97.5 percentile is $t_{(.975,412)}$ = -2.249625
Reject the null hypothesis if the calculated value of t ≥ 2.249625 or if t ≤ -2.249625
Not reject the null hypothesis if -2.249625 < t < 2.249625

The value of test statistic is given by

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

The value of the test statistic is -23.790365.

Since the test statistic is -23.790365 ≤ -2.249625 we reject the null hypothesis that $\hat{\beta}_1$ = 0 and conclude that there is a statistically significant relationship between logarithmic transformed independent variable and logarithmic transformed dependent variable.

2g. Calculate SSR, SSE & SST from your estimated model and comment on the model based upon your calculations.

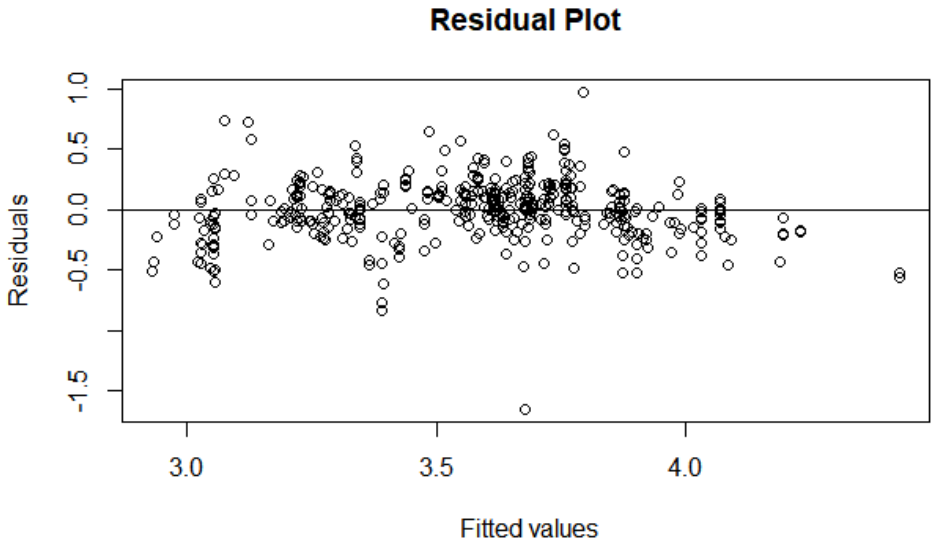
The value of SSR, SSE & SST estimated model are:

- SSE = 26.78778
- SSR = 36.8284
- SST = 63.61617

```
|                     |  Df|   Sum.Sq|    Mean.Sq|  F.value| Pr..F.|
|:--------------------|---:|--------:|----------:|--------:|------:|
|log(df$distance_mrt) |   1| 36.82840| 36.8283960| 566.4262|      0|
|Residuals            | 412| 26.78778|  0.0650189|       NA|     NA|
> SSE <- anov[2,2]
> SSE
[1] 26.78778
> SSR <- anov[1,2]
> SSR
[1] 36.8284
> SST <- SSR + SSE
> SST
[1] 63.61617
```

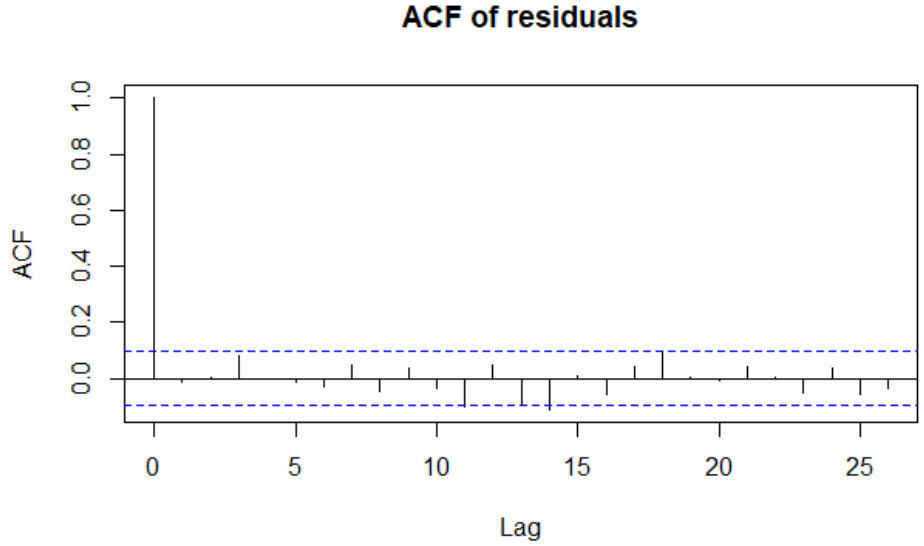
2f. On your estimated model, check whether the assumptions of homoscedasticity, no autocorrelation & normality of error term is satisfied. Discuss your findings. If the assumptions are not satisfied, suggest some modifications to the model so that the assumptions may get satisfied.

The assumption of homoscedasticity can be checked with the help of a residual plot, if the scatter plot is random and does not shown any signs of patterns then the homoscedasticity condition is satisfied



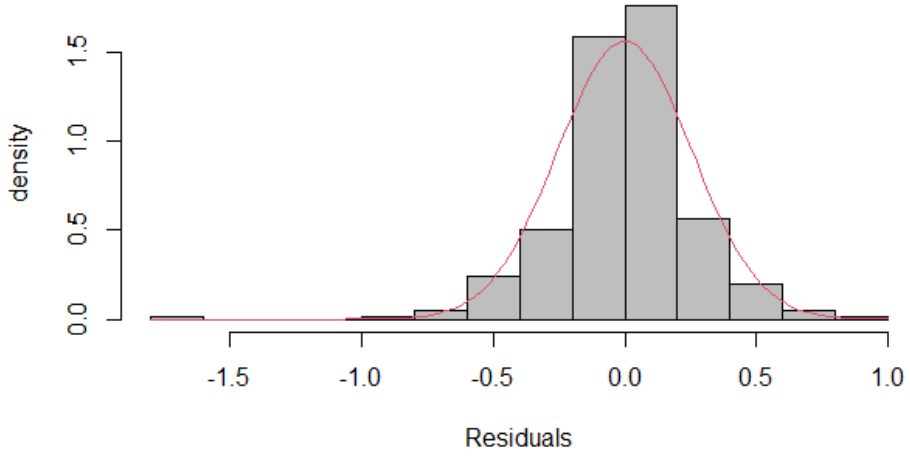
In the residual plot shown above the points are relatively concentrated in the middle and are not completely random to satisfy the homoscedasticity condition, so the model is assumed to have heteroskedasticity

To check if there is autocorrelation assumption the points should follow a particular pattern, in the residual plot the points don't follow a unique pattern. A ACF plot can be used to check if there is an autocorrelation among residuals.



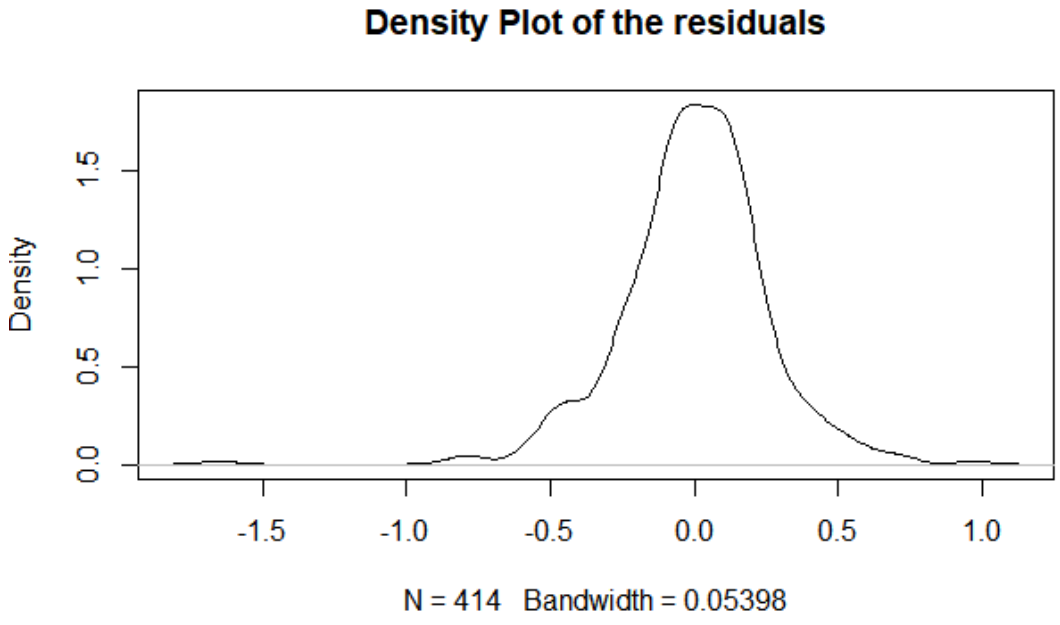
There is are spikes above the 0.1 significant value (the blue dotted line) so the model is assumed to have autocorrelation

The histogram of the residuals is shown below, the plot does not look normal there is a long left tail, this can be tested using Jarque-Bera test



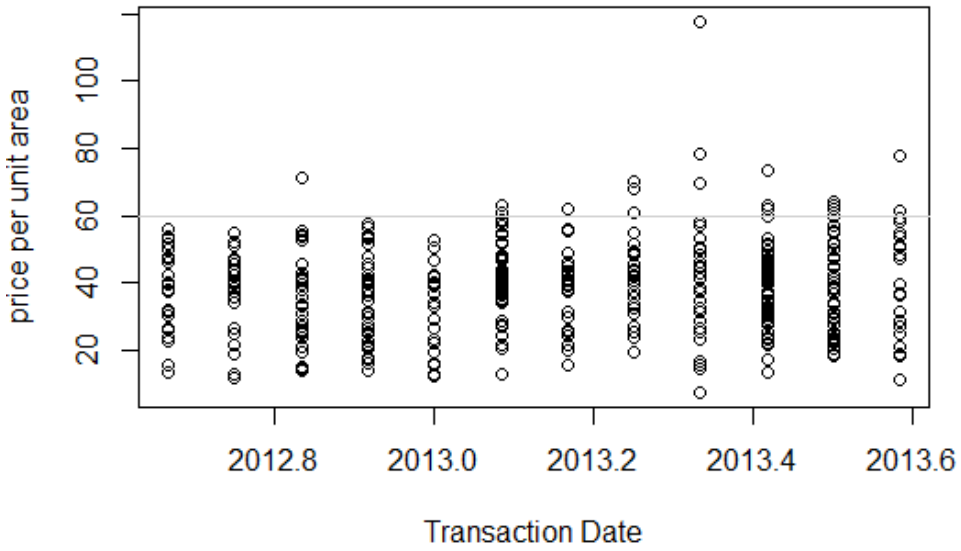

```
        Jarque Bera Test

data:  res
X-squared = 411.09, df = 2, p-value < 2.2e-16
```

This indicates that the test statistic is 441.09, with a p-value less than 2.2e-16. We would be able to reject the null hypothesis that the data is normally distributed in this scenario. As p-value is less than 0.05 (the significant value)



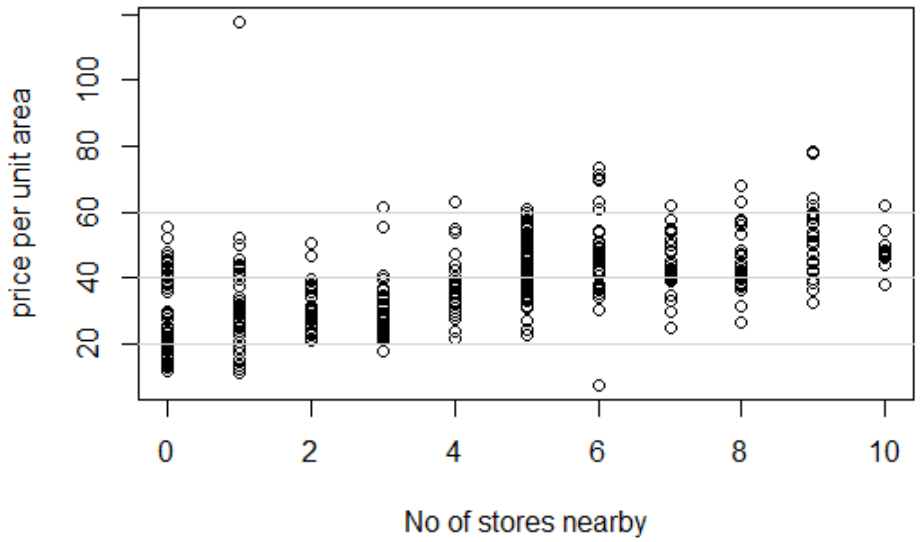
Even the density plot shows the residuals have a long left tail.

To improve the model more variables can be taken into to the model
The scatter plot between transaction date and house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) is shown below



It can be seen that the price per unit area was above 600,000 New Taiwan Dollar/Ping from nearly the first month of 2013 and was lower than 600,000 New Taiwan Dollar/Ping before that other than one exception

The scatter plot between the number of convenience stores in the living circle on foot (integer) and house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) is shown below



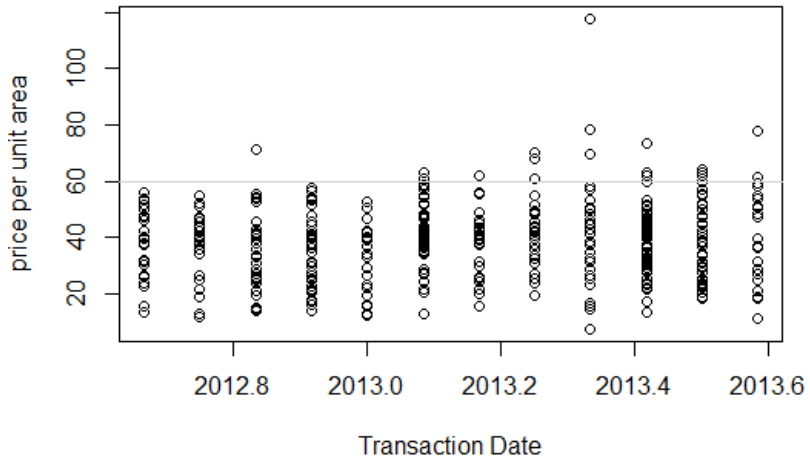
The price of house per unit is above 600,000 New Taiwan Dollar/Ping only if the number of convenience stores in the living circle on foot is above 6 and the it is less than 600,000 New Taiwan Dollar/Ping if they are below 2

The house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared also depends on the latitude and longitude of the location of the house
These variables can be taken into consideration for a better model

1. Modify your simple linear regression model (of Assignment-I) by including at least 2- 3 additional independent variables. Also, give justification for the selection of the additional variables.

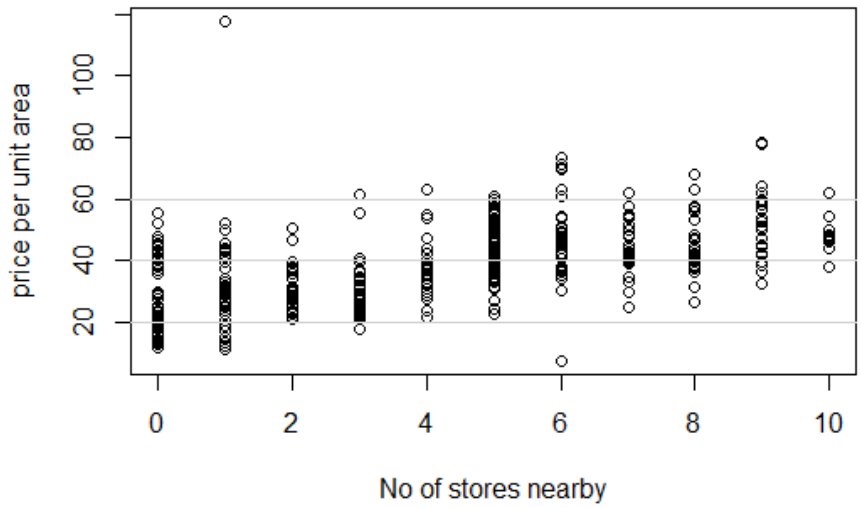
The additional parameters are the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.), the number of convenience stores in the living circle on foot (integer), the geographic coordinate latitude.(unit: degree) and the geographic coordinate longitude. (unit: degree)

The scatter plot between transaction date and house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) is shown below



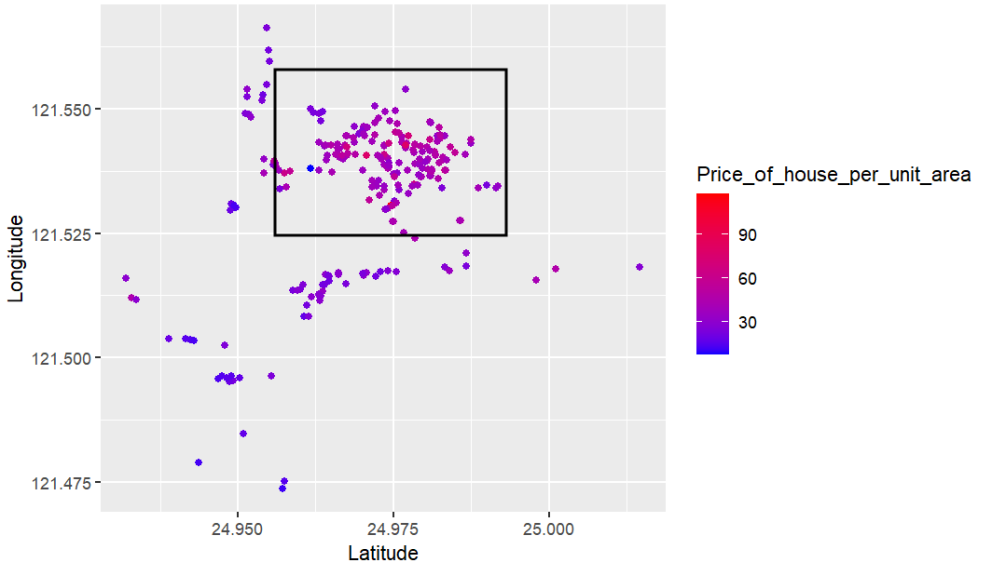
It can be seen that the price per unit area was above 600,000 New Taiwan Dollar/Ping from nearly the first month of 2013 and was lower than 600,000 New Taiwan Dollar/Ping before that other than one exception. This can be a useful variable to take into consideration

The scatter plot between the number of convenience stores in the living circle on foot (integer) and house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) is shown below



The price of house per unit is above 600,000 New Taiwan Dollar/Ping only if the number of convenience stores in the living circle on foot is above 6 and it is less than 600,000 New Taiwan Dollar/Ping if they are below 2. This can a variable that can be taken into consideration

Below is the scatter plot between the geographic coordinate, latitude. (unit: degree)
And the geographic coordinate, longitude. (unit: degree) with The house price of the unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter) being a heat map.
From the plot given below we can establish that there is a region where the prices are higher than other places, here red dots are expensive areas and blue dots are more cheaper areas



The rectangle area is more expensive, so the latitude and longitude variables are to be taken into consideration.
These are the other variables that affect The house price of the unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter) other than the distance to the nearest MRT station (unit: meter).
The general equation for multivariable regression for these variables is given by

$$lnY = \beta_0 + \beta_1 lnX_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

$$\beta_0 \; is \; intercept$$

$$\beta_1 \; is \; coefficient \; of \; the \; logarithmic \; transformed \; distance \; variable$$

$$\beta_2 \; is \; coefficient \; transaction \; date$$

$$\beta_3 \; is \; coefficient \; of \; the \; number \; of \; stores \; nearby$$

$$\beta_4 \; is \; coefficient \; the \; latitude$$

$$\beta_5 \; is \; coefficient \; the \; longitude$$

The equation for estimating β_0 and β_1 using Ordinary least squares is given by

$$\ln \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 lnX_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \epsilon$$

$$\hat{\beta}_i \; is \; estimate \; of \; \beta_i$$

2a. Calculate summary statistics of the variables & comment on the same.

Transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)

The Statistics are:

- Mean is 2013.5
- Minimum and maximum are 2012.67 and 2013.58 respectively
- There are no NULL or NA values in the dataset
- The variance and standard deviation are 0.08 and 0.28 respectively

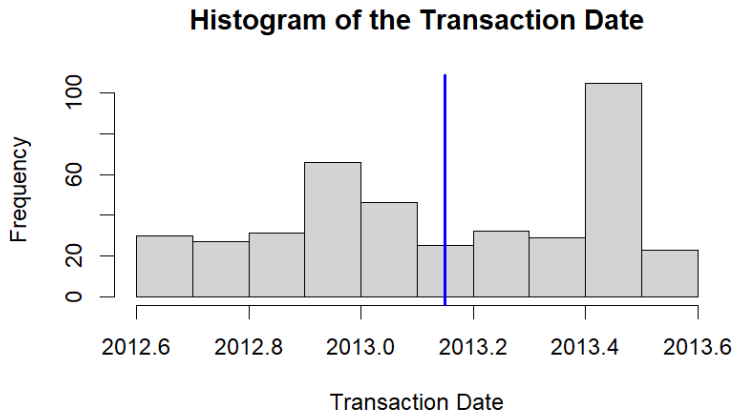
nbr.val	nbr.null	nbr.na	min
414.00	0.00	0.00	2012.67
max	range	sum	median
2013.58	0.92	833443.67	2013.17
mean	SE.mean	CI.mean.0.95	var
2013.15	0.01	0.03	0.08
std.dev	coef.var		
0.28	0.00		

The skew is -0.15 which is negative, meaning that the distribution has a longer left tail (left skewed). The kurtosis is -1.232 which is negative, tails are thinner than the normal distribution also called Platykurtic.

Table: Summary Statistics of Transaction Date

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	414	2013.149	0.282	2013.167	2013.157	0.371	2012.667	2013.583	0.917	-0.15	-1.232	0.014

The histogram of the Transaction date is:



The number of convenience stores in the living circle on foot (integer)

The Statistics are:

- Mean is 4.09
- Minimum and maximum are 0 and 10 respectively
- There are 67 NULL values and no NA values in the dataset
- The variance and standard deviation are 8.68 and 2.95 respectively

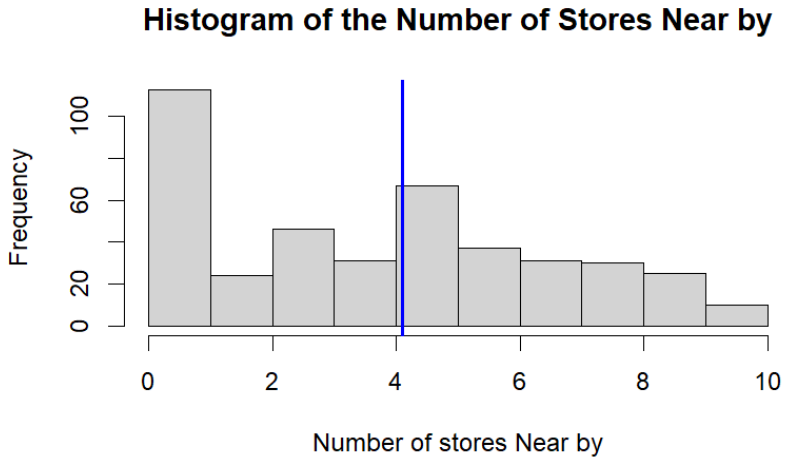
nbr.val	nbr.null	nbr.na	min	max	range
414.00	67.00	0.00	0.00	10.00	10.00
sum	median	mean	SE.mean	CI.mean.0.95	var
1695.00	4.00	4.09	0.14	0.28	8.68
std.dev	coef.var				
2.95	0.72				

The skew is 0.154 which is positive, meaning that the distribution has a longer right tail (right skewed) due to more 0 and 1 this arises. The kurtosis is -1.067 which is negative, tails are thinner than the normal distribution also called Platykurtic.

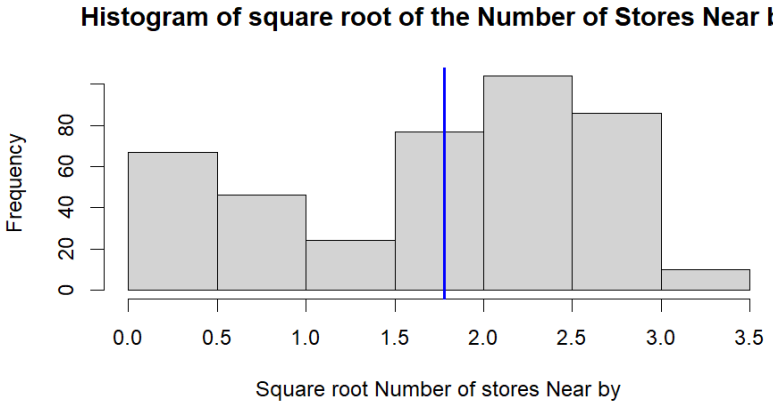
Table: Summary Statistics of Number of stores Near by

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	414	4.094	2.946	4	3.982	4.448	0	10	10	0.154	-1.067	0.145

The histogram of the Number of Stores Nearby is:



As the Number of stores nearby is heavily skewed to the right and a log transform would not be viable due to the large number of NULL values present, a square root is better to transform the variable.



The transformed Variable also represents that it is increasing at a decreasing rate which means after a certain increase in the stores nearby the marginal effect of that on the house price of the unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter) will decreasing.

The geographic coordinate, latitude. (unit: degree)

The Statistics are:

- Mean is 24.97
- Minimum and maximum are 24.93 and 25.01 respectively
- There are no NA or NULL values in the dataset
- The variance and standard deviation are 0.00 and 0.01 respectively

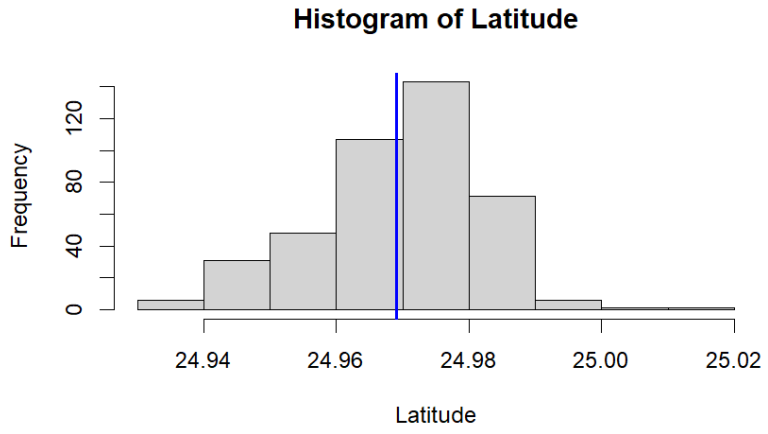
nbr.val	nbr.null	nbr.na	min	max	range	sum	median
414.00	0.00	0.00	24.93	25.01	0.08	10337.18	24.97
mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var		
24.97	0.00	0.00	0.00	0.01	0.00		

The skew is -0.437 which is negative, meaning that the distribution has a longer left tail (left skewed). The kurtosis is 0.251 which is positive, tails are fatter than the normal distribution also called Leptokurtic.

Table: Summary Statistics of Latitude

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	414	24.969	0.012	24.971	24.97	0.012	24.932	25.015	0.083	-0.437	0.251	0.001

The histogram of the Latitude is:



The geographic coordinate, longitude. (unit: degree)
The Statistics are:

- Mean is 121.53
- Minimum and maximum are 121.47 and 121.57 respectively
- There are no NA or NULL values in the dataset
- The variance and standard deviation are 0.00 and 0.02 respectively

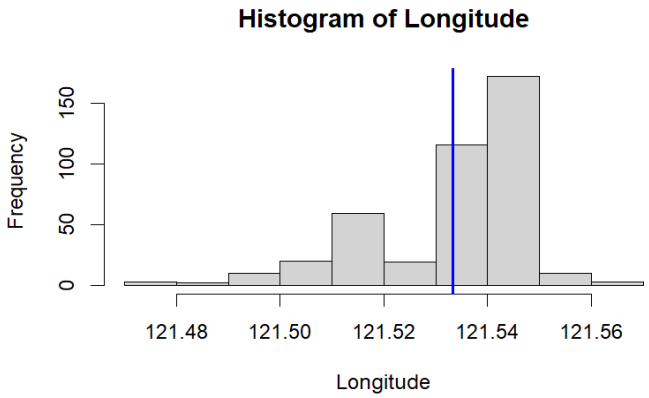
nbr.val	nbr.null	nbr.na	min	max	range	sum	median	mean
414.00	0.00	0.00	121.47	121.57	0.09	50314.81	121.54	121.53
SE.mean	CI.mean.0.95	var	std.dev	coef.var				
0.00	0.00	0.00	0.02	0.00				

The skew is -1.215 which is negative, meaning that the distribution has a longer left tail (left skewed). The kurtosis is 1.173 which is positive, tails are fatter than the normal distribution also called Leptokurtic.

Table: Summary Statistics of Longitude

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	414	121.533	0.015	121.539	121.535	0.009	121.474	121.566	0.093	-1.215	1.173	0.001

The histogram of the Longitude is:



If $(m_r = [\sum (X - mx)^r]/n)$ then the kurtosis formula is given by $(m_4/(m_2)^2 - 3)$ and skewness is given by $(m_3/(m_2)^{3/2})$

Taken the above into consideration the multivariable regression model is given by

$$lnY = \beta_0 + \beta_1 lnX_1 + \beta_2 X_2 + \beta_3 X_3^{1/2} + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$
$$\beta_0 \; is \; intercept$$
$$\beta_1 \; is \; coefficient \; of \; the \; logarithmic \; transformed \; distance \; variable$$
$$\beta_2 \; is \; coefficient \; transaction \; date$$
$$\beta_3 \; is \; coefficient \; of \; the \; transformed \; number \; of \; stores \; nearby$$
$$\beta_4 \; is \; coefficient \; the \; latitude$$
$$\beta_5 \; is \; coefficient \; the \; longitude$$

The equation for estimating β_0 and β_1 using Ordinary least squares is given by

$$\ln \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 lnX_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3^{1/2} + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \epsilon$$
$$\hat{\beta}_i \; is \; estimate \; of \; \beta_i$$

column	n	mean	sd	median	min	max
No	414	207.500000	119.6557562	207.5000	1.00000	414.00000
transaction_date	414	2013.148953	0.2819953	2013.1667	2012.66667	2013.58333
house_age	414	17.712560	11.3924845	16.1000	0.00000	43.80000
distance_mrt	414	1083.885689	1262.1095954	492.2313	23.38284	6488.02100
no_of_stores	414	4.094203	2.9455618	4.0000	0.00000	10.00000
latitude	414	24.969030	0.0124102	24.9711	24.93207	25.01459
longitude	414	121.533361	0.0153472	121.5386	121.47353	121.56627
price_of_unit_area	414	37.980193	13.6064877	38.4500	7.60000	117.50000

The above table has all the summary statistics of the variables that are relevant to the model which is considered.

2b. Estimate the model & interpret the result

The estimated regression equation using Ordinary least squares is given by:

$$\ln \hat{Y} = -899.448 - .182 ln X_1 + .166 X_2 + .030 X_3^{1/2} + \ 9.383 X_4 + \ 2.7557 X_5$$

The results of the estimated multivariable regression is given below:

	coefficient	Std. Error	t-value	p-value
(Intercept)	-899.448	134.041	-6.710	0.000
log(df$distance_mrt)	-0.182	0.015	-12.026	0.000
df$transaction_date	0.166	0.038	4.327	0.000
sqrt(df$no_of_stores)	0.030	0.015	1.975	0.049
df$latitude	9.383	1.026	9.142	0.000
df$longitude	2.755	0.933	2.954	0.003

The statistics of the of the multivariable model regression are:

$\hat{\beta}_0$ = -899.448 with standard error of 134.041

$\hat{\beta}_1$ = -0.182 with standard error of 0.015

$\hat{\beta}_2$ = 0.166 with standard error of 0.038

$\hat{\beta}_3$ = 0.030 with standard error of 0.015

$\hat{\beta}_4$ = 9.383 with standard error of 1.026

$\hat{\beta}_5$ = 2.755 with standard error of 0.933

The p-value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable, as the P-value for both the intercept estimate and independent variable coefficient estimate is less then 0.05 which rejects the null hypothesis for significance level of 0.05, which means there is non zero correlation, they are significant.

The R-squared of the model is 69.58%, whereas the adjusted R-square is 69.21%.
This shows that this is a better model than the previous model with only one variable.

```
Residual standard error: 0.2178 on 408 degrees of freedom
Multiple R-squared:  0.6958,    Adjusted R-squared:  0.6921
F-statistic: 186.7 on 5 and 408 DF,  p-value: < 2.2e-16
```

The F critical value for 5 and 408 degrees of freedom is for alpha 0.05 is 2.236, the F-statistic in this case is 186.7, which means at least on the variables are significant.

The coefficient of Transaction date, Latitude and Longitude are positive which means as they increase the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) increases

```
   SST    SSR    SSE
63.616 44.267 19.349
```

ANOVA of multiple regression gives the decomposition of the total sum of squares, showing by how much each predictor contributes to the reduction in the residual sum of squares is given below. The logarithmic transformed distance to the nearest MRT station (unit: meter) is the largest contribution to the reduction of variability in the residuals and the least is the Longitude variable.

```
Response: log(df$price_of_unit_area)
                     Df Sum Sq Mean Sq  F value    Pr(>F)
log(df$distance_mrt)    1 36.828  36.828 776.5723 < 2.2e-16 ***
df$transaction_date     1  1.443   1.443  30.4198 6.188e-08 ***
sqrt(df$no_of_stores)   1  1.031   1.031  21.7373 4.242e-06 ***
df$latitude             1  4.551   4.551  95.9685 < 2.2e-16 ***
df$longitude            1  0.414   0.414   8.7279  0.003315 **
Residuals             408 19.349   0.047
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The Covariance Matrix is given below:

Table: The coefficient covariance matrix

	(Intercept)	Log(Distance)	Date	(No of Stores)^(1/2)	Lat	Long
(Intercept)	17967.097	-0.754	-2.842	0.139	-1.325	-100.456
log(df$distance_mrt)	-0.754	0.000	0.000	0.000	0.002	0.007
df$transaction_date	-2.842	0.000	0.001	0.000	-0.002	-0.001
sqrt(df$no_of_stores)	0.139	0.000	0.000	0.000	-0.004	0.000
df$latitude	-1.325	0.002	-0.002	-0.004	1.053	-0.165
df$longitude	-100.456	0.007	-0.001	0.000	-0.165	0.870

2c. Calculate all the model selection criteria for the model. Comment on the goodness-of-fit of the model.

The criterias for model selection and the values are given below:

The Adjusted R square is better in this case than taking only one variable into consideration. Adjusted R square also takes into effect the increase in the number of variables in the regression and penalises it.
The formula for AIC is given below, it also penalises taking more variables into the regression, deals with the trade-off between the goodness of fit of the model and the simplicity of the model and the lower the value of AIC the better the model.
BIC also has the same properties of AIC but unlike the AIC, the BIC penalizes free parameters more strongly.F-statistic is greater than the F-critical value as shown above

Table: Function 'glance(mod1)' output

Rsq	AdjRsq	sig	F	pF	K	logL	AIC	BIC	dev	df.res	Obs
0.7	0.69	0.22	186.69	0	5	46.65	-79.29	-51.11	19.35	408	414

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}) \qquad \mathrm{BIC} = k\ln(n) - 2\ln(\hat{L}).$$

L(cap) is the maximum likelihood function, k is the number of variables taken into consideration and n is the number of data points.

The value of AIC is -79.3 and BIC is -51.1, these absolute values are useful only when compared to other models. The values are lower than the linear regression model.

2d. Use the RESET test to justify your selected model. If the model is not adequate, then try to adjust your model to build an appropriate model. Please discuss each step involved in the process. No marks will be given without proper interpretation and discussion of the steps.

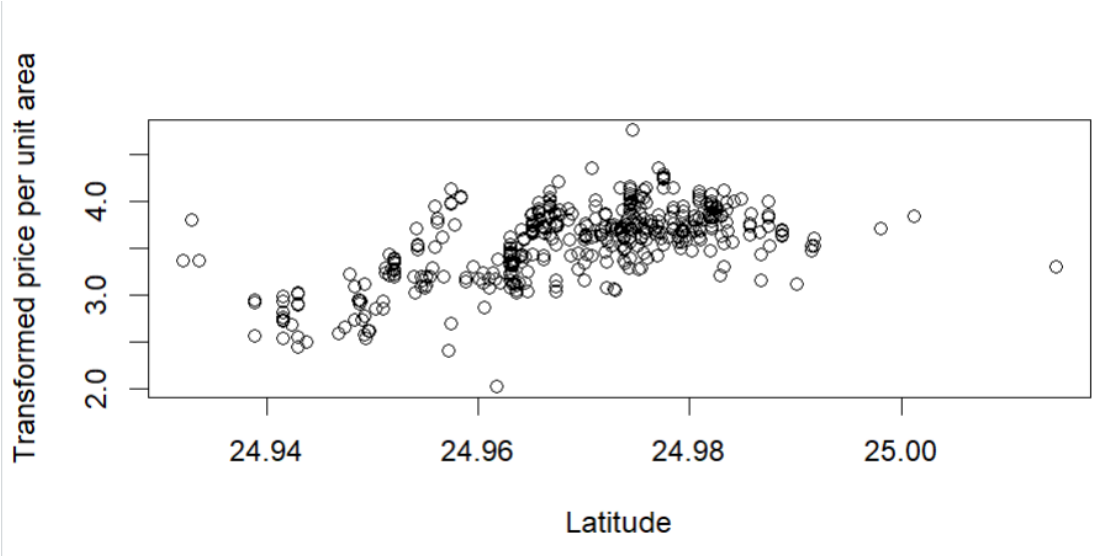
This method automatically adds higher-order polynomial terms to your model and tests the joint hypothesis that their coefficients are all zero.
Thus, the null hypothesis of the test is No higher-order polynomial terms are necessary if we reject the null hypothesis we need to consider including such terms.
In the first step using only quadratic terms of the fitted values, then using both quadratic and cubic terms.

```
        RESET test

data:  mod1
RESET = 0.045103, df1 = 1, df2 = 407, p-value = 0.8319
```

The p-value is greater than 0.05 which is alpha, so the null hypothesis is rejected.
So there is no need for any quadratic terms in order to with the data

```
        RESET test

data:  mod1
RESET = 3.6495, df1 = 2, df2 = 406, p-value = 0.02686
```

In the second case, which considers both the quadratic and cubic terms, the p-value is less than 0.05
Which means there needs to be a cubic term in order to fit the data properly
The plot given below is between the logarithmic transformed house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) and Longitude, the cubic term of latitude can be used inorder to fit the data.



The result of regression after including the cubic term of latitude is given below, the p-value has is greater than 0.05 for few of the variables but the p-value in RETEST improved from 0.02686 to 0.04458

Table: The multiple regression model

	coefficient	Std. Error	t-value	p-value
(Intercept)	17670.732	23120.642	0.764	0.445
log(df$distance_mrt)	-0.184	0.015	-11.961	0.000
df$transaction_date	0.165	0.039	4.282	0.000
sqrt(df$no_of_stores)	0.033	0.016	2.107	0.036
df$latitude	-1106.572	1389.387	-0.796	0.426
df$longitude	2.829	0.938	3.018	0.003
I(df$latitude^3)	0.597	0.743	0.803	0.422

```
        RESET test

data:  mod2
RESET = 1.416, df1 = 1, df2 = 406, p-value = 0.2348
```

```
        RESET test

data:  mod2
RESET = 3.1345, df1 = 2, df2 = 405, p-value = 0.04458
```

The plot given below is between the logarithmic transformed house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) and Longitude, the cubic term of longitude can be used inorder to fit the data but as the values of the longitude variable itself are in hundreds the value of the cubes will be large, then it is not possible to include them (Was not able to include them in R), so a transformation has to be done in order to include the cubic term.
The max and minimum value of the Longitude variable are respectively 121.47 and 121.57. The values of the longitude variable are subtracted with 120; this will not affect the distribution and then the cube of this is included for the regression

Table: The multiple regression model

	coefficient	Std. Error	t-value	p-value
(Intercept)	14222.981	23221.103	0.613	0.541
log(df$distance_mrt)	-0.178	0.016	-11.067	0.000
df$transaction_date	0.165	0.038	4.278	0.000
sqrt(df$no_of_stores)	0.035	0.016	2.225	0.027
df$latitude	-1114.903	1387.704	-0.803	0.422
df$longitude	32.420	20.972	1.546	0.123
I(df$latitude^3)	0.601	0.742	0.810	0.418
I(newlon^3)	-35.391	25.057	-1.412	0.159

Inclusion of the transformed latitude and longitude has made most of the variables insignificant but the p-value of RESET improved.
Quadratic term RESET Test:

```
        RESET test

data:  mod2
RESET = 0.55324, df1 = 1, df2 = 405, p-value = 0.4574
```

Cubic term RESET Test:

```
        RESET test

data:  mod2
RESET = 2.0395, df1 = 2, df2 = 404, p-value = 0.1314
```

The p-value has improved a lot from 0.02686 to 0.1314 but they made most of the variables insignificant which is not desirable, so sticking with the previous equation would be better.

In the quadratic RESET the NULL hypothesis is rejected for an alpha of 0.05 as p-value is 0.8319 and for the cubic and quadratic RESET the NULL hypothesis is rejected for an alpha of 0.01 as the p-value is 0.02686

2e.Test at least one joint hypothesis with linear combinations of regression coefficients from your model and comment on the results

A joint hypothesis is a set of relationships among regression parameters, relationships that need to be simultaneously true according to the null hypothesis.
To determine if being near to the MRT station by 1000 increases the house price of the unit area more than would increase int the number of stores nearby by 3 and if transaction date is increased by 2 months.

res.df	rss	df	sumsq	statistic	p.value
<db1>	<db1>	<db1>	<db1>	<db1>	<db1>
410	29.5	NA	NA	NA	NA
408	19.3	2	10.2	108.	3.14e-38

We fail to reject the null hypothesis as the p-value is less than 0.05, so being closer to the MRT station by 1000m increases the price more than increase in the number of stores by 3 or being closer to the MRT station by 1000m increases the price more than increase in the transaction by 2 months. Hence either of them is true.

2f. Test for the presence of collinearity and heteroskedasticity in your model and modify the model to control for them. Discuss the steps involved in the process with proper interpretation of the results.

The independent variables considered for this regression are:

- the distance to the nearest MRT station (unit: meter)
- the number of convenience stores in the living circle on foot (integer)
- the geographic coordinate, latitude. (unit: degree)
- the geographic coordinate, longitude. (unit: degree)

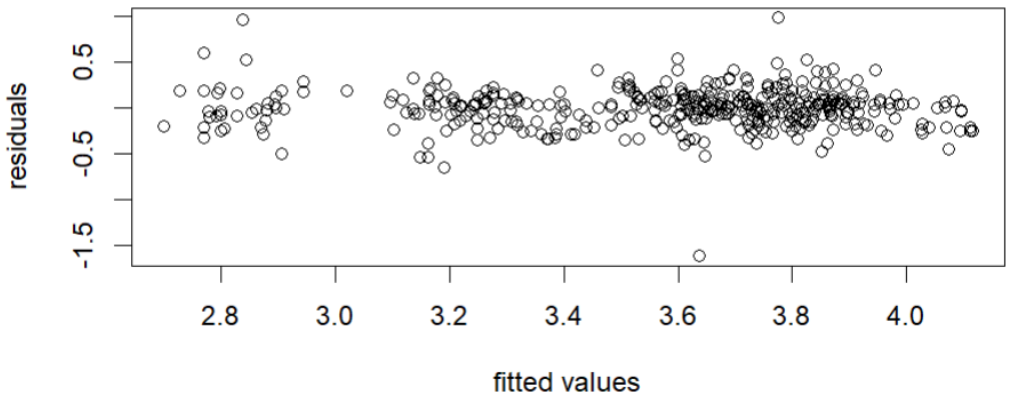
There maybe more stores near the MRT stations as people often roam around those areas, and latitude & longitude of the houses may depend on nearest MRT station and number of convenience store as people try to reside near areas with more facilities, so there might be collinearity among the variables, in order to check whether the are collinear are not VIF test is used

Table: Variance inflation factors for the 'price per unti area' regression model

regressor	VIF
log(df$distance_mrt)	2.499199
df$transaction_date	1.023980
sqrt(df$no_of_stores)	1.879595
df$latitude	1.412799
df$longitude	1.784425

If the VIF values are greater than 10 then there is that particular variable is collinear with other variables, in this case all are less than 10, so there is no collinearity among the variables

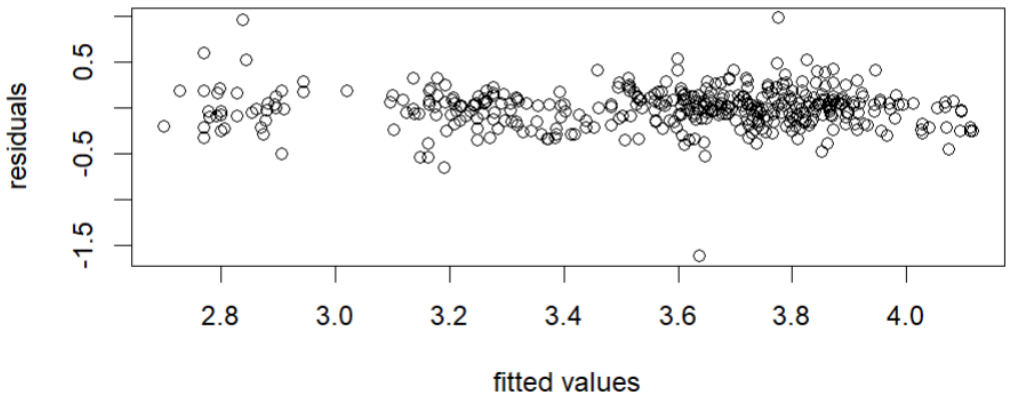
The residual plot is given below which is the residuals vs fitted values.



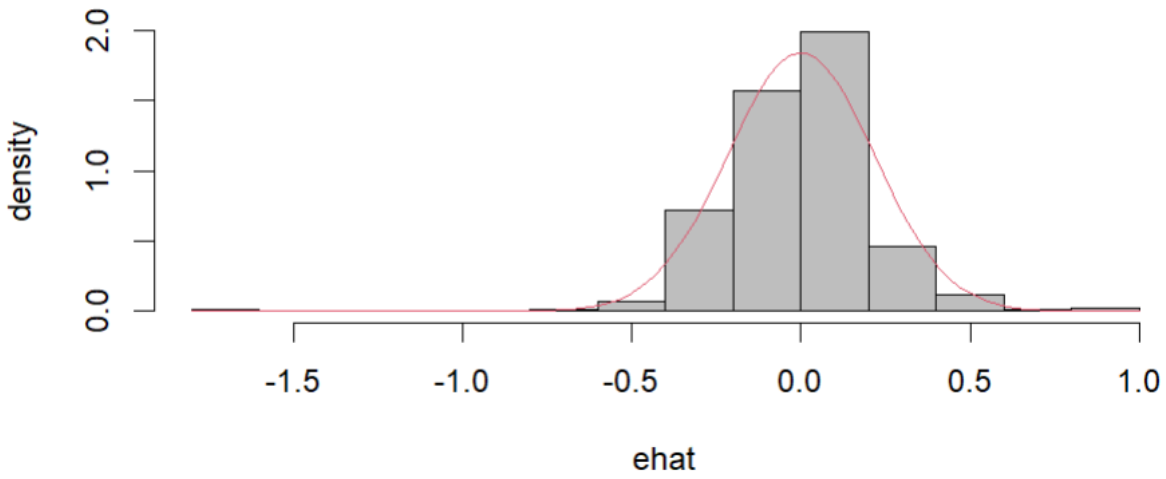
The residual plot does not show any particular pattern so it is assumed that the model is homoskedastic.

2g. For your estimated model, check whether the assumptions of no autocorrelation & normality of error term are satisfied by using scatter plots. Discuss your findings.

The residual plot does not show any specific pattern so can assume that there is no autocorrelation.



The histogram of the residuals along with density is plot is given below, the graph looks normal but due the abnormal values in the data it has a long tail towards the left.



The the results of JB test is given below the p-value is greater than 0.05 which means it is not be possible to reject the null hypothesis that the data is normally distributed in this scenario.

Table: Breusch-Pagan heteroskedasticity test

statistic	p.value	parameter	method
5.690983	0.3374579	5	studentized Breusch-Pagan test

2h. Lastly, make prediction based on your model and comment on the result

The prediction at the mean value of the variables is taken:

column	n	mean
No	414	207.500000
transaction_date	414	2013.148953
house_age	414	17.712560
distance_mrt	414	1083.885689
no_of_stores	414	4.094203
latitude	414	24.969030
longitude	414	121.533361
price_of_unit_area	414	37.980193

The prediction for the mean value is 31.99352 where the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) so the price of the house will be 319935.2. The model is able to predict the house price of unit area.