REGRESSION MODELING OF REAL ESTATE DATA

RAJ KHANDAGALE (215280002)

15/12/2021

REGRESSION MODEL OF REAL ESTAE DATASET

All About Dataset

Name of Dataset :- Real Estate data Source :- https://www.kaggle.com/dcw8161/real-estate-price-prediction/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 geographical position of property) 7) Y house price of unit area

Dimensions :- 414×8

```
setwd("C:/Users/admin/Desktop/Projects/R projects/REAL ESTATE")
```

Reading the data as Real_df

```
Real_df=read.csv('Real estate.csv' , header=TRUE)
head(Real_df)
```

```
##   No X1.transaction.date X2.house.age X3.distance.to.the.nearest.MRT.station
## 1  1            2012.917         32.0                                84.87882
## 2  2            2012.917         19.5                               306.59470
## 3  3            2013.583         13.3                               561.98450
## 4  4            2013.500         13.3                               561.98450
## 5  5            2012.833          5.0                               390.56840
## 6  6            2012.667          7.1                              2175.03000
##   X4.number.of.convenience.stores X5.latitude X6.longitude
## 1                              10    24.98298     121.5402
## 2                               9    24.98034     121.5395
## 3                               5    24.98746     121.5439
## 4                               5    24.98746     121.5439
## 5                               5    24.97937     121.5425
## 6                               3    24.96305     121.5125
##   Y.house.price.of.unit.area
## 1                       37.9
## 2                       42.2
## 3                       47.3
## 4                       54.8
## 5                       43.1
## 6                       32.1
```

```
str(Real_df)
```

```
## 'data.frame':    414 obs. of  8 variables:
##  $ No                                    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ X1.transaction.date                   : num  2013 2013 2014 2014 2013 ...
##  $ X2.house.age                          : num  32 19.5 13.3 13.3 5 7.1 34.5 20.3 31.7 17.9 ...
##  $ X3.distance.to.the.nearest.MRT.station: num  84.9 306.6 562 562 390.6 ...
##  $ X4.number.of.convenience.stores       : int  10 9 5 5 5 3 7 6 1 3 ...
##  $ X5.latitude                           : num  25 25 25 25 25 ...
##  $ X6.longitude                          : num  122 122 122 122 122 ...
##  $ Y.house.price.of.unit.area            : num  37.9 42.2 47.3 54.8 43.1 32.1 40.3 46.7 18.8 22.1 ..
```

```
dim(Real_df)
```

```
## [1] 414   8
```

```
summary(Real_df)
```

```
##        No        X1.transaction.date  X2.house.age
##  Min.   :  1.0   Min.   :2013        Min.   : 0.000
##  1st Qu.:104.2   1st Qu.:2013        1st Qu.: 9.025
##  Median :207.5   Median :2013        Median :16.100
##  Mean   :207.5   Mean   :2013        Mean   :17.713
##  3rd Qu.:310.8   3rd Qu.:2013        3rd Qu.:28.150
##  Max.   :414.0   Max.   :2014        Max.   :43.800
##  X3.distance.to.the.nearest.MRT.station X4.number.of.convenience.stores
##  Min.   :  23.38                        Min.   : 0.000
##  1st Qu.: 289.32                        1st Qu.: 1.000
##  Median : 492.23                        Median : 4.000
##  Mean   :1083.89                        Mean   : 4.094
##  3rd Qu.:1454.28                        3rd Qu.: 6.000
##  Max.   :6488.02                        Max.   :10.000
##   X5.latitude     X6.longitude   Y.house.price.of.unit.area
##  Min.   :24.93   Min.   :121.5   Min.   :  7.60
##  1st Qu.:24.96   1st Qu.:121.5   1st Qu.: 27.70
##  Median :24.97   Median :121.5   Median : 38.45
##  Mean   :24.97   Mean   :121.5   Mean   : 37.98
##  3rd Qu.:24.98   3rd Qu.:121.5   3rd Qu.: 46.60
##  Max.   :25.01   Max.   :121.6   Max.   :117.50
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1

## Warning: package 'tibble' was built under R version 4.1.1

## Warning: package 'tidyr' was built under R version 4.1.1

## Warning: package 'readr' was built under R version 4.1.1

## Warning: package 'purrr' was built under R version 4.1.1

## Warning: package 'dplyr' was built under R version 4.1.1

## Warning: package 'stringr' was built under R version 4.1.1

## Warning: package 'forcats' was built under R version 4.1.1

## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
sum(is.na(Real_df))
```

```
## [1] 0
```

our data doesnt contain any N.A. values hence it is good to define a linear regression model

model fitting

```
model1=lm(Y.house.price.of.unit.area~. , Real_df)
summary(model1)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ ., data = Real_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.003  -5.196  -0.990   4.181  75.384
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -1.404e+04  6.788e+03  -2.068  0.03927
## No                                    -3.593e-03  3.653e-03  -0.984  0.32590
## X1.transaction.date                    5.079e+00  1.559e+00   3.259  0.00121
## X2.house.age                          -2.708e-01  3.855e-02  -7.026 9.04e-12
## X3.distance.to.the.nearest.MRT.station -4.521e-03  7.189e-04  -6.289 8.28e-10
## X4.number.of.convenience.stores        1.129e+00  1.882e-01   6.000 4.37e-09
## X5.latitude                            2.247e+02  4.458e+01   5.040 7.02e-07
```

```
## X6.longitude                             -1.442e+01  4.863e+01  -0.297  0.76691
##
## (Intercept)                              *
## No
## X1.transaction.date                      **
## X2.house.age                             ***
## X3.distance.to.the.nearest.MRT.station ***
## X4.number.of.convenience.stores          ***
## X5.latitude                              ***
## X6.longitude
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.858 on 406 degrees of freedom
## Multiple R-squared:  0.5834, Adjusted R-squared:  0.5762
## F-statistic: 81.21 on 7 and 406 DF,  p-value: < 2.2e-16
```

considering a level of significance to be 1% It is found that No is just a observation number and also found to be insignificant for prediction of house price of unit area also transaction date can not be treated as predictor variable since it is not a numeric. transaction date defines time factor which we are not considering in our model. hence it should be removed from Model. so we define new model after removing **No.and transaction date**

```
model2=lm(Y.house.price.of.unit.area ~ .-No -X1.transaction.date , Real_df)
summary(model2)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ . - No - X1.transaction.date,
##     data = Real_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.546  -5.267  -1.600   4.247  76.372
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -4.946e+03  6.211e+03  -0.796    0.426
## X2.house.age                            -2.689e-01  3.900e-02  -6.896 2.04e-11
## X3.distance.to.the.nearest.MRT.station -4.259e-03  7.233e-04  -5.888 8.17e-09
## X4.number.of.convenience.stores          1.163e+00  1.902e-01   6.114 2.27e-09
## X5.latitude                              2.378e+02  4.495e+01   5.290 2.00e-07
## X6.longitude                            -7.805e+00  4.915e+01  -0.159    0.874
##
## (Intercept)
## X2.house.age                            ***
## X3.distance.to.the.nearest.MRT.station ***
## X4.number.of.convenience.stores          ***
## X5.latitude                              ***
## X6.longitude
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8.965 on 408 degrees of freedom
## Multiple R-squared:  0.5712, Adjusted R-squared:  0.5659
## F-statistic: 108.7 on 5 and 408 DF,  p-value: < 2.2e-16
```

again in newly defined model **X6 longitude** is **insignificant** hence we will remove it from model. and define our model which will give us better results than before

```
model3=lm(Y.house.price.of.unit.area ~ .-No -X1.transaction.date-X6.longitude , Real_df)
summary(model3)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ . - No - X1.transaction.date -
##     X6.longitude, data = Real_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.522  -5.292  -1.579   4.264  76.466
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            -5.916e+03  1.113e+03  -5.317 1.74e-07
## X2.house.age                           -2.687e-01  3.893e-02  -6.903 1.95e-11
## X3.distance.to.the.nearest.MRT.station -4.175e-03  4.928e-04  -8.473 4.37e-16
## X4.number.of.convenience.stores         1.165e+00  1.897e-01   6.141 1.94e-09
## X5.latitude                             2.386e+02  4.456e+01   5.355 1.43e-07
##
## (Intercept)                            ***
## X2.house.age                           ***
## X3.distance.to.the.nearest.MRT.station ***
## X4.number.of.convenience.stores        ***
## X5.latitude                            ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.954 on 409 degrees of freedom
## Multiple R-squared:  0.5711, Adjusted R-squared:  0.5669
## F-statistic: 136.2 on 4 and 409 DF,  p-value: < 2.2e-16
```

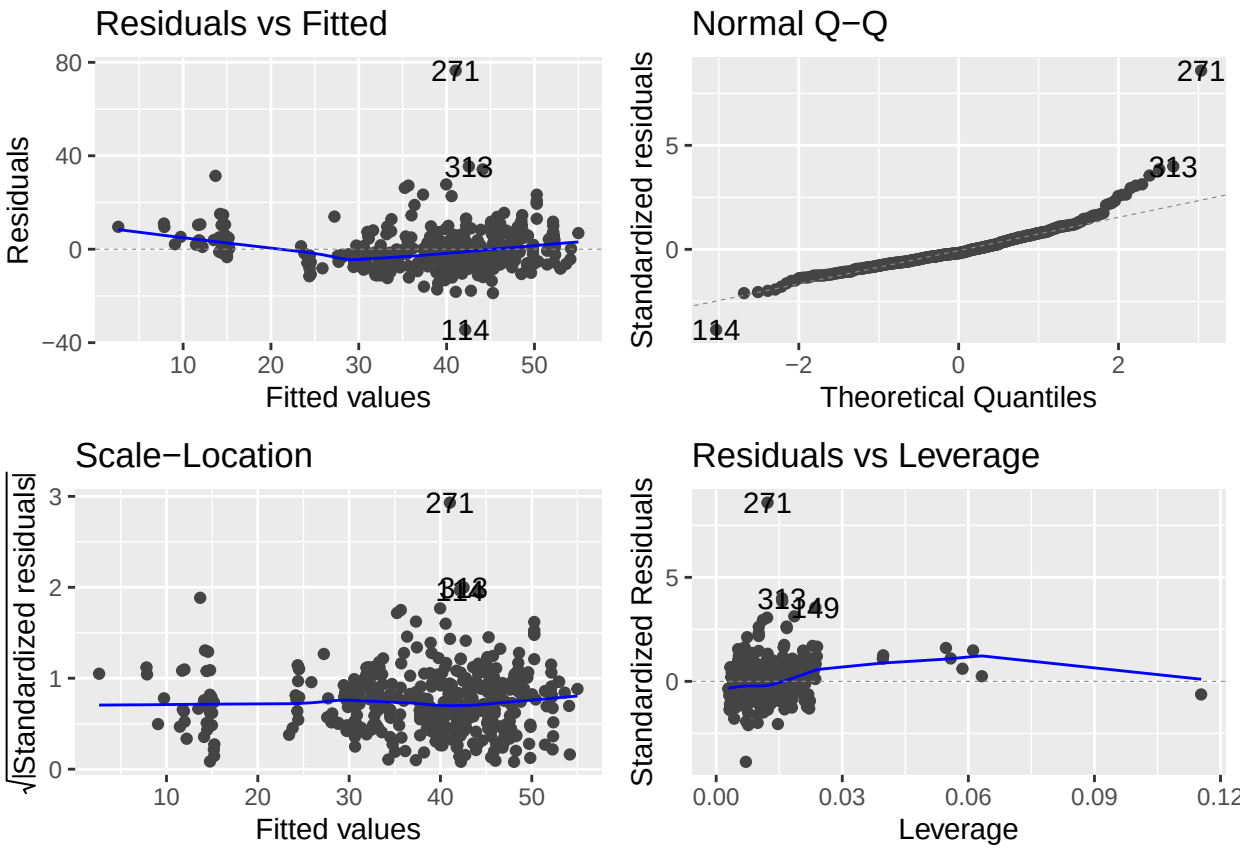
even though after removing **X6 longitude** our R-SQUARED value havent increased significantly but we got the model with all significant predictors
hence we will do some dignosis of model3 using plots

model diagnosis of model 3

```
library(ggplot2)
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 4.1.2
```

```
autoplot(model3)
```



from above plots it is clear that 1) except some earlier observations line's behovior is little non linear 2) From normal Q-Q plot we can conclude that normality is satisfied except extreme values 3) data is homoscedastic 4) some observations aer to be treated as liverage points

check for multicollineary

```
library(carData)
```

```
## Warning: package 'carData' was built under R version 4.1.1
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.1
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
vif(model3)
```

```
##                             X2.house.age X3.distance.to.the.nearest.MRT.station
##                                 1.013216                                 1.992371
##        X4.number.of.convenience.stores                            X5.latitude
##                                 1.607857                                 1.575344
```

since VIF (variance inflation factor) is less than 5 for all the predictors involved we can say that **Multicolinearity** is**not present** in the model 3

checking For Presence of interaction terms

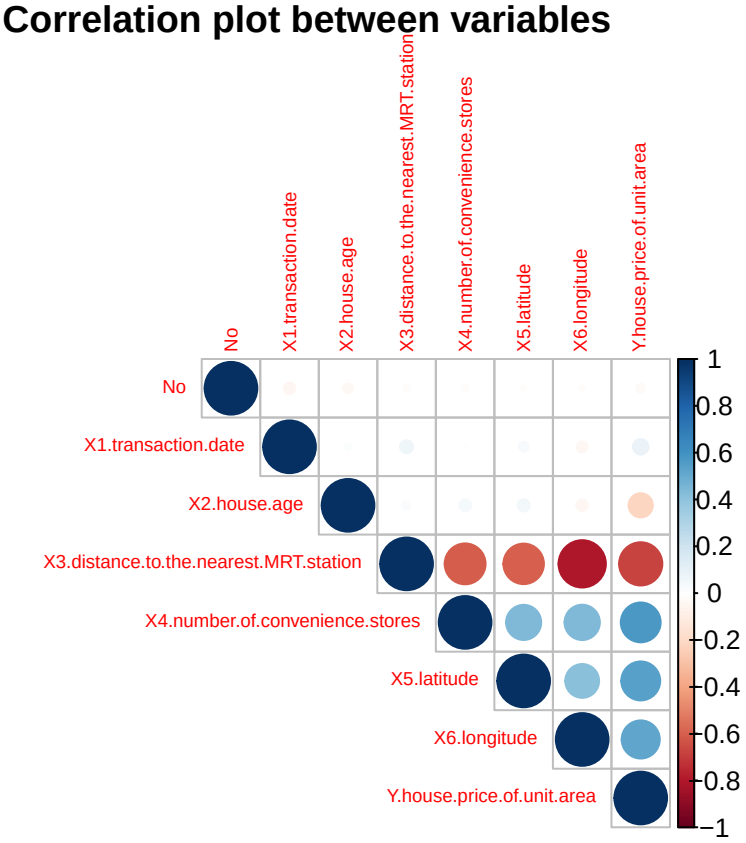
we will check the prescence of interaction terms using **Corrplot**

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.1
```

```
## corrplot 0.90 loaded
```

```
corrplot(cor(Real_df),type="upper",method="circle",title="Correlation plot between variables",
         mar=c(0.7,0.7,0.7,0.7),tl.cex = 0.6)
```



From above corrplot we can conclude that there is correlation bewteen following pairs of variables

1)X4.number.of.convenience.stores , X3.distance.to.the.nearest.MRT.station 2)X3.distance.to.the.nearest.MRT.station , X5.latitude 3)X3.distance.to.the.nearest.MRT.station , X6.longitude 4)X4.number.of.convenience.stores , X5.latitude

```
cor(Real_df$X4.number.of.convenience.stores , Real_df$X3.distance.to.the.nearest.MRT.station)
```

```
## [1] -0.6025191
```

```
cor(Real_df$X3.distance.to.the.nearest.MRT.station , Real_df$X5.latitude)
```

```
## [1] -0.5910666
```

```
cor(Real_df$X3.distance.to.the.nearest.MRT.station , Real_df$X6.longitude)
```

```
## [1] -0.8063168
```

```
cor(Real_df$X4.number.of.convenience.stores , Real_df$X5.latitude)
```

```
## [1] 0.4441433
```

1)cor(X4.number.of.convenience.stores , X3.distance.to.the.nearest.MRT.station) is $-\mathbf{0.6025191}$ 2)cor(X3.distance.to.the.nea , X5.latitude) is $-\mathbf{0.5910666}$ 3)cor(X3.distance.to.the.nearest.MRT.station , X6.longitude) is $-\mathbf{0.8063168}$ 4)cor(X4.number.of.convenience.stores , X5.latitude) is $\mathbf{0.4441433}$

correlation between **X3.distance.to.the.nearest.MRT.station and X6.longitude** is very high but X6.longitude is not included in model so their interaction term is not added in the forthcoming model Interaction terms of following terms are added in model4 1)X4.number.of.convenience.stores , X3.distance.to.the.nearest.MRT.station 2)X3.distance.to.the.nearest.MRT.station , X5.latitude

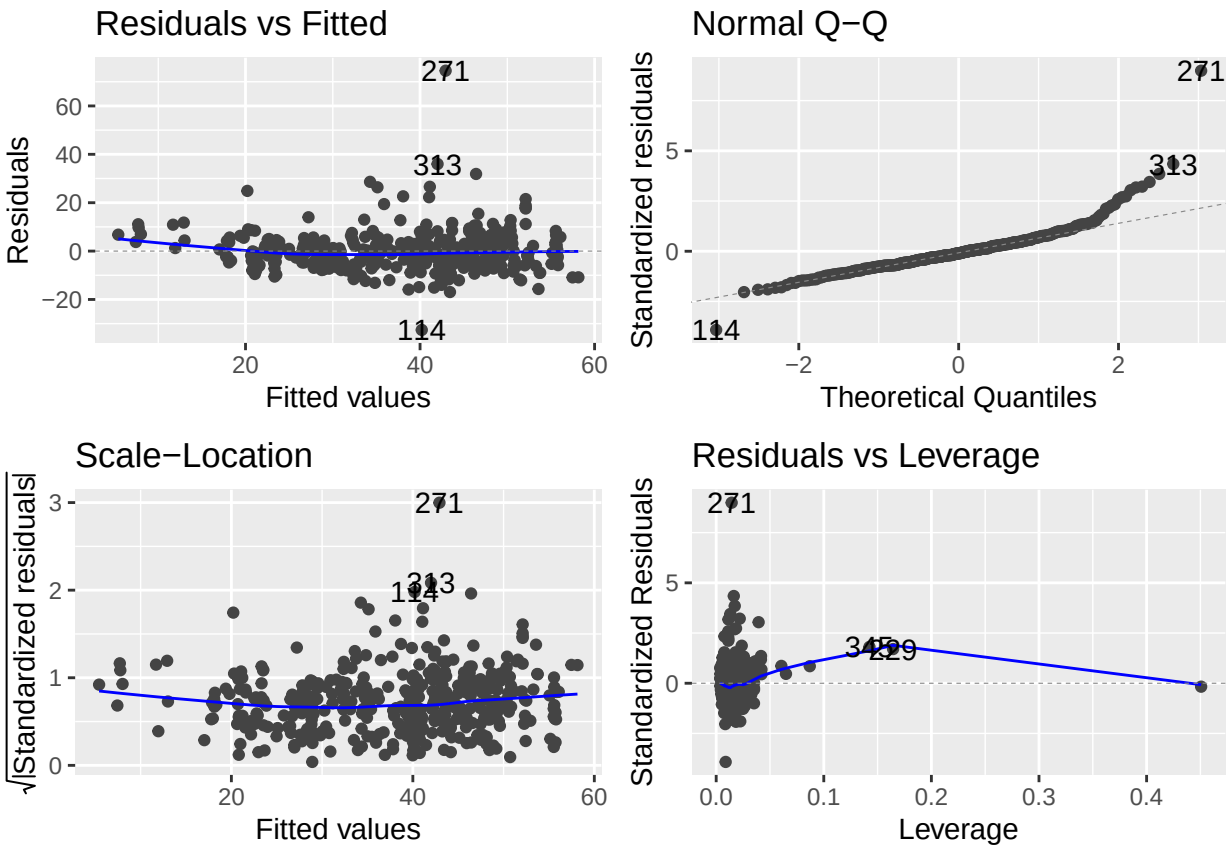
```
model4=lm(Y.house.price.of.unit.area ~ .-No -X1.transaction.date-X6.longitude + X4.number.of.convenienc
summary(model4)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ . - No - X1.transaction.date -
##     X6.longitude + X4.number.of.convenience.stores:X3.distance.to.the.nearest.MRT.station +
##     X3.distance.to.the.nearest.MRT.station:X5.latitude, data = Real_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.595  -4.758  -0.856   3.454  74.569
##
## Coefficients:
##                                                                    Estimate
## (Intercept)                                                      -1.100e+04
## X2.house.age                                                     -2.815e-01
## X3.distance.to.the.nearest.MRT.station                            2.130e+00
## X4.number.of.convenience.stores                                   1.456e+00
## X5.latitude                                                       4.424e+02
```

```
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores -1.434e-03
## X3.distance.to.the.nearest.MRT.station:X5.latitude                     -8.546e-02
##                                                                        Std. Error
## (Intercept)                                                             1.529e+03
## X2.house.age                                                            3.640e-02
## X3.distance.to.the.nearest.MRT.station                                  6.991e-01
## X4.number.of.convenience.stores                                         2.055e-01
## X5.latitude                                                             6.125e+01
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores  2.525e-04
## X3.distance.to.the.nearest.MRT.station:X5.latitude                      2.802e-02
##                                                                        t value
## (Intercept)                                                             -7.195
## X2.house.age                                                            -7.735
## X3.distance.to.the.nearest.MRT.station                                   3.047
## X4.number.of.convenience.stores                                          7.083
## X5.latitude                                                              7.223
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores  -5.680
## X3.distance.to.the.nearest.MRT.station:X5.latitude                      -3.050
##                                                                        Pr(>|t|)
## (Intercept)                                                            3.02e-12
## X2.house.age                                                           8.22e-14
## X3.distance.to.the.nearest.MRT.station                                  0.00246
## X4.number.of.convenience.stores                                        6.23e-12
## X5.latitude                                                            2.52e-12
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores 2.57e-08
## X3.distance.to.the.nearest.MRT.station:X5.latitude                      0.00244
##
## (Intercept)                                                            ***
## X2.house.age                                                           ***
## X3.distance.to.the.nearest.MRT.station                                 **
## X4.number.of.convenience.stores                                        ***
## X5.latitude                                                            ***
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores ***
## X3.distance.to.the.nearest.MRT.station:X5.latitude                     **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.354 on 407 degrees of freedom
## Multiple R-squared:  0.6285, Adjusted R-squared:  0.623
## F-statistic: 114.8 on 6 and 407 DF,  p-value: < 2.2e-16
```

model4 is model after addition of interaction terms

```
autoplot(model4)
```



from liverage plot observations 271 , 345 and 229 are found to be livearage points and removing them will yeild some efficiency in the model

```
Real_df1=Real_df[-c(271 , 345 , 229),]
dim(Real_df1)
```

```
## [1] 411   8
```

```
model5=lm(Y.house.price.of.unit.area ~ .-No -X1.transaction.date-X6.longitude + X4.number.of.convenienc
summary(model5)
```

so clearly liverage points are removed in newly stored datset Real_df1 , so we will redefine model using this dataset Real_df1

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ . - No - X1.transaction.date -
##     X6.longitude + X4.number.of.convenience.stores:X3.distance.to.the.nearest.MRT.station +
##     X3.distance.to.the.nearest.MRT.station:X5.latitude, data = Real_df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.815  -4.432  -0.470   3.635  35.506
```

```
##
## Coefficients:
##                                                                          Estimate
## (Intercept)                                                            -1.065e+04
## X2.house.age                                                           -2.767e-01
## X3.distance.to.the.nearest.MRT.station                                  2.736e+00
## X4.number.of.convenience.stores                                         1.529e+00
## X5.latitude                                                             4.284e+02
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores -1.204e-03
## X3.distance.to.the.nearest.MRT.station:X5.latitude                     -1.098e-01
##                                                                        Std. Error
## (Intercept)                                                             1.364e+03
## X2.house.age                                                            3.247e-02
## X3.distance.to.the.nearest.MRT.station                                  6.909e-01
## X4.number.of.convenience.stores                                         1.860e-01
## X5.latitude                                                             5.464e+01
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores  2.389e-04
## X3.distance.to.the.nearest.MRT.station:X5.latitude                      2.769e-02
##                                                                        t value
## (Intercept)                                                             -7.809
## X2.house.age                                                            -8.523
## X3.distance.to.the.nearest.MRT.station                                   3.960
## X4.number.of.convenience.stores                                          8.221
## X5.latitude                                                              7.839
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores  -5.042
## X3.distance.to.the.nearest.MRT.station:X5.latitude                      -3.964
##                                                                        Pr(>|t|)
## (Intercept)                                                            5.01e-14
## X2.house.age                                                           3.12e-16
## X3.distance.to.the.nearest.MRT.station                                 8.84e-05
## X4.number.of.convenience.stores                                        2.78e-15
## X5.latitude                                                            4.07e-14
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores 6.99e-07
## X3.distance.to.the.nearest.MRT.station:X5.latitude                     8.72e-05
##
## (Intercept)                                                            ***
## X2.house.age                                                           ***
## X3.distance.to.the.nearest.MRT.station                                 ***
## X4.number.of.convenience.stores                                        ***
## X5.latitude                                                            ***
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores ***
## X3.distance.to.the.nearest.MRT.station:X5.latitude                     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.426 on 404 degrees of freedom
## Multiple R-squared:  0.6819, Adjusted R-squared:  0.6772
## F-statistic: 144.4 on 6 and 404 DF,  p-value: < 2.2e-16
```

we can clearly see 5% increament in the efficeincy of model5 after removal of liverage points which indicates improvement in model

Adding Polynomial terms in the model

since X2.house.age & X3.distance.to.the.nearest.MRT.station have least p values we tested models adding higher powers of these two varible , I defined some test models and expremented with them (all those test models are not mentioned here) and got following model with appropriate polynomial terms. we will define that model as Test_odel

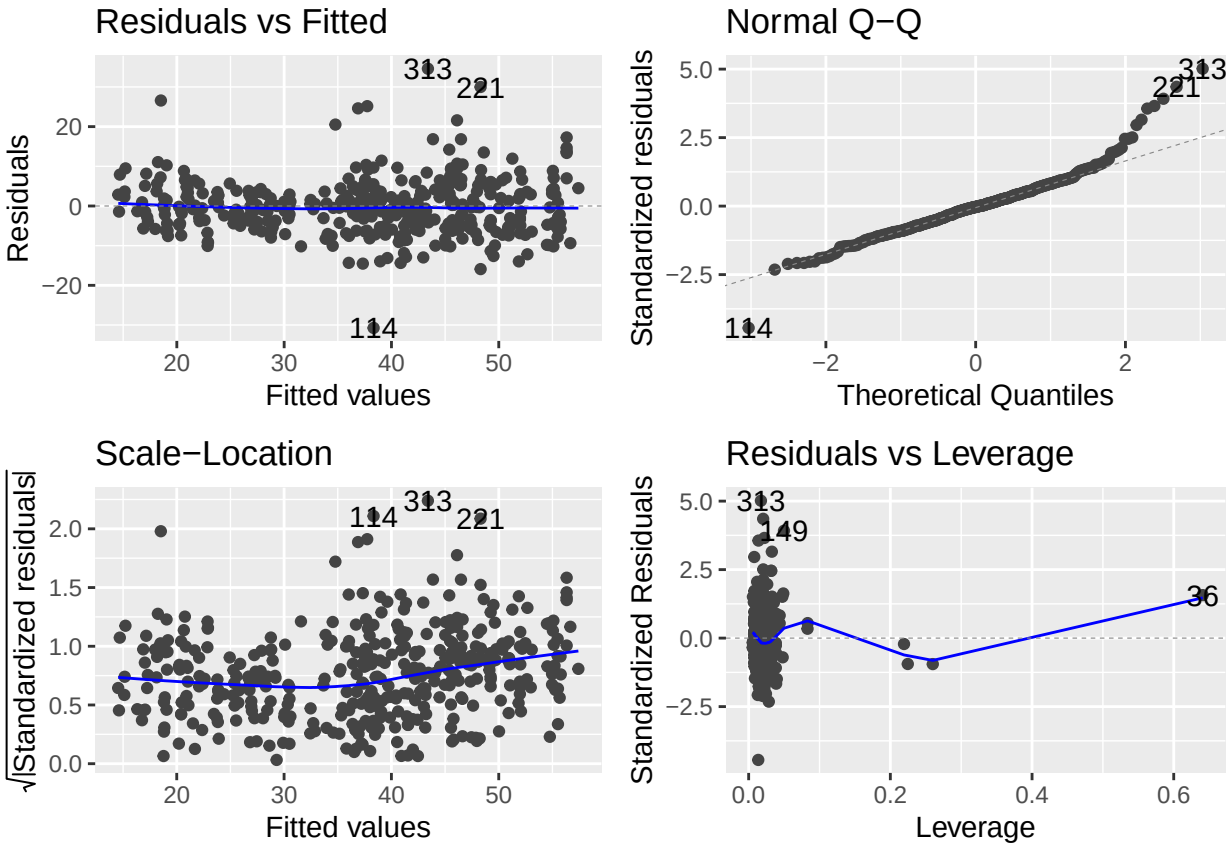
```
Test_model=lm(Y.house.price.of.unit.area ~ .-No -X1.transaction.date-X6.longitude + X4.number.of.conveni

summary(Test_model)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ . - No - X1.transaction.date -
##     X6.longitude + X4.number.of.convenience.stores:X3.distance.to.the.nearest.MRT.station +
##     X3.distance.to.the.nearest.MRT.station:X5.latitude + I(X2.house.age^2) +
##     I(X3.distance.to.the.nearest.MRT.station^2), data = Real_df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.738  -4.339  -0.218   3.588  34.602
##
## Coefficients:
##                                                                   Estimate
## (Intercept)                                                     -1.001e+04
## X2.house.age                                                    -1.019e+00
## X3.distance.to.the.nearest.MRT.station                           2.408e+00
## X4.number.of.convenience.stores                                  1.144e+00
## X5.latitude                                                      4.030e+02
## I(X2.house.age^2)                                                1.857e-02
## I(X3.distance.to.the.nearest.MRT.station^2)                      6.692e-07
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores -7.870e-04
## X3.distance.to.the.nearest.MRT.station:X5.latitude               -9.674e-02
##                                                                  Std. Error
## (Intercept)                                                      1.325e+03
## X2.house.age                                                     1.189e-01
## X3.distance.to.the.nearest.MRT.station                           6.626e-01
## X4.number.of.convenience.stores                                  1.959e-01
## X5.latitude                                                      5.308e+01
## I(X2.house.age^2)                                                2.891e-03
## I(X3.distance.to.the.nearest.MRT.station^2)                      2.205e-07
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores  2.452e-04
## X3.distance.to.the.nearest.MRT.station:X5.latitude               2.655e-02
##                                                                  t value
## (Intercept)                                                       -7.555
## X2.house.age                                                      -8.570
## X3.distance.to.the.nearest.MRT.station                             3.635
## X4.number.of.convenience.stores                                    5.843
## X5.latitude                                                        7.592
## I(X2.house.age^2)                                                  6.423
## I(X3.distance.to.the.nearest.MRT.station^2)                        3.035
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores  -3.210
## X3.distance.to.the.nearest.MRT.station:X5.latitude                 -3.644
##                                                                  Pr(>|t|)
```

```
## (Intercept)                                                          2.85e-13
## X2.house.age                                                         2.25e-16
## X3.distance.to.the.nearest.MRT.station                               0.000315
## X4.number.of.convenience.stores                                      1.06e-08
## X5.latitude                                                          2.22e-13
## I(X2.house.age^2)                                                    3.78e-10
## I(X3.distance.to.the.nearest.MRT.station^2)                          0.002564
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores 0.001434
## X3.distance.to.the.nearest.MRT.station:X5.latitude                   0.000303
##
## (Intercept)                                                          ***
## X2.house.age                                                         ***
## X3.distance.to.the.nearest.MRT.station                               ***
## X4.number.of.convenience.stores                                      ***
## X5.latitude                                                          ***
## I(X2.house.age^2)                                                    ***
## I(X3.distance.to.the.nearest.MRT.station^2)                          **
## X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores **
## X3.distance.to.the.nearest.MRT.station:X5.latitude                   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.96 on 402 degrees of freedom
## Multiple R-squared:  0.7219, Adjusted R-squared:  0.7164
## F-statistic: 130.5 on 8 and 402 DF,  p-value: < 2.2e-16
```

analysis from plots of Test model

```
autoplot(Test_model)
```



since value of $\mathbf{R-Squared}$ is 71.64 % we conclude that Test_model is our Final model

Final model is

Y.house.price.of.unit.area = -1.001e+04 -(1.019e+00)X2.house.age+(2.408e+00)X3.distance.to.the.nearest.MRT.station
+ (1.144e+00)X4.number.of.convenience.stores+(4.030e+02)X5.latitude+(1.857e-02)(X2.house.age$^{2)+(6.692e-07)(X3.distance.to.t}$
(7.870e-04)[X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores]-(9.674e-02)[X3.distance.to.the.nearest.M

Conclusion of the final model:

1) R-Squared value of our final model is 71.64% .
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