Introduction to Data Science (Assignment 1)

Quang Dong Nguyen

2022-08-25

Question 01- Regression

```
library(r2symbols)
```

```
## Warning: package 'r2symbols' was built under R version 4.1.3
```

```
kc <- read.csv("kc_house.csv")
View(kc)
head(kc)
```

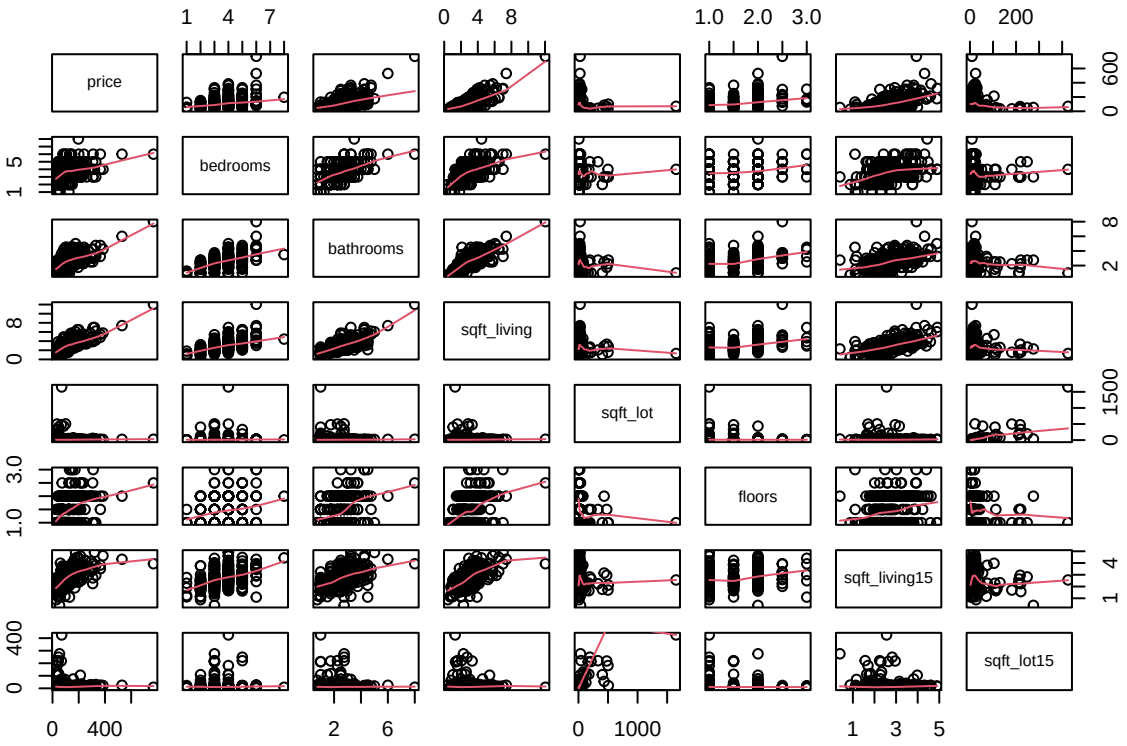
```
##           id  price bedrooms bathrooms sqft_living sqft_lot floors waterfront
## 1 7922800400  95.10        5      3.25        3.25   14.342      2          0
## 2 1516000055  65.00        3      2.25        2.15   21.235      1          0
## 3 2123039032  36.99        1      0.75        0.76   10.079      1          1
## 4 9297300045  55.00        3      2.00        1.97    4.166      2          0
## 5 1860600135 238.40        5      2.50        3.65    9.050      2          0
## 6 1560930070  84.00        4      3.50        2.84   40.139      1          0
##   sqft_living15 sqft_lot15
## 1          2.96     11.044
## 2          2.57     18.900
## 3          1.23     14.267
## 4          2.39      4.166
## 5          2.88      5.400
## 6          3.18     36.852
```

```
str(kc)
```

```
## 'data.frame':    341 obs. of  10 variables:
##  $ id           : num  7.92e+09 1.52e+09 2.12e+09 9.30e+09 1.86e+09 ...
##  $ price        : num  95.1 65 37 55 238.4 ...
##  $ bedrooms     : int  5 3 1 3 5 4 3 3 3 2 ...
##  $ bathrooms    : num  3.25 2.25 0.75 2 2.5 3.5 3 2.5 3 2.25 ...
##  $ sqft_living  : num  3.25 2.15 0.76 1.97 3.65 ...
##  $ sqft_lot     : num  14.34 21.23 10.08 4.17 9.05 ...
##  $ floors       : num  2 1 1 2 2 1 2 2 2 1.5 ...
##  $ waterfront   : int  0 0 1 0 0 0 1 1 0 0 ...
##  $ sqft_living15: num  2.96 2.57 1.23 2.39 2.88 3.18 2.28 3.98 3.08 2.13 ...
##  $ sqft_lot15   : num  11.04 18.9 14.27 4.17 5.4 ...
```

1) Construct the matrix plot and correlation matrix (consider only relevant variables). Comment on the relationship among variable

```
#Matrix plot of relevant variables
pairs(price ~
      bedrooms + bathrooms + sqft_living  + sqft_lot + floors + sqft_living15 +sqft_lot15,
      data=kc,
      panel=panel.smooth)
```



In the matrix plot of relevant variables, there are some observable positive linear trend between these following variables:

- price and bedrooms ; price and bathrooms; price and sqft_living; price and sqft_living15.
- bedroom and bathrooms; bedrooms and sqft_living; bedrooms and sqft_living15.
- bathrooms and sqft_living; bathrooms and sqft_living15.
- sqft_living and sqft_living15.

There are also some negative trends in between some variables, but are not significantly big enough to consider a linear relationship between them.

```
#Correlation matrix of relevant variables
cor(kc[c(2,3,4,5,6,7,9,10)])
```

```
##                    price     bedrooms  bathrooms sqft_living     sqft_lot
## price          1.00000000  0.365855613  0.6499829   0.7880793 -0.089022456
## bedrooms       0.36585561  1.000000000  0.5544973   0.5579069 -0.006739024
## bathrooms      0.64998292  0.554497306  1.0000000   0.7809491 -0.121073782
## sqft_living    0.78807930  0.557906874  0.7809491   1.0000000 -0.104880686
## sqft_lot      -0.08902246 -0.006739024 -0.1210738  -0.1048807  1.000000000
## floors         0.37379631  0.192062371  0.4189076   0.3581140 -0.076466191
## sqft_living15  0.60567854  0.392195987  0.4941674   0.6493149 -0.050714790
## sqft_lot15    -0.14040054 -0.021601830 -0.1275580  -0.1292349  0.735502746
##                    floors sqft_living15  sqft_lot15
## price          0.37379631    0.60567854 -0.14040054
## bedrooms       0.19206237    0.39219599 -0.02160183
## bathrooms      0.41890757    0.49416739 -0.12755804
## sqft_living    0.35811398    0.64931494 -0.12923490
## sqft_lot      -0.07646619   -0.05071479  0.73550275
## floors         1.00000000    0.19516827 -0.06666012
## sqft_living15  0.19516827    1.00000000 -0.12127202
## sqft_lot15    -0.06666012   -0.12127202  1.00000000
```

By assessing on the correlation between variables, we can see:

- A strong positive correlation price and sqft_living , price and sqft_living15; bathrooms and sqft_living.
- An intermediate positive correlation price and bathrooms; bedrooms and bathrooms, bedrooms and sqft_living; bathrooms and sqft_living15; sqft_living and sqft_living15.
- A weak positive correlation price and bedrooms; bedrooms and sqft_living15.

2) Simple Linear Regression

i/ Fit a model to predict price in terms of sqft_living

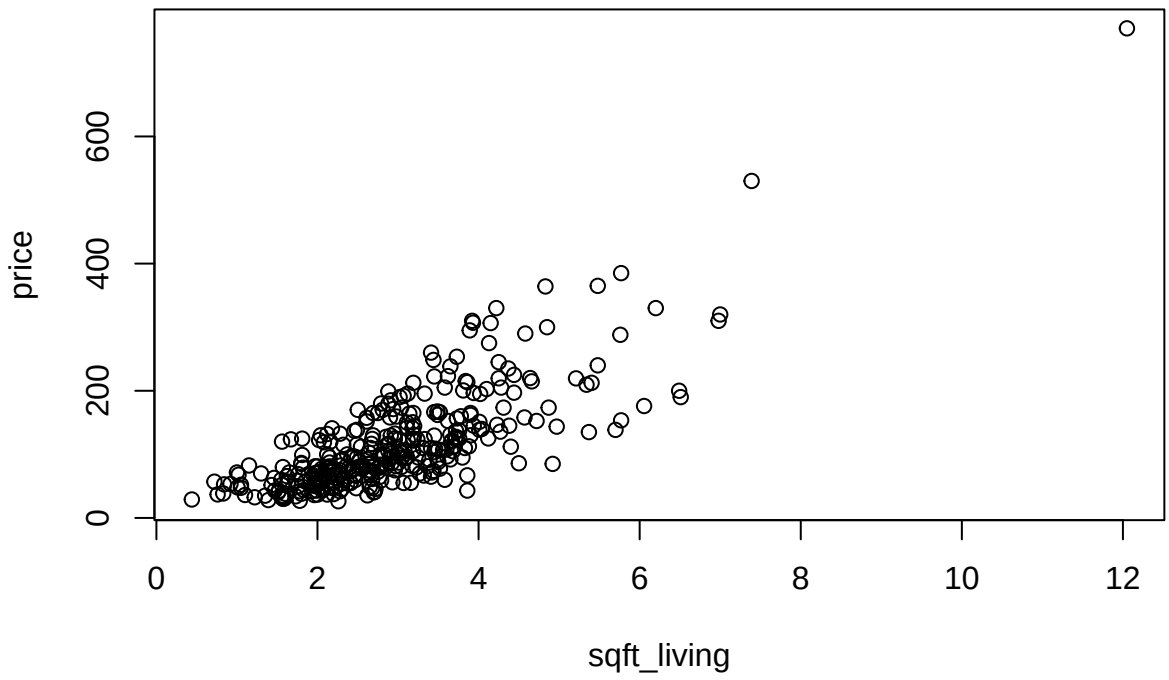
```
#shows first 6 elements of kc
head(kc)
```

```
##           id  price bedrooms bathrooms sqft_living sqft_lot floors waterfront
## 1 7922800400  95.10        5      3.25        3.25   14.342      2          0
## 2 1516000055  65.00        3      2.25        2.15   21.235      1          0
## 3 2123039032  36.99        1      0.75        0.76   10.079      1          1
## 4 9297300045  55.00        3      2.00        1.97    4.166      2          0
## 5 1860600135 238.40        5      2.50        3.65    9.050      2          0
## 6 1560930070  84.00        4      3.50        2.84   40.139      1          0
##   sqft_living15 sqft_lot15
## 1          2.96     11.044
## 2          2.57     18.900
## 3          1.23     14.267
## 4          2.39      4.166
## 5          2.88      5.400
## 6          3.18     36.852
```

```
nrow(kc) #population size is large enough to predict a model
```

```
## [1] 341
```

```
plot(price~sqft_living, data=kc)
```




```
#the scatterplot seems to have a linear trend between the response variable (price)
#and the explanatory variable (sqft_living)

#Fit the model of price in terms of sqft_living
model1 <- lm(price~sqft_living, data=kc)
model1
```

```
##
## Call:
## lm(formula = price ~ sqft_living, data = kc)
##
## Coefficients:
## (Intercept)  sqft_living
##      -33.98        50.95
```

The graph shows a sense of evidence of a linear relationship between the price variable and sqft_living

ii/ Discuss the significance of the slope parameter estimate. Write down the relevant hypothesis

H0: B = 0. There is no evidence of a linear relationship between sqft_living and price

HA: B != 0. There is evidence of a linear relationship between sqft_living and price

```
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ sqft_living, data = kc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -131.704  -29.885   -6.956   24.696  190.005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -33.982      6.956  -4.885 1.59e-06 ***
## sqft_living   50.952      2.162  23.572  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.33 on 339 degrees of freedom
## Multiple R-squared:  0.6211, Adjusted R-squared:   0.62
## F-statistic: 555.6 on 1 and 339 DF,  p-value: < 2.2e-16
```

Based on summary made above about the linear model:

- P-value is significant, (p-value is less than 0.05)

Therefore, there is a strong evidence to reject the null hypothesis at 5% level of significance, and support the alternative hypothesis: B != 0

The slope estimate of the parameter is not equal to 0.

There is a linear relationship between the two variables, which are price and sqft_living.

iii/ Discuss the accuracy of the parameter estimates. (Standard errors/ confidence intervals)

```
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ sqft_living, data = kc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -131.704  -29.885   -6.956   24.696  190.005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -33.982      6.956  -4.885 1.59e-06 ***
## sqft_living   50.952      2.162  23.572  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 49.33 on 339 degrees of freedom
## Multiple R-squared:  0.6211, Adjusted R-squared:   0.62
## F-statistic: 555.6 on 1 and 339 DF,  p-value: < 2.2e-16
```

On average, the estimated value for the intercept can vary from the true value by 6.956 units.

On average, the estimated value for the slope B of sqft_living can vary from its true value by 2.162 units.

```
#confidence interval:
confint(model1, level=0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) -47.66432 -20.30013
## sqft_living  46.70061  55.20429
```

Therefore, the confidence interval of the variable's coefficient shown its ranging of 95% CI is in between 46.70061 and 55.20429. Meaning on average, for each 1 unit increased in sqft_living will have the price in between 46.70061 and 55.20429. In other words, for each 1000 sq.ft increased in the square footage of the apartments interior living will have the price in between 46700.61 dollars and 55204.29 dollars

For the intercept, in the absence of sqft_living variables in the parameter, the price will , on average, lay between -47.66432 and -20.30013 units on the level of confidence of 95% accurate. Meaning no money is charged when the sqft_living of the apartment's interior living space does not exist.

iv/ Discuss the model accuracy. (R-squared, residual standard error, etc)

```
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ sqft_living, data = kc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -131.704  -29.885   -6.956   24.696  190.005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -33.982      6.956  -4.885 1.59e-06 ***
## sqft_living   50.952      2.162  23.572  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.33 on 339 degrees of freedom
## Multiple R-squared:  0.6211, Adjusted R-squared:   0.62
## F-statistic: 555.6 on 1 and 339 DF,  p-value: < 2.2e-16
```

Based on the parameter shown above, we see:

- p-value of sqft_living is significantly small (1.59e-06, less than 0.05) and its t-value is large, meaning there shown to have a linear relationship with the response variable(price).

- However the Residual Standard error is large (49.33 on 339 df), which explains the fan-shaped trend (meaning the scatterplot of the residual model explained later on, will spread out) when plotting.
- Multiple R-squared refers to 62% of the variance in the response variable (price) is explained by the model, supporting moderate linear relationship

```
qf(0.95, 1, 339) #quantified value
```

```
## [1] 3.869036
```

- F-statistic of the model1 is much larger than 3.869036. As a result, the model is appropriate, and between sqft_living and price, there is a significant linear relationship.

Therefore, the model has shown the evidence of a linear trend among the response variable and the explanatory variable.

```
anova(model1)
```

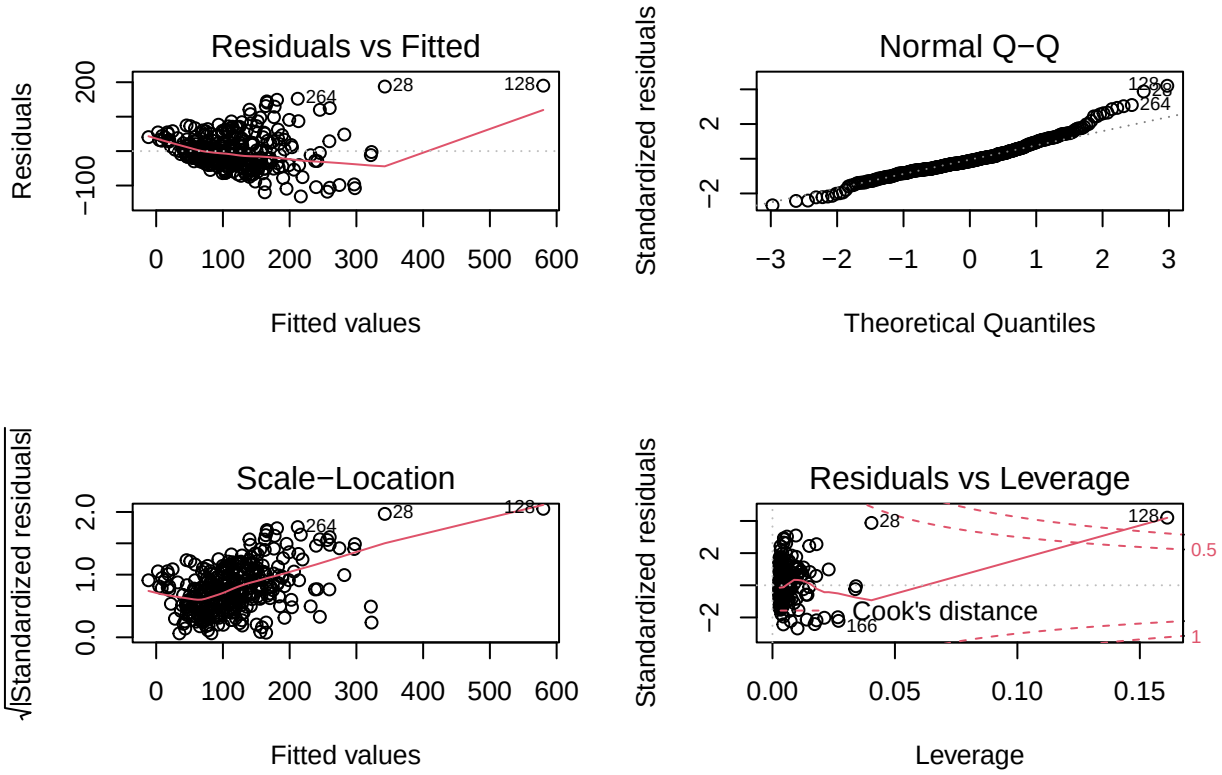
```
## Analysis of Variance Table
##
## Response: price
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## sqft_living   1 1351998 1351998  555.62 < 2.2e-16 ***
## Residuals   339  824891    2433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the overall assess on the accuracy of the model, we observe:

- High sum of square suggests the high variability among the observations and their fitted model1; this also demonstrates the fan-shaped of the dataset when plotting it out in the residual vs fitted model as in the below section (v/ check for the model assumptions)

v/ Check for the model assumptions

```
par(mfrow=c(2,2))
plot(model1)
```



Assumption checking:

- In graph 1, the horizontal line is not straightly flat, it scales upward when the fitted value reaches approximately 350 on the x - axis. However, the residual equally scatters out, like a fan equally to both sides. Thus, it suggests that the error variances are not equal.
- Furthermore, in the scale-location plot, the standardised residuals seems to increase for every increasing in fitted values, which created the fan-shaped model.
- In the Q-Q plot: points don't lie on the straight line. Hence, the data does not meet the normality assumption

vi/ Write down the model equation

```
model1
```

```
##
## Call:
## lm(formula = price ~ sqft_living, data = kc)
##
## Coefficients:
## (Intercept)  sqft_living
##      -33.98        50.95
```

Therefore, the fitted model is:

E(Y) = -33.98222 + 50.95245 *sqft_living

E(price) = -33.98222 + 50.95245* sqft_living

Where:

- price - measured in ten thousand dollars
- sqft_living - measured in thousand sq.ft

vii/ Predict the price of a house with 10,000 sq.ft of the apartments interior living space (sqft_living)

```
predict(model1, newdata= data.frame(sqft_living = 10))
```

```
##        1
## 475.5423
```

Hence, with a house of 10,000 sq.ft interior living space. The predicted price would be around 475.5423 (in ten thousand dollar). Meaning, the actual price would be 475.5423 * 10,000 = 4,755,423 dollars.

3) Multiple Linear Regression

i/ Fit a model to predict price in terms of all the other quantitative predictors (numerical predictors)

```
str(kc) #to check all the quantitative predictors
```

```
## 'data.frame':    341 obs. of  10 variables:
##  $ id           : num  7.92e+09 1.52e+09 2.12e+09 9.30e+09 1.86e+09 ...
##  $ price        : num  95.1 65 37 55 238.4 ...
##  $ bedrooms     : int  5 3 1 3 5 4 3 3 3 2 ...
##  $ bathrooms    : num  3.25 2.25 0.75 2 2.5 3.5 3 2.5 3 2.25 ...
##  $ sqft_living  : num  3.25 2.15 0.76 1.97 3.65 ...
##  $ sqft_lot     : num  14.34 21.23 10.08 4.17 9.05 ...
##  $ floors       : num  2 1 1 2 2 1 2 2 2 1.5 ...
##  $ waterfront   : int  0 0 1 0 0 0 1 1 0 0 ...
##  $ sqft_living15: num  2.96 2.57 1.23 2.39 2.88 3.18 2.28 3.98 3.08 2.13 ...
##  $ sqft_lot15   : num  11.04 18.9 14.27 4.17 5.4 ...
```

All quantitative variables include: id, bedrooms, bathrooms, sqft_living, sqft_lot, floors, sqft_living15, sqft_lot15.

```
#Fit the model price in terms of other quantitative predictors
model3 <- lm(price~ .,data= kc)
```

ii/ Remove the insignificant variables and fit a model including the rest of the variables

```
summary(model3)
```

```
##
## Call:
## lm(formula = price ~ ., data = kc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -109.950  -27.671   -0.099   21.052  238.156
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.848e+01  1.289e+01  -6.090 3.12e-09 ***
## id            -8.375e-10  8.304e-10  -1.009   0.3139
## bedrooms      -3.309e+00  2.862e+00  -1.156   0.2484
## bathrooms      6.981e+00  4.263e+00   1.638   0.1024
## sqft_living    3.946e+01  3.495e+00  11.290  < 2e-16 ***
## sqft_lot       3.552e-02  3.190e-02   1.114   0.2663
## floors         1.171e+01  5.017e+00   2.335   0.0201 *
## waterfront     5.295e+01  6.147e+00   8.614 2.93e-16 ***
## sqft_living15  1.913e+01  4.222e+00   4.532 8.18e-06 ***
## sqft_lot15    -1.288e-01  8.235e-02  -1.563   0.1189
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.55 on 331 degrees of freedom
## Multiple R-squared:  0.7247, Adjusted R-squared:  0.7172
## F-statistic: 96.83 on 9 and 331 DF,  p-value: < 2.2e-16
```

Some insignificant variables are:

- id: p-value is insignificant (0.3139, which is >0.05) to fit the variable into the linear regression model.
- bathrooms: p-value is large (0.2484, which is >0.05). Thus, it cannot fit into the multiple linear regression model.
- sqft_lot: p-value is insignificant (0.2663, which is >0.05). Thus, the variable will be removed from the model.
- sqft_lot15: p-value is large (0.1189, which is >0.05). Thus, it will be removed from the model.

Thus, the fitted model is shown as below:

```
model4 <- lm(price~ sqft_living + floors + sqft_living15, data= kc)
summary(model4)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + floors + sqft_living15, data = kc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.823  -26.521   -4.337   20.673  236.481
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -80.762     11.520  -7.011  1.3e-11 ***
## sqft_living     41.227      2.893  14.251  < 2e-16 ***
## floors          17.504      5.407   3.237  0.00133 **
## sqft_living15   18.715      4.697   3.984  8.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.73 on 337 degrees of freedom
## Multiple R-squared:  0.6473, Adjusted R-squared:  0.6442
## F-statistic: 206.2 on 3 and 337 DF,  p-value: < 2.2e-16
```

iii/ Add the Interaction term bedrooms*floors to the model above (part ii)

```
model5 <- lm(price~
               sqft_living + floors + sqft_living15 + I(bedrooms*floors),
             data=kc)
```

iv/ Comment on the significance of the parameter estimates of the model above (part iii)

```
summary(model5)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + floors + sqft_living15 + I(bedrooms *
##     floors), data = kc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.431  -27.146   -4.252   21.327  233.480
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -85.930     12.188  -7.050 1.02e-11 ***
## sqft_living             43.155      3.255  13.259  < 2e-16 ***
## floors                  26.852      9.049   2.968  0.00322 **
## sqft_living15           18.392      4.699   3.914  0.00011 ***
## I(bedrooms * floors)    -2.477      1.924  -1.288  0.19872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.68 on 336 degrees of freedom
## Multiple R-squared:  0.649,  Adjusted R-squared:  0.6449
## F-statistic: 155.3 on 4 and 336 DF,  p-value: < 2.2e-16
```

We do the hypothesis testing:

- H0 : Bi = 0. There is no evidence of linear relationship between the response variable and the explanatory variable
- HA: Bi != 0 . There is evidence of a linear relationship between the response variable and the explanatory variable

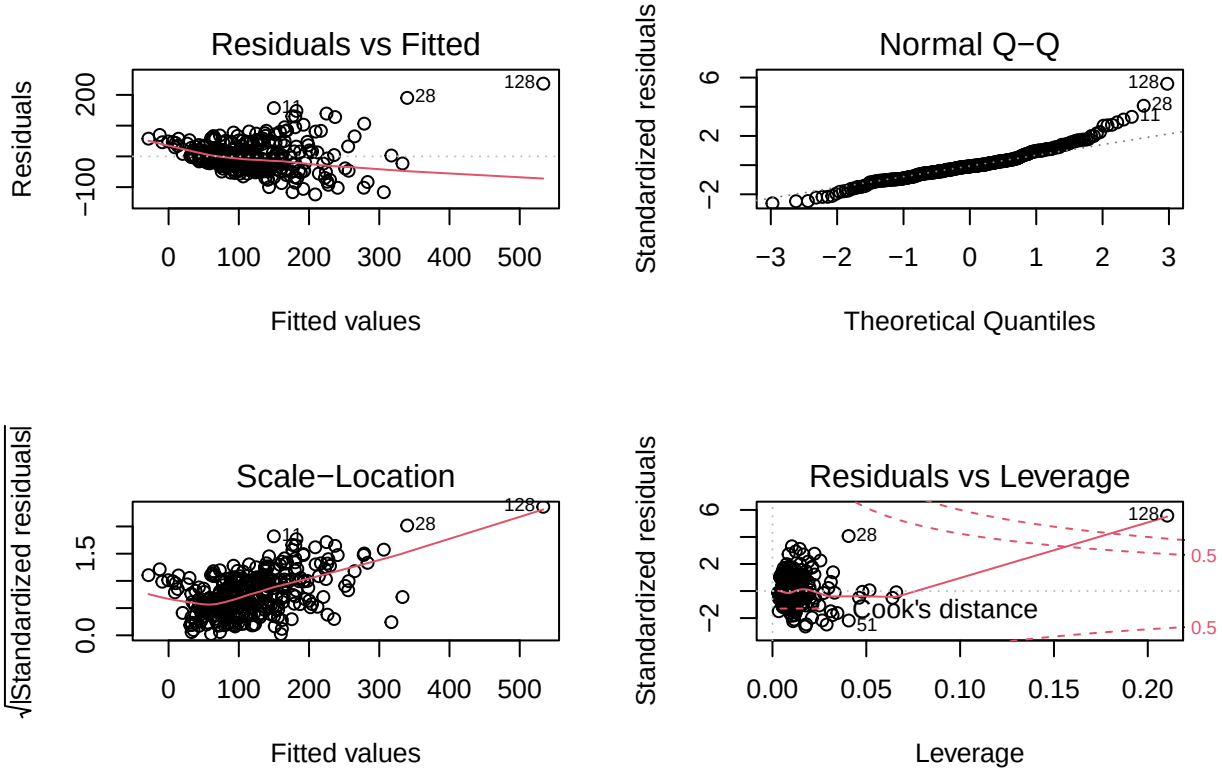
With the interaction term bedrooms * floors being added, the model become less accurate. As the interaction term along, though has a small std.error, it still has small t-value and its p-value is insignificant (0.19872, which is >0.05).

Therefore, there is enough evidence at 5% level of significance to not reject the null hypothesis. There is no evidence of a linear relationship between price and the interaction term of bedrooms*floors.

v/ Check for the model assumptions (model in part iii)

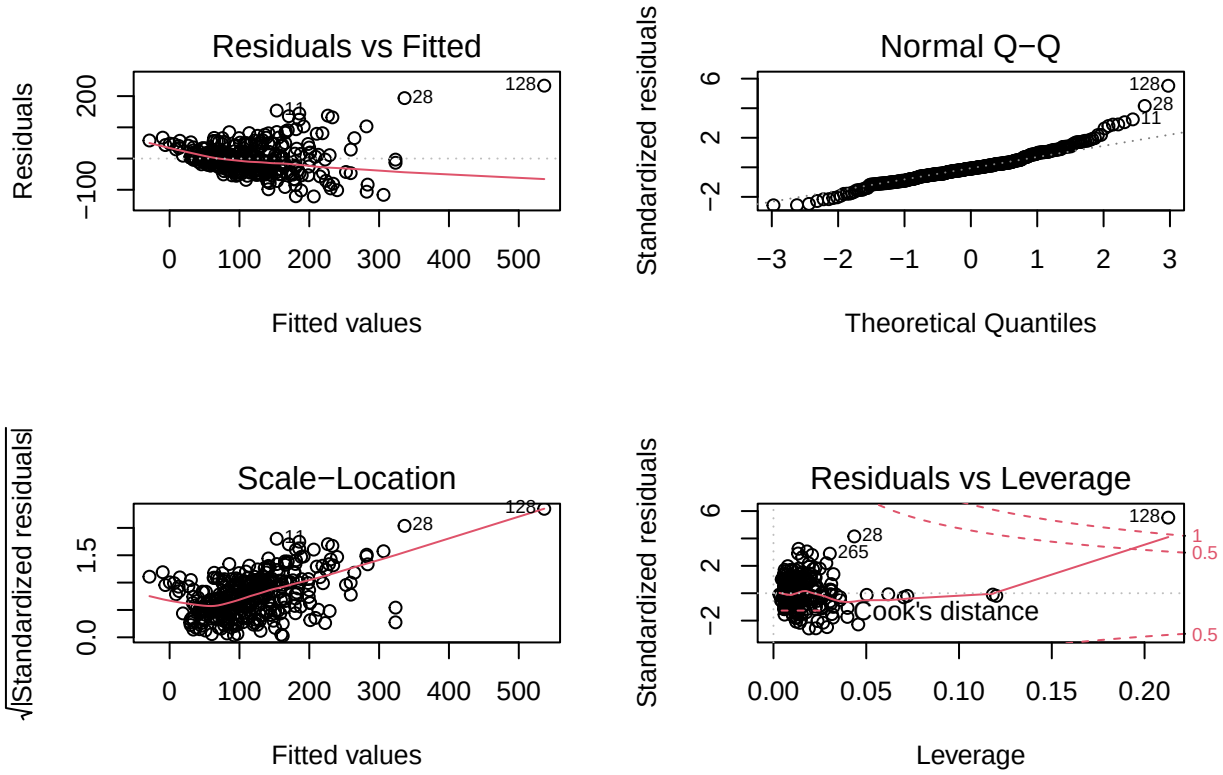
Model4's assumption checking:

```
par(mfrow=c(2,2))
plot(model4)
```



Model5's assumption checking:

```
par(mfrow=c(2,2))
plot(model5)
```



As we can observe from both plots a very similar trending between each graphs. In details:

- Homoscedasticity - both graphs also have a fan-shaped trend. Therefore, both model4 and model5 have their constant variance assumption not met the standard.
- Normality - Both graphs of Q-Q plot is not normal, the standardised residuals data tends to bends upward toward the end of the graph. This suggests both model4 and model5 are not normally distributed.
- In graph of the scale-location of both models, their standardised residuals increase as the fitted values increase, which supports the fan-shaped trend in the model.

vi/ Compare and comment on the accuracy of the models in part ii and part iii. Suggest the best model

Final comparison:

```
summary(model4)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + floors + sqft_living15, data = kc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -123.823  -26.521   -4.337   20.673  236.481
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -80.762     11.520  -7.011  1.3e-11 ***
## sqft_living     41.227      2.893  14.251  < 2e-16 ***
## floors          17.504      5.407   3.237  0.00133 **
## sqft_living15   18.715      4.697   3.984  8.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.73 on 337 degrees of freedom
## Multiple R-squared:  0.6473, Adjusted R-squared:  0.6442
## F-statistic: 206.2 on 3 and 337 DF,  p-value: < 2.2e-16
```

```
summary(model5)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + floors + sqft_living15 + I(bedrooms *
##     floors), data = kc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.431  -27.146   -4.252   21.327  233.480
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -85.930     12.188  -7.050 1.02e-11 ***
## sqft_living             43.155      3.255  13.259  < 2e-16 ***
## floors                  26.852      9.049   2.968  0.00322 **
## sqft_living15           18.392      4.699   3.914  0.00011 ***
## I(bedrooms * floors)    -2.477      1.924  -1.288  0.19872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.68 on 336 degrees of freedom
## Multiple R-squared:  0.649,  Adjusted R-squared:  0.6449
## F-statistic: 155.3 on 4 and 336 DF,  p-value: < 2.2e-16
```

Regarding to both model accuracy:

- Model4 has these following characteristics:
  - There are only 3 explanatory variables: sqft_living, floors +sqft_living15, but all 3 have low std.errors, large t-values and significant p-values (all the explanatory variables' p-value is smaller than 0.05)
  - For the parameter accuracy: RSE of the model is just a fraction higher than RSE of the model5. As well as for the Multiple R-squared, model5 has the variance in the response variable explained better only for a bit than model4's.
  - Multiple R-squared refers to 64.73% of the variation in the response variables is explained by the model, supporting adequate linear relationship

- Model5 has these following characteristics:
  - There are 4 explanatory variables to build up the fitted model; yet one variable unfitted into the model is the Interaction term of bedrooms*floor, in which its p-value is insignificant, suggesting that there is no evidence of a linear relationship between the two X and Y variable
  - F-statistic of model5 is on a lower scale than the F-statistic of model4;
  - Multiple R-squared refers to 64.49% of the variation in the response variables is explained by the model.

Hence, the better model of multiple regression is model4.

vii/ Fit a polynomial regression model to predict price using sqft_living of order 2 and test the model significance

```
model6 <- lm(price ~ poly(sqft_living, 2), data=kc)
pred_mod6 <- predict(model6)
head(pred_mod6) #price prediction with 6 newdata on the fitted polynomial model.
```

```
##         1         2         3         4         5         6
## 125.29892  78.51938  29.97793  71.56823 144.14218 106.99887
```

```
summary(model6)
```

```
##
## Call:
## lm(formula = price ~ poly(sqft_living, 2), data = kc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -125.447  -27.185   -7.469   21.488  162.383
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            117.406      2.585  45.427  < 2e-16 ***
## poly(sqft_living, 2)1 1162.754     47.726  24.363  < 2e-16 ***
## poly(sqft_living, 2)2  234.526     47.726   4.914 1.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.73 on 338 degrees of freedom
## Multiple R-squared:  0.6463, Adjusted R-squared:  0.6442
## F-statistic: 308.9 on 2 and 338 DF,  p-value: < 2.2e-16
```

We use hypothesis testing to test the model significance:

- H0: Bi = 0
- HA: Bi != 0

By looking at the summary of the model, we observe the following characteristics of the parameters:

- std.errors of both of polynomial order 1 and 2 of the explanatory variables are quite strongly deviated (polynomial 1= 47.726 and polynomial 2 std.error = 47.726) from the average estimate coefficient.
- Both of polynomial of order 1 and 2 has large t-value (t-value of polynomial of order 1 = 24.363 and t-value of polynomial of order 2 = 4.914).
- Both p-values corresponding to the coefficient parameter of poly 1 and poly 2 are 2e-16 and 1.39e-06, which are smaller than 0.05 and are significantly small enough to reject the null hypothesis at 5% of the significance level. There is evidence of a linear relationship among the polynomial order in the explanatory variables with the explanatory variable (price).

```
#anova table to check overall of model significance:
anova(model6)
```

```
## Analysis of Variance Table
##
## Response: price
##                        Df  Sum Sq Mean Sq F value    Pr(>F)
## poly(sqft_living, 2)    2 1407001  703500  308.85 < 2.2e-16 ***
## Residuals             338  769888    2278
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By assessing on the model's overall accuracy, we observe:

- Extreme high value of sum of square, which suggests the high variability from the mean value of the linear model
- Significant p-value (2.2e-16<0.05), meaning the model is accurately explained and strongly supporting strong linear relationship.
- Multiple R-squared refers 64.63% of the response variable's variation is moderately explained by the polynomial model, supporting moderate linear relationship

```
qf(0.95,2,338)
```

```
## [1] 3.022441
```

- Large F-statistic: 308 on 2 and 338 df, which is larger than 3.022441.

Question 02 - Classification

Use the same data set.

Create a new categorical variable "price_cat" by assigning value of "High" if price > median(price) else "Low". Remove the variable price from the dataset.

```
#QUESTION 2- CLASSIFCATION
kc <- read.csv("kc_house.csv")
kc_cncat <- kc
attach(kc_cncat)

kc_cncat$price_cat[kc_cncat$price > median(kc$price)] <- "High"
kc_cncat$price_cat[kc_cncat$price <= median(kc$price)] <- "Low"
kc_cncat <- subset(kc_cncat, select= -price)
str(kc_cncat$price_cat)
```

```
##  chr [1:341] "Low" "Low" "Low" "Low" "High" "Low" "Low" "High" "High" "Low" ...
```

Divide the dataset into two sets namely training set and test set by assigning 75% of the observations to training set and the rest of the observations to the test set. (Hint Use set.seed(100) for reproducible results)

```
set.seed(100)
training_1 <- sample(1:nrow(kc_cncat), 256) #75% of observations
head(training_1)
```

```
## [1] 202 112 206   4 311 326
```

- 75% of the data frame containing 341 observations is 256 observations.

1) Logistic Regression

i/ Construct Logistic regression model for "price_cat" in terms of all the other variables (Use training dataset)

```
model7<- glm(as.factor(price_cat) ~ . ,
             data=kc_cncat, subset=training_1,
             family= binomial) # using the training set

summary(model7)
```

```
##
## Call:
## glm(formula = as.factor(price_cat) ~ ., family = binomial, data = kc_cncat,
##     subset = training_1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48908  -0.44670  -0.01208   0.42126   3.03289
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.122e+01  1.646e+00   6.816 9.34e-12 ***
## id             8.079e-11  6.724e-11   1.202  0.22954
## bedrooms       1.798e-01  2.419e-01   0.743  0.45734
## bathrooms     -7.661e-01  3.852e-01  -1.989  0.04671 *
## sqft_living   -1.244e+00  4.036e-01  -3.081  0.00206 **
## sqft_lot       9.302e-03  2.571e-02   0.362  0.71749
## floors        -5.103e-01  4.208e-01  -1.213  0.22532
## waterfront    -3.308e+00  7.169e-01  -4.615 3.93e-06 ***
## sqft_living15 -2.517e+00  5.011e-01  -5.023 5.10e-07 ***
## sqft_lot15     7.469e-02  3.289e-02   2.271  0.02316 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 354.5  on 255  degrees of freedom
## Residual deviance: 166.9  on 246  degrees of freedom
## AIC: 186.9
##
## Number of Fisher Scoring iterations: 7
```

ii/ Comment on the significance of the parameter estimates

We use hypothesis testing on the parameter estimates:

- H0: Bi = 0. There is no evidence of logistic relationship between the response variable with the explanatory variables
- HA: Bi != 0. There is evidence of logistic relationship between the response variable with the explanatory variables

By testing the parameter significance, we can observe some unfitted variables within the parameter estimates with insignificant p-values, which suggests no logistic relationship between these variables with the response variable (price_cat), including:

- id: p-value of 0.708460, which is >0.05.
- bedrooms: p-value of 0.960256 , which is >0.05
- sqft_lot: p-value of 0.466618, which is >0.05
- floors: p-value of 0.221906, which is >0.05

Hence, we will not reject these variables at 5% level of significance, there is strong evidence not to reject the null hypothesis. There is no evidence supporting the logistic relationship between these variables with the response variable (price_cat).

Therefore, we can remove these insignificant variables to make the logistic model more fitting.

iii/ Improve the model based on the output in part i. (Hint Consider the significance of the parameter estimates)

Thus, the improved logistic model based on the output in part ii/ is:

```
model7<- glm(
  as.factor(price_cat)~ bathrooms + sqft_living + waterfront + sqft_living15 + sqft_lot15,
  data=kc_cncat,
  subset= training_1,
  family = binomial)

summary(model7)
```

```
##
## Call:
## glm(formula = as.factor(price_cat) ~ bathrooms + sqft_living +
##     waterfront + sqft_living15 + sqft_lot15, family = binomial,
##     data = kc_cncat, subset = training_1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.58138  -0.50965  -0.01417   0.47231   3.10416
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   11.14821    1.47034   7.582 3.40e-14 ***
## bathrooms     -0.84098    0.36082  -2.331 0.019767 *
## sqft_living   -1.13167    0.36953  -3.062 0.002195 **
## waterfront    -3.39682    0.69920  -4.858 1.18e-06 ***
## sqft_living15 -2.37112    0.47479  -4.994 5.91e-07 ***
## sqft_lot15     0.07444    0.02194   3.393 0.000691 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 354.50  on 255  degrees of freedom
## Residual deviance: 171.38  on 250  degrees of freedom
```

```
## AIC: 183.38
##
## Number of Fisher Scoring iterations: 7
```

iv/ Predict the outputs for the test dataset using the model in part iii and construct misclassification table

```
#prediction of the output of the test dataset using model in part iii
pred_prob <- predict(model7,
                     newdata = data.frame(kc_cncat[-training_1,]),
                     type = "response")
head(pred_prob)
```

```
##          6          9         10         21         22         23
## 0.54831740 0.04466833 0.89672068 0.43923928 0.37178822 0.66944889
```

```
#Construct the misclassifcation table:
pred_class <- rep(NA, nrow(kc_cncat[-training_1,]))
length(pred_class) # the 25% data left from the dataset
```

```
## [1] 85
```

```
pred_class[pred_prob >= 0.5] <- "High"
pred_class[pred_prob < 0.5] <- "Low"
MisClass <- table( "Predicted" = pred_class , "Actual" = kc_cncat$price_cat[-training_1])

MisClass # Misclassification table
```

```
##           Actual
## Predicted High Low
##      High    8  38
##      Low    29  10
```

v/ Calculate the misclassification rate (Use test dataset)

```
#The Misclassification rate using the test dataset:
sum(MisClass[1,2], MisClass[2,1])/sum(MisClass)
```

```
## [1] 0.7882353
```

This is a very high rate of misclassification for a test dataset.

2) Decision Tree

i/ Build a classification tree model for the training dataset

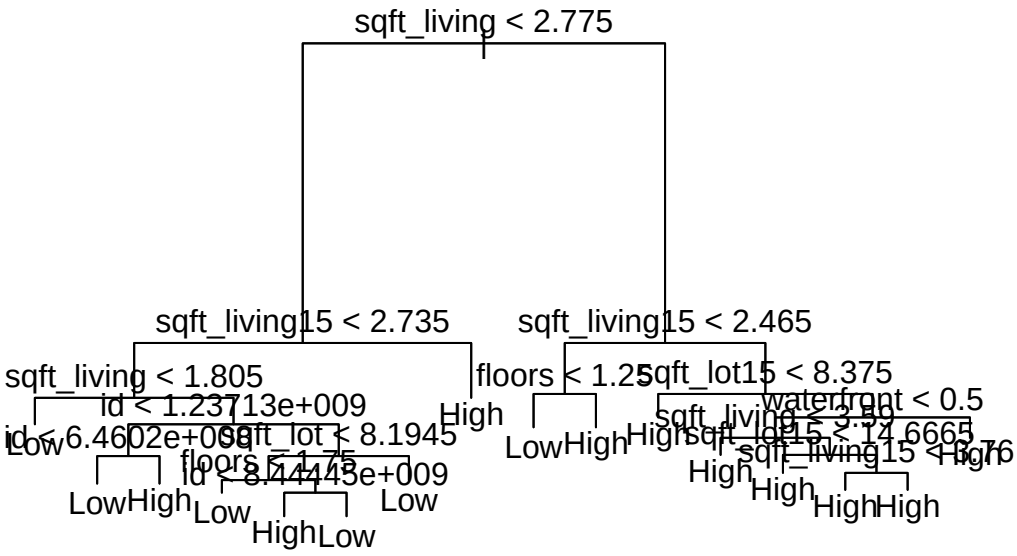
```
par(mfrow= c(1,1))
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.1.3
```

```
head(kc_cncat)
```

```
##           id bedrooms bathrooms sqft_living sqft_lot floors waterfront
## 1 7922800400        5      3.25        3.25   14.342      2          0
## 2 1516000055        3      2.25        2.15   21.235      1          0
## 3 2123039032        1      0.75        0.76   10.079      1          1
## 4 9297300045        3      2.00        1.97    4.166      2          0
## 5 1860600135        5      2.50        3.65    9.050      2          0
## 6 1560930070        4      3.50        2.84   40.139      1          0
##   sqft_living15 sqft_lot15 price_cat
## 1          2.96     11.044       Low
## 2          2.57     18.900       Low
## 3          1.23     14.267       Low
## 4          2.39      4.166       Low
## 5          2.88      5.400      High
## 6          3.18     36.852       Low
```

```
tree_model1 <- tree(as.factor(price_cat)~ . ,
                    data= kc_cncat,
                    subset= training_1) #tree model using training dataset
plot(tree_model1)
text(tree_model1, pretty = 0)
```



ii/ Use cross-validation and choose the best size for the tree part i

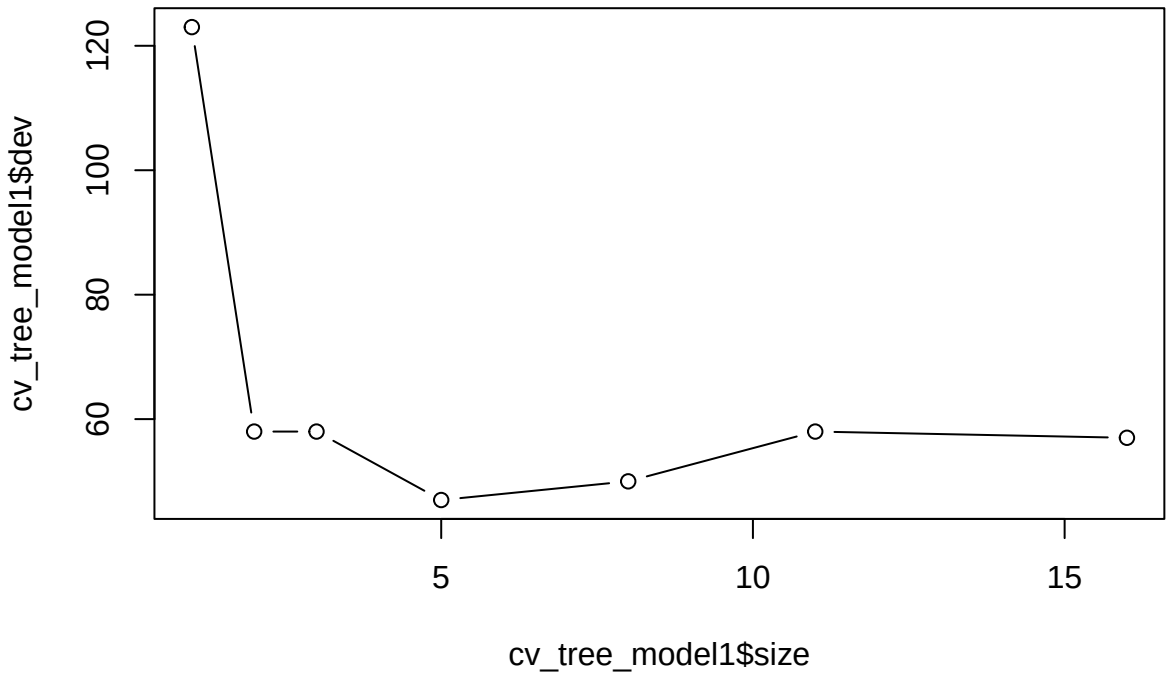
```
cv_tree_model1 <- cv.tree(tree_model1, FUN = prune.misclass)
names(cv_tree_model1)
```

```
## [1] "size"   "dev"    "k"      "method"
```

```
cv_tree_model1
```

```
## $size
## [1] 16 11  8  5  3  2  1
##
## $dev
## [1]  57  58  50  47  58  58 123
##
## $k
## [1]      -Inf  0.000000  1.000000  1.666667  4.500000  5.000000 75.000000
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```
plot(cv_tree_model1$size , cv_tree_model1$dev , type = "b")
```

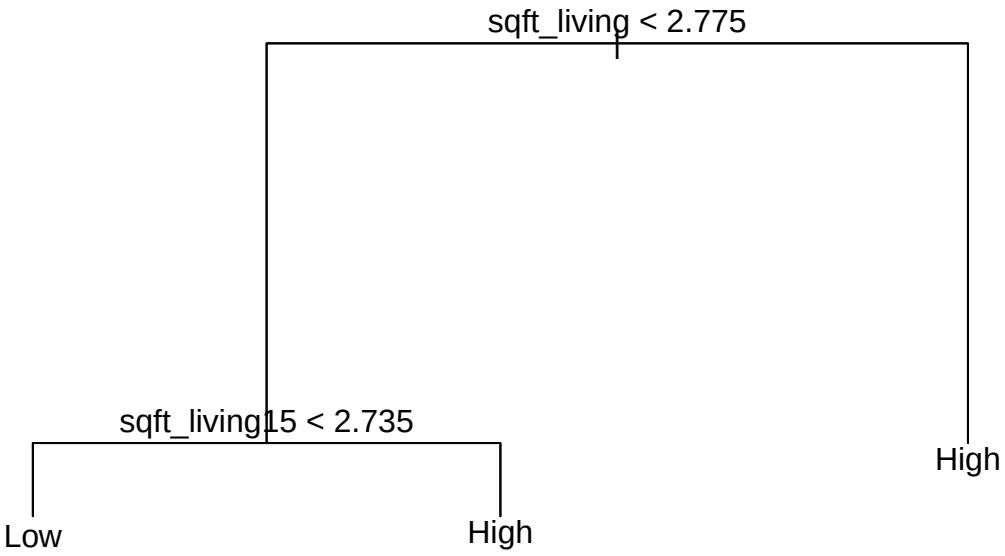


The graph demonstrates the estimated validation for different sizes of the tree model. And there is a sudden drop in between the size of 3 and 4 suggesting a huge improvement in the model size However, choosing beyond the size of 4 may lead to overfitting the model as we might try to fit every single observation into the model, which we would not want that but rather a simple generalised model for easier prediction and classification of data in the future.

Therefore, the best model represented, is the cv_tree_model with the size of 3 as it has the greatest explanation for the variance in the dev, and is more representative of the change in variables.

iii/ Build the tree with the best size (pruning) obtained in part ii

```
prune_tree_model1 <- prune.misclass(tree_model1, best = 3)
plot(prune_tree_model1)
text(prune_tree_model1, pretty = 0)
```



iv/ Predict the outputs for the test dataset using the model in part iii and construct misclassification table

```
pred_tree_model1 <- predict(prune_tree_model1,
                            newdata= data.frame(kc_cncat[-training_1,]),
                            type = "class")
head(pred_tree_model1)
```

```
## [1] High High Low  High High Low
## Levels: High Low
```

```
#misclass table:
table("predicted_test" = pred_tree_model1, "actual" = kc_cncat$price_cat[-training_1])
```

```
##               actual
## predicted_test High Low
##           High   34  17
##           Low     3  31
```

v/ Calculate the misclassification rate (Use test dataset)

```r
tab2 <- table("predicted_test" = pred_tree_model1,
              "actual" = kc_cncat$price_cat[-training_1])
Misclass_rate2 <- (tab2[1,2] + tab2[2,1])/sum(tab2)
Misclass_rate2
```

```
## [1] 0.2352941
```

```r
#lower misclassification rate
```

3) Compare the models in part 1 and part 2 and suggest the best model (Give reasons)

Based on the dataset given that we have trained the model on with methods such as fitting model using Logistic regression methods or using Decisions tree method to suggest the best model. There are reasons suggesting why model 2 is more applicable and better:

Models part 2 has lower Misclassfication rate than model part 1 ($0.2352941 < 0.7882353$), suggesting the overall model in part 2 is accurate enough to give out high possibility rate of true results. Meanwhile the model in part 1 using logistic regression to approach price prediction, is less predictable

However, Model part 2 which using the tree classification method, only has the sqft_living15 variable in its fitted tree model to predict the prices; while Model part 1 using logistic regression which fits multiple explanatory variable to predict the fittest response variable, which makes it more "reliable". Model part 1 is more accurate as though it has less corresponded dependent variables to predict the price.

Therefore, model part 2 using the tree classification method to predict the 25% price from the test dataset is suggested to be better than model part 1

The end