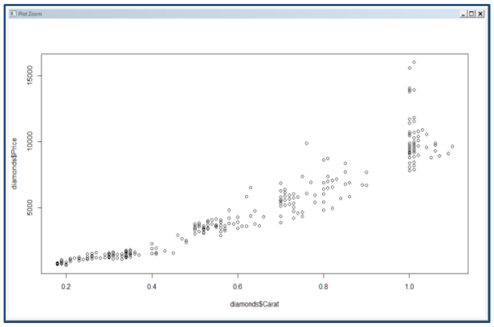
Data Analysis – EIT – Data Science

UPM

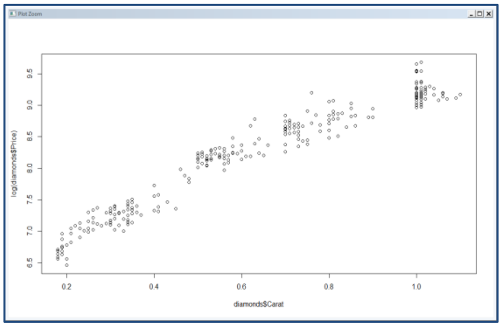
Assignment 2.1

Dina Mohamed

1. For linear regression models, we want to make sure that there is a linear relationship between the input and output variables. Taking the log to the price makes the relationship between price and carat looks more linear. This is our main objective for linear regression, As we can see in the below graphs, plotting the log price gives a better representation of the variables.



Price vs Carat



Log (Price) vs Carat

2. The parameters in the given model are represented as below:

```
Call:
lm(formula = log(Price) ~ Weight + ColourPurity + Clarity + Certifier,
    data = diamonds)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31236 -0.11520  0.01613  0.10833  0.36339

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.086094   0.041873 145.346  < 2e-16 ***
Weight         2.855013   0.036968  77.230  < 2e-16 ***
ColourPurityD  0.416557   0.041382  10.066  < 2e-16 ***
ColourPurityE  0.387047   0.030824  12.557  < 2e-16 ***
ColourPurityF  0.310198   0.027479  11.288  < 2e-16 ***
ColourPurityG  0.210207   0.028359   7.412 1.32e-12 ***
ColourPurityH  0.128681   0.028523   4.511 9.31e-06 ***
ClarityIF      0.298541   0.033303   8.964  < 2e-16 ***
ClarityVS1     0.096609   0.024919   3.877  0.00013 ***
ClarityVVS1    0.297835   0.028102  10.598  < 2e-16 ***
ClarityVVS2    0.201923   0.025344   7.967 3.56e-14 ***
CertifierHRD  -0.008856   0.020864  -0.424  0.67155
CertifierIGI  -0.182711   0.024952  -7.323 2.33e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1382 on 295 degrees of freedom
Multiple R-squared:  0.9723,     Adjusted R-squared:  0.9712
F-statistic: 863.6 on 12 and 295 DF, p-value: < 2.2e-16
```

Interpretation:

- As we can see in the above summary, the overall model has a Multiple R-squared of 0.9723, so the model is able to predict any Y using X with accuracy of 97%.

- Test of significance for the overall model is 2.2e-16, We can reject the null hypothesis which assumes that all the B_i is equal to 1. We can conclude that the variables are significant for the model.

- For the weight coefficient: Since the dependent variable is expressed in log, then we can say either:
  - In log scale: If you increase one unit of cartage , log(price) will increase by 2.855

  In Price scale: If you increase one unit of cartage the price of diamond is multiplied by a factor of exp (2.885) =17.37466

  For the color purity coefficient: Since the dependent variable is expressed in log and reference category is “I”, then we can say that the price of “ColourPurityD” is exp(0.41)=1.53 times the price of “ColourPurityI” , we can conclude that “ColourPurityD” is much better than the reference category

- For the color purity coefficient: Since the dependent variable is expressed in log and reference category is “I”, then we can say that the price of “ColourPurityD” is exp(0.41)=1.53 times the price of “ColourPurityI” , we can conclude that “ColourPurityD” is much better than the reference category

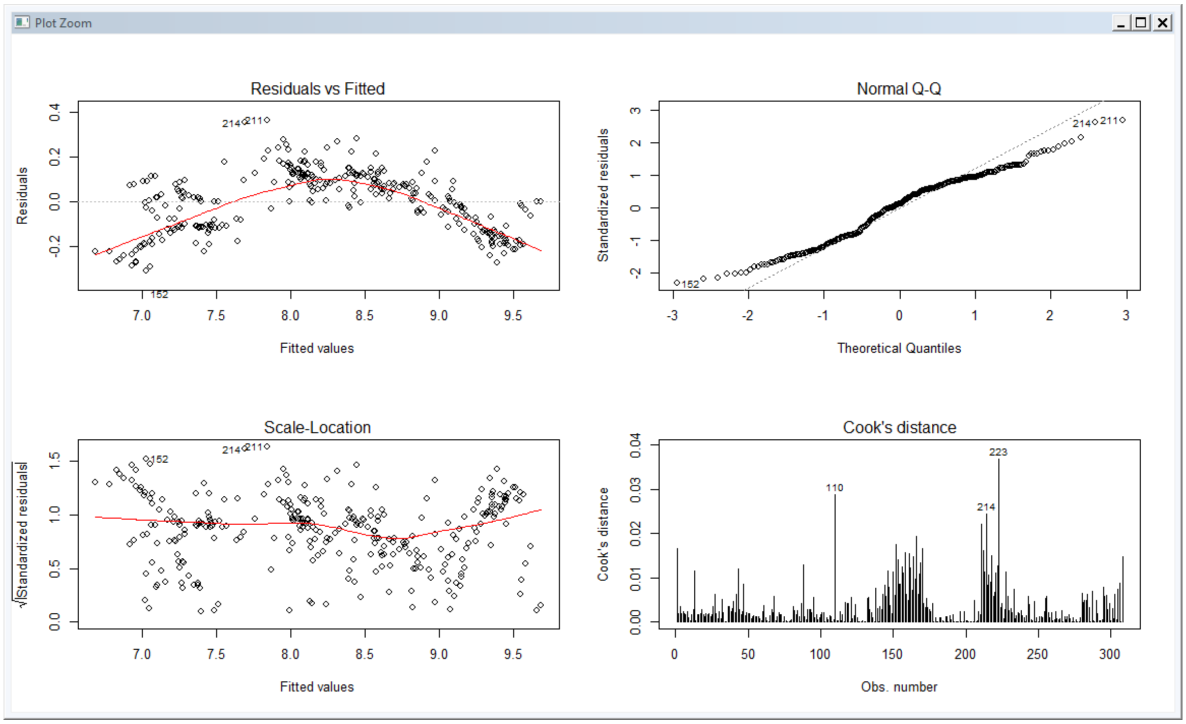
- For the certifiers coefficient, we can see that the HRD and GIA are better than IGI but with a very small difference

Plot Analysis:

In order to make some interpretation on the residuals, we started by plotting a graph using the plot function for the model.

From the graph we can see that:

1. The residuals don't behave nicely , there is a nonlinear relationship between the outcome and predictor (seem to be following a curve plot not a linear plot)
2. Residuals come from a normal distribution which is a good
3. In the scale location plot, we can see that the residuals are spread somehow equally along with the range of the predictor.
4. In Cooks distance plot, we can see that there are 3 influential values , which are : 110, 214 and 223



Statistical Analysis:

- ➢ We can test residuals dependency using "Durbin Watson test" , the p-value for the test is p-value (< 2.2e-16) , so we can reject the null hypothesis which means that the residuals are dependent and have correlation (not a good interpretation, residuals should be independent)

- ➢ Jarque-Bera test has been used to check for residuals normality and the p-value of the residuals is p-value = 0.1952 , which means accepting the null hypothesis , so residuals follow a normal distribution (good interpretation about the model)

- To check for variance equality , Breusch-Pagan test have been used , the p-value equals to 0.3507 , so we will accept the null hypothesis that means variances are constant (good interpretation about the model)

- We can see that there are 3 values that can be considered as outliers: 152,214,110

diamonds [152,]

Weight	ColourPurity	Clarity	Certifier	Price	Carat_Size	Carat_Size
0.18	F	VVS1	IGI	823	Small	Small

(Average price of diamonds having ColourPurity=F is 4786.402

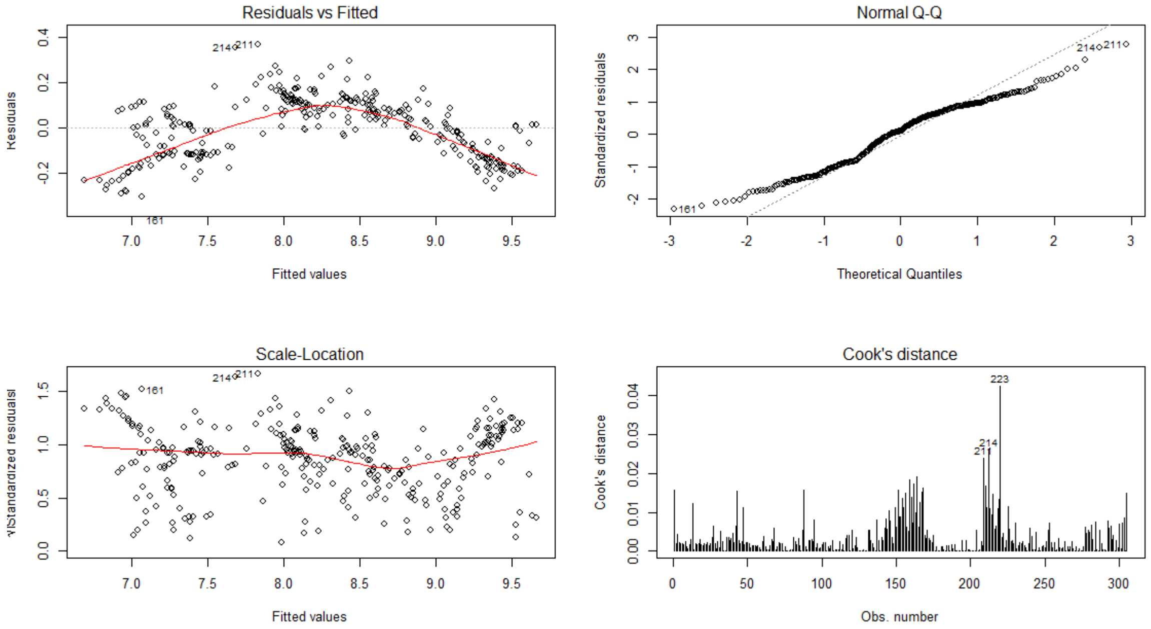
diamonds[214,]

Weight	ColourPurity	Clarity	Certifier	Price	Carat_Size	Carat_Size
0.52	I	IF	IGI	3095	Medium	Medium

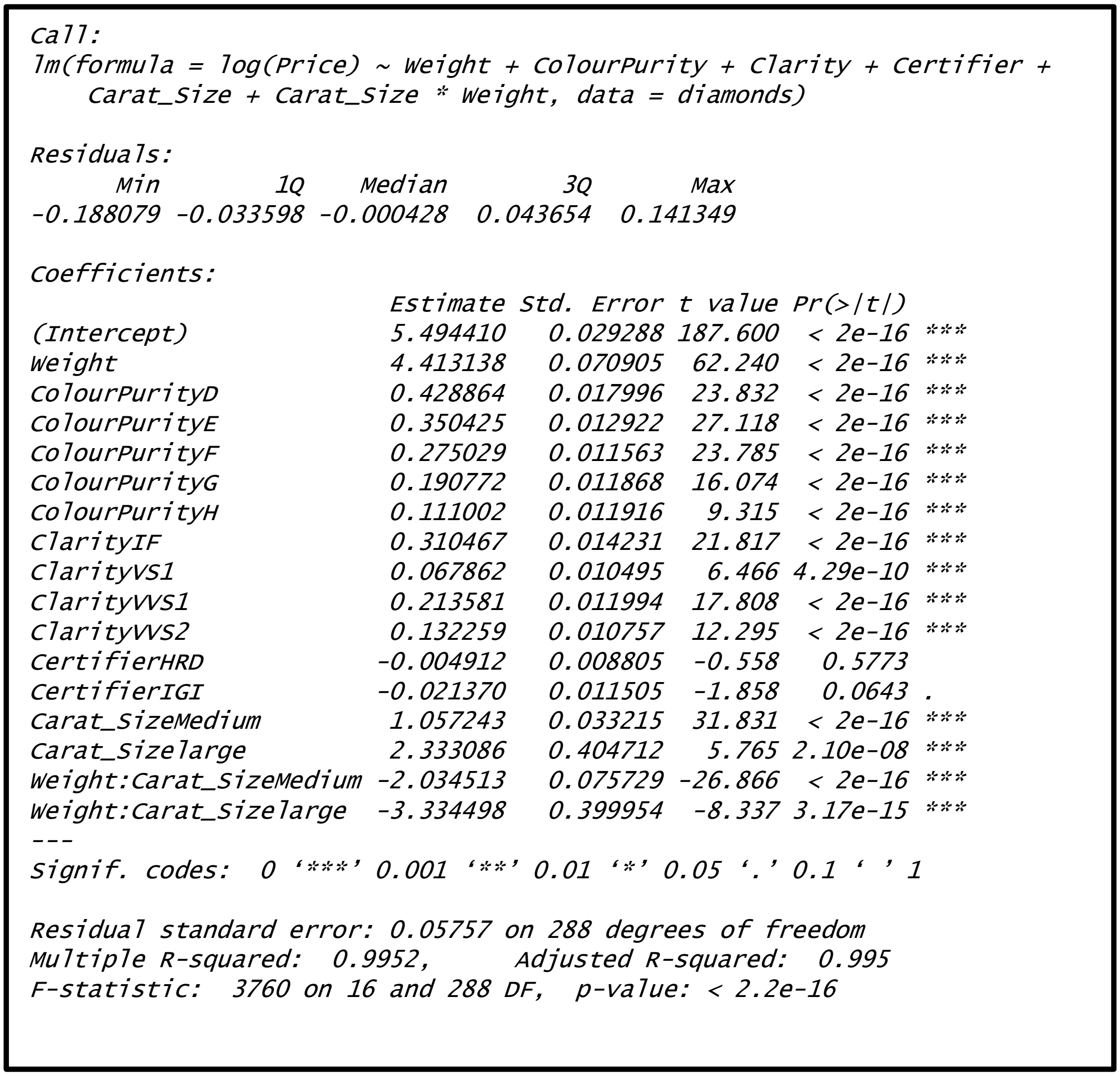
diamonds[110,]

Weight	ColourPurity	Clarity	Certifier	Price	Carat_Size	Carat_Size
0.76	D	IF	GIA	9885	Medium	Medium

After removing the above outliers from the model, there was no enhancement done and other outliers appeared, we can see that from the below



3. After adding the new categorical variable below are the answers to the questions
    a. The parameters are given as below:



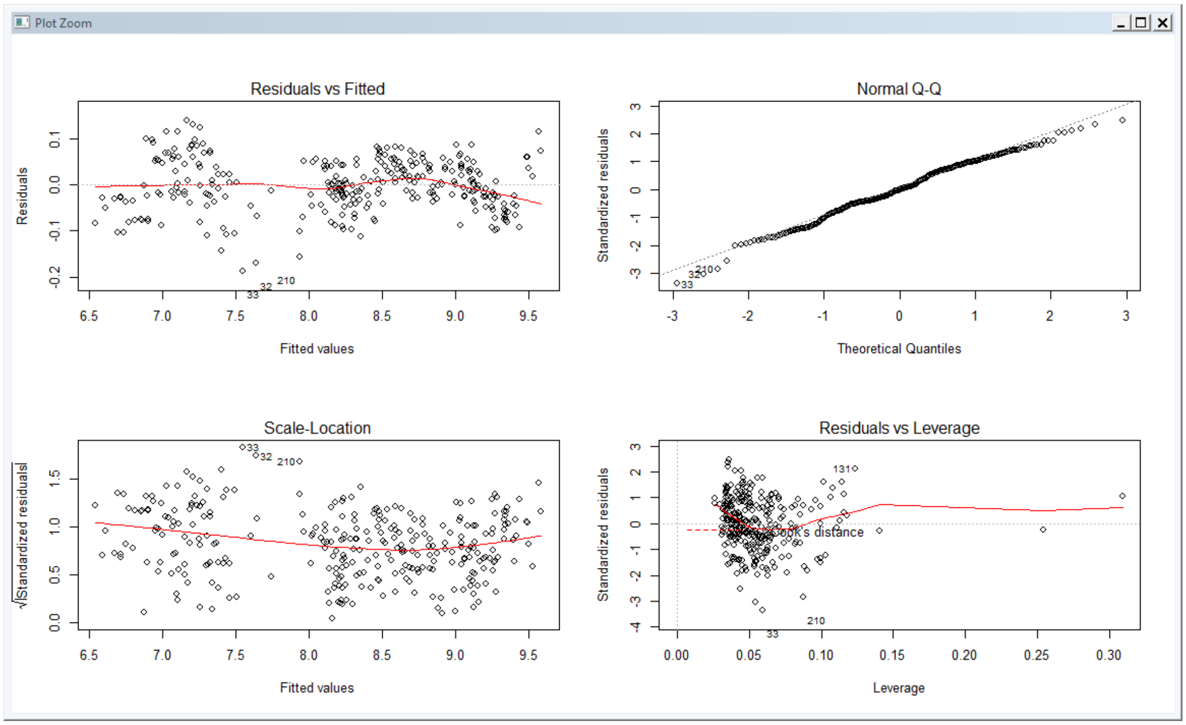
```
Call:
lm(formula = log(Price) ~ Weight + ColourPurity + Clarity + Certifier +
    Carat_Size + Carat_Size * Weight, data = diamonds)

Residuals:
      Min        1Q    Median        3Q       Max
-0.188079 -0.033598 -0.000428  0.043654  0.141349

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.494410   0.029288 187.600  < 2e-16 ***
Weight                   4.413138   0.070905  62.240  < 2e-16 ***
ColourPurityD            0.428864   0.017996  23.832  < 2e-16 ***
ColourPurityE            0.350425   0.012922  27.118  < 2e-16 ***
ColourPurityF            0.275029   0.011563  23.785  < 2e-16 ***
ColourPurityG            0.190772   0.011868  16.074  < 2e-16 ***
ColourPurityH            0.111002   0.011916   9.315  < 2e-16 ***
ClarityIF                0.310467   0.014231  21.817  < 2e-16 ***
ClarityVS1               0.067862   0.010495   6.466 4.29e-10 ***
ClarityVVS1              0.213581   0.011994  17.808  < 2e-16 ***
ClarityVVS2              0.132259   0.010757  12.295  < 2e-16 ***
CertifierHRD            -0.004912   0.008805  -0.558   0.5773
CertifierIGI            -0.021370   0.011505  -1.858   0.0643 .
Carat_SizeMedium         1.057243   0.033215  31.831  < 2e-16 ***
Carat_Sizelarge          2.333086   0.404712   5.765 2.10e-08 ***
Weight:Carat_SizeMedium -2.034513   0.075729 -26.866  < 2e-16 ***
Weight:Carat_Sizelarge  -3.334498   0.399954  -8.337 3.17e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05757 on 288 degrees of freedom
Multiple R-squared:  0.9952,	Adjusted R-squared:  0.995
F-statistic:  3760 on 16 and 288 DF,  p-value: < 2.2e-16
```

- The regression model is satisfactory as the overall test of significance for the model is 2.2e-16 and $R^2 = 0.9953$ (there is an obvious improvement in the model compared to model 1)

Standard Assumptions of linear regression check:

Plot Analysis:

In order to make some interpretation on the model, we started by plotting a graph using the plot function for the model, as in the below picture.



- Residuals vs Fitted: we can see that the residuals do follow somehow a linear relationship
- Normal Q-Q: Residuals come from a normal distribution which is a good
- In the scale location plot: we can see that the residuals are divided into 2 groups along with the range of the predictor and there is a gap between those 2 groups.
- Residuals vs Leverage: We can see that there are some outliers like: 32, 33 and 210 that are seen far away from the concentrated data on the left

Statistical Analysis:

- Durbin Watson test: The p-value < 2.2e-16, which means that the residuals have a correlation (not good interpretation for linear regression models)
- Jarque-Bera: The p-value = 0.162, which means accepting the null hypothesis, so residuals follow a normal distribution (good interpretation about the model)
- Breusch-Pagan: The p-value equals to 0.0008164, so we will reject the null hypothesis that means variances are not constant.

- We can interpret that the interaction term for “med*weight” is significant as p-value is less than 0.05 (5% level)

Since the dependent variable (price) is expressed in logs, then we can say that if all the variables are equal, the increase in price for a diamond of medium size is exp(-2.04)= 0.13 times the increase in the price of a small diamond.
This means that the price of diamond for small ones increases faster than medium ones, when you increase the carat by one

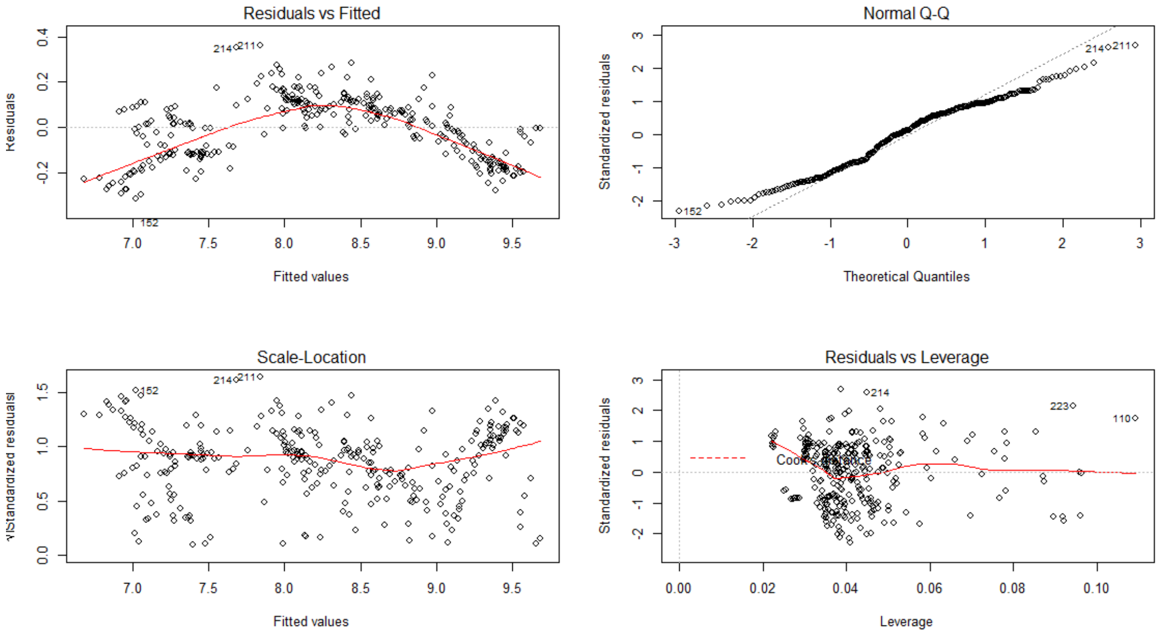
- Color Purity is more valued
  - Fit D value = 9.342298
    Fit I value = 8.906038
    Fit E value = 9.25695

- Average price distance between D and I is: 0.436
  Average price distance between D and E is: 0.085

- Fit GIA value = 9.342298
  Fit IGI value = 9.31861
  Fit HRD value = 9.336692

Prices do not differ a lot among different certifiers

b. After adding the square of the carat weights we plotted the below graph for the new model (model3):



4. I would prefer the first remedial
   In both models we have a good R^2 and a good p-value, the main difference in my point of view is:
   - The linearity of residuals , model2 is more linear than model3
   - Normality , model2 behaves better compared to model3
   - Variance equality, model 3 is better than model2

	Model2	Model3
R^2	0.9952	0.9723
Durbin-Watson test p-value (linearity)	< 2.2e-16	NA
Jarque Bera Test p-value	0.1723	0.01775
studentized Breusch-Pagan test p-value	0.0007143	7.195e-06