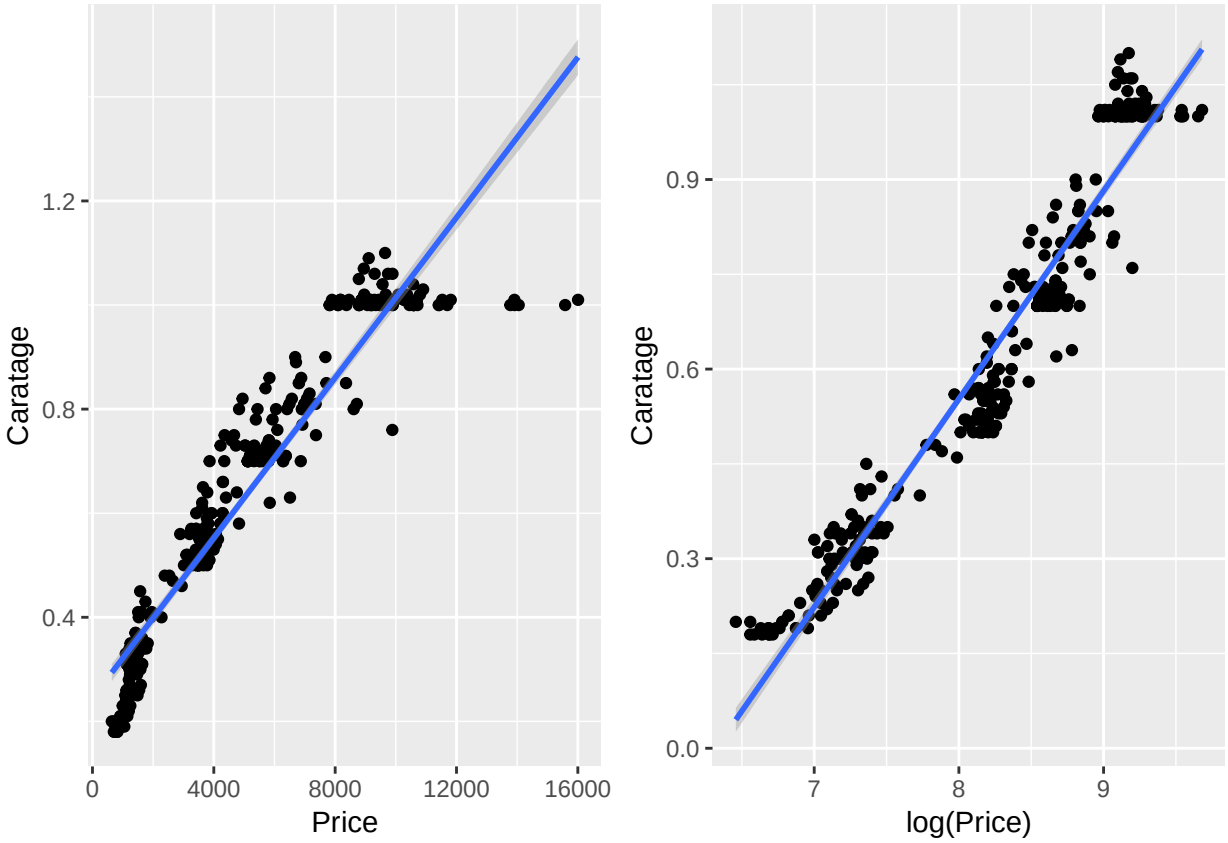
Homework 2.1 Report

Monika Wysoczanska, 180817

Manuel Barbas, 180832

Diogo Oliveira, 180832

Question 1




```
## [1] 0.9447266
```

```
## [1] 0.9672327
```

Looking at the plots above, we can observe the logaritmic transformation 'normalizes' the realtion between the two variables, meaning it seems to be more linear. The correlation between log_price and Caratage is also positively higher than between Price and Caratage. Since our goal is to deploy a Linear Regression Model, this is the transformation we would apply.

QUestion 2

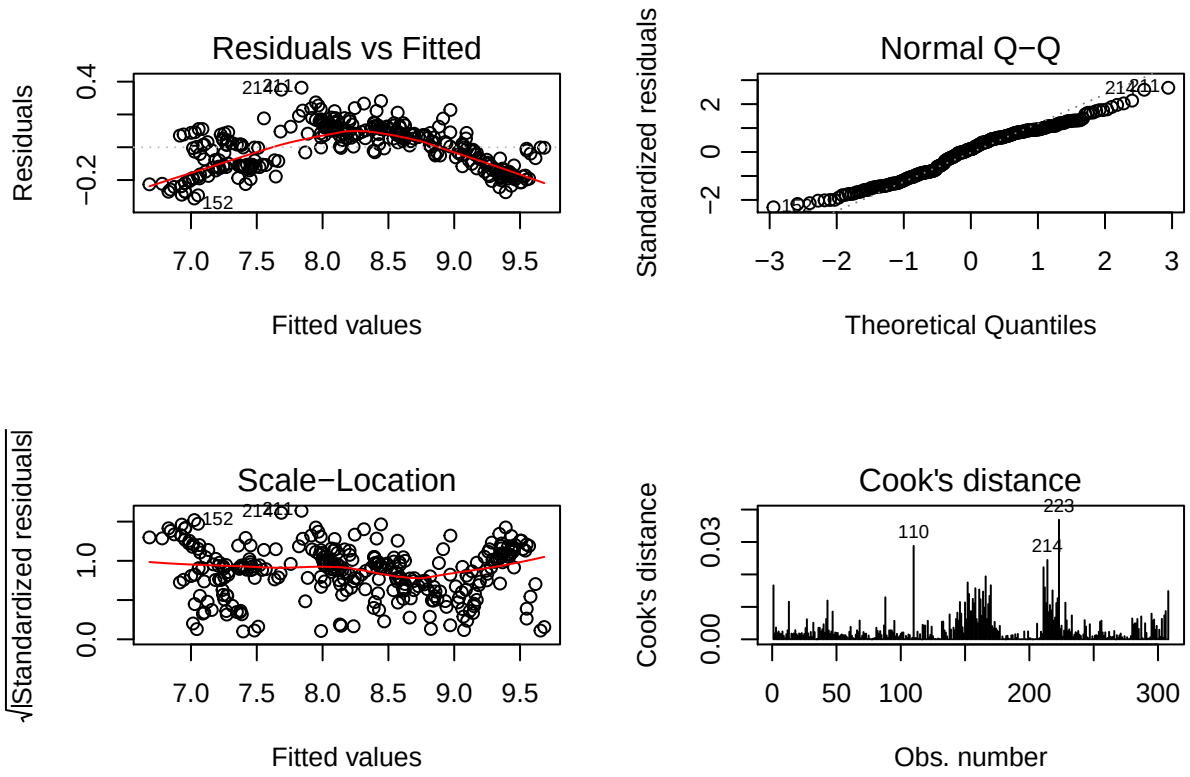
The summary of the obtained model is presented below.

```
##
```

```
## Call:
## lm(formula = log_price ~ Caratage + Purity + Clarity + Certificate,
##     data = db)
##
## 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)     6.077239   0.048091 126.369  < 2e-16 ***
## Caratage        2.855013   0.036968  77.230  < 2e-16 ***
## PurityD         0.416557   0.041382  10.066  < 2e-16 ***
## PurityE         0.387047   0.030824  12.557  < 2e-16 ***
## PurityF         0.310198   0.027479  11.288  < 2e-16 ***
## PurityG         0.210207   0.028359   7.412 1.32e-12 ***
## PurityH         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 ***
## CertificateGIA  0.008856   0.020864   0.424  0.67155
## CertificateIGI -0.173855   0.028673  -6.063 4.07e-09 ***
## ---
## 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
```

It can be seen that p-value of the F-statistic is $<$ 2.2e-16, which is highly significant - at least one of the predictor variables is significantly related to the outcome variable. Looking deeper, we can observe that there is indeed a significant association between all of the explanatory variables and log_price, besides 'CertificateGIA'. \

Leaving only our reference variables, the cost of the diamond is 436 Singapore Dollars. Every one unit of caratage increase results in 17 dollars total price increase. One level higher in Purity, which is 'PurityH' increases the total price of 14%, leaving the rest of variables the same. In comparison, the diamond having the highest Purity rank results in almost 52% price increase. The interesting observation on Clarity of a given stone, is the difference between the percentage of price increase of VVS1 and Internal Flawless (which is the highest possible) is only around 0.1 point percent. We can also conclude that Certificate IGI means less than our base HRD Certificate in terms of price as it results in almost 16% decrease in total price, leaving the rest variables the same. ### Model plots



The residuals behaviour is not constant, in fact it resembles a bit of a parabole - quadratic function. When it comes to outlier analysis we detected 3 major ones (by Cook's distance) which are 110, 214, 223 and also 211 appearing on each one of the plots. We conduct Bonferonni test for outlier detection.

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 211 2.710409          0.0071149           NA
```

The test revealed that the actual outlier is example 211, so we :

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 211 2.710409          0.0071149           NA
```

	Caratage	Purity	Clarity	Certificate	Price	log_price
211	0.5	G	IF	IGI	3652	8.20303

Residuals normality

We also check residuals normality by applying Jarque Bera Test.

```
##
##  Jarque Bera Test
##
## data:  lm1$residuals
## X-squared = 8.0626, df = 2, p-value = 0.01775
```

The p-value being <0.05 for this normality test makes us reject the null hypothesis and state that residuals

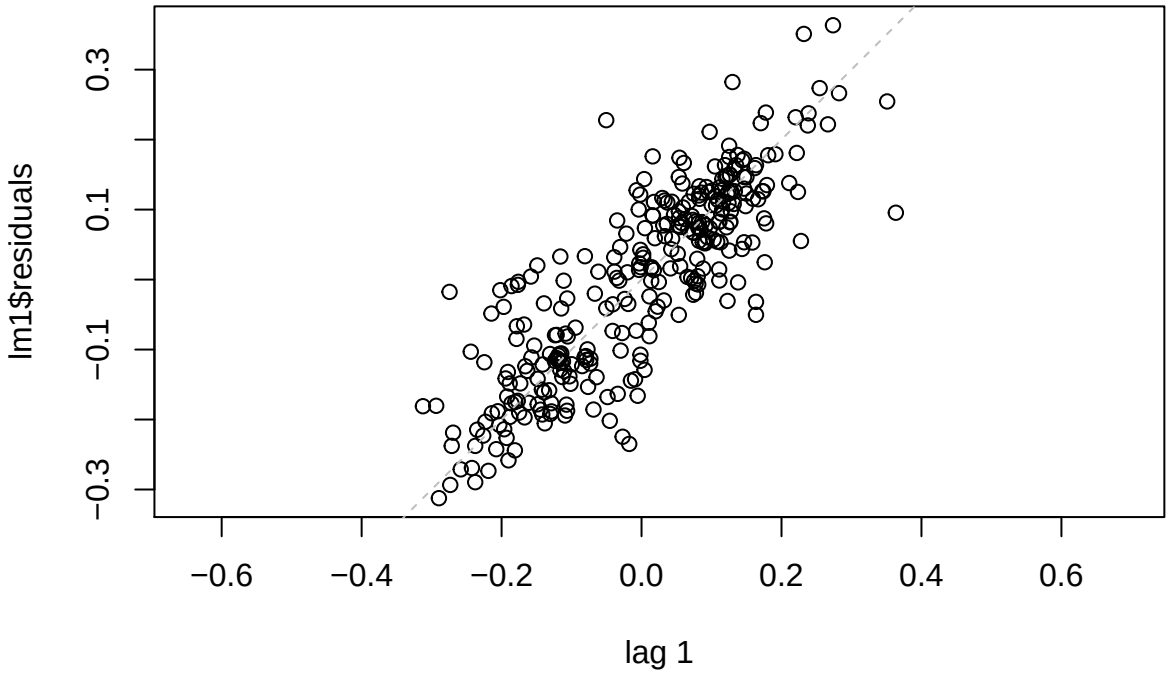
are not normally distributed.

Constant variance

```
##
##  studentized Breusch-Pagan test
##
## data:  lm1
## BP = 47.223, df = 12, p-value = 4.265e-06
```

The test on constant variance of the residuals results in the fail of homogenity hypothesis and leaves us with the conclusion that the variance is not constant.

Independence of the residuals




```
##
##  Durbin-Watson test
##
## data:  lm1
## DW = 0.31422, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

We conducted a Durbin-Watson test for residuals' autocorrelation and rejected the null hypothesis, leaving conclusion that it is greater than 0.

Question 3

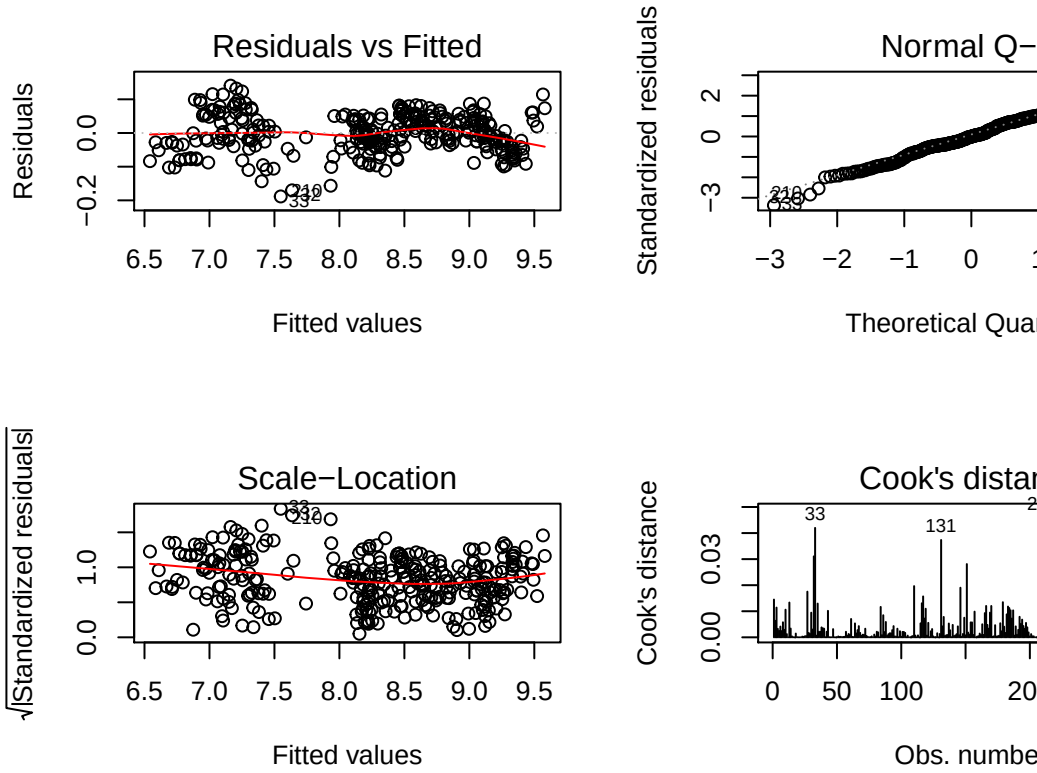
We create a new variable based on Caratage and assign 'small' as a reference level for the next model.

Caratage	Purity	Clarity	Certificate	Price	log_price	Caratage_cat
0.30	D	VS2	GIA	1302	7.171657	small
0.30	E	VS1	GIA	1510	7.319865	small
0.30	G	VVS1	GIA	1510	7.319865	small
0.30	G	VS1	GIA	1260	7.138867	small
0.31	D	VS1	GIA	1641	7.403061	small
0.31	E	VS1	GIA	1555	7.349231	small

Now we feed the model with our new variable as well as with the interaction term between this new variable and caratage.

```
##
## Call:
## lm(formula = log_price ~ Caratage + Purity + Clarity + Certificate +
##     Caratage_cat + Caratage:Caratage_cat, data = db)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.188358 -0.031815 -0.000249  0.043143  0.140535
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.483149   0.029919 183.268  < 2e-16 ***
## Caratage                       4.427061   0.069811  63.415  < 2e-16 ***
## PurityD                        0.436261   0.017465  24.979  < 2e-16 ***
## PurityE                        0.350912   0.012927  27.146  < 2e-16 ***
## PurityF                        0.275010   0.011535  23.841  < 2e-16 ***
## PurityG                        0.191449   0.011869  16.131  < 2e-16 ***
## PurityH                        0.111067   0.011923   9.316  < 2e-16 ***
## ClarityIF                      0.315793   0.013935  22.662  < 2e-16 ***
## ClarityVS1                     0.067530   0.010498   6.432 5.15e-10 ***
## ClarityVVS1                    0.213448   0.011932  17.889  < 2e-16 ***
## ClarityVVS2                    0.132373   0.010756  12.307  < 2e-16 ***
## CertificateGIA                 0.005606   0.008794   0.637    0.524
## CertificateIGI                -0.018082   0.012684  -1.426    0.155
## Caratage_catmedium             1.062001   0.032653  32.523  < 2e-16 ***
## Caratage_catlarge              2.340691   0.404861   5.781 1.91e-08 ***
## Caratage:Caratage_catmedium -2.047162   0.074556 -27.458  < 2e-16 ***
## Caratage:Caratage_catlarge  -3.350469   0.399880  -8.379 2.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0576 on 291 degrees of freedom
## Multiple R-squared:  0.9953, Adjusted R-squared:  0.995
## F-statistic:  3816 on 16 and 291 DF,  p-value: < 2.2e-16
```

For this particular model there is no significant difference between all of the certificates in terms of diamond



price. Let's take a look at residuals.
Judging only by plots, we assume that residuals still don't behave the way we wanted them to behave, meaning they probably fail all of the tests for assumptions of linear regression. We want to make sure about that.

```
##
##  Jarque Bera Test
##
## data:  lm2$residuals
## X-squared = 3.517, df = 2, p-value = 0.1723
```

Our new model actually passes the test for residuals normality with the p-value of Jarque Bera Test above 0.05.

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 33 -3.432994         0.00068392      0.21065
```

We detected new outlier for this model. We also check the constant variance assumption and independence of residuals.

```
##
##  studentized Breusch-Pagan test
##
## data:  lm2
## BP = 40.256, df = 16, p-value = 0.0007143
```

```
##
##  Durbin-Watson test
##
```

```
## data:  lm2
## DW = 0.96989, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

As we assumed, residuals of our new improved model fail the constant variance test and homogenity tests.

Interpretation of medcar

Leaving the rest the same, having a diamond from medium caratage cluster the price rises 189%, but then for each carat unit price decreases about 87% comparing to our reference 'small' cluster. In case of 'large' cluster, initially price increases almost 940% but for each caratage unit price decreases 96% comparing to 'small' cluster. We conclude that each caratage unit increase is highly valued for diamonds only up to 0.5 caratage ('small' cluster). By includin cluster variable we definitely introduced some kind of bias, which makes the model harder to interpret.

Clarity vs Purity

At the first glance it seems like Purity is higher valued than Clarity having the model's coefficient generally higher. Nevertheless, we compute the mean of coefficients and assure our observation having 0.27 for Clarity vs 0.18 of Purity average increase in log price (leaving the rest variables the same).

Average price difference between grade D and higher

Setting 'D' grade as our reference we observe on average diamond graded 'I' price is 35% lower than of those the highest graded. When it comes to 'E' grade it's on average 8% lower than 'D' grade (leaving the rest the same).

Price differences amongst Certificates

The significance t-test revealed that particular certificats do not impact our response variable. Moreover, having certificate 'HDR' as our reference we observe slight differences, such as 0.6% increase in price when particualar diamond is certified with 'GIA' rather than 'HDR', and decrease of around 1.8% for 'IGI', leaving us with the conclusion, that there are no significant differences amongst Certificates.

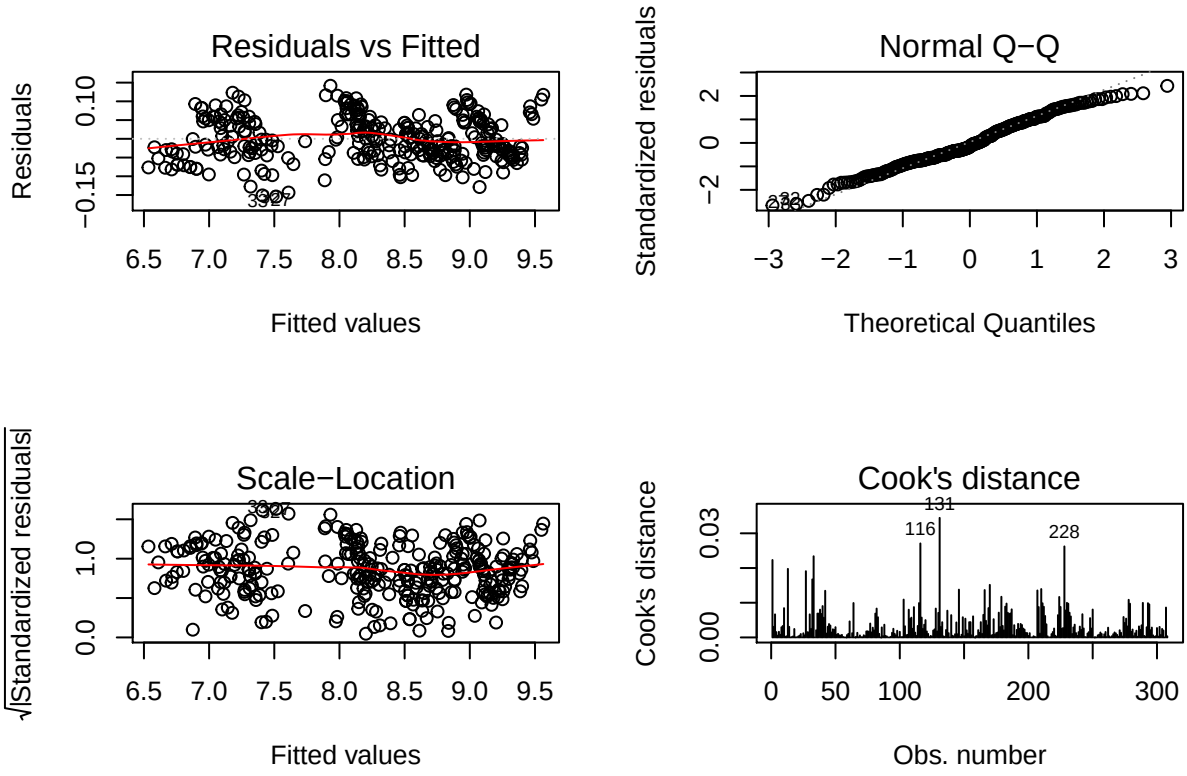
3b - Including squared carat

```
##
## Call:
## lm(formula = log_price ~ Caratage + I(Caratage^2) + Purity +
##     Clarity + Certificate, data = db)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15411 -0.04120 -0.00911  0.04543  0.14158
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.748945   0.030874 186.209  < 2e-16 ***
## Caratage        5.670616   0.079284  71.523  < 2e-16 ***
## I(Caratage^2)  -2.102922   0.058022 -36.243  < 2e-16 ***
## PurityI        -0.442606   0.017742 -24.947  < 2e-16 ***
## PurityE        -0.079247   0.017387  -4.558 7.59e-06 ***
## PurityF        -0.155991   0.016328  -9.554  < 2e-16 ***
## PurityG        -0.245033   0.016734 -14.643  < 2e-16 ***
## PurityH        -0.339098   0.016969 -19.983  < 2e-16 ***
```

```
## ClarityIF       0.320183   0.014279  22.424  < 2e-16 ***
## ClarityVS1      0.075713   0.010690   7.083 1.05e-11 ***
## ClarityVVS1     0.226174   0.012199  18.540  < 2e-16 ***
## ClarityVVS2     0.143481   0.010976  13.072  < 2e-16 ***
## CertificateGIA  0.006223   0.008938   0.696    0.487
## CertificateIGI -0.019190   0.013003  -1.476    0.141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0592 on 294 degrees of freedom
## Multiple R-squared:  0.9949, Adjusted R-squared:  0.9947
## F-statistic:  4445 on 13 and 294 DF,  p-value: < 2.2e-16
```

Our new variable is significant for the model. We investigate if it meets the linear model assumptions.

Model plots



Statistical tests

```
##
##  Jarque Bera Test
##
## data:  lm3$residuals
## X-squared = 3.8901, df = 2, p-value = 0.143
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  lm3
## BP = 16.094, df = 13, p-value = 0.2441
```

```
##
##  Durbin-Watson test
##
## data:  lm3
## DW = 0.98039, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

Looking at the plots, we cannot conclude much, so we get straight to statistical tests. It seems like residuals are normally distributed and have constant variance. On the other hand, the model still fails the residual homogenity hypothesis.

4 Conclusion

We definitely prefer the second remedial action as it results in meeting two linear model assumptions (residuals normality and constance in variance). In term of intepretability, it's true that square of a variable makes the model hard to intepret, but in our opinion this approach outperforms the bias introduced by artificially created clusters - in this approach we could also have some interpretability difficulties especially with the values very close to 'breaks'.