Tasks (*90pt*)

Please download the data set `diamonds.RData` that is provided with the quiz and that contains the prices and other attributes of almost 54,000 diamonds.

Here is a description of variables contained in the data set: `price`, price in US dollars; `carat`, weight of the diamond; `cut`, quality of the cut (Fair, Good, Very Good, Premium, Ideal); `color`, diamond colour, from D (best) to J (worst); `clarity`, a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)); `x`, length in mm; `y`, width in mm; `z`, depth in mm; `depth`, total depth.

You will build linear models whose the response variable is `price`. For this, first you need to load the data set using the `load` command, and then for the command `lm`, you can specify `data = diamonds`. Please make sure each predictor has the correct type such as being categorical or not; otherwise your model may be insensible. Note that you need to install R library `car` in order to use the command `vif` to obtain variance inflation factor.

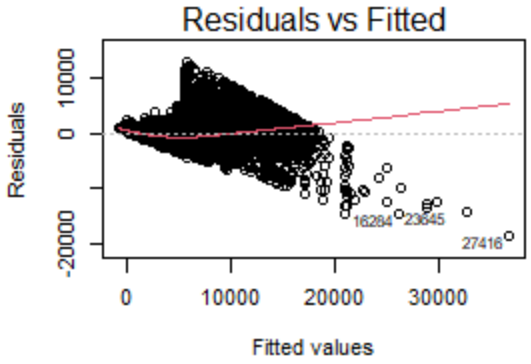
```r
## auto printing may break plots ##
par(mfrow=c(3,3))

## require packages ##
library(MASS)
library(ISLR)
library(car)
library(ggplot2)

## load the data ##
str(diamonds)
names(diamonds)

```

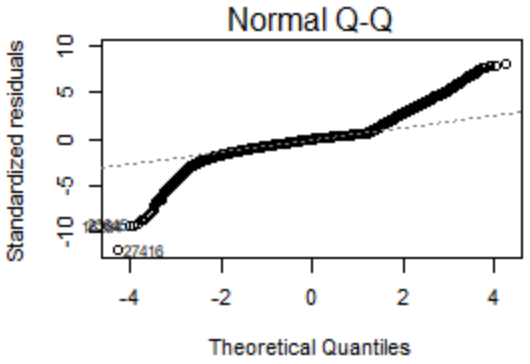
```
> ## auto printing may break plots ##
> par(mfrow=c(3,3))
>
> ## require packages ##
> library(MASS)
> library(ISLR)
> library(car)
Loading required package: carData
> library(ggplot2)
>
> ## load the data ##
> str(diamonds)
tibble [53,940 x 10] (S3: tbl_df/tbl/data.frame)
 $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
 $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
 $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
 $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
 $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
> names(diamonds)
 [1] "carat"   "cut"     "color"   "clarity" "depth"   "table"   "price"   "x"       "y"       "z"
>
```

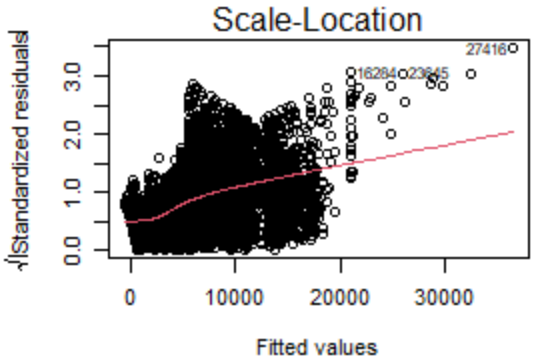
(1) Build a simple linear regression model with `carat` as the predictor, provide a summary of the fitted model (i.e., the estimated model), and show diagnostic plots for the model. (*20pt*)

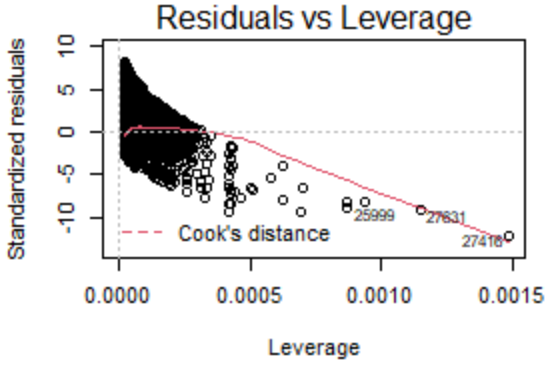
```r
## Simple Linear Regression Model with carat predictor ##
model1<-lm(price~carat, data=diamonds)

summary(model1)
plot(model1)

```

```
> ## Simple Linear Regression Model with carat predictor ##
> model1<-lm(price~carat, data=diamonds)
>
> summary(model1)

Call:
lm(formula = price ~ carat, data = diamonds)

Residuals:
     Min       1Q   Median       3Q      Max
-18585.3   -804.8    -18.9    537.4  12731.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2256.36      13.06  -172.8   <2e-16 ***
carat        7756.43      14.07   551.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1549 on 53938 degrees of freedom
Multiple R-squared:  0.8493,    Adjusted R-squared:  0.8493
F-statistic: 3.041e+05 on 1 and 53938 DF,  p-value: < 2.2e-16

> plot(model1)
>
```


Residuals vs Fitted
Residuals
Fitted values
16284
23645
27416


Normal Q-Q
Standardized residuals
Theoretical Quantiles
27416

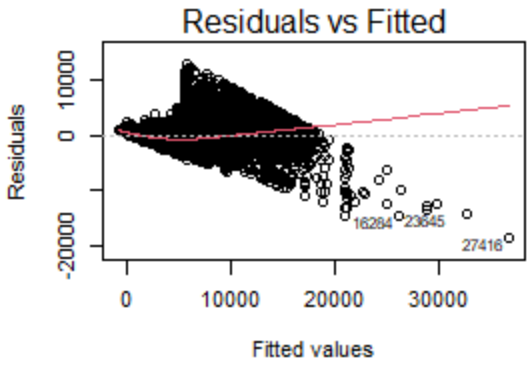

Scale-Location
√|Standardized residuals|
Fitted values
16284
23645
27416

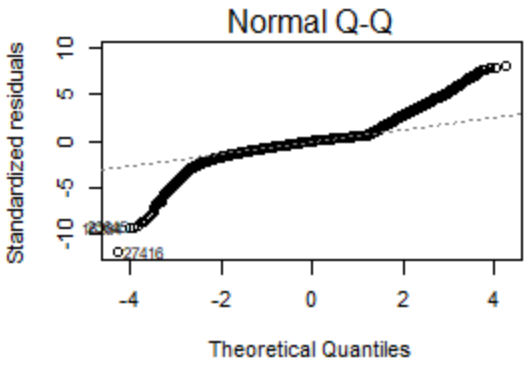

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
25999
27631
27416

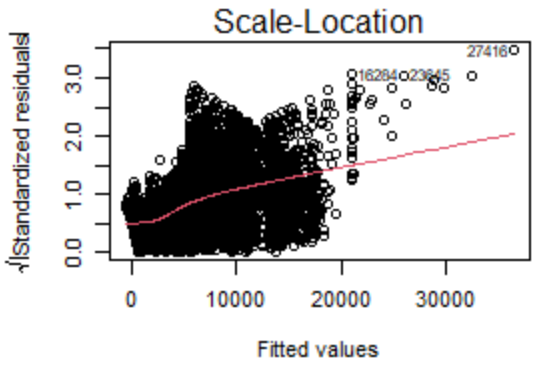
(2) Build a multiple linear regression model with `carat`, `cut`, `x`, `y`, `z` and `depth` as the predictors, provide a summary of the fitted model, provide the variance inflation factor for each predictor, and show diagnostic plots for the model. (*30pt*)

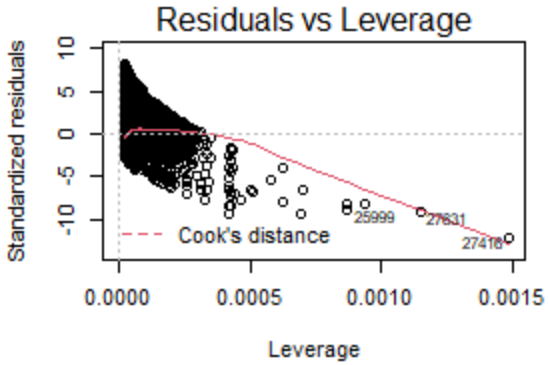
```r
## Multiple Linear Regression Model with carat, cut, x, y, z, and depth predictors ##
model2<-lm(price~carat+cut+x+y+z+depth, data = diamonds)

summary(model2)
vif(model2)
plot(model2)

```

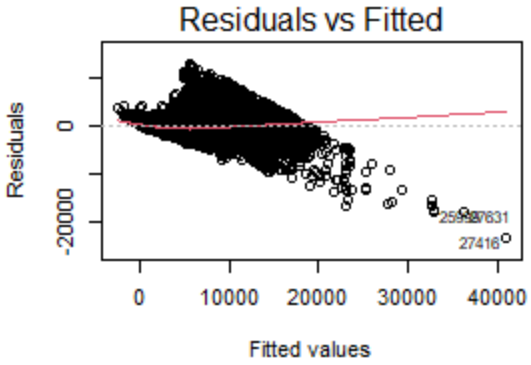
```
> ## Multiple Linear Regression Model with carat, cut, x, y, z, and depth predictors ##
> model2<-lm(price~carat+cut+x+y+z+depth, data = diamonds)
>
> summary(model2)

Call:
lm(formula = price ~ carat + cut + x + y + z + depth, data = diamonds)

Residuals:
     Min       1Q   Median       3Q      Max
-22938.7   -620.4    -60.0    344.2  12966.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8343.256    377.675  22.091  < 2e-16 ***
carat       10618.091     62.571 169.697  < 2e-16 ***
cut.L        1107.544     27.035  40.967  < 2e-16 ***
cut.Q        -449.603     23.334 -19.268  < 2e-16 ***
cut.C         353.852     19.865  17.813  < 2e-16 ***
cut^4          73.151     15.995   4.573 4.81e-06 ***
x           -1260.702     42.804 -29.453  < 2e-16 ***
y              47.990     25.350   1.893   0.0584 .
z              38.522     43.910   0.877   0.3803
depth        -103.420      5.527 -18.711  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1482 on 53930 degrees of freedom
Multiple R-squared:  0.862,     Adjusted R-squared:  0.8619
F-statistic: 3.742e+04 on 9 and 53930 DF,  p-value: < 2.2e-16

> vif(model2)
          GVIF Df GVIF^(1/(2*Df))
carat 21.59526  1        4.647070
cut    1.19662  4        1.022691
x     56.59828  1        7.523183
y     20.57979  1        4.536495
z     23.57263  1        4.855165
depth  1.53920  1        1.240645
> plot(model2)
>
```

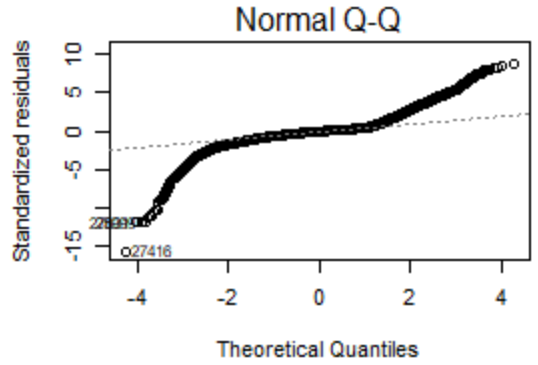

Residuals vs Fitted
Residuals
Fitted values
16284
23645
27416

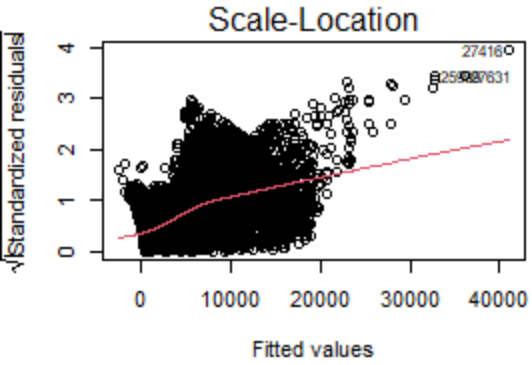

Normal Q-Q
Standardized residuals
Theoretical Quantiles
27416

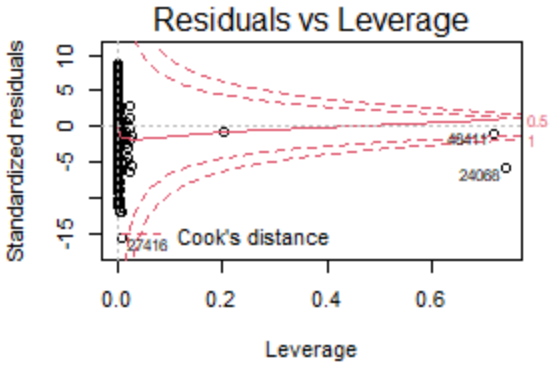

Scale-Location
√|Standardized residuals|
Fitted values
27416
16284
23645


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
25999
27631
27416


Residuals vs Fitted
Residuals
Fitted values
27631
27416


Normal Q-Q
Standardized residuals
Theoretical Quantiles
27416

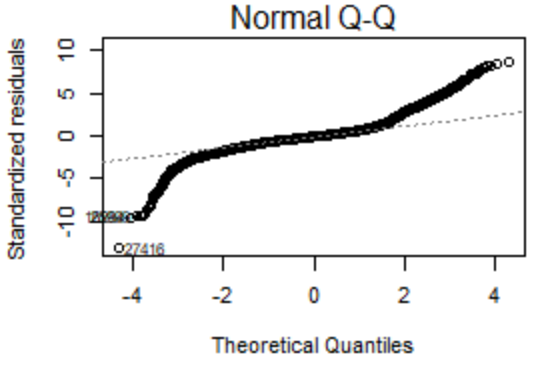

Scale-Location
√|Standardized residuals|
Fitted values
27416
27631

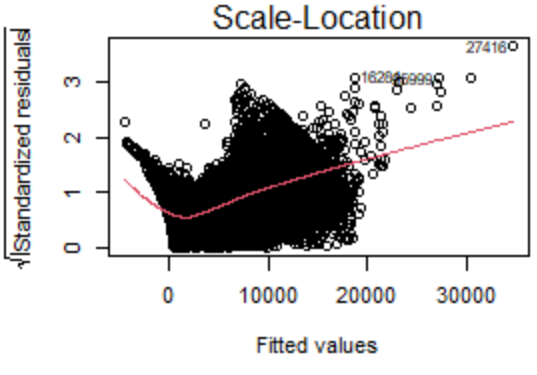

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
1
24068
27416

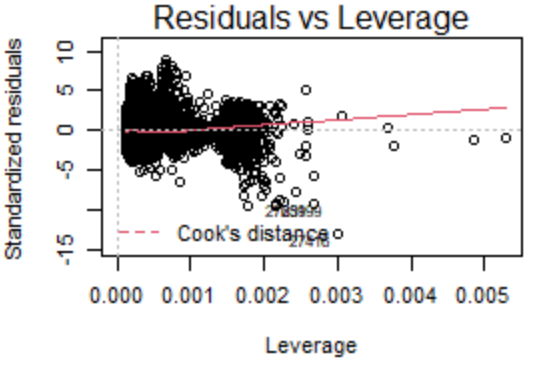
(3) Build a multiple linear regression model with `carat`, `cut`, `depth`, `clarity`, the interaction between `cut` and `clarity`, the interaction bewteen `depth` and `clarity` as the predictors, provide a summary of the fitted model, provide the variance inflation factor for each predictor, and show diagnostic plots for the model. (*40pt*)

```r
## Multiple Linear Regression Model with carat, cut, depth, clarity, the interaction between cut and clarity, the interaction bewteen depth and clarity as the predictors, provide a summary of the fitted model, provide the variance inflation factor for each predictor, and show diagnostic plots for the model ##
model3<-lm(price~carat+cut+depth+clarity, data = diamonds)
model4<-lm(price~cut*clarity, data = diamonds)
model5<-lm(price~depth*clarity, data = diamonds)

summary(model3)
vif(model3)
plot(model3)

```

```
> ## Multiple Linear Regression Model with carat, cut, depth, clarity, the interaction between cut and clarity, the interacti
on bewteen depth and clarity as the predictors, provide a summary of the fitted model, provide the variance inflation factor
 for each predictor, and show diagnostic plots for the model ##
> model3<-lm(price~carat+cut+depth+clarity, data = diamonds)
> model4<-lm(price~cut*clarity, data = diamonds)
> model5<-lm(price~depth*clarity, data = diamonds)
>
> summary(model3)

Call:
lm(formula = price ~ carat + cut + depth + clarity, data = diamonds)

Residuals:
     Min       1Q   Median       3Q      Max
-16814.2   -638.1   -114.8    471.7  11230.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1555.379    256.669  -6.060 1.37e-09 ***
carat        8472.278     12.610 671.847  < 2e-16 ***
cut.L         667.393     23.653  28.215  < 2e-16 ***
cut.Q        -305.271     20.345 -15.005  < 2e-16 ***
cut.C         187.942     17.211  10.920  < 2e-16 ***
cut^4           8.544     13.831   0.618   0.5368
depth         -26.229      4.118  -6.369 1.92e-10 ***
clarity.L    4001.317     33.958 117.833  < 2e-16 ***
clarity.Q   -1818.872     31.862 -57.086  < 2e-16 ***
clarity.C     913.445     27.311  33.446  < 2e-16 ***
clarity^4    -427.771     21.826 -19.599  < 2e-16 ***
clarity^5     253.553     17.823  14.226  < 2e-16 ***
clarity^6      28.692     15.536   1.847   0.0648 .
clarity^7     186.316     13.680  13.620  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1280 on 53926 degrees of freedom
Multiple R-squared:  0.897,     Adjusted R-squared:  0.897
F-statistic: 3.613e+04 on 13 and 53926 DF,  p-value: < 2.2e-16

> vif(model3)
            GVIF Df GVIF^(1/(2*Df))
carat   1.175452  1        1.084182
cut     1.249584  4        1.028243
depth   1.145076  1        1.070082
clarity 1.236851  7        1.015299
> plot(model3)
```


Normal Q-Q
Standardized residuals
Theoretical Quantiles
27416

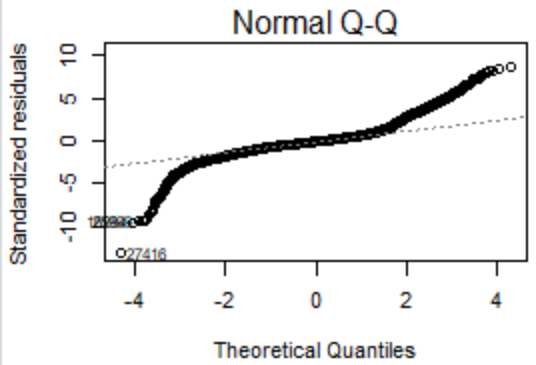

Scale-Location
√|Standardized residuals|
Fitted values
27416

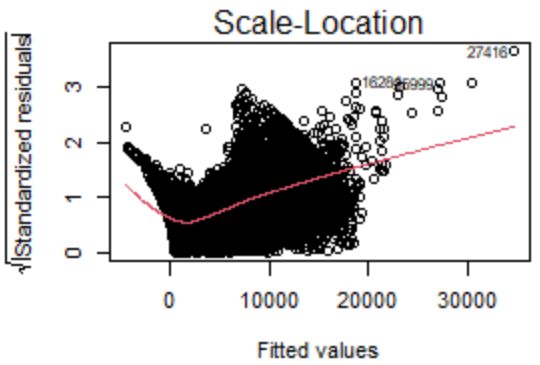

Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
27416

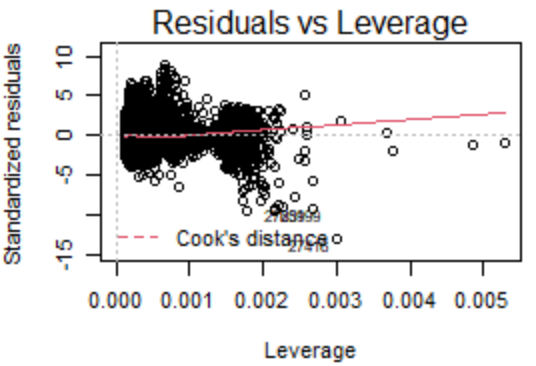
```
summary(model4)
vif(model4)
plot(model4)
```

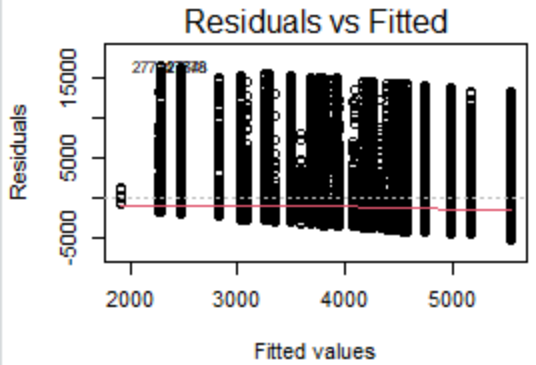
```
> summary(model4)

Call:
lm(formula = price ~ cut * clarity, data = diamonds)

Residuals:
   Min     1Q Median     3Q    Max
 -5201  -2605  -1375   1215  16533

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      3797.786     52.106  72.886  < 2e-16 ***
cut.L             -70.944    147.073  -0.482 0.629543
cut.Q            -277.947    129.540  -2.146 0.031906 *
cut.C            -441.082     99.898  -4.415 1.01e-05 ***
cut^4            -114.517     76.878  -1.490 0.136333
clarity.L       -1373.222    195.734  -7.016 2.31e-12 ***
clarity.Q        -285.636    182.458  -1.565 0.117473
clarity.C         748.651    164.016   4.565 5.02e-06 ***
clarity^4          -1.935    147.978  -0.013 0.989569
clarity^5         700.815    126.200   5.553 2.82e-08 ***
clarity^6        -377.618     99.864  -3.781 0.000156 ***
clarity^7          91.743     75.778   1.211 0.226023
cut.L:clarity.L  -413.175    548.489  -0.753 0.451275
cut.Q:clarity.L  -905.960    486.080  -1.864 0.062354 .
cut.C:clarity.L   150.202    375.003   0.401 0.688764
cut^4:clarity.L  -214.334    297.478  -0.721 0.471218
cut.L:clarity.Q   542.647    504.522   1.076 0.282126
cut.Q:clarity.Q  -813.975    450.230  -1.808 0.070626 .
cut.C:clarity.Q   698.529    353.600   1.975 0.048219 *
cut^4:clarity.Q   410.545    289.022   1.420 0.155478
cut.L:clarity.C   -63.756    459.448  -0.139 0.889635
cut.Q:clarity.C -1372.789    406.520  -3.377 0.000734 ***
cut.C:clarity.C   116.213    315.966   0.368 0.713021
cut^4:clarity.C  -147.698    248.670  -0.594 0.552547
cut.L:clarity^4   272.906    427.074   0.639 0.522817
cut.Q:clarity^4 -1230.670    372.072  -3.308 0.000942 ***
cut.C:clarity^4   613.739    277.581   2.211 0.027038 *
cut^4:clarity^4  -140.666    200.172  -0.703 0.482231
cut.L:clarity^5   473.570    367.912   1.287 0.198035
cut.Q:clarity^5  -552.232    318.034  -1.736 0.082500 .
cut.C:clarity^5    69.378    236.462   0.293 0.769218
cut^4:clarity^5  -224.152    161.580  -1.387 0.165371
cut.L:clarity^6   520.131    286.703   1.814 0.069656 .
cut.Q:clarity^6   -67.911    248.176  -0.274 0.784361
cut.C:clarity^6   217.580    193.457   1.125 0.260724
cut^4:clarity^6   -51.029    135.061  -0.378 0.705561
cut.L:clarity^7     2.306    211.864   0.011 0.991317
cut.Q:clarity^7  -378.093    185.211  -2.041 0.041215 *
cut.C:clarity^7   173.475    151.559   1.145 0.252377
cut^4:clarity^7   -86.033    112.636  -0.764 0.444980
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3914 on 53900 degrees of freedom
Multiple R-squared:  0.03797,   Adjusted R-squared:  0.03728
F-statistic: 54.55 on 39 and 53900 DF,  p-value: < 2.2e-16
```

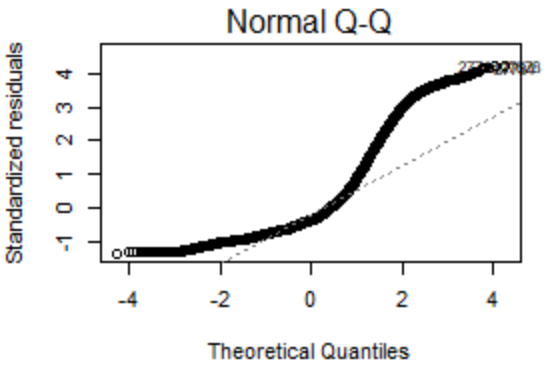
```
> vif(model4)
                    GVIF Df GVIF^(1/(2*Df))
cut             157.2219  4        1.881760
clarity        4723.2359  7        1.829977
cut:clarity  226269.8859 28        1.246288
> plot(model4)
>
```

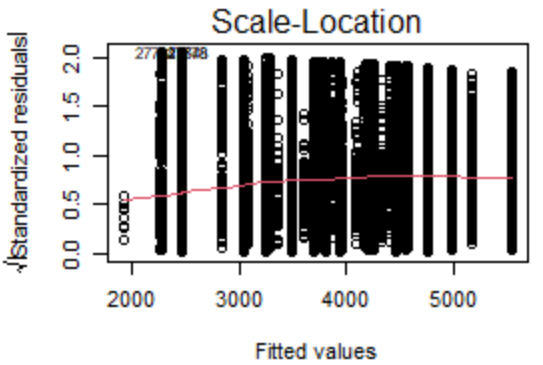

Normal Q-Q
Standardized residuals
Theoretical Quantiles
27416

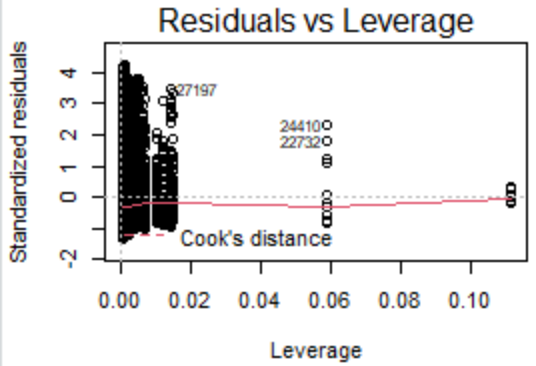

Scale-Location
√|Standardized residuals|
Fitted values
27416


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Residuals vs Fitted
Residuals
Fitted values


Normal Q-Q
Standardized residuals
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Standardized residuals
27197
244101
227320
Cook's distance
Leverage

```
summary(model5)
vif(model5)
plot(model5)
```

```
> summary(model5)

Call:
lm(formula = price ~ depth * clarity, data = diamonds)

Residuals:
   Min     1Q Median     3Q    Max
 -4846  -2730  -1423   1267  16317

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6372.84    1076.74   5.919 3.27e-09 ***
depth             -43.69      17.43  -2.506   0.0122 *
clarity.L         536.01    3851.00   0.139   0.8893
clarity.Q         998.27    3674.29   0.272   0.7859
clarity.C        5041.82    3354.98   1.503   0.1329
clarity^4       -1131.42    2953.87  -0.383   0.7017
clarity^5        -188.28    2635.72  -0.071   0.9431
clarity^6        -758.74    2366.26  -0.321   0.7485
clarity^7       -4140.42    2016.62  -2.053   0.0401 *
depth:clarity.L   -36.92      62.31  -0.593   0.5535
depth:clarity.Q   -23.29      59.44  -0.392   0.6951
depth:clarity.C   -71.24      54.30  -1.312   0.1895
depth:clarity^4    16.25      47.85   0.340   0.7342
depth:clarity^5    16.02      42.72   0.375   0.7077
depth:clarity^6     7.99      38.35   0.208   0.8350
depth:clarity^7    68.38      32.68   2.092   0.0364 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3935 on 53924 degrees of freedom
Multiple R-squared:  0.02758,   Adjusted R-squared:  0.02731
F-statistic: 101.9 on 15 and 53924 DF,  p-value: < 2.2e-16

> vif(model5)
                    GVIF Df GVIF^(1/(2*Df))
depth       2.173061e+00  1        1.474131
clarity     9.381687e+22  7       43.739842
depth:clarity 9.382963e+22  7       43.740267
> plot(model5)
>
```