PROJECT SAS-R – EVALUATION 4 DU 01 NOVEMBRE 2025

DINH THI THAO NHUNG – NGUYEN LINH CHI

Exercise 1. Overview

1.1.Load and desribe the data:

```
diamonds <- read.csv("C:\\Users\\Thinkpad\\Downloads\\diamonds\\diamonds.csv")

str(diamonds)

summary(diamonds)
```

Result

```
'data.frame':   53940 obs. of  11 variables:
 $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut    : chr  "Ideal" "Premium" "Good" "Premium" ...
 $ color  : chr  "E" "E" "E" "I" ...
 $ clarity: chr  "SI2" "SI1" "VS1" "VS2" ...
 $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
 $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
 $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

```
       X             carat             cut                color              clarity              depth      
 Min.   :    1   Min.   :0.2000   Length:53940       Length:53940       Length:53940       Min.   :43.00  
 1st Qu.:13486   1st Qu.:0.4000   Class :character   Class :character   Class :character   1st Qu.:61.00  
 Median :26971   Median :0.7000   Mode  :character   Mode  :character   Mode  :character   Median :61.80  
 Mean   :26971   Mean   :0.7979                                                           Mean   :61.75  
 3rd Qu.:40455   3rd Qu.:1.0400                                                           3rd Qu.:62.50  
 Max.   :53940   Max.   :5.0100                                                           Max.   :79.00  
     table           price             x                y                z          
 Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910  
 Median :57.00   Median : 2401   Median : 5.700   Median : 5.710   Median : 3.530  
 Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735   Mean   : 3.539  
 3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040  
 Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900   Max.   :31.800  
```

Overview of the dataset:

The dataset contains 53,940 diamonds and 11 variables describing their characteristics, including size, cut, color, clarity, and price. Most variables are numeric (carat, depth, table, price, x, y, z), while cut, color, and clarity are categorical descriptors of quality.

Exercise 2. Fitted model using continuous variables

We first explored correlations to assess predictive power and multicollinearity. Only after this exploratory step did we fit regression models to select the most informative variables.

2.1 Correlation with price

```
num_vars <- diamonds %>% select(price, carat, depth, table, x, y, z)

cor_with_price <- cor(num_vars)[, "price"]

round(cor_with_price, 2)
```

result:

```
price carat depth table     x     y     z
 1.00  0.92 -0.01  0.13  0.88  0.87  0.86
```

Correlation with price shows that carat (0.92) and the dimensions x, y, z (0.88–0.86) are strong predictors of diamond price. Depth (-0.01) and table (0.13) are weakly correlated and contribute little

2.2 check multicollinearity between predictors

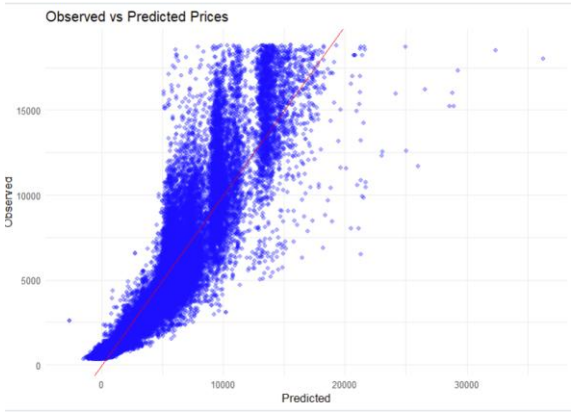
```
num_vars <- diamonds %>% select(carat, depth, table, x, y, z)

cor(num_vars)
```

```
            carat       depth      table          x          y          z
carat 1.00000000  0.02822431  0.1816175  0.97509423  0.95172220 0.95338738
depth 0.02822431  1.00000000 -0.2957785 -0.02528925 -0.02934067 0.09492388
table 0.18161755 -0.29577852  1.0000000  0.19534428  0.18376015 0.15092869
x     0.97509423 -0.02528925  0.1953443  1.00000000  0.97470148 0.97077180
y     0.95172220 -0.02934067  0.1837601  0.97470148  1.00000000 0.95200572
z     0.95338738  0.09492388  0.1509287  0.97077180  0.95200572 1.00000000
```

Carat and the dimensions x, y, z are highly correlated (0.95–0.98), causing multicollinearity. Depth and table are weakly correlated with other variables. To avoid redundancy while keeping three predictors, we recommend carat, depth, and table.

```
final_model <- lm(price ~ carat + depth + table, data = diamonds)

summary(final_model)

AIC(final_model)

vif(final_model)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13003.441    390.918   33.26   <2e-16 ***
carat        7858.771     14.151  555.36   <2e-16 ***
depth        -151.236      4.820  -31.38   <2e-16 ***
table        -104.473      3.141  -33.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1526 on 53936 degrees of freedom
Multiple R-squared:  0.8537,    Adjusted R-squared:  0.8537
F-statistic: 1.049e+05 on 3 and 53936 DF,  p-value: < 2.2e-16

> AIC(final_model)
[1] 943891.9
> vif(final_model)
   carat    depth    table
1.042039 1.104275 1.141032
```



The fitted model:

$$\widehat{\{price\}} = 13003.44 + 7858.77 \times carat - 151.24 \times depth - 104.47 \times table$$

Interpretation:

The linear regression model uses carat, depth, and table to predict diamond price. Carat has the largest positive effect, increasing price by ~7,859 per additional carat, and is highly significant (p < 2e-16). Depth and table have small negative effects (-151 and -104, respectively), also highly significant, indicating minor adjustments to price. The model explains ~85% of the variation in price (Adjusted R^2 = 0.854), with a highly significant F-statistic. The AIC (943,892) provides a measure of model quality, and all predictors have low VIFs (~1-1.14), indicating minimal multicollinearity.

Exercise 3: Fitted model using discrete variables

3.1 select discrete variables

```
> str(diamonds)
'data.frame':   53940 obs. of  12 variables:
 $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ carat    : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut      : Factor w/ 5 levels "Fair","Good",..: 3 4 2 4 2 5 5 5 1 5 ...
 $ color    : Factor w/ 7 levels "D","E","F","G",..: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity  : Factor w/ 8 levels "I1","IF","SI1",..: 4 3 5 6 4 8 7 3 6 5 ...
 $ depth    : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table    : num  55 61 65 58 58 57 57 55 61 61 ...
 $ price    : int  326 326 327 334 335 336 336 337 337 338 ...
 $ x        : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y        : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z        : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
 $ predicted: num  -236 -763 -585 -214 -193 ...
```

⇨ 3 discrete variables are cut, color and clarity.

We run ANOVA table to quantify significance of the predictors

```
> anova_cut <- aov(price ~ cut, data = diamonds); summary(anova_cut)
              Df    Sum Sq   Mean Sq F value Pr(>F)
cut            4 1.104e+10 2.760e+09   175.7 <2e-16 ***
Residuals  53935 8.474e+11 1.571e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova_color <- aov(price ~ color, data = diamonds); summary(anova_color)
              Df    Sum Sq   Mean Sq F value Pr(>F)
color          6 2.685e+10 4.475e+09   290.2 <2e-16 ***
Residuals  53933 8.316e+11 1.542e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova_clarity <- aov(price ~ clarity, data = diamonds); summary(anova_clarity)
              Df    Sum Sq   Mean Sq F value Pr(>F)
clarity        7 2.331e+10 3.330e+09     215 <2e-16 ***
Residuals  53932 8.352e+11 1.549e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

All three categorical variables—cut, color, and clarity—significantly influence diamond price ($p < 0.001$ for all). Among them, color has the largest F value, indicating it explains the most variation in price, followed by clarity and then cut. Although cut explains less variation, it is still statistically significant. These results show that discrete characteristics of diamonds meaningfully affect price and justify their inclusion in a categorical-only regression model.

3.2 Fitting the categorical variables and grouping

```
model_cat <- lm(price ~ cut + clarity + color, data = diamonds)

summary(model_cat)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2882.983    165.966  17.371  < 2e-16 ***
cutGood      -214.866    112.052  -1.918   0.0552 .
cutIdeal     -403.871    101.998  -3.960 7.52e-05 ***
cutPremium    414.170    103.090   4.018 5.89e-05 ***
cutVery Good  -35.957    104.187  -0.345   0.7300
clarityIF    -784.217    171.527  -4.572 4.84e-06 ***
claritySI1    316.532    147.940   2.140   0.0324 *
claritySI2   1354.896    149.194   9.081  < 2e-16 ***
clarityVS1     92.755    150.463   0.616   0.5376
clarityVS2    293.973    148.428   1.981   0.0476 *
clarityVVS1 -1068.271    158.305  -6.748 1.51e-11 ***
clarityVVS2  -238.377    154.536  -1.543   0.1229
colorE         -8.097     61.050  -0.133   0.8945
colorF        689.236     61.539  11.200  < 2e-16 ***
colorG       1059.873     59.995  17.666  < 2e-16 ***
colorH       1371.528     63.362  21.646  < 2e-16 ***
colorI       2000.190     70.468  28.385  < 2e-16 ***
colorJ       2126.119     86.785  24.499  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3856 on 53922 degrees of freedom
Multiple R-squared:  0.06587,   Adjusted R-squared:  0.06557
F-statistic: 223.7 on 17 and 53922 DF,  p-value: < 2.2e-16
```

Based on the categorical regression results, I grouped **cut** and **clarity** because some levels have very similar effects on diamond price. For cut, Fair and Good have small, non-significant differences, while Premium and Ideal show similar positive effects, so grouping simplifies the model without losing information. For clarity, levels such as SI1 and VS2 show similar price effects, and the extremes (I1, SI2 for low, VS1, VVS1, VVS2, IF for high) can be combined into Low, Medium, and High groups. **Color** was kept separate because each level has a distinct and significant impact on price, so grouping would reduce interpretability.

Fitted model:

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            2427.09      55.46  43.763   <2e-16 ***
cut_groupedLow         -103.40      52.76  -1.960    0.050 *
cut_groupedMedium        62.91      40.97   1.536    0.125
clarity_groupedLow     1701.07      48.83  34.836   <2e-16 ***
clarity_groupedMedium   693.43      37.94  18.279   <2e-16 ***
colorE                  -15.21      61.43  -0.248    0.804
colorF                  669.62      61.87  10.824   <2e-16 ***
colorG                 1048.83      60.25  17.408   <2e-16 ***
colorH                 1377.74      63.65  21.644   <2e-16 ***
colorI                 2017.50      70.85  28.475   <2e-16 ***
colorJ                 2211.04      87.19  25.358   <2e-16 ***
---
```

$$\widehat{price} = 2427.09 - 103.40 \cdot Cut\_Low + 62.91 \cdot Cut\_Medium + 1701.07 \cdot Clarity\_Low + 693.43 \cdot Clarity\_Medium + (Color\ Effects)$$

Interpretation :

Cut
Grouping reduced levels effectively. Low-quality cuts slightly reduce price (~-103), and medium cuts show a small positive effect (~63). Overall, cut has a modest but statistically significant effect (ANOVA p = 0.009).

Clarity
Clarity is the strongest categorical determinant of price. Compared to high clarity, low clarity increases price by ~1701 and medium clarity by ~693 (p < 2e-16). Clarity differences remain economically and statistically meaningful.

Color
Color shows a clear and consistent price gradient: diamonds with lower color grades (toward J) have significantly higher prices compared to D, ranging ~670 to ~2211. Color has a strong overall effect (p < 2e-16).

Model fit
The grouped categorical model explains ~5% of price variation (R^2 ≈

0.053). This is expected since key numeric factors are not yet included; categorical features provide modest explanatory power alone.

Exercise 4:

4.1.Fitting the model with bot continuous and discrete variables

```
model_combined <- lm(price ~ carat + depth + table + cut_group + clarity_group + color, data = diamonds)

summary(model_combined)
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1960.380    362.103   5.414 6.19e-08 ***
carat               8782.605     12.715 690.754  < 2e-16 ***
depth                -58.852      4.221 -13.942  < 2e-16 ***
table                -46.174      2.745 -16.821  < 2e-16 ***
cut_groupMedium      373.892     19.667  19.012  < 2e-16 ***
cut_groupHigh        433.435     18.569  23.341  < 2e-16 ***
clarity_groupMedium 1430.173     15.013  95.265  < 2e-16 ***
clarity_groupHigh   2348.737     16.581 141.655  < 2e-16 ***
colorE              -200.087     19.453 -10.285  < 2e-16 ***
colorF              -286.359     19.636 -14.583  < 2e-16 ***
colorG              -474.238     19.207 -24.691  < 2e-16 ***
colorH              -981.956     20.450 -48.017  < 2e-16 ***
colorI             -1428.257     22.986 -62.136  < 2e-16 ***
colorJ             -2306.614     28.375 -81.290  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1229 on 53926 degrees of freedom
Multiple R-squared:  0.9051,    Adjusted R-squared:  0.905
F-statistic: 3.954e+04 on 13 and 53926 DF,  p-value: < 2.2e-16
```

Fitted model:

$Price\hat{} = \text{intercept} + \sum(\text{coefficient} * \text{variables})$

Interpretation:

Numeric variables:

Carat is the strongest predictor — each additional carat increases price by ~8783.

Depth and table have small negative effects.

Categorical variables:

Cut:

Medium (+374) and High (+433) increase price relative to Low.

Clarity:

Medium (+1430) and High (+2349) increase price relative to Low — very strong effect.

Color:

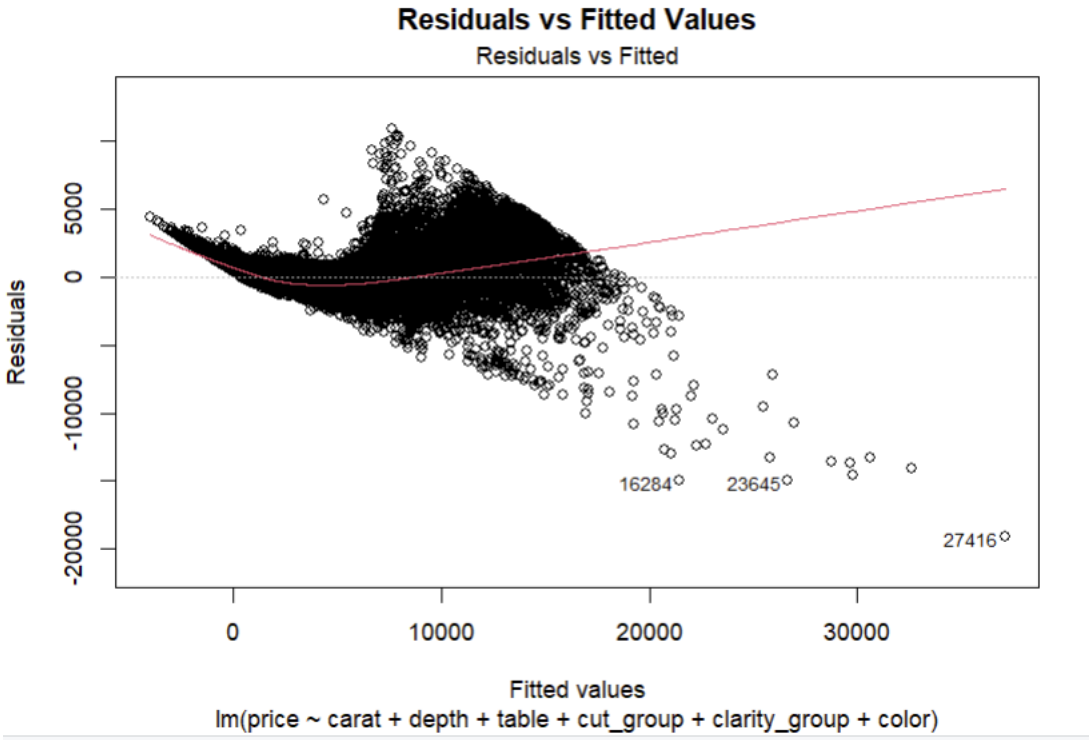
Surprisingly, E-J have negative coefficients compared to D — check factor reference, but it reflects adjustments relative to D.

Adjusted R^2 = 0.905 → numeric + categorical variables together explain ~90.5% of price variation.

In conclusion, our model explains diamond prices using both physical characteristics (carat, depth, table) and key categorical attributes (cut, color, clarity) that reflect quality. Carat weight is the strongest driver of price, showing that larger diamonds are significantly more expensive. Cut, clarity, and color groups also matter: higher-quality categories are associated with higher prices, even after controlling for size. Depth and table have smaller effects, meaning geometric proportions are less influential than weight and quality grades. Overall, the model captures major pricing dimensions and shows that diamond value is determined by a combination of size and perceived quality.

Exercise 5. Verify model assumptions

5.1.Linearity check



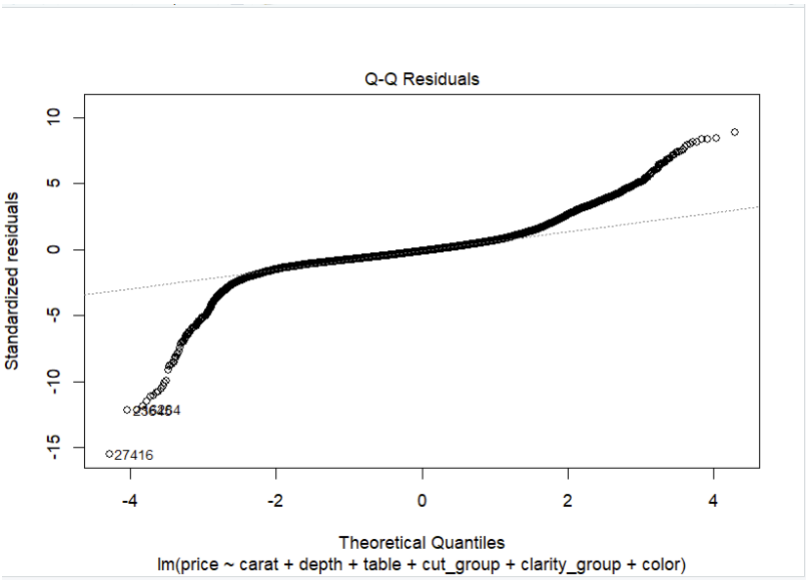
The plot shows that the current linear model violates key regression assumptions, making its predictions unreliable.

Non-Linearity: The data exhibits a strong curved (U-shaped) pattern around the zero line, indicated by the red line following the curve

instead of remaining flat. This means a linear model is inappropriate for this data.

Heteroscedasticity: The vertical spread of the residuals is not constant (unequal variance), violating the assumption of homoscedasticity.

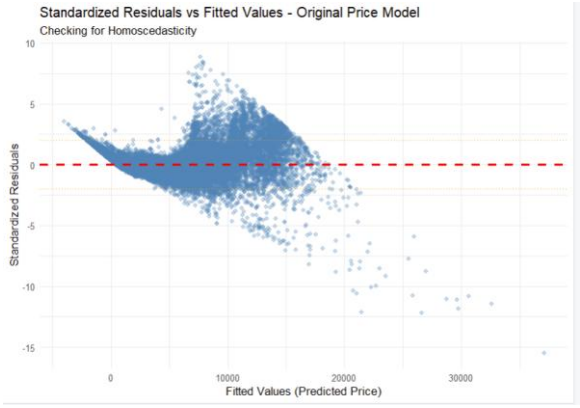
5.2. Normality of residual



The Q-Q Plot shows a severe violation of the normality assumption, as the residuals deviate sharply from the straight line, especially in the tails (S-shape). This indicates the model errors are negatively skewed.

5.3. Homoskedasticity test

```
        studentized Breusch-Pagan test

data:  model_combined
BP = 7075.2, df = 13, p-value < 2.2e-16
```



Since the p-value is extremely low , we **reject the null hypothesis** (H0: Variance is constant (Homoscedasticity))

The test formally confirms the presence of severe **Heteroscedasticity**.

5.4. Independence of residuals

```
        Durbin-Watson test

data:  model_combined
DW = 0.95941, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

The test shows that there are autocorrelations among residuals.

5.5.Multicollinearity

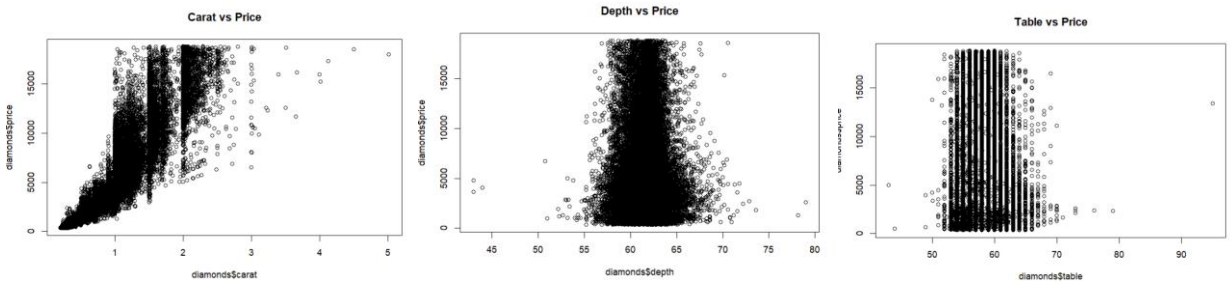
```
> vif(model_combined)
                    GVIF Df GVIF^(1/(2*Df))
carat           1.296214  1        1.138514
depth           1.305157  1        1.142435
table           1.342654  1        1.158729
cut_group       1.280769  2        1.063819
clarity_group 1.197396  2        1.046067
color           1.152980  6        1.011933
>
```

The model shows no evidence of problematic multicollinearity as all VIF values are smaller than 5. The predictor variables are not highly correlated with each other, meaning their individual effects on the price can be reliably estimated.

MODEL LIMITATION:

We fully acknowledge that the diagnostic analysis, including the Residuals vs. Fitted Plot, the Q-Q Plot, and the Breusch-Pagan Test, strongly indicated that the relationship is non-linear and heteroscedastic.

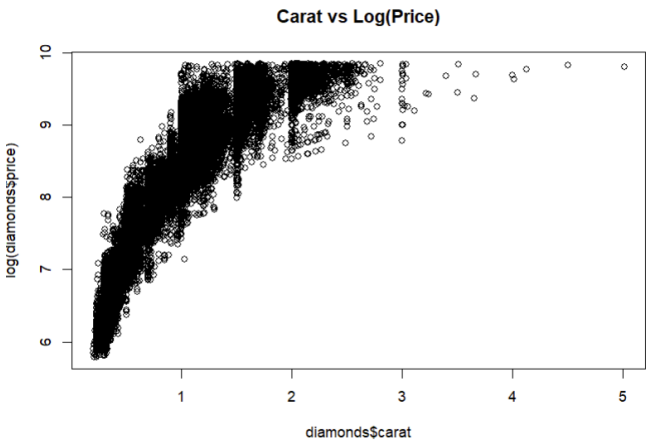
We can clearly see from the plots, only carat shows some upward trend towards price.

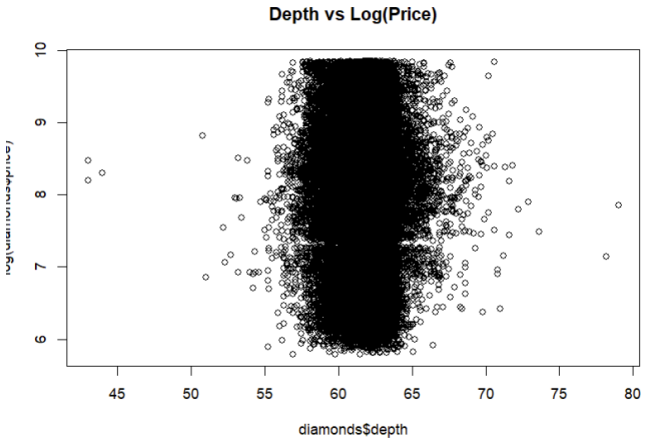


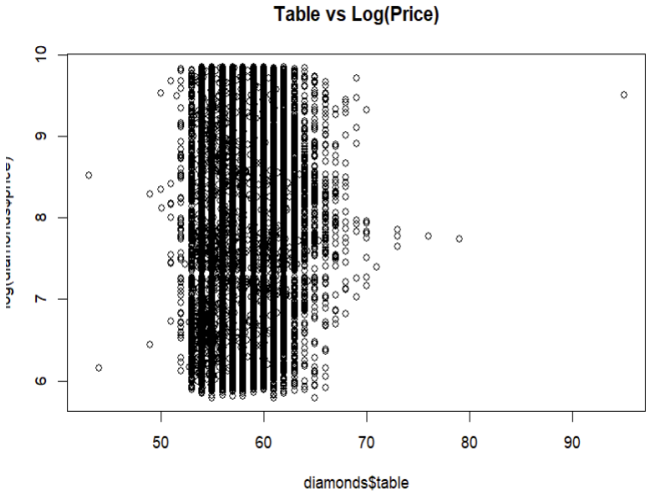
Alternative model check, using logarithics transformation of price to prevent right-skewed distribution.

Brief comparision

```
Original Price Model:
> cat("R-squared:", summary(model_combined)$r.squared, "\n")
R-squared: 0.9050545
> cat("Adjusted R-squared:", summary(model_combined)$adj.r.squared, "\n")
Adjusted R-squared: 0.9050316
> cat("AIC:", AIC(model_combined), "\n")
AIC: 920581.8
>
> cat("\nLog-transformed Price Model:\n")

Log-transformed Price Model:
> cat("R-squared:", summary(model_combined_log)$r.squared, "\n")
R-squared: 0.8831163
> cat("Adjusted R-squared:", summary(model_combined_log)$adj.r.squared, "\n")
Adjusted R-squared: 0.8830881
> cat("AIC:", AIC(model_combined_log), "\n")
AIC: 38886.68
```







The logarit transformation model is better in model selection, as shown by significant lower AIC, though the explanatory power is slight reduced (adjusted Rsquared is slightly lower compared to the original model). However, the plots show non-linearity relationshi between numeric explantory variables and price/log(price). That explains why almost assumptions were violated.