126-final-project

Chengwei Zhang, Ji Qi and Lin Fang

2024-06-07

Part 1: Data Description and Descriptive Statistics

1.1 Indroduction

In the diamond market, the price of a diamond is influenced by various factors such as carat, cut, color, clarity, and other physical dimensions. Understanding these relationships can help in predicting diamond prices and making informed purchasing or selling decisions. In this report, we analyze diamond pricing data using linear regression to understand how different attributes of a diamond affect its price.

The primary purpose of this report is to explore the relationships between the listed price of a diamond and its various attributes, including carat, cut, color, clarity, depth, and dimensions. By analyzing these relationships, we aim to provide insights into the key factors that determine diamond pricing and develop a predictive model.

This report is structured into four sections. Section 2 covers exploratory data analysis (EDA), providing a general understanding of the dataset and forming expectations about our models. In Section 3, we select and fit the most reasonable model through variable selection and regression techniques. Section 4 interprets our model in the context of real-world scenarios and draws conclusions based on the findings.

1.2 Describe All Variables

In this dataset, there are ten variables: nine independent variables and one response variable. Below is a detailed description of these variables:

- **Carat (0.2 - 5.01):** The carat is the diamond's physical weight measured in metric carats. One carat equals 0.20 grams and is subdivided into 100 points. The weight of the diamond is a critical factor influencing its price.
- **Cut (Fair, Good, Very Good, Premium, Ideal):** The quality of the diamond's cut. The cut grade ranges from Fair to Ideal, with Ideal being the highest quality. The precision of the cut affects the diamond's brilliance and overall appearance, thus impacting its price.
- **Color (J (worst) to D (best)):** The color grade of the diamond, ranging from J (worst) to D (best). Colorless diamonds (D) are the most valuable, while diamonds with a yellowish tint (J) are less valuable. The color grading significantly influences the diamond's price.
- **Clarity (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)):** This measures the presence of internal characteristics (inclusions) or external characteristics (blemishes) within the diamond. Clarity grades range from I1 (included) to IF (internally flawless), with higher clarity grades being more valuable. Clarity affects the diamond's appearance and value.

- **Depth (43 - 79):** The total depth percentage of the diamond, calculated as 2 * z / (x + y). Depth measures the height of the diamond from the culet (bottom tip) to the table (top flat surface). The depth percentage affects the diamond's brilliance and is a factor in pricing.
- **Table (43 - 95):** The width of the top of the diamond relative to its widest point. A larger table size can enhance the diamond's ability to reflect light, affecting its visual appeal and price.
- **Price ($326 - $18,826):** The listed price of the diamond in US dollars. This is the target variable for our analysis, representing the value we aim to predict based on other attributes.
- **X (0 - 10.74):** Length of the diamond in millimeters. The dimensions of the diamond (length, width, and depth) contribute to its overall size and weight.
- **Y (0 - 58.9):** Width of the diamond in millimeters. The width, along with the length and depth, defines the diamond's size, which is correlated with its carat weight and price.
- **Z (0 - 31.8):** Depth of the diamond in millimeters. The depth dimension, together with length and width, affects the diamond's proportions and overall appearance.

The dataset contains nearly 54,000 observations with no missing data, providing a comprehensive overview of the diamond market. Through EDA, we aim to summarize the key characteristics of the data, visualize the distributions and relationships among variables, and identify any interesting patterns or anomalies. This analysis will set the stage for building and evaluating predictive models in the subsequent sections.

1.3 Random Sampling

To begin our analysis, we selected a random sample of 500 diamonds from the original dataset. Random sampling is crucial to ensure that our sample is representative of the overall population and to avoid any bias that might arise from non-random selection.

The steps involved in selecting the random sample were as follows:

```
##   carat       cut color clarity depth table price    x    y    z
## 1  0.23     Ideal     E     SI2  61.5    55   326 3.95 3.98 2.43
## 2  0.21   Premium     E     SI1  59.8    61   326 3.89 3.84 2.31
## 3  0.23      Good     E     VS1  56.9    65   327 4.05 4.07 2.31
## 4  0.29   Premium     I     VS2  62.4    58   334 4.20 4.23 2.63
## 5  0.31      Good     J     SI2  63.3    58   335 4.34 4.35 2.75
## 6  0.24 Very Good     J    VVS2  62.8    57   336 3.94 3.96 2.48
```

To understand the sample data, we performed exploratory data analysis (EDA) using the skimr package, which provides a concise summary of the dataset. The summary statistics for the sample data were as follows:

- `Number of rows: 500`
- `Number of columns: 10`
- `Column types: 3 factor variables and 8 numeric variables`

1.4 Summary Statistics

```r
#Pick a random sample of 500 data, get the skim of the sample data
library(skimr)
set.seed(126)
diamond$cut <- as.factor(diamond$cut)
diamond$color <- as.factor(diamond$color)
diamond$clarity <- as.factor(diamond$clarity)
diamond <- diamond[sample(nrow(diamond),500),]
skim(diamond)
```

Table 1: Data summary

Name	diamond
Number of rows	500
Number of columns	10
Column type frequency:	
factor	3
numeric	7
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
cut	0	1	FALSE	5	Ide: 211, Pre: 124, Ver: 110, Goo: 43
color	0	1	FALSE	7	E: 95, F: 92, G: 86, H: 83
clarity	0	1	FALSE	8	VS2: 121, SI1: 116, SI2: 92, VS1: 73

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
carat	0	1	0.80	0.48	0.22	0.40	0.71	1.02	3.05	
depth	0	1	61.70	1.46	56.40	61.00	61.70	62.50	72.20	
table	0	1	57.50	2.21	53.00	56.00	57.00	59.00	68.00	
price	0	1	3847.89	3864.40	379.00	1018.00	2432.50	4889.00	18281.00	
x	0	1	5.74	1.11	3.90	4.74	5.69	6.52	9.38	
y	0	1	5.74	1.10	3.85	4.77	5.71	6.50	9.31	
z	0	1	3.54	0.68	2.38	2.93	3.55	4.01	5.66	

1.5 Initial MLR Model Fitting

Backward Stepwise Selection

To build the regression model, we used backward stepwise selection, a method that iteratively removes the least significant variables from the model to find the optimal subset of predictors. The process involved fitting an initial model with all predictors and then removing variables step-by-step based on their significance.

We started with a full model including all predictor variables: carat, cut, color, clarity, depth, table, x, y, and z.

1.6 Model Summary and Interpretation

```
model_total <- lm(price ~ ., data = diamond)
summary(model_total)
```

```
##
## Call:
## lm(formula = price ~ ., data = diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9000.7  -603.5  -191.3   406.5  5772.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6769.61   13829.33  -0.490   0.6247
## carat         9638.32     552.53  17.444  < 2e-16 ***
## cutGood        105.77     406.63   0.260   0.7949
## cutIdeal       689.21     402.60   1.712   0.0876 .
## cutPremium     653.81     389.40   1.679   0.0938 .
## cutVery Good   533.26     396.02   1.347   0.1788
## colorE        -316.00     186.74  -1.692   0.0913 .
## colorF        -142.16     190.29  -0.747   0.4554
## colorG        -489.82     197.48  -2.480   0.0135 *
## colorH       -1124.03     199.18  -5.643 2.87e-08 ***
## colorI       -1307.80     233.02  -5.613 3.39e-08 ***
## colorJ       -2354.42     309.89  -7.598 1.61e-13 ***
## clarityIF     6291.10     510.56  12.322  < 2e-16 ***
## claritySI1    4680.00     416.86  11.227  < 2e-16 ***
## claritySI2    3826.63     417.91   9.156  < 2e-16 ***
## clarityVS1    5509.98     432.51  12.739  < 2e-16 ***
## clarityVS2    5094.27     419.20  12.152  < 2e-16 ***
## clarityVVS1   5871.67     477.76  12.290  < 2e-16 ***
## clarityVVS2   5618.29     448.17  12.536  < 2e-16 ***
## depth           67.74     217.78   0.311   0.7559
## table          -61.32      32.58  -1.882   0.0604 .
## x              409.95    1578.86   0.260   0.7952
## y             -337.32    1589.65  -0.212   0.8320
## z             -840.33    3469.89  -0.242   0.8087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1181 on 476 degrees of freedom
## Multiple R-squared:  0.9109, Adjusted R-squared:  0.9066
## F-statistic: 211.6 on 23 and 476 DF,  p-value: < 2.2e-16
```

Backward Stepwise Selection We used the step function with the direction set to "backward" to perform the stepwise selection. The AIC (Akaike Information Criterion) was used as the criterion for variable selection.

```
step(model_total, direction = "backward")
```

The backward stepwise selection process identified the following five significant predictors: carat, color, clarity, table, and z, which are two categorical variables and three numerical variables.

```r
model_initial <- lm(formula = price ~ carat + color + clarity + table + z, data = diamond)
summary(model_initial)
```

```
##
## Call:
## lm(formula = price ~ carat + color + clarity + table + z, data = diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8872.1  -630.7  -188.0   382.3  5938.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -375.13    1767.08  -0.212 0.831970
## carat        9685.89     498.15  19.444  < 2e-16 ***
## colorE       -312.90     186.61  -1.677 0.094237 .
## colorF       -107.61     189.66  -0.567 0.570715
## colorG       -478.79     196.73  -2.434 0.015309 *
## colorH      -1085.62     197.64  -5.493 6.41e-08 ***
## colorI      -1290.33     230.80  -5.591 3.79e-08 ***
## colorJ      -2426.16     306.04  -7.928 1.56e-14 ***
## clarityIF    6476.99     495.39  13.074  < 2e-16 ***
## claritySI1   4820.02     405.29  11.893  < 2e-16 ***
## claritySI2   3992.58     407.41   9.800  < 2e-16 ***
## clarityVS1   5690.20     419.50  13.564  < 2e-16 ***
## clarityVS2   5277.23     405.11  13.027  < 2e-16 ***
## clarityVVS1  6046.07     465.75  12.981  < 2e-16 ***
## clarityVVS2  5820.87     432.88  13.447  < 2e-16 ***
## table         -93.29      25.01  -3.730 0.000214 ***
## z            -727.59     345.62  -2.105 0.035792 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1183 on 483 degrees of freedom
## Multiple R-squared:  0.9093, Adjusted R-squared:  0.9063
## F-statistic: 302.5 on 16 and 483 DF,  p-value: < 2.2e-16
```

Part 2: Simple Linear Regression

2.1 Initial SLR Model Fitting

After comparing the adjusted R-squared value of each SLR model, the model with carat as IV have the highest adjusted R-square. Therefore we choose carat as the IV for our SLR model. This initial fitting will help us understand the relationship between the weight of the diamond (carat) and its price.

The initial model was: [$\log(price) = \beta_0 + \beta_1 log(carat) + \varepsilon_i$] Where:

- (β_0) is the intercept
- (β_1) is the coefficient for carat
- (ε_i) is the error term.

Model Summary and Interpretation

```
##
## Call:
## lm(formula = price ~ carat, data = diamond)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10135.5   -748.7    -32.4    453.9   7552.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2074.5      136.0  -15.26   <2e-16 ***
## carat         7430.5      146.5   50.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1558 on 498 degrees of freedom
## Multiple R-squared:  0.8378, Adjusted R-squared:  0.8375
## F-statistic:  2572 on 1 and 498 DF,  p-value: < 2.2e-16
```

The model summary provides key statistics:

- **Coefficients**: The coefficient for 'carat' is significant with a p-value much less than 0.05, indicating a strong linear relationship between 'carat' and 'price'.

- R^2 **:** The R^2 value is 0.8375, suggesting that 83.75% of the variability in the diamond prices can be explained by the carat weight.

- **Residual Standard Error**: The residual standard error is 1558, indicating the average amount that the observed values deviate from the regression line.

- **F-statistic**: The F-statistic value is extremely high with a significant p-value, further confirming that the model explains a significant amount of the variance in the response variable.

- **Confidence Intervals**: The confidence intervals for the coefficients show that the effect of 'carat' on 'price' is significant and does not include zero.

- **Diagnostic Plots**: The residual plots suggest a reasonably good fit but show some indication of heteroscedasticity.

Assumption Testing and Transformations

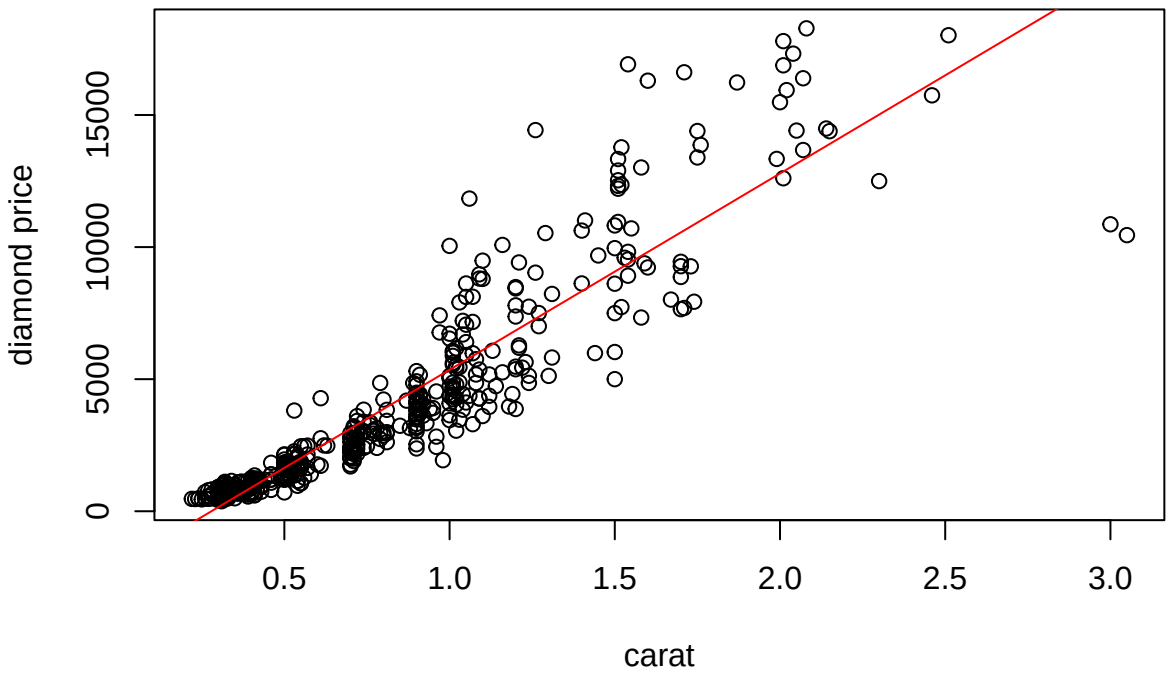
The initial regression model was subjected to diagnostic checks to evaluate the following assumptions:

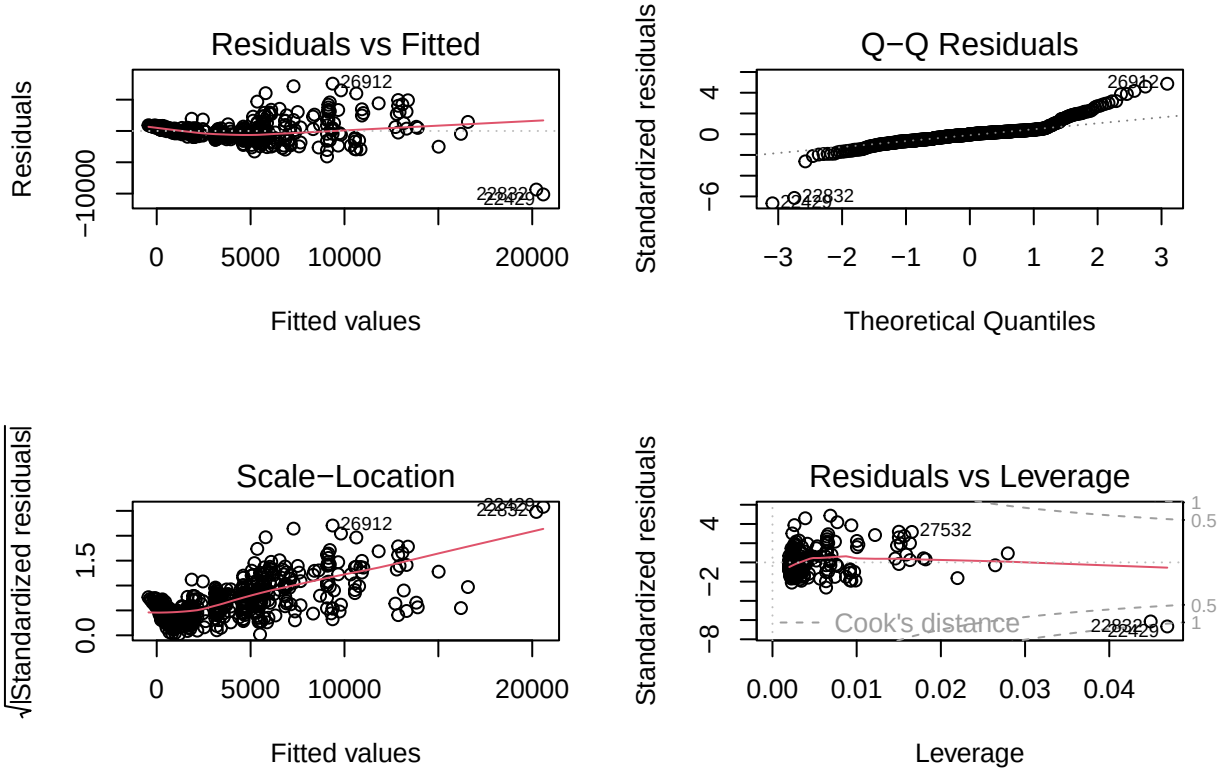
1.Linearity: The relationship between the IV and DV should be linear.

2.Independence: The residuals (errors) should be independent.

3.Homoscedasticity: The residuals should have constant variance.

4.Normality: The residuals should be normally distributed.


diamond price
0
5000
10000
15000
0.5
1.0
1.5
2.0
2.5
3.0
carat



Adjusted R^2 value:

```
## [1] 0.8374587
```

Initial Findings:

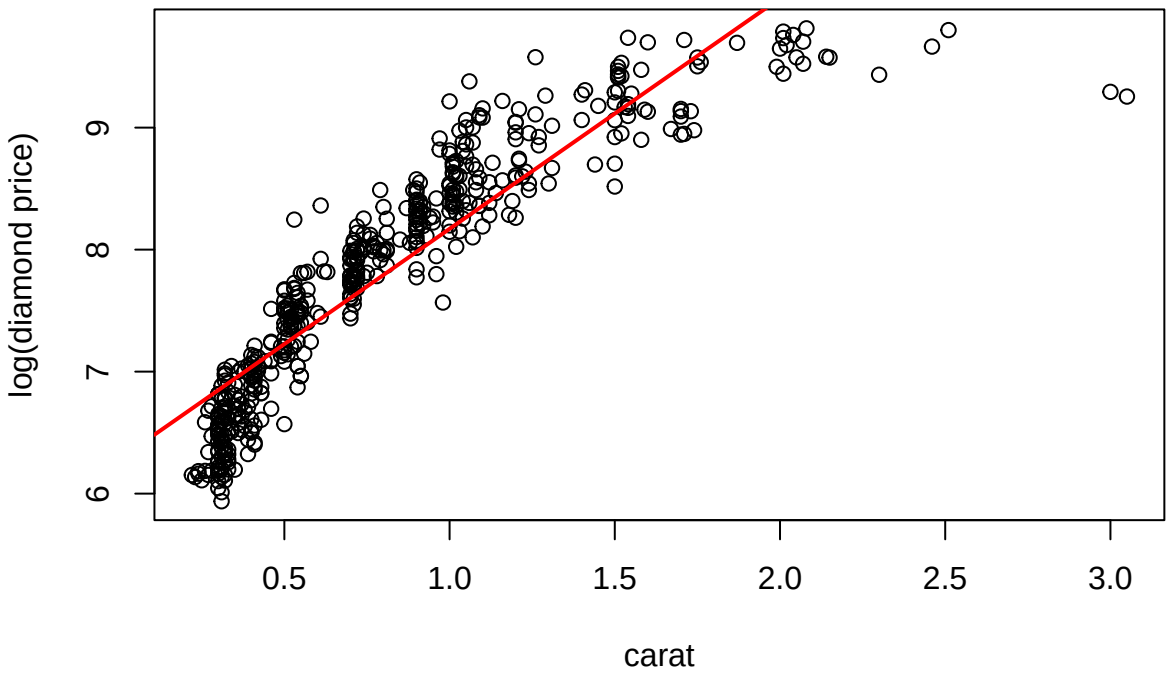
- **Linearity**: The residuals vs fitted plot is more diverged as the value of carat get large, the position of the dots indicates some non-linearity, suggesting that *a transformation might be necessary.*
- **Expected Value of Residuals**: The expected value of the residuals was approximately zero, satisfying the requirement $\mathrm{E}(\varepsilon_i) = 0$.
- **Homoscedasticity**: The plot of residuals versus fitted values indicated a pattern, suggesting *non-constant variance* (heteroscedasticity).
- **Independence**: The residuals do not show any apparent pattern, implying that the independence assumption holds.
- **Normality**: The residuals were roughly normally distributed, as assessed by a Q-Q plot.

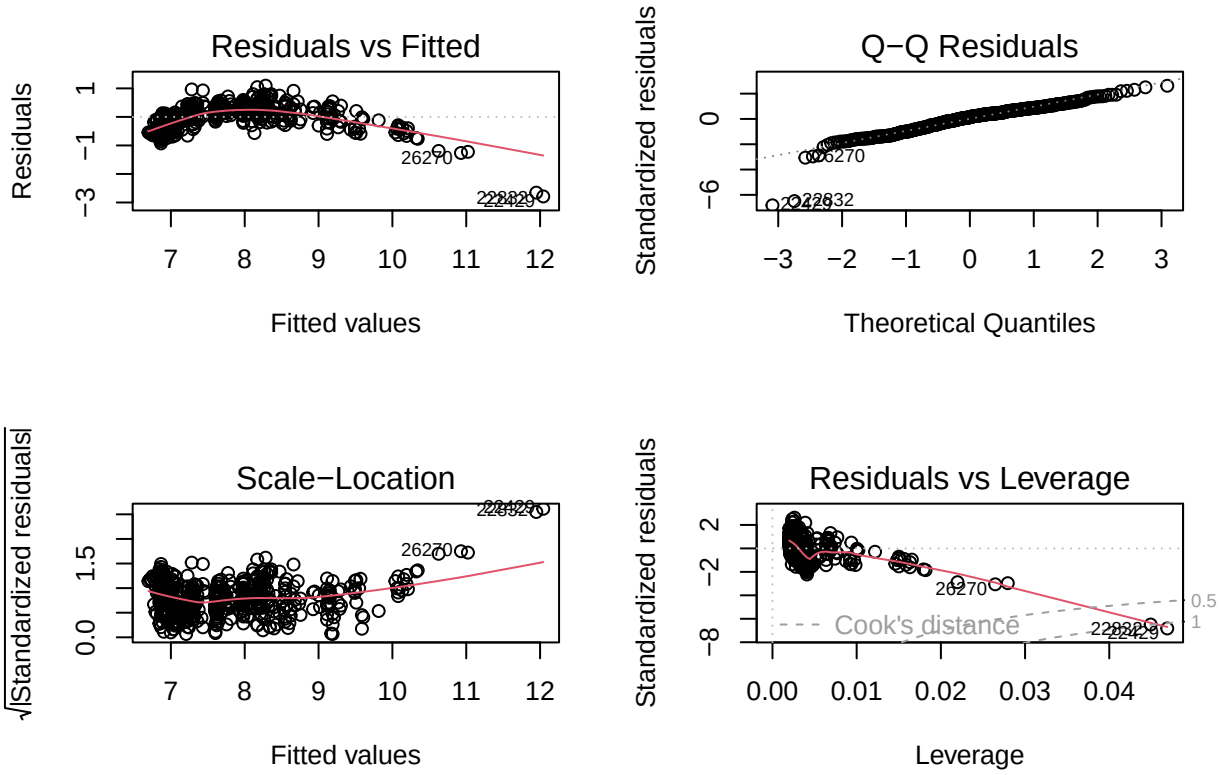
2.2 Model Transformation

To address the issue of heteroscedasticity, we applied a logarithmic transformation to price. The transformed model was:

$\log(price) = \beta_0 + \beta_1 carat + \varepsilon_i$

This transformation aimed to stabilize the variance of the residuals and improve model fit.

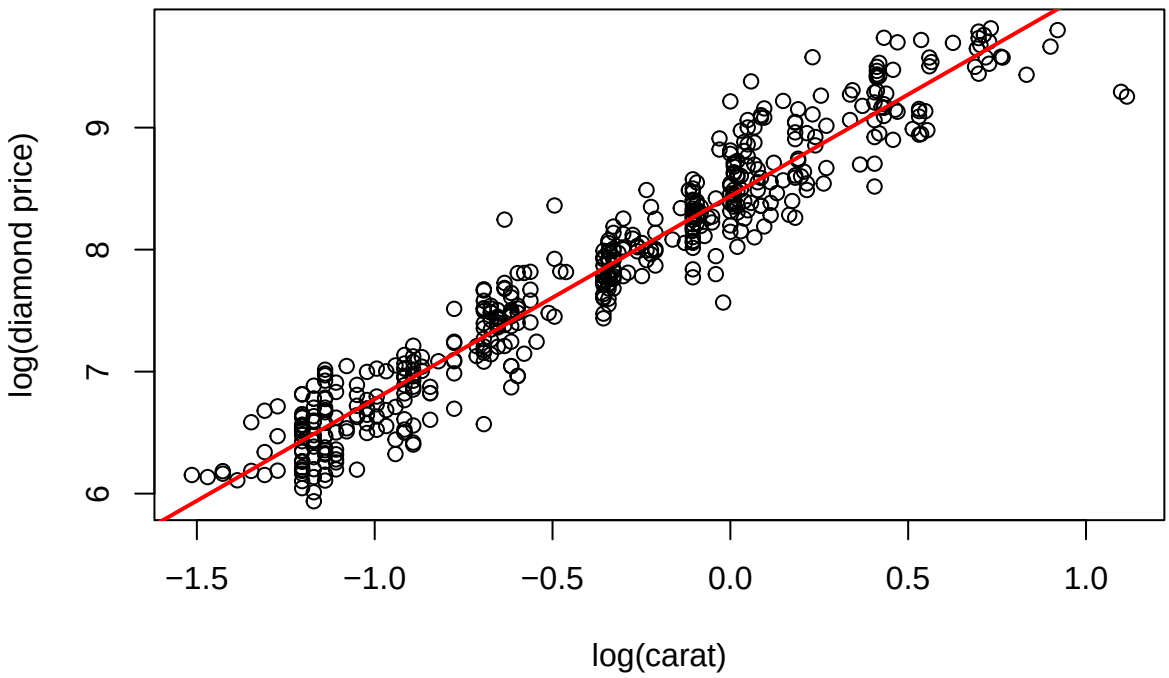

log(diamond price)
6
7
8
9
0.5
1.0
1.5
2.0
2.5
3.0
carat

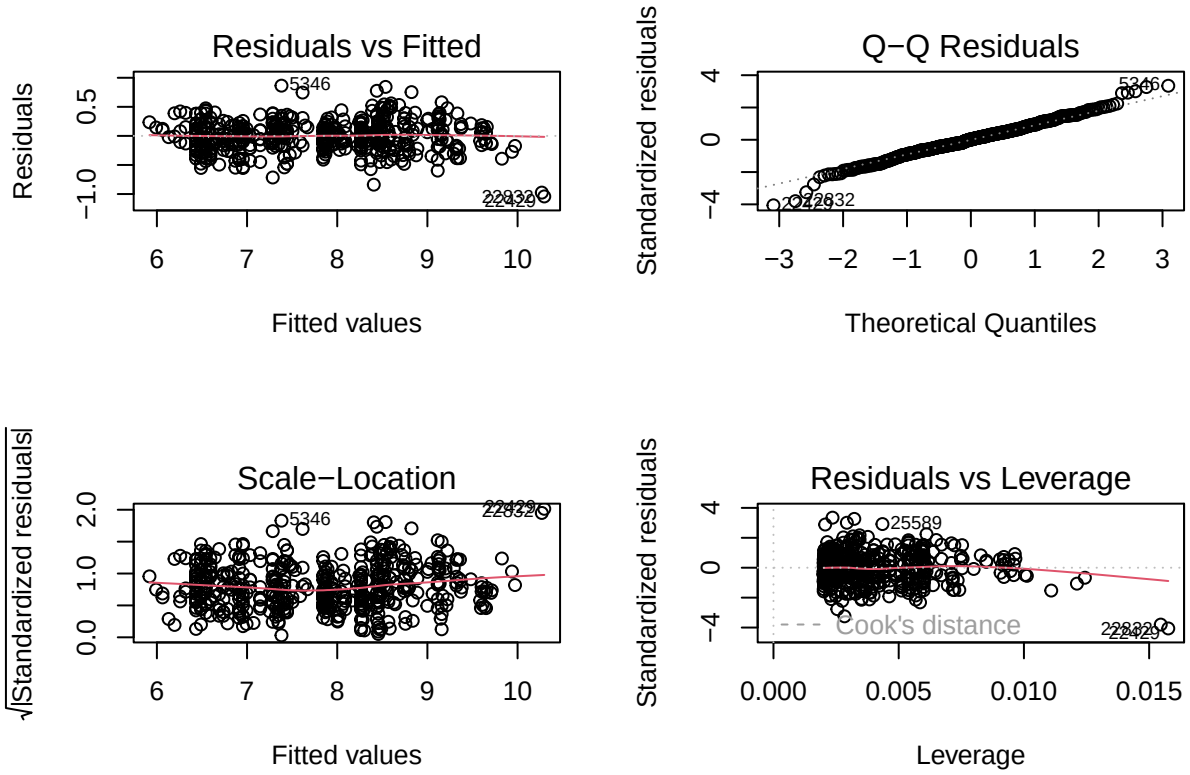


From the plot of log(diamond price) vs carat, we found out that they didn't have a linear relationship. To address the issue, we further applied a logarithmic transformation to carat. The transformed model was:

$\log(price) = \beta_0 + \beta_1 log(carat) + \varepsilon_i$

This transformation aimed to form a linear relationship between our DV and IV.


log(diamond price)
6
7
8
9
−1.5
−1.0
−0.5
0.0
0.5
1.0
log(carat)



- Linearity: The residuals vs fitted plot indicates linearity.
- Expected Value of Residuals: The expected value of the residuals was approximately zero, satisfying the requirement $\mathrm{E}(\varepsilon_i) = 0$.
- Homoscedasticity: The scale-location plot reveals no heteroscedasticity, indicating that the variance of the residuals is constant.
- Independence: The residuals do not show any apparent pattern, implying that the independence assumption holds.
- Normality: The Q-Q plot shows some deviation from normality, especially in the tails.

Transformed Model

```
##
## Call:
## lm(formula = price ~ carat, data = diamond_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04145 -0.15848  0.00422  0.15539  0.86404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.43911    0.01400  602.85   <2e-16 ***
## carat        1.66525    0.02013   82.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2586 on 498 degrees of freedom
## Multiple R-squared:  0.9321, Adjusted R-squared:  0.932
## F-statistic:  6841 on 1 and 498 DF,  p-value: < 2.2e-16
```

The transformed model's summary statistics show:

- Coefficients: The coefficient for 'log_carat' remains significant.
- R^2 : The adjusted R^2 increased from 0.8375 in the original model to 0.932 in the model after transformation, indicating improvement in the model fit.
- Residual Standard Error: The residual standard error has decreased, suggesting a better fit.
- Diagnostic Plots: The diagnostic plots of the transformed model show improved normality and homoscedasticity.

Transformed Model Diagnostics

```
##             Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 8.439106 0.01399862 602.85281  0.000000e+00
## carat       1.665252 0.02013370  82.70969 4.329134e-293
```

```
##                2.5 %   97.5 %
## (Intercept) 8.411602 8.466609
## carat       1.625694 1.704809
```

The model from this 500 data sample is:

$\log(price) = 8.439106 + 1.665252 log(carat) + \varepsilon_i$

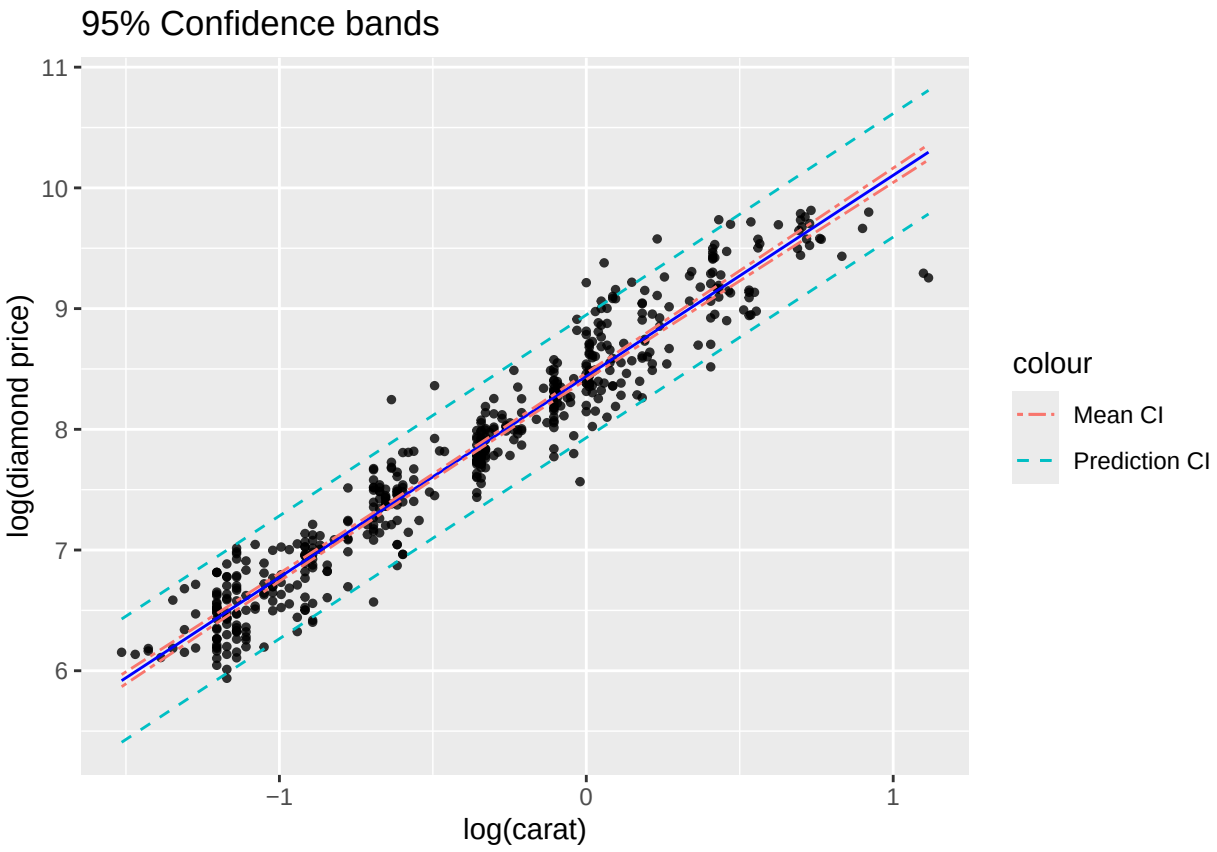
After the logarithmic transformation, the model diagnostics were revisited:

- Expected Value of Residuals: The expected value of the residuals remained approximately zero.
- Normality: The residuals were normally distributed as indicated by the Q-Q plot.
- Homoscedasticity: The residual plot showed no apparent pattern, suggesting that the issue of heteroscedasticity was resolved.

Confidence Intervals:

The 95% confidence intervals for the model coefficients are as follows:

- Intercept: [8.411602, 8.466609]
- log(carat): [1.625694, 1.704809]



The transformed regression model, with both price and carat logged, satisfied all the key assumptions of linear regression, making it a robust model for predicting price based on carat. The final model can be represented as: $\log(\text{price}) = \beta_0 + \beta_1 \cdot \log(\text{carat}) + \epsilon$

2.3 Initial MLR Model Fitting

Started from the simple linear model obtained above, we incrementally added four remaining variables into it, and compare the R-square of each models to identify the best-performing model:

```
mfit_adj <- lm(formula = price ~ carat, data = diamond_2)
summary(mfit_adj)$adj.r.squared
```

```
## [1] 0.9320061
```

```
mfit_adj_1 <- lm(formula = price ~ carat + color, data = diamond_2)
summary(mfit_adj_1)$adj.r.squared
```

```
## [1] 0.9421566
```

```
mfit_adj_2 <- lm(formula = price ~ carat + color + clarity, data = diamond_2)
summary(mfit_adj_2)$adj.r.squared
```

```
## [1] 0.9805211
```

```r
mfit_adj_3 <- lm(formula = price ~ carat + color + clarity + table, data = diamond_2)
summary(mfit_adj_3)$adj.r.squared
```

```
## [1] 0.9807495
```

```r
mfit_adj_4 <- lm(formula = price ~ carat + color + clarity + table + z, data = diamond_2)
summary(mfit_adj_4)$adj.r.squared
```

```
## [1] 0.9807312
```

After evaluating each model, we observed that the addition of 'z' to the model resulted in a very slight decrease in the adjusted R^2 by 0.0001. This small marginal change suggests that 'z' does not significantly affect the model's explanatory power, and such a decrement further confirms our claims to remove 'z' .

Therefore, 'z' was excluded from the final multiple linear regression model. Assessing the adjusted R^2 values at each step ensured a robust model. Given the negligible contribution of 'z', it was excluded, ensuring that the final model remains both parsimonious and effective in predicting 'price' using the selected variables: 'carat', 'color', 'clarity', and 'table'.

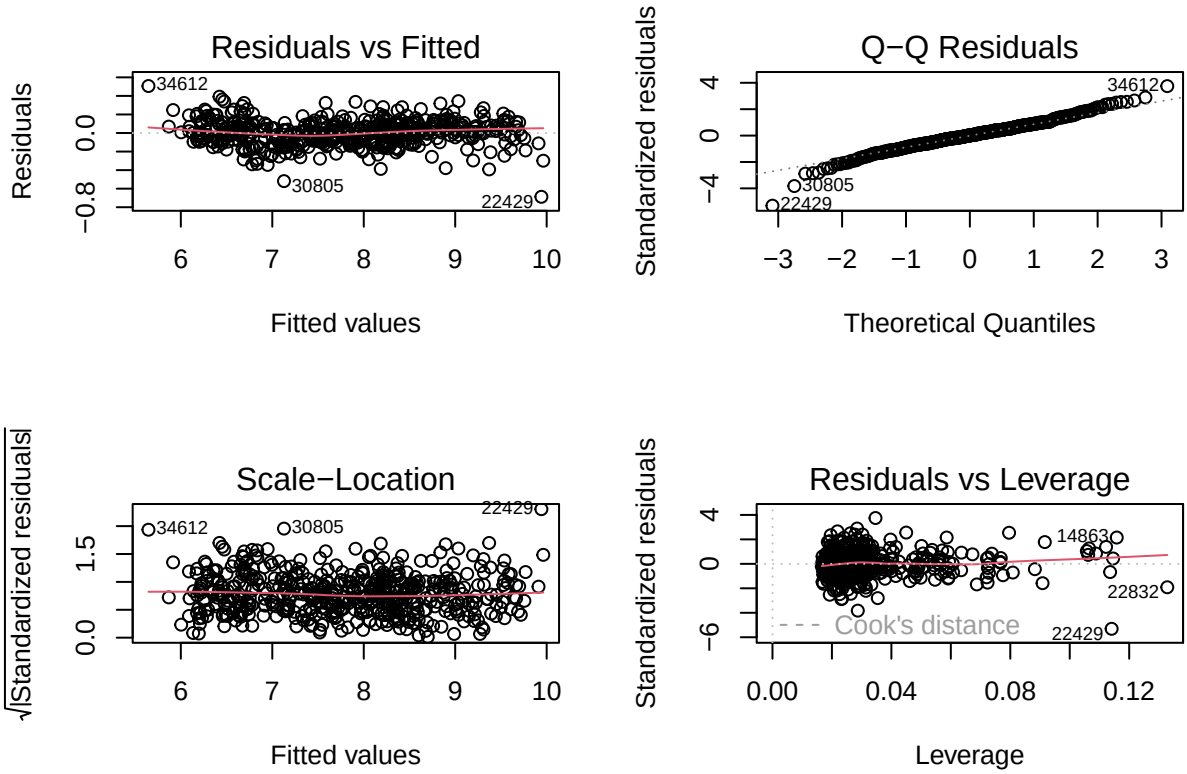
2.4 Multicollinearity Assessing

```
##             GVIF Df GVIF^(1/(2*Df))
## carat   1.408382  1        1.186753
## color   1.316583  6        1.023185
## clarity 1.391121  7        1.023859
## table   1.080964  1        1.039694
```

The results show that all predictor variables ('carat', 'color', 'clarity', and 'table') have adjusted GVIF values significantly below the threshold of 10. This indicates that there is low multicollinearity among the predictor variables in the model. Therefore, we can conclude that multicollinearity is not a concern in this multiple linear regression model, and the coefficients can be interpreted with confidence.

Part 3: MLR Observation

3.1 Diagnostic Plots & Coefficients and Confidence Intervals




```
##
## Call:
## lm(formula = price ~ carat + color + clarity + table, data = diamond_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68771 -0.08115  0.00147  0.07862  0.50652
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.325051   0.175447  47.451  < 2e-16 ***
## carat        1.888308   0.012714 148.526  < 2e-16 ***
## colorE      -0.052474   0.021674  -2.421  0.01584 *
## colorF      -0.067624   0.022036  -3.069  0.00227 **
## colorG      -0.182627   0.022837  -7.997 9.47e-15 ***
## colorH      -0.255382   0.022898 -11.153  < 2e-16 ***
## colorI      -0.380033   0.026675 -14.247  < 2e-16 ***
## colorJ      -0.545596   0.034967 -15.603  < 2e-16 ***
## clarityIF    1.197452   0.057150  20.953  < 2e-16 ***
## claritySI1   0.698908   0.046409  15.060  < 2e-16 ***
## claritySI2   0.541614   0.046496  11.649  < 2e-16 ***
```

```
## clarityVS1   0.946566   0.048016  19.713  < 2e-16 ***
## clarityVS2   0.829756   0.046329  17.910  < 2e-16 ***
## clarityVVS1  1.103261   0.053741  20.529  < 2e-16 ***
## clarityVVS2  1.028560   0.049631  20.724  < 2e-16 ***
## table       -0.007516   0.002892  -2.599  0.00964 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1376 on 484 degrees of freedom
## Multiple R-squared:  0.9813, Adjusted R-squared:  0.9807
## F-statistic:  1696 on 15 and 484 DF,  p-value: < 2.2e-16
```

Diamond_2 here is the copy of the original 500 data sample that tranformed the price to log(price) and carat to log(carat).

The multiple linear regression model effectively uses 'carat', 'color', 'clarity', and 'table' to explain the variance in diamond 'price'. The diagnostic plots suggest that the assumptions of linear regression are reasonably met, although slight deviations and potential influential points should be considered. The coefficients and their confidence intervals indicate the significant impact of these predictors on diamond price, with 'carat' being the most influential factor, followed by 'clarity' and 'color'. The 'table' variable has a minor negative effect on price.

3.2 Confirm Best Model

The stepwise regression analysis confirms that the best model for predicting diamond price includes all the initial predictors: 'carat', 'color', 'clarity', and 'table'. None of these predictors should be removed, as their exclusion increases the AIC, indicating a worse model fit.

```
## Start:  AIC=-1967.64
## price ~ carat + color + clarity + table
##
##           Df Sum of Sq    RSS      AIC
## <none>                   9.16 -1967.64
## - table    1      0.13   9.29 -1962.71
## - color    6      8.70  17.87 -1645.79
## - clarity  7     18.34  27.50 -1432.13
## - carat    1    417.70 426.87   -49.07
```

```
##
## Call:
## lm(formula = price ~ carat + color + clarity + table, data = diamond_2)
##
## Coefficients:
## (Intercept)        carat       colorE       colorF       colorG       colorH
##    8.325051     1.888308    -0.052474    -0.067624    -0.182627    -0.255382
##      colorI       colorJ    clarityIF   claritySI1   claritySI2   clarityVS1
##   -0.380033    -0.545596     1.197452     0.698908     0.541614     0.946566
##  clarityVS2  clarityVVS1  clarityVVS2        table
##    0.829756     1.103261     1.028560    -0.007516
```

This model ensures that the predictors included provide the best explanatory power for the price of diamonds, as indicated by the lowest AIC value.

Cross Validation for SLR and MLR

In order to evaluate the performance of each model on unseen data, the diamond dataset is divided into training and testing subsets. In other words, models will be trained on the training data and resultant models are applied to the testing sets. We use a random sample that include 80% of the whole data as the training set and the rest of the data is the testing set. During the test stage, the value of dependent variable is predicted, and the model accuracy is calculated on the basis of prediction error:

- R-square
- Root Mean Squared Error(RMSE)
- Mean Absolute Error(MAE)

```
#loading data
library(caret)
```

```
##      lattice
```

```
library(lattice)
set.seed(126)

#generating the training and testing data sets
diamond_test <- read.csv("Diamonds Prices2022.csv")
diamond_test$cut <- as.factor(diamond_test$cut)
diamond_test$color <- as.factor(diamond_test$color)
diamond_test$clarity <- as.factor(diamond_test$clarity)
diamond_test$price <- log(diamond_test$price)
diamond_test$carat <- log(diamond_test$carat)
diamond_test <- diamond_test[,-1]
random_sample <- sample(nrow(diamond_test), nrow(diamond_test) * 0.8)
training_dataset  <- diamond_test[random_sample, ]
testing_dataset <- diamond_test[-random_sample, ]

#simple linear regression model
slr_model <- lm(formula = price ~ carat, data = training_dataset)
slr_predictions <- predict(slr_model, testing_dataset)
#computing the performance metrics
data.frame( R2 = R2(slr_predictions, testing_dataset$price),
            RMSE = RMSE(slr_predictions, testing_dataset$price),
            MAE = MAE(slr_predictions, testing_dataset$price))
```

```
##          R2      RMSE       MAE
## 1 0.9329815 0.2650388 0.2062229
```

```
#multiple linear regression model
mlr_model <- lm(formula = price ~ carat + color + clarity + table, data = training_dataset)
mlr_predictions <- predict(mlr_model, testing_dataset)
data.frame( R2 = R2(mlr_predictions, testing_dataset$price),
            RMSE = RMSE(mlr_predictions, testing_dataset$price),
            MAE = MAE(mlr_predictions, testing_dataset$price))
```

```
##          R2      RMSE       MAE
## 1 0.9818737 0.1378395 0.1069274
```

The comparison between the simple linear regression model and the multiple linear regression model shows that the MLR model performs significantly better. The MLR model, which includes 'carat', 'color', 'clarity', and 'table' as predictors, has a much higher R-square value(0.9818737) and lower RMSE(0.1378395) and MAE(0.1069274) values, indicating that it provides a much more accurate and reliable prediction of diamond prices.

This analysis confirms that adding additional variables ('color', 'clarity', and 'table') improves the explanatory power of the model significantly, compared to using 'carat' alone. Therefore, the multiple linear regression model should be preferred for predicting diamond prices.

Decide on the Final Model

The stepwise regression analysis confirms that the best model for predicting diamond price includes all the initial predictors: 'carat', 'color', 'clarity', and 'table', that is:

Best Model: `log(price) ~ log(carat) + color + clarity + table`

None of these predictors should be removed, as their exclusion increases the AIC, indicating a worse model fit.

```
## Start:  AIC=-1967.64
## price ~ carat + color + clarity + table
##
##           Df Sum of Sq    RSS      AIC
## <none>                   9.16 -1967.64
## - table    1      0.13   9.29 -1962.71
## - color    6      8.70  17.87 -1645.79
## - clarity  7     18.34  27.50 -1432.13
## - carat    1    417.70 426.87   -49.07
```

```
##
## Call:
## lm(formula = price ~ carat + color + clarity + table, data = diamond_2)
##
## Coefficients:
## (Intercept)        carat       colorE       colorF       colorG       colorH
##    8.325051     1.888308    -0.052474    -0.067624    -0.182627    -0.255382
##      colorI       colorJ    clarityIF   claritySI1   claritySI2   clarityVS1
##   -0.380033    -0.545596     1.197452     0.698908     0.541614     0.946566
##  clarityVS2  clarityVVS1  clarityVVS2        table
##    0.829756     1.103261     1.028560    -0.007516
```

Diamond_2 here is the copy of the original 500 data sample that tranformed the price to log(price) and carat to log(carat).

This model ensures that the predictors included provide the best explanatory power for the price of diamonds, as indicated by the lowest AIC value.

3.3 CI and PI for one combination of the IV

Here is the CI and PI for a diamond that has a size of 1 carat, a color level of I, clarity level of VS2, and a table of 60 mm.

```
x_combination <- data.frame(carat = 1, color = "I", clarity = "VS2", table = 60)
x_combination$carat <- log(x_combination$carat)
x_combination$table <- log(x_combination$table)
p3 = predict(mfit, x_combination, se.fit = TRUE, interval = "confidence", level = 0.95)
p4 = predict(mfit, x_combination, se.fit = TRUE, interval = "prediction", level = 0.95)
p3$fit
```

```
##     fit     lwr     upr
## 1 8.744 8.43496 9.05304
```

```
p4$fit
```

```
##     fit      lwr     upr
## 1 8.744 8.333381 9.15462
```

The 95% confidence intervals for the log of the price of the diamond that has a size of 1 carat, a color level of I, clarity level of VS2, and a table of 60 mm is [8.43496, 9.05304]

The 95% prediction intervals for the log of the price of the diamond that has a size of 1 carat, a color level of I, clarity level of VS2, and a table of 60 mm is [8.333381, 9.15462]

3.4 Observations

Comment on interesting findings during this part:

1. The initial simple linear regression showed a strong relationship between 'carat' and 'price'. However, diagnostic plots indicated the violations of model assumptions and the need for transformation.
2. Log transformations of 'price' and 'carat' improved the model fit and addressed the issues of heteroscedasticity and normality of residuals.
3. Including additional categorical variables like 'color' and 'clarity' further improved the model, significantly increasing the R^2 value and providing a more comprehensive understanding of the factors affecting diamond prices.
4. The analysis revealed the importance of multiple factors in determining diamond prices, with 'carat' being the most influential predictor, followed by 'color', 'clarity' and 'table'.

3.5 Summary

Model Building and Evaluation

$\log(price) = \beta_0 + \beta_1 log(carat) + \varepsilon_i$

```
1. **Simple Linear Regression (SLR)**:
•   Model: log(price) ~ log(carat)
•   Performance:
   $R^2$ : 0.9329815
   RMSE: 0.2650388
   MAE: 0.2062229
•   Findings: The SLR model showed a strong relationship between carat and price, explaining 93.3% of tl
```

$$\log(price) = \beta_0 + \beta_1 log(carat) + \beta_2 I_{color=E} + \beta_3 I_{color=I} + \beta_4 I_{color=J} + \beta_5 I_{color=H} + \beta_6 I_{color=F} + \beta_7 I_{color=G} + \beta_8 I_{color=D} + \beta_9 I_{clarity=I1} + \beta_{10} I_{clarity=VS2} + \beta_{11} I_{clarity=SI1} + \beta_{12} I_{clarity=VVS1} + \beta_{13} I_{clarity=VVS2} + \beta_{14} I_{clarity=VS1} + \beta_{15} I_{clarity=SI2} + \beta_{16} I_{clarity=IF} + \beta_{16} table + \varepsilon_i$$

2. **Multiple Linear Regression (MLR)**:
- Model: log(price) ~ log(carat) + color + clarity + table
- Performance:
  R^2 : 0.9818737
  RMSE: 0.1378395
  MAE: 0.1069274
- Findings: The MLR model significantly improved performance, explaining 98.2% of the variance in dia

Stepwise Regression and Model Selection

Stepwise regression using AIC confirmed that the best model includes 'carat', 'color', 'clarity', and 'table'. Removing any of these predictors increased the AIC, indicating a worse model fit.

Key Observations

1. Impact of Carat: Carat weight remains the most influential predictor of diamond prices.
2. Role of Color and Clarity: Lower color grades decrease prices, while higher clarity grades increase
3. Minor Role of Table: Table percentage has a minor but significant negative impact on price.
4. Model Robustness: Diagnostic checks confirmed that the MLR model is robust and reliable.

Conclusion

The comprehensive analysis revealed the multifaceted nature of diamond pricing, with carat, color, and clarity being the most critical factors. The final MLR model, including 'carat', 'color', 'clarity', and 'table', provides an excellent fit to the data, making it a reliable tool for predicting diamond prices. This model can be effectively used for accurate price predictions and offers valuable insights for the diamond industry.