B105 Applied Statistical Modelling Final Assessment

GH1016532 Youyoung Seo

Diamond Pricing Analysis

1. Introduction

As a jewelry enthusiast I have always been fascinated by the way diamonds are priced and valued. This fascination along with my data analysis profession motivated me to dig into the determinants of diamond prices. I am analyzing actual data to determine how attributes such as carat, cut, clarity and color combine so as to value a diamond in the market.

The goal of this analysis is to answer the following business question. What are the factors which affect diamond price most and how can we use this information to model/optimize our pricing strategy? Going from there, I came up with three business hypotheses:

Hypothesis 1: Carat is positively correlated with the price of the diamond.

Hypothesis 2: The price rises or falls with the change of cut quality.

Hypothesis 3: The price of the diamond can be predicted using a linear regression model considering multiple attributes like the carat, cut, clarity and color based on the X,Y,Z dimensions.

I am working on this project to give jewelers targeted information that can help them price diamonds better based on their characteristics.

2. Data Loading and Cleaning

Before we can analyze data, our first steps are always to clean the dataset, get rid of errors, missing values etc.

Here, I have to check whether all data is available and give clear names. I also removed all rows containing any missing information, a step that ensures our dataset is as clean and complete as possible for downstream analysis.

I used to read csv() to load the dataset and renamed columns associated with diamond dimensions (x, y, z) to length, width and depth_mm respectively for ease of understanding.

2-1. Load required libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(car)
```

```
## Loading required package: carData
##
```

```
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(corrplot)
```

```
## corrplot 0.94 loaded
```

2-2. Load the dataset and display few columns

```
getwd()
```

```
## [1] "/Users/youyoungseo/Desktop"
```

```
setwd("/Users/youyoungseo/Downloads/")
diamonds_data <- read.csv("/Users/youyoungseo/Downloads/Diamonds Prices2022.csv")

head(diamonds_data)
```

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

2-3. Checking the column names

```
colnames(diamonds_data)
```

```
##  [1] "X"       "carat"   "cut"     "color"   "clarity" "depth"   "table"
##  [8] "price"   "x"       "y"       "z"
```

2-4. Renaming the columns for clarity

```
diamonds_data_cleaned <- diamonds_data %>%
  rename(
    length = x,
    width = y,
    depth_mm = z
  )
```

2-5. Checking missing values and cleaning the dataset

I next checked for missing values and using the na. Removing incomplete entries with omit().

```r
colSums(is.na(diamonds_data_cleaned))
```

```
##        X    carat      cut    color  clarity    depth    table    price
##        0        0        0        0        0        0        0        0
##   length    width depth_mm
##        0        0        0
```

```r
diamonds_data_cleaned <- na.omit(diamonds_data_cleaned)
summary(diamonds_data_cleaned)
```

```
##        X             carat            cut               color
##  Min.   :    1   Min.   :0.2000   Length:53943       Length:53943
##  1st Qu.:13486   1st Qu.:0.4000   Class :character   Class :character
##  Median :26972   Median :0.7000   Mode  :character   Mode  :character
##  Mean   :26972   Mean   :0.7979
##  3rd Qu.:40458   3rd Qu.:1.0400
##  Max.   :53943   Max.   :5.0100
##    clarity              depth           table           price
##  Length:53943       Min.   :43.00   Min.   :43.00   Min.   :  326
##  Class :character   1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  950
##  Mode  :character   Median :61.80   Median :57.00   Median : 2401
##                     Mean   :61.75   Mean   :57.46   Mean   : 3933
##                     3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5324
##                     Max.   :79.00   Max.   :95.00   Max.   :18823
##      length           width           depth_mm
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
##  Median : 5.700   Median : 5.710   Median : 3.530
##  Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
##  3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
##  Max.   :10.740   Max.   :58.900   Max.   :31.800
```

3. Exploratory Data Analysis (EDA)

EDA is a substantial exploratory data analysis to plot the correlation among the variables in visually appealing behind science form. That in turn, drives discussion and influences pricing patterns between nearby/similar diamonds.

3-1. Scatterplot of Carat vs Price of the diamond

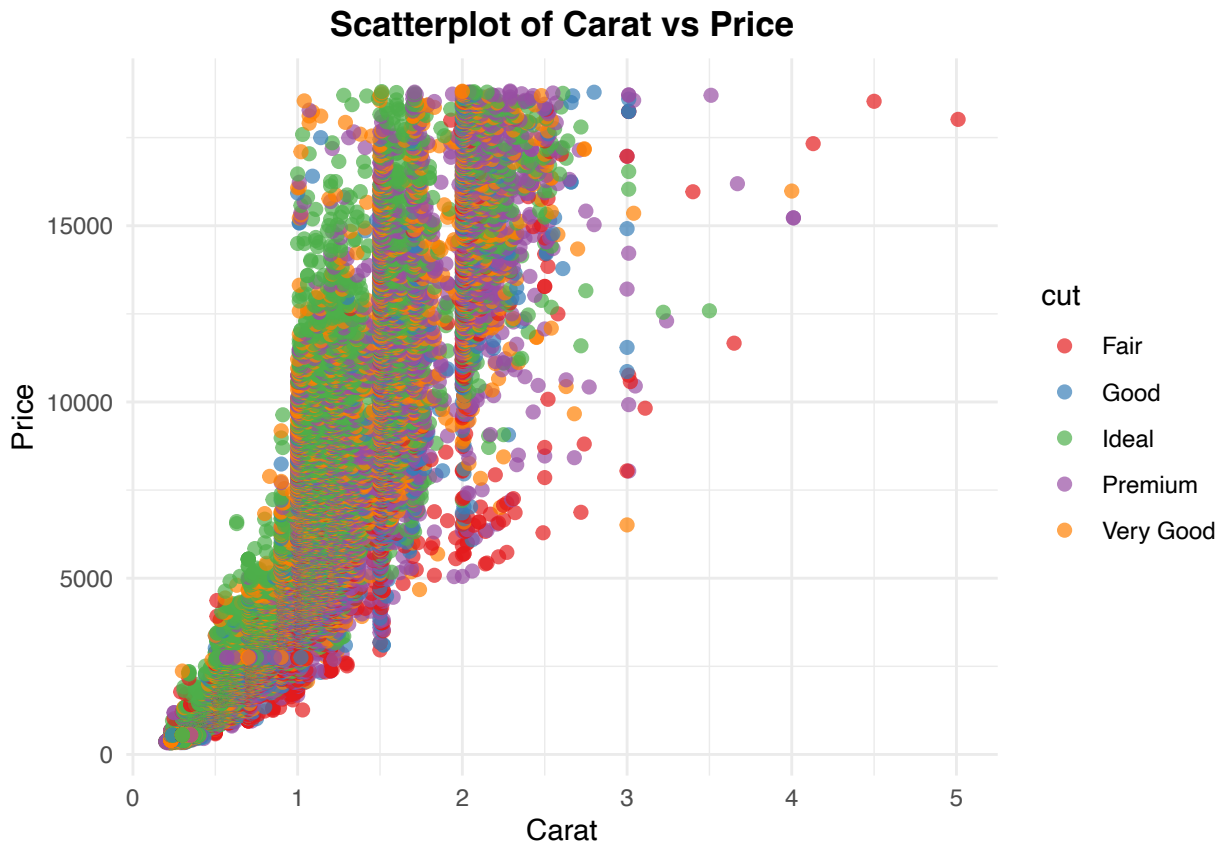
I made a scatterplot to analyze the relationship between carat and price, adding cut quality as a variable distinguished by color.

It shows that price and carat are positively related: the larger the diamond, the more expensive it is. Moreover, this relationship is visible in all categories of cut quality.

Thus, this graph supports Hypothesis 1.

```r
ggplot(diamonds_data_cleaned, aes(x = carat, y = price, color = cut)) +
  geom_point(alpha = 0.7, size = 2) +
  scale_color_brewer(palette = "Set1") +
  ggtitle("Scatterplot of Carat vs Price") +
  xlab("Carat") + ylab("Price") +
```

```
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



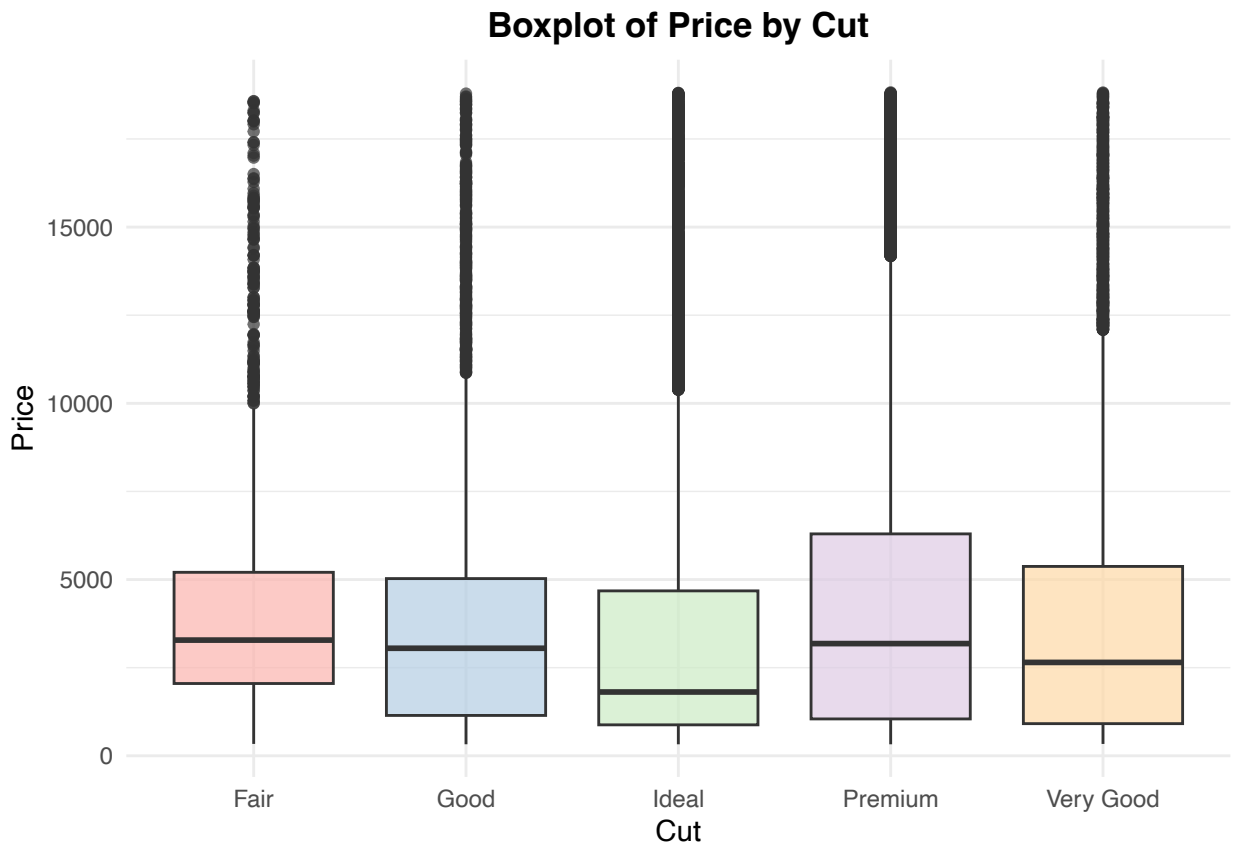
3-2. Boxplot of **Price** by **Cut**

I created a boxplot to check if the median price of diamond differs depending on cut quality.

Category "Ideal" has the highest median, closely followed by "Premium" cut quality.

It is apparent that the boxplot supports Hypothesis 2.

```
ggplot(diamonds_data_cleaned, aes(x = cut, y = price, fill = cut)) +
  geom_boxplot(alpha = 0.7) +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "Boxplot of Price by Cut", x = "Cut", y = "Price") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        legend.position = "none")
```



3-3. Boxplot of **Price** by **Clarity**

I used another boxplot to examine how **clarity** affects **price**.

Easy to understand the graph, I will explain about Diamond clarity briefly.

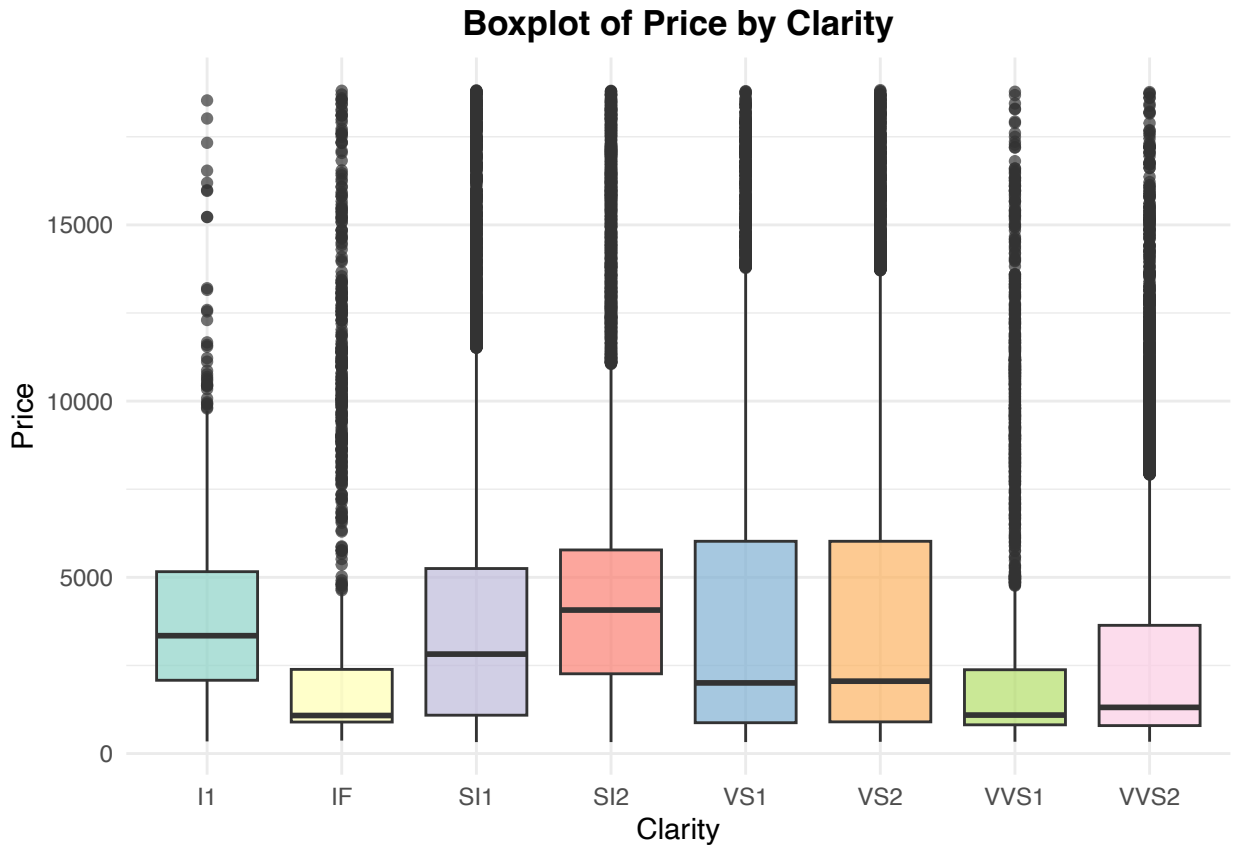
Diamond clarity is graded depending on the number, size, color, and location of internal flaws or inclusions, as well as external ones or blemishes. The lowest blemish used in the grading process is "included". There are seven classes of clarity:

- FL : Flawless. No inclusions or blemishes are visible under 10 times magnification.
- IF : Internally Flawless. No inclusions, but might have blemishes.
- VVS1, VS2 and VVS2 : Very, Very Slightly Included: hardly visible inclusions for a naked eye.
- VS1 and VS2 : Very slightly included: can be recognized at 10 times magnification
- SI1 and SI2 : Slightly Included
- I1, I2, and I3 : Included: inclusions visible to the human eye, which may affect transparency.

Diamonds with higher clarity levels (e.g., IF, VVS1) generally command higher prices, showing clarity is an important pricing factor.

```
ggplot(diamonds_data_cleaned, aes(x = clarity, y = price, fill = clarity)) +
  geom_boxplot(alpha = 0.7) +
  scale_fill_brewer(palette = "Set3") +
  labs(title = "Boxplot of Price by Clarity", x = "Clarity", y = "Price") +
```

```
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        legend.position = "none")
```



4. Hypothesis Testing

4-1. Business Hypothesis 1: Correlation Between **Carat** and **Price**

Hypothesis: There is a strong positive correlation between carat and diamond price.

I used the correlation test to check whether a positive relationship exists between carat and price significantly.

The Pearson correlation coefficient** = 0.9216, **p-value** **< 2.2e-16**, and r = **0.77**.

The relationship is strong. Hence, I find support that Hypothesis 1 is valid.

```
cor_test_result <- cor.test(diamonds_data_cleaned$carat, diamonds_data_cleaned$price)
print(cor_test_result)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diamonds_data_cleaned$carat and diamonds_data_cleaned$price
## t = 551.42, df = 53941, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.9203098 0.9228529
## sample estimates:
##       cor
## 0.9215913
```

4-2. Business Hypothesis 2: **Price Differences** by **Cut**

Hypothesis: There are significant differences in price across different cut qualities.

I run an ANOVA test to see if the quality of the cut significantly affects the price.

The result of the ANOVA test: F(4, 53938) = 175.6, p-value < 2e-16 confirmed significant differences in price by cut.

Hence, the test confirmed the hypothesis and the quality of the cut significantly affects the price.

```
anova_cut <- aov(price ~ cut, data = diamonds_data_cleaned)
summary(anova_cut)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## cut             4 1.104e+10 2.759e+09   175.6 <2e-16 ***
## Residuals   53938 8.474e+11 1.571e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4-3. Business Hypothesis 3: **Predicting Price** with Linear Regression

Hypothesis: A linear regression model can predict diamond prices based on carat, cut, clarity, and other attributes.

I built a linear regression model to predict diamond prices using multiple attributes.

The model's R-squared value was 0.9202, meaning it explained approximately 92% of the variance in diamond prices. Most predictors were significant, with carat having the largest effect on price.

The high R-squared value supports Hypothesis 3, meaning the model can accurately predict diamond prices based on key attributes.

```
diamonds_data_cleaned$cut <- factor(diamonds_data_cleaned$cut)
diamonds_data_cleaned$color <- factor(diamonds_data_cleaned$color)
diamonds_data_cleaned$clarity <- factor(diamonds_data_cleaned$clarity)

reg_model <- lm(price ~ carat + cut + clarity + color + depth +
                  table + length + width + depth_mm, data = diamonds_data_cleaned)

summary(reg_model)
```

```
##
## Call:
## lm(formula = price ~ carat + cut + clarity + color + depth +
##     table + length + width + depth_mm, data = diamonds_data_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21376.3   -592.4   -183.4    376.5  10694.3
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2183.404    408.175   5.349 8.87e-08 ***
## carat        11257.127     48.626 231.504  < 2e-16 ***
## cutGood        579.792     33.591  17.260  < 2e-16 ***
## cutIdeal       832.975     33.406  24.935  < 2e-16 ***
## cutPremium     762.161     32.227  23.650  < 2e-16 ***
## cutVery Good   726.771     32.240  22.543  < 2e-16 ***
## clarityIF     5345.101     51.023 104.759  < 2e-16 ***
## claritySI1    3665.451     43.633  84.006  < 2e-16 ***
## claritySI2    2702.611     43.818  61.679  < 2e-16 ***
## clarityVS1    4578.415     44.545 102.782  < 2e-16 ***
## clarityVS2    4267.181     43.852  97.308  < 2e-16 ***
## clarityVVS1   5007.771     47.159 106.190  < 2e-16 ***
## clarityVVS2   4950.832     45.854 107.970  < 2e-16 ***
## colorE        -209.237     17.892 -11.694  < 2e-16 ***
## colorF        -272.877     18.092 -15.083  < 2e-16 ***
## colorG        -482.046     17.716 -27.210  < 2e-16 ***
## colorH        -980.271     18.835 -52.044  < 2e-16 ***
## colorI       -1466.251     21.162 -69.287  < 2e-16 ***
## colorJ       -2369.401     26.130 -90.676  < 2e-16 ***
## depth          -63.790      4.534 -14.068  < 2e-16 ***
## table          -26.469      2.911  -9.092  < 2e-16 ***
## length       -1008.321     32.897 -30.651  < 2e-16 ***
## width            9.606     19.332   0.497    0.619
## depth_mm       -50.123     33.486  -1.497    0.134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 53919 degrees of freedom
## Multiple R-squared:  0.9198, Adjusted R-squared:  0.9198
## F-statistic: 2.688e+04 on 23 and 53919 DF,  p-value: < 2.2e-16
```

5. Model Diagnostics and Assumptions

The purpose of this analysis is to validate the regression model used for predicting diamond prices. I performed several diagnostic checks for this: these include checking the residuals for nonlinearity, for multicollinearity with the VIF values, and identifying influential outliers in the data.

```
sapply(diamonds_data_cleaned[, c("carat", "cut", "clarity",
                                 "color", "depth", "table", "length",
                                 "width", "depth_mm")], length)
```

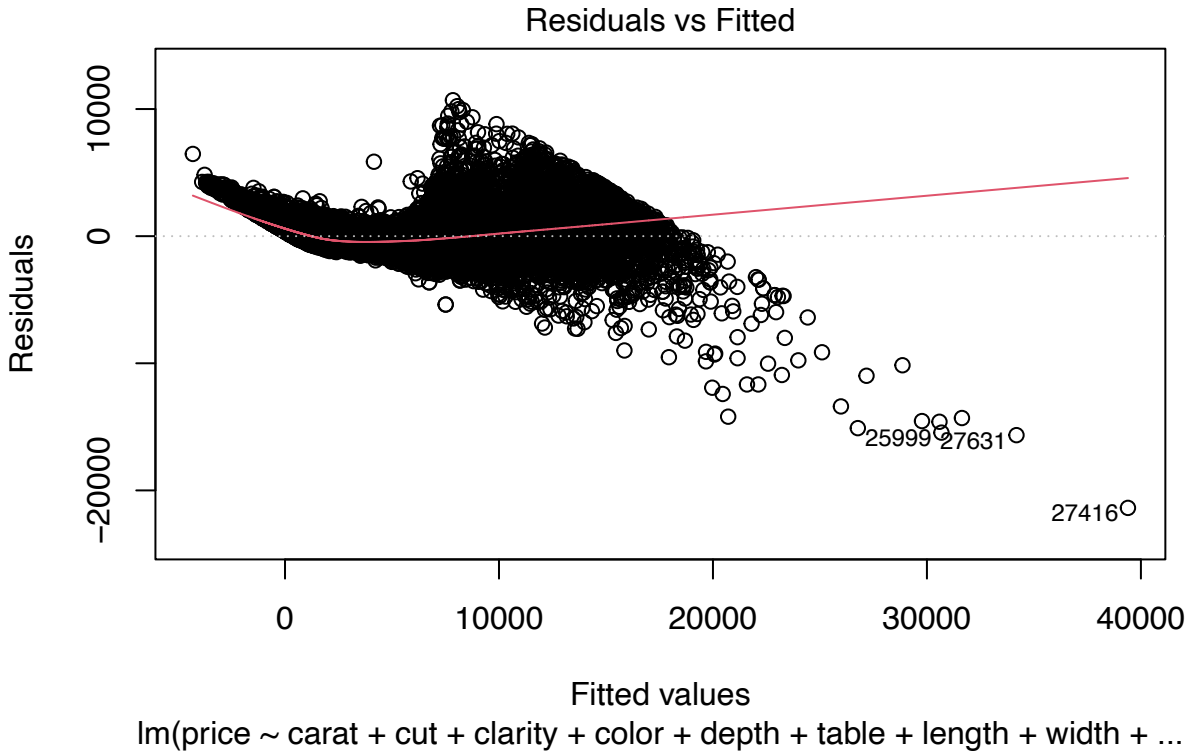
```
##    carat      cut  clarity    color    depth    table   length    width
##    53943    53943    53943    53943    53943    53943    53943    53943
## depth_mm
##    53943
```

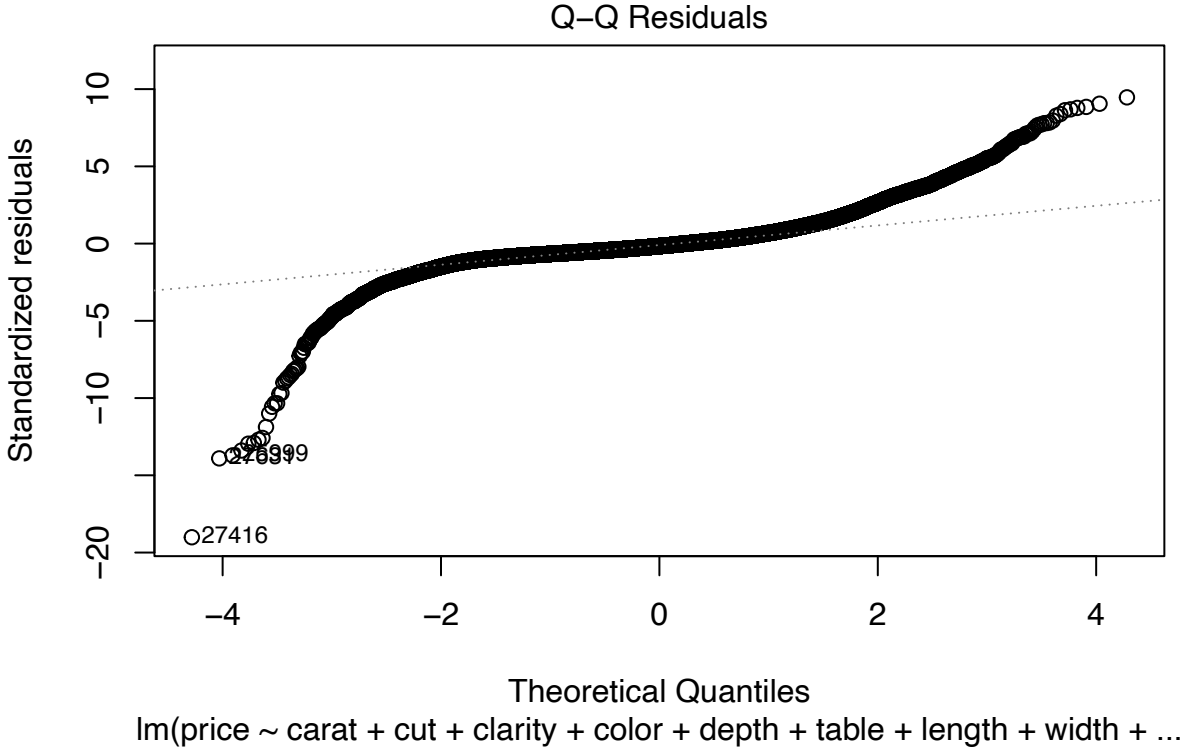
```
vif(reg_model)
```

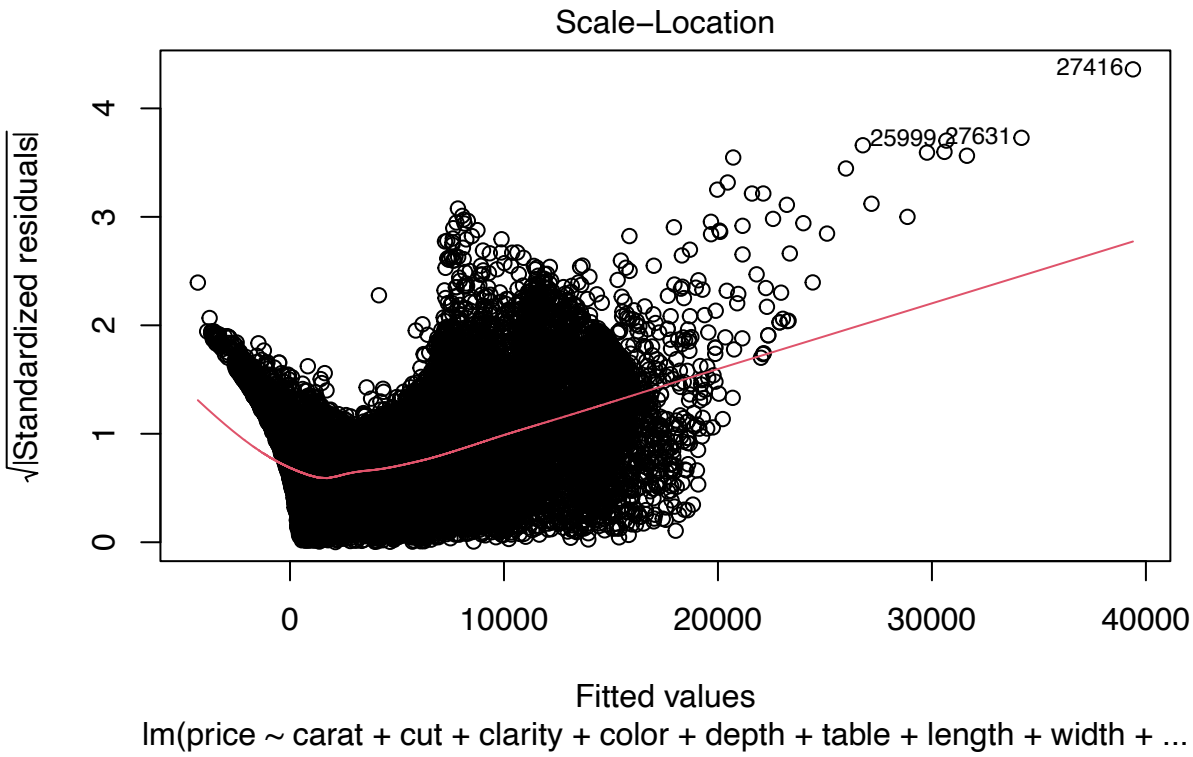
```
##               GVIF Df GVIF^(1/(2*Df))
```

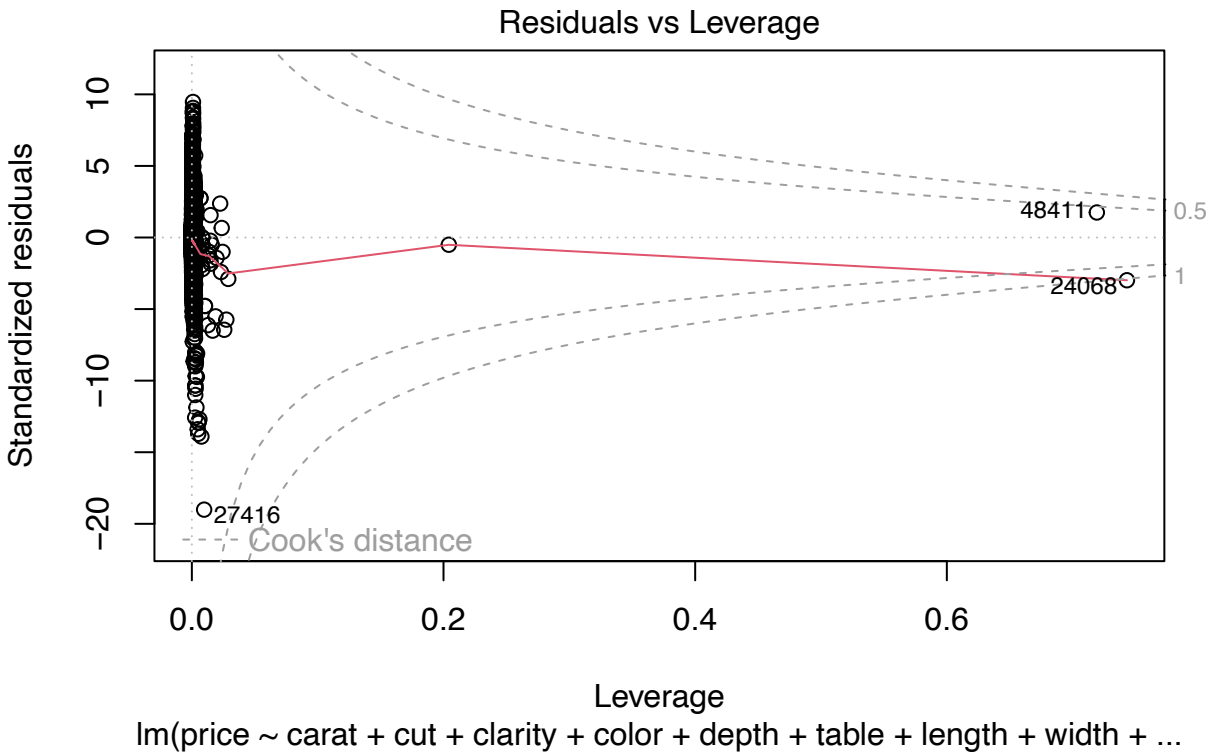
```
## carat    22.439246  1        4.737008
## cut       1.945570  4        1.086753
## clarity   1.347475  7        1.021531
## color     1.178333  6        1.013769
## depth     1.782449  1        1.335084
## table     1.787674  1        1.337039
## length   57.517927  1        7.584057
## width    20.592150  1        4.537857
## depth_mm 23.585595  1        4.856500
```

```
plot(reg_model)
```




Q–Q Residuals
Standardized residuals
10
5
0
−5
−10
−20
27416
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ carat + cut + clarity + color + depth + table + length + width + ...


Scale−Location
27416
25999
27631
√|Standardized residuals|
0
1
2
3
4
0
10000
20000
30000
40000
Fitted values
lm(price ~ carat + cut + clarity + color + depth + table + length + width + ...



What the analysis found:

I implemented all diagnostic checks and present them below:

- **Residuals vs. Fitted plot**: It is evident there is a slight nonlinearity as residuals are not quite random, but there is a concentration of these, near zero. This suggests the model captures most of the price variation.
- **Scale-Location plot**: this suggests there is heteroscedasticity, as the variance of residuals increases with the size of fitted values.
- **Q-Q plot**: this suggests the residuals mostly come from a normal distribution, but deviations appear in the tails.
- **Residuals vs. Leverage plot**: there seem to be few high-leverage points, identified by being outside of the dashed line. These may be influential outliers.
- **VIF values**: it is apparent all values suggest severe multicollinearity, especially with the carat and length.

6. Conclusion

In this project, I have successfully answered the business question about the key factors that affect prices of diamonds using a meaningful dataset.

As a result of data preparation and cleaning, as well as the use of exploratory data analysis, I have found that carat and cut are the most important factors for such determination. At the same time, using the regression model, I also confirmed Hypothesis 3, because it was demonstrated that price can be predicted

very well according to multiple attributes. With the help of diagnostics, I have also found some minor issues with multicollinearity and heteroscedasticity.

Although improvements can be made, valuable conclusions can be made for the jewelry sector. Since the price was predictable, the current approach to such a parameter is optimal.