Diamond Prices 2022

Leslie Cervantes Rivera & Valerie De La Fuente

2024-06-11

Introduction

Diamonds have traditionally been associated with luxury, romance, and wealth. Due to their allure, they are in great demand in the jewelry market and investment assets. However, the process to determining their price is based on several factors such as carat, cut, color, clarity, and etc. This report aims to examine the relationship between the explanatory variables (cut, color, length, and width of a diamond) and its effect on the diamond price.

Part 1

This report utilizes a dataset obtained from Kaggle, which includes 53,943 observations across 11 variables related to diamond prices in 2022. From this big data set, a random sample of 300 observations was selected, focusing on 5 variables. Each observation (row) corresponds to a randomly selected diamond, providing insight to its many attributes.

```r
#diamondprices <- read.csv("~/Desktop/Diamonds Prices2022.csv")
#diamond_samples <- diamondprices[sample(nrow(diamondprices), 300),]
diamond_samples <- read.csv("~/Desktop/diamond_samples.csv")
```

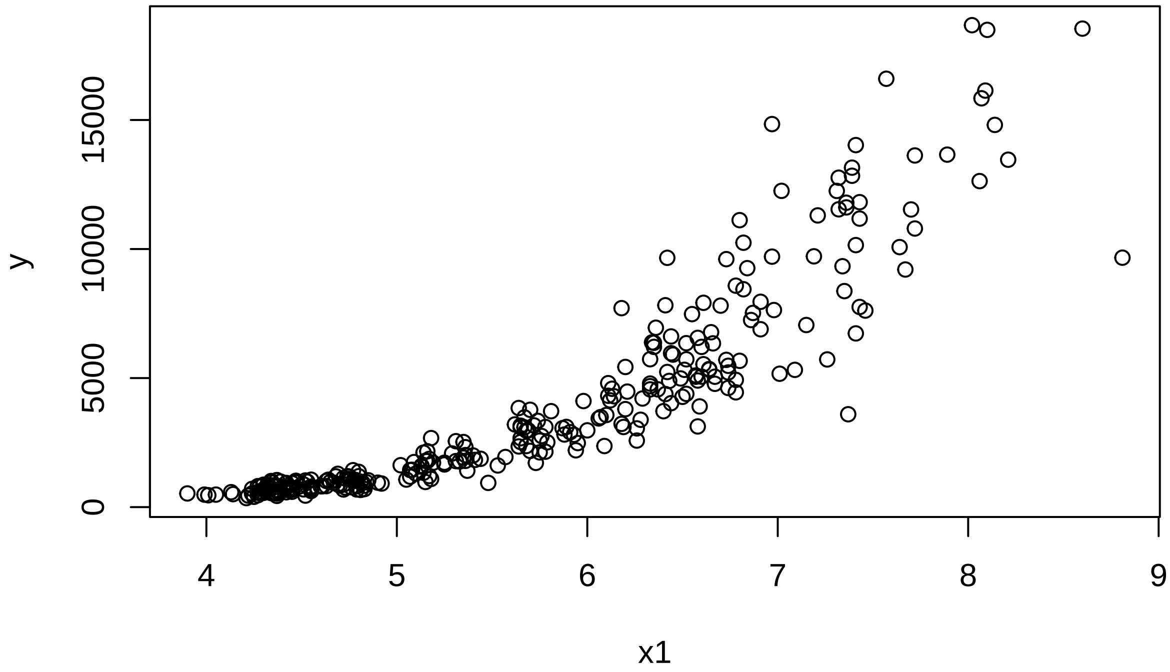
The examination revolves around two categorical independent variables: "Cut" and "Color." The "Cut" variable includes 5 input values: Fair, Good, Ideal, Premium, and Very Good, while "Color" includes 7 input values: Color D, Color E, Color F, Color G, Color H, Color I, and Color J.

Additionally, the examination incorporates two quantitative independent variables: "x" (renamed to "x1"), representing the length of the diamond in millimeters (mm), and "y" (renamed to "x2"), signifying the width of the diamond in millimeters. Together, these factors help to contribute to understanding the physical characteristics of the diamonds under examination.

Other variables that are not included in the examination are "Carat", "Clarity", "Depth", and "Table." "Carat" is used to describe the weight of diamonds. "Clarity" includes SI1, VS2, and others which describes the presence of internal and external imperfections. "Depth" represents the height of the diamond in millimeters (mm). "Table" describes the largest flat facet on the top of the diamond.

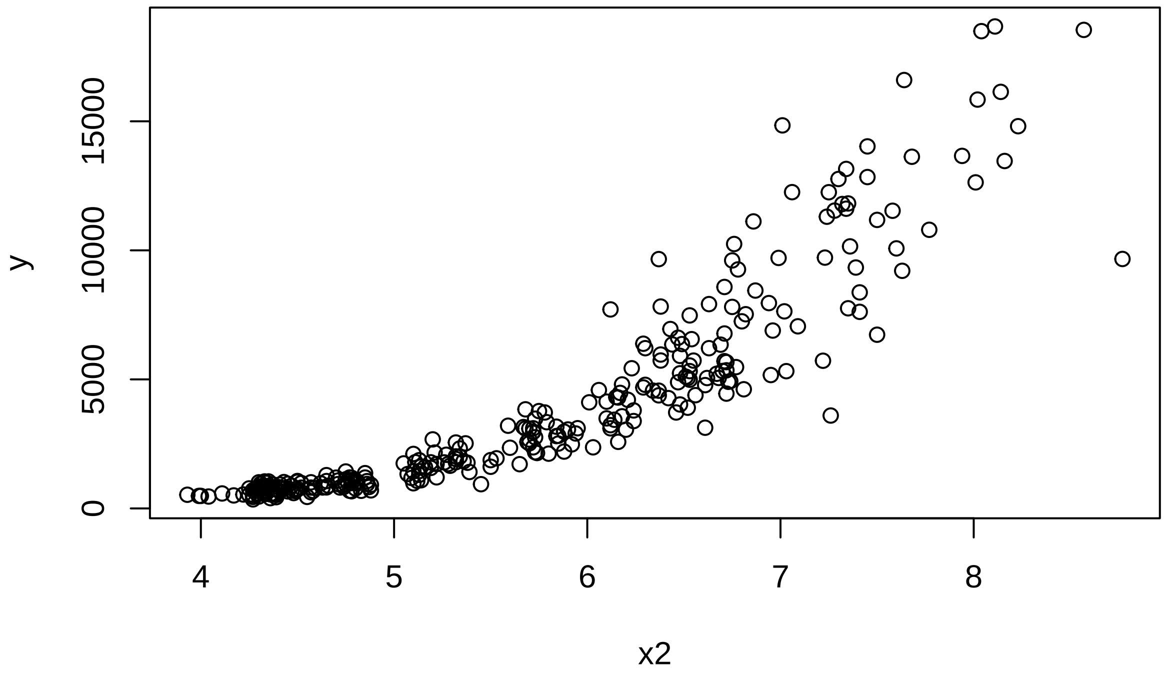
Finally, the report's main focus is on the one quantitative dependent variable, the price, which serves as the response variable in this analysis. We examine the interplay between these variables to explain how they affect the price of diamonds.

```r
diamonds <- diamond_samples[c("color", "cut", "x", "y", "price")]
colnames(diamonds) <- c("color", "cut", "x1", "x2", "y")
x1vy <- subset(diamonds, select = c(x1, y))
plot(x1vy)
```



There is not a correlation between length (x1) and price (y).

```
x2vy <- subset(diamonds, select = c(x2, y))
plot(x2vy)
```



There is not a correlation between width (x2) and price (y).

```
mlr_summary <- lm(y~., diamonds)
summary(mlr_summary)
```

```
##
```

```
## Call:
## lm(formula = y ~ ., data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5272.1 -1062.5  -180.7   964.5  6528.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -15559.17     919.35 -16.924  < 2e-16 ***
## colorE         -252.63     374.82  -0.674  0.50086
## colorF          136.78     360.73   0.379  0.70484
## colorG          -38.57     345.41  -0.112  0.91118
## colorH         -376.35     388.67  -0.968  0.33372
## colorI        -1308.52     413.40  -3.165  0.00172 **
## colorJ        -1191.74     654.97  -1.820  0.06987 .
## cutGood          76.40     820.57   0.093  0.92589
## cutIdeal       1260.97     710.76   1.774  0.07710 .
## cutPremium      937.16     698.45   1.342  0.18073
## cutVery Good   1013.84     754.56   1.344  0.18014
## x1             -263.07    2352.26  -0.112  0.91103
## x2             3554.86    2353.83   1.510  0.13208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1705 on 287 degrees of freedom
## Multiple R-squared:  0.8226, Adjusted R-squared:  0.8152
## F-statistic: 110.9 on 12 and 287 DF,  p-value: < 2.2e-16
```

It is interesting to see there are 3 significant variables in the summary table.

Part 2

For our simple linear regression model, we have chosen x1 (diamond length in mm) to be our chosen predictor.

```r
length <- lm(y~x1, diamonds)
summary(length)
```

```
##
## Call:
## lm(formula = y ~ x1, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5624.8 -1163.6  -167.0   947.6  7399.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14025.96     531.82  -26.37   <2e-16 ***
## x1            3154.64      91.71   34.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1782 on 298 degrees of freedom
## Multiple R-squared:  0.7988, Adjusted R-squared:  0.7981
## F-statistic:  1183 on 1 and 298 DF,  p-value: < 2.2e-16
```

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Our alpha is 0.05 and the p-value is less than 2e-16. We reject the null hypothesis since 2e-16 < 0.05.

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 < 0$$

Our alpha is the same and our p-value is (2e-16)/2 = 1e-16. We reject the null hypothesis since 1e-16 < 0.05.

Therefore, the intercept and x1 are significant to the simple linear regression model. This tells us diamond length in mm is significant to the price value and should be included in the multiple linear regression model.

Our adjusted R-squared is at 79.81% which indicates the model is a good fit.

```r
predict(length, newdata = data.frame(x1=7),
         level = .95,
         interval = "confidence")
```

```
##        fit      lwr      upr
## 1 8056.542 7745.129 8367.955
```

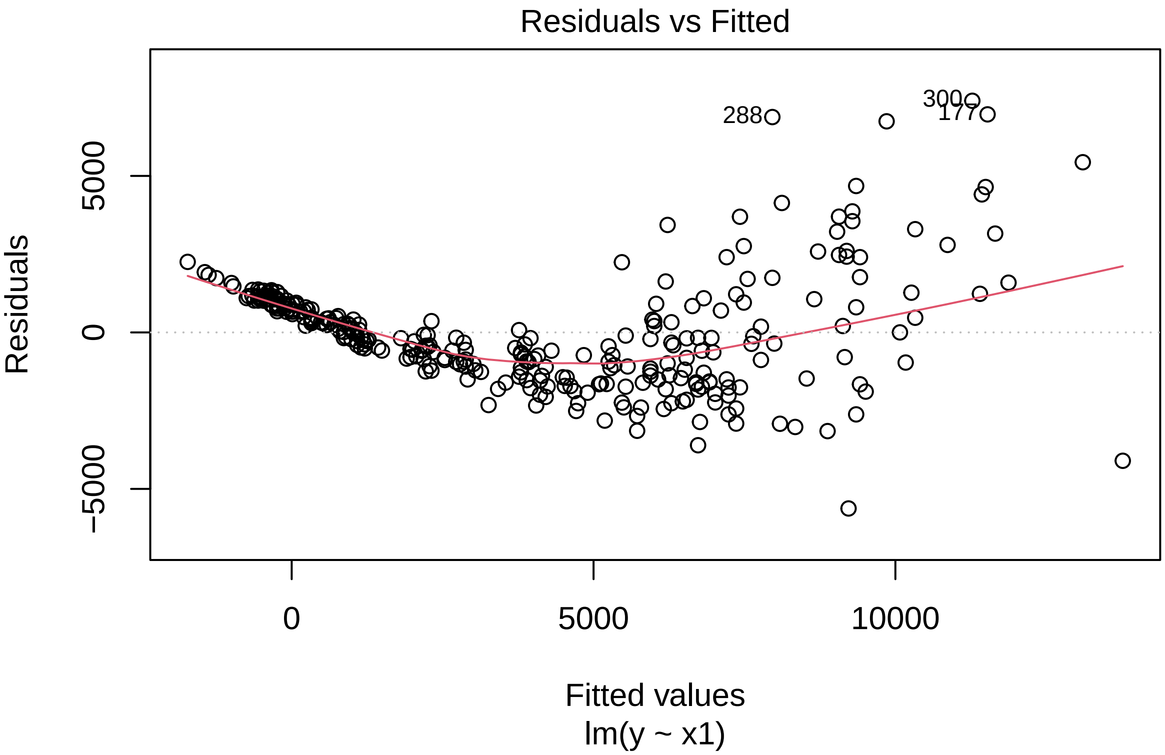
```r
predict(length, newdata = data.frame(x1=7),
         level = .95,
         interval = "prediction")
```

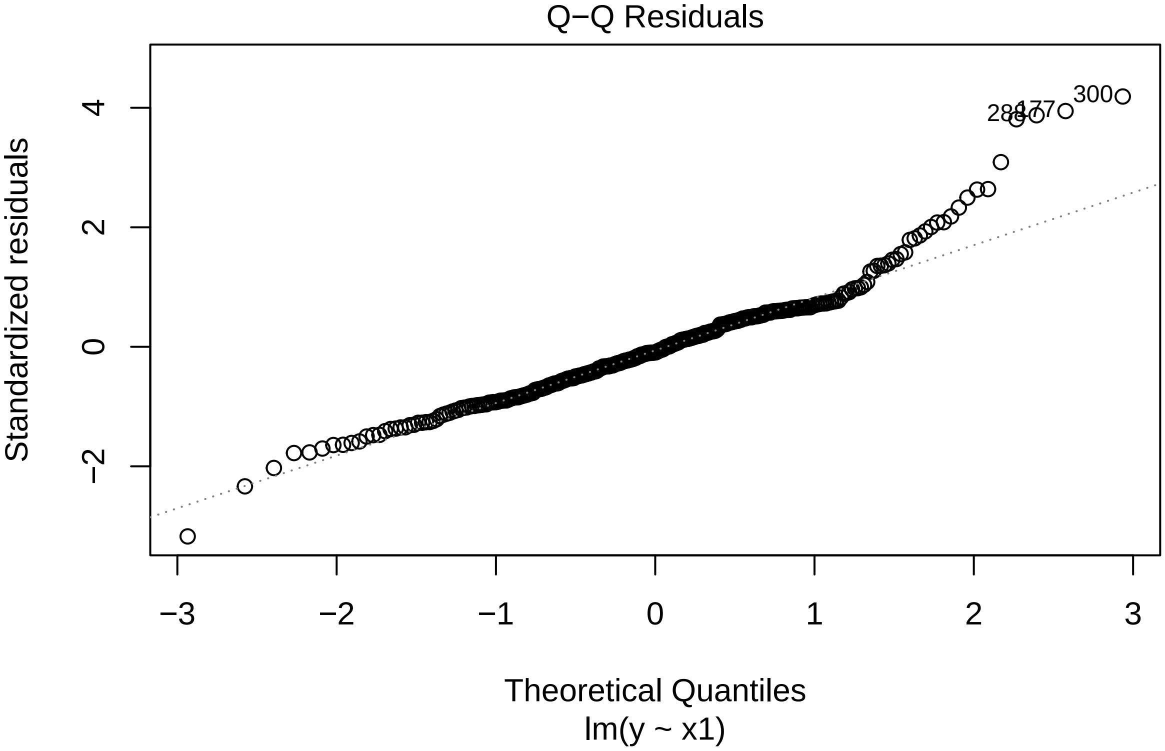
```
##        fit      lwr      upr
## 1 8056.542 4535.139 11577.94
```

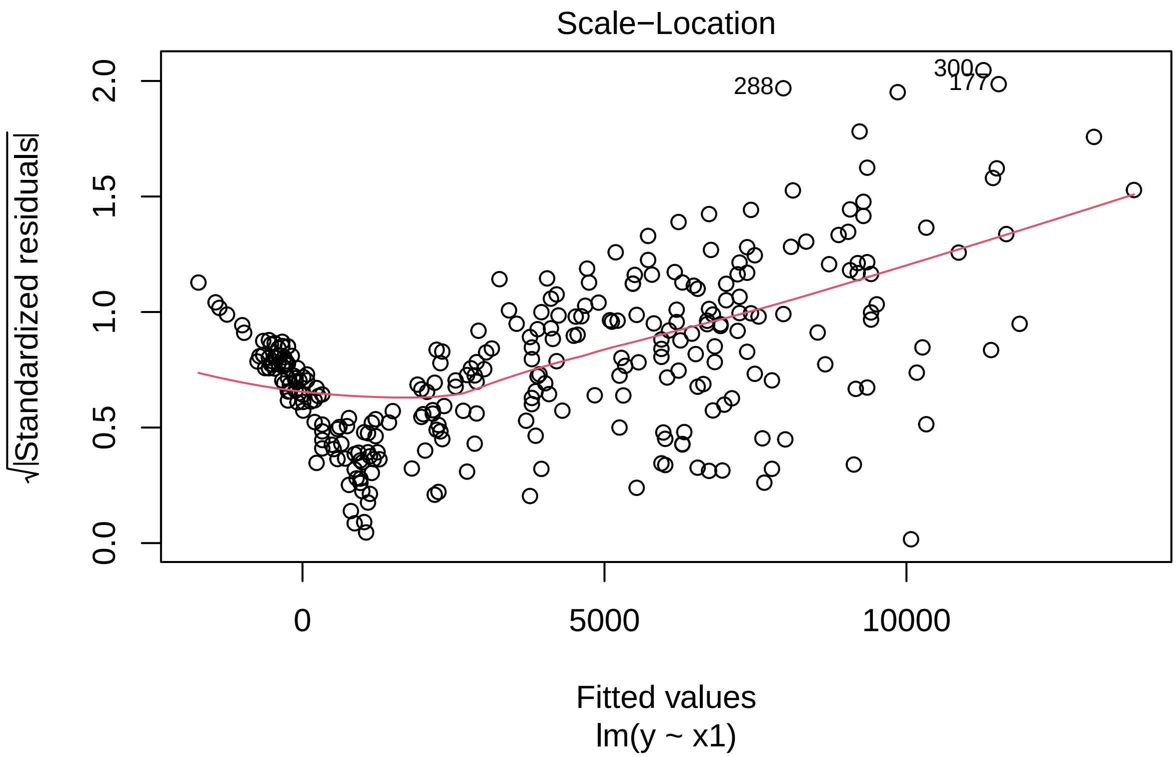
We are 95% confident that the true mean value of price lies within $7,745.13 and $8,367.96. This interval is wide which indicates the sample does not provide a precise representation of the population mean.

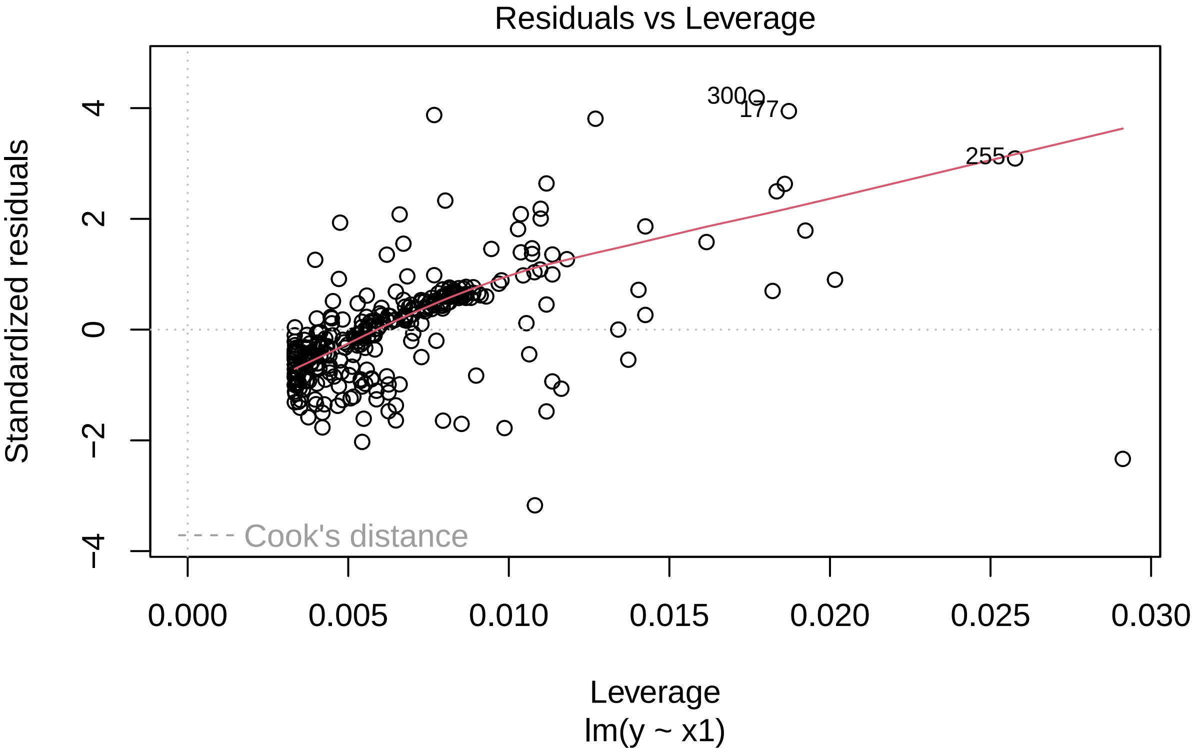
We are 95% confident that a new observation of price lies within $4,535.14 and $11,577.94. This interval is wider than the confidence interval because it accounts for uncertainty in estimating the mean and the variability of individual observations.

```r
plot(length)
```


Residuals vs Fitted
Residuals
5000
0
−5000
0
5000
10000
288
300
177
Fitted values
lm(y ~ x1)

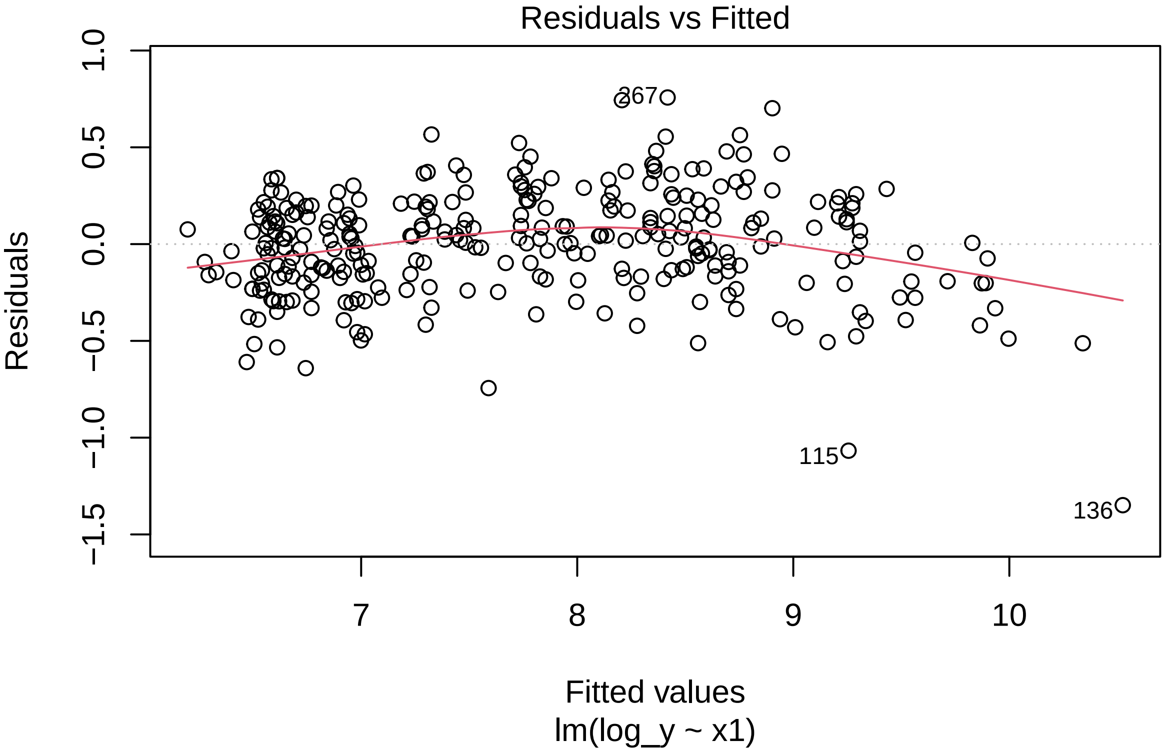

Q−Q Residuals
Standardized residuals
4
2
0
−2
−3
−2
−1
0
1
2
3
288
177
300
Theoretical Quantiles
lm(y ~ x1)

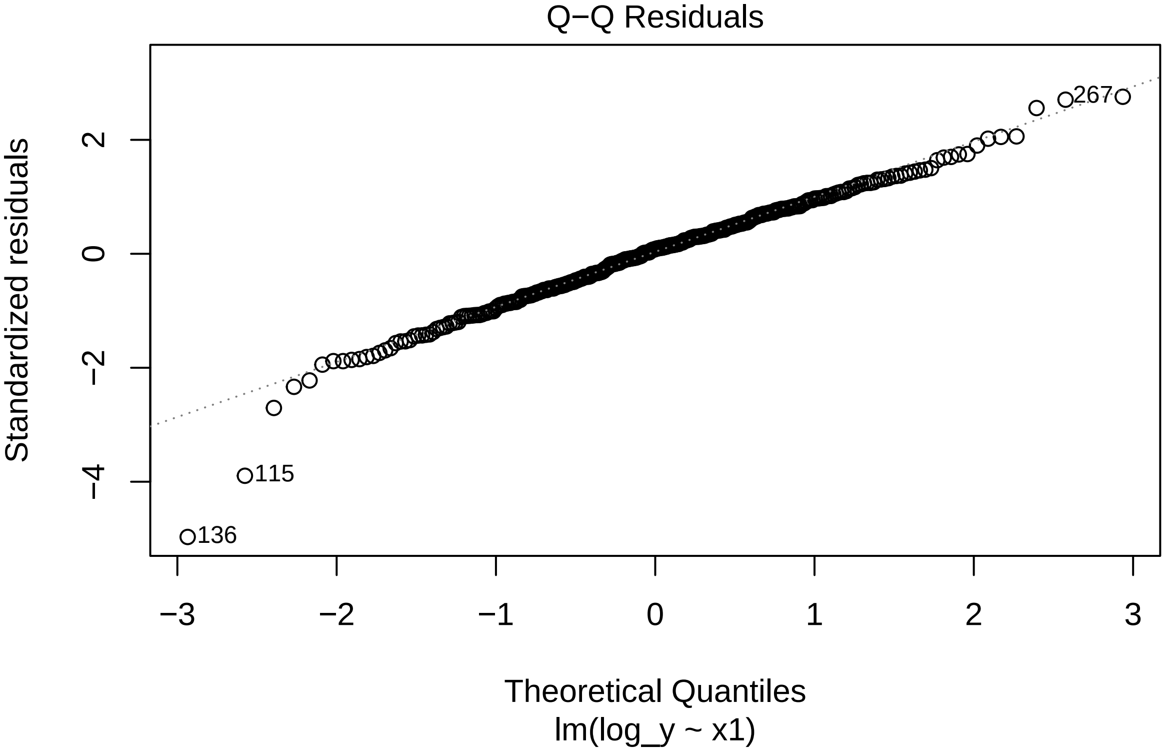


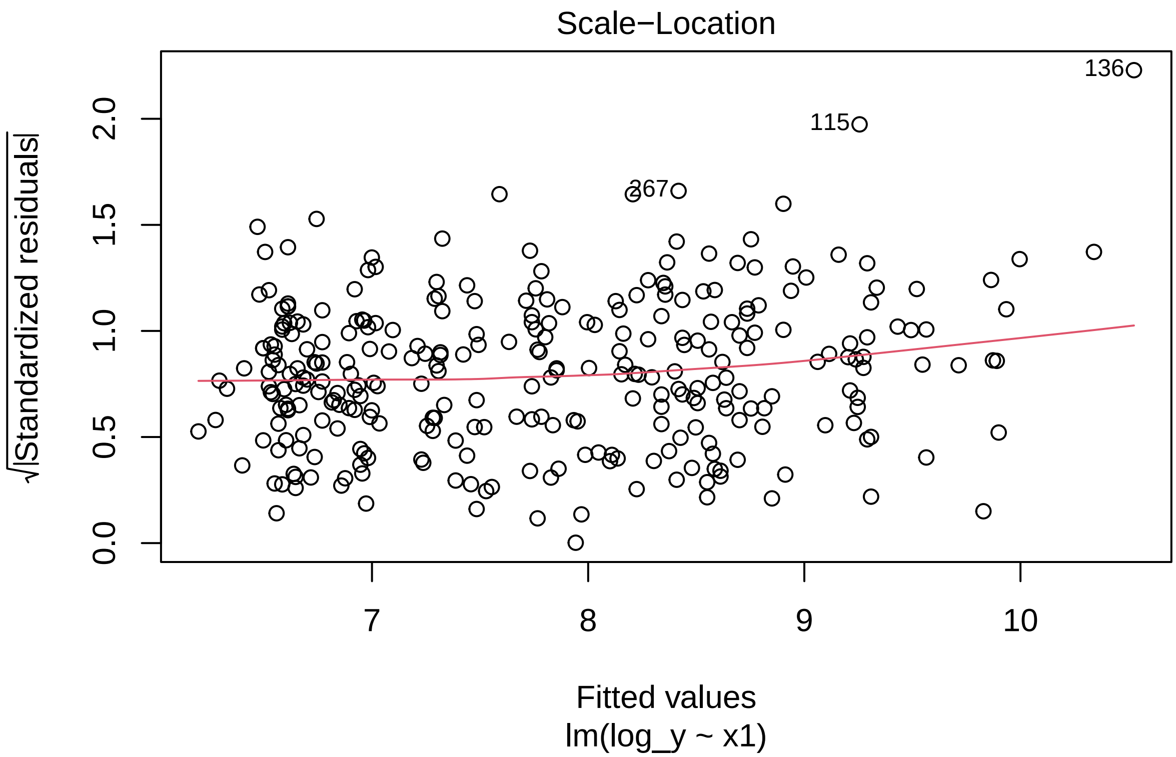


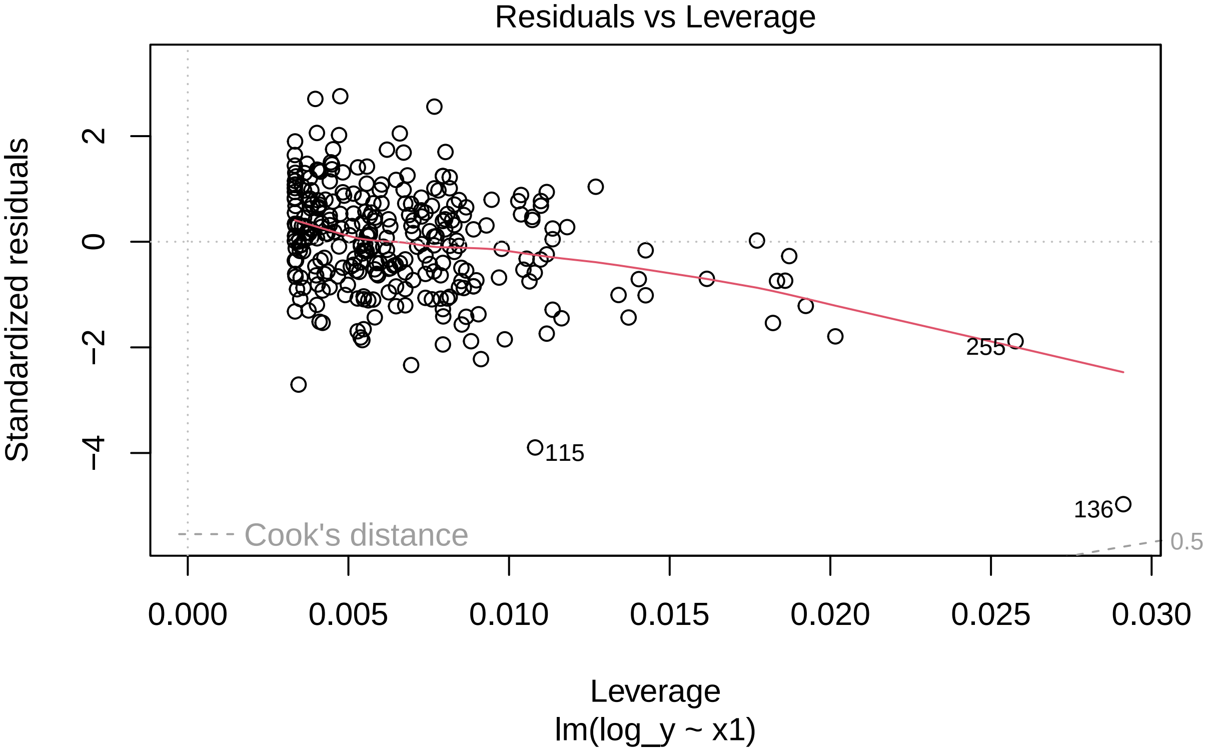
Using x1 as our predictor, the residuals versus fitted plot indicates a non-linear relationship. For the QQ plot, some of the observations are way off the line, thus the normality assumption might be violated. In the scale-location plot, the points are not randomly spread around the line and widens as the fitted values increase. This means the variability of the residuals and variance are not constant.

```
log_y <- log(diamonds$y)
length_log <- lm(log_y~x1, diamonds)
plot(length_log)
```









After we transformed our response variable y by taking the log of it, we can see our plots are more linear in the residuals verses fitted plot. Additionally, there are more observations on the line in the QQ plot and on scale-location, the points are more constant for the variability of the residuals and variance.

```
summary(length_log)
```

```
##
## Call:
## lm(formula = log_y ~ x1, data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34884 -0.16949  0.02256  0.18838  0.75739
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.75924    0.08220   33.57   <2e-16 ***
## x1           0.88148    0.01418   62.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2755 on 298 degrees of freedom
## Multiple R-squared:  0.9284, Adjusted R-squared:  0.9282
## F-statistic:  3867 on 1 and 298 DF,  p-value: < 2.2e-16
```

Looking at the summary of the transformed model, we noticed an increase in adjusted R-squared which indicates this model is a better fit than the original simple linear regression model.

Conclusion

We have demonstrated that x1 (length in mm) is significant to y (price value of diamonds) and therefore should be kept in our simple linear regression model. In the background, we have added predictors to our simple linear regression model and examined adjusted R-squared to determine if they should be kept in our multiple linear regression model. We have concluded that x1 (length in mm), x2 (width in mm), Color, and Cut are significant and should not be dropped based on adjusted R-squared.

```
updatedmodel <- lm(y~ x1 + x2 + color + cut, diamonds)
library(car)
```

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

```
vif(updatedmodel)
```

```
##             GVIF Df GVIF^(1/(2*Df))
## x1    718.723442  1       26.809018
## x2    717.164477  1       26.779927
## color   1.260383  6        1.019472
## cut     1.729545  4        1.070882
```

The results of the vif function indicate there is multicollinearity between x1 and x2.

Part 3

```
start_model <- lm(y~1, diamonds)
mlr <- lm(y~., diamonds)
step(start_model, direction="forward", scope=formula(mlr))
```

```
## Start:  AIC=4972.46
## y ~ 1
##
##          Df  Sum of Sq        RSS    AIC
## + x2      1 3771311897  934109954 4489.4
## + x1      1 3758733829  946688022 4493.4
## + color   6  265361131 4440060720 4967.0
## <none>                 4705421851 4972.5
## + cut     4   25359331 4680062520 4978.8
##
## Step:  AIC=4489.4
## y ~ x2
##
##         Df Sum of Sq       RSS    AIC
## + color  6  68794786 865315167 4478.4
## + cut    4  43754092 890355862 4483.0
## <none>               934109954 4489.4
## + x1     1    646314 933463640 4491.2
##
## Step:  AIC=4478.45
## y ~ x2 + color
##
##        Df Sum of Sq       RSS    AIC
## + cut   4  30770058 834545109 4475.6
## <none>              865315167 4478.4
## + x1    1    840312 864474855 4480.2
##
## Step:  AIC=4475.58
## y ~ x2 + color + cut
##
##        Df Sum of Sq       RSS    AIC
## <none>              834545109 4475.6
## + x1    1     36368 834508741 4477.6
```

```
##
## Call:
## lm(formula = y ~ x2 + color + cut, data = diamonds)
##
## Coefficients:
##  (Intercept)            x2        colorE        colorF        colorG
##    -15584.49       3291.83       -254.92        135.42        -37.52
##       colorH        colorI        colorJ       cutGood      cutIdeal
##      -380.41      -1307.77      -1194.86        108.17       1287.58
##   cutPremium  cutVery Good
##       953.79       1049.39
```

Despite using the adjusted R-squared comparison method to see which predictors are best to use, we used forward selection as a more robust way to check which predictors work best for the multiple linear regression model. Forward selection tells us that x2 (width in mm), Color, and Cut are significant and should remain in our best model. It is logical that we dropped x1 (length in mm) from our multiple linear regression model because there was multicollinearity present between x1 and x2.

```r
bestmodel <- lm(y~ x2 + color + cut, diamonds)
summary(bestmodel)
```

```
## 
## Call:
## lm(formula = y ~ x2 + color + cut, data = diamonds)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5288.5 -1062.1  -186.9   963.0  6522.0 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -15584.49     889.51 -17.520  < 2e-16 ***
## x2             3291.83      93.56  35.184  < 2e-16 ***
## colorE         -254.92     373.62  -0.682  0.49559    
## colorF          135.42     359.91   0.376  0.70699    
## colorG          -37.52     344.69  -0.109  0.91340    
## colorH         -380.41     386.31  -0.985  0.32558    
## colorI        -1307.77     412.64  -3.169  0.00169 ** 
## colorJ        -1194.86     653.26  -1.829  0.06842 .  
## cutGood         108.17     768.51   0.141  0.88817    
## cutIdeal       1287.58     668.58   1.926  0.05511 .  
## cutPremium      953.79     681.27   1.400  0.16258    
## cutVery Good   1049.39     683.17   1.536  0.12562    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1702 on 288 degrees of freedom
## Multiple R-squared:  0.8226, Adjusted R-squared:  0.8159 
## F-statistic: 121.4 on 11 and 288 DF,  p-value: < 2.2e-16
```

From the summary, we can see that adjusted R-squared is high which indicates the predictors are a good fit for the model. However, we can see that the only significant colors are Color I and J, and the only significant cut is Cut Ideal. This leads us to conclude that our final best model includes x2, Color I and J, and Cut Ideal as our predictors.

```r
diamonds_bestfinal <- subset(diamonds, !(color %in% c("E", "F", "G", "H")))
diamonds_bestfinal <- subset(diamonds, !(cut %in% c("Good", "Premium", "Very Good")))
bestfinal <- lm(y~ x2 + color, diamonds_bestfinal)

predict(bestfinal, newdata = data.frame( x2 = 8, color = "I"),
         level = .95,
         interval = "confidence")
```

```
##        fit      lwr      upr
## 1 11366.25 10161.14 12571.36
```

```r
predict(bestfinal, newdata = data.frame(x2 = 8, color = "I"),
         level = .95,
         interval = "prediction")
```

```
##        fit      lwr      upr
## 1 11366.25 7853.982 14878.51
```

We are 95% confident that the true mean value of price lies within $10,161.14 and $12,571.36. This interval is wide which indicates the sample does not provide a precise representation of the population mean.

We are 95% confident that a new observation of price lies within $7,853.98 and $14,878.51. This interval is wider than the confidence interval because it accounts for uncertainty in estimating the mean and the variability of individual observations.

Conclusion

We have concluded that x2 (width in mm), Color I and J, and Cut Ideal have the largest impact on diamond prices in 2022. This is logical because width of a diamond is often the best indicator of its size. Additionally, Color I and J are near colorless, which are one of the most popular types of diamonds to choose and Cut Ideal is also a common choice.