PSTAT 126 Final Project

Daniel Larson, Emanuel Rodriguez

2025-12-09

Contents

Part 1: Data Description and Descriptive Statistics

Requirements: Include at least 2 categorical variables and 3 independent quantities.

1. Select a random sample (must have 500 observations):

```
filepath <- "/Users/daniellarson/Downloads/Diamonds Prices2022.csv"
df <- read.csv(filepath)

set.seed(777)
sample_df <- df[sample(nrow(df), 500), ]
```

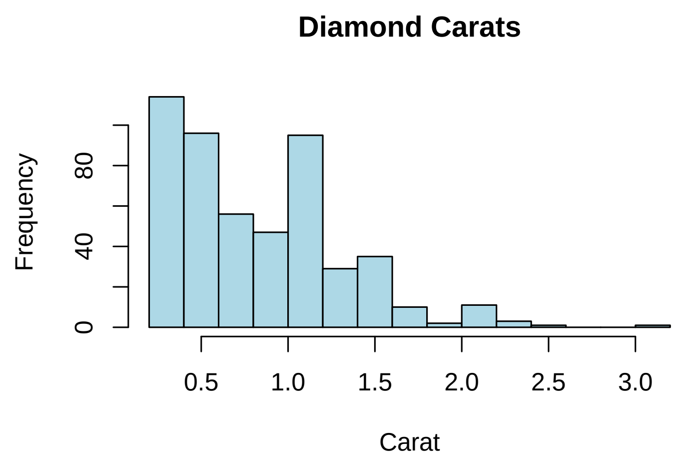
2. Describe all the variables(call summary function on the dataset, see the structure, create histograms for continuous random variable, comment on their distribution, bar plots for categorical random variable)

summary(sample_df)

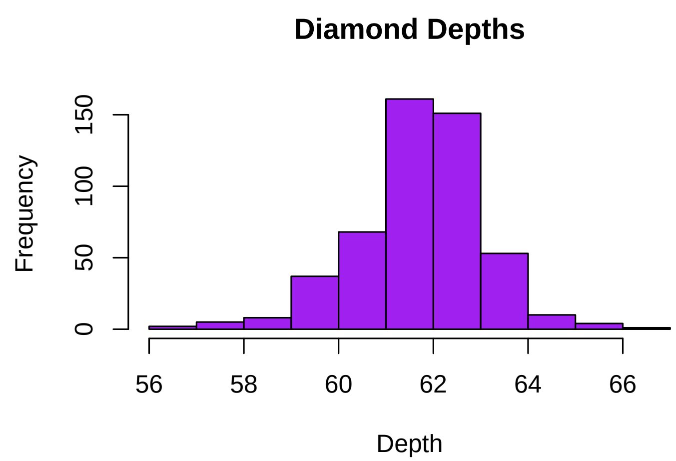
```
summary(sample_df)
```

```
##        X             carat            cut               color          
##  Min.   :  130   Min.   :0.2100   Length:500         Length:500        
##  1st Qu.:13193   1st Qu.:0.4200   Class :character   Class :character  
##  Median :25246   Median :0.7250   Mode  :character   Mode  :character  
##  Mean   :25858   Mean   :0.8303                                        
##  3rd Qu.:38800   3rd Qu.:1.0600                                        
##  Max.   :53853   Max.   :3.0100                                        
##    clarity              depth           table           price      
##  Length:500         Min.   :56.10   Min.   :53.00   Min.   :  386  
##  Class :character   1st Qu.:61.17   1st Qu.:56.00   1st Qu.: 1008  
##  Mode  :character   Median :61.90   Median :57.00   Median : 2858  
##                     Mean   :61.78   Mean   :57.43   Mean   : 4207  
##                     3rd Qu.:62.60   3rd Qu.:59.00   3rd Qu.: 5884  
##                     Max.   :66.60   Max.   :65.00   Max.   :17904  
##        x               y               z        
##  Min.   :3.870   Min.   :3.830   Min.   :2.330  
##  1st Qu.:4.800   1st Qu.:4.810   1st Qu.:2.990  
##  Median :5.810   Median :5.830   Median :3.590  
##  Mean   :5.823   Mean   :5.823   Mean   :3.596  
##  3rd Qu.:6.565   3rd Qu.:6.590   3rd Qu.:4.050  
##  Max.   :9.540   Max.   :9.380   Max.   :5.340  
```

```r
hist(sample_df$carat,
     main = "Diamond Carats",
     xlab = "Carat",
     ylab = "Frequency",
     col = "lightblue",
     border = "black")
```

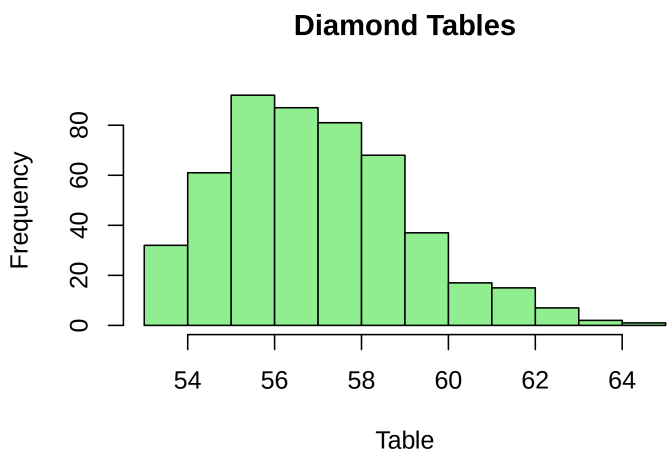



```r
hist(sample_df$depth,
     main = "Diamond Depths",
     xlab = "Depth",
     ylab = "Frequency",
     col = "purple",
     border = "black")
```

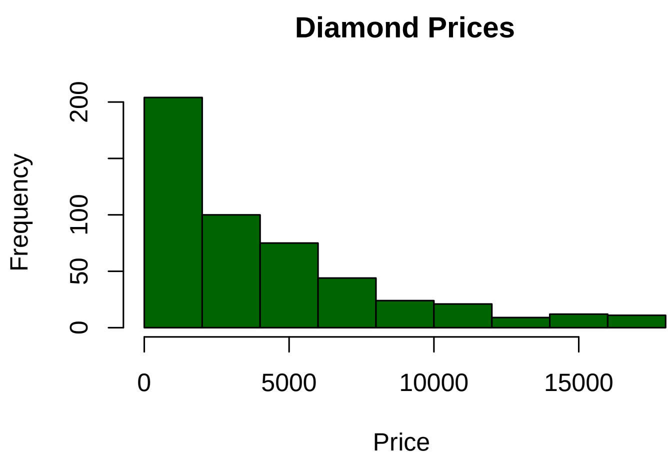



```r
hist(sample_df$table,
     main = "Diamond Tables",
     xlab = "Table",
```

```
      ylab = "Frequency",
      col = "lightgreen",
      border = "black")
```

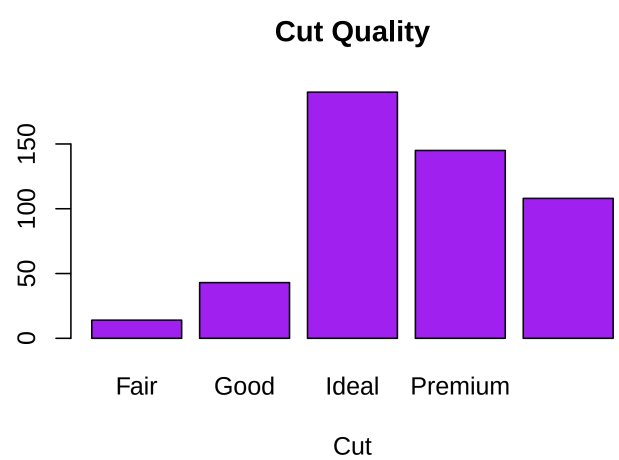



```
hist(sample_df$price,
      main = "Diamond Prices",
      xlab = "Price ",
      ylab = "Frequency",
      col = "darkgreen",
      border = "black")
```

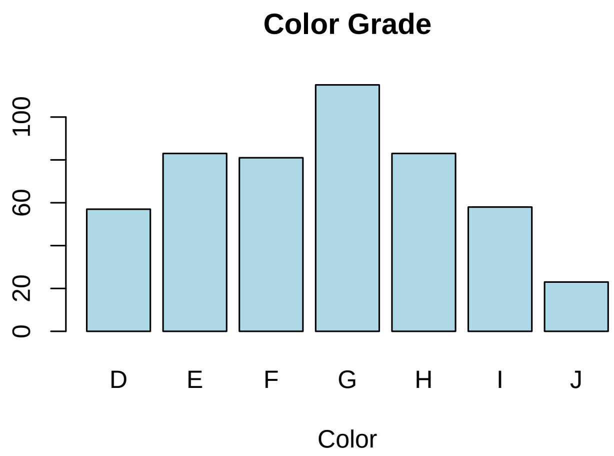


Histogram Comments: The carat distribution is clearly right-skewed. Most observations fall in the 0–1 range, with a peak around the 1.0 and 1.5 interval with a smaller increase near 1.5. We have an outlier.The distribution matches with real world expectation that the larger the carat the rarer the diamond is. Depth values look roughly symmetric and close to a normal shape, centered around the typical range, with one visible outlier slightly above 70. Besides that, the distribution is pretty regular without major irregularities. The table distribution is slightly right-skewed. The table numbers show a tight pattern, with most of the diamonds being clustered around the middle, between 54–60, with around 55 being the most common, meaning the majority of the diamonds are of a common cut. This suggests that most diamonds fall within an average range for the table proportions meaning they share an ordinary shape. Prices on the other hand, are heavily skewed to the right. This means that the majority of the diamonds of the sample are priced below 5,000 dollars, however, there are many diamonds priced at 10,000 dollars or more causing a long upper tail. This is normal and is to be expected because a small number of diamonds are expensive. There are little to no outliers to be seen beyond the tail, so the overall distribution is realistic for this data.

```
#Barplot's

#Cut
barplot(table(sample_df$cut),
        main="Cut Quality",
        xlab="Cut",
        col = "purple")
```




```
#Color
barplot(table(sample_df$color),
        main="Color Grade",
        xlab="Color",col = "lightblue")
```

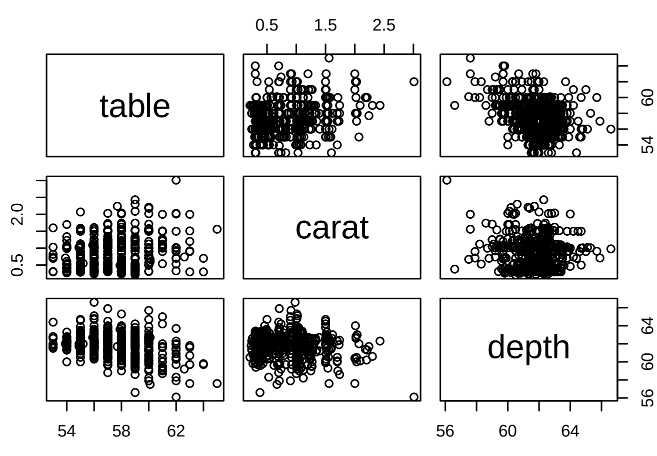


3. Choose 3 quantitative and 2 categorical variables appropriately and determine if there is any correlation between these variables.

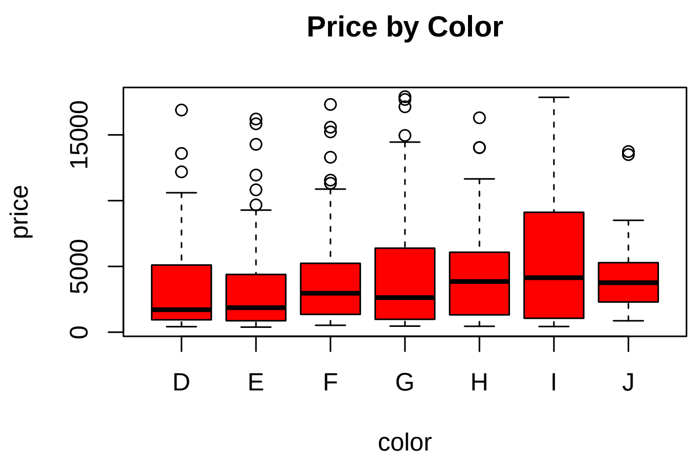
```
#Variables
quant_vars <- sample_df[, c("table", "carat", "depth")]

cor(quant_vars)
```

```
##             table       carat       depth
## table  1.0000000  0.25164918 -0.33747879
## carat  0.2516492  1.00000000 -0.08903546
## depth -0.3374788 -0.08903546  1.00000000
```

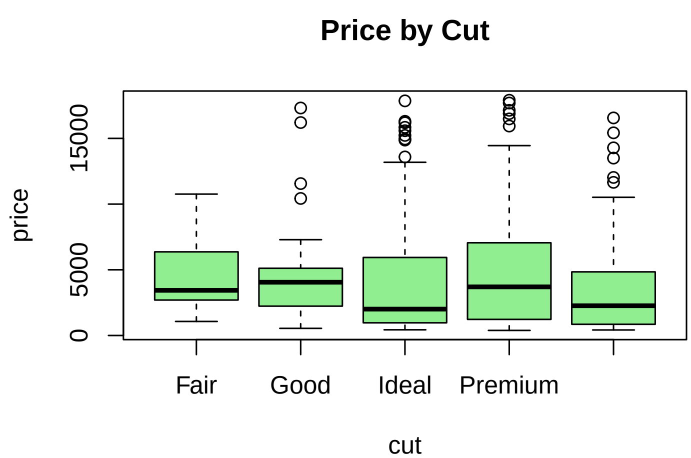
```
pairs(quant_vars)
```



```r
boxplot(price ~ color,
        data = sample_df,
        main = "Price by Color",
        col = "red")
```




```r
boxplot(price ~ cut,
        data = sample_df,
        main = "Price by Cut",
        col = "lightgreen")
```



Comments on correlation

Between table, cut, and depth, there doesn't appear to be any strong correlation between the quantitative variables. The only real sense of what could appear to be correlated is between table and depth, with a (-0.337). As shown on the pairs scatterplot (bottom left), the graph seems to go slightly downwards, indicating a slight correlation, but all of these plots show a generally weak correlation.Then, we made boxplots pertaining to categorical variables, sorting price by color, and price by cut. We can see that looking at the boxplots that the median prices can change depending on the quality of the cut and the grade of the color. Additionally, some categories have noticeably higher or lower ranges in pricing, so there appears to be an association in price when sorted by both cut and color, even if those variables are not numerical.As a whole, carat seems to be the biggest impact on price generally, whereas on the other side, depth and table do not show too many correlations in the scatterplots. However, the color and cut, categorial variables, also appear to influence the price depending on how the boxplots different across groups.

4.Run multiple linear regression model using all these variables and observe the summary statistics. (No need to explain hypothesis testing or other things)

```r
diamondmodel <- lm(price ~ color + cut + carat + depth + table, data = sample_df)
summary(diamondmodel)
```

```
##
## Call:
## lm(formula = price ~ color + cut + carat + depth + table, data = sample_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8521.0  -723.0  -112.3   551.1  6573.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8151.41    5127.63  -1.590  0.11255
## colorE        -136.74     250.51  -0.546  0.58544
## colorF        -254.14     252.22  -1.008  0.31414
## colorG         -84.36     235.21  -0.359  0.72001
## colorH       -1078.81     251.85  -4.283 2.22e-05 ***
## colorI       -1158.57     276.12  -4.196 3.23e-05 ***
## colorJ       -2184.47     362.49  -6.026 3.32e-09 ***
## cutGood       1451.49     463.94   3.129  0.00186 **
## cutIdeal      2329.69     444.17   5.245 2.34e-07 ***
## cutPremium    1922.74     436.21   4.408 1.29e-05 ***
## cutVery Good  1967.59     434.22   4.531 7.39e-06 ***
## carat         8432.71     150.31  56.102  < 2e-16 ***
## depth           98.01      56.98   1.720  0.08606 .
## table          -38.20      42.25  -0.904  0.36627
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1444 on 486 degrees of freedom
## Multiple R-squared:  0.8751, Adjusted R-squared:  0.8718
## F-statistic: 261.9 on 13 and 486 DF,  p-value: < 2.2e-16
```

5. comment on anything of interest that occurred in this part. Were the data approximately what you expected, or did some of the results surprise you?

The descriptive analysis revealed patterns that aligned closely with typical diamond market behavior. Carat and price were both heavily right-skewed, with most observations concentrated in the lower range and a long tail of larger stones and high priced diamonds extending outward. Depth and table measurements were far more constrained, clustering around common industry standards, and their distributions showed no unusual irregularities. The boxplots highlighted strong and consistent relationships between the categorical variables and price, where higher cut and color grades corresponded to noticeably higher median price levels. Among the numeric predictors, correlations were generally weak, though there was a moderate negative association between depth and table, which is reasonable given the geometric trade offs in diamond proportions. Aside from a few expected high end values, no influential anomalies appeared in the sample, suggesting the data are representative of real world pricing patterns. Overall, even before model fitting, it was clear that carat, cut, and color play major roles in explaining price variation.

Part 2: Simple Linear Regression (continuation of Part 1)

1.Start with one predictor and one response from the variables you chose in Part I. For instance, you can start with the predictor 'carat' and the response 'price', and conduct a simple linear regression analysis on it.

We will use carat predictors and price.

2. Run the model and examine the summary statistics, interpreting everything (hypothesis testing, adjusted R squared as discussed in class, confidence interval, prediction interval, plot, etc.).

```r
model_simple <- lm(price ~ carat, data = sample_df)
summary(model_simple)
```

```
##
## Call:
## lm(formula = price ~ carat, data = sample_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10692.6   -853.7    -36.6    586.0   7563.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2362.5      144.9  -16.30   <2e-16 ***
## carat         7912.3      152.1   52.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1591 on 498 degrees of freedom
## Multiple R-squared:  0.8446, Adjusted R-squared:  0.8443
## F-statistic:  2707 on 1 and 498 DF,  p-value: < 2.2e-16
```

Full interpretation:

Model: $price_i = \beta_0 + \beta_1(carat_i) + \epsilon_i$

Hypothesis Tests: H0 = $\beta_1 = 0$ (carat has no linear effect on price) HA = β_1 =/= 0 (carat affects price)

The estimated slope for carat is 7912.3 , indicating that for each additional carat, the expected diamond price increases by approximately $7912 on average. The hypothesis test for the slope yields a p value less than 2*(10^-16), leading to rejection of the null hypothesis because the p value is less than any reasonable significance level. This confirms that carat is a statistically significant predictor of price.

R^2 adjusted: The adjusted R^2 is 0.8446, meaning that roughly 84% of the variability in price is explained by carat alone. This is remarkably high for a single predictor model and aligns well with expectations, given the dominant role of size in diamond valuation. The residual standard error of about 1591 indicates that, while carat explains most of the variation, substantial price variability remains due to other characteristics such as cut, color, and clarity.

The Confidence Interval:

```r
new_carat <- data.frame(carat = 1)

predict(model_simple, new_carat, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 5549.867 5401.129 5698.605
```

Interpretation for CI: For a diamond with the specified carat value, the estimated mean price is approximately 5550 dollars, with the 95% confidence interval for this mean price being (5401, 5699). This relatively narrow interval reflects the precision of the estimated regression line and indicates that the average price at this carat level is estimated with fairly high confidence.

The Prediction Interval:

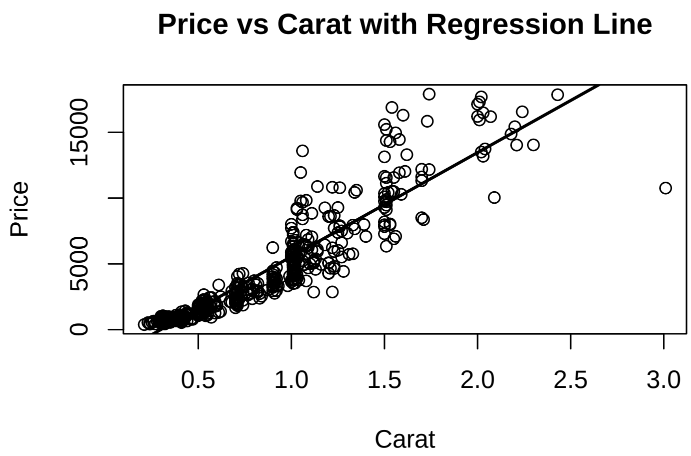
```r
predict(model_simple, new_carat, interval = "prediction")
```

```
##        fit      lwr     upr
## 1 5549.867 2419.655 8680.08
```

Interpretation for PI: For a single new diamond of the same carat size, the predicted price is also 5550 dollars, but the 95% prediction interval is much wider: (2420, 8680). This wider interval is accounting not only for uncertainty in estimating the regression line but also for the natural variability in individual diamond prices. This highlights that while the mean price can be estimated precisely, individual diamond prices can vary substantially despite having the same carat level.

Plot with fitted regression line:

```r
plot(sample_df$carat, sample_df$price,
     xlab = "Carat",
     ylab = "Price",
     main = "Price vs Carat with Regression Line")

abline(model_simple, lwd = 2)
```

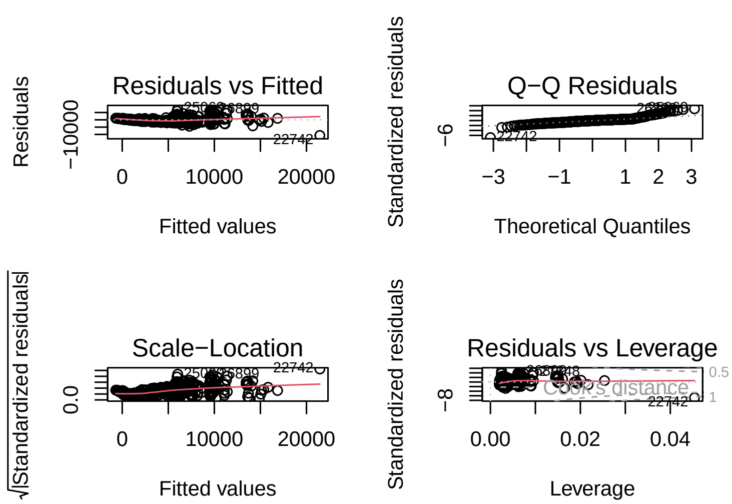


The scatterplot line shows that there is a strong positive linear relationship between carat and price. Additionally, points are centered close around the regression line, which shows a high R^2 support. Variability tends to increase for larger diamond with more vertical spread, typical among heteroscedastic pricing data.

3.Test the assumptions and apply any necessary transformations to the response variable y or the predictor.

Original model diagnostics:

```
par(mfrow=c(2,2))
plot(model_simple)
```




```
par(mfrow=c(1,1))
log_model <- lm(log(price) ~ carat, data = sample_df)
```

4. Call the summary function on the transformed variables, observe the summary, and note any changes.

```r
summary(log_model)
```

```
##
## Call:
## lm(formula = log(price) ~ carat, data = sample_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.90905 -0.23163  0.02578  0.25810  1.18424
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.23443    0.03549  175.65   <2e-16 ***
## carat        1.97950    0.03724   53.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3897 on 498 degrees of freedom
## Multiple R-squared:  0.8502, Adjusted R-squared:  0.8499
## F-statistic:  2826 on 1 and 498 DF,  p-value: < 2.2e-16
```

Note changes

After log-transforming the response variable, the model shows a modest improvement in fit, with the adjusted R^2 increasng from 0.8443 to 0.8499. The residual standard error is substantially reduced on the log scale (1591 to 0.3897), indicating more stable variance and improved adherence to model assumptions. In both models, the carat coefficient remains highly statistically significant, confirming the strong relationship between carat and price. However, the interpretation of the slope changes: in the transformed model it represents an approximate percentage change in price per unit increase in carat rather than a dollar change. Overall, the log transformed model provides a better fitting and more appropriate description of the data, while still preserving the strong relationship between carat and price observed in the original model.

5. Add other variables to the model and assess if the model improves. For step 5, run the code in the background and include all interpretations in the file. For instance, if adding depth to the simple linear regression model (carat and price) increases the adjusted R^2, include it in the model; if it decreases, exclude it. Do not include the code for step 5 in the submitted file; only write the conclusions.

```
##      base        m2        m3        m4        m5
## 0.8498601 0.8519611 0.8678846 0.8714167 0.8713216
```

All the variables that were added to this model always kept improving the adjusted R^2. Overall, the final model with carat, depth, color, and cut explains substantially more variation in price than carat alone. These results confirm that diamond price depends not only on size but also on quality characteristics. The model balances improved fit with interpretability and satisfies regression assumptions.

6. Comment anything of interest while doing this.

While coding the models, it was clear that carat is the strongest predictor of price, as expected. Adding depth, color, and cut steadily improved the adjusted R-squared, showing that quality characteristics also play a meaningful role. Interestingly, adding table had almost no effect, suggesting it is less important for predicting price. The log transformation of price helped stabilize variance and made the residuals more symmetric. Overall, the results matched expectations, but it was notable how much cut and color contributed beyond just size.

Part 3: Part 2 Continutation...

1. Based on the best model obtained from Part II (you would have more than one variable now), run it and call the summary function to analyze how it works and what you observe.

2. Detect multicollinearity among the variables using the variance inflation factor (VIF)

3. Give CIs for a mean predicted value and the PIs of a future predicted value for at least one combination of X's (from your final linear model).

Using approach 1, we answered #1 and #2 and #3 of Part 3:

Add all remaining variables to the simple regression model you have obtained in part-2 after transformation. Use AIC/BIC to avoid overfitting, check variance inflation factor (VIF) for multicollinearity, and ensure model assumptions are satisfied (you have to check model assumptions again in this part after adding all the other variables).

Compute confidence intervals, prediction intervals, and summarize your report clearly.

```r
best_model <- lm(log(price) ~ carat + depth + table + color + cut + clarity, data = sample_df)
summary(best_model)
```

```
##
## Call:
## lm(formula = log(price) ~ carat + depth + table + color + cut +
##     clarity, data = sample_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77887 -0.19978  0.05101  0.22326  0.80092
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.500615   1.164762   3.005  0.00279 **
## carat        2.217139   0.036143  61.343  < 2e-16 ***
## depth        0.028203   0.013198   2.137  0.03310 *
## table       -0.007769   0.009534  -0.815  0.41551
## colorE      -0.045015   0.056615  -0.795  0.42694
## colorF       0.010798   0.057067   0.189  0.85000
## colorG      -0.090524   0.054586  -1.658  0.09790 .
## colorH      -0.212145   0.057490  -3.690  0.00025 ***
```

```
## colorI       -0.451494   0.062719  -7.199 2.37e-12 ***
## colorJ       -0.380062   0.081839  -4.644 4.42e-06 ***
## cutGood       0.207604   0.105644   1.965  0.04998 *
## cutIdeal      0.208290   0.101938   2.043  0.04157 *
## cutPremium    0.191020   0.099698   1.916  0.05596 .
## cutVery Good  0.171059   0.099499   1.719  0.08622 .
## clarityIF     1.541263   0.190415   8.094 4.79e-15 ***
## claritySI1    1.136220   0.172848   6.574 1.29e-10 ***
## claritySI2    0.957326   0.171872   5.570 4.25e-08 ***
## clarityVS1    1.293815   0.174882   7.398 6.23e-13 ***
## clarityVS2    1.229676   0.171928   7.152 3.21e-12 ***
## clarityVVS1   1.284112   0.181152   7.089 4.88e-12 ***
## clarityVVS2   1.378685   0.178530   7.722 6.72e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3242 on 479 degrees of freedom
## Multiple R-squared:  0.9002, Adjusted R-squared:  0.8961
## F-statistic: 216.1 on 20 and 479 DF,  p-value: < 2.2e-16
```

```r
AIC_model <- stepAIC(best_model,
                          direction = "both",  # Could use "backward" or "forward"
                          trace = TRUE)
```

```
## Start:  AIC=-1105.73
## log(price) ~ carat + depth + table + color + cut + clarity
##
##           Df Sum of Sq    RSS     AIC
## - cut      4      0.50  50.86 -1108.8
## - table    1      0.07  50.43 -1107.0
## <none>                  50.36 -1105.7
## - depth    1      0.48  50.84 -1103.0
## - color    6     10.17  60.53 -1025.7
## - clarity  7     12.90  63.26 -1005.7
## - carat    1    395.60 445.95   -17.2
##
## Step:  AIC=-1108.76
## log(price) ~ carat + depth + table + color + clarity
##
##           Df Sum of Sq    RSS      AIC
## <none>                  50.86 -1108.76
## - depth    1      0.23  51.09 -1108.46
## - table    1      0.26  51.12 -1108.20
## + cut      4      0.50  50.36 -1105.73
## - color    6     10.24  61.10 -1029.07
## - clarity  7     13.90  64.76 -1001.99
## - carat    1    400.79 451.66   -18.84
```

```r
#Summary
summary(AIC_model)
```

```
##
## Call:
```

```
## lm(formula = log(price) ~ carat + depth + table + color + clarity,
##     data = sample_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94381 -0.19746  0.04745  0.22215  0.85099
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.511062   0.969831   4.651 4.26e-06 ***
## carat        2.215589   0.035912  61.694  < 2e-16 ***
## depth        0.017865   0.011978   1.491 0.136484
## table       -0.011865   0.007536  -1.574 0.116043
## colorE      -0.045285   0.056323  -0.804 0.421786
## colorF       0.007045   0.056670   0.124 0.901110
## colorG      -0.089574   0.054573  -1.641 0.101371
## colorH      -0.219122   0.057444  -3.815 0.000154 ***
## colorI      -0.450733   0.062703  -7.188 2.51e-12 ***
## colorJ      -0.372692   0.081801  -4.556 6.61e-06 ***
## clarityIF    1.599255   0.188610   8.479 2.78e-16 ***
## claritySI1   1.187105   0.171062   6.940 1.27e-11 ***
## claritySI2   1.013746   0.169692   5.974 4.49e-09 ***
## clarityVS1   1.353502   0.172332   7.854 2.63e-14 ***
## clarityVS2   1.284038   0.169898   7.558 2.08e-13 ***
## clarityVVS1  1.345675   0.178754   7.528 2.55e-13 ***
## clarityVVS2  1.437370   0.176010   8.166 2.79e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3245 on 483 degrees of freedom
## Multiple R-squared:  0.8992, Adjusted R-squared:  0.8959
## F-statistic: 269.4 on 16 and 483 DF,  p-value: < 2.2e-16
```

The variable cut was eliminated from the model using the step wise AIC approach. The improved model fit under the AIC criterion is demonstrated by the modified model's lower AIC value (AIC = −1108.76). The model still includes the predictors carat, depth, table, color, and clarity. Carat continues to be the best predictor still following real world expectations, and the majority of clarity and color levels have statistically significant influence on log(price). With fewer predictors, the final model retains a high adjusted R^2 value (0.8959), indicating that the reduced model explains almost as much variability as the full model. The chosen model strikes a balance between model fit and parsimony because AIC penalizes model complexity.

4. Summarize your report (for the final deliverable).

In this project, our group took a deep insight at the Diamond price dataset and from that learned and interpreted how different attributes influenced pricing. Using a sample of about 500 different diamonds, our group took a look at the data through statistics and visualizations. We noticed a rightward skew in both carat and price, as well as normal depth, and price differences across the cuts and color grades. By using simple linear regression, we solidified carat as being one of the most powerful predictors of price, explaining 84% of a diamond's price. Adding the logarithm price attributed to further improving the models and making the patterns a lot clearer. From that, we added other factors such as cut, color, depth, and clarity. We then used an AIC-based backward elimination statistical model to strip away anything that didn't pull its weight. We then landed on an efficient model based on the previous mentioned factors, deriving the following:

log(price) ∼ carat + depth + table + color + clarity

The stated model explains the 0.8959 adjusted R^2 value, which explains 90% of the price variation. We demonstrated its application by constructing confidence and prediction intervals for a sample diamond, showing how it estimates both the average price for a type and the plausible range for a specific stone. In short: price is predominantly a function of size, but is meaningfully adjusted by the quality signals of color and clarity. The process highlighted the value of transformations and disciplined model selection.