STAT 6021: Project 1

Taylor Tucker, Mahin Ganesan, Wyatt Priddy

2023-10-18

Setup

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## 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
##
## Attaching package: 'MASS'
##
##
## The following object is masked from 'package:dplyr':
##
##     select

##   carat clarity color       cut price
## 1  0.51     SI2     I Very Good   774
## 2  0.93      IF     H     Ideal  6246
## 3  0.50    VVS2     D Very Good  1146
## 4  0.30     VS1     F     Ideal   538
## 5  0.31     SI1     F     Ideal   502
## 6  1.00     VS1     F     Ideal  7046
```

Section 1: High Level Results of the Analysis

Our report studied the relationship between various factors associated with the grading of diamonds and the price of the diamonds. We found that the carat weight, roughly equal to 0.2 grams, is the most important factor in determining the price of the diamond. Thus, we analyzed the relationship between the carat weight of the diamond and its price. We found that, if the carat weight of a diamond increases by 1%, then the price of the diamond will increase by between 1.92% and 1.97%.

Section 2: Data Description

The information in this section mostly comes from the following source: Blue Nile. “The 4Cs of Diamonds,” n.d. https://www.bluenile.com/education/diamonds.

Dataset Variables

The quality, and consequently cost, of diamonds is measured with the "4Cs of Diamonds": the `cut`, `color`, `clarity`, and `carat` weight, which all appear in the dataset.

The `cut` of a diamond measures how "well-proportions a diamond's dimension are", as well as, once being processed, the "balance and brilliance of its facets". The facets of a diamond are the sides which are formed on the diamonds surface to give it its reflective, glittery appearance. The `color` of a diamond refers to the colorlessness of the diamond. Diamonds can often be yellowish, or off-white, depending on impurities in the source material. Therefore, the purest of diamonds are the ones which are completely colorless. The `clarity` of a diamond measures the cloudiness or impurities that can occur during diamond formation, due to impurities or foreign material. The `carat` weight refers to the physical weight of the diamond, and is often the most important indicator for showing how large, and subsequently how valuable, a diamond is. One carat is roughly equivalent to 0.2g, or 0.00705 oz. Finally, he `price` that the particular diamond sold for.

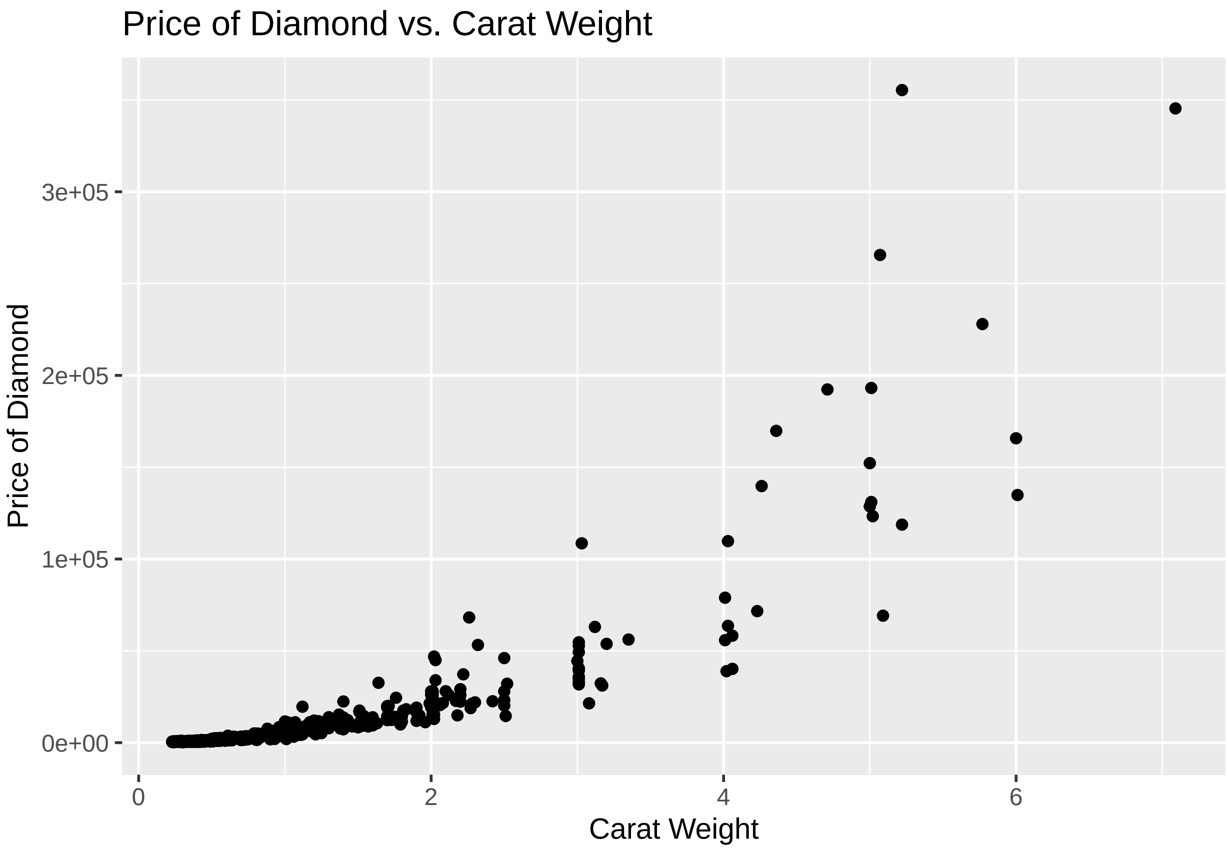
Created Variables

During the analysis portion, we had to create a few new variables, namely `xstar` (x^*) and `ystar` (y^*). The variable `xstar` is the log of the `carat` variable, calculated as $x^* = ln(carat)$. The variable `ystar` is the log of the `price` variable, similarly calculated as $y^* = ln(price)$

Visualizations and Descriptions

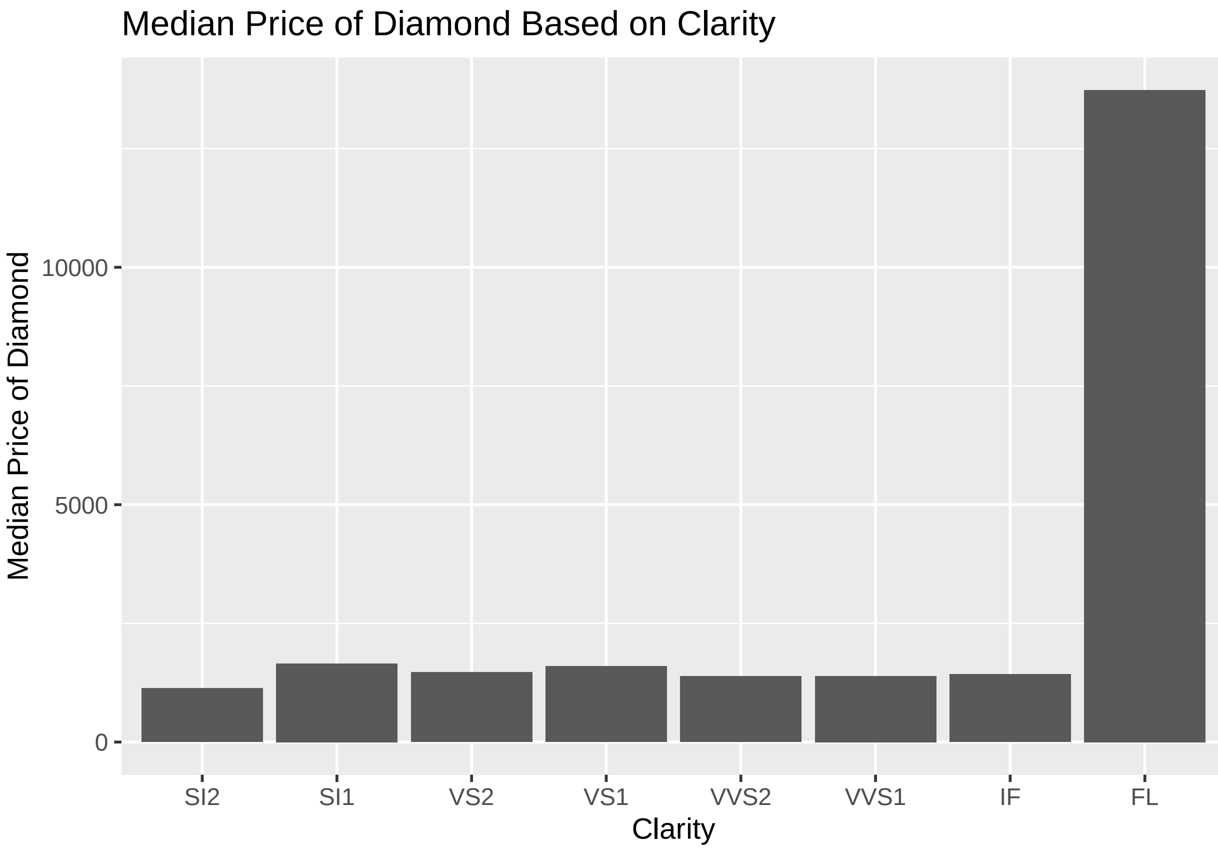
To better understand how price is related to the other variables in the dataset, and how the other variables are related to each other, we will create visualizations to see the appropriate scatterplots and boxplots of each variable against all of the others.

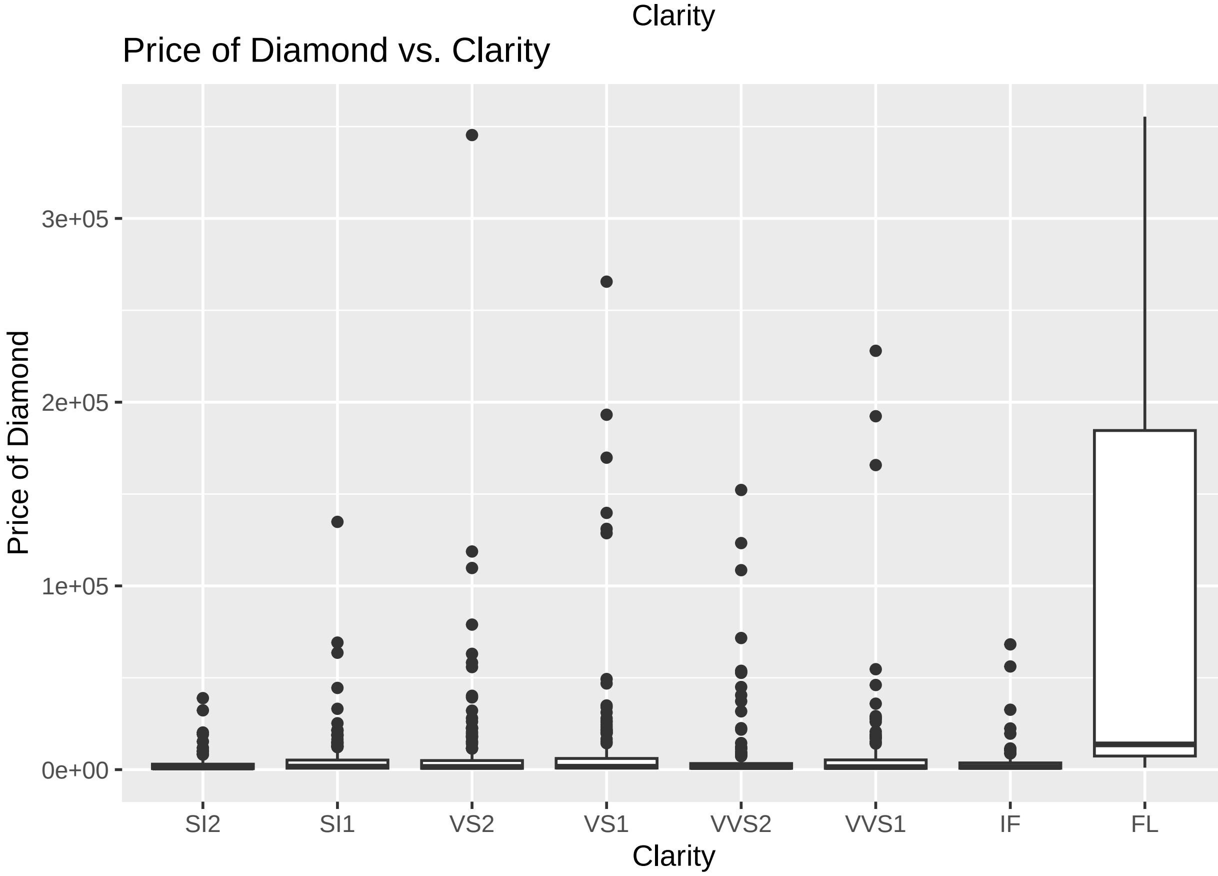
Price vs. Clarity



We can see from this scatterplot that the relationship between the price of a diamond and the carat weight appears exponential of some sort. As the carat weight increases, the price substantially increases.

Price vs. Clarity

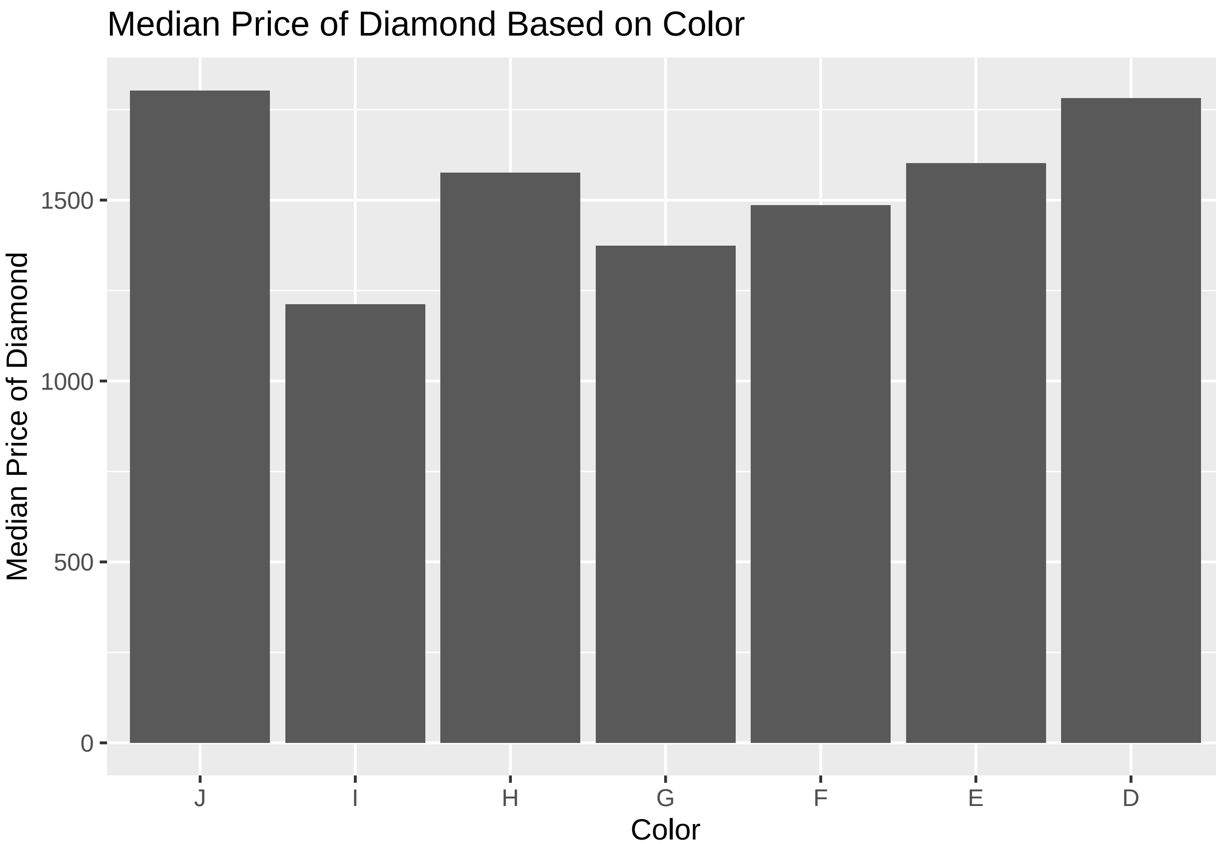


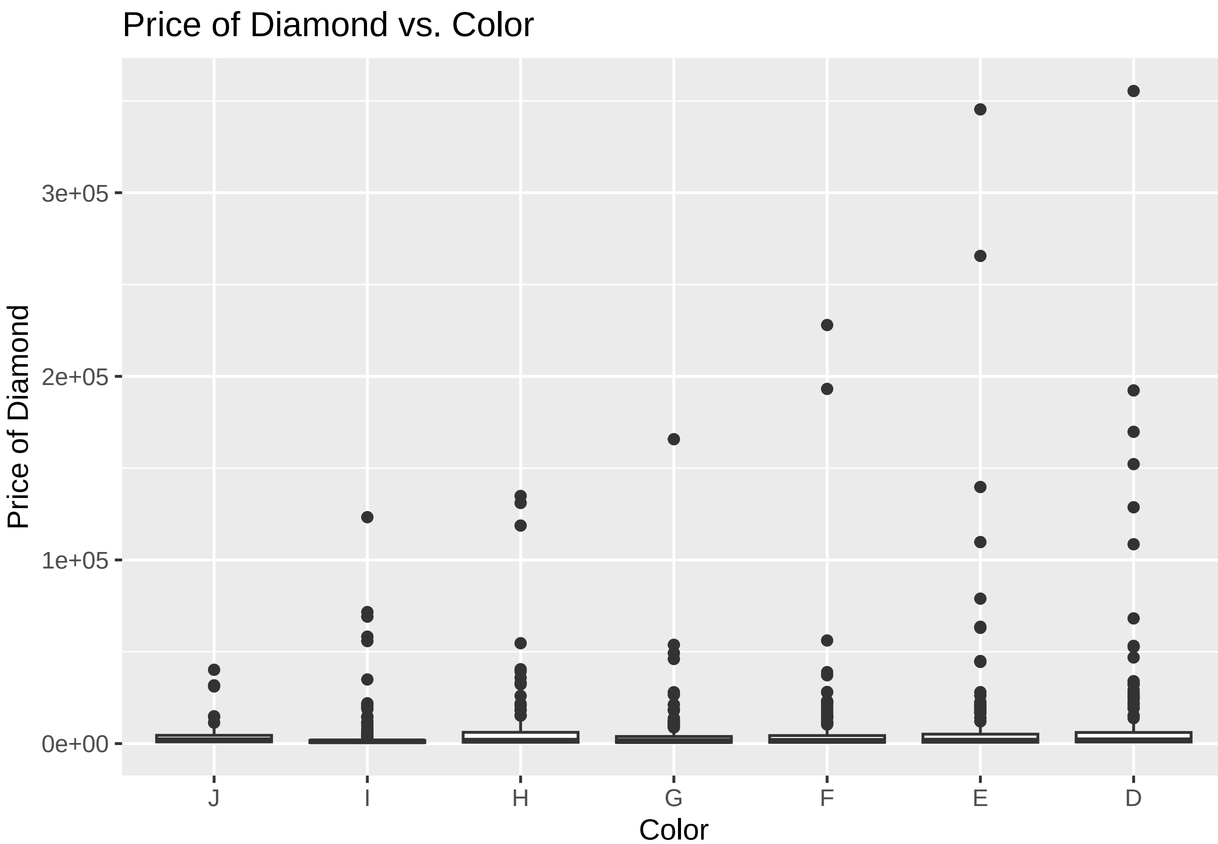


As we can see from the barplot above, the price increases dramatically when the clarity reaches the level of FL, which is flawless. Therefore, we can see that the median price of the diamond is not particularly affected by the clarity. However, based on the boxplot, the upper end of the prices for each quality do appear to have an upwards trend.

Price vs. Color

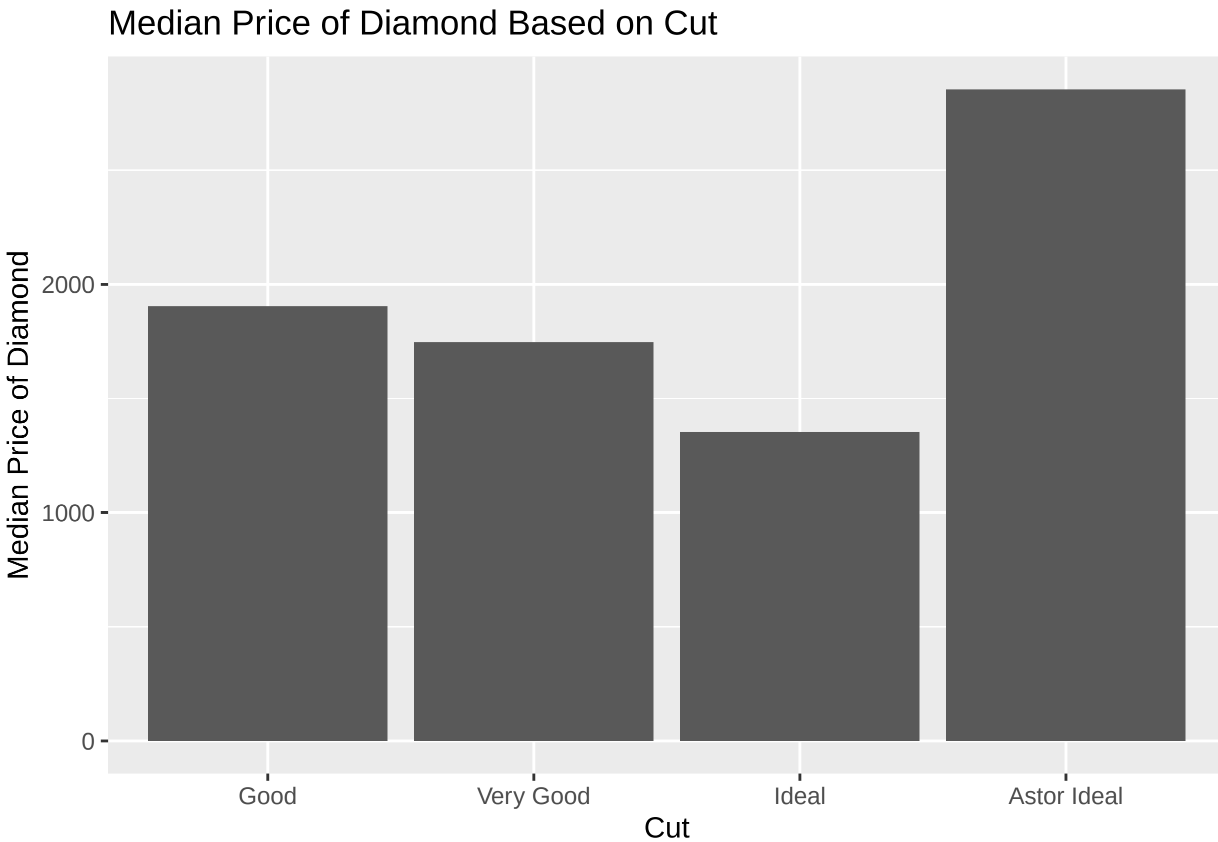
```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `color = color %>% fct_relevel(c("K", "J", "I", "H", "G", "F",
##   "E", "D"))`.
## Caused by warning:
## ! 1 unknown level in `f`: K
```

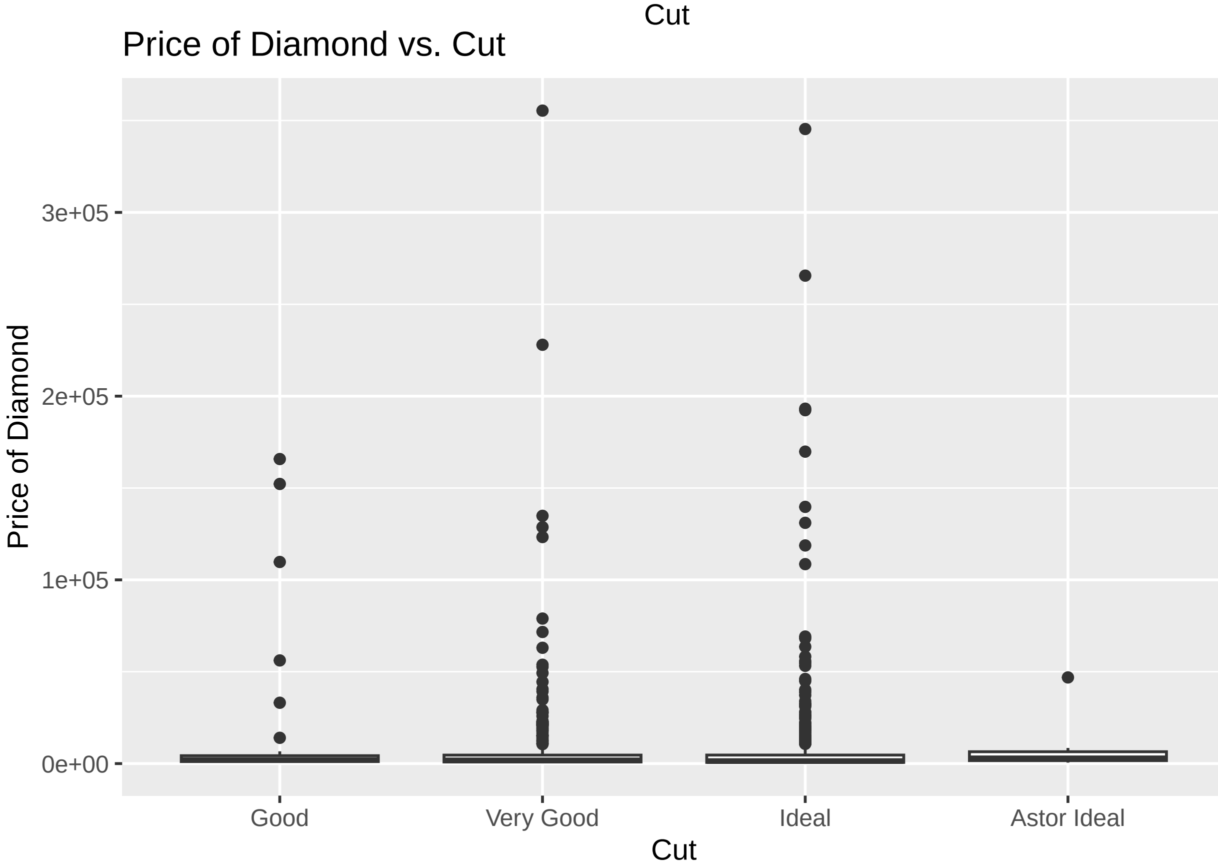




Based on the bar plot, we do not see much of a relationship between color and the price of the median diamond of each color. In fact, the highest median price for a diamond appears for the lowest color quality on the scale: J. However, when looking at the boxplots, the upper end of the diamonds sold at each color appear to have an upwards trajectory. That being said, the color does not appear to be a major factor in determining the price of a diamond.

Price vs. Cut


Median Price of Diamond Based on Cut
Median Price of Diamond
0
1000
2000
Good
Very Good
Ideal
Astor Ideal
Cut


Price of Diamond vs. Cut
Price of Diamond
0e+00
1e+05
2e+05
3e+05
Good
Very Good
Ideal
Astor Ideal
Cut

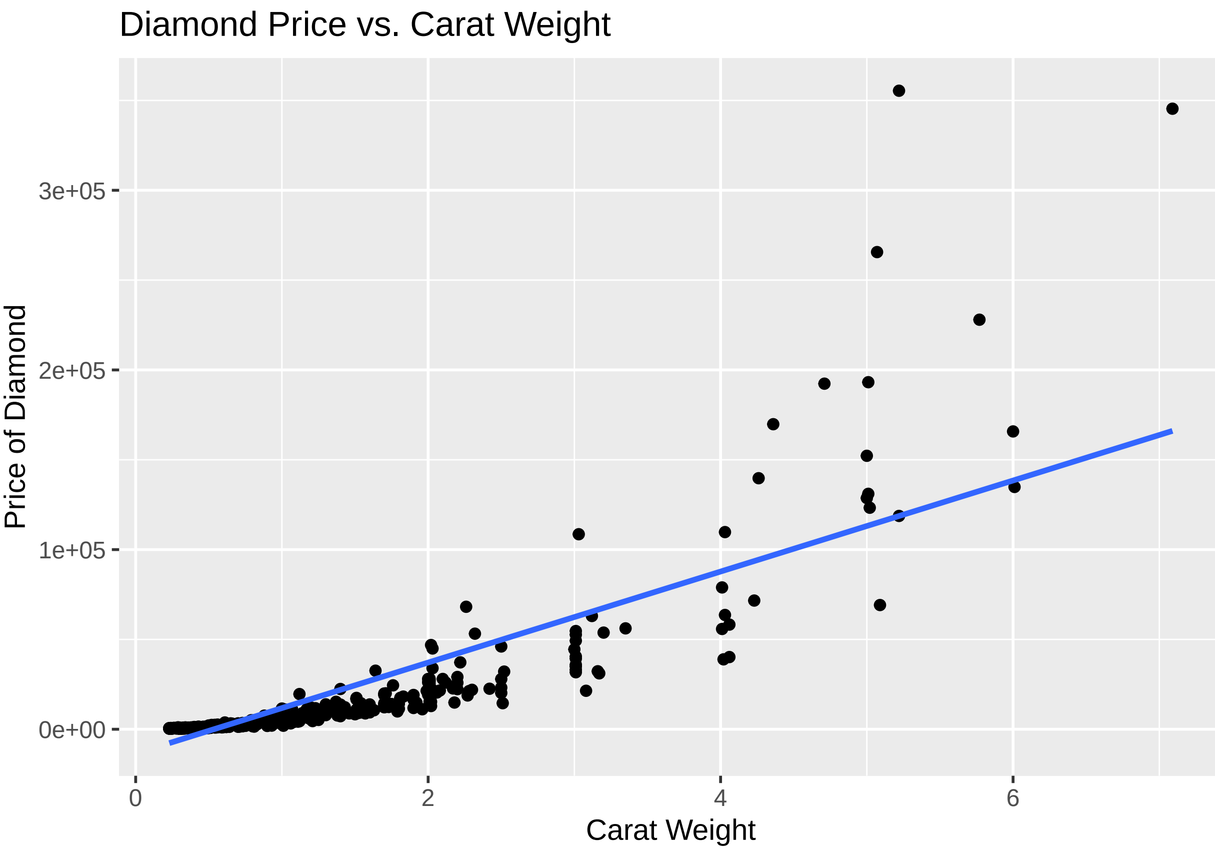
In a pattern very similar to the last two variables, we see that the bar plot does not show much of an upwards trend in median prices based on the cut of the diamond. In fact, with the exception of the Astor Ideal cut, the median prices decline as the cut improves. This trend continues when looking at the boxplot. There is no considerable upwards trend in even the upper end of prices as the cut of the diamonds improve. Based on this information, it appears that the cut is not a significant variable when it comes to determining the price of a diamond

Based on the analysis above, the variable that best appears to predict the price of a diamond is its carat weight, which agrees with the claims made on the BlueNile diamond education page. We will use the carat weight variable as the predictor variable moving forward in our regressions. However, based on the exponential appearance of the graph, some data transformations may be needed.

Section 4: Regression Description

First, we must take a look at the scatterplot of the response variable *price* against the designated predictor variable *carat*.

```
## `geom_smooth()` using formula = 'y ~ x'
```

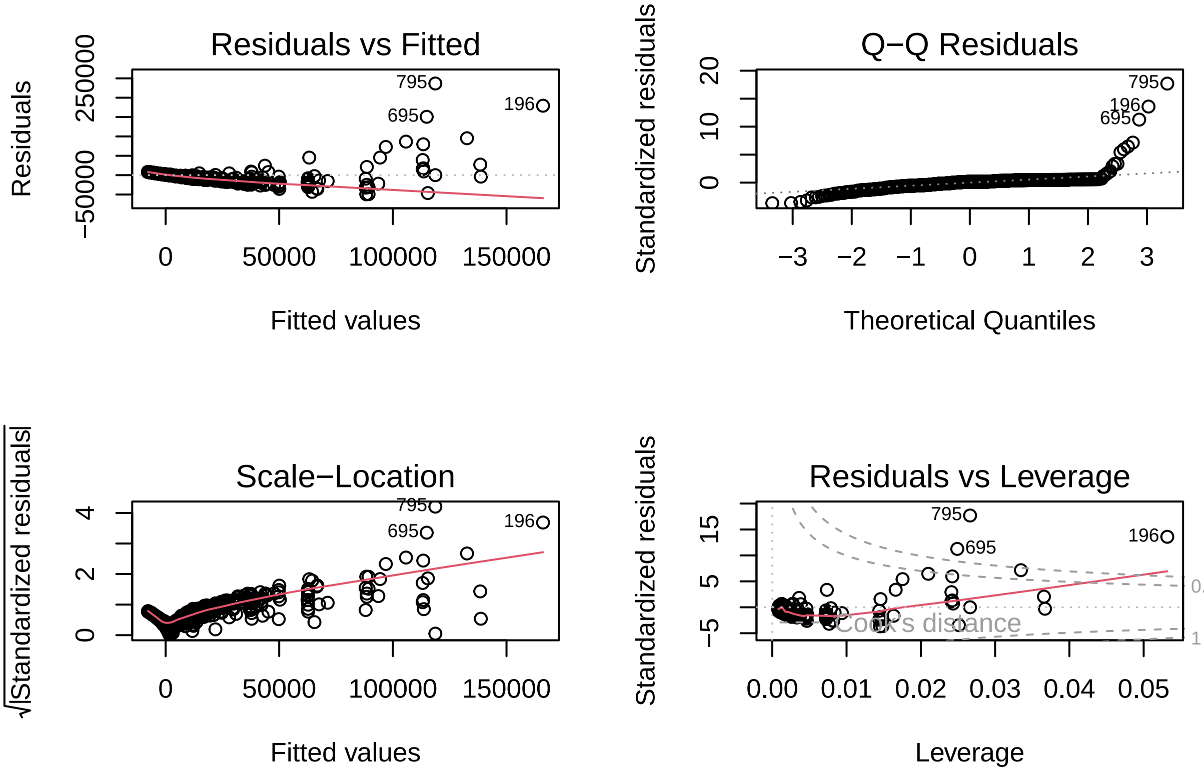


The scatterplot comparing carat and price looks exponential rather than linear. This would indicate that a transformation is needed, likely a log transformation.

As a reminder, the regression assumptions are as follows:

- The residuals have mean 0.
- The errors have constant variance.
- The errors are independent.
- The errors are normally distributed

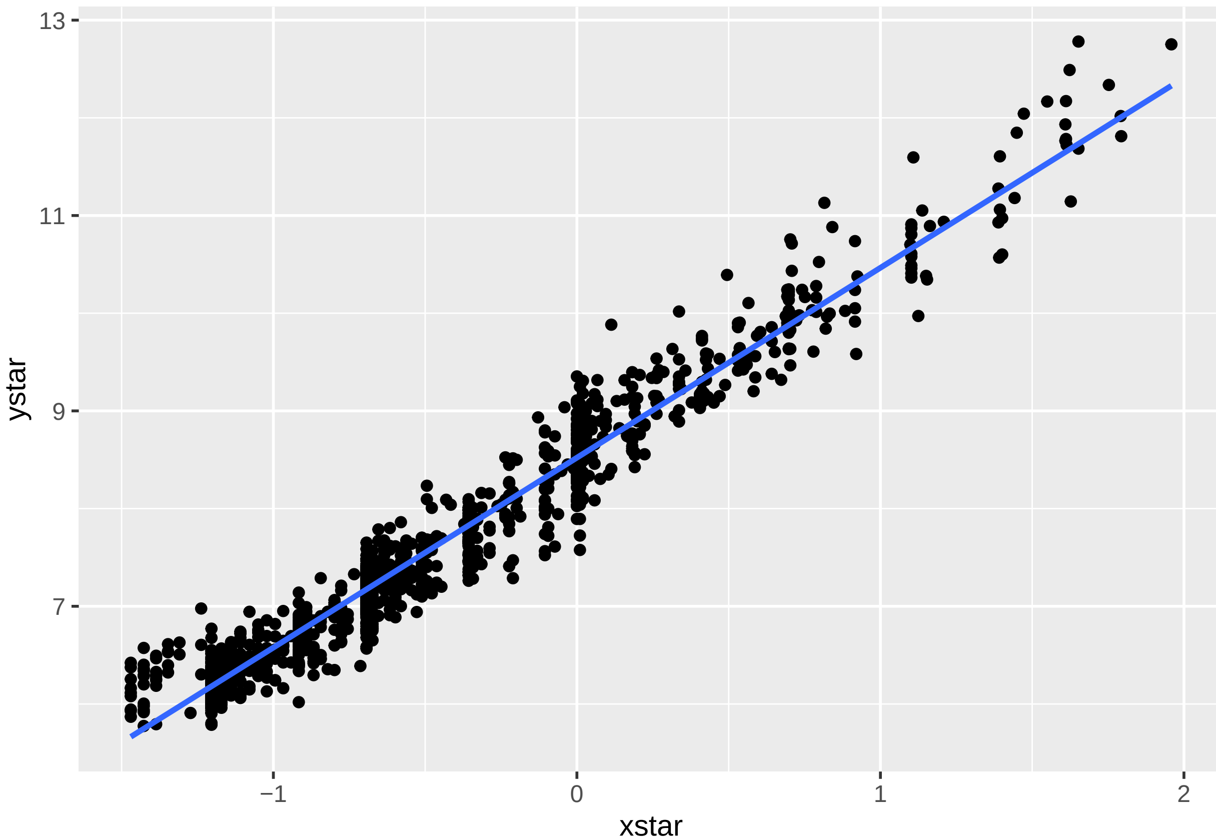
To confirm, we will take a look at the diagnostic plots.



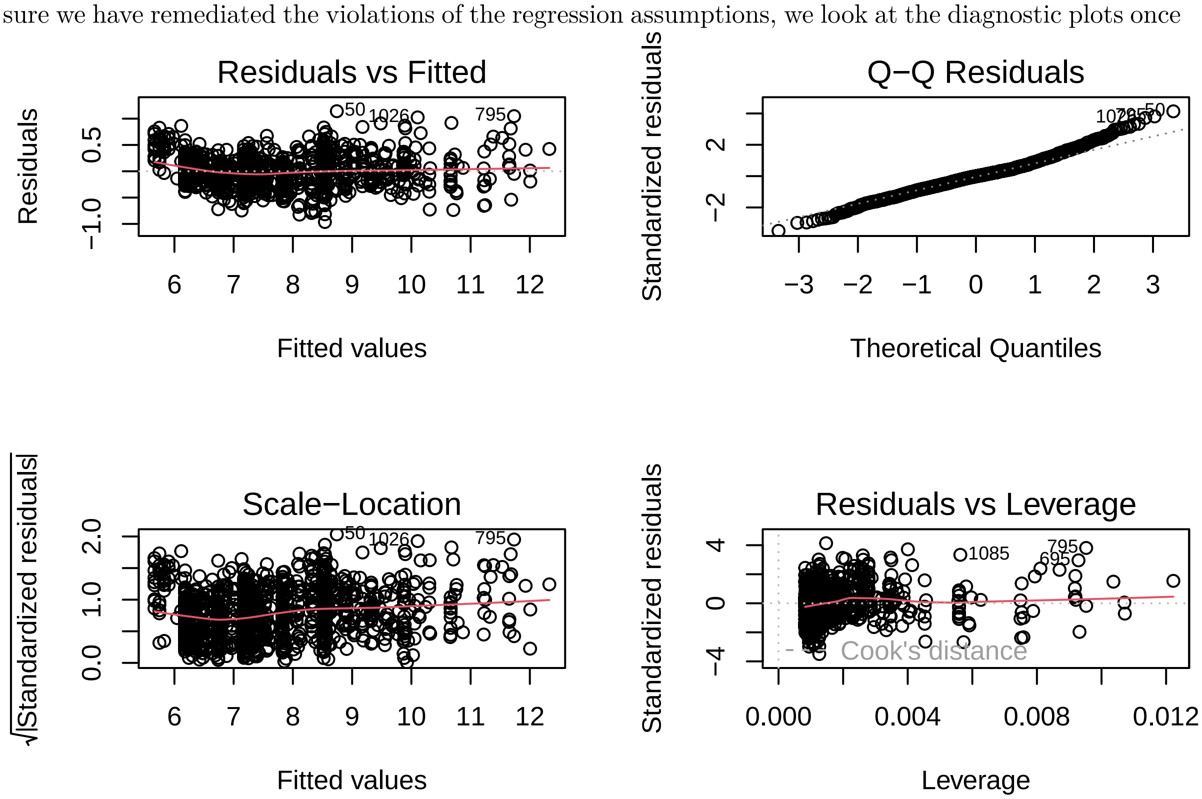
Based on the diagnostic plots as well as the Box-Cox plots, it seems that both regression assumptions 1 and 2 have been violated since the errors do not have mean 0 nor constant variance. The mean of the residuals appears to decrease, and the residuals appear to get more spread out. If both assumptions are violated, the best solution is to log transform both the x and y variables. The Box-Cox plot agrees with this since the confidence interval does not include 1, indicating the y variable must be transformed.

We will create the variable `xstar`, which is defined as $x^* = ln(carat)$ and the variable `ystar`, defined as $y^* = ln(price)$:

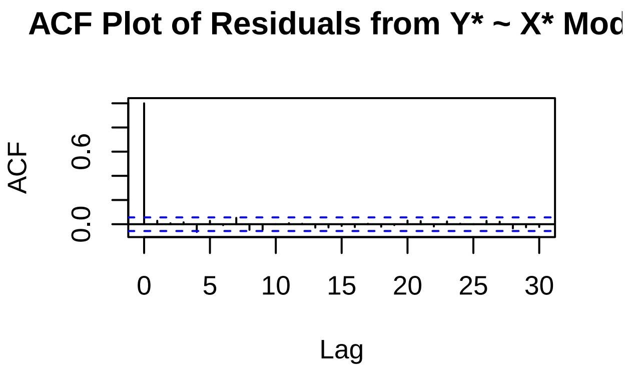
```
## `geom_smooth()` using formula = 'y ~ x'
```



As we can see from this scatterplot, the relationship appears much more linear than the first scatterplot before. To be sure we have remediated the violations of the regression assumptions, we look at the diagnostic plots once



more:



After log transforming both the `x` and `y` variables, the diagnostic plots also pass the assumption, with errors appearing to have both a mean of 0 and constant variance. Further, the Q-Q plot shows that the residuals have a fairly normal distribution. Based on the ACF plot, we can also see that the errors are independent. Thus, all four regression assumptions are met.

So, our hypotheses are:

$H_0 : \beta_1 = 0$

$H_0 : \beta_1 \neq 0$

When we run the regression, we see:

```
##
## Call:
## lm(formula = ystar ~ xstar, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96394 -0.17231 -0.00252  0.14742  1.14095
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.521208   0.009734   875.4   <2e-16 ***
## xstar       1.944020   0.012166   159.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2761 on 1212 degrees of freedom
## Multiple R-squared:  0.9547, Adjusted R-squared:  0.9546
## F-statistic: 2.553e+04 on 1 and 1212 DF,  p-value: < 2.2e-16
```

Based on the results of the model with both the response and the predictor being log transformed, `xstar`, $x^* = ln(carat)$, is statistically significant with a p-value of 2.2×10^{-16} much lower than 0.05. This means that the log of carat is significant in predicting the log of the price of the diamond.

The regression equation is $ln(price) = 8.521 + 1.944 ln(carat)$. This means that for every 1% increase in the carat weight, the price of the diamond is expected to increase by around 1.944%.

The 95% confidence interval is:

```
##                 2.5 %   97.5 %
## (Intercept) 8.502110 8.540306
## xstar       1.920152 1.967888
```

Which shows that we are 95% confident that for each 1% increase in the carat weight of a diamond, its price will increase between 1.92% and 1.97%.