Stat 6021: Project 1

Group 2: Ben G. Ballard, Sirish Kumar Desai, Kevin Kuc, Tom Lever

10/01/22

Executive Summary

Blue Nile, Inc. contacted our team of data analysts to review a data set containing data relating to 1214 various diamonds for sale on Blue Nile's website. Our team was tasked with exploring the relationships between key attributes of the diamonds: weight (carat); clarity; color; cut; and price. Below, we further describe these data and variables and present visualizations and analysis. Our visualizations illustrate proportions of diamonds across groups of variables. Our analysis describes trends within the visualizations and evaluates claims on Blue Nile's website.

Cut refers to how ideally a diamond is shaped, and consists of a range from *Good* to $Astor\ Ideal$. $Astor\ Ideal$ diamonds have the highest median price followed by $Ideal$, $Very\ Good$, and *Good* diamonds. Blue Nile implies that $Astor\ Ideal$ diamonds gather and reflect the most light possible, and may on this basis give $Astor\ Ideal$ diamonds relatively high prices.

Color is the second most significant factor that influences the price of a diamond. As the diamond becomes more colorful the diamond generally decreases in value. The Gemological Institute of America (GIA) color scale is the industry standard for diamond grading. The scale starts with D for least colorful diamonds and alphabetically proceeds to Z for most colorful diamonds (although Blue Nile only carries diamonds with color identifiers in the range D through J).

The third factor that affects a diamonds price is clarity, which is an assessment of small imperfections on the surface and within the stone. The clarity of a diamond ranges from $SI2$, which has inclusions observable by a human under $10x$ magnification, to an FL diamond, which is flawless. Flawless diamonds have the highest median price among diamonds grouped by clarity. Their lack of abundance drives up their price; less than 1 percent of all diamonds have FL clarity. The $VS1$ clarity identifier corresponds to a group of diamonds with the second highest median price, although four groups of diamonds have greater clarity.

The weight of a diamond is measured in carats; Blue Nile and others refer to the weight of a diamond as carat weight and carat. The weight of a diamond is not always indicative of a diamond's size. That being said, we found that as weight increases, the price of a diamond also increases, at an increasing rate, with increasing variability. In "Simple Linear Regression of Price versus Weight", we determine a linear model that relates transformed price to transformed weight. This linear model may be used to predict diamond price given weight. For an increase in weight of 10 percent, price increases by a factor of 1.2. For example, for an increase in weight from 1 *carat* to 1.1 *carat*, price increases from $\$5,019.07$ by a factor of 1.2 to $\$6,040.75$. These prices have a ratio of 1.2.

Overall, the cut, color, clarity, and carat of a diamond are all predictors for a diamond's market value, with various qualifications.

Presentation and Analysis of Data

Description of Data and Variables

Our data set describes 1214 difference diamonds that are for sale at http://www.bluenile.com. Our data set describes diamonds with carat, clarity, color, cut, and price data. Our data set describes a subset of the diamonds that are for sale with a subset of features. Table 1 presents data for three diamonds.

Table 1: first three diamonds in our data set

weight	clarity	color	cut	price
0.51	SI2	I	Very Good	774
0.93	IF	H	Ideal	6246
0.50	VVS2	D	Very Good	1146

Weight measures a diamond's weight in carats. Clarity assesses small imperfections within a diamond and quantifies and specifies inclusions. Color refers to how colorless a diamond is. Cut measures how well-proportioned a diamond's dimensions are.

General Analysis, and Visualizations and Visualization-Specific Analysis

General Analysis

This report explores the relationships between key attributes of the diamonds, specifically, carat, clarity, color, cut, and finally price. Price is the attribute we are most interested in. We are interested in the other attributes in as far as they can be used to evaluate price. We first explored the relationship of the attributes, or variables, by exploring one variable at a time, known as univariate analysis. Then we explored relationships between two or more variables at a time, known as multivariate analysis. We did this through data visualization and descriptive statistics.

We start by analyzing the data systematically, one variable at a time, through data visualization. As the saying goes, a picture is worth a thousand words. An effective graphic can quickly communicate a lot about a dataset's describe statistics. Common univariate data visualizations graphics are histograms, density plots, and boxplots. These visualizations show the count or proportion of the variable in question as well as the dispersion or spread of the data. We use multivariate graphics to quickly understand the relationship of two or more variables. Common data visualizations include scatterplots, boxplots, and histograms.

We identified high-level findings relating to cut, color, clarity, and weight and their relationships to price. In some attributes, characteristics clearly had higher prices. However, in some attributes, the way those attributes related to price was less obvious. For example, within Clarity, there is a designation 'FL' that has much higher prices than any other clarity designation. Whereas some attributes, such as Color, one color attribute does not stand out as having a higher price than the others.

Over the course of the analysis, we came across claims on the BlueNile website, and we evaluated them against what the data says. Xxx Finally, our team fit a linear model to the data with price as the response variable. Our data analysis team then fit the data to a simple linear regression model. A Simple linear regression is a statistical method that allows us to summarize and study relationships between variables. Our intent is to use this simple linear regression to better understand the relationship between the attributes of a diamond and price. We can also use this technique to predict the price of a diamond, given knowledge of the other attributes. To use a linear regression model, five assumptions in the data must be met. Initially these assumptions were not met. However, there are methods an analyst can employ to allow the 'transform' the data such that these assumptions are then met.

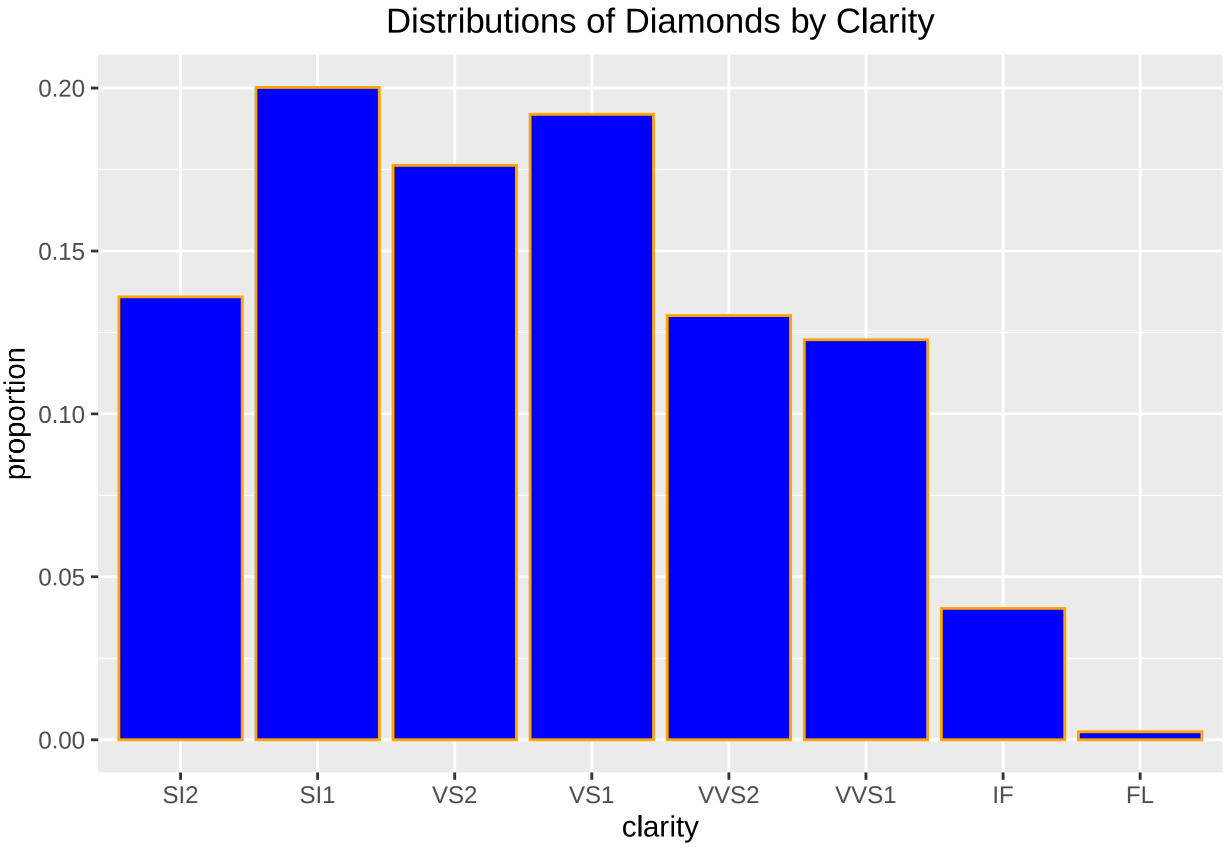
Considering the claim at https://www.bluenile.com/education/diamonds that "The higher [in idealness] a diamond is, the more expensive it will be", this claim is supported by a trend that as diamonds improve in cut, third-quartile price increases; a trend that as diamonds improve in cut from good to very good to ideal, maximum price increases; and the fact that Astor ideal diamonds have the highest minimum / first-quartile / median / third-quartile price. However, this claim is refuted by a trend that as diamonds improve in cut from good to very good to ideal, minimum / first-quartile / median price decreases.

Considering the claim that "The higher [in color] a diamond is, the more expensive it will be", this claim is supported in by a trend that for a transition from a group of diamonds with a color identifier in the set {"G", "F", "E", "D"} to a group of diamonds with a color identifier closer to "D", the min / first-quartile / median / third-quartile / max price increases.

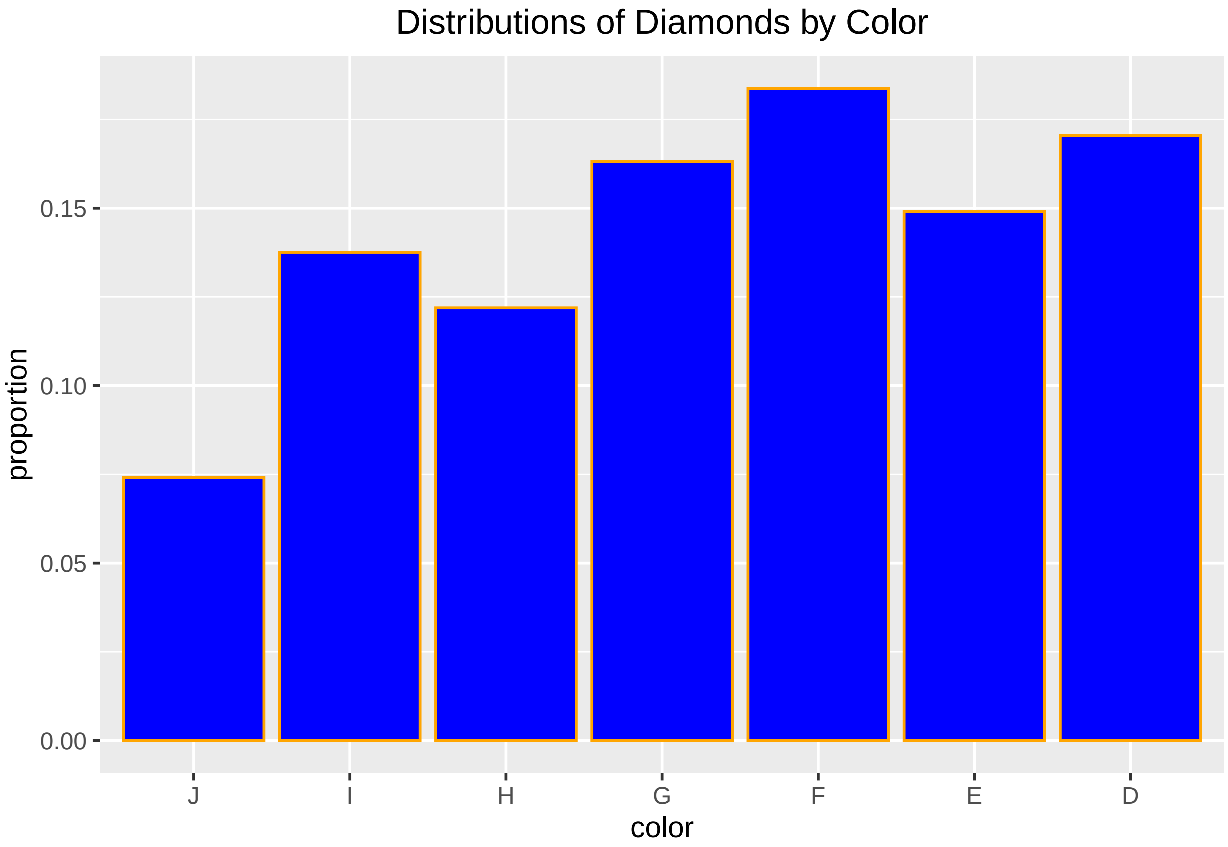
Considering the claim that "The higher [in weight] a diamond is, the more expensive it will be," we find in a scatterplot of price versus weight a trend that s weight of a diamond increases, price of the diamond increases, at an increasing rate.

Detailed Visualizations and Plot-Specific Analysis

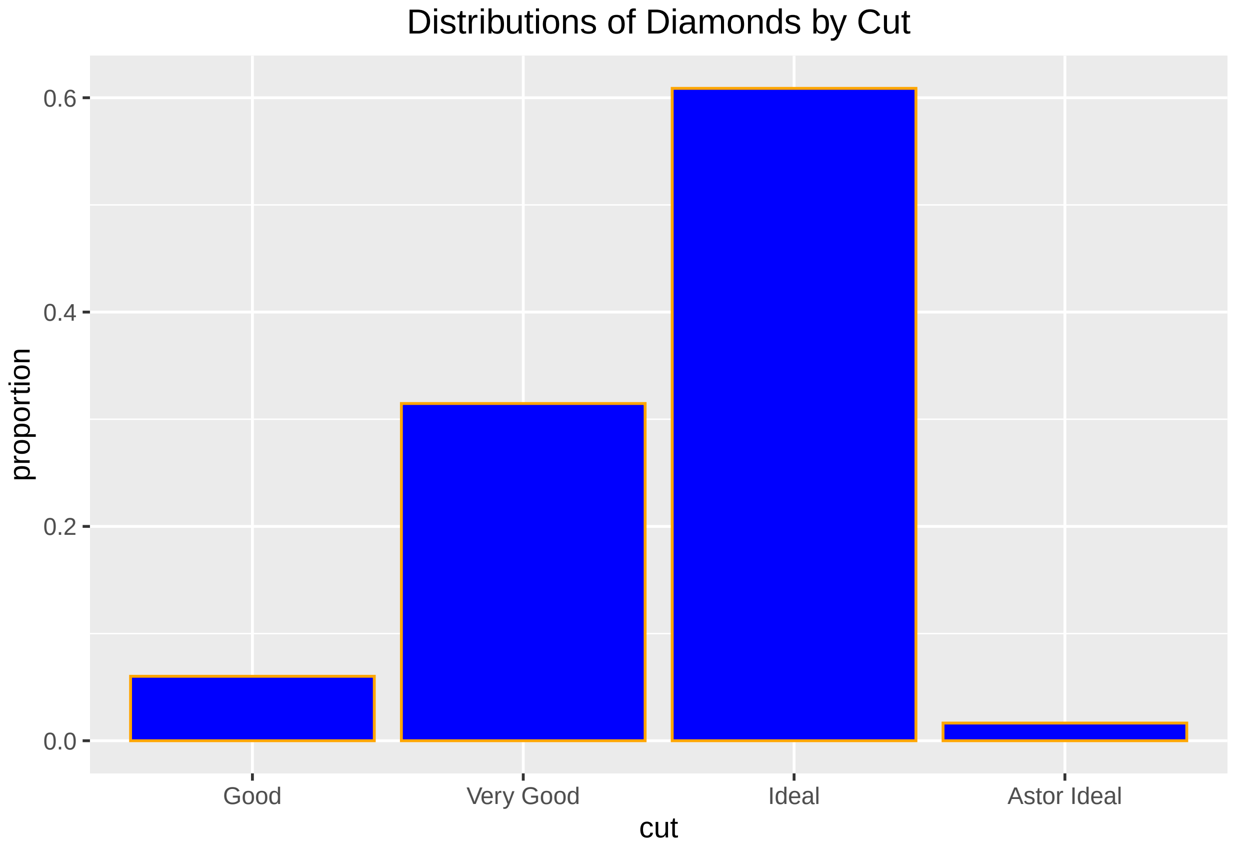
Considering proportions of diamonds by clarity, the group of diamonds with clarity identifier $SI1$ has the highest proportion of about 0.2. For a transition from a group of diamonds with a clarity identifier in the set $\{VS1, VVS2, VVS1, IF, FL\}$ to a group of diamonds with clarity identifier closer to FL, proportion of diamonds decreases. The group of diamonds with clarity identifier FL has the lowest proportion of less than 0.01.



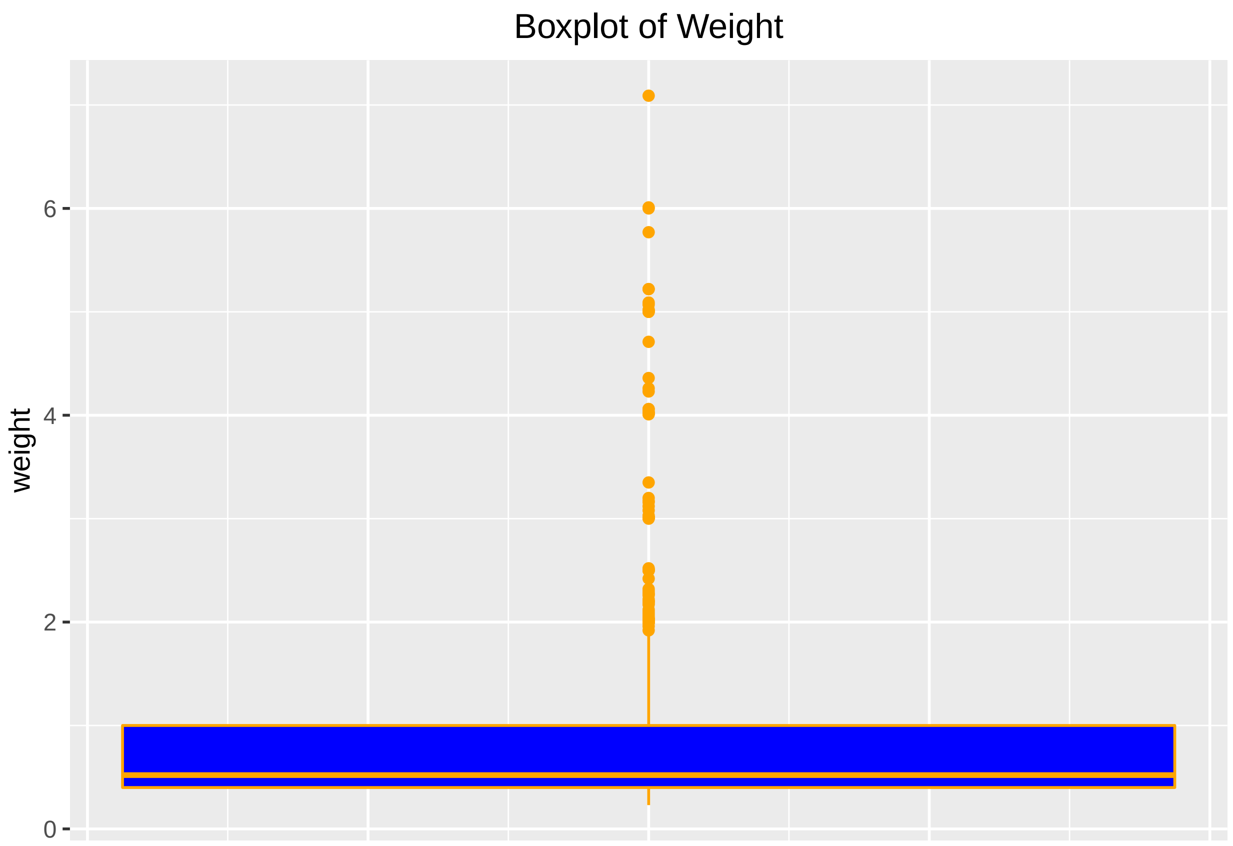
Considering proportions of diamonds by color, the group of diamonds with color identifier F has the highest proportion of over 0.175. The group of diamonds with color identifier J has the lowest proportion of about 0.075. There seems to be a general trend of proportion increasing as color identifier approaches D.

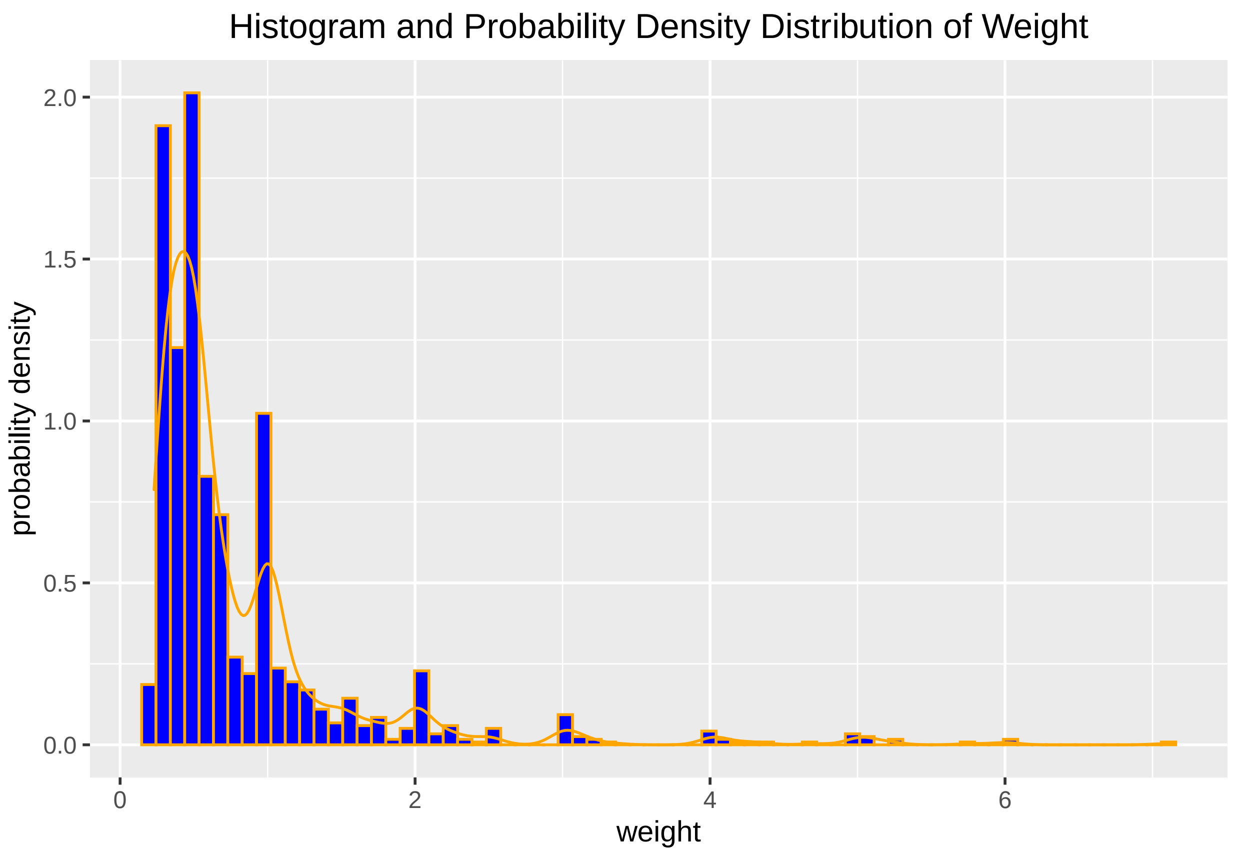


Considering proportions of diamonds by cut, the group of ideal diamonds has the highest proportion of over 0.6. The group of Astor ideal diamonds has the lowest proportion of well under 0.005. As cut improves from good to ideal, proportion of diamonds increases.

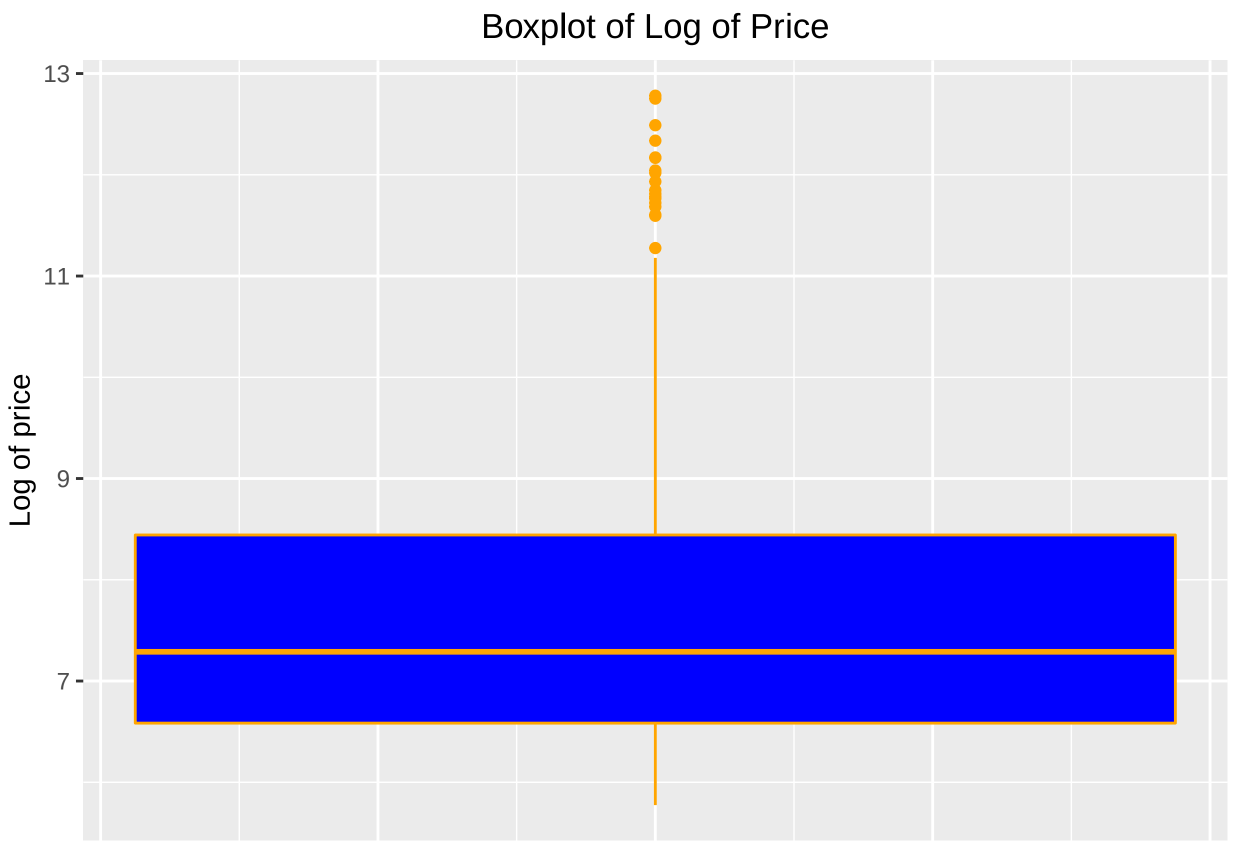


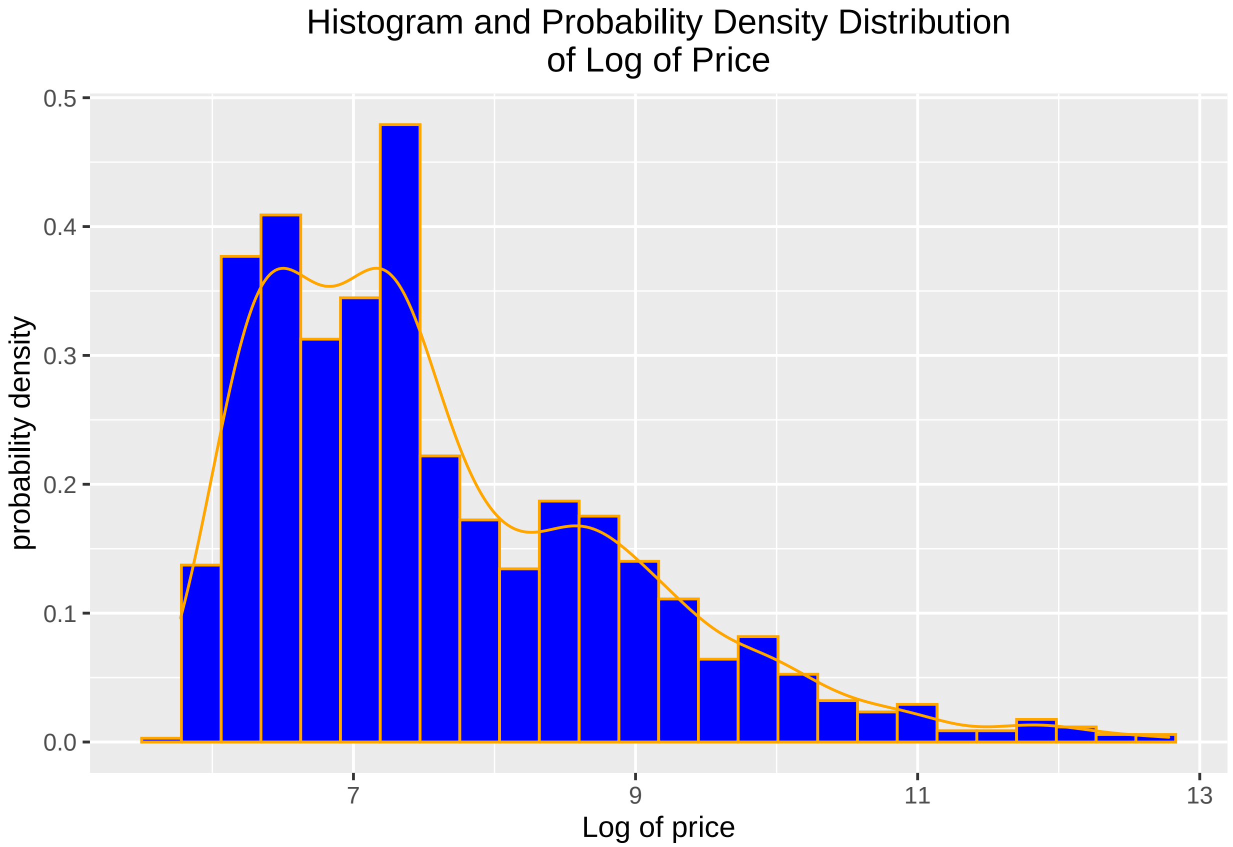
Considering a boxplot and a probability density distribution for weight, the middle 50 percent of weights fall between about 0.4 *carat* and 1.0 *carat*. The distribution is non-normal and positively / right skewed.


Boxplot of Weight
weight
0
2
4
6

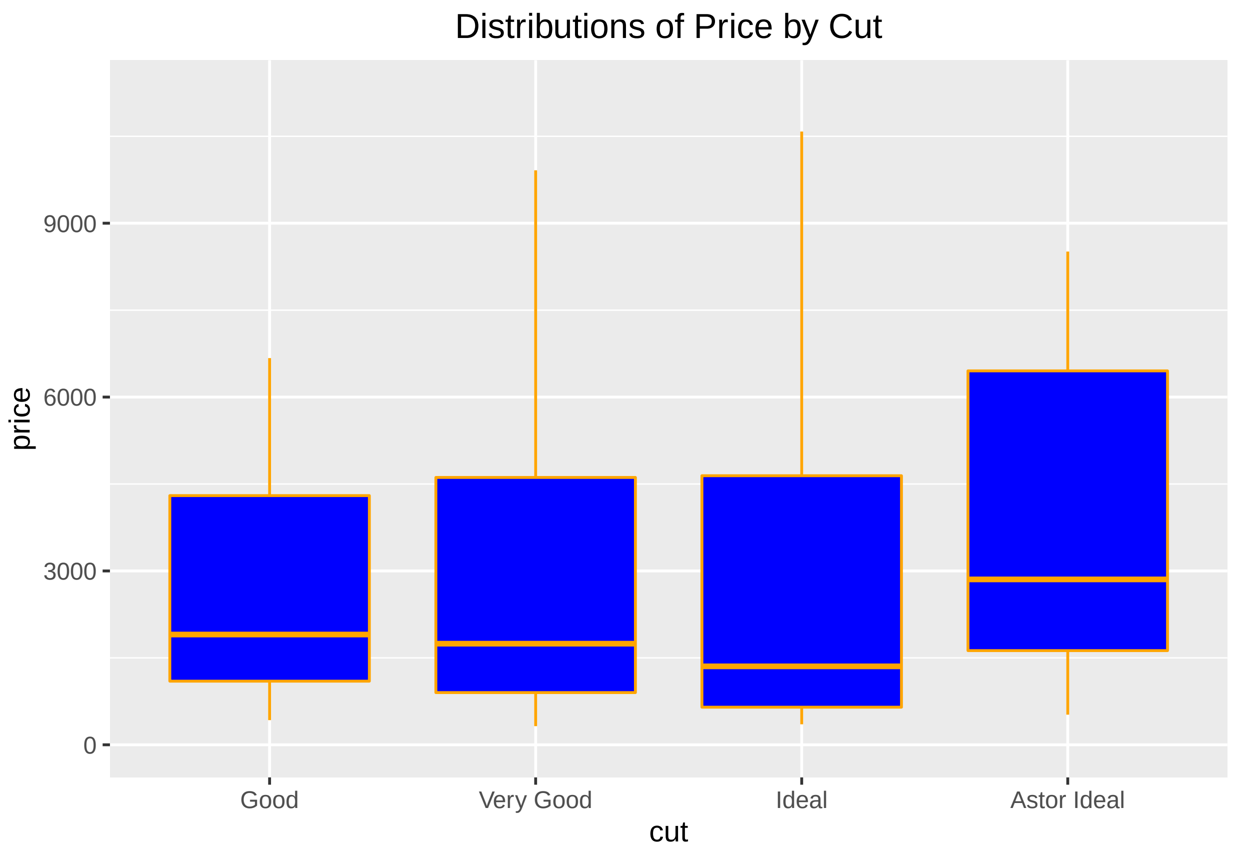


Considering a boxplot and a probability density distribution for price, the middle 50 percent of logarithms of prices fall between 6.5 and 7.5. The distribution is non-normal and positively / right skewed.

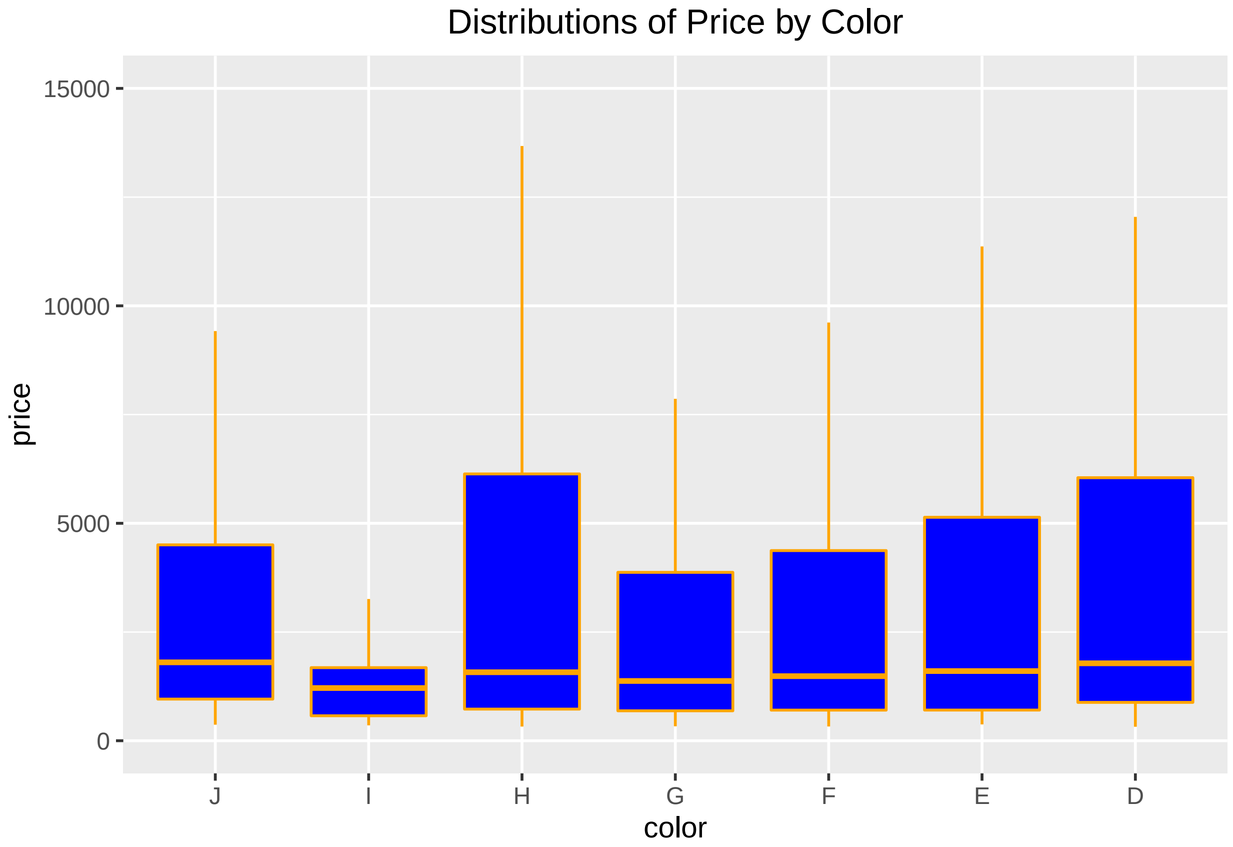

Boxplot of Log of Price
Log of price
13
11
9
7



Considering the relationship of price of a diamond in our data set and the cut of the diamond, we construct boxplots of price versus cut with and without outliers and present the boxplot without outliers. Including outliers, a very good diamond has the highest price of over $350, 000. Excluding outliers, an ideal diamond has the highest price at about $10, 500. A very good diamond has the lowest price at $322. The minimum / first-quartile / median price and interquartile range of prices of Astor ideal diamonds are highest. The minimum / first-quartile / median prices of good, very good, and ideal diamonds decrease in that order, and are less than the minimum / first-quartile / median price of Astor ideal diamonds. The third-quartile price of Astor ideal diamonds is highest at about $6, 500. The third-quartile prices of good, very good, and ideal diamonds increase in that order, and are less than the third-quartile price of Astor ideal diamonds.

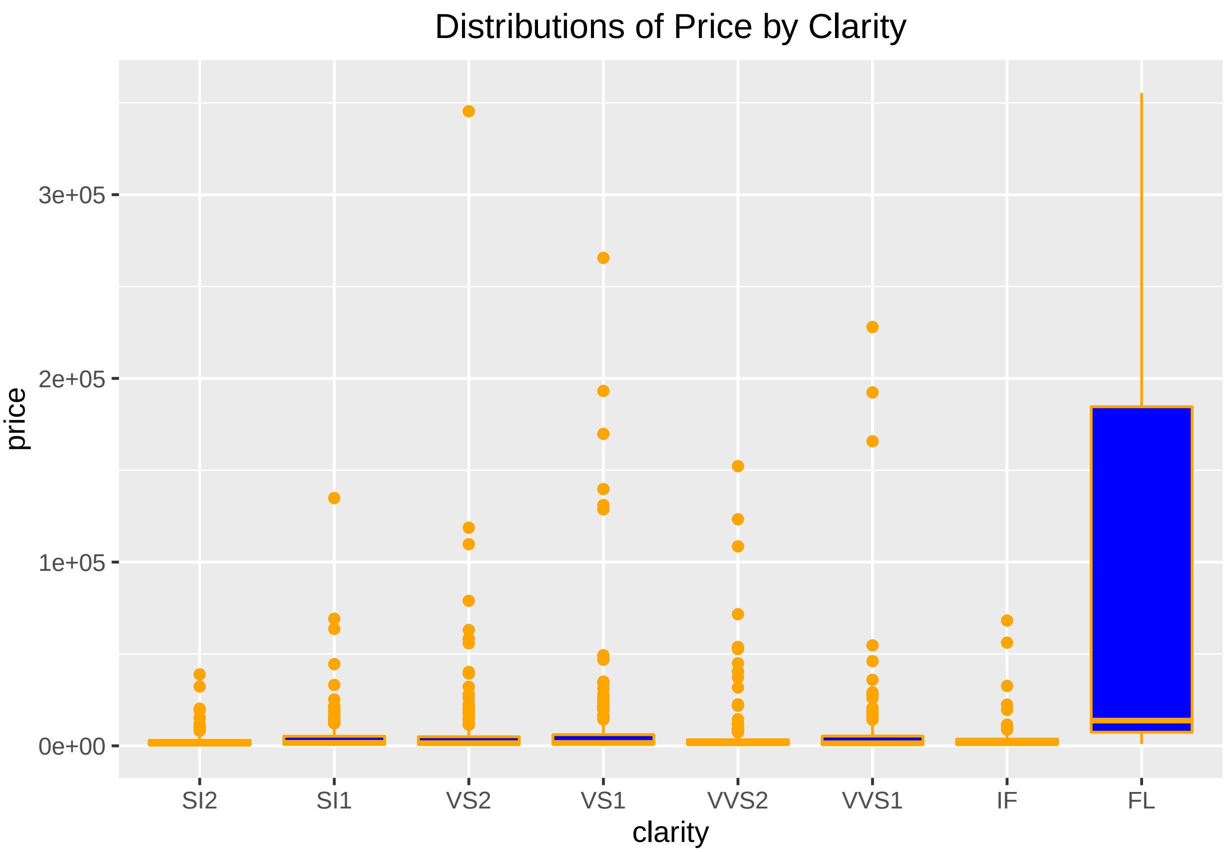


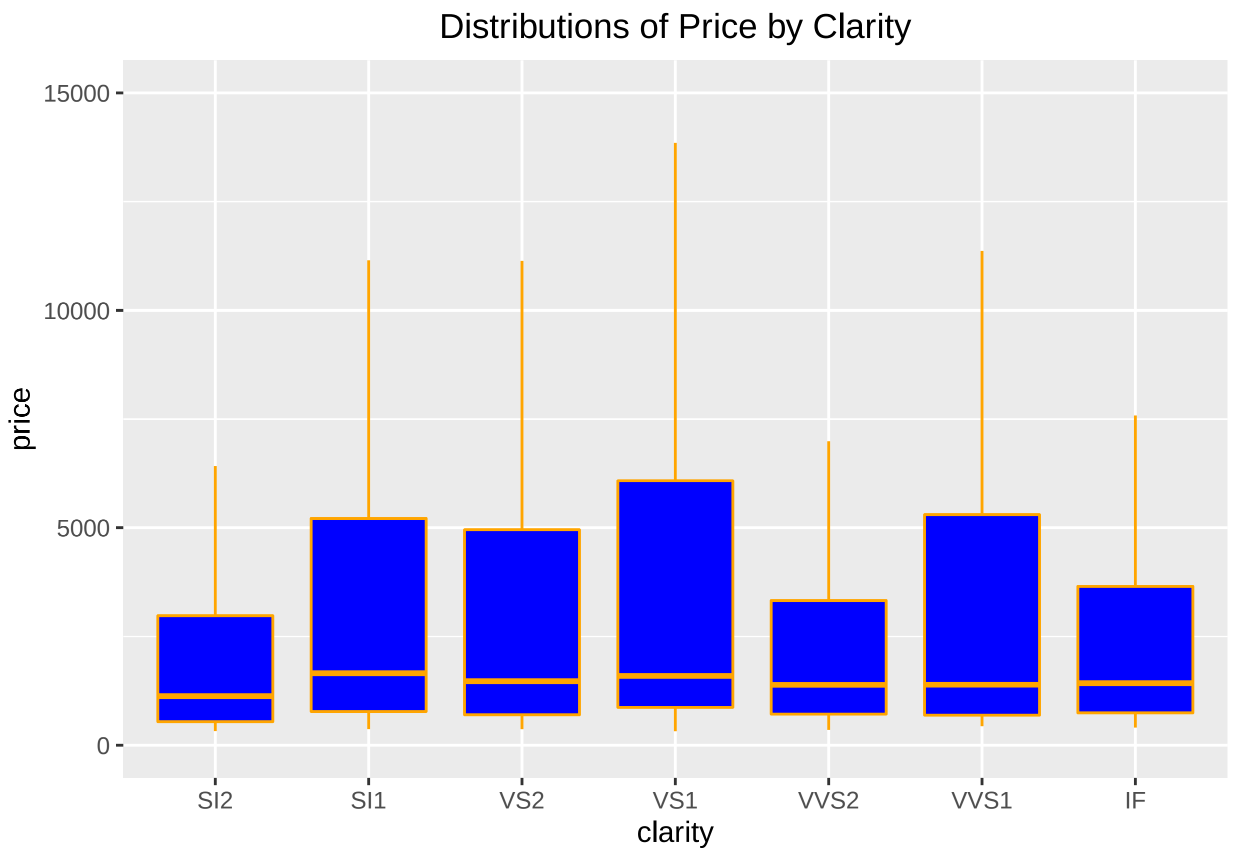
Considering the relationship of price of a diamond in our data set and the color of the diamond, we construct boxplots of price versus color with and without outliers and present the boxplot without outliers. Including outliers, for a transition from a group of diamonds with a color identifier in the English alphabet to a group of diamonds with a color identifier of the next letter closer to the beginning of the alphabet, maximum price of a diamond increases from around $\$50,000$ to around $\$350,000$. Excluding outliers, a group of diamonds with color identifier H has the highest price of about $\$13,750$ and the highest interquartile range of prices. A diamond with color identifier D has the lowest price of $322. For a transition from a group of diamonds with a color identifier G to a group of diamonds with a color identifier of a letter closer to the beginning of the alphabet, the first-quartile / median / third-quartile price of a diamond increases.



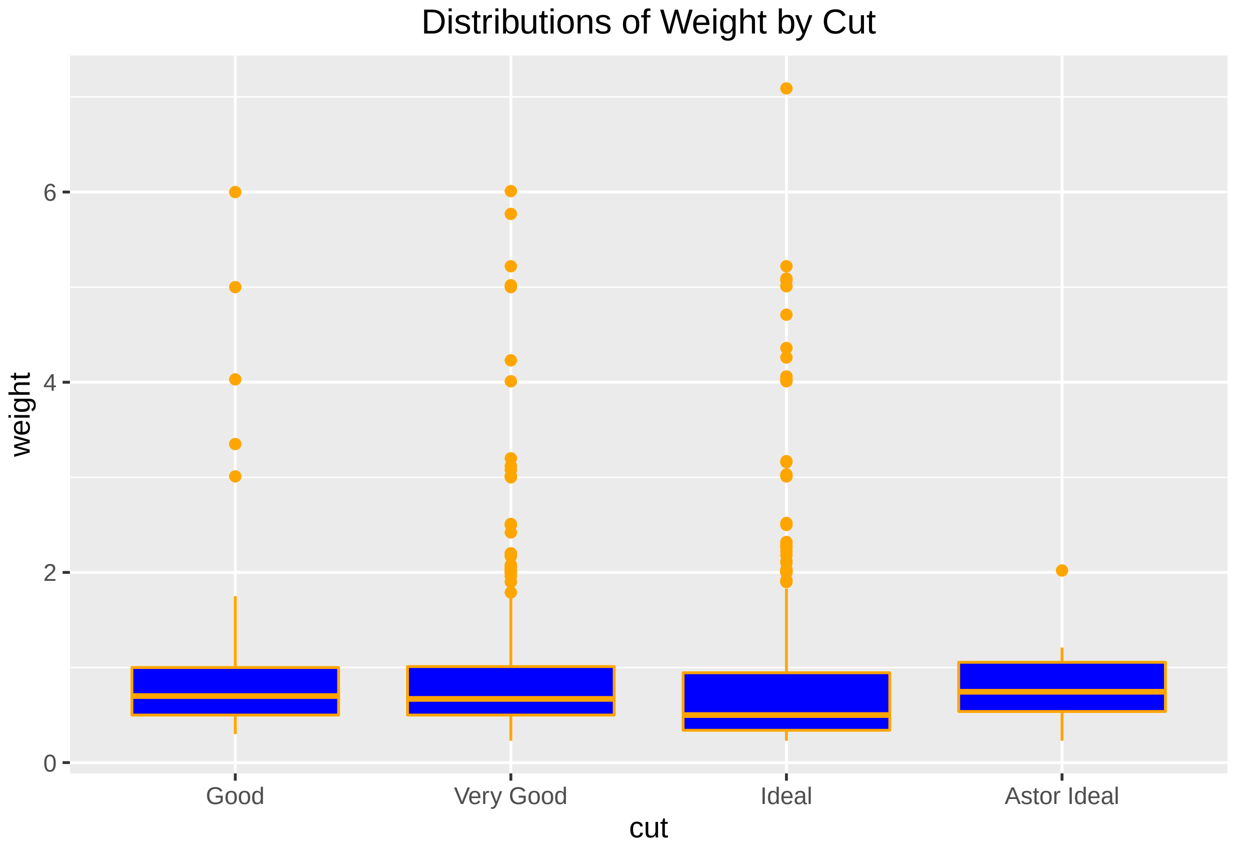
Considering the relationship of price of a diamond in our data set and the clarity of the diamond, we construct boxplots of price versus clarity with outliers and group of diamonds with clarity identifier FL, and without outliers and without the group of diamonds with clarity identifier FL. Including outliers and the group of diamonds with clarity identifier FL, a diamond with clarity identifier FL has the highest first-quartile / median / third-quartile / maximum price and interquartile range of prices. Excluding outliers and the group of diamonds with clarity identifier FL, a diamond with clarity identifier $VS1$ has the lowest price. A diamond with clarity identifier $VS1$ has the highest first-quartile / third-quartile / maximum price and interquartile range of prices. A diamond with clarity identifier $SI1$ has the highest median price.

Considering the claim at https://www.bluenile.com/education/diamonds that “The higher [in clarity] a diamond is, the more expensive it will be”, this claim is supported in that flawless diamonds have the highest first-quartile / median / third-quartile / maximum price. However, neither when including nor excluding outliers is this claim supported by a trend that price increases with clarity.

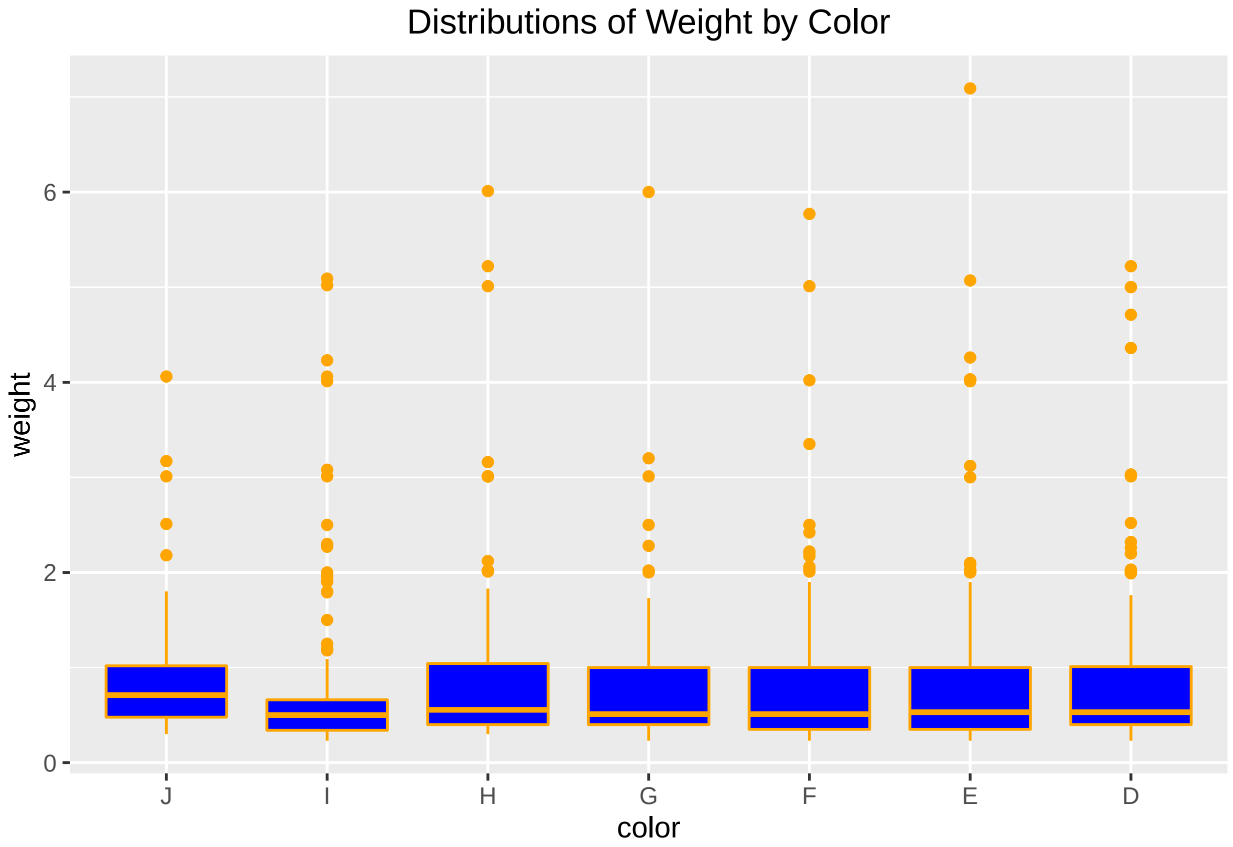

Distributions of Price by Clarity
price
3e+05
2e+05
1e+05
0e+00
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
FL
clarity



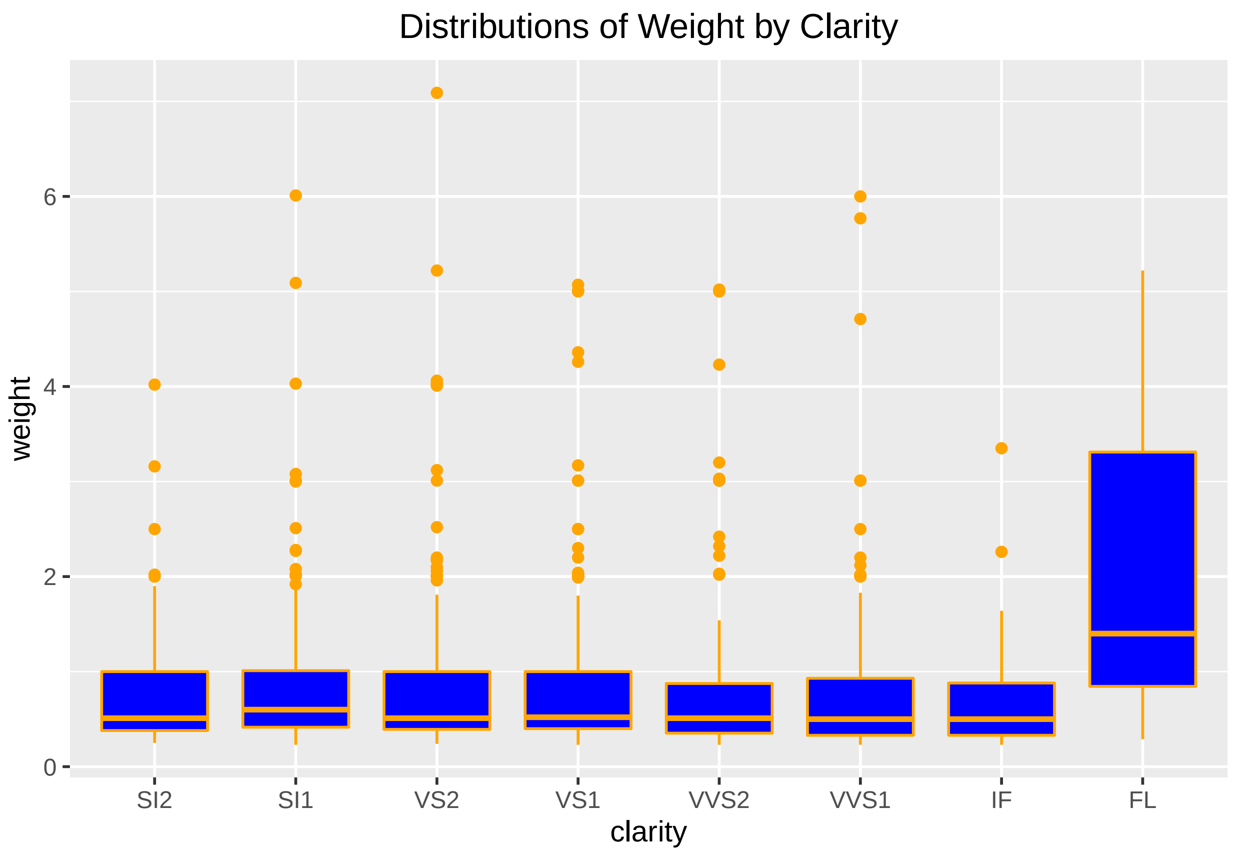
Considering the relationship of weight of a diamond in our data set and the cut of the diamond, we construct a boxplot of weight versus cut. Including outliers, an ideal diamond has the highest weight of over 5 *carat*. Excluding outliers, an ideal diamond the highest weight of about 2 *carat*. An Astor ideal diamond has the lowest weight of 0.23 *carat*. Astor ideal diamonds have the highest first-quartile, median, and third-quartile weights. Ideal diamonds have the greatest interquartile range of weights. The median weights of good, very good, and ideal diamonds decrease in that order.



Considering the relationship of weight of a diamond in our data set and the color of the diamond, we construct a boxplot of weight versus color. Including outliers, a diamond with color identifier E has the highest weight of over 7 $carat$. Excluding outliers, a diamond with color identifier F or E has the highest weight of around 2 $carat$. A diamond with color identifier F has the lowest weight of 0.23 $carat$.

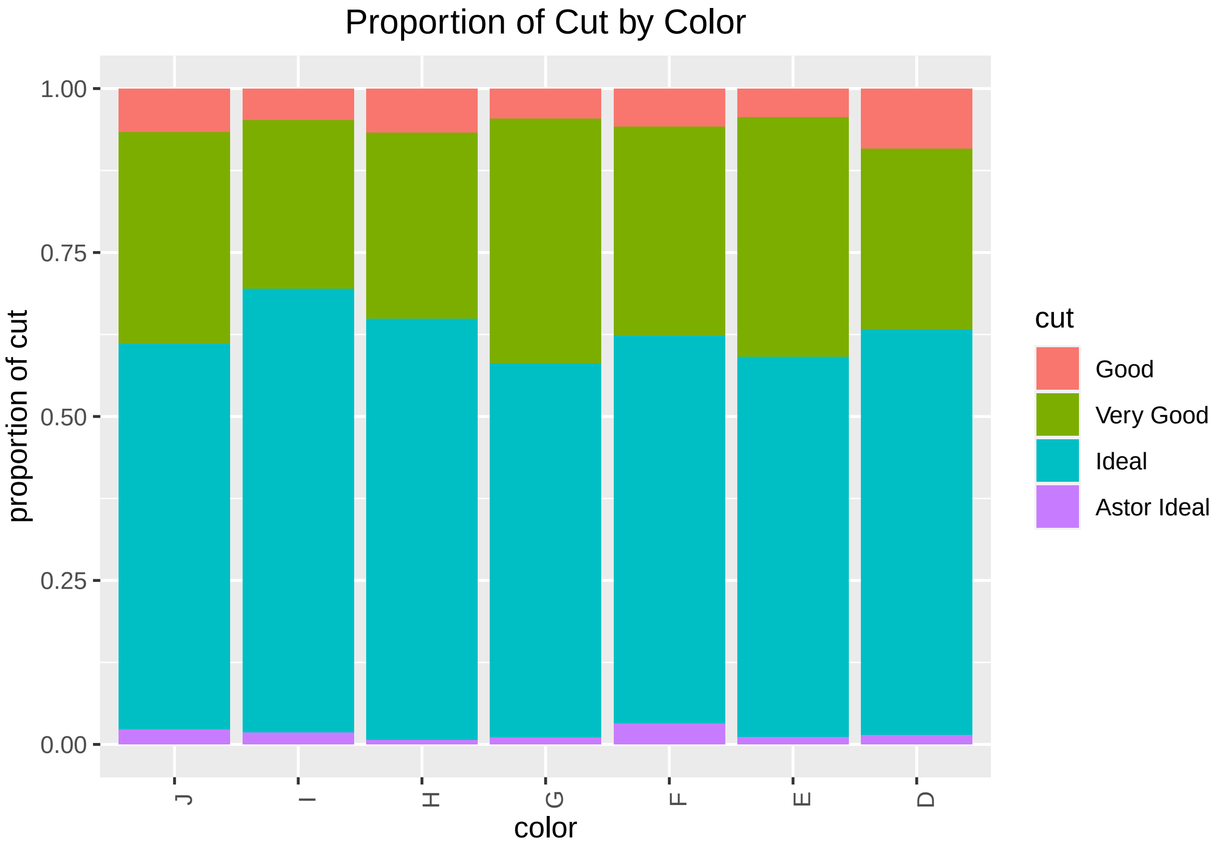


Considering the relationship of weight of a diamond in our data set and the clarity of the diamond, we construct a boxplot of weight versus color. Including outliers, a diamond with clarity identifier $VS2$ has the highest weight of over 7 $carat$. Excluding outliers, a diamond with clarity identifier FL has the highest weight over about 5.3 $carat$. A diamond with clarity identifier $VS1$ has the lowest weight of 0.23 $carat$. Diamonds with clarity identifier FL have the highest first-quartile / median / third-quartile weights.

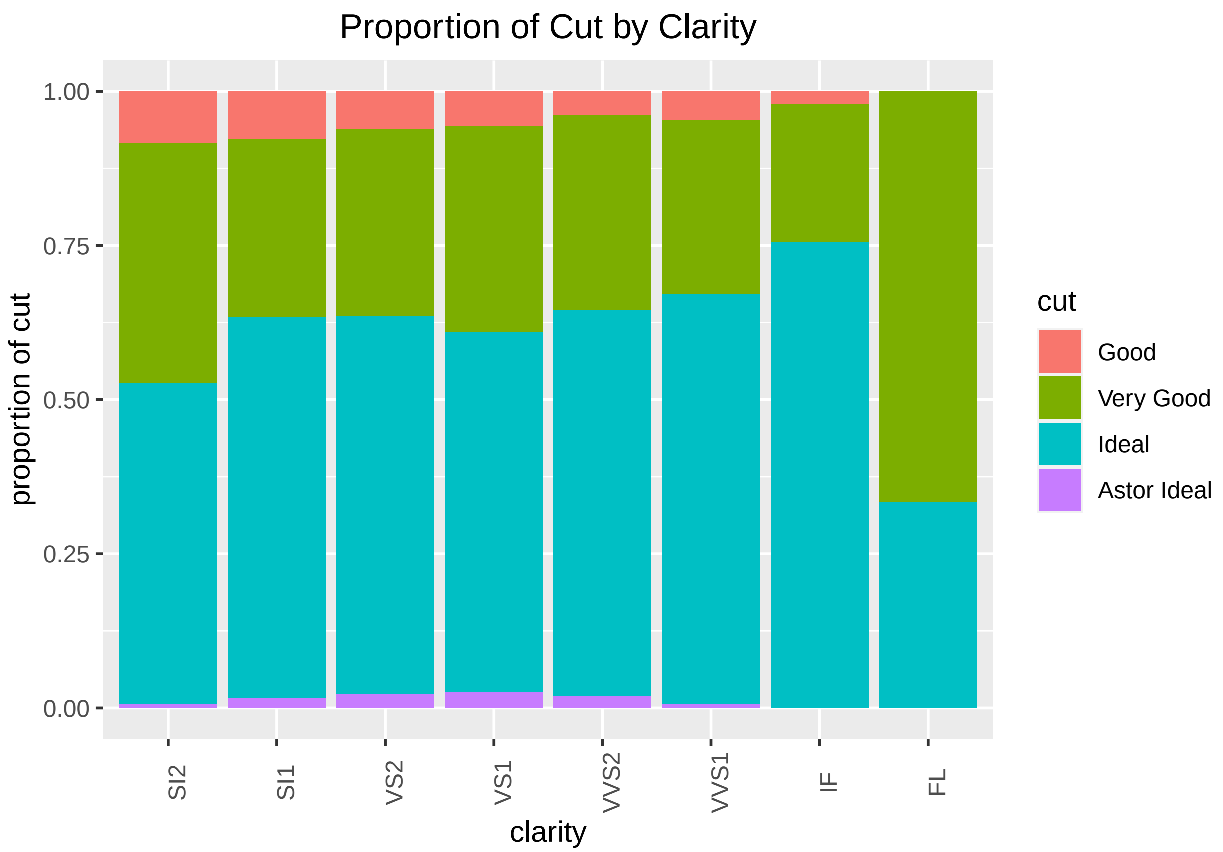


```
## [1] VS1
## Levels: SI2 SI1 VS2 VS1 VVS2 VVS1 IF FL
```

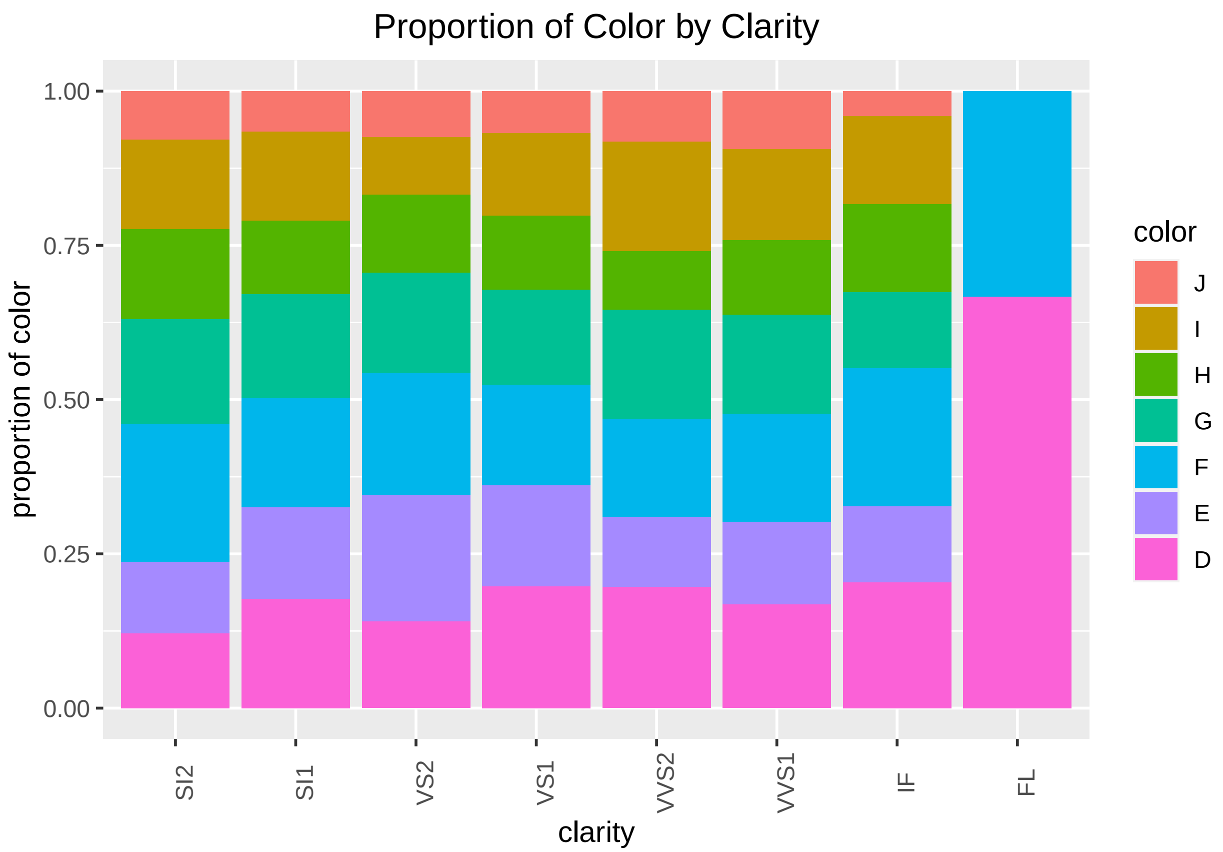
Considering the relationship of cut of a diamond in our data set and the color of the diamond, we construct a bar chart of proportion of cut by color. For each group of diamonds with a unique color identifier, most diamonds were ideal. The proportions of ideal, very good, good, and Astor ideal diamonds decreased in that order. The group of diamonds with color identifier F had the highest proportion of Astor ideal diamonds. The group of diamonds with color identifier I had the highest proportion of ideal diamonds. The group of diamonds with color identifier G had the highest proportion of very good diamonds. The group of diamonds with color identifier D had the highest proportion of good diamonds.



Considering the relationship of cut of a diamond in our data set and the clarity of the diamond, we construct a bar chart of proportion of cut by clarity. For a group of diamonds with a clarity identifier other than FL, most diamonds are ideal. For a transition from a group of diamonds with a clarity identifier in the set of identifiers $\{VS1, VVS2, VVS1, IF\}$ to a group of diamonds with a clarity identifier in that set closer to IF, the proportion of ideal diamonds increases and the proportion of very good diamonds decreases, and the proportion of Astor ideal diamonds decreases. For a transition from a group of diamonds with a clarity identifier in the set of identifiers $\{SI2, SI1, VS2, VS1, VVS1\}$ to a group of diamonds with a clarity identifier in that set closer to $VVS1$, the proportion of good diamonds decreases. The majority of diamonds with clarity identifier FL are very good, while all other diamonds with clarity identifier FL are ideal.

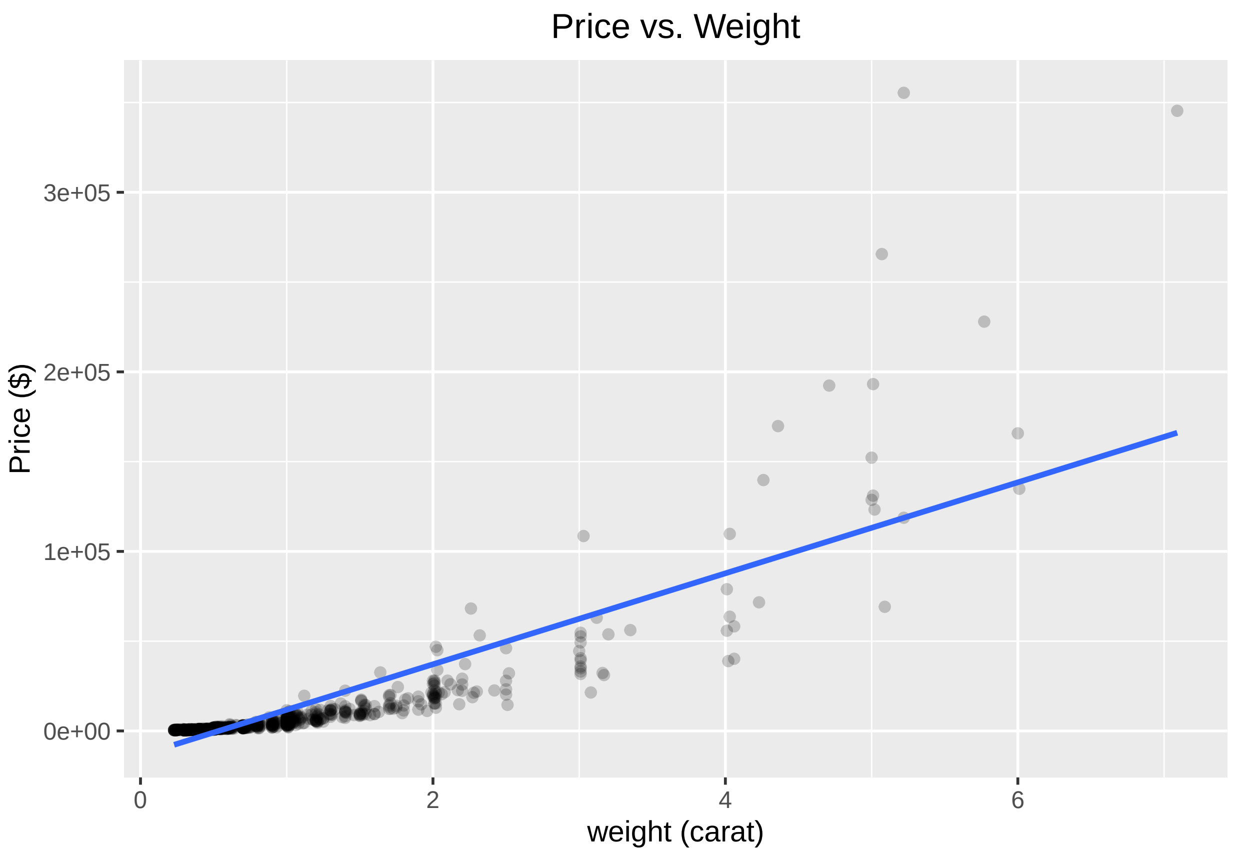


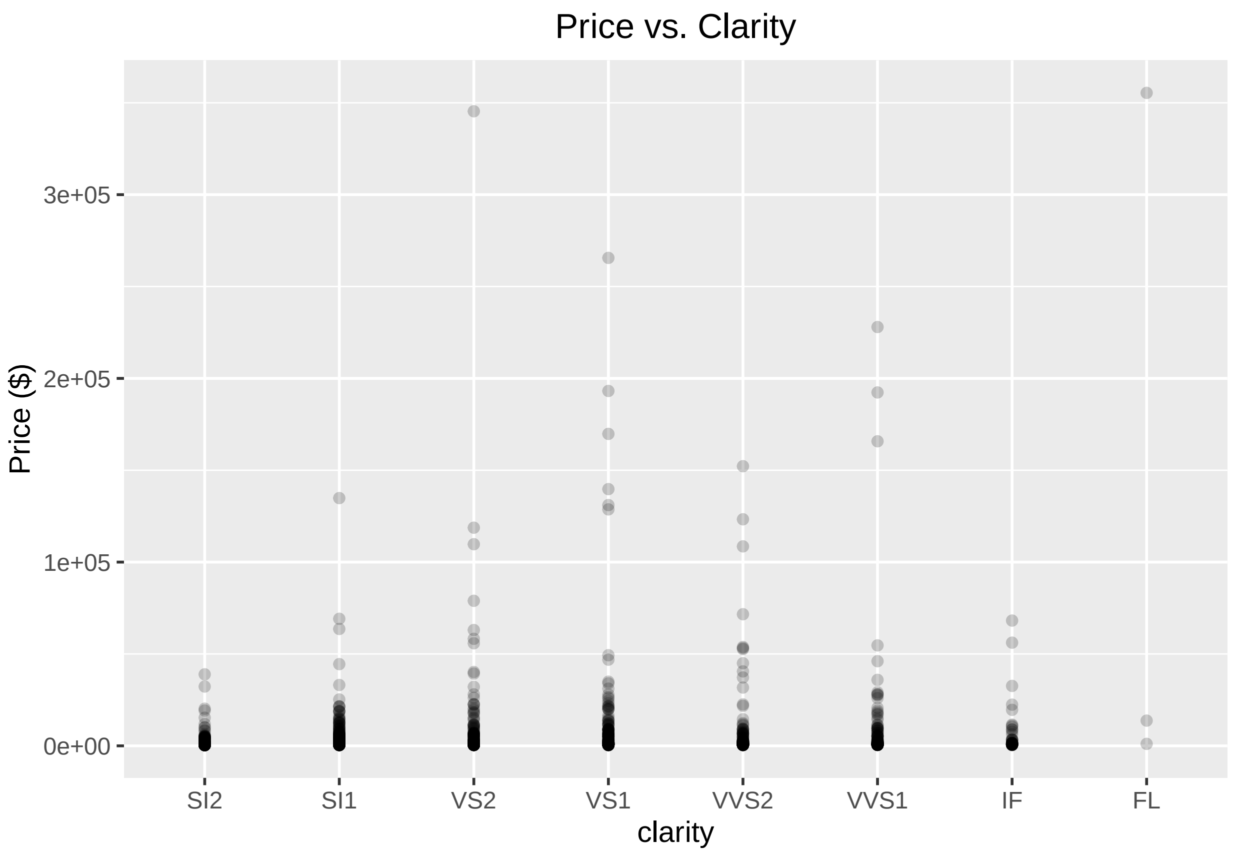
Considering the relationship of color of a diamond in our data set and the clarity of the diamond, we construct a bar chart of proportion of color by clarity. The proportion of diamonds with color identifier D for the group of diamonds with clarity identifier FL is greater than 0.5, is the only proportion greater than 0.5, and is significantly greater than the proportion of diamonds with color identifier D for any other group of diamonds by clarity identifier. For the group of diamonds with clarity identifier FL, all diamonds other than diamonds with color identifier D have color identifier F. For a transition from a group of diamonds with a clarity identifier in the set of identifiers $\{VVS2, VVS1, IF, FL\}$ to a group of diamonds with a clarity identifier in that set closer to FL, the proportion of diamonds with color identifier F increases.

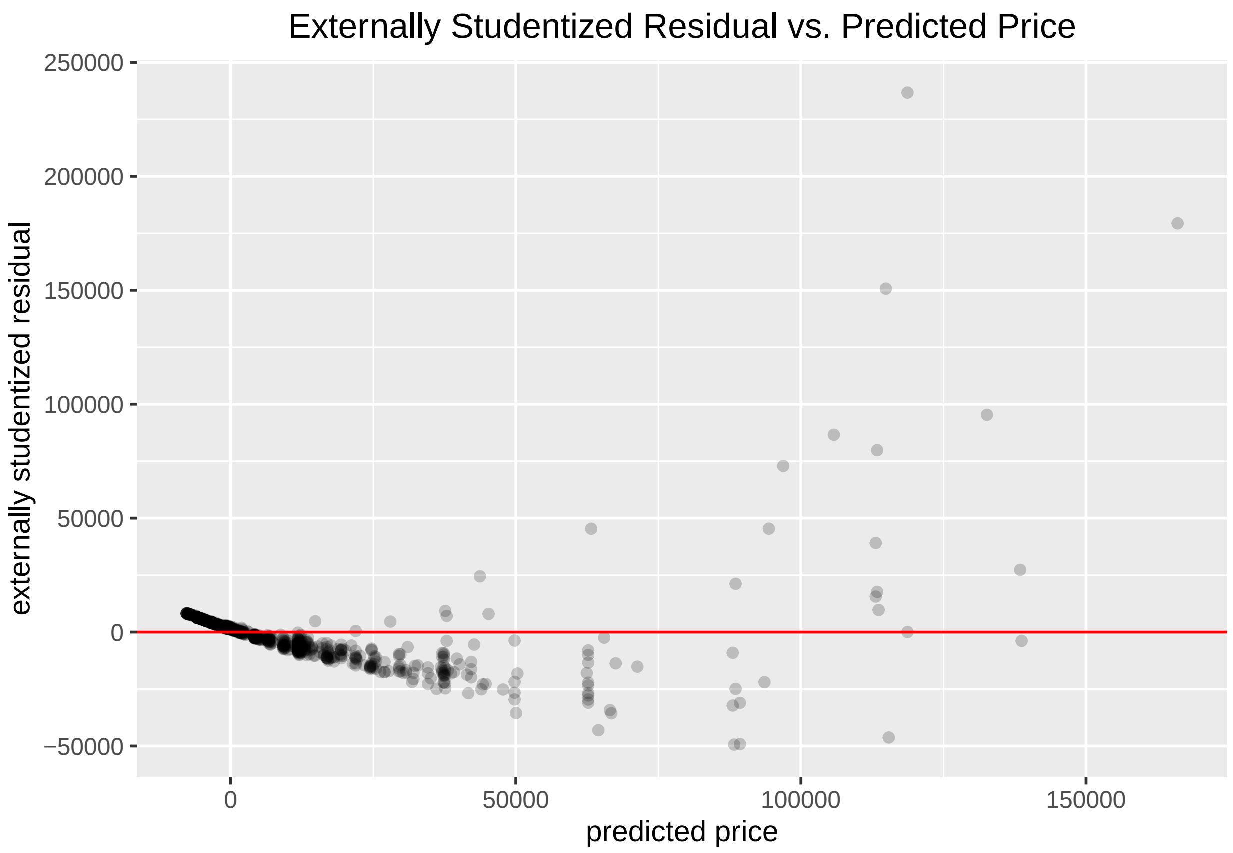


Considering the relationship of price of a diamond in our data set and the weight of the diamond, we construct a scatterplot for a linear model of price versus weight, as well as a plot of externally studentized residuals versus predicted prices for the model, a plot of AutoCorrelation Function values versus lag for the model, and a plot of sample quantiles versus theoretical quantiles for the residuals of the model.

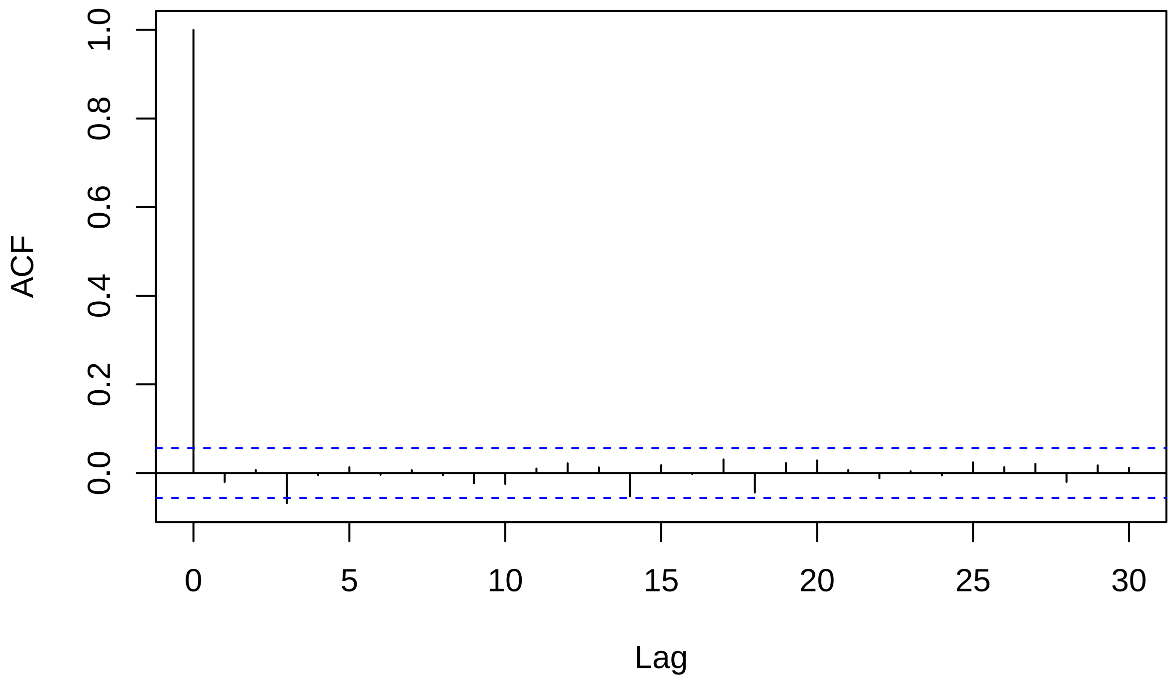
As weight of a diamond increases, price of the diamond increases, at an increasing rate. As weight of a diamond increases, the variance of residuals of a linear model of price versus weight increases. There is a significant correlation between $(weight, \ price)$ observations and observations three away. Given a moderate downward and a sharp upward curve at extremes of a plot of sample quantiles versus theoretical quantiles for the residuals of the linear model of price versus weight, the tails of a probability vs. externally studentized residuals plot / distribution are too light for this distribution to be considered normal.

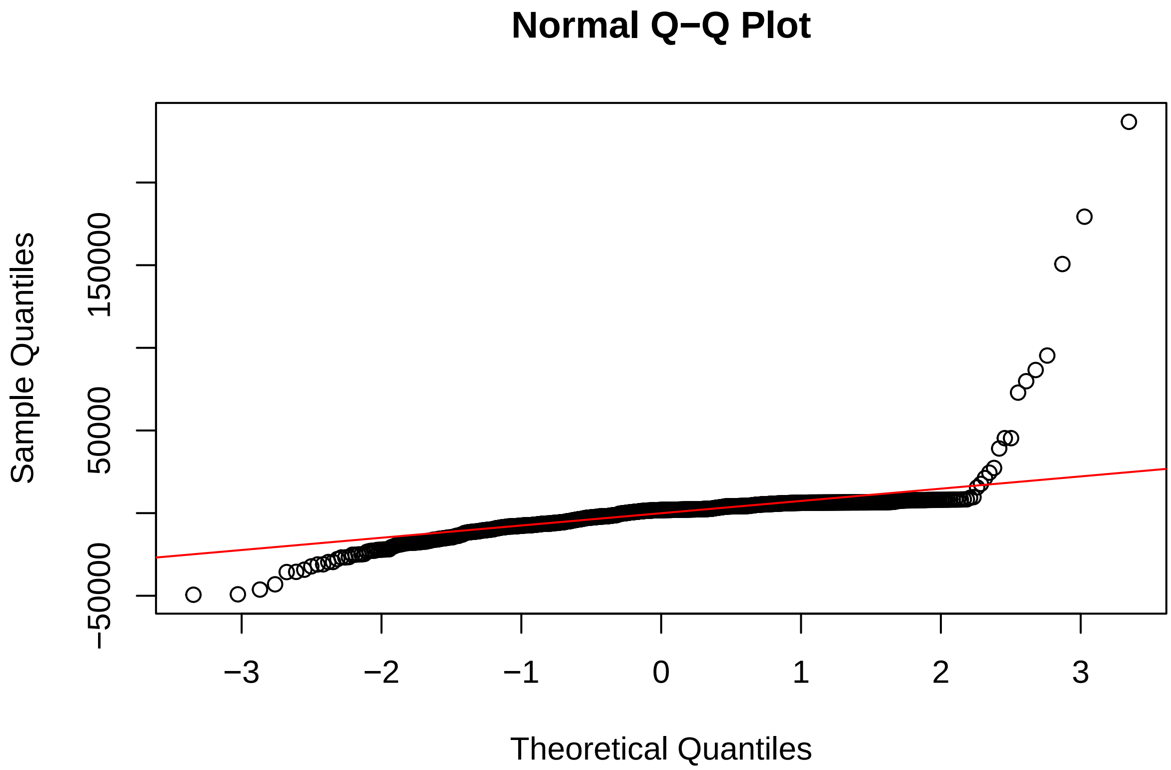

Price vs. Weight
Price ($)
3e+05
2e+05
1e+05
0e+00
0
2
4
6
weight (carat)


Price vs. Clarity
Price ($)
3e+05
2e+05
1e+05
0e+00
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
FL
clarity


Externally Studentized Residual vs. Predicted Price
250000
200000
150000
100000
50000
0
−50000
externally studentized residual
0
50000
100000
150000
predicted price

ACF Value vs. Lag for Transformed Linear Model





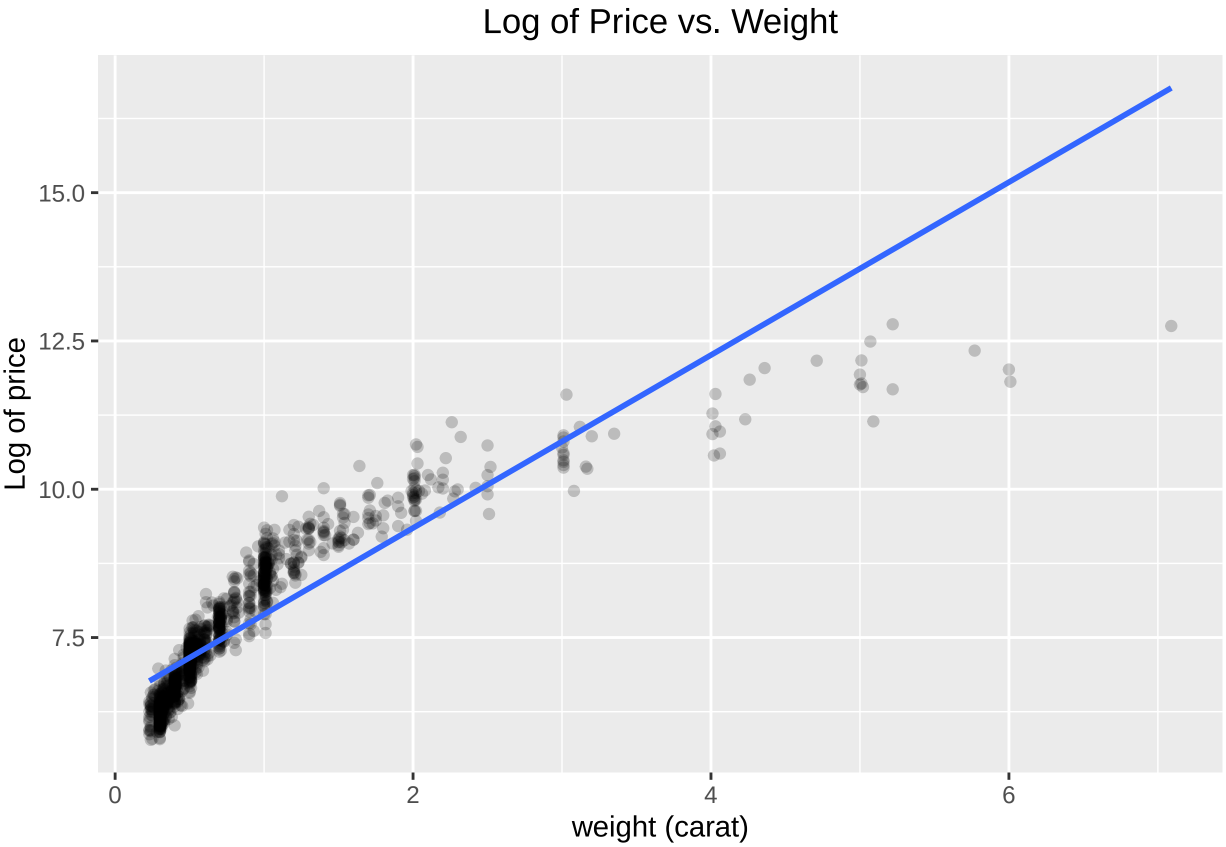
Simple Linear Regression of Price versus Weight

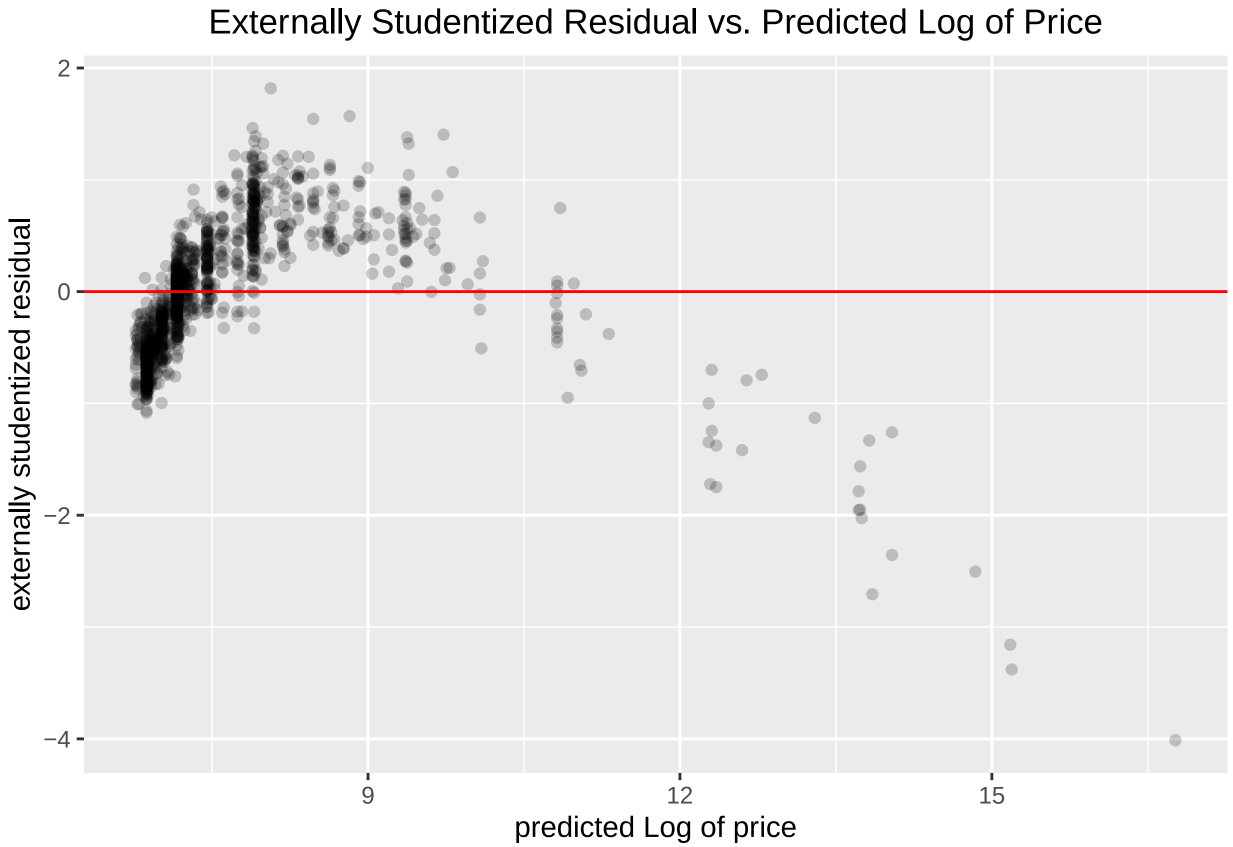
Simple linear regression depends of the following five assumptions being met.

1. The relationship between response / price and predictor / regressor / weight is linear, at least approximately. This assumption is not met. The relationship appears to be nonlinear.
2. The residuals of the linear model of price versus weight have mean 0. This assumption is not met. observations are not scattered evenly around the fitted line. Residuals are not evenly scattered around $e = 0$.
3. The distributions of residuals of the linear model of price versus weight for different weights have constant variance. This assumption is not met. The vertical variation of observations is not constant. Residuals are not evenly scattered around $e = 0$.
4. The residuals of the linear model of price versus weight are uncorrelated. This assumption is not met. The ACF value for lag 0 is always 1; the correlation of the vector of residuals with itself is always 1. Since the ACF value for lag 3 is significant, we have sufficient evidence to reject a null hypothesis that the residuals of the linear model of price versus weight are uncorrelated. We have sufficient evidence to conclude that the residuals of the linear model of price versus weight are correlated. We have sufficient evidence to conclude that the assumption that the residuals are uncorrelated is not met.
5. The residuals of the linear model of price versus weight are normally distributed. This assumption is not met. A linear model is robust to these assumptions. Given a moderate downward and a sharp upward curve at extremes of a plot of sample quantiles versus theoretical quantiles for the residuals of the linear model of price versus weight, the tails of a probability vs. externally studentized residuals plot / distribution are too light.

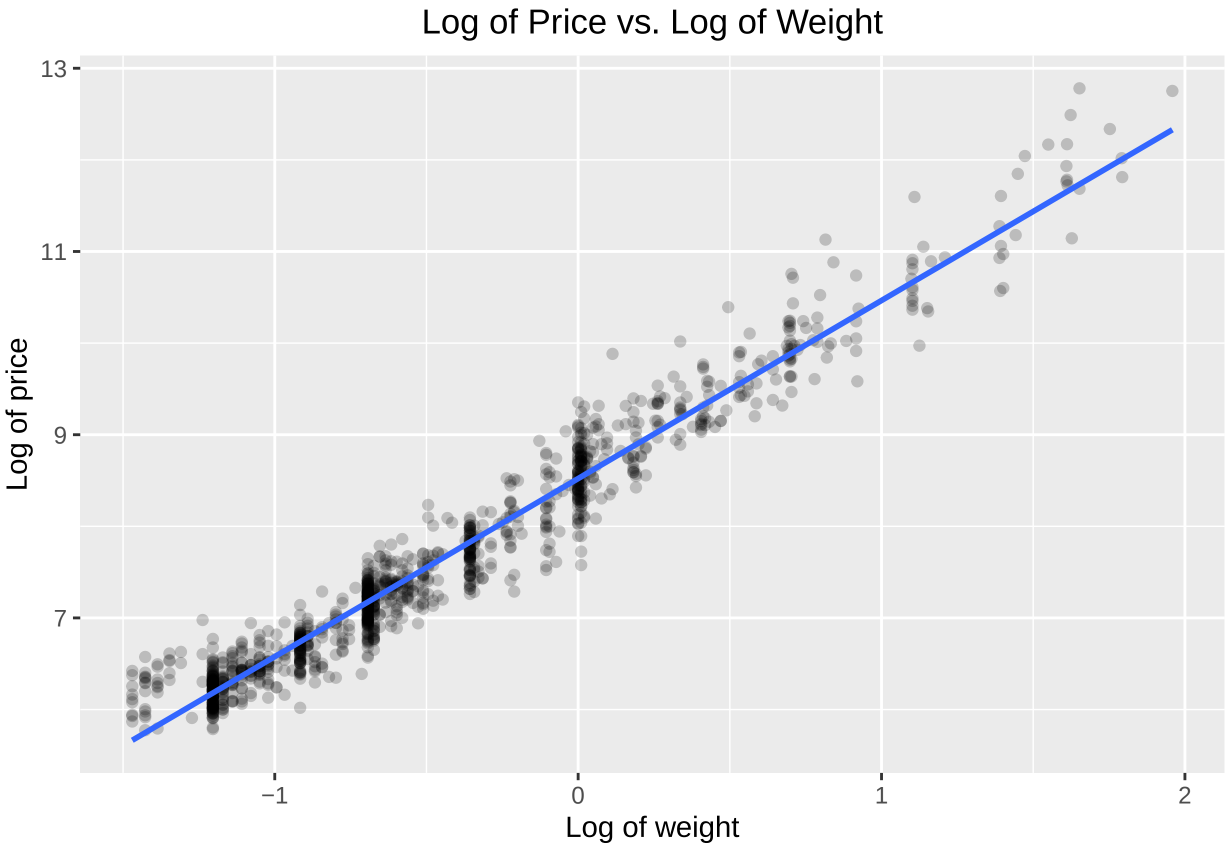
```
#{r, eval=TRUE, echo=FALSE, results="show", message=FALSE, warning = FALSE} #library(TomLeversRPackage)
#linear_model <- lm(price ~ weight, data = data_set) #
```

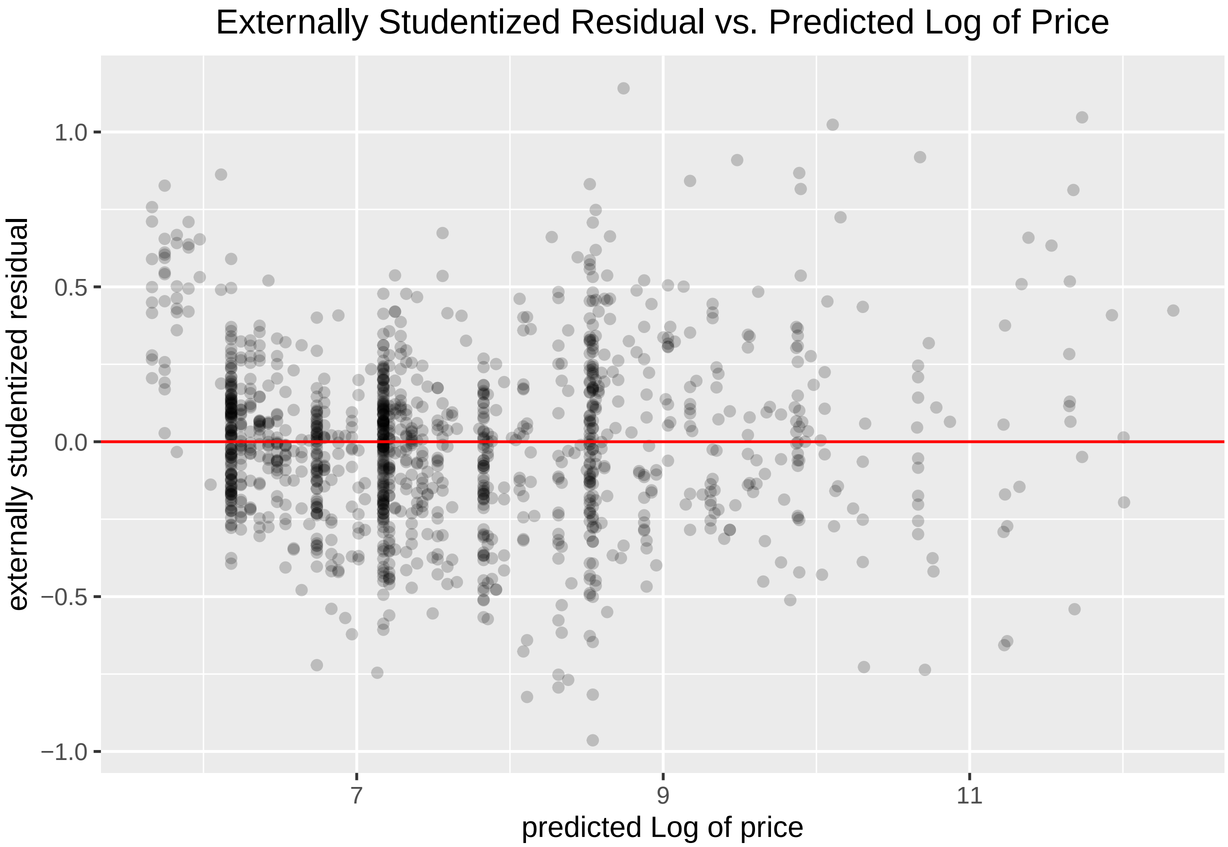
Given that the above assumptions for simple linear regression are not met, we generate a data set of transformed price and/or transformed weight such that all of the assumptions are met. The Box-Cox Method is presented in section 5.4.1 of *Introduction to Linear Regression Analysis* (Sixth Edition) by Douglas C. Montgomery et al.. To generate a data set of transformed price and/or transformed weight for which the assumption that the relationship between transformed price and/or transformed weight is linear is met, and the assumption that the distributions of residuals of a linear model of transformed price and/or transformed weight for different weights or transformed weights have constant variance, using R, we perform the Box-Cox Method to determine a maximum-likelihood estimate of a parameter $\lambda = 0.311$ to be used in a power transformation $y' = y^\lambda$ of a price y. The maximum likelihood estimate of λ is close to a whole parameter $\lambda = 0$. Given a power-transformation parameter $\lambda = 0$, We transform price values according to $y' = ln(y)$. We present a scatterplot of Log of price versus weight. We construct a linear model of Log of price versus weight and present a scatterplot of externally studentized residuals versus predicted Log of price for the linear model of Log of price versus weight.

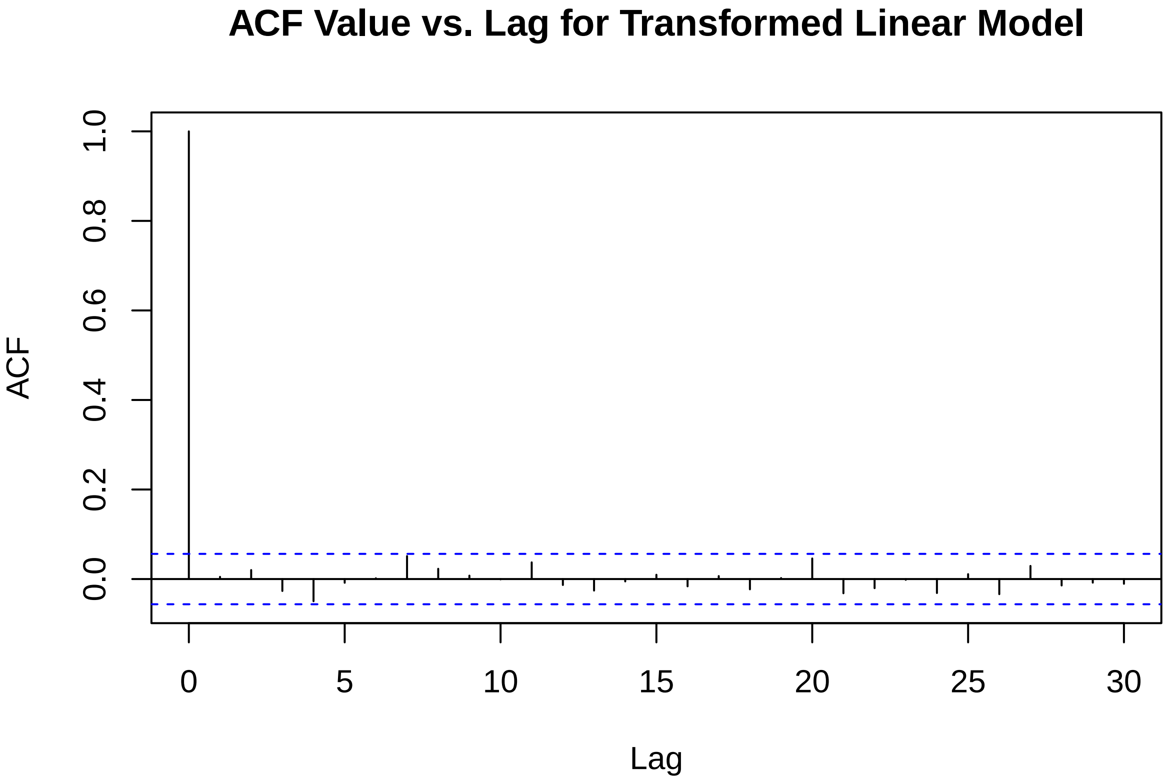


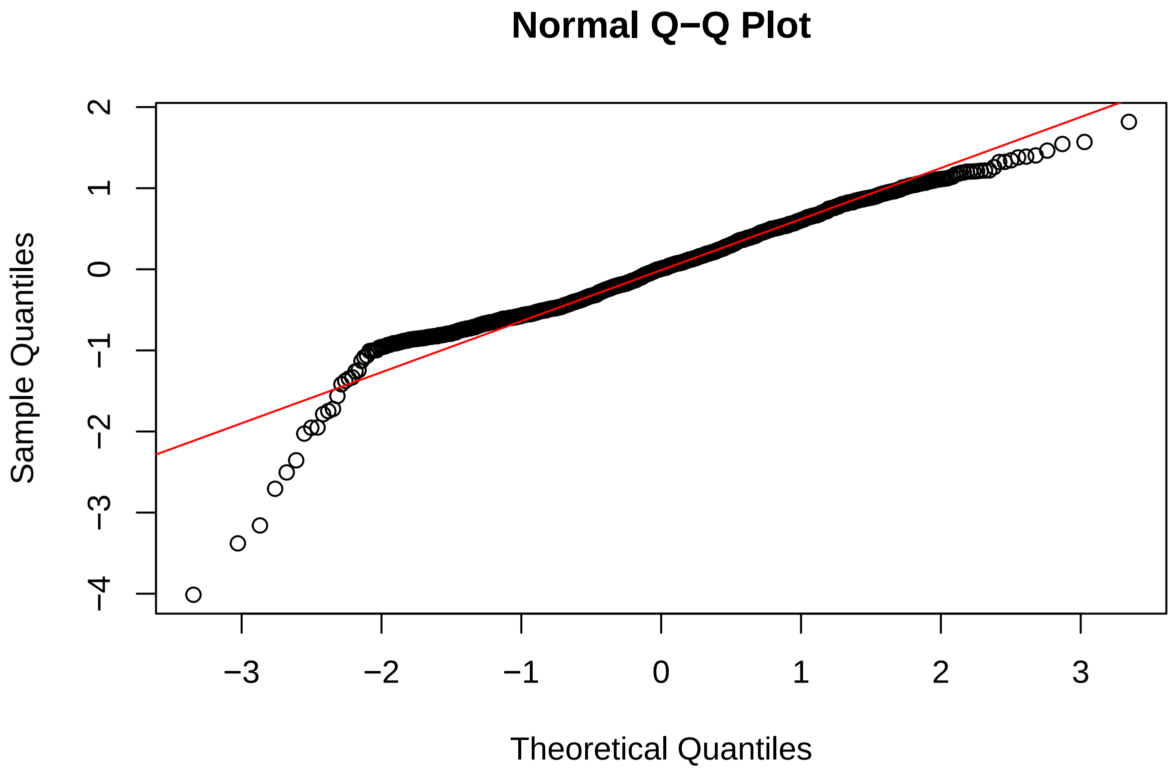


We have generated a data set of Log of price versus weight for which the assumption that the distributions of residuals of a linear model of Log of price and/or weight for different weights have constant variance. To generate a data set of transformed price and/or transformed weight for which the assumption that the relationship between transformed price and/or transformed weight is linear is met, and the assumption that the mean of the residuals of the linear model of transformed price and/or transformed weight is 0, we take our Log of-price data and additionally logarithmicize weight. We present a scatterplot of Log of price versus Log of weight. We construct a linear model of Log of price versus Log of weight. We present a scatterplot of externally studentized residuals versus predicted Log of price for the linear model of Log of price versus Log of weight. Additionally, we construct a plot of AutoCorrelation Function values versus lag for the linear model, and a plot of sample quantiles versus theoretical quantiles for the residuals of the model.


Log of Price vs. Log of Weight
Log of price
13
11
9
7
Log of weight
−1
0
1
2


Externally Studentized Residual vs. Predicted Log of Price
externally studentized residual
1.0
0.5
0.0
−0.5
−1.0
7
9
11
predicted Log of price


ACF Value vs. Lag for Transformed Linear Model
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
Lag



1. The assumption that the relationship between response / Log of price and predictor / regressor / Log of weight is linear, at least approximately, is met. The relationship appears to be linear.
2. The assumption that the residuals of the linear model of Log of price versus Log of weight have mean 0 is met. observations are scattered evenly around the fitted line. Residuals are evenly scattered around $e = 0$.
3. The assumptions that the distributions of residuals of the linear model of Log of price versus logarithmized weight for different weights have constant variance is met. The vertical variation of observations is constant. Residuals are evenly scattered around $e = 0$.
4. The assumption that the residuals of the linear model of Log of price versus Log of weight are uncorrelated is met. The ACF value for lag 0 is always 1; the correlation of the vector of residuals with itself is always 1. Since all ACF value are insignificant, we have insufficient evidence to reject a null hypothesis that the residuals of the linear model of Log of price versus Log of weight are uncorrelated. We have insufficient evidence to conclude that the residuals of the linear model of Log of price versus Log of weight are correlated. We have insufficient evidence to conclude that the assumption that the residuals of the linear model of Log of price versus Log of weight are uncorrelated is not met.
5. The assumption that the residuals of the linear model of Log of price versus Log of weight are normally distributed is not met. However, a linear model is robust to these assumptions. Given a sharp downward curve at the bottom left of a plot of sample quantiles versus theoretical quantiles for the residuals of the linear model of Log of price versus Log of weight, a probability vs. externally studentized residuals plot / distribution is not normal.

We determine an estimated linear-regression equation

$$\hat{\beta}_0 = 8.521$$

$$\hat{\beta}_1 = 1.944$$

$$ln(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 ln(x)$$

where $\hat{y} = E(y|x)$ is the expected price given a weight x.

$$\hat{y} = exp[\hat{\beta}_0 + \hat{\beta}_1 ln(x)]$$

Consider weights x and $x_+ = (1+p)x$ and corresponding predicted prices $\hat{y}$ and $\hat{y}_+$.

$$ln(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 ln(x)$$

$$ln\left(\hat{y}_+\right) = \hat{\beta}_0 + \hat{\beta}_1 ln\left(x_+\right)$$

$$ln\left(\hat{y}_+\right) - ln\left(\hat{y}\right) = \left[\hat{\beta}_0 + \hat{\beta}_1 ln\left(x_+\right)\right] - \left[\hat{\beta}_0 + \hat{\beta}_1 ln\left(x\right)\right]$$

$$ln\left(\frac{\hat{y}_+}{\hat{y}}\right) = \hat{\beta}_1 ln\left(x_+\right) - \hat{\beta}_1 ln\left(x\right)$$

$$ln\left(\frac{\hat{y}_+}{\hat{y}}\right) = \hat{\beta}_1\left[ln\left(x_+\right) - ln\left(x\right)\right]$$

$$ln\left(\frac{\hat{y}_+}{\hat{y}}\right) = \hat{\beta}_1 ln\left(\frac{x_+}{x}\right)$$

$$ln\left(\frac{\hat{y}_+}{\hat{y}}\right) = \hat{\beta}_1 ln\left(\frac{(1+p)x}{x}\right)$$

$$ln\left(\frac{\hat{y}_+}{\hat{y}}\right) = \hat{\beta}_1 ln\left(1+p\right)$$

$$ln\left(\frac{\hat{y}_+}{\hat{y}}\right) = ln\left[(1+p)^{\hat{\beta}_1}\right]$$

$$\frac{\hat{y}_+}{\hat{y}} = (1+p)^{\hat{\beta}_1}$$

For an increase in weight by proportion p, the predicted price increases by a factor of $(1+p)^{\hat{\beta}_1}$. For example, for an increase in weight by proportion 0.1, or 10 percent, the predicted price increases by a factor of $(1 + 0.1)^{1.944} = 1.2$. For an increase in weight from 1 *carat* to 1.1 *carat*, price increases from $exp\left[8.521 + 1.944\ ln(1)\right]\ dollars = \$5,019.07$ by a factor of 1.2 to $exp\left[8.521 + 1.944\ ln(1.1)\right]\ dollars = \$6,040.75$. These prices have a ratio of 1.2.