Stat 6021: Project 1

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

10/01/22

Executive Summary

Present high-level results of analysis.

Present key findings.

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.

Presentation of Motivations, Visualizations, and Analysis

Present univariate visualizations (e.g., boxplots for categorical variables and boxplots and histograms for quantitative variables)

Present multivariate visualizations.

Address the various claims on at https://www.bluenile.com/education/diamonds.

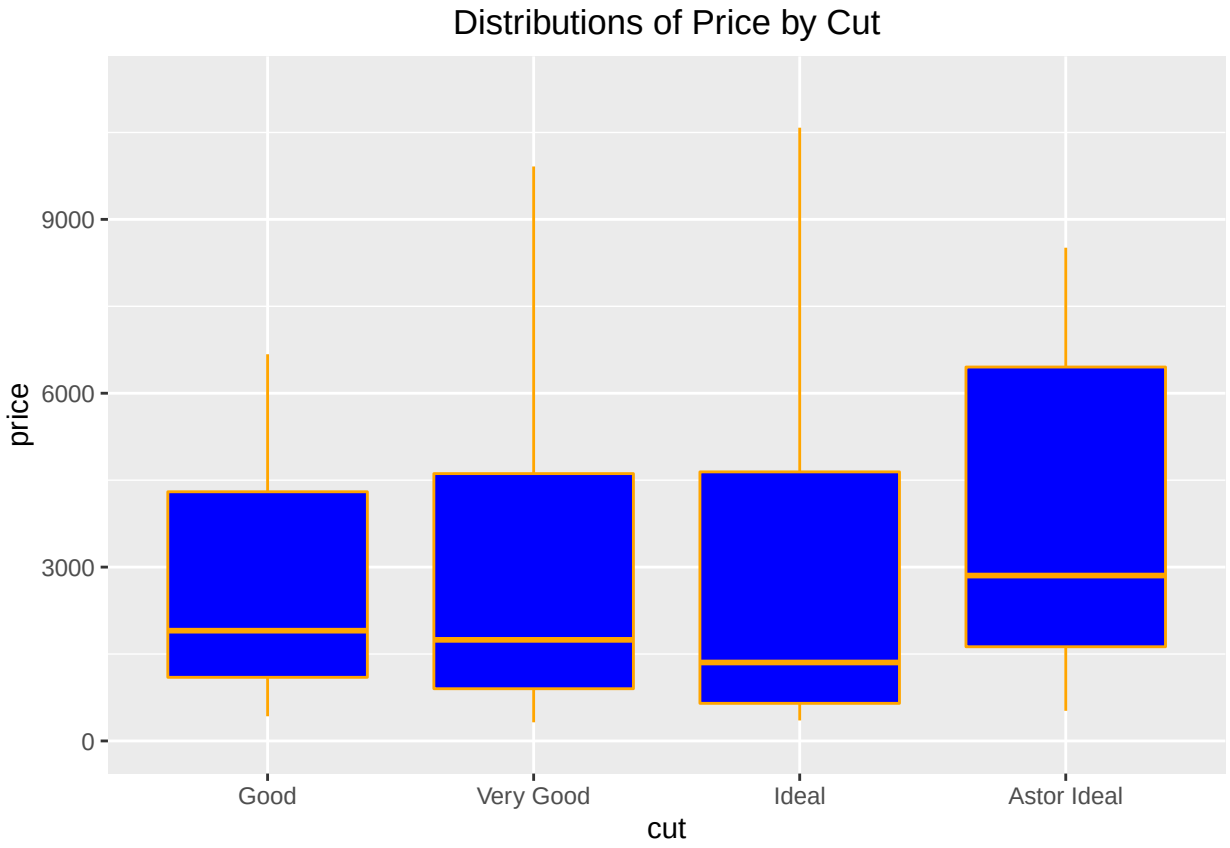
1. "The higher quality a diamond is, the more expensive it will be."
    a. The higher the weight of a diamond, the higher the price of the diamond.
    b. The closer the clarity identifier of a diamond is to "FL", the higher the price.
    c. The closer the color identifier of a diamond is to "D", the higher the price.
    d. The more ideal a diamond is, the higher the price.
2. How ideal a diamond is most significant in determining price.
3. How colorlessness a diamond is second most significant in determining price.

Determine other claims at https://www.bluenile.com/education/diamonds.

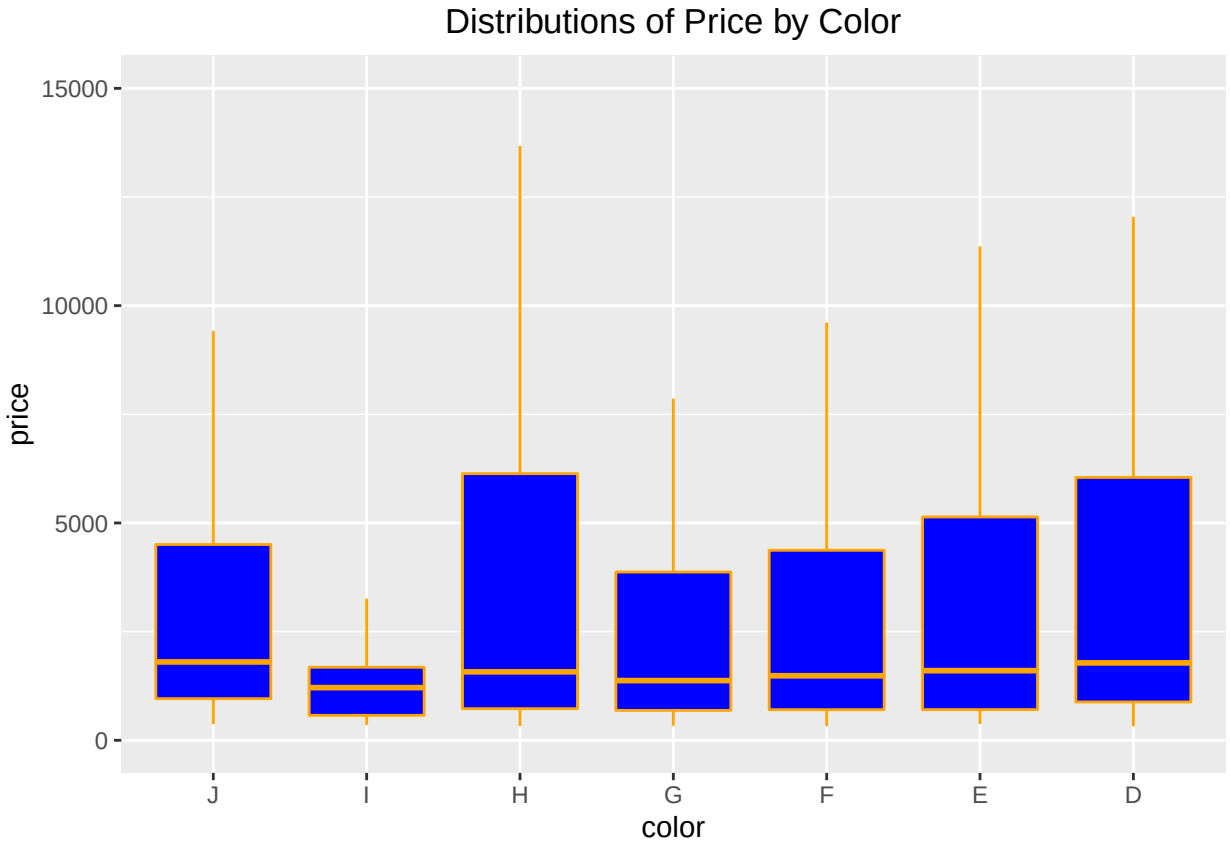
Add reasons why each visualization is presented.

Resize visualizations to condense report.

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.

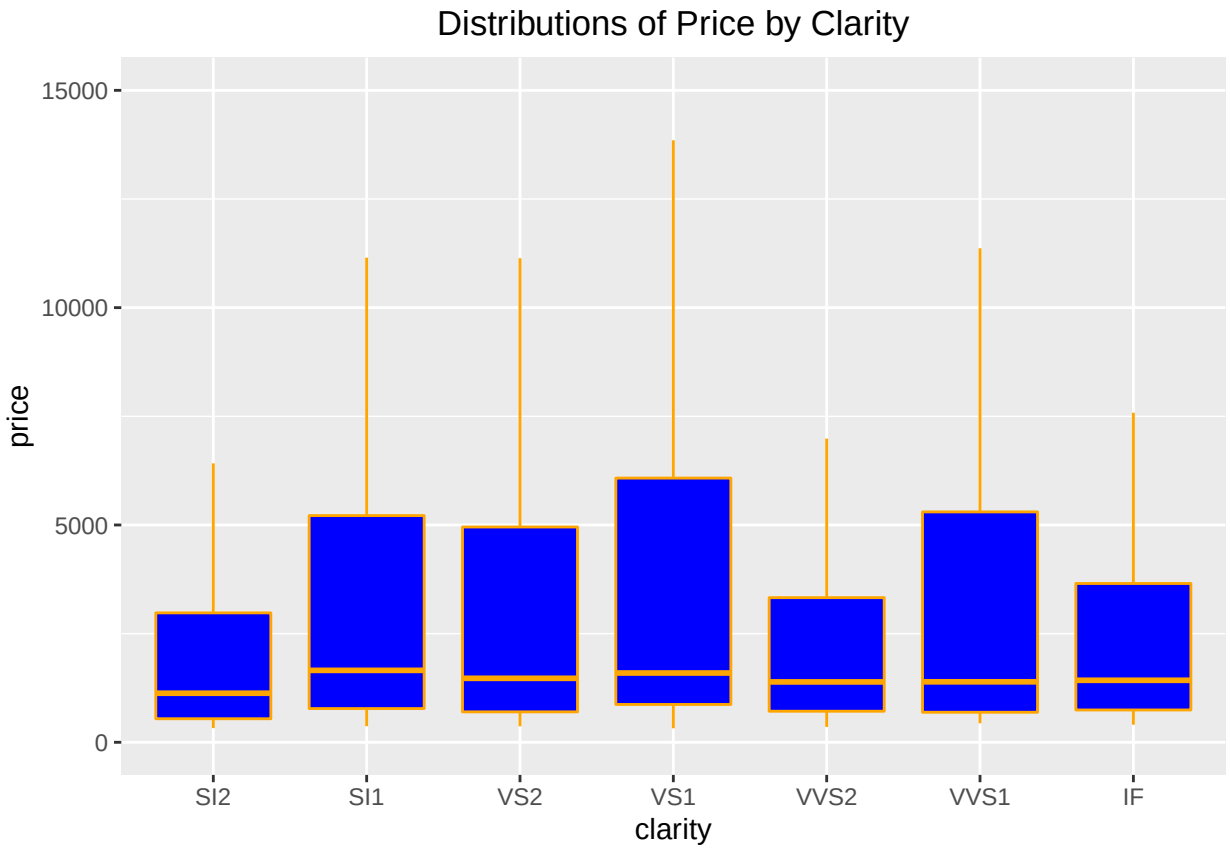


Considering the relationship of price of a diamond in our data set and the color of the diamond, we construct boxplots of price versus color with and without outliers and present the boxplot without outliers. Including outliers, for a transition from a group of diamonds with a color identifier in the 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.

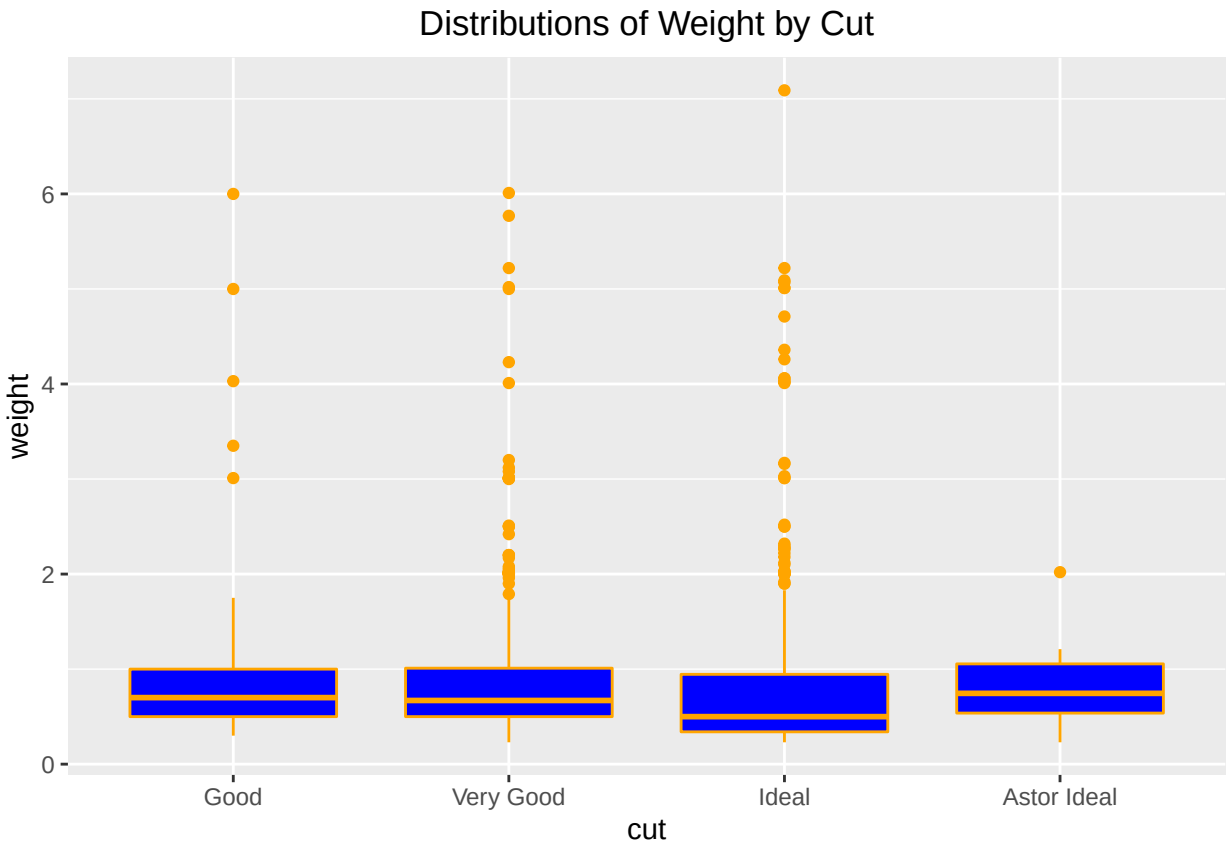


Considering the relationship of price of a diamond in our data set and the clarity of the diamond, we construct boxplots of price versus clarity with outliers and group of diamonds with clarity identifier FL, and without outliers and without the group of diamonds with clarity identifier FL. Including 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.

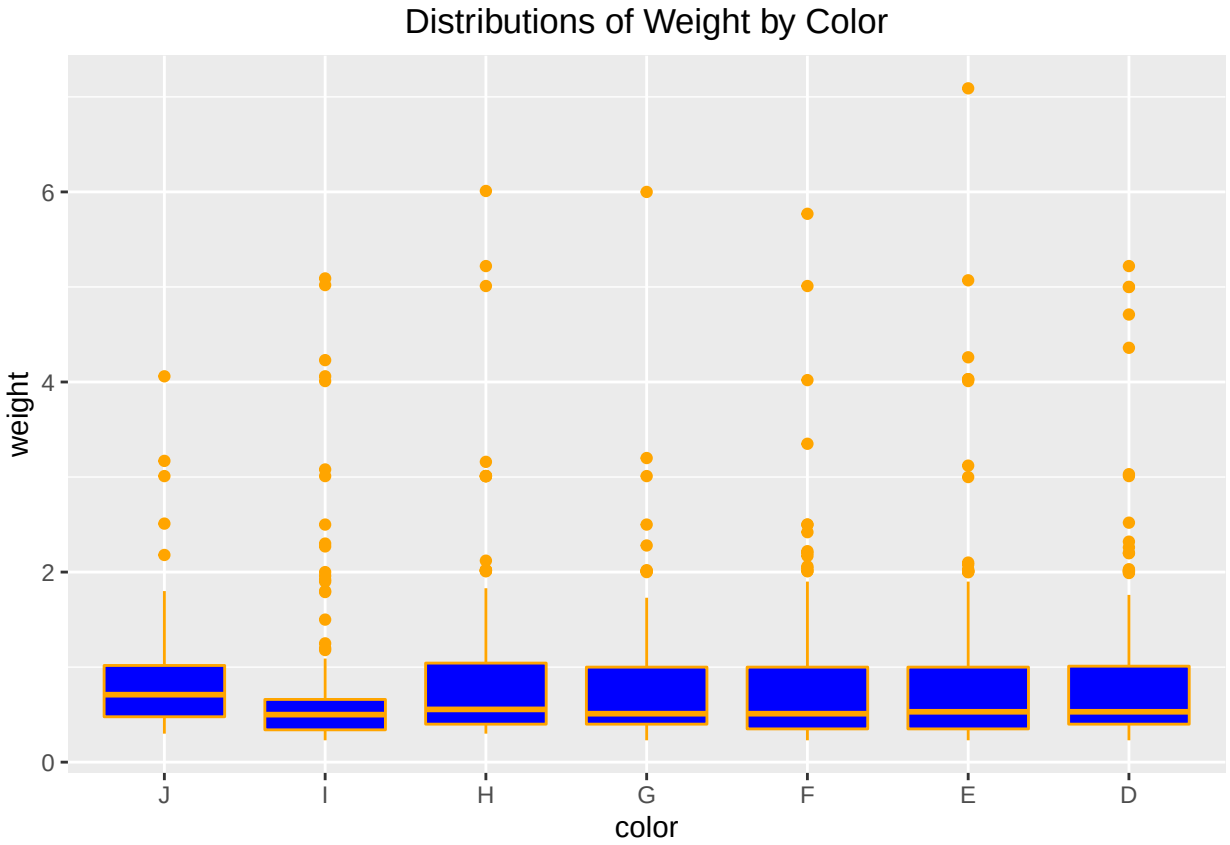

Distributions of Price by Clarity
price
3e+05
2e+05
1e+05
0e+00
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
FL
clarity



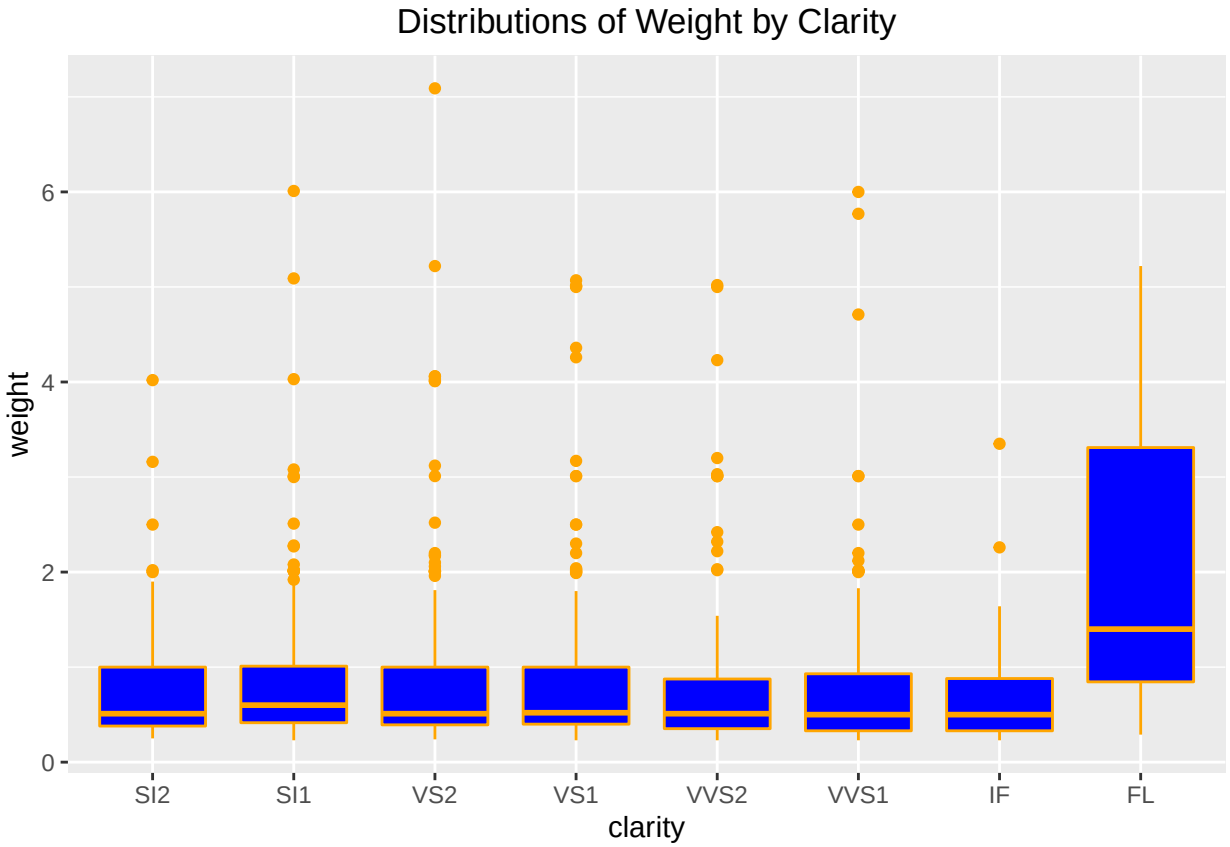
Considering the relationship of weight of a diamond in our data set and the cut of the diamond, we construct a boxplot of weight versus cut. Including outliers, an ideal diamond has the highest weight of over 5 *carat*. Excluding outliers, an ideal diamond the highest weight of about 2 *carat*. An Astor 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 $carat$.

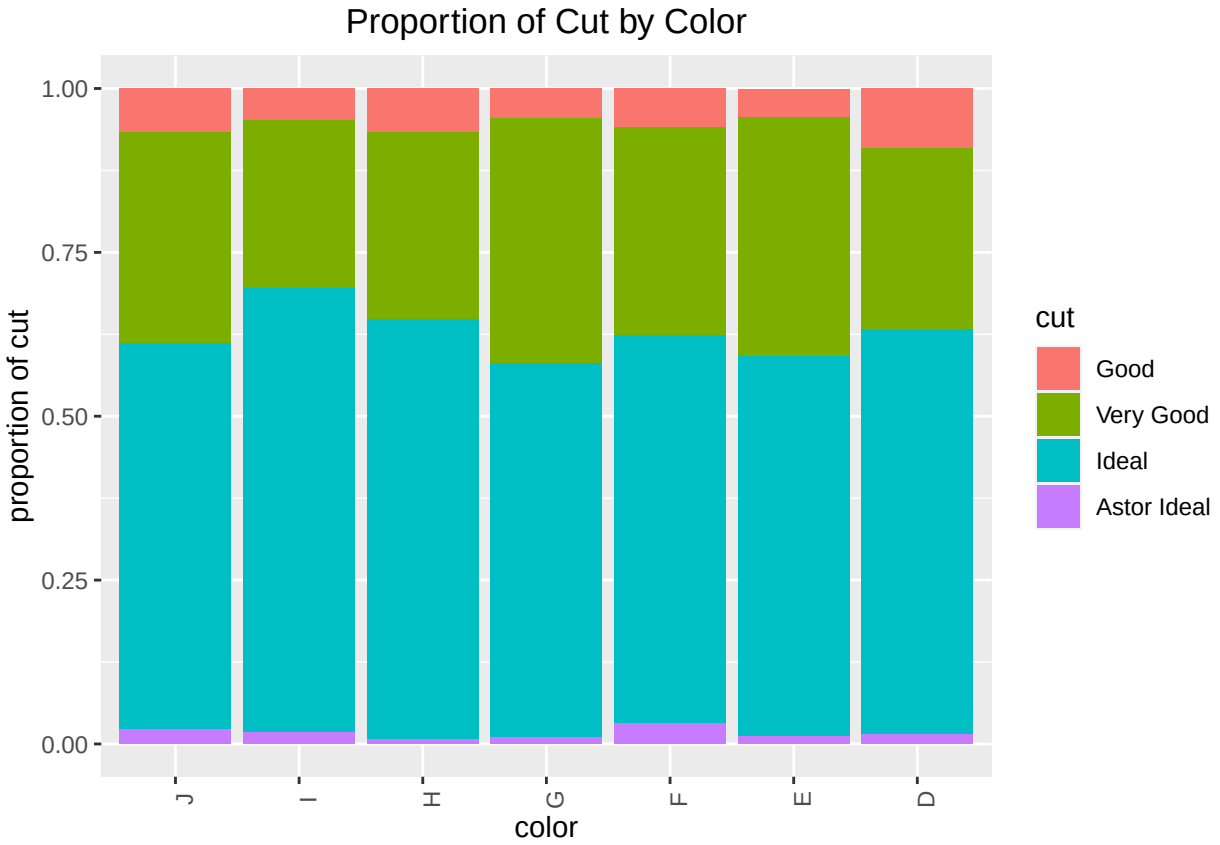


Considering the relationship of weight of a diamond in our data set and the clarity of the diamond, we construct a boxplot of weight versus color. Including outliers, a diamond with clarity identifier $VS2$ has the highest weight of over 7 *carat*. Excluding outliers, a diamond with clarity identifier 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.

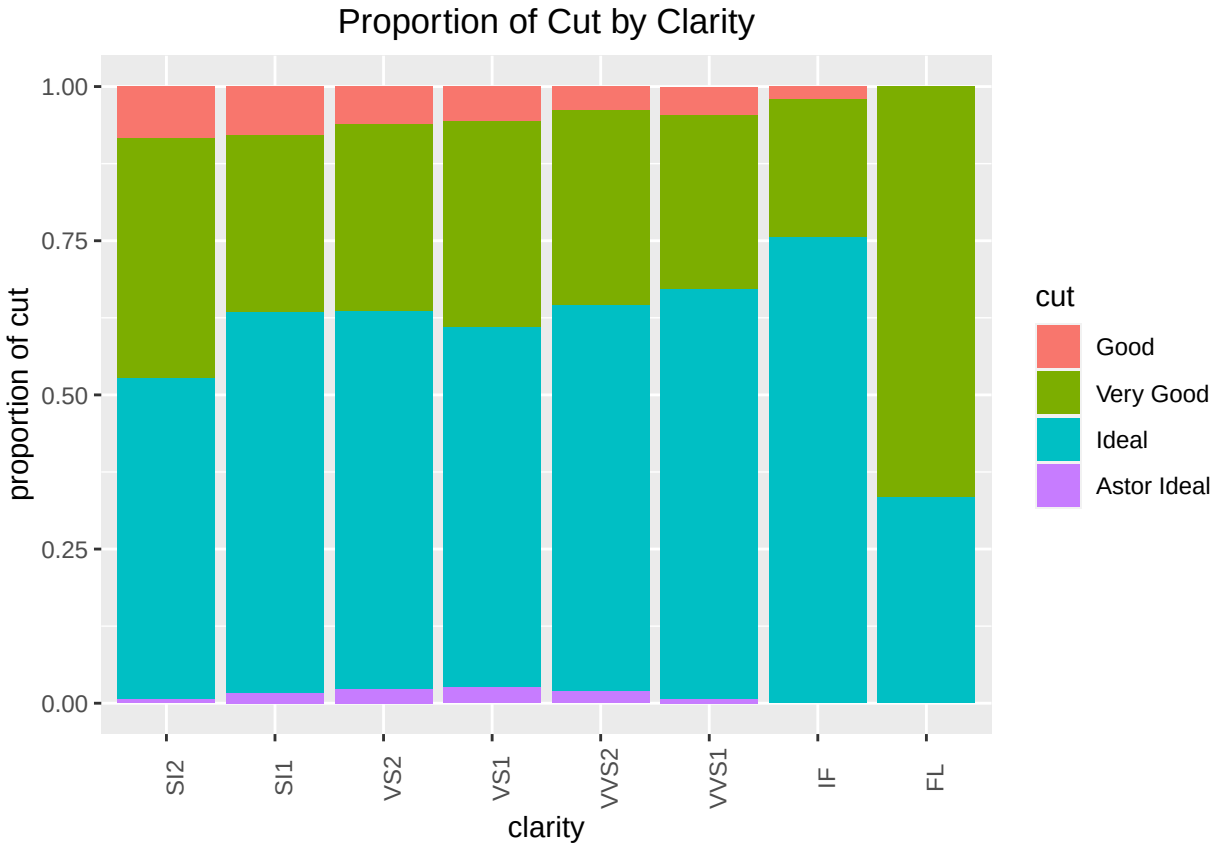



```
## [1] VS1
## Levels: SI2 SI1 VS2 VS1 VVS2 VVS1 IF FL
```

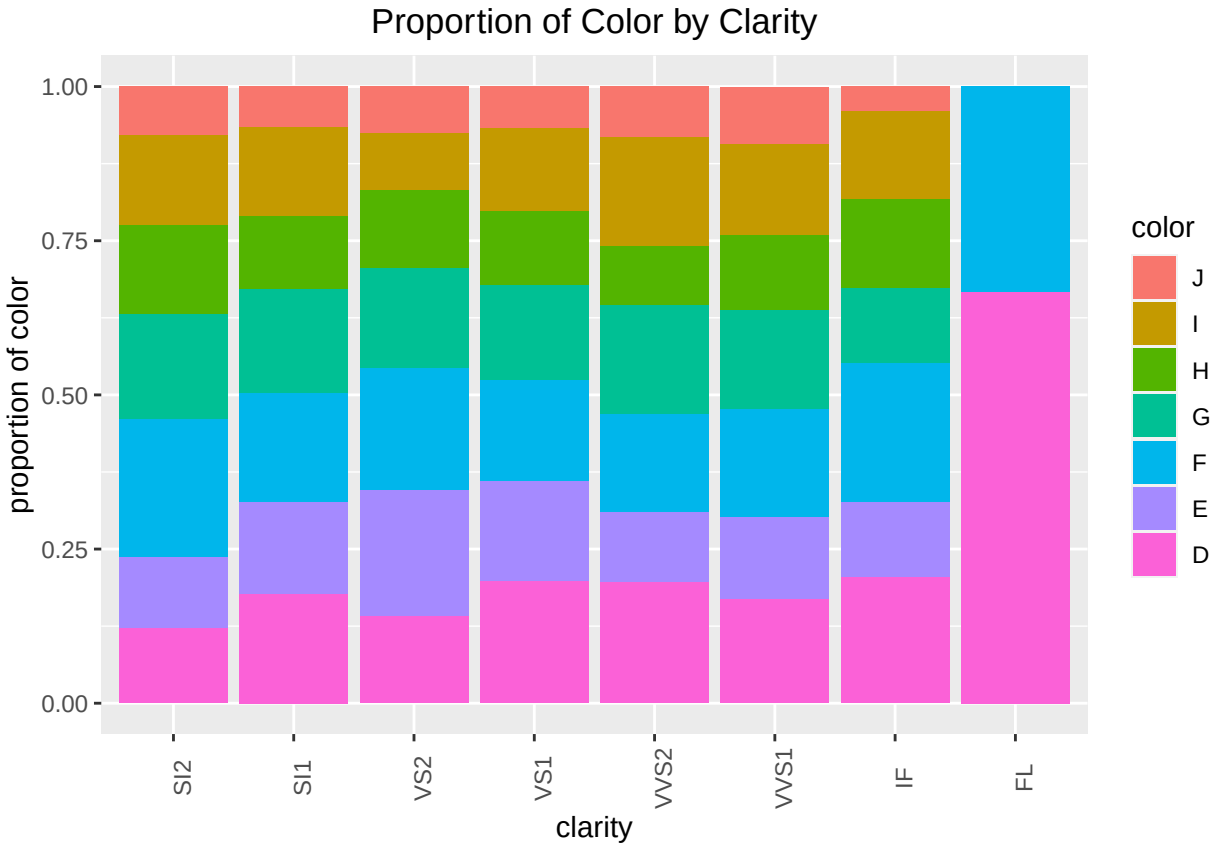
Considering the relationship of cut of a diamond in our data set and the color of the diamond, we construct a bar chart of proportion of cut by color. For each group of diamonds with a unique color identifier, most diamonds were ideal. The proportions of ideal, very good, good, and Astor ideal diamonds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

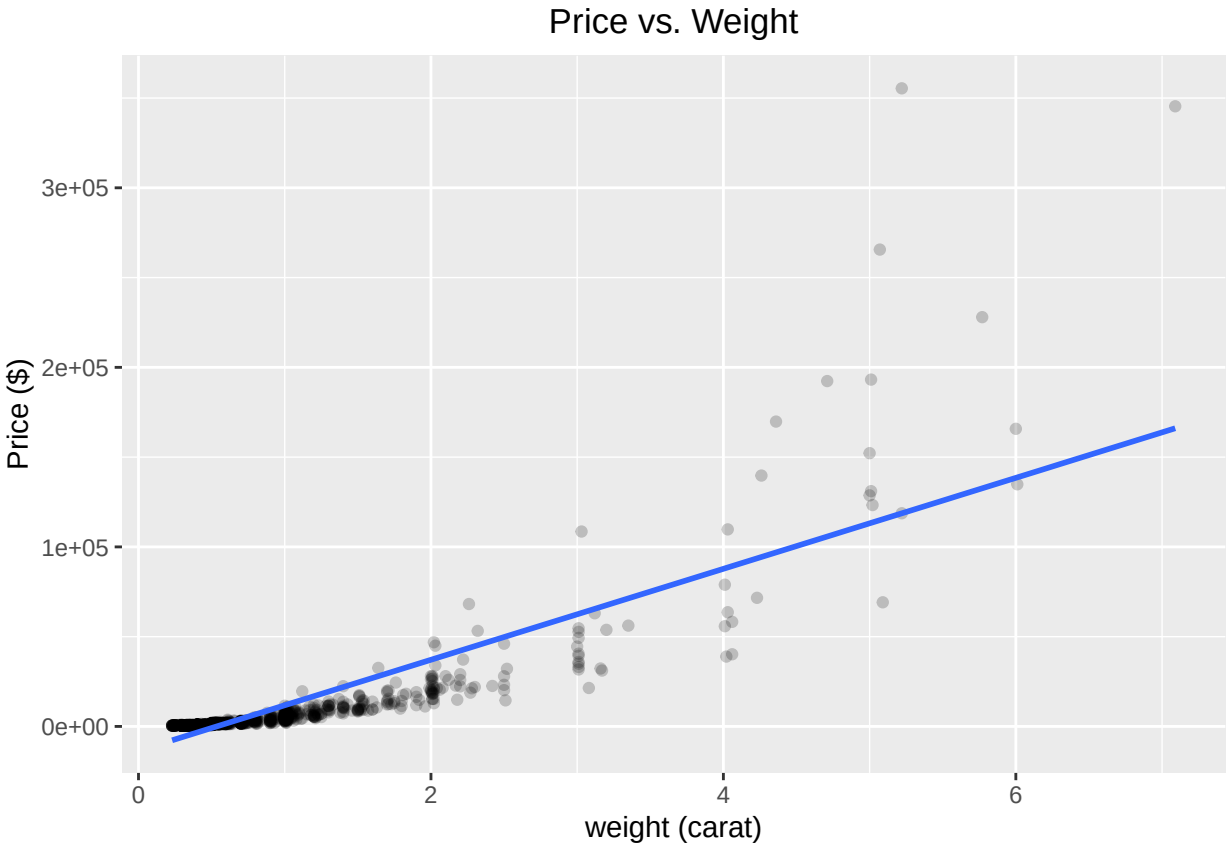


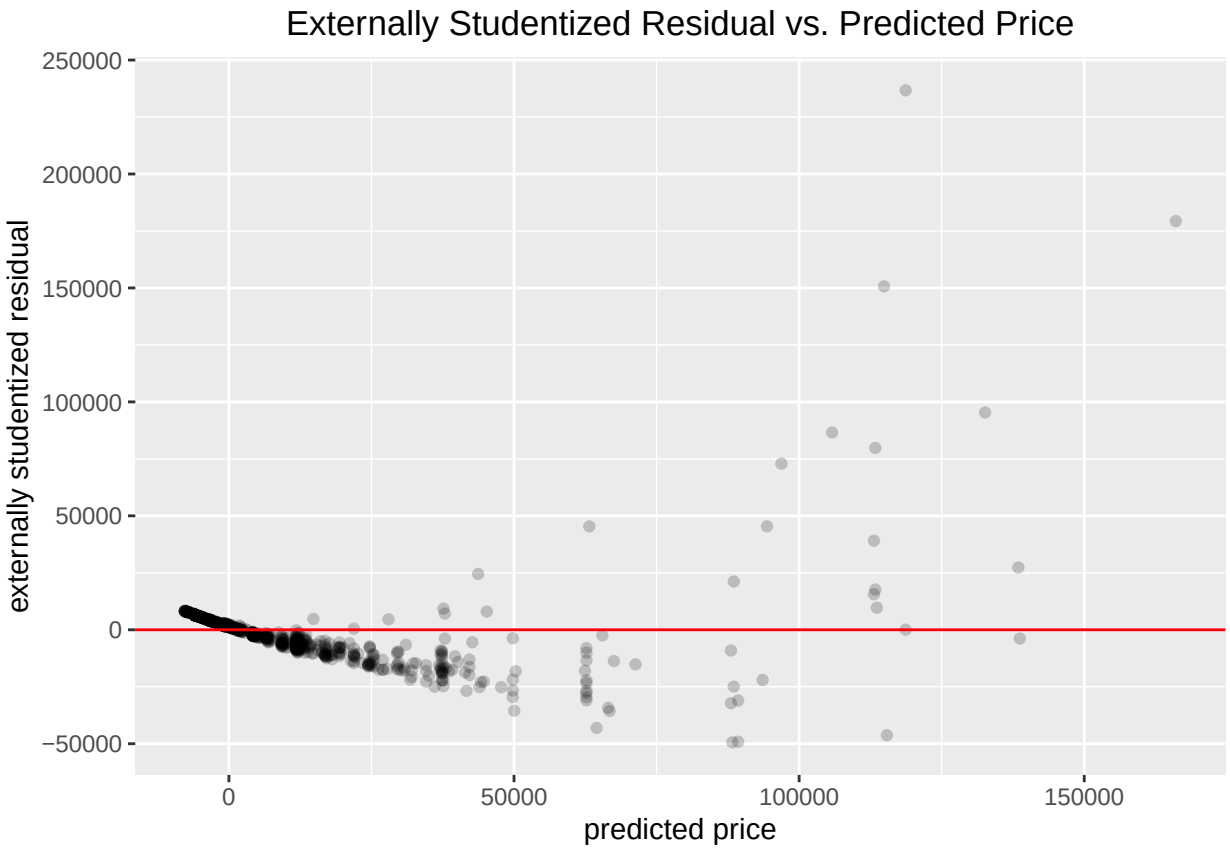
Considering the relationship of color of a diamond in our data set and the clarity of the diamond, we construct a bar chart of proportion of color by clarity. The proportion of diamonds with color identifier D for the group of diamonds with clarity identifier FL is greater than 0.5, is the only proportion 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.

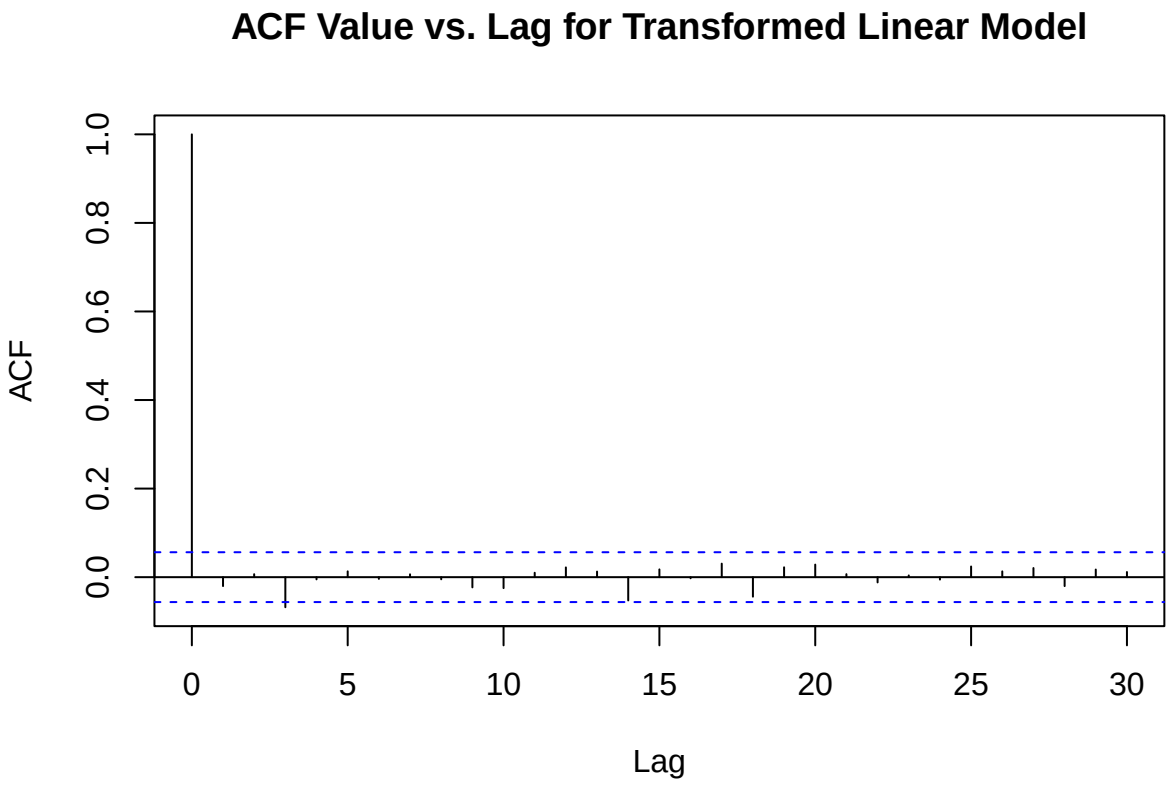


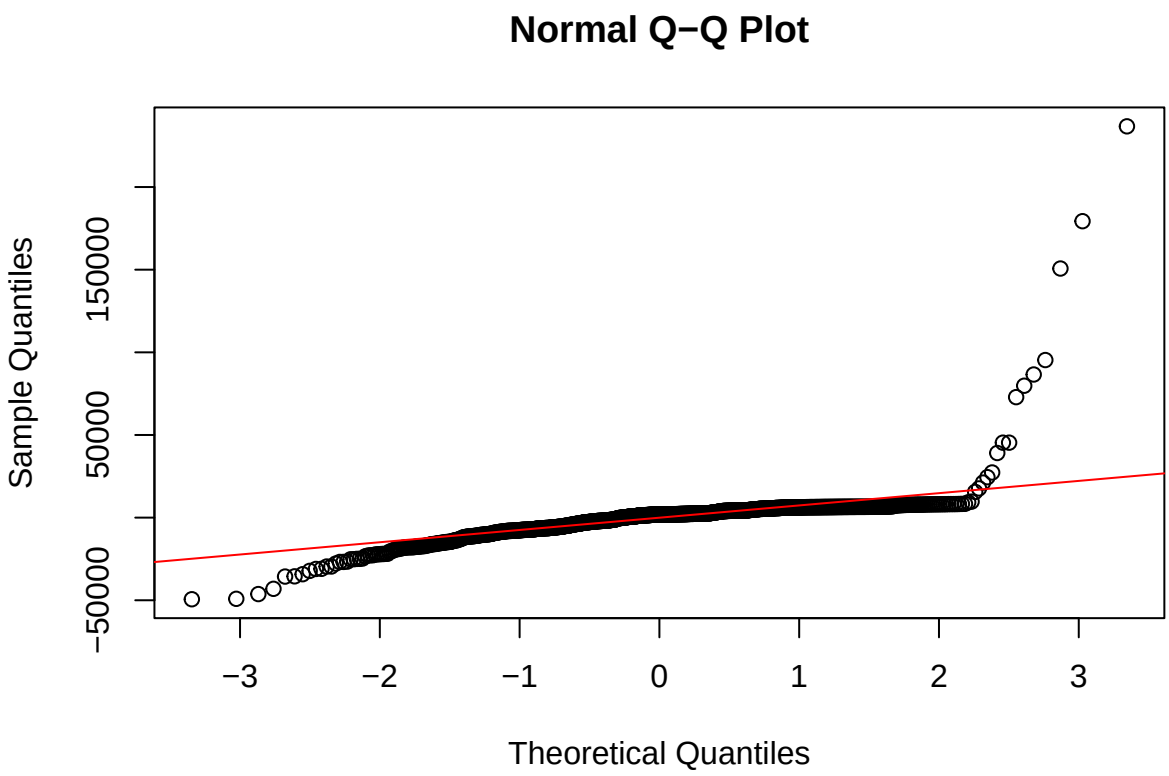
Considering the relationship of price of a diamond in our data set and the weight of the diamond, we construct a scatterplot for a linear model of price versus weight, as well as a plot of externally studentized residuals versus predicted prices for the model, a plot of AutoCorrelation Function values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


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


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



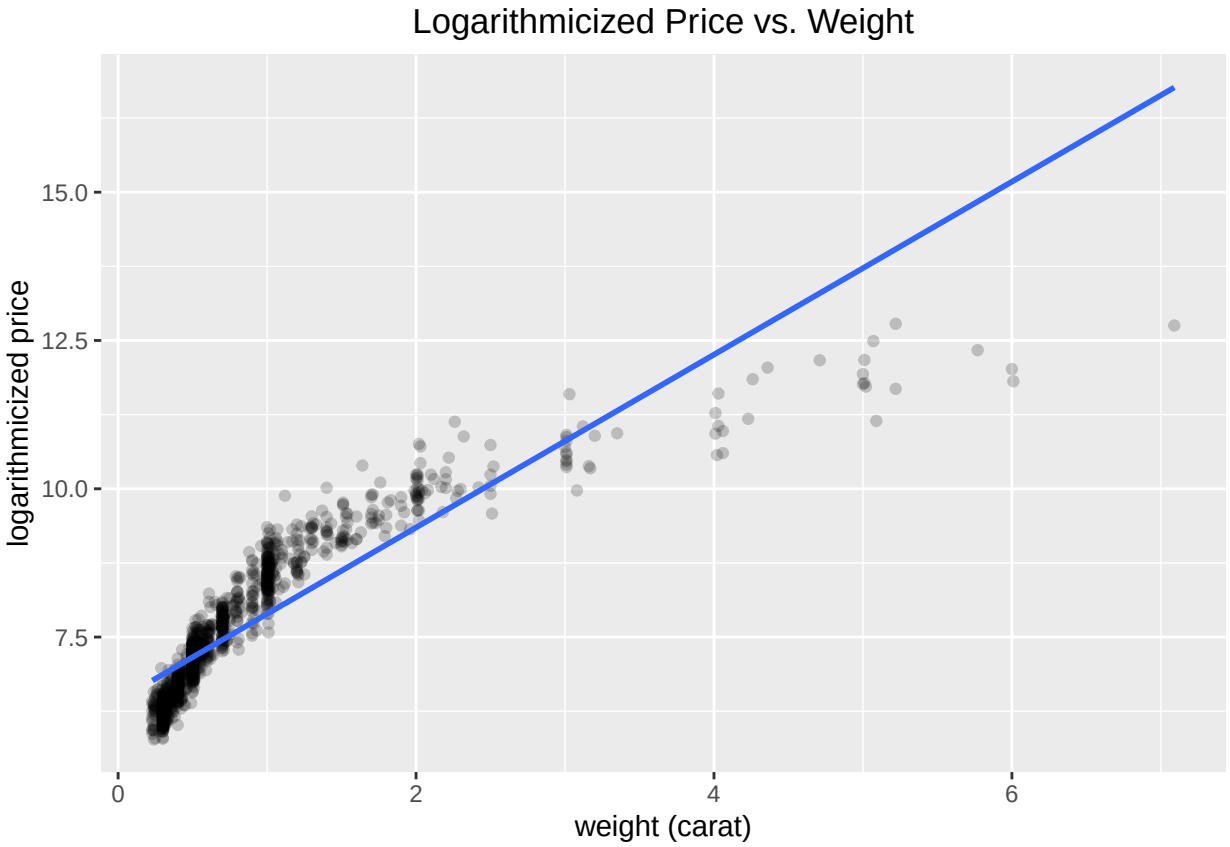
Simple Linear Regression of Price versus Weight

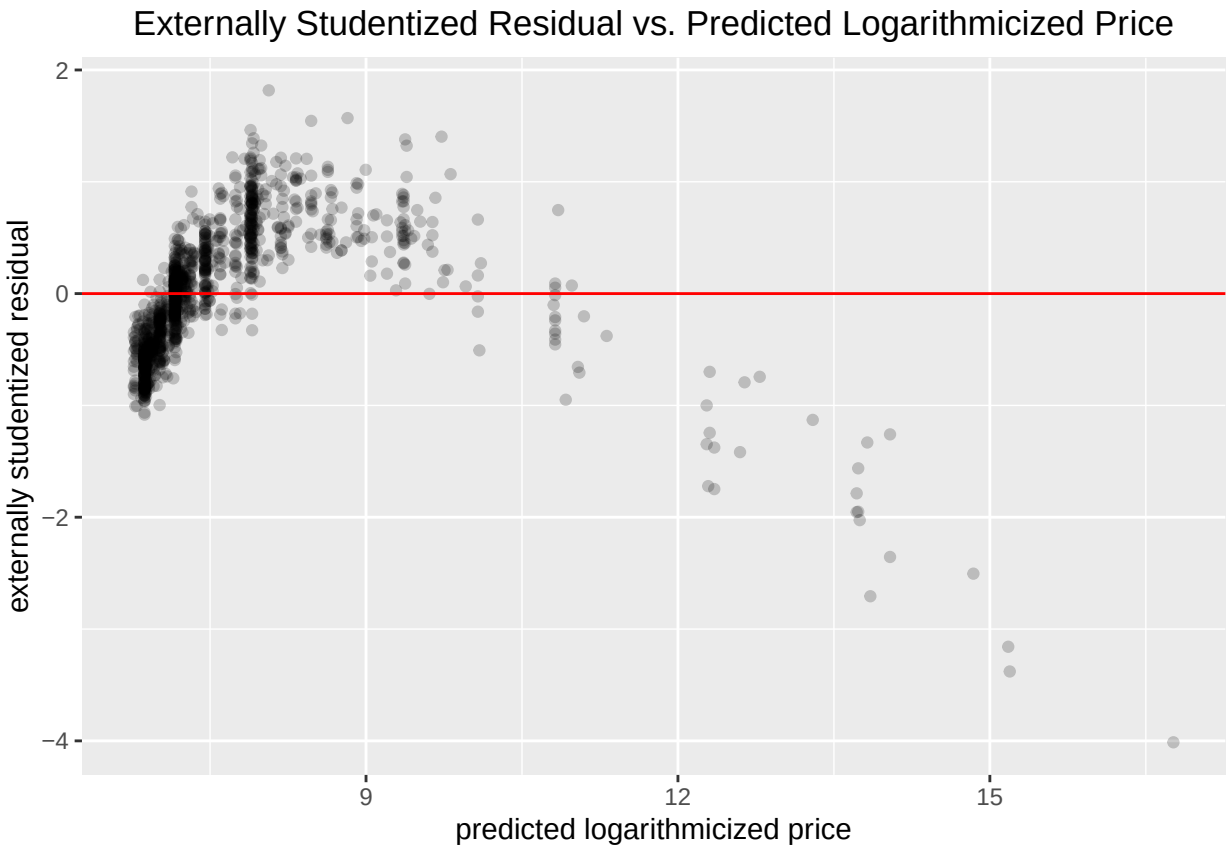
Simple linear regression depends of the following five assumptions being met.

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.

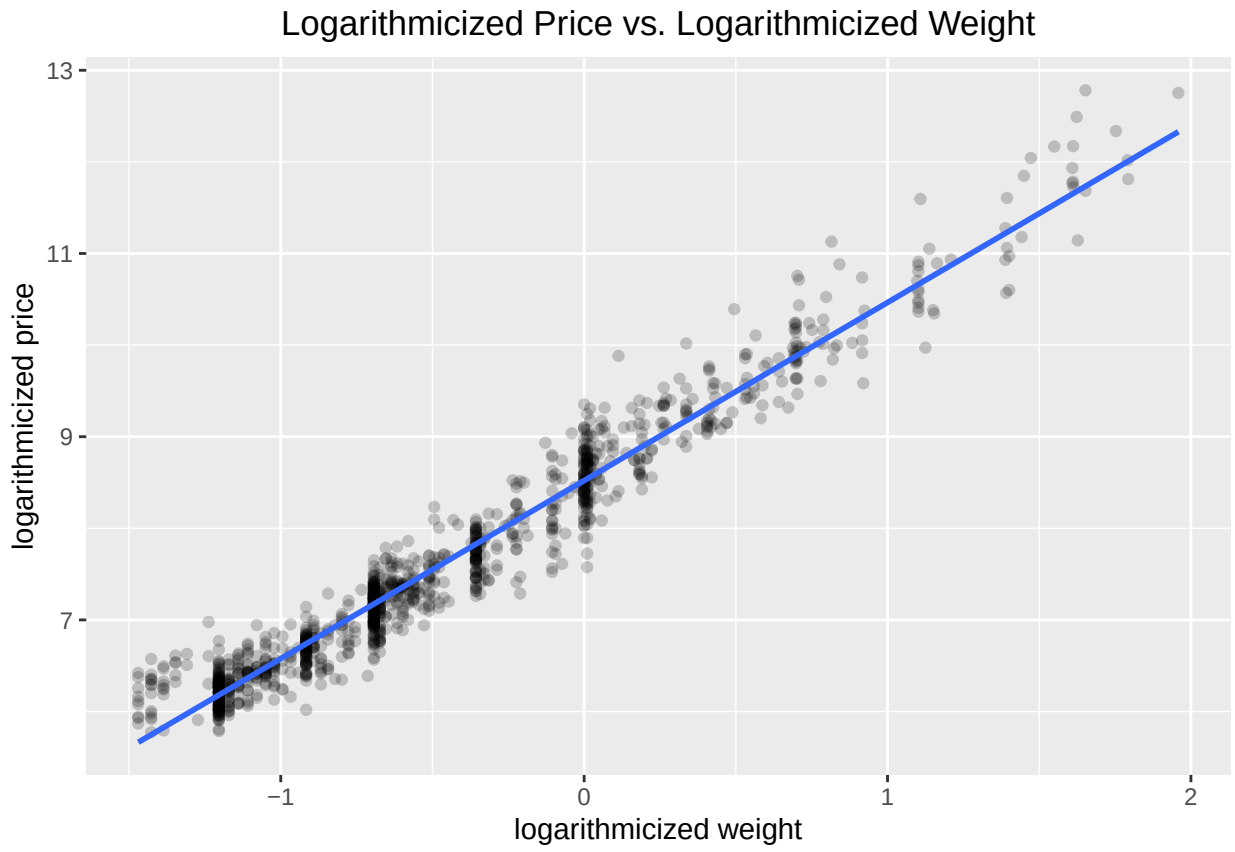
Given that the above assumptions for simple linear regression are not met, we generate a data set of transformed price and/or transformed weight such that all of the assumptions are met. The Box-Cox Method

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 logarithmicized price versus weight. We construct a linear model of logarithmicized price versus weight and present a scatterplot of externally studentized residuals versus predicted logarithmicized price for the linear model of logarithmicized price versus weight.

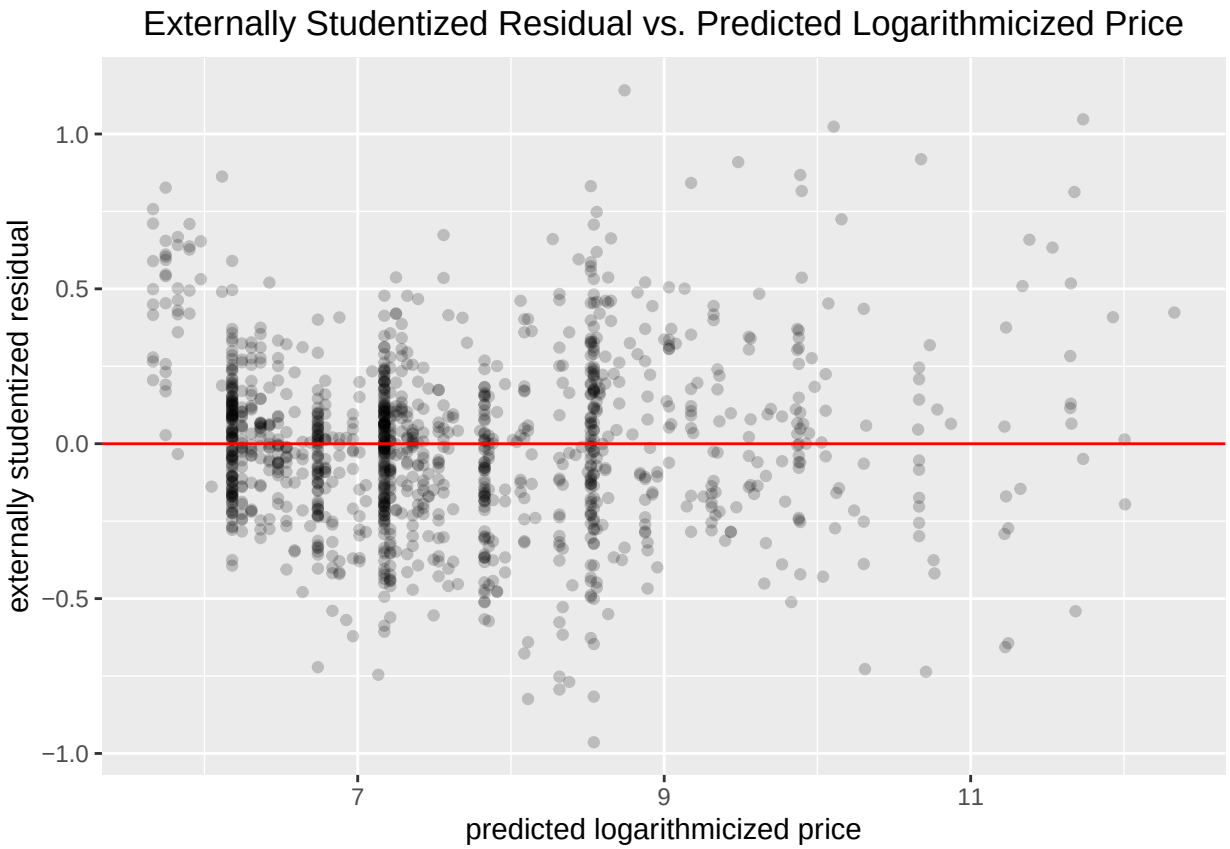


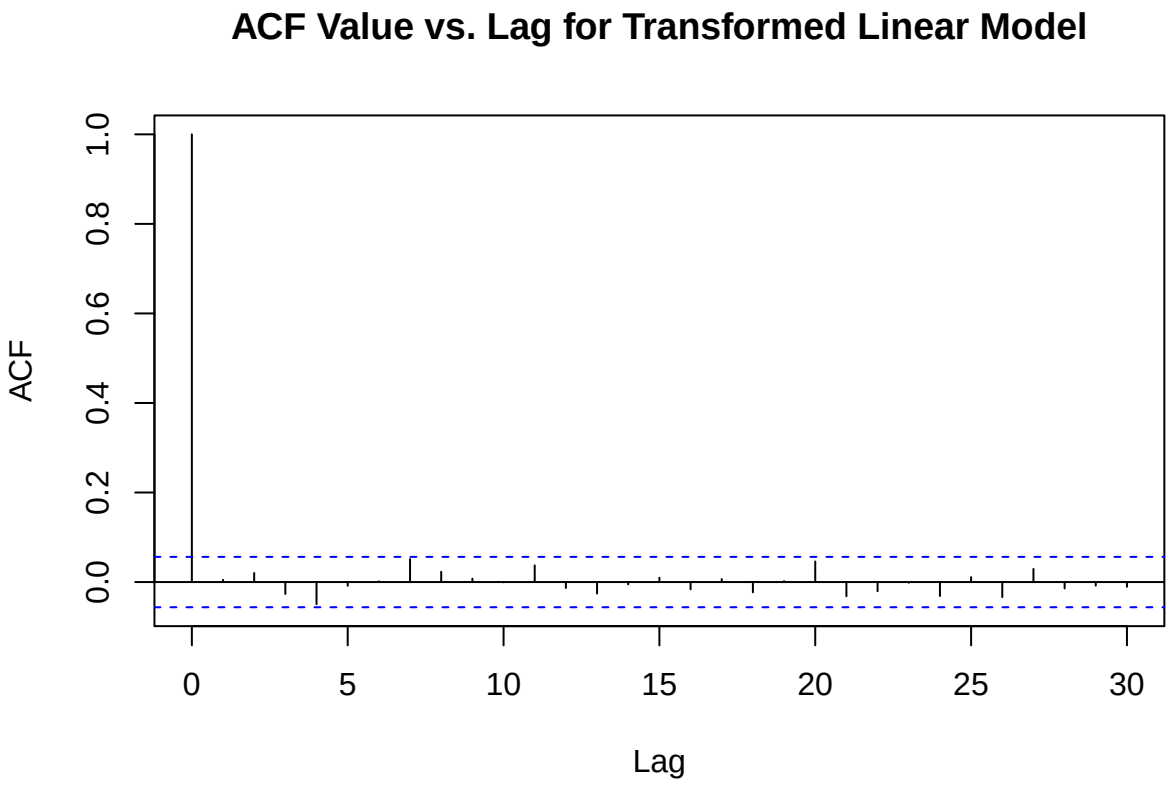


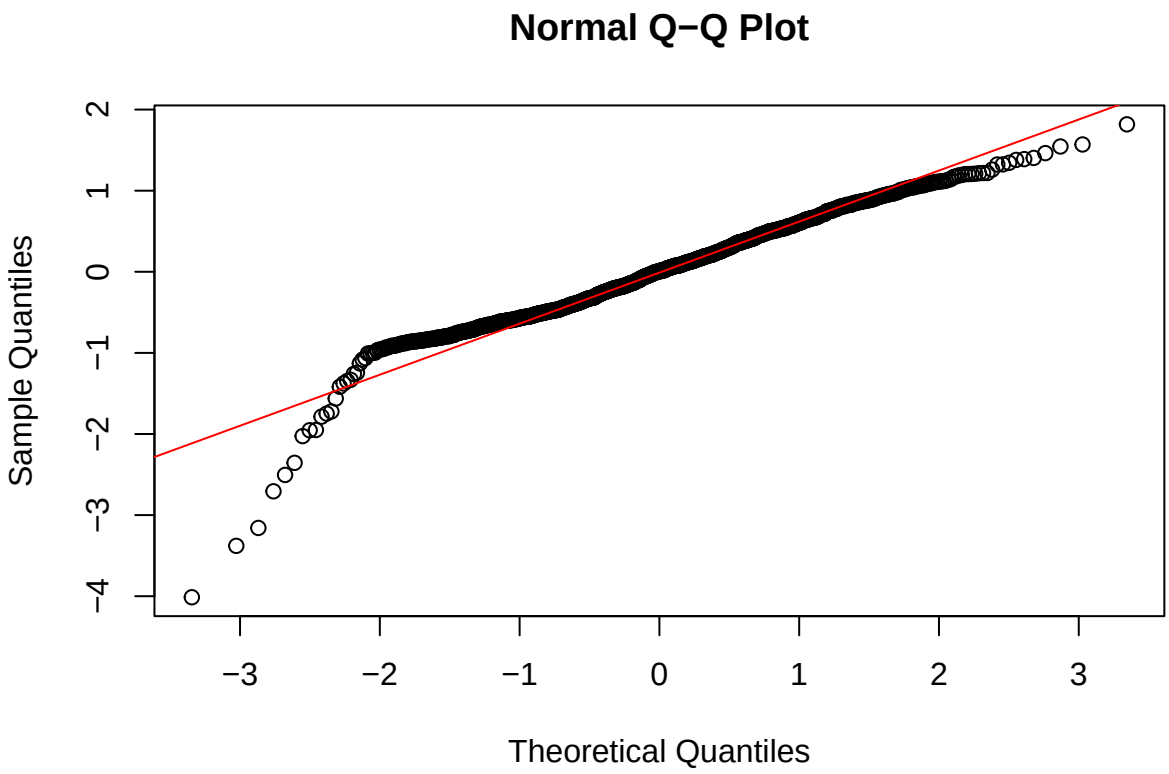
We have generated a data set of logarithmicized price versus weight for which the assumption that the distributions of residuals of a linear model of logarithmicized price and/or weight for different weights have constant variance. To generate a data set of transformed price and/or transformed weight 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 logarithmicized-price data and additionally logarithmicize weight. We present a scatterplot of logarithmicized price versus logarithmicized weight. We construct a linear model of logarithmicized price versus logarithmicized weight. We present a scatterplot of externally studentized residuals versus predicted logarithmicized price for the linear model of logarithmicized price versus logarithmicized 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.




```
##
## Call:
## lm(formula = logarithmicized_price ~ logarithmicized_weight,
##     data = data_set)
##
## 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 ***
## logarithmicized_weight 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
##
## E(y | x) = B_0 + B_1 * x = 8.52120787371624 + 1.94401996290127 * x
## Number of observations: 1214
## Estimated variance of errors: 0.0762394014101852
## Multiple R:  0.977080521362286   Adjusted R:  0.977061388921729
```


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


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



1. The assumption that the relationship between response / logarithmicized price and predictor / regressor / logarithmicized 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 logarithmicized price versus logarithmicized 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 logarithmicized 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 logarithmicized price versus logarithmicized 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 logarithmicized price versus logarithmicized weight are uncorrelated. We have insufficient evidence to conclude that the residuals of the linear model of logarithmicized price versus logarithmicized weight are correlated. We have insufficient evidence to conclude that the assumption that the residuals of the linear model of logarithmicized price versus logarithmicized weight are uncorrelated is not met.
5. The assumption that the residuals of the linear model of logarithmicized price versus logarithmicized 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 logarithmicized price versus logarithmicized 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 $1+p$, the predicted price increases by a factor of $(1+p)^{\hat{\beta}_1}$.