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.

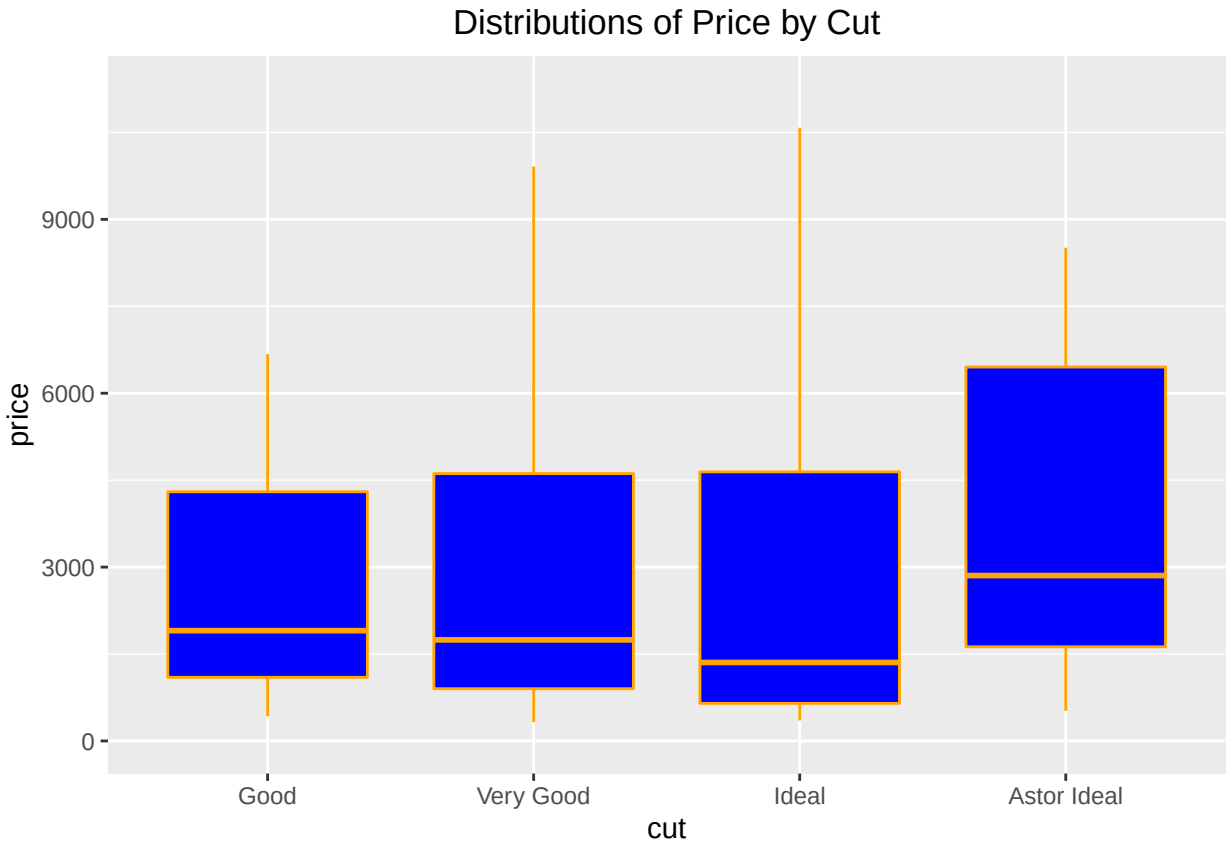
1. "The higher quality a diamond is, the more expensive it will be."
    a. The higher the weight of a diamond, the higher the price of the diamond.
    b. The closer the clarity identifier of a diamond is to "FL", the higher the price.
    c. The closer the color identifier of a diamond is to "D", the 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.

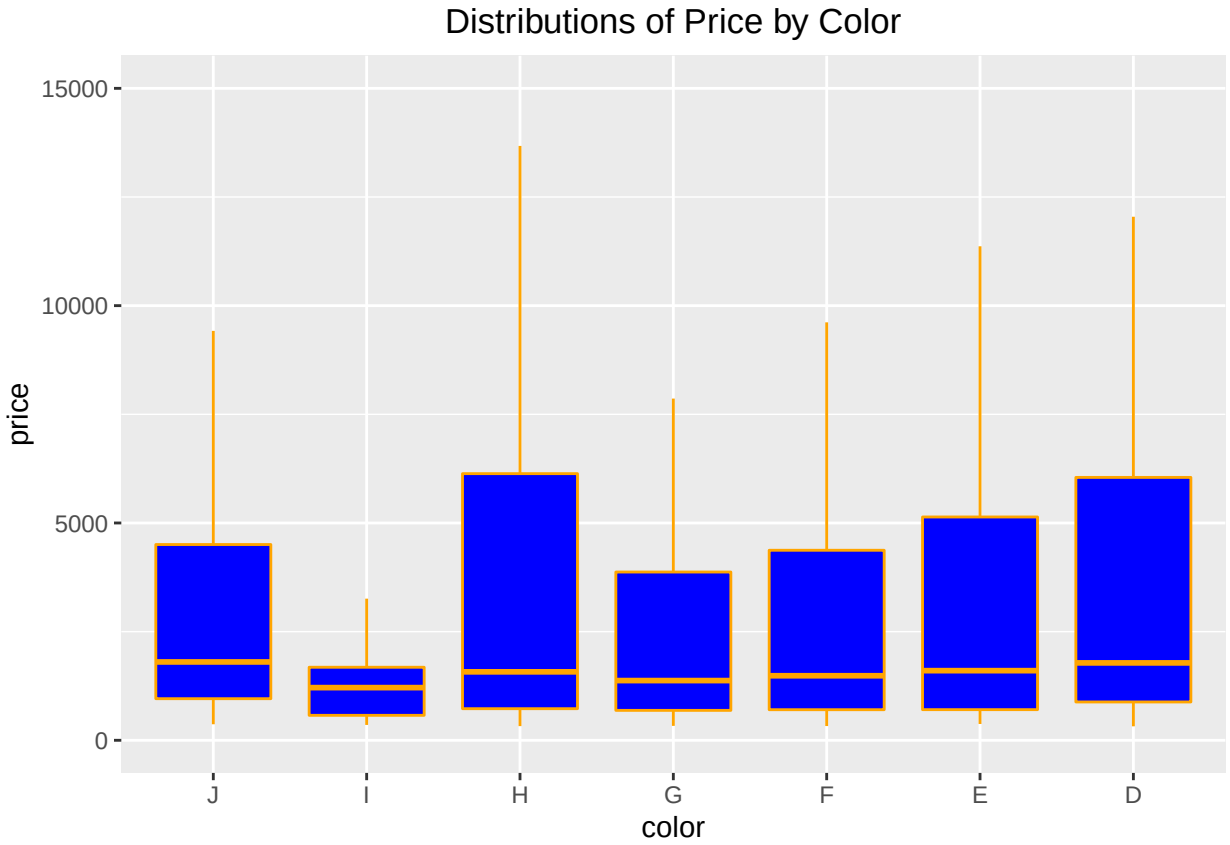
Add reasons why each visualization is presented.

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.

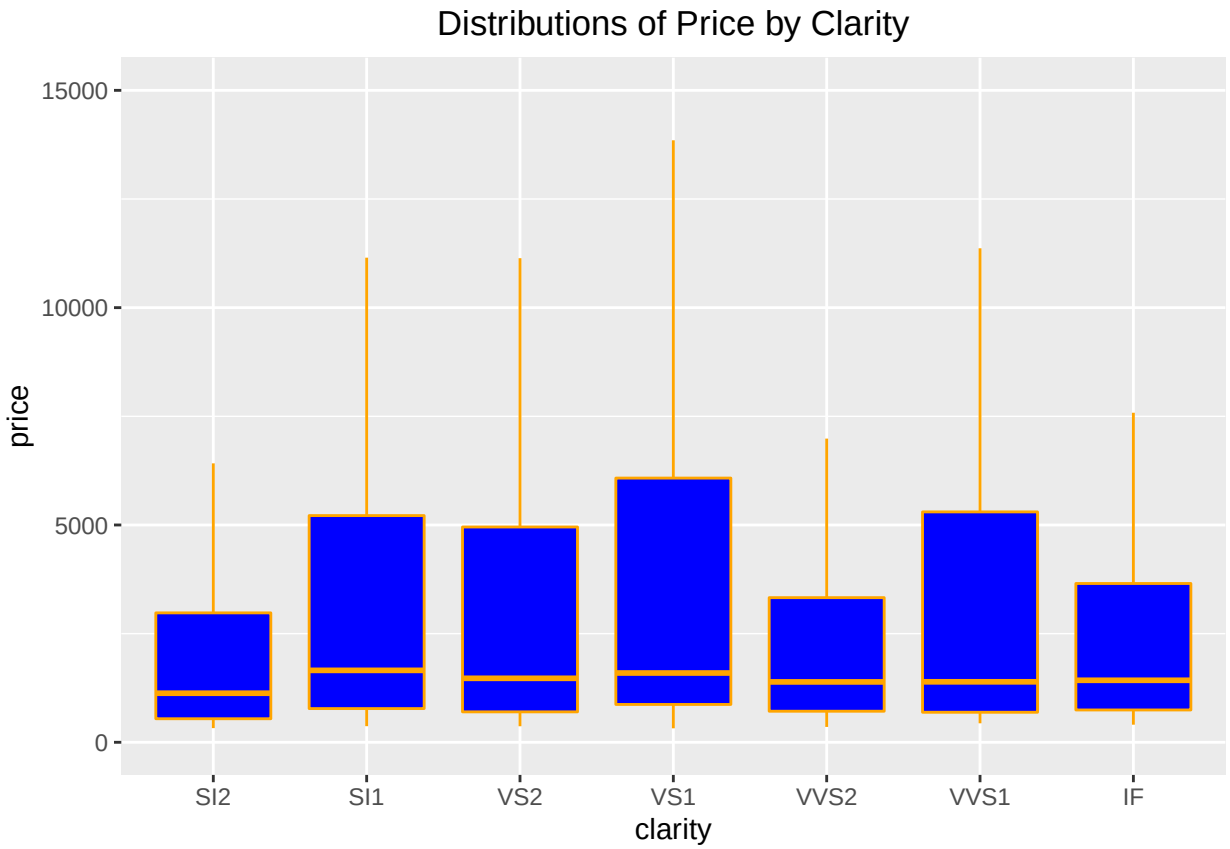


Considering the relationship of price of a diamond in our data set and the color of the diamond, we construct boxplots of price versus color with and without outliers and present the boxplot without outliers. Including outliers, for a transition from a group of diamonds with a color identifier in the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

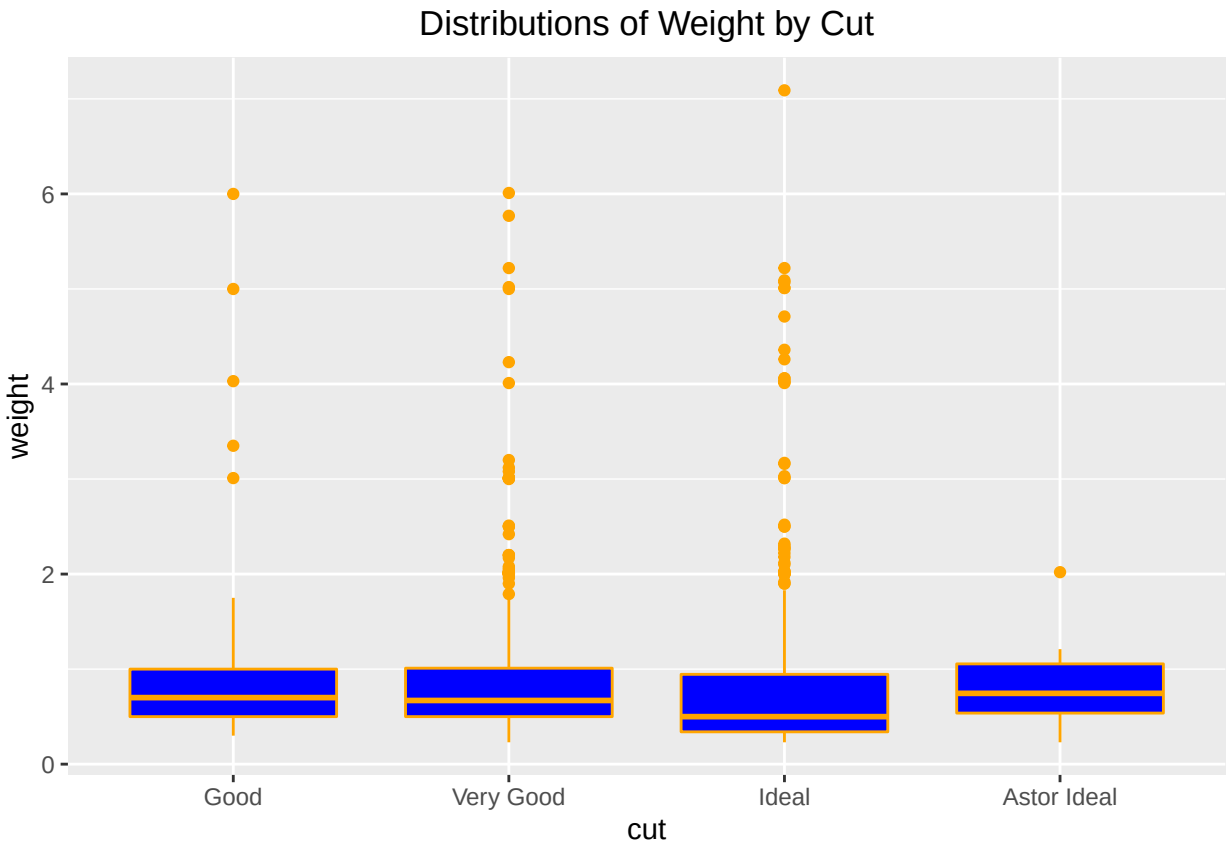


Considering the relationship of price of a diamond in our data set and the clarity of the diamond, we construct boxplots of price versus clarity with outliers and group of diamonds with clarity identifier FL, and without outliers and without the group of diamonds with clarity identifier FL. Including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

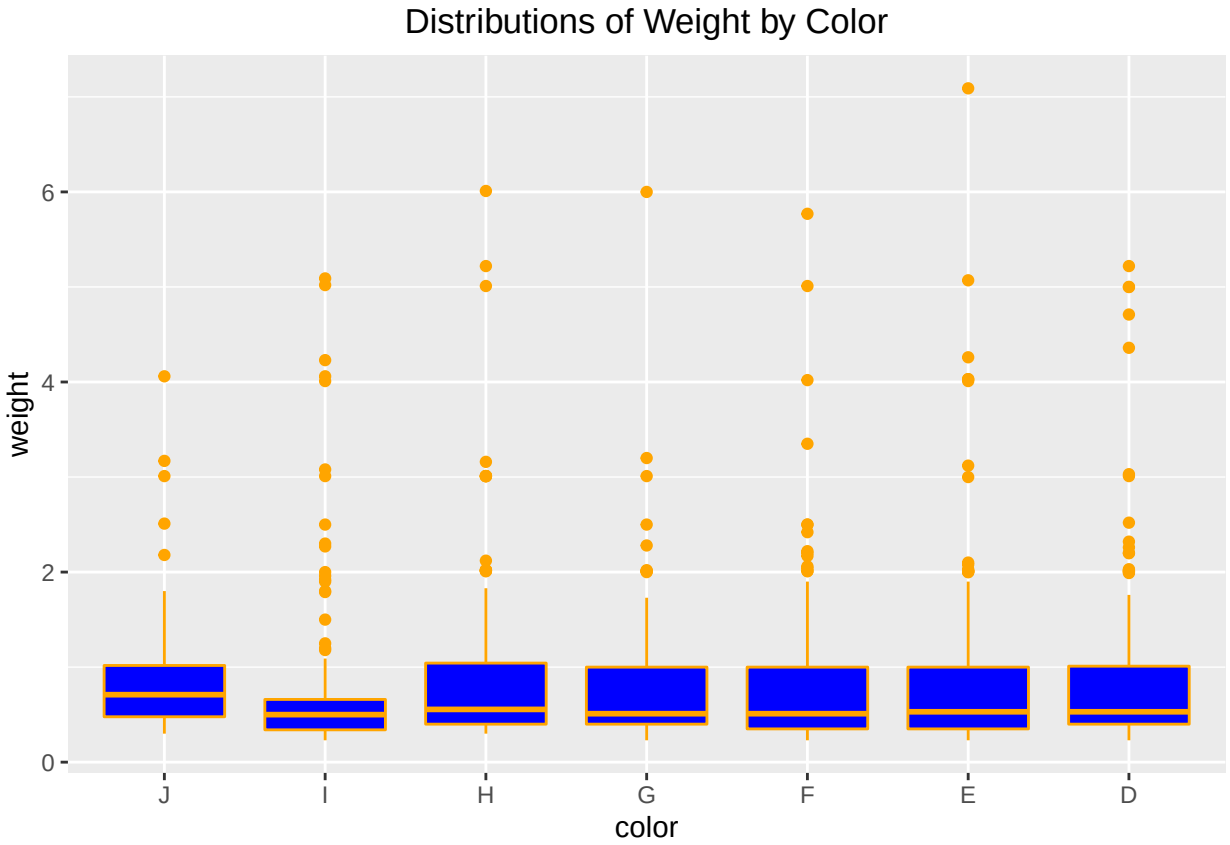

Distributions of Price by Clarity
price
3e+05
2e+05
1e+05
0e+00
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
FL
clarity



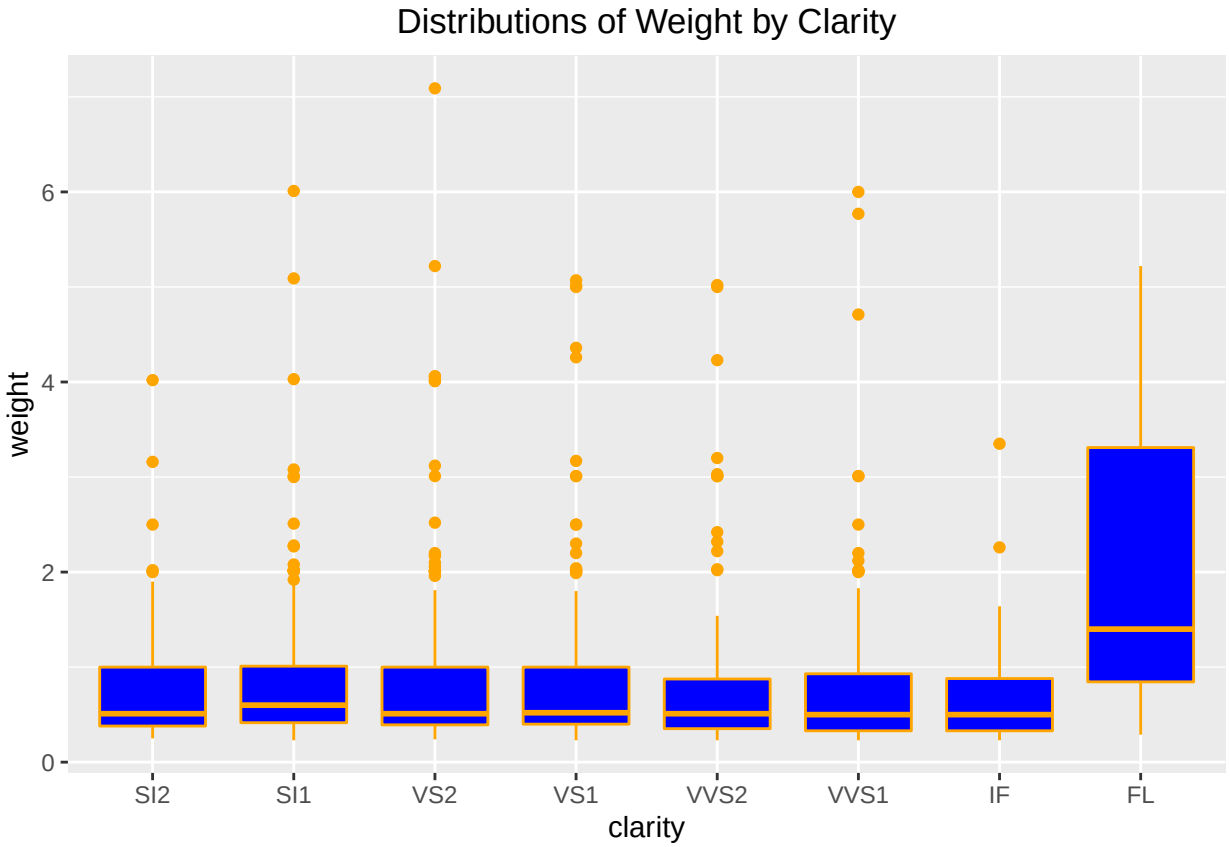
Considering the relationship of weight of a diamond in our data set and the cut of the diamond, we construct a boxplot of weight versus cut. Including outliers, an ideal diamond has the highest weight of over 5 *carat*. Excluding outliers, an ideal diamond the highest weight of about 2 *carat*. An Astor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

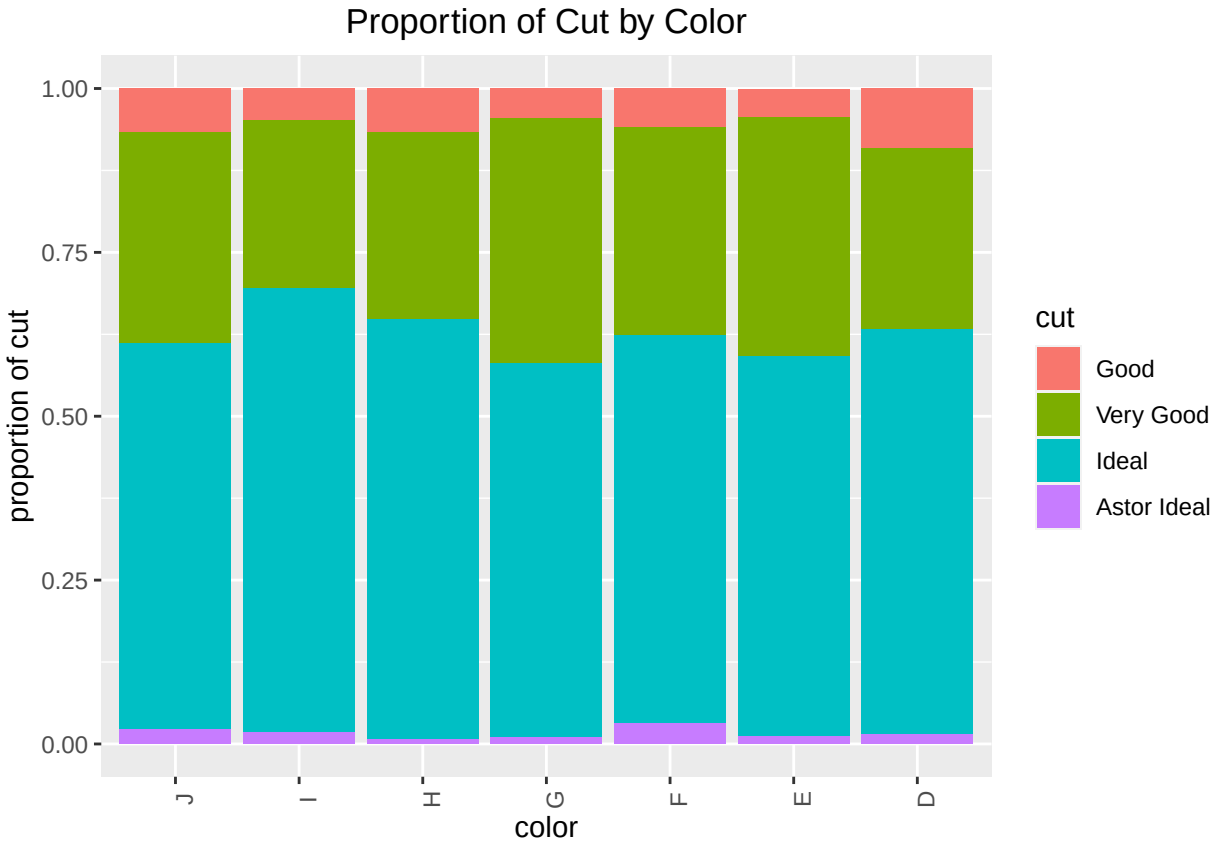


Considering the relationship of weight of a diamond in our data set and the clarity of the diamond, we construct a boxplot of weight versus color. Including outliers, a diamond with clarity identifier $VS2$ has the highest weight of over 7 *carat*. Excluding outliers, a diamond with clarity identifier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

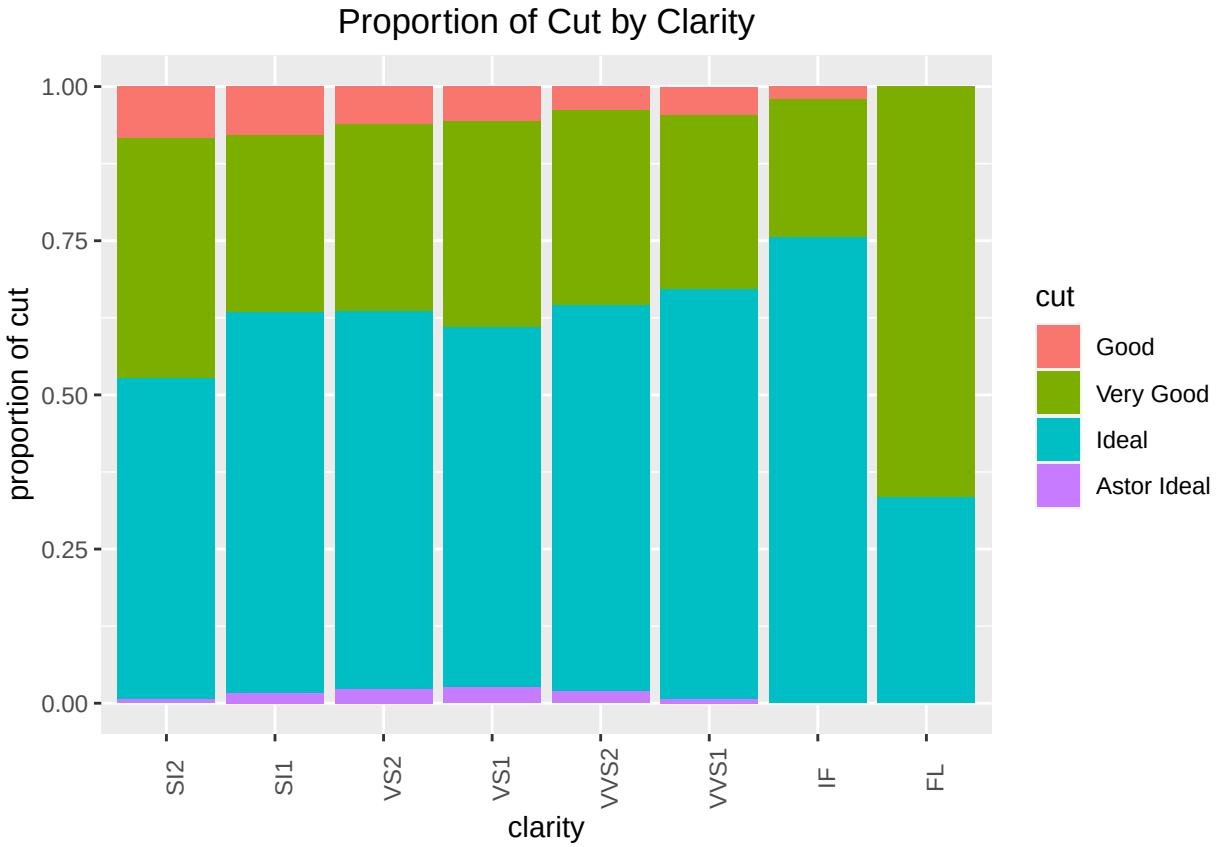



```
## [1] VS1
## Levels: SI2 SI1 VS2 VS1 VVS2 VVS1 IF FL
```

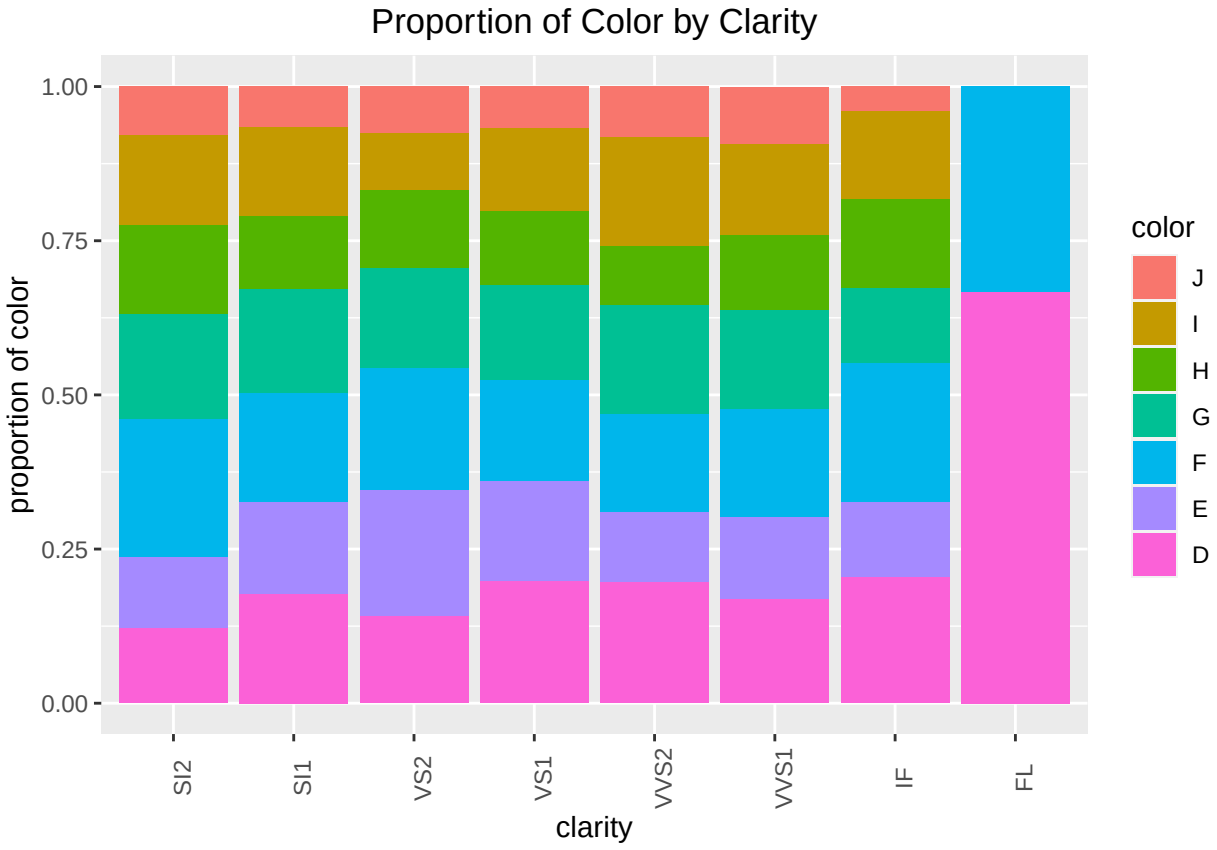
Considering the relationship of cut of a diamond in our data set and the color of the diamond, we construct a bar chart of proportion of cut by color. For each group of diamonds with a unique color identifier, most diamonds were ideal. The proportions of ideal, very good, good, and Astor ideal diamonds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



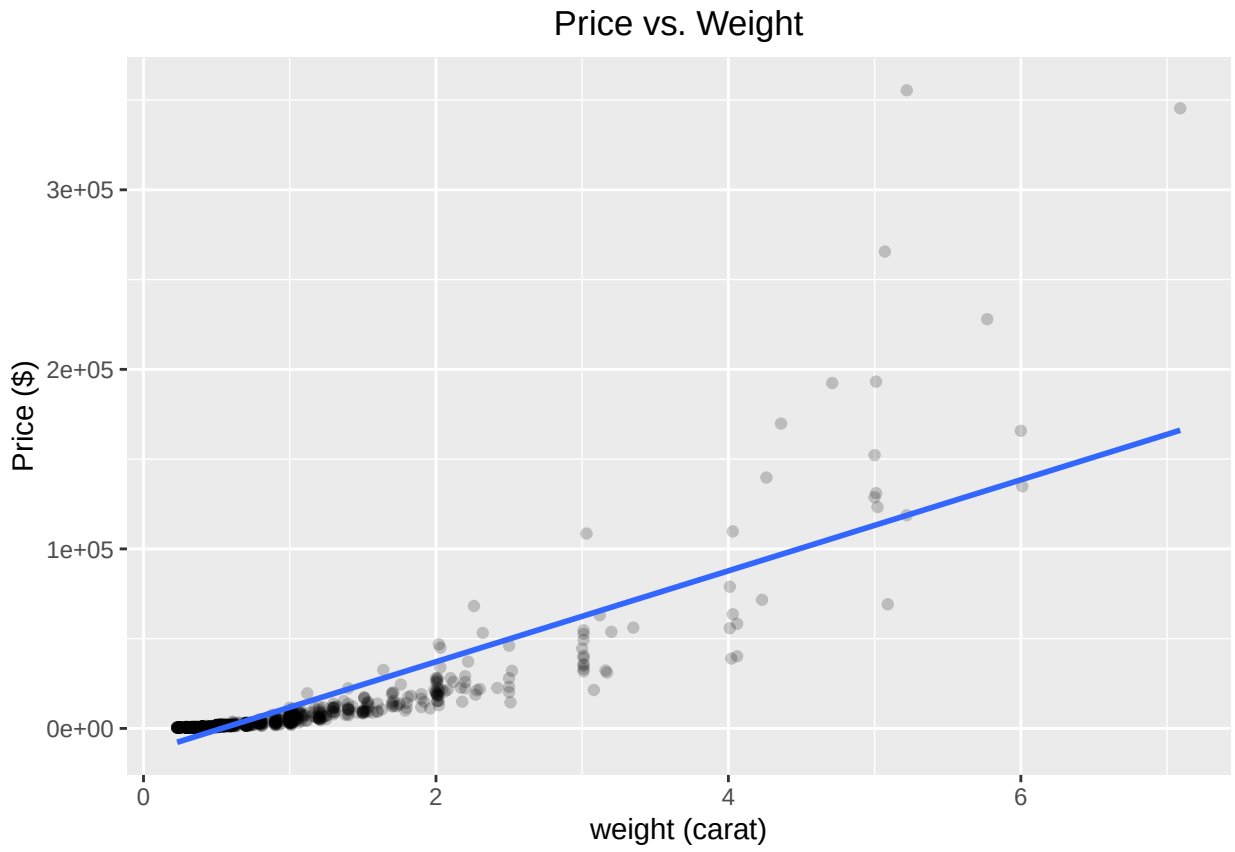
Considering the relationship of cut of a diamond in our data set and the clarity of the diamond, we construct a bar chart of proportion of cut by clarity. For a group of diamonds with a clarity identifier other than FL, most diamonds are ideal. For a transition from a group of diamonds with a clarity 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 of price versus weight.



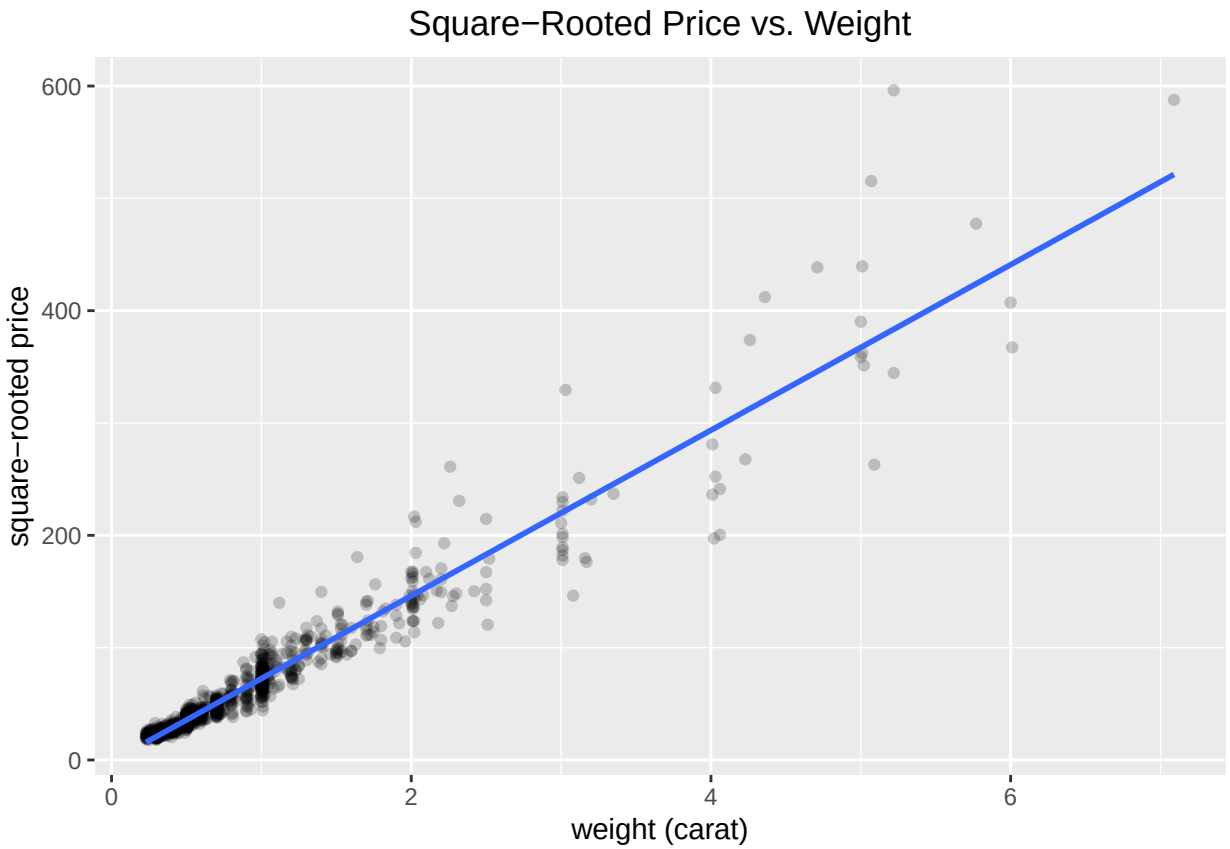
The relationship between price and weight seems quadratic. This intuition is supported by our below linear regression and choice of price transformation parameter $\lambda = 0.5$.

Simple Linear Regression of Price versus Weight

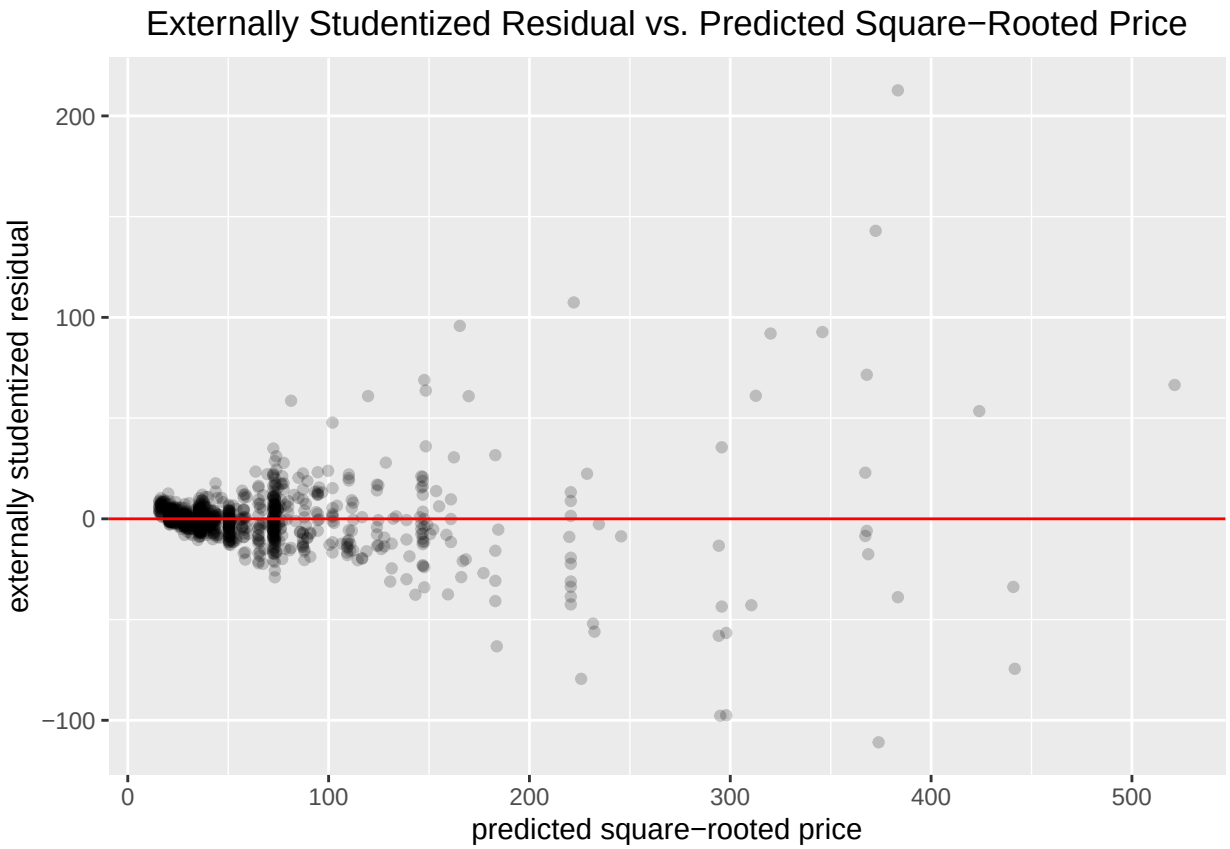
Present commentary and conclusions on an end linear-regression equation and how the SLR model informs us how price of diamonds are related to weight.

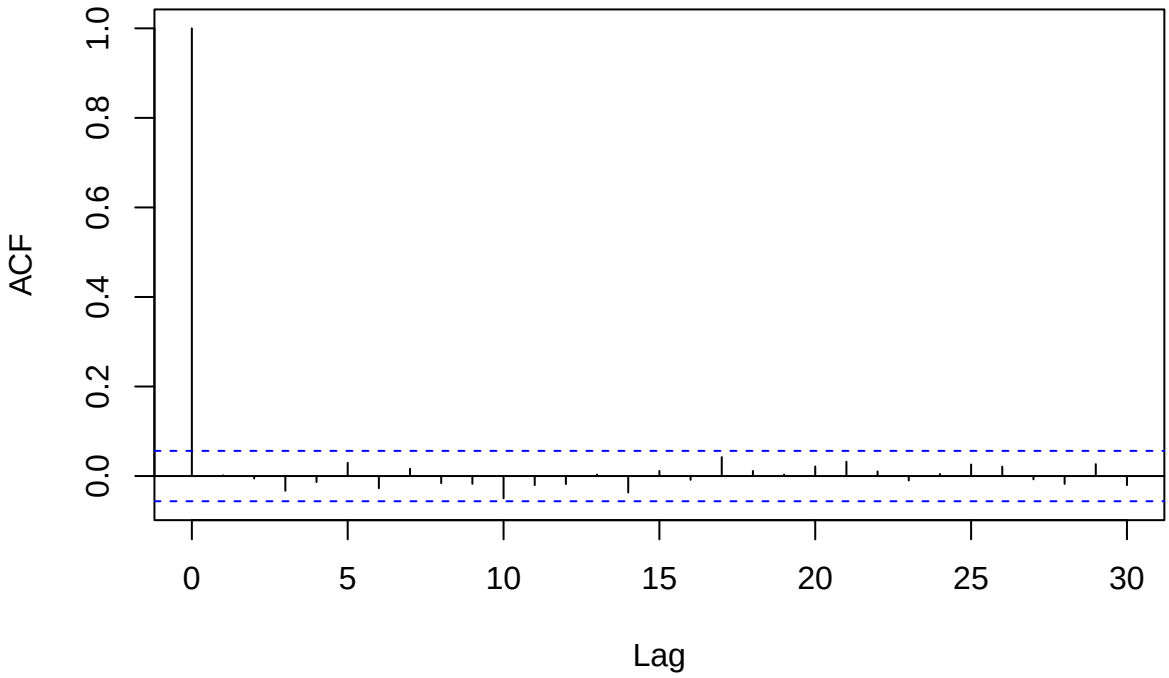
Using R, we construct a linear model of price versus weight. We assume that our sample of diamonds in simple random. We assume that (*weight*, *price*) matched pairs of data for the diamonds are independent. We assume that residuals between actual and predicted prices are normally and independently distributed with mean 0 and constant variance. Based on the above scatterplot of price versus weight, the relationship between price and weight seems quadratic and nonlinear.

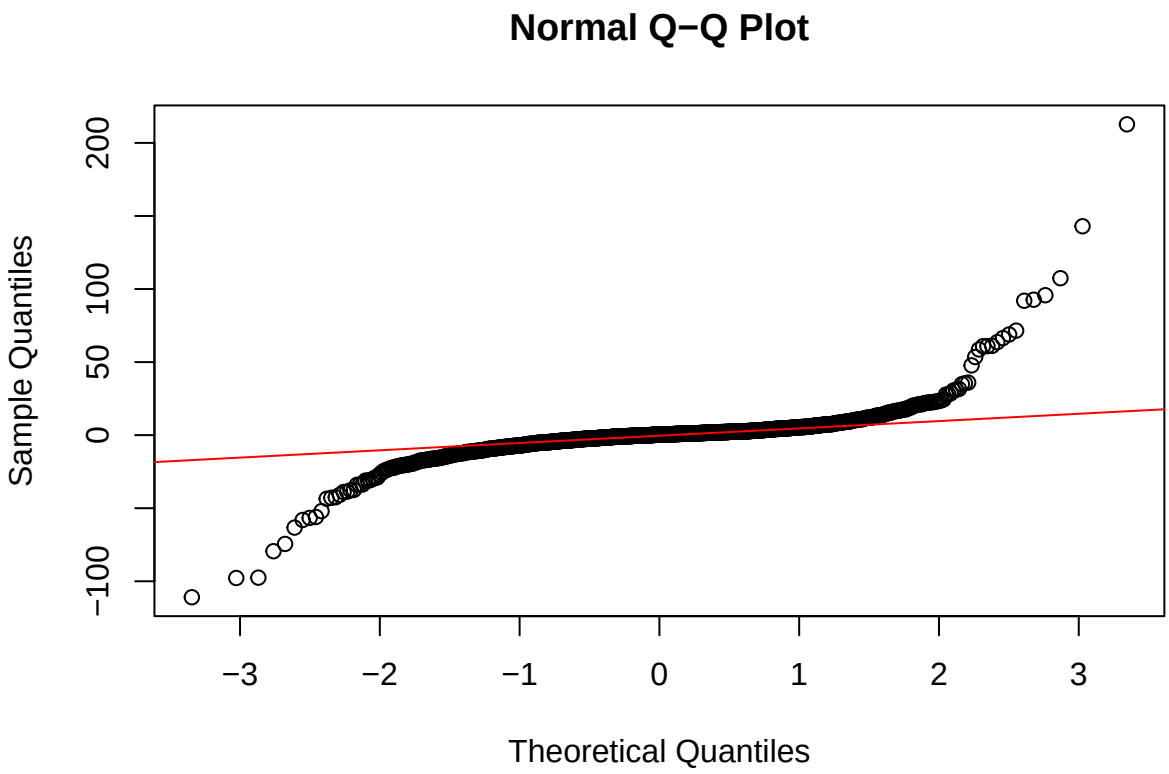
The Box-Cox Method is presented in section 5.4.1 of *Introduction to Linear Regression Analysis* (Sixth Edition) by Douglas C. Montgomery et al.. Because the above assumption that the error term of the linear model has constant variance is not met, using R, we perform the Box-Cox Method to determine a maximum-likelihood estimate of a parameter $\lambda = 0.311$ to be used in a power transformation $\hat{y}' = \hat{y}^{\lambda}$ of the predicted prices $\hat{y}$. After some experimentation, we choose a parameter $\lambda = 0.5$. This choice supports our intuition that the relationship between price and weight seems quadratic. We transform predicted price and present a scatterplot of transformed price versus weight.



Using R, we construct a linear model of squared-rooted price versus weight. We present a scatterplot of externally studentized residual versus predicted squared-rooted price.


Externally Studentized Residual vs. Predicted Square−Rooted Price
externally studentized residual
200
100
0
−100
0
100
200
300
400
500
predicted square−rooted 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



We determine an estimated linear-regression equation

$$\sqrt{E(y|x)} = \beta_0 + \beta_1 x = (-1.216\ \$) + \left(73.692\ \frac{\$}{carat}\right) x$$

where $E(y|x)$ is the expected price given a weight x.

Some assumptions for simple linear regression appear to be met.

1. The assumption that the relationship between response / yields y and predictor / regressor / nitrogen fertilizer applications x is linear, at least approximately, is met. The relationship appears to be linear.
2. The assumption that the error term ϵ of the linear model has mean 0 is met. Observations are scattered evenly around the fitted line. Residuals may be evenly scattered aroun $e = 0$.
3. The assumption that the error term ϵ of the linear model has constant variance is not met. The vertical variation of observations is not constant. Residuals are not evenly scattered around $e = 0$.
4. The assumption that the errors ϵ_i / residuals e_i 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 values are insignificant, we have insufficient evidence to reject a null hypothesis that the residuals of the linear model are uncorrelated. We have insufficient evidence to conclude that the residuals of the linear model are correlated. We have insufficient evidence to conclude that the assumption that the errors ϵ_i / residuals e_i are uncorrelated is not met.
5. Assumptions that the errors ϵ_i / residuals e_i are normally distributed is not met. A linear model is robust to these assumptions. Given sharp downward and upward curves at extremes at a plot of sample quantiles versus theoretical quantiles for the residuals of the transformed linear model, the tails of the probability vs. externally studentized residuals plot / distribution are too light for this distribution to be considered normal. The assumption that the residuals are normally distributed is not met.