PSTAT126 - Regression Analysis

Final Project - Diamonds Analyis

Quinn Harrington and Landen Kurtz

2024-06-03

Summary Statistics - Entire Dataset

Table 1: Data summary

Name	diamonds
Number of rows	53943
Number of columns	11
Column type frequency:	
character	3
numeric	8
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
cut	0	1	4	9	0	5	0
color	0	1	1	1	0	7	0
clarity	0	1	2	4	0	8	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
X	0	1	26972.00	15572.15	1.0	13486.50	26972.00	40457.50	53943.00
carat	0	1	0.80	0.47	0.2	0.40	0.70	1.04	5.01
depth	0	1	61.75	1.43	43.0	61.00	61.80	62.50	79.00
table	0	1	57.46	2.23	43.0	56.00	57.00	59.00	95.00
price	0	1	3932.73	3989.34	326.0	950.00	2401.00	5324.00	18823.00
x	0	1	5.73	1.12	0.0	4.71	5.70	6.54	10.74
y	0	1	5.73	1.14	0.0	4.72	5.71	6.54	58.90
z	0	1	3.54	0.71	0.0	2.91	3.53	4.04	31.80

Using the Summary command from the “Skimr” Package, we find that the “Diamonds” data-set contains data on 53,943 diamonds. Each diamond observation contains data stored in three categorical/character

variables and eight quantitative/numeric variables. This includes an additional numeric variable, X, used to assign index numbers to each observation of diamond.

The original diamonds data set is very large, which may make statistical analysis more difficult. Therefore, we wish the acquire a smaller subset of the data to analyze. To accomplish this, we will take a random sample of 500 observations from the diamonds dataset using R. By taking a random sample, it ensures that the observational units(diamonds) of our sample will be independent of one another.

Preview - Sample of 500 Diamonds

```
##       X carat   cut color clarity depth table price    x    y    z
## 1 51663  0.73 Ideal     I     VS1  60.7    56  2397 5.85 5.81 3.54
## 2  2986  0.70 Ideal     G     VS1  60.8    56  3300 5.73 5.80 3.51
## 3 29925  0.31 Ideal     D     VS1  61.6    55   713 4.30 4.33 2.66
## 4 29710  0.31 Ideal     H    VVS1  62.2    56   707 4.34 4.37 2.71
## 5 37529  0.31 Ideal     E      IF  60.9    55   987 4.39 4.41 2.68
## 6  2757  0.83  Good     E     SI1  63.7    59  3250 5.95 5.89 3.77
```

Summary Statistics - Sample of 500 Diamonds

Table 4: Data summary

Name	diamonds_sample
Number of rows	500
Number of columns	11
Column type frequency:	
character	3
numeric	8
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
cut	0	1	4	9	0	5	0
color	0	1	1	1	0	7	0
clarity	0	1	2	4	0	8	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
X	0	1	26969.30	15903.03	31.00	13324.25	26637.50	41057.00	53899.00
carat	0	1	0.80	0.46	0.23	0.38	0.71	1.05	2.30
depth	0	1	61.83	1.36	57.80	61.00	61.90	62.50	67.70
table	0	1	57.38	2.21	52.00	56.00	57.00	59.00	66.00
price	0	1	3922.04	3970.49	368.00	956.75	2537.00	5283.75	18118.00
x	0	1	5.73	1.12	3.93	4.65	5.73	6.56	8.43
y	0	1	5.73	1.11	3.96	4.64	5.72	6.55	8.46

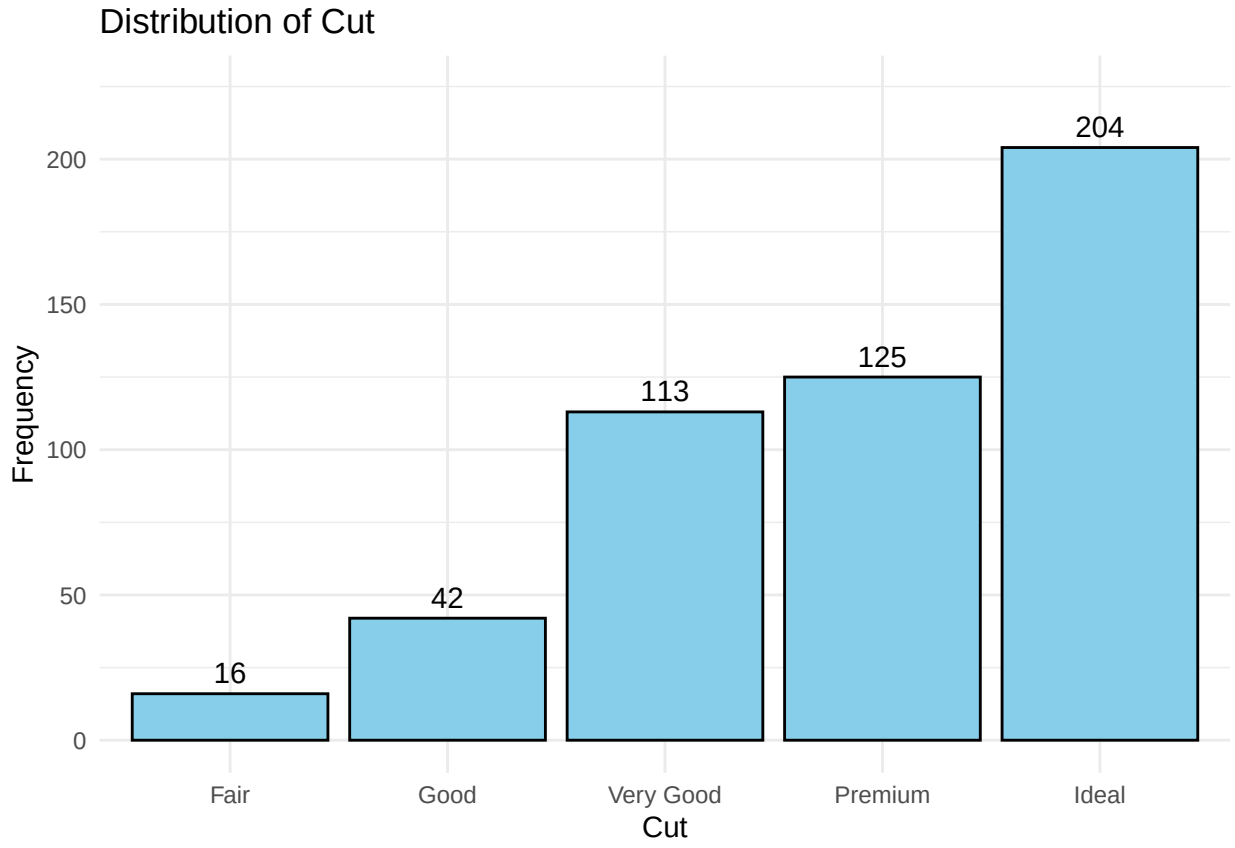
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
z	0	1	3.54	0.69	2.38	2.85	3.53	4.03	5.28

```
## Possible Values for Categorical Variables:

## Diamond Cut :
## [1] "Ideal"     "Premium"   "Good"      "Very Good" "Fair"
##
## Diamond Color :
## [1] "E" "I" "J" "H" "F" "G" "D"
##
## Diamond Clarity :
## [1] "SI2"  "SI1"  "VS1"  "VS2"  "VVS2" "VVS1" "I1"   "IF"
```

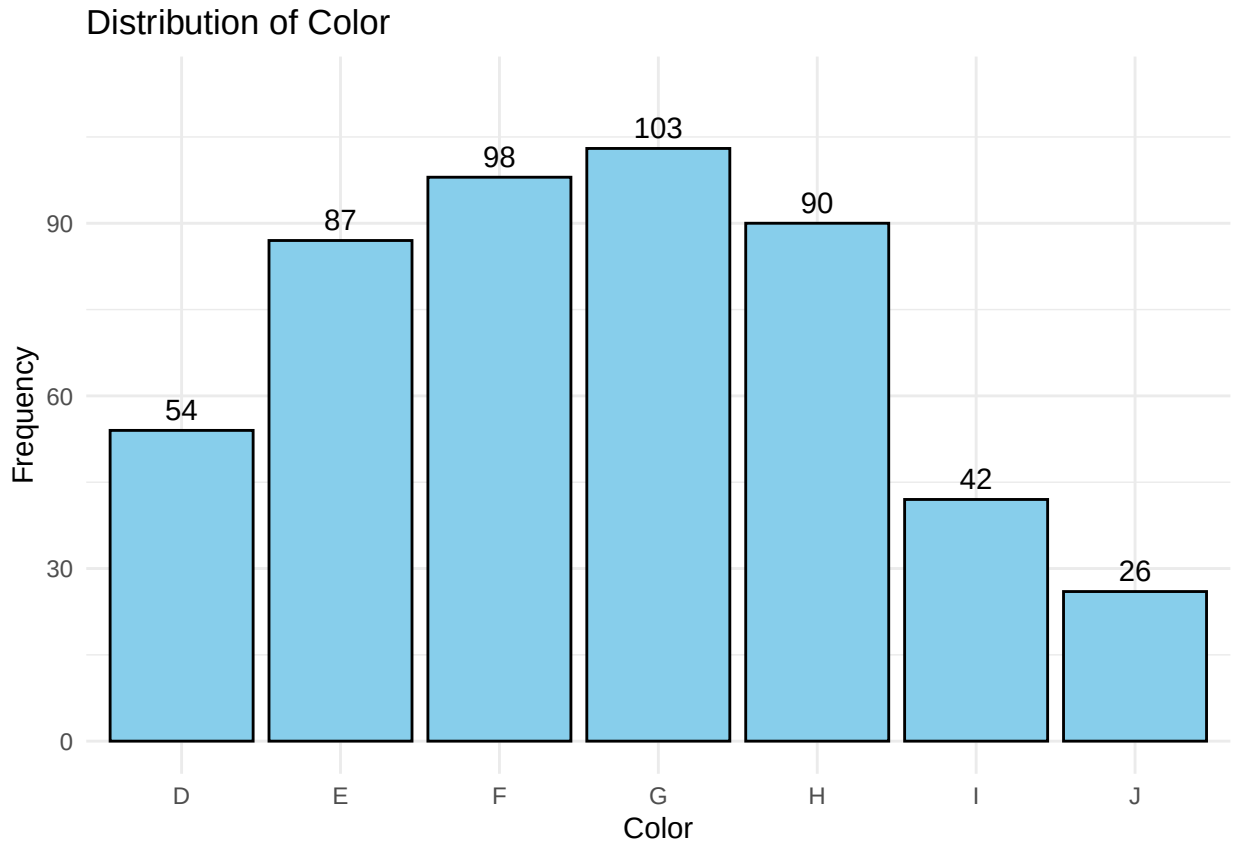
Qualitative/Categorical Variables:

Cut - Describes the quality of the diamond's cut. Possible values for the "Cut" variable are Fair, Good, Very Good, Premium, and Ideal. Fair is considered the lowest quality, while Ideal is considered the highest quality.



The distribution of the values of "Cut" in our random sample of 500 diamonds is shown below. We can observe that the distribution of "Cut" is skewed to the left. This skew in the distribution tells us that we have many more observations with cut values of "Very good, premium, and ideal" while many fewer observations have cut values of "Fair and Good". Therefore, it appears that our random sample has many more diamonds that are considered to be of medium to the highest quality of cut.

Color - Describes the color of the diamond. Possible color values are D, E, F, G, H, I, and J. Colors D, E, and F are considered colorless diamonds and are rarer and typically of higher value. Colors D, H, I, and J are considered nearly colorless- meaning they may have slight color that may not be visible to the naked eye, but this color is visible when viewed using different forms of magnification. These diamonds are considered of less value than colors D-F and are therefore typically more accessible.



Examining the distribution of "Color", we find that the possible values of "Color" for the 500 observations of diamonds in our sample are roughly normally distributed with a very slight skew to the right. This tells us that, in our random sample, the majority of the observations have color values of medium color rarity- while there are less diamonds with color values nearing the extremities of the possible color values on both the high and low ends of rarity.

Clarity - Describes how clear the diamond is. Possible values in our data are I1, SI1, SI2, VS1, VS2, VVS1, VVS2, and IF. I1 is considered the lowest quality, while IF is considered the highest quality.

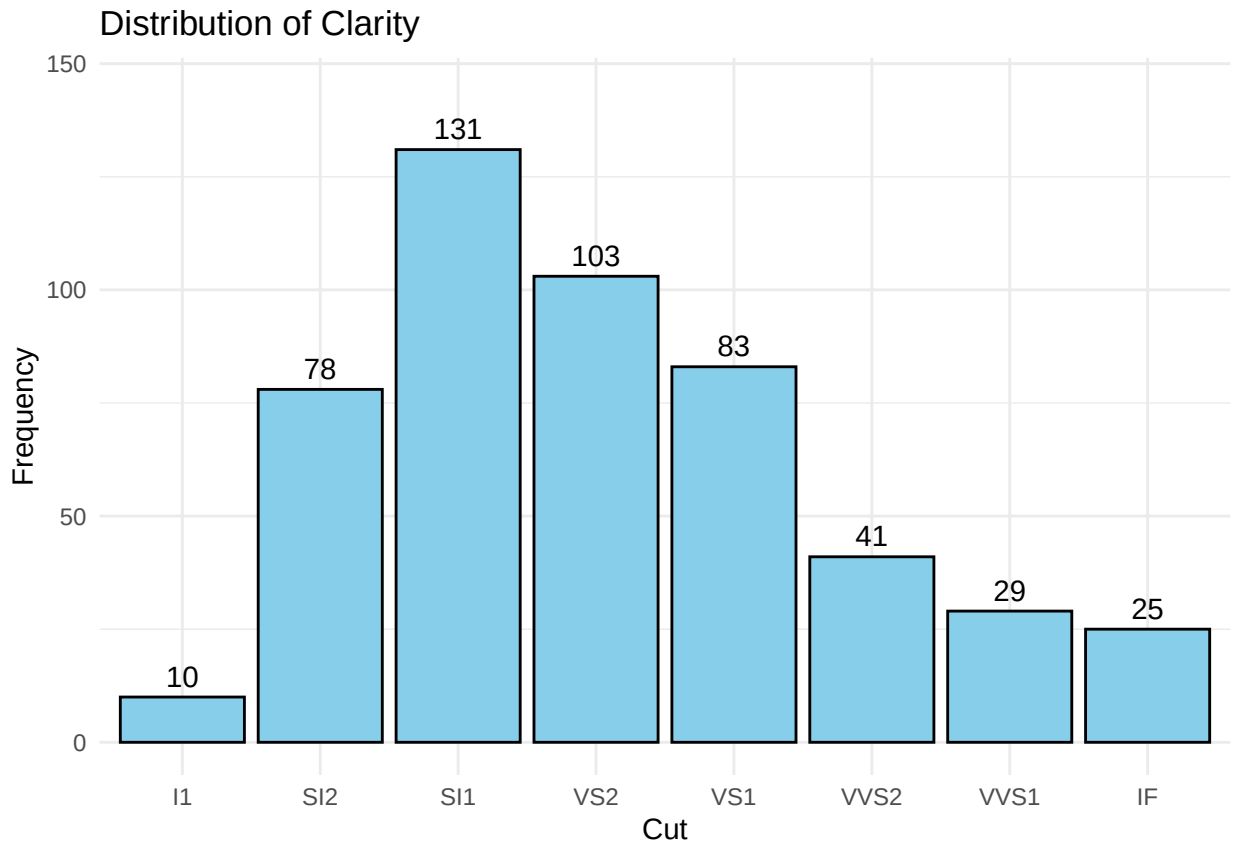
The clarity of diamonds is determined based on the presence or absence of internal qualities called "inclusions" and external qualities called "blemishes". The grading of clarity is determined, from best to worst, as follows:

1. IF (Internally Flawless) - Under 10x magnification, there are no inclusions visible, but there are blemishes. Diamonds of internally flawless clarity are considered of very high rarity and value.

*Note: There is technically a grade above IF, FL(Flawless), but there are no diamonds in this dataset, nor the sample, of this clarity

2. VVS1 (Very Very Slightly Included 1) - There are inclusions present, but they are extremely difficult to see, even under 10x magnification
3. VVS2 (Very Very Slightly Included 2) - There are inclusions present, but they are very difficult to see, even under 10x magnification, although slightly more visible than VVS1.
4. VS1 (Very Slightly Included 1) - Inclusions are present, but are difficult to see under 10x magnification.
5. VS2 (Very Slightly Included 2) - Inclusion are present and somewhat easily visible to see when viewed under 10x magnification. These inclusions are still not visible to the naked eye and are considered minor.
6. SI1 (Slightly Included 1) - Inclusions are present and visible when viewed under 10x magnification, but may still be non-visible to the naked eye.
7. SI2 (Slightly Included 2) - Inclusions are present and easily visible when viewed under 10x magnification and may be visible to the naked eye.
8. I1 (Included 1) - Inclusions are obviously present when viewed under 10x magnification. These inclusions may interfere with the diamonds transparency and brilliance.

*Note: There are two further grades below I1, I2 and I3, but there are no diamonds of this clarity in the larger dataset, nor the sample.



Examining the distribution of "Clarity", we can observe that the possible values of clarity for our 500 observation sample are roughly normally distributed with a slight right-skew. This tells us, in our random sample, the majority of diamonds have clarity values on the lower-to-medium end of the grading scale described above. There are very few diamonds with clarity values in the far extremities of the scale. Although amongst these extreme values, there are more diamonds with the highest clarity gradings than there are with the the lowest clarity gradings.

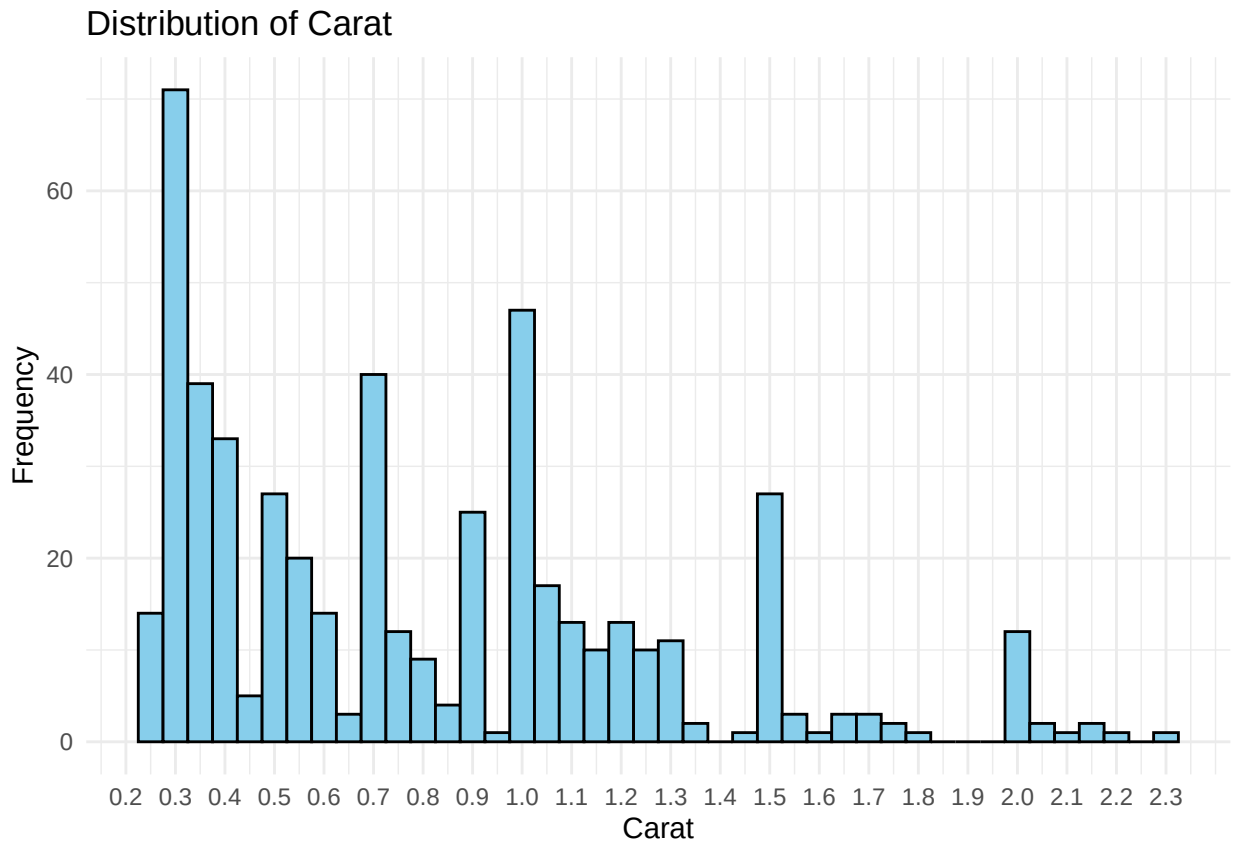
Quantitative/Numerical Variables:

X - The X variable is used to assign indexing numbers to each observation of diamond. Each observation of diamond, X, is contained in its own row.

- It is not meaningful to consider the individual distribution of X, as it is just a index variable.

Carat - Contains the weight of the diamond in carats. One carat is equal to 200 milligrams.

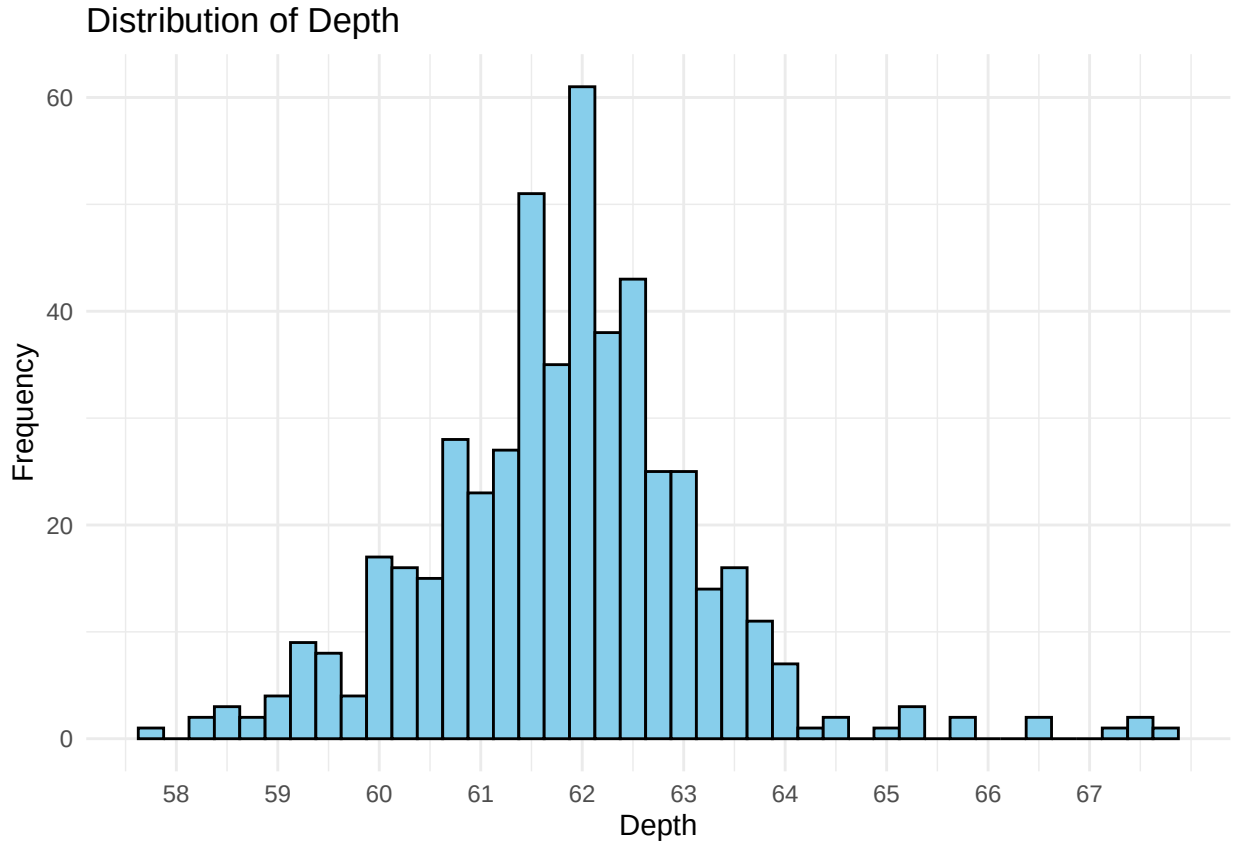
From the summary statistics displayed above, we can observe that the mean weight of diamonds in our 500 observation sample is 0.80 carats with a standard deviation of 0.23 carats. This mean is identical to the population mean of 0.80 carats. The population standard deviation is 0.47 carats, but this discrepancy is explainable due to more extreme values comparable to the mean being included in the population. The sample minimum is 0.23 carats, while the sample maximum is 2.30 carats. The population minimum is 0.20 carats and the population maximum is 5.01 carats.



Examining the distribution of "Carat" displayed above, we can observe that the possible values of carat are heavily right skewed. This corroborates the summary statistics for carat, as our mean of 0.80 indicates that the majority of the diamond's weights should be grouped on the left-hand side amongst the lighter weights. Additionally, we can observe that there a very few diamonds (compared to the rest of the sample), with weights greater than roughly 1.40 carats.

Depth - Contains the distance(in millimeters) from the bottom tip of the diamond to its flat top-surface. This distance is expressed as a percentage of the average diameter of the diamond at its widest point, called the girdle. The girdle is usually found by taking the average of the length, x, and the width, y. Depth is a key factor in a diamond's cut and its ability to refract light.

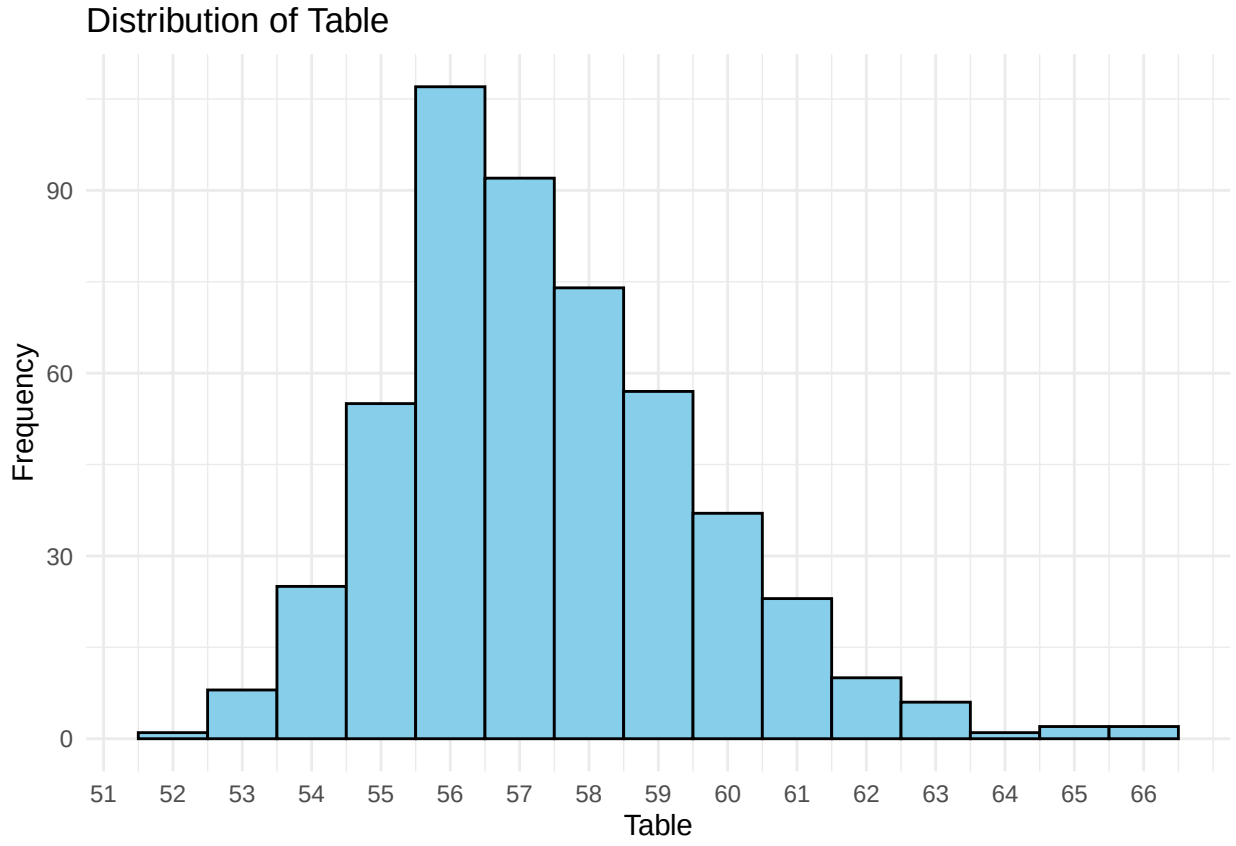
From the summary statistics displayed above, we can observe that the sample mean for depth is 61.83 with a standard deviation of 1.36. These sample statistics are extremely similar to the population mean of 61.75 with a standard deviation of 1.43. The sample minimum and maximum are 57.80 and 67.70, respectively, while the population minimum and maximum are 43.00 and 79.00.



Examining the distribution of depth displayed above, we can observe that the values of depth for the diamonds in our 500 observation sample are normally distributed with a very slight right skew. This corroborates the sample summary statistics, as the observations are fairly tightly grouped around the sample mean, 61.83, with less observations present the farther we move left or right from the sample mean.

Table - A diamond's table refers to whe width of the largest facet(flat surface) at the top of the diamond. This variable contains the table expressed as a percentage of the diamonds average diameter at its widest point, called the girdle. The girdle is calculated in the same manner as it is for depth.

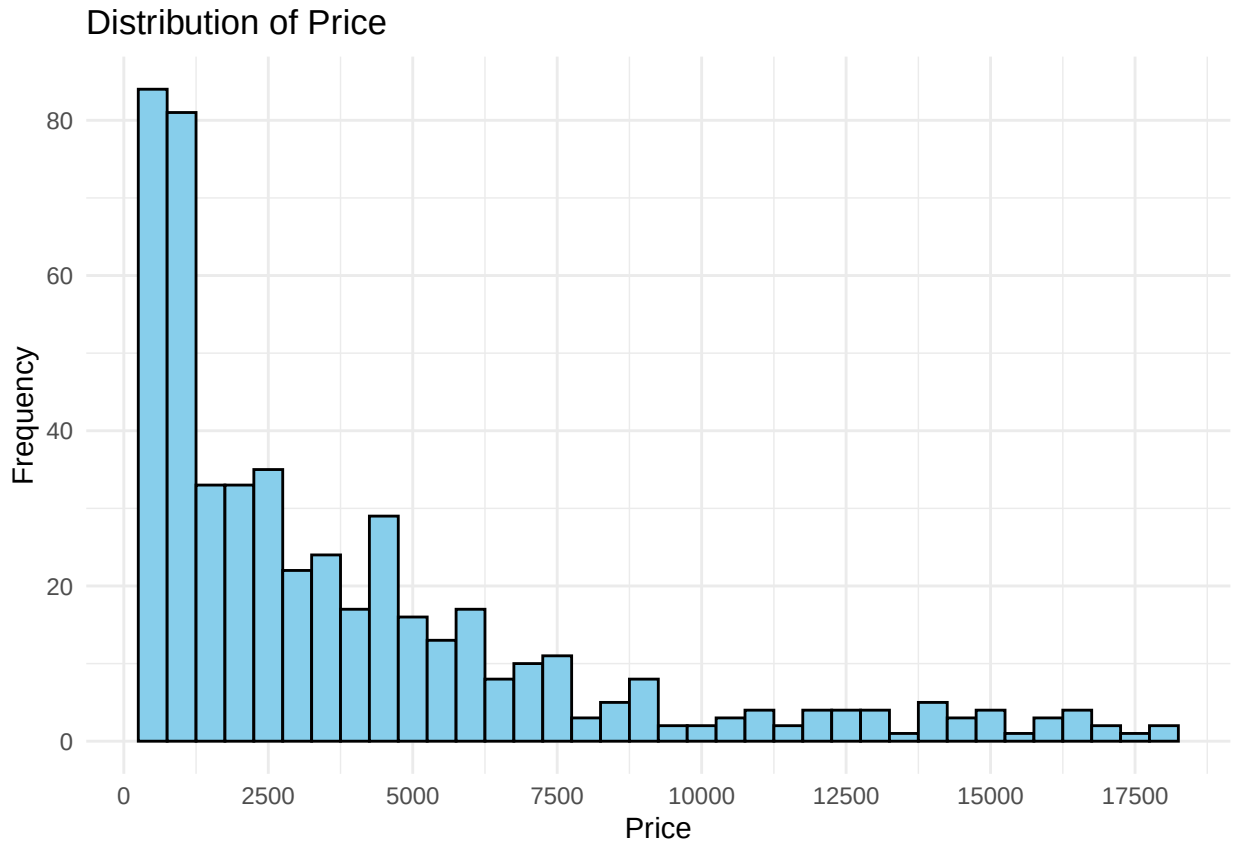
From the summary statistics displayed above, we can observe that the mean table value of diamonds in our 500 observation sample is 57.38 with a standard deviation of 1.36. These sample statistics are mostly similar to the population mean and standard deviation, which are 57.46 and 2.23 respectively. The discrepancy in the standard deviation can be explained by the difference between our sample minimum and maximum, 52.00 and 66.00 respectively, and our population minimum and maximum, 43.00 and 95.00.



Examining the distribution of table displayed above, we can observe that the values of table for the diamonds in our 500 observation sample are approximately normally distributed with a slight right-skew. This corroborates the values we observed in the sample statistics, as the table values are tightly grouped about the sample mean, 57.38, with less values/less density the farther we move up or down from the mean.

Price - The price of the diamond in dollars($).

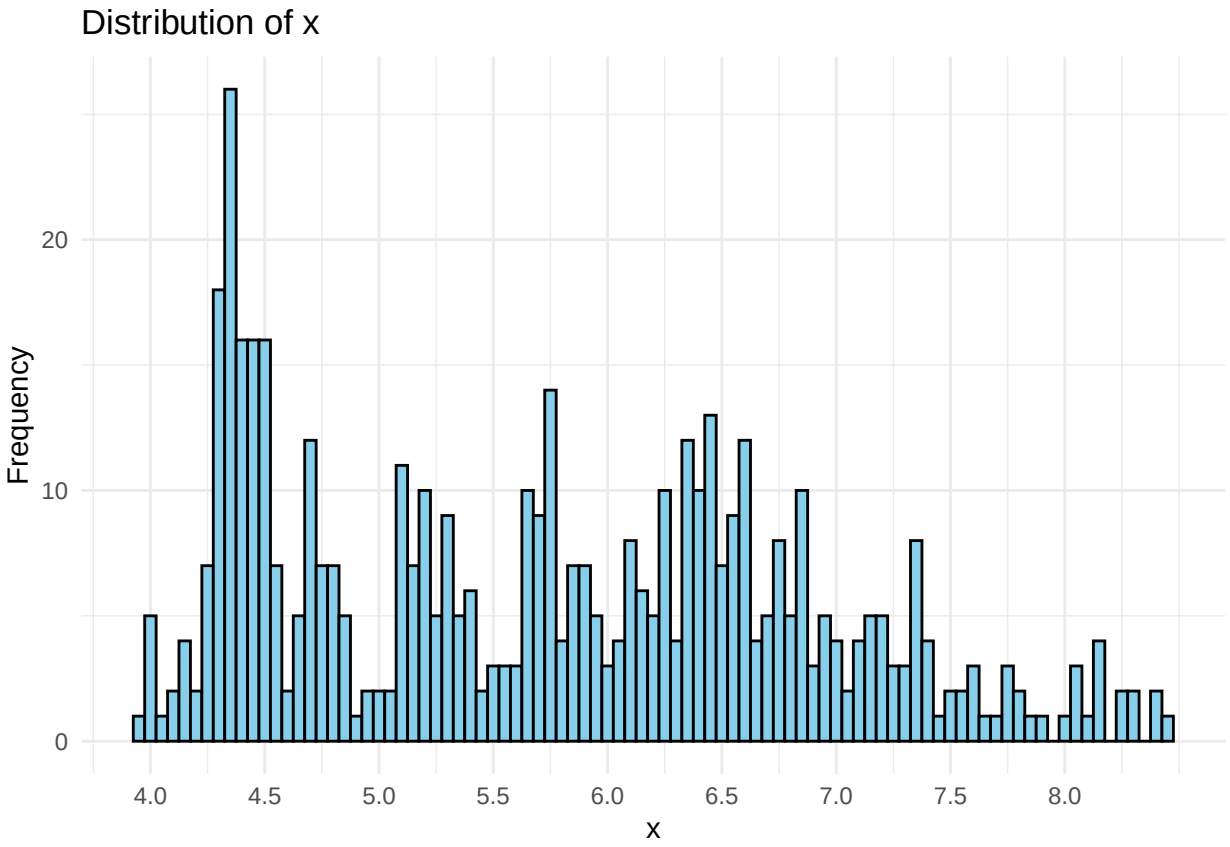
From the summary statistics displayed above, we can observe that the mean price of diamonds in our 500 observation sample is $3922.04 with a standard deviation of $3970.49. These sample statistics are extremely similar to the population mean and standard deviation, which are $3932.73 and $3989.34 respectively.



Examining the distribution of price displayed above, we can observe that the values of price of the diamonds in our 500 observation sample are heavily right skewed. This aligns with the values of our summary statistics, as the vast majority of price values are less than the sample mean, $3922.04, but we have a fair amount of price values in the right-tail. This balance between most of the diamond prices being relatively small, with some extremely high-priced values, results in our sample mean.

x - Denotes the length of the diamond in millimeters.

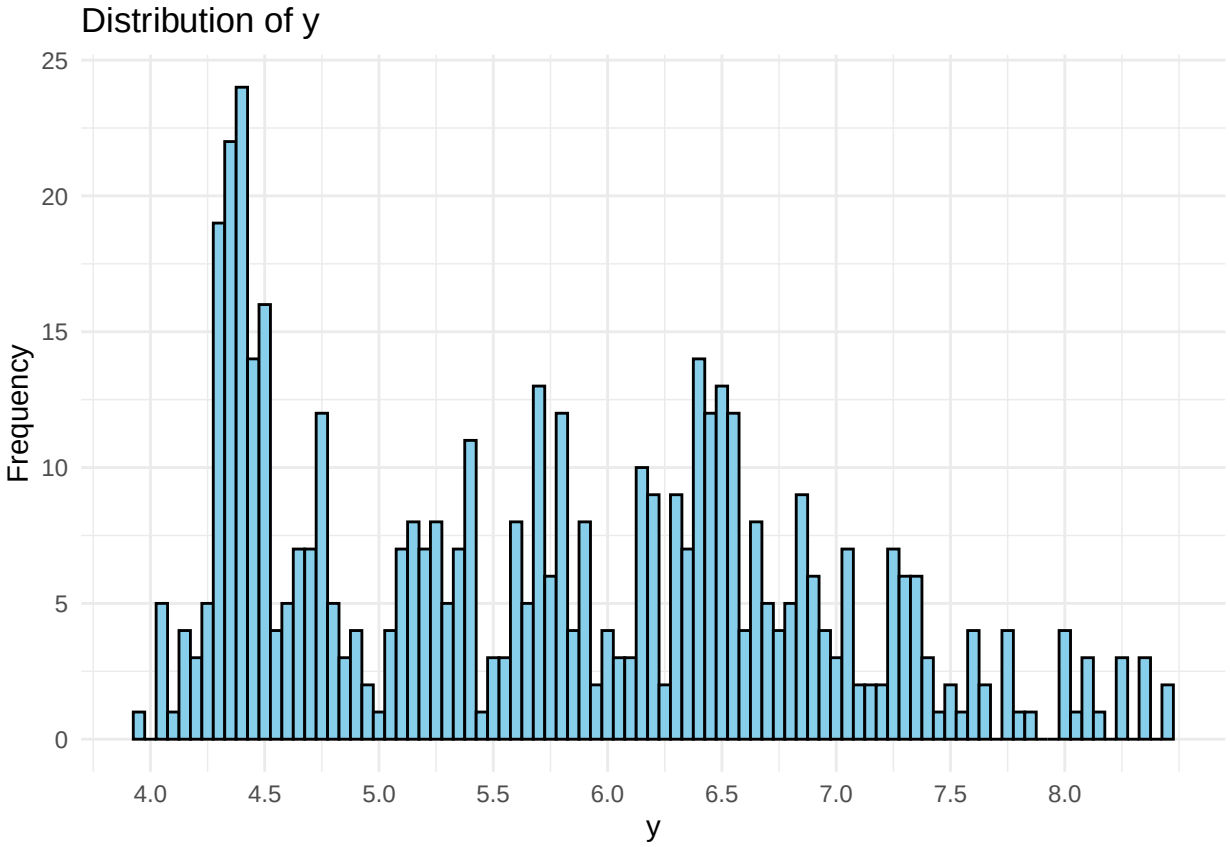
From the summary statistics, we find that the sample mean diamond length is 5.73mm with a standard deviation of 1.12mm. These sample statistics are identical to the population mean and standard deviation, 5.73mm and 1.12mm respectively.



We can observe that the values of x(length) for the diamonds in our 500 observation sample are skewed right. This tells us that many of the diamonds in our sample have smaller lengths comparatively to one another, rather than larger.

y - Denotes the width of the of the diamond in millimeters.

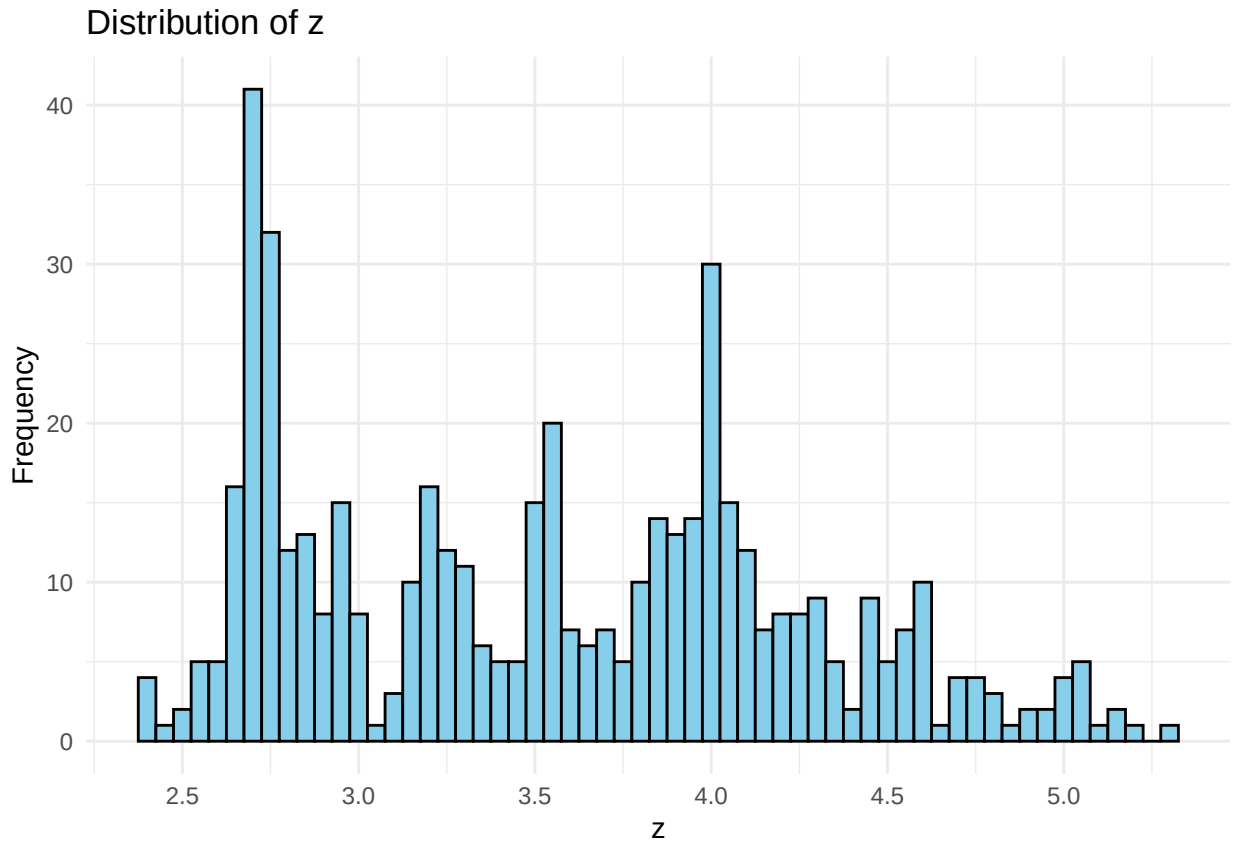
From the summary statistics, we find the that sample mean diamond length is 5.73mm with a standard deviation of 1.11mm. This is extremely similar to the population mean and standard deviation of 5.73mm and 1.14mm respectively.



We can observe that the values of y(width) for the diamonds in our 500 observation sample are skewed right. This tells us that many of the diamonds in our sample have smaller widths comparatively to one another, rather than larger.

z - Denotes the depth of the diamond in millimeters.

From the summary statistics, we find that the sample mean diamond depth for the diamonds in our 500 observation sample is 3.54mm with a standard deviation of 0.69mm. This is extremely similar to the population mean and standard deviation, 3.54mm and 0.71mm, respectively.



We can observe that the values of z(depth) are slightly right skewed with denser distributions of values around 2.75mm and 4.00mm. This implies that the depth values of the diamonds in our sample are relatively equally distributed with many depths being around 2.75mm and 4.00mm as well as between these values, but few greater than 4.00mm.

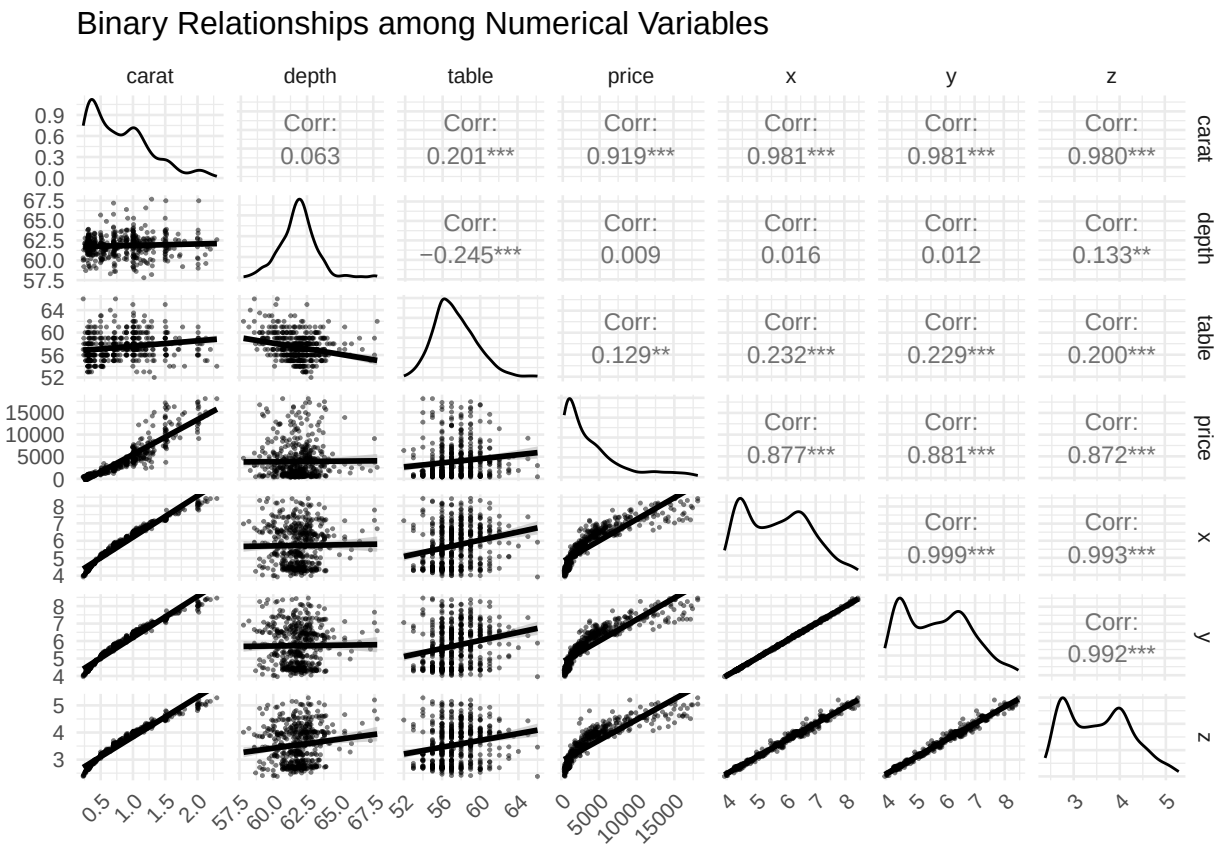
Notes on Summary Statistics

We can observe that the sample from the diamonds dataset of size 500, "diamonds_sample", appears to be representative sample of the population. Note that the values of the sample statistics for "diamonds_sample" are comparatively similar to the population parameters for the original "diamonds" dataset.

*There are discrepancies in the percentiles of some of the quantitative variables in the sample data versus the population data, as well as the standard deviation values for some variables- but this is explainable by noting that the sample has failed to include some of the more extreme values in both the low and high directions, so the percentiles have been skewed as a result. This is not any cause for concern, as these statistics will not be relevant to our regression analysis.

Binary Comparisons

```
numvars<-subset(diamonds_sample,select=c(carat,depth,table,price,x,y,z))
ggpairs(numvars,
        upper = list(continuous = wrap("cor", size = 3)),
        lower = list(continuous = wrap("smooth", alpha = 0.5, size = 0.2)),
        diag = list(continuous = wrap("densityDiag", alpha = 0.5)),
        title = "Binary Relationships among Numerical Variables") +
  theme_minimal(base_size = 10) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Notes on Binary Comparisons

Our current end-goal is to conduct analysis on the relationship between price, as the independent variable, and a subset of the other variables in the diamonds dataset as predictors.

Based on these relationships, we will decide which variables to proceed with to construct a model.

Binary Comparisons: Price vs. Quantitative Variables

Strong Positive Correlation

We can observe that there appears to be a strong positive linear relationship between price and the following variables:

- carat (correlation: +0.919)
- x/length (correlation: +0.877)

- y/width (correlation: +0.881)
- z/depth (correlation: +0.872).

Weak Positive Correlation

We can observe that there appears to be a weak or very weak positive linear relationship between price and the following variables:

- table (correlation: +0.129)
- depth (correlation: +0.009)

Binary Comparisons: Carat vs. Quantitative Variables

We can observe the following relationships between Carat and:

- Depth (correlation: +0.063)
- Table (correlation: +0.201)
- x/length (correlation: +0.981)
- y/width (correlation: +0.980)
- z/depth (correlation: +0.872).

Binary Comparisons: Depth vs. Quantitative Variables

We can observe the following relationships between Depth and:

- Carat (correlation: +0.063)
- Table (correlation: -0.245)
- x/length (correlation: +0.016)
- y/width (correlation: +0.012)
- z/depth (correlation: +0.133).

Binary Comparisons: Table vs. Quantitative Variables

We can observe the following relationships between Table and:

- Carat (correlation: +0.201)
- Depth (correlation: -0.245)
- x/length (correlation: +0.232)
- y/width (correlation: +0.229)
- z/depth (correlation: +0.200.

Binary Comparisons: x/length vs. Quantitative Variables

We can observe the following relationships between x/length and:

- Carat (correlation: +0.981)
- Depth (correlation: -0.016)
- Table (correlation: +0.232)
- y/width (correlation: +0.999)
- z/depth (correlation: +0.993).

Binary Comparisons: y/width vs. Quantitative Variables

We can observe the following relationships between y/width and:

- Carat (correlation: +0.981)
- Depth (correlation: -0.012)
- Table (correlation: +0.229)
- x/length (correlation: +0.999)
- z/depth (correlation: +0.992)

Binary Comparisons: z/depth vs. Quantitative Variables

We can observe the following relationships between z/depth and:

- Carat (correlation: +0.980)
- Depth (correlation: 0.133)
- Table (correlation: 0.200)
- x/length (correlation: 0.993)
- y/width (correlation: +0.992)

Conclusions on Binary Comparisons

We can observe that the relationship between price and depth is not significant, so including depth in the model would not be meaningful(over-fitting).

This leaves us with, Price vs. carat, table, x/length, y/width, and z/depth.

Amongst these variables, we can observe a significant linear relationship between x, y, z, and carat. It follows that only one of these variables should be included in our model to prevent multi-colinearity. Carat has the highest correlation with price, so this is the variable we will include.

Of the quantitative variable candidates for predictors in our regression model listed above, we are left with price vs. carat and table.

Approaching our first MLR model

It is not feasible to conduct binary comparisons between price and and the categorical variables in the same way as the quantitative variables. To examine the value of the categorical variables in terms of our regression model, it is useful to try running a multiple linear regression model with the various categorical variables included vs not included and examining the results.

Note: The index variable, X, will not be included in any regression modeling, as it is only referencing the original position of the diamond observation in the larger dataset, which is not meaningful for our purposes.

Let us begin with all categorical variables included. Let y denote price, x_1 denote carat, x_2 denote table, x_3 denote cut, x_4 denote clarity, and x_5 denote color. Then the beginning MLR model is:

$\mathbb{E}[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$

```
MLR1 <- lm(formula = price ~ carat + table + cut + clarity + color, data = diamonds_sample)
summary(MLR1)
```

```
##
## Call:
## lm(formula = price ~ carat + table + cut + clarity + color, data = diamonds_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2594.5  -643.6  -196.7   391.1  5340.5
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3959.91    1787.32  -2.216 0.027190 *
## carat         9137.55     130.00  70.291  < 2e-16 ***
## table          -63.06      29.03  -2.172 0.030339 *
## cutGood        747.16     355.53   2.102 0.036114 *
## cutIdeal      1044.05     329.65   3.167 0.001638 **
## cutPremium     774.29     321.46   2.409 0.016387 *
## cutVery Good  1179.76     325.00   3.630 0.000314 ***
## clarityIF     4866.44     462.06  10.532  < 2e-16 ***
## claritySI1    3473.40     398.95   8.706  < 2e-16 ***
## claritySI2    2466.34     406.53   6.067 2.65e-09 ***
## clarityVS1    4571.16     411.19  11.117  < 2e-16 ***
## clarityVS2    4480.66     407.41  10.998  < 2e-16 ***
## clarityVVS1   5041.11     450.88  11.181  < 2e-16 ***
## clarityVVS2   4939.65     432.79  11.413  < 2e-16 ***
## colorE        -265.00     201.85  -1.313 0.189854
## colorF        -287.41     198.72  -1.446 0.148743
## colorG        -587.24     198.50  -2.958 0.003246 **
## colorH       -1111.07     205.46  -5.408 1.01e-07 ***
## colorI       -1552.68     244.73  -6.344 5.17e-10 ***
## colorJ       -2106.72     281.19  -7.492 3.28e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1151 on 480 degrees of freedom
## Multiple R-squared:  0.9192, Adjusted R-squared:  0.916
## F-statistic: 287.5 on 19 and 480 DF,  p-value: < 2.2e-16
```

Running the initial multiple linear regression model described above, we observe that this initial model has an R^2_{adj} of 0.916. We will now try eliminating the categorical variables one by one to see which are meaningful to the model.

Model 2:

We will begin by attempting a model without x_5, color. The proposed second model becomes:

$\mathbb{E}[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

```
MLR2 <- lm(formula = price ~ carat + table + cut + clarity, data = diamonds_sample)
summary(MLR2)
```

```
##
## Call:
## lm(formula = price ~ carat + table + cut + clarity, data = diamonds_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3151.0  -676.0   -75.1   433.9  6321.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5192.25    1921.32  -2.702 0.007124 **
## carat         8761.07     136.10  64.373  < 2e-16 ***
## table          -42.81      31.41  -1.363 0.173602
```

```
## cutGood        976.77     387.05   2.524 0.011933 *
## cutIdeal      1112.20     357.92   3.107 0.001998 **
## cutPremium     737.80     350.52   2.105 0.035816 *
## cutVery Good  1233.60     353.66   3.488 0.000531 ***
## clarityIF     4334.37     503.36   8.611  < 2e-16 ***
## claritySI1    3154.64     435.30   7.247 1.68e-12 ***
## claritySI2    2266.50     444.49   5.099 4.90e-07 ***
## clarityVS1    4183.24     447.98   9.338  < 2e-16 ***
## clarityVS2    4088.09     443.13   9.226  < 2e-16 ***
## clarityVVS1   4607.35     491.30   9.378  < 2e-16 ***
## clarityVVS2   4564.07     470.42   9.702  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1264 on 486 degrees of freedom
## Multiple R-squared:  0.9013, Adjusted R-squared:  0.8986
## F-statistic: 341.3 on 13 and 486 DF,  p-value: < 2.2e-16
```

Running the second multiple linear regression model, we find that model 2 has an R^2_{adj} of 0.8986. This means that this model explains slightly less of the total variance in price, so Model 2 is of worse fit than Model 1. Therefore, we will continue to include x_5, color, in subsequent models.

Model 3

We will proceed by attempting a model without x_4, clarity. The proposed second model becomes:

$\mathbb{E}[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5$

```
MLR3 <- lm(formula = price ~ carat + table + cut + color, data = diamonds_sample)
summary(MLR3)
```

```
##
## Call:
## lm(formula = price ~ carat + table + cut + color, data = diamonds_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4762.1  -752.6   -62.7   650.2  6503.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    441.21    2216.12   0.199 0.842274
## carat         8299.64     152.84  54.301  < 2e-16 ***
## table          -75.31      36.82  -2.046 0.041326 *
## cutGood       1492.97     434.74   3.434 0.000645 ***
## cutIdeal      1927.57     404.83   4.761 2.54e-06 ***
## cutPremium    1604.38     392.71   4.085 5.15e-05 ***
## cutVery Good  2051.39     398.62   5.146 3.86e-07 ***
## colorE        -512.63     254.79  -2.012 0.044771 *
## colorF        -259.98     249.73  -1.041 0.298377
## colorG        -451.76     248.64  -1.817 0.069850 .
## colorH       -1014.14     259.46  -3.909 0.000106 ***
## colorI       -1181.04     308.75  -3.825 0.000148 ***
## colorJ       -1680.69     355.60  -4.726 3.00e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1466 on 487 degrees of freedom
## Multiple R-squared:  0.867,  Adjusted R-squared:  0.8637
## F-statistic: 264.6 on 12 and 487 DF,  p-value: < 2.2e-16
```

Running the third multiple linear regression model, we find that model 3 has an R^2_{adj} of 0.8637. This means that this model explains less of the total variance in price than Models 1 and 2, so Model 3 is of worse fit than both Models 1 and 2. Therefore, we will continue to include x_5, color, as well as x_4, clarity in subsequent models.

Model 4

We will proceed by attempting a model without x_3, cut. The proposed second model becomes:

$\mathbb{E}[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5$

```
MLR4 <- lm(formula = price ~ carat + table + color + clarity, data = diamonds_sample)
summary(MLR4)
```

```
##
## Call:
## lm(formula = price ~ carat + table + color + clarity, data = diamonds_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2938.7  -640.9  -190.9   439.7  5565.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1617.09    1501.27  -1.077  0.28195
## carat        9099.75     130.72  69.612  < 2e-16 ***
## table         -91.77      24.80  -3.701  0.00024 ***
## colorE       -227.74     203.83  -1.117  0.26443
## colorF       -311.97     200.94  -1.553  0.12118
## colorG       -598.64     200.00  -2.993  0.00290 **
## colorH      -1168.60     206.29  -5.665 2.53e-08 ***
## colorI      -1516.26     246.94  -6.140 1.72e-09 ***
## colorJ      -2112.27     283.45  -7.452 4.26e-13 ***
## clarityIF    5194.05     453.75  11.447  < 2e-16 ***
## claritySI1   3760.10     388.44   9.680  < 2e-16 ***
## claritySI2   2751.61     394.94   6.967 1.06e-11 ***
## clarityVS1   4860.13     401.00  12.120  < 2e-16 ***
## clarityVS2   4785.49     396.21  12.078  < 2e-16 ***
## clarityVVS1  5358.68     440.86  12.155  < 2e-16 ***
## clarityVVS2  5264.19     424.76  12.393  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1167 on 484 degrees of freedom
## Multiple R-squared:  0.9162, Adjusted R-squared:  0.9137
## F-statistic:   353 on 15 and 484 DF,  p-value: < 2.2e-16
```

Running the fourth multiple linear regression model, we find that model 4 has an R^2_{adj} of 0.9137. This means that this model explains less of the total variance in price than Model 1, so Model 4 is of worse fit than Model 1. Therefore, we will continue to include x_5, color, x_4, clarity, and x_3, cut, in subsequent models.

Conclusions on our first MLR model

In essence, we will proceed with Model 1, which includes all categorical variables, in addition to the quantitative variables carat and table. The model is as follows:

$\mathbb{E}[\mathbf{y}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$

β_0 - Intercept term.

$\beta_1 x_1$

- β_1 - Carat coefficient, denotes the expected dollar amount change in price per one unit increase in carat, when all other predictors are held constant.
- x_1 - Carat.

$\beta_2 x_2$

- β_2 - Table coefficient. Denotes the expected dollar amount change in price per one unit increase in table, when all other predictors are held constant.
- x_2 - Table.

Note: When categorical variables are included in a multiple linear regression model, they are encoded with regard to dummy variables to indicate which value of the categorical variable is present. A reference value for each categorical variable is chosen, then the other possible values of the categorical variable are referred to in terms of the reference value. As a result, there are different values of each $\beta_{i,k}$ for each possible value, k, of the corresponding x_i. The $\beta_{i,k}$ refers to the change in price when compared to the reference variable's change in price.

$\beta_3 x_3$

- β_3 - Cut coefficient, denotes the expected change in diamond price, when all other predictors are held constant, when the cut takes on a certain value, in comparison to the reference value, cut=Fair.
- x_3 - Cut, possible values are Fair(reference), Good, Very Good, Ideal, and Premium

$\beta_4 x_4$

- β_4 - Clarity coefficient, denotes the expected change in diamond price, when all other predictors are held constant, when clarity takes on a certain value, in comparison to the reference value, clarity=I1.
- x_4 - Clarity, possible values are I1(reference), SI1, SI2, VS1, VS2, VVS1, and VVS2.

$\beta_5 x_5$

- β_5 - Color coefficient, denotes the expected change in diamond price, when all other predictors are held constant, when color takes on a certain value, in comparison to the reference value, color=D.
- x_5 - Color, possible values are D(reference), E, F, G, H, I, and J.

==========

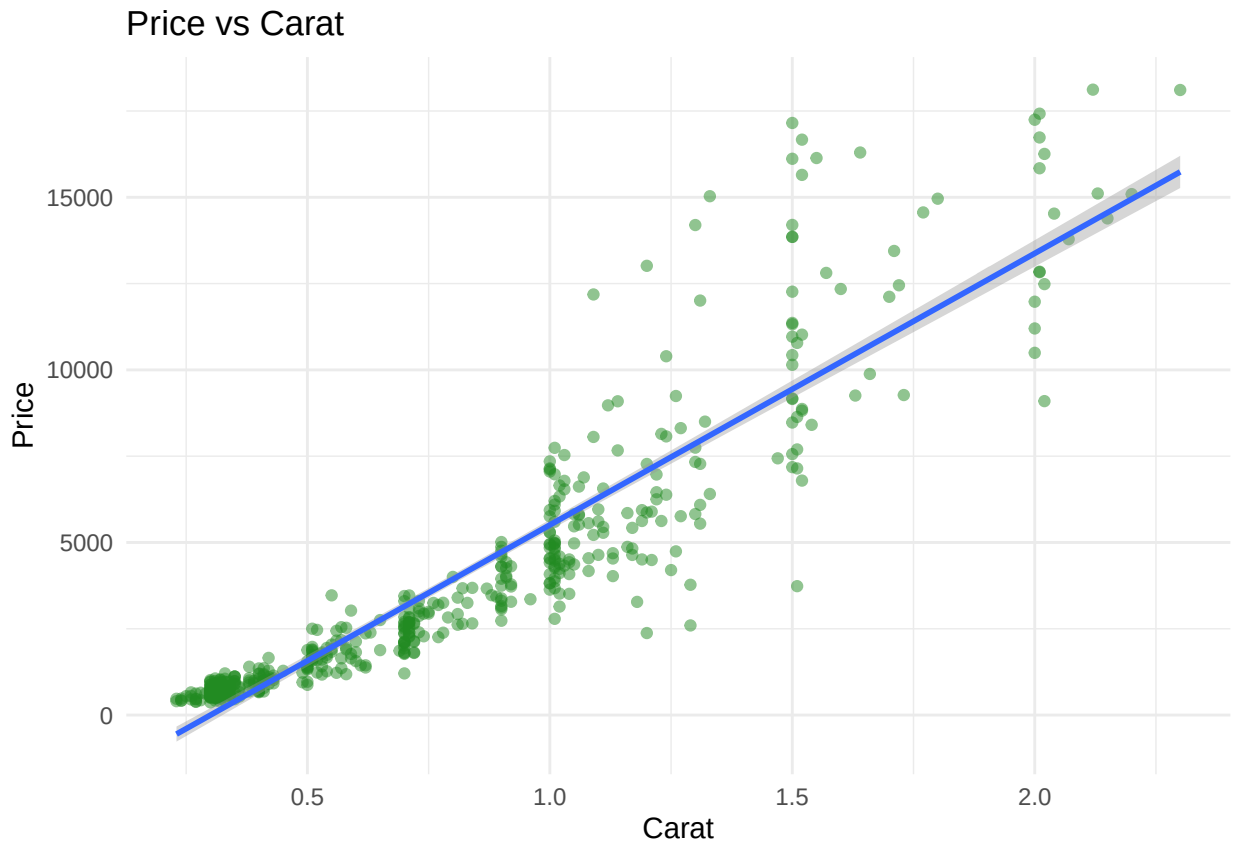
PART 2

Hypothesis:

There is a linear relationship between the dependent variable of price and the predictor carat. We believe that there is a positive correlation between the variables.

Simple Linear Regression price by carat (1):

```
simple_reg<- ggplot(diamonds_sample, aes(x = carat, y = price)) +
  geom_point(alpha = 0.5, color = "forestgreen") +
  theme_minimal() +
  labs(title = "Price vs Carat", x = "Carat", y = "Price")
simple_reg + geom_smooth(method = "lm", formula = y~x)
```




```
model11 <- lm(formula= price~carat , diamonds_sample)
summary(model11)
```

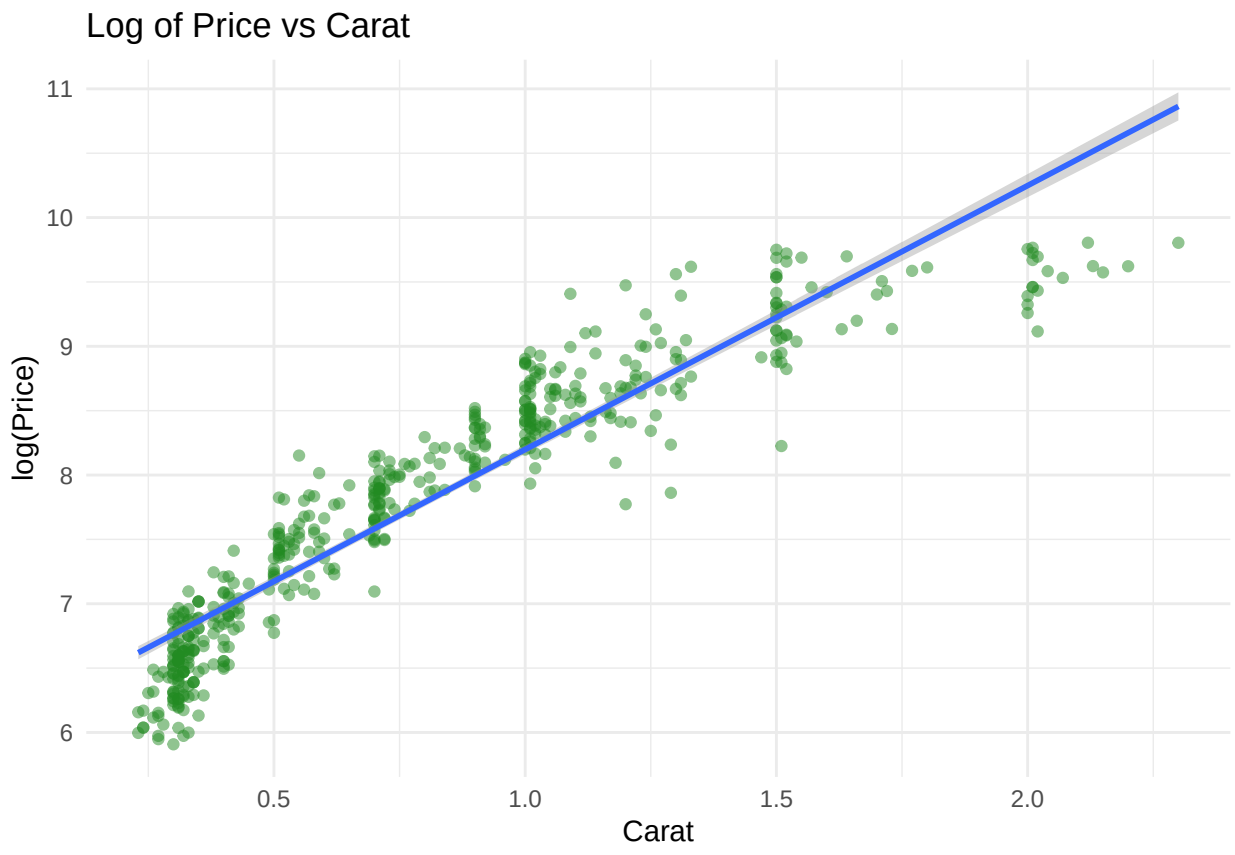
```
##
## Call:
## lm(formula = price ~ carat, data = diamonds_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5785.6  -850.9     1.8   619.4  7712.1
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2361.6      139.6  -16.92   <2e-16 ***
## carat         7868.4      151.2   52.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1566 on 498 degrees of freedom
## Multiple R-squared:  0.8447, Adjusted R-squared:  0.8443
## F-statistic:  2708 on 1 and 498 DF,  p-value: < 2.2e-16
```

The graph indicates a strong positive correlation between carat and price. The model predicts that for every unit increase in carat there is a price increases $7868.4 with small margin of error being 151.2. The p-value is very small being 2e-16 meaning if we were to test the null hypothesis that the beta value equals 0 vs the beta not equal to 0 at a very small significance level we would reject the null in favor of the alternative. This implies that carat is an essential predictor of price meaning we can't get rid of it. 84.47% of the variance in price is explained by the carat. However, the residuals do appear to violate the model assumption of constant variance as they increase as carat increases.

Simple Linear Regression log(price) by carat 2:

```
simple_reg<- ggplot(diamonds_sample, aes(x = carat, y = log(price))) +
  geom_point(alpha = 0.5, color = "forestgreen") +
  theme_minimal() +
  labs(title = "Log of Price vs Carat", x = "Carat", y = "log(Price)")
simple_reg + geom_smooth(method = "lm", formula = y~x)
```



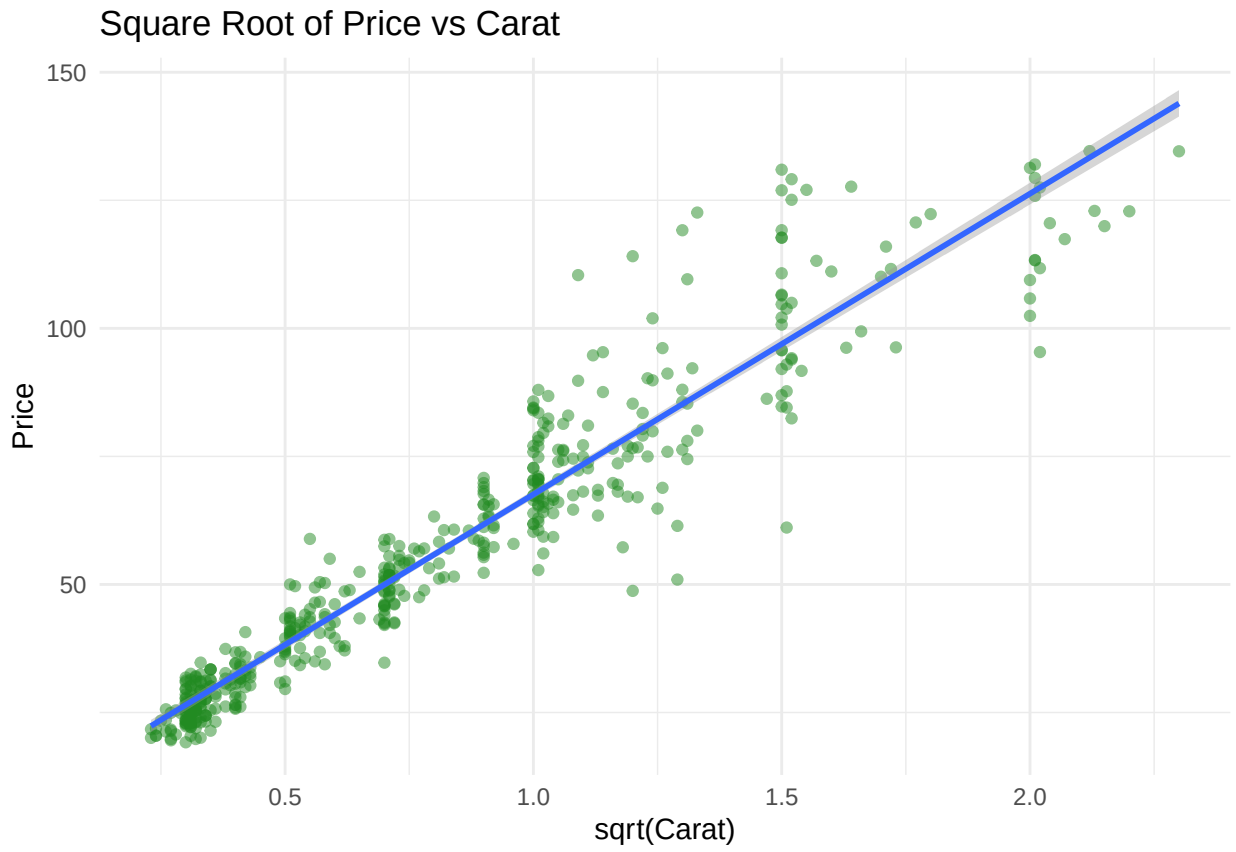
```
model22 <- lm(formula= log(price)~carat , diamonds_sample)
summary(model22)
```

```
##
## Call:
## lm(formula = log(price) ~ carat, data = diamonds_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17398 -0.21853  0.04734  0.26813  1.02514
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.14817    0.03273  187.82   <2e-16 ***
## carat        2.05015    0.03546   57.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3673 on 498 degrees of freedom
## Multiple R-squared:  0.8704, Adjusted R-squared:  0.8701
## F-statistic:  3344 on 1 and 498 DF,  p-value: < 2.2e-16
```

The graph indicates a strong positive correlation between carat and price. The p-value remains the same, 2e-16, meaning if we were to test the null hypothesis that the beta value equals 0 vs the beta not equal to 0 at a very small significance level we would reject the null in favor of the alternative. 87.01% of the variance in price is explained by the carat which implies that taking the log of price produces a better fit. The residuals on this model appear to deviate from the model at a more constant rate compared to the first model with no adjustments to price. Towards the tail of the regression line the residuals start to fall below the line.

Simple Linear Regression sqrt(price) by carat 3:

```
simple_reg<- ggplot(diamonds_sample, aes(x = carat, y = sqrt(price))) +
  geom_point(alpha = 0.5, color = "forestgreen") +
  theme_minimal() +
  labs(title = "Square Root of Price vs Carat", x = "sqrt(Carat)", y = "Price")
simple_reg + geom_smooth(method = "lm", formula = y~x)
```




```
model <- lm(formula= sqrt(price)~carat , diamonds_sample)
summary(model)
```

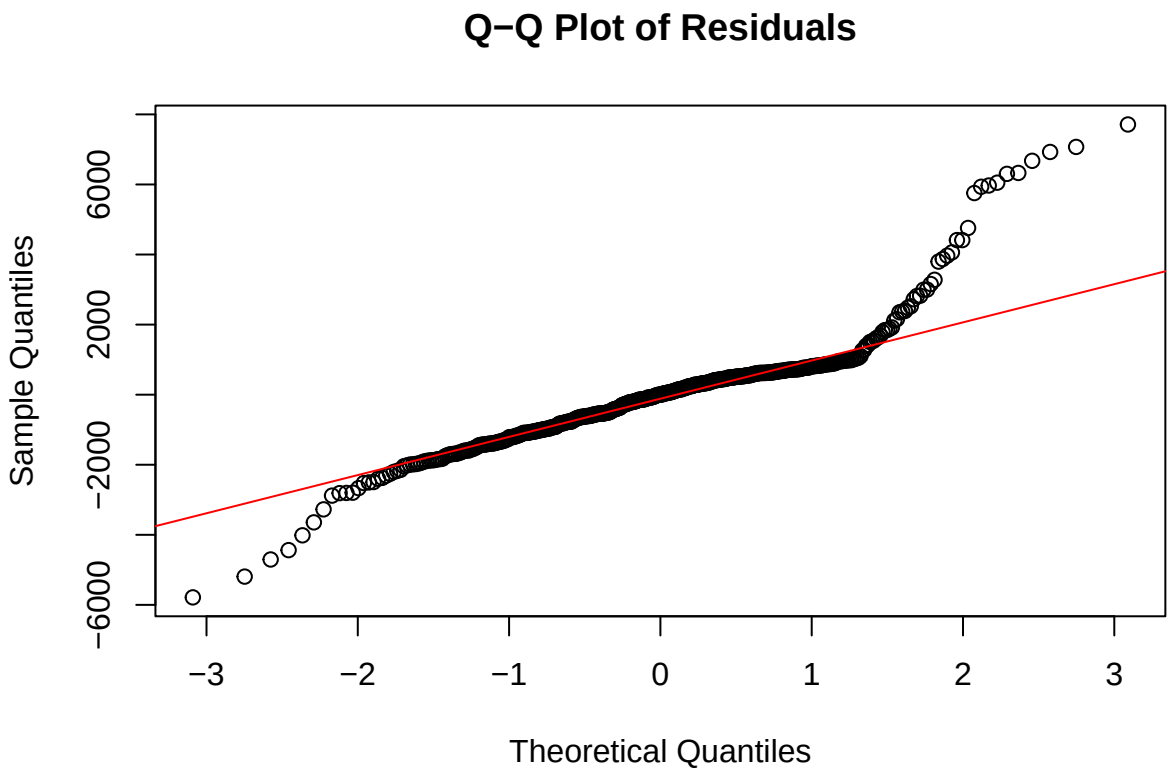
```
##
## Call:
## lm(formula = sqrt(price) ~ carat, data = diamonds_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.414  -4.316  -0.476   3.052  37.523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.8683     0.7662   11.57   <2e-16 ***
## carat        58.7103     0.8299   70.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.598 on 498 degrees of freedom
## Multiple R-squared:  0.9095, Adjusted R-squared:  0.9093
## F-statistic:  5005 on 1 and 498 DF,  p-value: < 2.2e-16
```

The graph indicates a strong positive correlation between carat and price. The p-value remains the same, 2e-16, meaning if we were to test the null hypothesis that the beta value equals 0 vs the beta not equal to 0 at a very small significance level we would reject the null in favor of the alternative. 90.93% of the variance

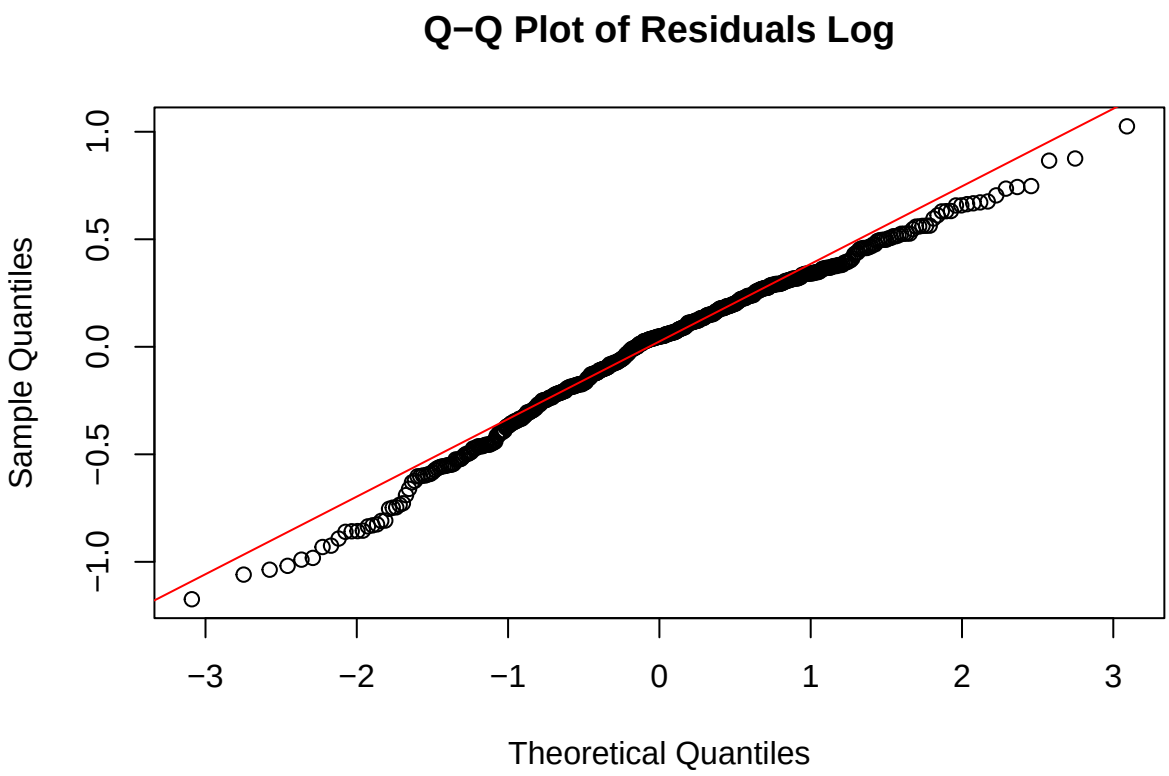
in price is explained by the carat which implies that taking the square root of price produces a better fit than both other models. The residuals on this model appear to deviate from the model in the same form as the first model with slightly less deviation, so overall better than the first

Q-Q plots for all models:

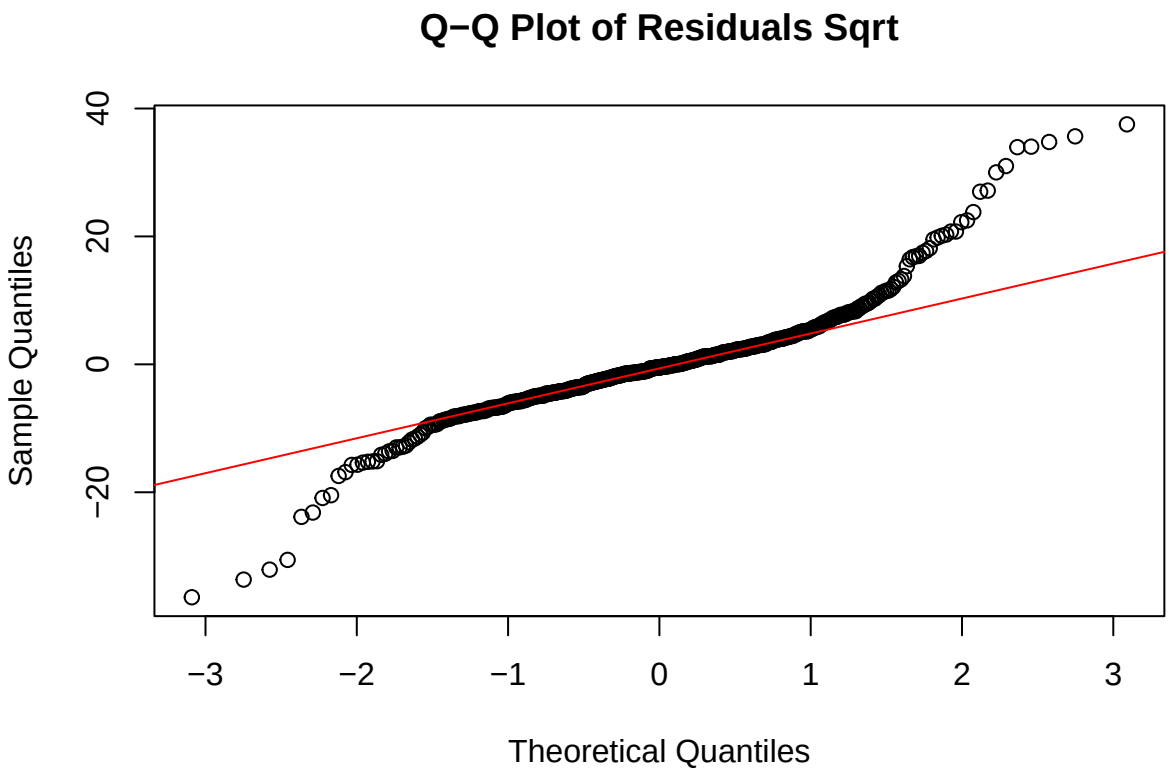
```
slr_none <- lm(price~carat, diamonds_sample)
nnone <- resid(slr_none)
qqnorm(nnone, main="Q-Q Plot of Residuals")
qqline(nnone, col="red")
```




```
slr_log <- lm(log(price)~carat, diamonds_sample)
nlog <- resid(slr_log)
qqnorm(nlog, main="Q-Q Plot of Residuals Log")
qqline(nlog, col="red")
```




```
slr_sqrt <- lm(sqrt(price)~carat, diamonds_sample)
nsqrt <- resid(slr_sqrt)
qqnorm(nsqrt, main="Q-Q Plot of Residuals Sqrt")
qqline(nsqrt, col="red")
```



The Q-Q plots depict how well the residuals follow the model assumption of normality and constant variance. Both the models with no change to price and taking the square root of the price have similar Q-Q plots with the residuals close to the mean following the model, but the outer tails begin to deviate more. The Q-Q plot with the log of price shows a more constant line of residuals throughout.

Conclusion:

After conducting simple linear regression on price and carat using no adjustment, log of price, and square root of price, we have concluded that using the log of price allows for the best fitting model. We decided this based on the fact that it fits the model assumption of constant variance better than the other methods. The R squared on the square root of price was the highest, but just by a small margin.

Examining MLR with log(price):

Similar to part 1, we shall now examine how adding different predictors to the MLR of log(price) influences R^2_{adj}.

Our base model of regression of log(price) as the independent vs. only carat as the sole predictor yields an $R^2_{adj} = 0.8701$

We acquire the following R^2_{adj} by including the following variables:

log(price) ~ carat + table :
$R^2_{adj} = 0.8704$

log(price) ~ carat + table + clarity : $R^2_{adj} = 0.8857$

log(price) ~ carat + table + clarity + cut :
$R^2_{adj} = 0.886$

log(price) ~ carat + table + clarity + cut + color : $R^2_{adj} = 0.9076$

log(price) ~ carat + table + clarity + cut + color + depth :
$R^2_{adj} = 0.9077$

log(price) ~ carat + table + clarity +cut + color + depth + x:
$R^2_{adj} = 0.9848$

log(price) ~ carat + table + clarity +cut + color + depth + x + y:
$R^2_{adj} = 0.985$

log(price) ~ carat + table + clarity +cut + color + depth + x + y + z:
$R^2_{adj} = 0.9851$

With each additional predictor added, we saw an increase in R^2_{adj}. This makes sense, as the more predictors that are included in the model typically results in an increase in R^2. Although, we should be cautious about the extremely high $R^2_{adj} = 0.9851$ that we have obtained.

We can note that certain variables, table, cut, depth, y, and z, did not improve the R^2_{adj} by very much (i.e. an improvement <0.0005). These predictors could be at risk of introducing overfitting to the model due to their negligible additions to the amount of variance explained in the data.

Additionally, we observed in Part 1 that many of our predictors were highly correlated with one another. As a result, when using the model above that includes all predictors, we should proceed with caution as this model may not be reliable due to multicolinearity.

The goal of part 2 is to produce the model with the highest R^2_{adj}, so we will proceed to part 3 with the model described above, which includes all of the original predictors.

==========

PART 3

We have now observed that the MLR model created using all possible predictors(not including X, the index variable) yields the highest possible R^2_{adj}. We are still concerned with whether or not this is the best model with respect to other factors besides R^2_{adj}. To continue zeroing in on the best possible model, we will proceed with with both backward elimination using AIC as well as step-wise regression using AIC.

```
aic_model <- lm(log(price)~.-X,diamonds_sample)
backward<- step(aic_model, direction="backward")
```

```
## Start:  AIC=-2059.6
## log(price) ~ (X + carat + cut + color + clarity + depth + table +
##     x + y + z) - X
##
##           Df Sum of Sq     RSS     AIC
## - y        1    0.0078  7.3924 -2061.1
## - depth    1    0.0121  7.3968 -2060.8
## <none>                  7.3846 -2059.6
## - z        1    0.0623  7.4469 -2057.4
## - table    1    0.0955  7.4801 -2055.2
## - cut      4    0.3347  7.7193 -2045.4
## - x        1    0.2827  7.6673 -2042.8
## - carat    1    3.2543 10.6390 -1879.0
## - color    6    9.1766 16.5613 -1667.8
## - clarity  7   17.0052 24.3898 -1476.2
##
## Step:  AIC=-2061.08
```

```
## log(price) ~ carat + cut + color + clarity + depth + table +
##     x + z
##
##           Df Sum of Sq     RSS     AIC
## - depth    1    0.0047  7.3972 -2062.8
## <none>                  7.3924 -2061.1
## - table    1    0.0951  7.4875 -2056.7
## - z        1    0.1730  7.5654 -2051.5
## - cut      4    0.3908  7.7832 -2043.3
## - x        1    0.2987  7.6911 -2043.3
## - carat    1    3.2532 10.6457 -1880.7
## - color    6    9.1774 16.5699 -1669.5
## - clarity  7   17.2383 24.6308 -1473.3
##
## Step:  AIC=-2062.76
## log(price) ~ carat + cut + color + clarity + table + x + z
##
##           Df Sum of Sq     RSS     AIC
## <none>                  7.3972 -2062.8
## - table    1    0.0904  7.4876 -2058.7
## - cut      4    0.4016  7.7987 -2044.3
## - x        1    2.3920  9.7891 -1924.7
## - z        1    2.5299  9.9270 -1917.7
## - carat    1    3.2959 10.6931 -1880.5
## - color    6    9.1986 16.5957 -1670.7
## - clarity  7   17.7297 25.1269 -1465.3
```

```
summary(backward)
```

```
##
## Call:
## lm(formula = log(price) ~ carat + cut + color + clarity + table +
##     x + z, data = diamonds_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37845 -0.08277 -0.00156  0.08560  0.37921
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.652054   0.237467  -2.746 0.006263 **
## carat        -0.961345   0.065873 -14.594  < 2e-16 ***
## cutGood       0.124323   0.039780   3.125 0.001885 **
## cutIdeal      0.186822   0.040829   4.576 6.06e-06 ***
## cutPremium    0.137005   0.039753   3.446 0.000618 ***
## cutVery Good  0.163807   0.038172   4.291 2.15e-05 ***
## colorE       -0.051031   0.021895  -2.331 0.020183 *
## colorF       -0.112154   0.021535  -5.208 2.84e-07 ***
## colorG       -0.143904   0.021591  -6.665 7.31e-11 ***
## colorH       -0.276855   0.022307 -12.411  < 2e-16 ***
## colorI       -0.404797   0.026539 -15.253  < 2e-16 ***
## colorJ       -0.556602   0.030584 -18.199  < 2e-16 ***
## clarityIF     1.175455   0.050096  23.464  < 2e-16 ***
## claritySI1    0.725152   0.043135  16.811  < 2e-16 ***
```

```
## claritySI2    0.541680   0.043956  12.323  < 2e-16 ***
## clarityVS1    0.931561   0.044555  20.908  < 2e-16 ***
## clarityVS2    0.862750   0.044072  19.576  < 2e-16 ***
## clarityVVS1   1.139818   0.048904  23.307  < 2e-16 ***
## clarityVVS2   1.030875   0.046863  21.997  < 2e-16 ***
## table         0.008350   0.003455   2.417 0.016024 *
## x             0.687591   0.055306  12.433  < 2e-16 ***
## z             1.123926   0.087903  12.786  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1244 on 478 degrees of freedom
## Multiple R-squared:  0.9857, Adjusted R-squared:  0.9851
## F-statistic:  1572 on 21 and 478 DF,  p-value: < 2.2e-16
```

```
start_model <- lm(log(price)~1, diamonds_sample)
forward <- step(start_model, direction = "forward", scope = formula(aic_model))
```

```
## Start:  AIC=19.96
## log(price) ~ 1
##
##           Df Sum of Sq    RSS      AIC
## + y        1    483.63  34.66 -1330.57
## + x        1    482.36  35.92 -1312.62
## + z        1    478.09  40.19 -1256.47
## + carat    1    451.09  67.19  -999.57
## + cut      4     31.36 486.93    -3.25
## + clarity  7     37.02 481.26    -3.10
## + table    1     22.95 495.33    -0.69
## <none>                 518.28    19.96
## + color    6     11.54 506.74    20.69
## + depth    1      0.41 517.87    21.56
##
## Step:  AIC=-1330.57
## log(price) ~ y
##
##           Df Sum of Sq    RSS     AIC
## + clarity  7   11.7853 22.871 -1524.4
## + color    6    6.8256 27.830 -1428.2
## + carat    1    2.8617 31.794 -1371.7
## + depth    1    0.1467 34.509 -1330.7
## + z        1    0.1388 34.517 -1330.6
## <none>                 34.656 -1330.6
## + table    1    0.0690 34.587 -1329.6
## + x        1    0.0307 34.625 -1329.0
## + cut      4    0.4300 34.226 -1328.8
##
## Step:  AIC=-1524.38
## log(price) ~ y + clarity
##
##         Df Sum of Sq    RSS     AIC
## + color  6    9.7834 13.087 -1791.5
## + carat  1    3.1211 19.749 -1595.7
## + z      1    1.0323 21.838 -1545.5
```

```
## + depth  1    0.8721 21.998 -1541.8
## + x      1    0.3296 22.541 -1529.6
## + cut    4    0.3917 22.479 -1525.0
## <none>               22.870 -1524.4
## + table  1    0.0003 22.870 -1522.4
##
## Step:  AIC=-1791.49
## log(price) ~ y + clarity + color
##
##         Df Sum of Sq    RSS     AIC
## + z      1   1.81523 11.272 -1864.2
## + carat  1   1.77949 11.308 -1862.6
## + depth  1   1.56548 11.522 -1853.2
## + x      1   0.64313 12.444 -1814.7
## <none>               13.087 -1791.5
## + table  1   0.01984 13.067 -1790.2
## + cut    4   0.14991 12.937 -1789.2
##
## Step:  AIC=-1864.15
## log(price) ~ y + clarity + color + z
##
##         Df Sum of Sq     RSS     AIC
## + carat  1    3.2282  8.0436 -2030.9
## + x      1    0.2556 11.0162 -1873.6
## + depth  1    0.0507 11.2212 -1864.4
## + cut    4    0.1839 11.0880 -1864.4
## <none>               11.2719 -1864.2
## + table  1    0.0305 11.2414 -1863.5
##
## Step:  AIC=-2030.86
## log(price) ~ y + clarity + color + z + carat
##
##         Df Sum of Sq    RSS     AIC
## + x      1  0.316138 7.7275 -2048.9
## + depth  1  0.106402 7.9372 -2035.5
## + cut    4  0.197591 7.8461 -2035.3
## <none>               8.0436 -2030.9
## + table  1  0.005318 8.0383 -2029.2
##
## Step:  AIC=-2048.91
## log(price) ~ y + clarity + color + z + carat + x
##
##         Df Sum of Sq    RSS     AIC
## + cut    4  0.243656 7.4838 -2056.9
## <none>               7.7275 -2048.9
## + depth  1  0.007798 7.7197 -2047.4
## + table  1  0.000089 7.7274 -2046.9
##
## Step:  AIC=-2056.93
## log(price) ~ y + clarity + color + z + carat + x + cut
##
##         Df Sum of Sq    RSS     AIC
## + table  1  0.087090 7.3968 -2060.8
## <none>               7.4838 -2056.9
```

```
## + depth  1  0.003759 7.4801 -2055.2
##
## Step:  AIC=-2060.78
## log(price) ~ y + clarity + color + z + carat + x + cut + table
##
##         Df Sum of Sq    RSS     AIC
## <none>               7.3968 -2060.8
## + depth  1  0.012125 7.3846 -2059.6
```

```
summary(forward)
```

```
##
## Call:
## lm(formula = log(price) ~ y + clarity + color + z + carat + x +
##     cut + table, data = diamonds_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37802 -0.08234 -0.00160  0.08539  0.37902
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.646368   0.240365  -2.689 0.007415 **
## y             0.021823   0.136741   0.160 0.873271
## clarityIF     1.174667   0.050390  23.312  < 2e-16 ***
## claritySI1    0.724791   0.043238  16.763  < 2e-16 ***
## claritySI2    0.541568   0.044007  12.306  < 2e-16 ***
## clarityVS1    0.930980   0.044749  20.805  < 2e-16 ***
## clarityVS2    0.862124   0.044291  19.465  < 2e-16 ***
## clarityVVS1   1.139110   0.049155  23.174  < 2e-16 ***
## clarityVVS2   1.030157   0.047127  21.859  < 2e-16 ***
## colorE       -0.050855   0.021945  -2.317 0.020906 *
## colorF       -0.112033   0.021570  -5.194 3.06e-07 ***
## colorG       -0.143824   0.021619  -6.653 7.90e-11 ***
## colorH       -0.276638   0.022372 -12.366  < 2e-16 ***
## colorI       -0.404476   0.026642 -15.182  < 2e-16 ***
## colorJ       -0.556470   0.030626 -18.170  < 2e-16 ***
## z             1.122292   0.088587  12.669  < 2e-16 ***
## carat        -0.961953   0.066050 -14.564  < 2e-16 ***
## x             0.666978   0.140528   4.746 2.74e-06 ***
## cutGood       0.122646   0.041184   2.978 0.003049 **
## cutIdeal      0.184874   0.042654   4.334 1.78e-05 ***
## cutPremium    0.136072   0.040220   3.383 0.000776 ***
## cutVery Good  0.161631   0.040571   3.984 7.84e-05 ***
## table         0.008273   0.003491   2.370 0.018192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1245 on 477 degrees of freedom
## Multiple R-squared:  0.9857, Adjusted R-squared:  0.9851
## F-statistic:  1498 on 22 and 477 DF,  p-value: < 2.2e-16
```

```
anova(forward,aic_model)
```

```
## Analysis of Variance Table
##
## Model 1: log(price) ~ y + clarity + color + z + carat + x + cut + table
## Model 2: log(price) ~ (X + carat + cut + color + clarity + depth + table +
##     x + y + z) - X
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    477 7.3968
## 2    476 7.3846  1  0.012125 0.7815 0.3771
```

```
anova(backward,aic_model)
```

```
## Analysis of Variance Table
##
## Model 1: log(price) ~ carat + cut + color + clarity + table + x + z
## Model 2: log(price) ~ (X + carat + cut + color + clarity + depth + table +
##     x + y + z) - X
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    478 7.3972
## 2    476 7.3846  2   0.01252 0.4035 0.6682
```

AIC Interpretation

Using AIC forward and backward resulted in giving us different predictive values for the model. The forward method outputted the predictors carat, cut, clarity, color, table, y, z, x . The backwards method gave carat, cut, color, clarity, table, x, z.

Running the summary of both of these models we receive the same R^2_{adj} of 0.9851. We then checked the ANOVA tables for both backward and forward and found the RSS for the forwards and and backwards AIC models to be 7.3968 and 7.3972, respectively.

Multicolinearity Check

During our analysis of the quantitative variables in part 1, we found that there was significant correlation between price and certain variables. Additionally, among these candidate variables for predictors in our model, we found there to be significant correlation. As a result, we should test the models produced through forward and backward AIC for multicolinearity to ensure the results produced by the model are of the utmost accuracy and reliability.

```
vif(forward)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## y       745.973447  1       27.312514
## clarity   1.676994  7        1.037619
## color     1.373081  6        1.026774
## z       121.975904  1       11.044270
## carat    30.194324  1        5.494936
## x       795.276239  1       28.200643
## cut       3.722380  4        1.178562
## table     1.923682  1        1.386969
```

```
vif(backward)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## carat    30.093855  1        5.485787
## cut       2.807837  4        1.137749
## color     1.361875  6        1.026073
## clarity   1.600887  7        1.034183
## table     1.887568  1        1.373888
## x       123.427980  1       11.109815
## z       120.346693  1       10.970264
```

Using the VIF function to test for multicolinearity, we found that several variables were highly correlated for both the forward and backward AIC models.

For the forward model x, y, z, and carat showed relationships and for the backward model x, z, and carat showed relationships.

From part 1 and 2, we found that carat is a more significant predictor for price than the other predictors with which it had interaction. Therefore, we decided carat should remain in the model. As a result, we removed x, y, and z from the forwards AIC model, as well as x and z from the backwards AIC model.

These changes yield a single model with predictors carat, cut, clarity, color, table. From here, we noticed that including cut as a predictor did not influence the model in a productive way. Therefore we suspect that including cut as a predictor may introduce overfitting to the model.

Observe the following data on the model when cut is included, then when cut is not included:

Cut included:

```
aic_modified <- lm(formula = log(price) ~ carat + clarity + color + table + cut, data = diamonds_sample)
summary(aic_modified)
```

```
##
## Call:
## lm(formula = log(price) ~ carat + clarity + color + table + cut,
##     data = diamonds_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15083 -0.20735  0.05493  0.21817  0.66444
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.382962   0.481338   9.106  < 2e-16 ***
## carat         2.253990   0.035009  64.384  < 2e-16 ***
## clarityIF     0.999822   0.124436   8.035 7.33e-15 ***
## claritySI1    0.720651   0.107441   6.707 5.59e-11 ***
## claritySI2    0.539343   0.109483   4.926 1.16e-06 ***
## clarityVS1    0.795566   0.110736   7.184 2.60e-12 ***
## clarityVS2    0.791018   0.109718   7.210 2.20e-12 ***
## clarityVVS1   0.957193   0.121425   7.883 2.17e-14 ***
## clarityVVS2   0.916088   0.116554   7.860 2.55e-14 ***
## colorE       -0.011244   0.054360  -0.207   0.8362
## colorF       -0.046320   0.053517  -0.866   0.3872
## colorG       -0.046657   0.053459  -0.873   0.3832
```

```
## colorH       -0.235650   0.055333  -4.259 2.47e-05 ***
## colorI       -0.352176   0.065909  -5.343 1.41e-07 ***
## colorJ       -0.613215   0.075727  -8.098 4.65e-15 ***
## table         0.015909   0.007818   2.035   0.0424 *
## cutGood       0.090392   0.095747   0.944   0.3456
## cutIdeal      0.080788   0.088778   0.910   0.3633
## cutPremium    0.079056   0.086572   0.913   0.3616
## cutVery Good  0.022900   0.087524   0.262   0.7937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3099 on 480 degrees of freedom
## Multiple R-squared:  0.9111, Adjusted R-squared:  0.9076
## F-statistic: 258.8 on 19 and 480 DF,  p-value: < 2.2e-16
```

Notes on model including cut:

If we look at the summary above, we can note that none of the possible values of cut have significant influence on the model.

Cut not included:

```
final_model <- lm(formula = log(price) ~ carat + clarity + color + table, data = diamonds_sample)
summary(final_model)
```

```
##
## Call:
## lm(formula = log(price) ~ carat + clarity + color + table, data = diamonds_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13334 -0.20044  0.05404  0.22258  0.68604
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.527154   0.398542  11.359  < 2e-16 ***
## carat        2.252284   0.034703  64.903  < 2e-16 ***
## clarityIF    1.027791   0.120458   8.532  < 2e-16 ***
## claritySI1   0.746054   0.103120   7.235 1.84e-12 ***
## claritySI2   0.568385   0.104845   5.421 9.36e-08 ***
## clarityVS1   0.822745   0.106453   7.729 6.33e-14 ***
## clarityVS2   0.818336   0.105182   7.780 4.41e-14 ***
## clarityVVS1  0.983301   0.117036   8.402 4.93e-16 ***
## clarityVVS2  0.938012   0.112762   8.319 9.12e-16 ***
## colorE      -0.008207   0.054112  -0.152   0.8795
## colorF      -0.045773   0.053343  -0.858   0.3913
## colorG      -0.039084   0.053094  -0.736   0.4620
## colorH      -0.234994   0.054765  -4.291 2.15e-05 ***
## colorI      -0.351428   0.065555  -5.361 1.29e-07 ***
## colorJ      -0.612974   0.075248  -8.146 3.23e-15 ***
## table        0.014066   0.006583   2.137   0.0331 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3097 on 484 degrees of freedom
## Multiple R-squared:  0.9104, Adjusted R-squared:  0.9076
## F-statistic: 327.9 on 15 and 484 DF,  p-value: < 2.2e-16
```

Notes on model not including cut:

From the summary above, we can observe that the model including cut and without cut both have the same R^2_{adj}, 0.9076. Although both models have the same R^2_{adj}, the cut-not-included model has a lower residual standard error, 0.3097, than the cut-included model, 0.3099. Additionally, the F-statistic produced by the cut-not-included model, 327.9, is much more significant than the model including cut, 258.8. This implies that the values produced from the model without cut are more significant than the model including cut.

As a result of the observations above, we decided to proceed with the model without cut. Our final model will be focused on the relation between price, as the independent variable, and carat, clarity, color, and table as predictors.

Notes on final model vs. initial model:

While the initial model we created that included all predictors produced a significantly higher R^2_{adj}, this model may produce unreliable conclusions due to overfitting and multicolinearity.

We can observe below that the initial model has extremely high GVIF values for several predictors, while the final model that we produced does not share this issue:

```
vif(aic_model)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## carat     30.299132  1        5.504465
## cut        3.817495  4        1.182285
## color      1.390697  6        1.027865
## clarity    1.692832  7        1.038316
## depth     35.359426  1        5.946379
## table      1.974571  1        1.405194
## x       1294.015043  1       35.972421
## y       1346.917318  1       36.700372
## z       2384.669223  1       48.833075
```

```
vif(final_model)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## carat    1.347233  1        1.160704
## clarity 1.352362  7        1.021795
## color    1.244095  6        1.018367
## table    1.105606  1        1.051478
```

Although we typically consider models with higher R^2_{adj} to be superior in their explanation of the data, in this case proceeding with the higher R^2_{adj} model would result in any results or conclusions produced using this model to be unreliable and potentially misleading.

Confidence/Prediction interval:

```
final_model <- lm(formula = log(price) ~ carat + clarity + color + table, data = diamonds_sample)
new_data <- data.frame(carat = 0.5, clarity = "VS2", color = "E", table = 55)
conf_int <- exp(predict(final_model, newdata = new_data, interval = "confidence"))
print(conf_int)
```

```
##        fit      lwr      upr
## 1 1390.002 1269.121 1522.397
```

```
pred_int <- exp(predict(final_model, newdata = new_data, interval = "prediction"))
print(pred_int)
```

```
##        fit      lwr      upr
## 1 1390.002 751.2274 2571.932
```

Interpretation:

From the final model, we took a data point assigning values to the predictors (carat = 0.5, clarity = "VS2", color = "E", table = 55).

Using these values we obtained a confidence interval of (1269.121, 1522.397), with a fitted value of 1390.002. In the context of our model and data, this means that we are 95% confident that the true mean price for diamonds with carat=0.05, clarity="VS2", color='E', and table=55 is between $1269.121 and $1522.397.

We also obtained a prediction interval of (751.2274, 2571.932), with a fitted value of 1390.002. In the context of our model and data, this means that we are 95% confident that the true price of an individual diamond with carat=0.05, clarity="VS2", color='E', and table=55 will fall between $751.2274 and $2571.932.

Both intervals have a fitted value of 1390.002, implying that the predicted price for a diamond with the above characteristics is $1390.002.

The confidence interval provides a range for the mean predicted price for a diamond with our predetermined characteristics, while the prediction interval provides a range for the price of one singular diamond with these characteristics. The sizable difference in the range of these intervals(PI is much wider than CI) is due to the fact that prediction intervals account for the variability and error/uncertainty in individual diamond prices.

Overall summary

For part one, we described all of the variables and gave their distributions. We decided to proceed using all of the variables and eliminate them as we acquired more information.

For part two, we analyzed the relationship between carat and price. As expected, we found carat to be a good predictor of price. We then performed transformations on the model of price vs. carat to see if we could improve the goodness of fit provided by the model. Additionally, we wanted to ensure the rigidity of assumptions necessary to conducting multiple linear regression.

We used the square root and log functions and found that the log worked the best, so moving forward we used log as our transformation throughout. We then proceeded to add all the variables one by one and obtain the summary statistics. With each added predictor, the adjusted R^2_{adj} would increase with some predictors only increasing by the slightest of margins. We also found that there was multicollinearity in some of the variables, specifically in x, y, z, carat, and depth.

Part 3 was where we really hammered down on obtaining a final model. We started by performing both the backwards and forwards AIC functions and obtained two similar models. We then checked for multicollinearity from these models using the VIF function and found several variables to be correlated. Taking the results of the VIF multicolinearity check into account, we eliminated predictors with a high GVIF value and formed a new model.

Due to observed values in the summary of this new model, we suspected including 'cut' as a predictor introduced overfitting to the model, so we decided to remove it.

These modifications led us to the following model, which we believe to be the best model with respect to R^2_{adj}, multicolinearity, and overfitting. The model can be described as such:

$\mathbb{E}[\mathbf{y}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

```
summary(final_model)
```

```
##
## Call:
## lm(formula = log(price) ~ carat + clarity + color + table, data = diamonds_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13334 -0.20044  0.05404  0.22258  0.68604
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.527154   0.398542  11.359  < 2e-16 ***
## carat        2.252284   0.034703  64.903  < 2e-16 ***
## clarityIF    1.027791   0.120458   8.532  < 2e-16 ***
## claritySI1   0.746054   0.103120   7.235 1.84e-12 ***
## claritySI2   0.568385   0.104845   5.421 9.36e-08 ***
## clarityVS1   0.822745   0.106453   7.729 6.33e-14 ***
## clarityVS2   0.818336   0.105182   7.780 4.41e-14 ***
## clarityVVS1  0.983301   0.117036   8.402 4.93e-16 ***
## clarityVVS2  0.938012   0.112762   8.319 9.12e-16 ***
## colorE      -0.008207   0.054112  -0.152   0.8795
## colorF      -0.045773   0.053343  -0.858   0.3913
## colorG      -0.039084   0.053094  -0.736   0.4620
## colorH      -0.234994   0.054765  -4.291 2.15e-05 ***
## colorI      -0.351428   0.065555  -5.361 1.29e-07 ***
## colorJ      -0.612974   0.075248  -8.146 3.23e-15 ***
## table        0.014066   0.006583   2.137   0.0331 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3097 on 484 degrees of freedom
## Multiple R-squared:  0.9104, Adjusted R-squared:  0.9076
## F-statistic: 327.9 on 15 and 484 DF,  p-value: < 2.2e-16
```

The above summary is provided for reference for the following predictors and their coefficients.

β_0 - Intercept term.

$\beta_1 x_1$

- β_1 - Carat coefficient, denotes the expected dollar amount change in price per one unit increase in carat, when all other predictors are held constant.
- x_1 - Carat.

$\beta_2 x_2$

- β_2 - Clarity coefficient, denotes the expected change in diamond price, when all other predictors are held constant, when clarity takes on a certain value, in comparison to the reference value, clarity=I1.
- x_2 - Clarity, possible values are I1(reference), SI1, SI2, VS1, VS2, VVS1, and VVS2.

$\beta_3 x_3$

- β_3 - Color coefficient, denotes the expected change in diamond price, when all other predictors are held constant, when color takes on a certain value, in comparison to the reference value, color=D.
- x_3 - Color, possible values are D(reference), E, F, G, H, I, and J.

$\beta_4 x_4$

- β_4 - Table coefficient. Denotes the expected dollar amount change in price per one unit increase in table, when all other predictors are held constant.
- x_4 - Table.

We then constructed confidence interval and prediction intervals given a specified value for each of our predictors. This provides an example of how the final model that we produced can be used in practical application.