Global HW Part 1

Taylor Stone and Carolina Suarez

2024-10-12

- **Universidad Carlos III de Madrid**
- **Programming in R**

The Diamonds data set, found in the ggplot2 package in R, contains information about the physical properties and pricing of almost 54,000 diamonds. The goal of the dataset is to determine if physical characteristics, such as cut, color, and clarity, strongly influence the pricing of diamonds. There are 10 variables within this dataset, a mixture of categorical and continuous variables. Among initial observations of the dataset and variables, there were several outliers found and removed for analytical purposes. However, it is important to keep these outliers in mind and acknowledge the effects they imposed on the data. ### Variables in the dataset:

1. carat (continuous): The weight of the diamond.
2. cut (categorical): Quality of the diamond's cut (Fair, Good, 3. Very Good, Premium, Ideal).
3. color (categorical): Diamond color, from D (best) to J (worst).
4. clarity (categorical): Diamond clarity, from I1 (worst) to IF (best).
5. depth (continuous): Total depth percentage.
6. table (continuous): Width of the top of the diamond relative to the widest point.
7. price (continuous): Price in US dollars.
8. x (continuous): Length of the diamond in mm.
9. y (continuous): Width of the diamond in mm.
10. z (continuous): Depth of the diamond in mm.

All points of the analysis will be based on the Diamonds dataset provided by R:

```
library(ggplot2)
data(diamonds)
```

Outlier removal and data organization:

```
new_diamonds = diamonds[-c(24068,48411,49190),] #Removing outliers
num_vars = new_diamonds[,-c(2,3,4)] #Creating a data frame of numerical variables
categ_vars = new_diamonds[,c(2,3,4)] #Creating a data frame of categorical variables
```

i) Frequency Tables Continuous Variables

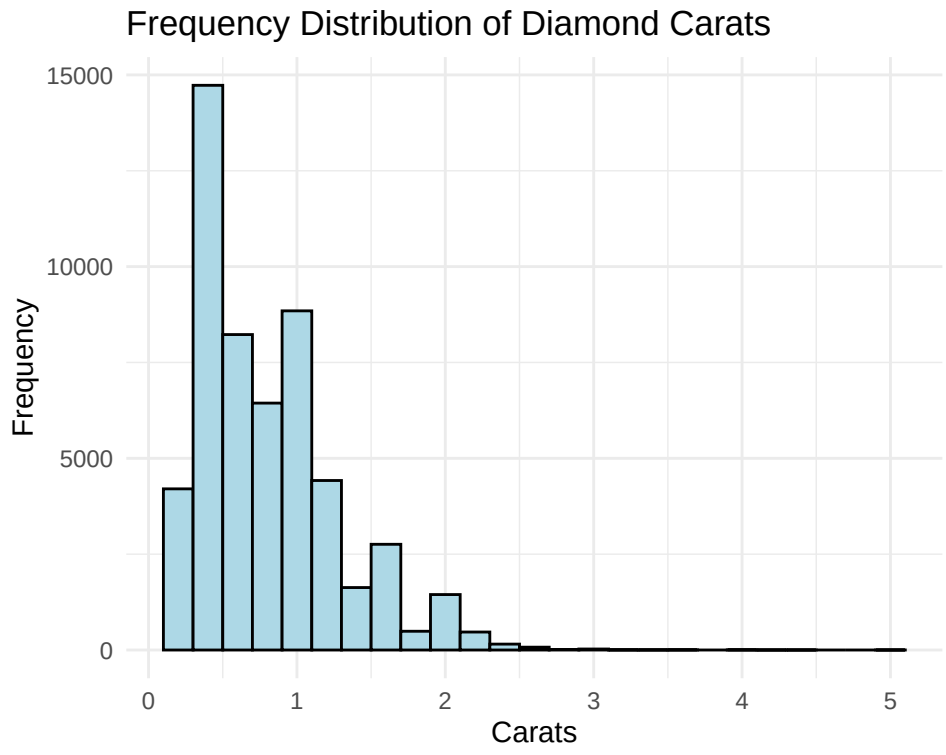
Carat (The weight of the diamond) variable

```
freq_table = table(cut(num_vars$carat, breaks=10))
knitr::kable(freq_table, caption = "Frequency Table for Carat")
```

Table 1: Frequency Table for Carat

Var1	Freq
(0.195,0.681]	25153
(0.681,1.16]	18626
(1.16,1.64]	7129
(1.64,2.12]	2348
(2.12,2.6]	614
(2.6,3.09]	53
(3.09,3.57]	6
(3.57,4.05]	5
(4.05,4.53]	2
(4.53,5.01]	1

```
ggplot(new_diamonds, aes(x = carat)) +
  geom_histogram(binwidth = 0.2, fill = "lightblue", color = "black") +
  ggtitle("Frequency Distribution of Diamond Carats") +
  xlab("Carats") +
  ylab("Frequency") +
  theme_minimal()
```



The carat variable represents a measurement of weight for diamonds where one carat indicates 200 milligrams. The diamonds are distributed between 0.2 to 5.01 carats, where 75% of the observations are between 0.2 and 1.05 carats The left side of the histogram has higher bars, indicating that smaller carat weights are much more common than larger ones. The peak level is between 0.2 to 0.48 carats, with 17,629 observations, indicating the prevalence of diamonds under half a carat. The remaining 15% of values fall above 1.05, however they significantly decrease after 2.18 carats, where just 525 diamonds out of the 53,940 are found above that level. At the maximum carat, 5.01, there is only one observation, indicating this weight is not common. The data

indicates more popularity between smaller carat diamonds, and will indicate more preference or availability for customers. As shown in the very right part of the chart as carats increase there are fewer diamonds, so they become rare to find, less demanded by the public, and can be considered luxury items. As the analysis advances forward, the data will show a relation between the carats of a diamond and its price, supporting that those with greater carats are treated as luxury, therefore the price will increase as well.

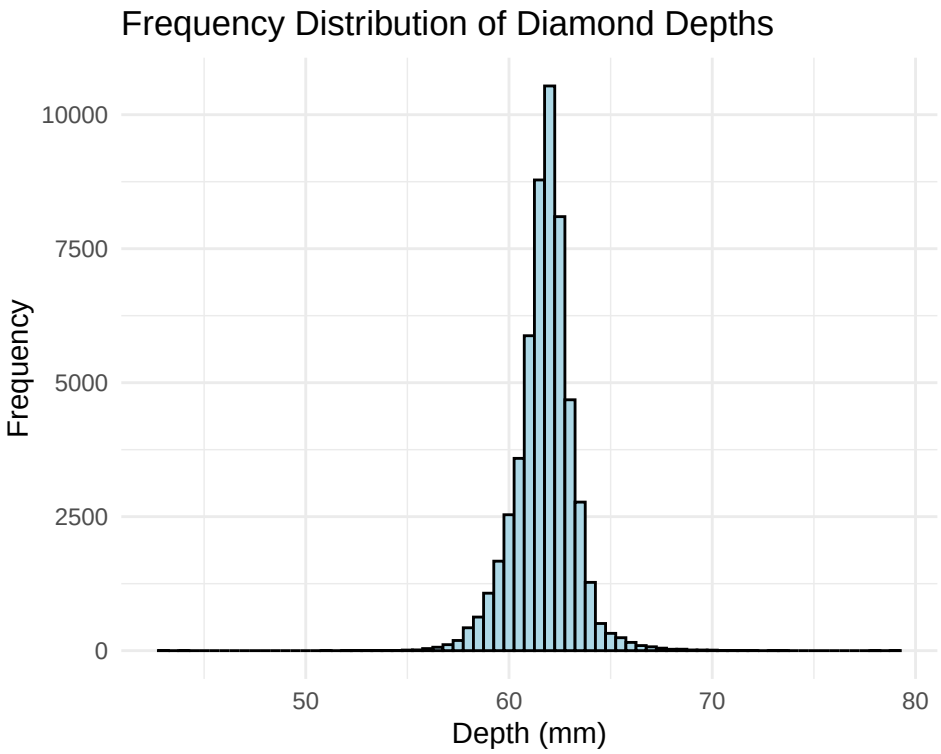
Depth (Total depth percentage) Variable

```
freq_table = table(cut(num_vars$depth, breaks=10))
knitr::kable(freq_table, caption = "Frequency Table for Depth")
```

Table 2: Frequency Table for Depth

Var1	Freq
(43,46.6]	3
(46.6,50.2]	0
(50.2,53.8]	12
(53.8,57.4]	304
(57.4,61]	13245
(61,64.6]	39261
(64.6,68.2]	1032
(68.2,71.8]	74
(71.8,75.4]	3
(75.4,79]	3

```
ggplot(new_diamonds, aes(x = depth)) +
    geom_histogram(binwidth = 0.5, fill = "lightblue", color = "black") +
    ggtitle("Frequency Distribution of Diamond Depths") +
    xlab("Depth (mm)") +
    ylab("Frequency") +
    theme_minimal()
```



The frequency chart categorizes depth percentage, referring to the ratio of the diamond's total depth to its average diameter. With a minimum of 43 and a maximum of 79, the frequency shows that 87% of the diamonds have a depth percentage between 59.9 and 64.2, making the depth between those values the more common for diamonds. After this peak, the ranges below 51.5 and above 72.6 have low frequencies, indicating that diamonds with such depth percentages are rare in the data given. Additionally after a value of diamonds depth of 64.2, there is a noticeable drop-off as depth percentage increases. There are very few and rare diamonds in the ranges above 70, difficult to distinguish in the chart, making them not common around availability and demand.

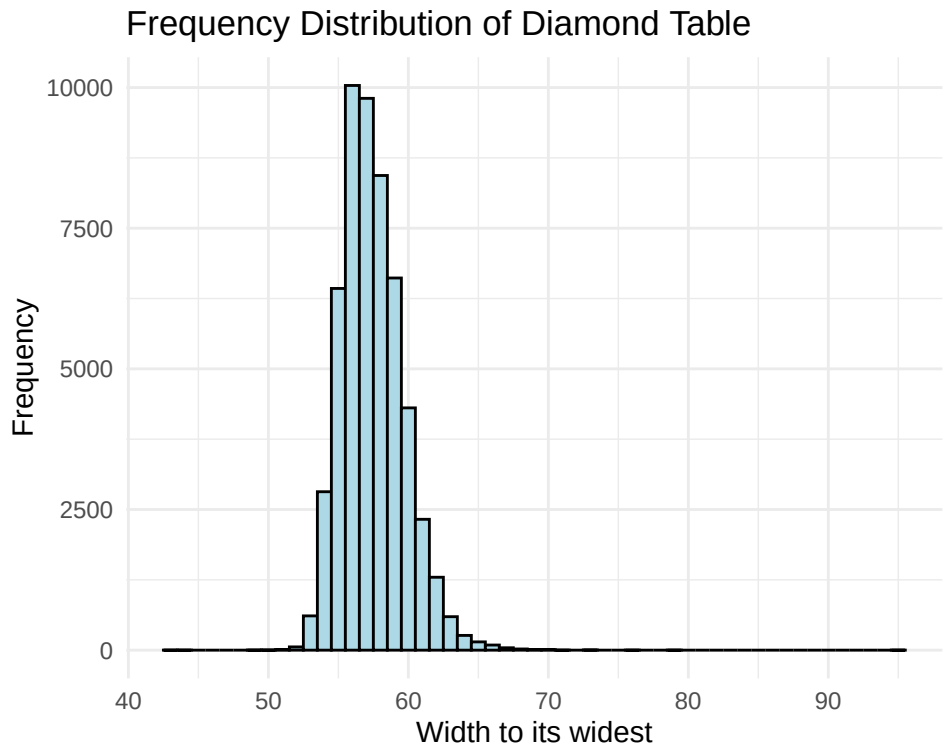
Table (Width of the top of the diamond) variable

```
freq_table = table(cut(num_vars$table, breaks=10))
knitr::kable(freq_table, caption = "Frequency Table for Table")
```

Table 3: Frequency Table for Table

Var1	Freq
(42.9,48.2]	2
(48.2,53.4]	669
(53.4,58.6]	37546
(58.6,63.8]	15131
(63.8,69]	572
(69,74.2]	14
(74.2,79.4]	2
(79.4,84.6]	0
(84.6,89.8]	0
(89.8,95.1]	1

```
ggplot(new_diamonds, aes(x = table)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  ggtitle("Frequency Distribution of Diamond Table") +
  xlab("Width to its widest") +
  ylab("Frequency") +
  theme_minimal()
```



The variable Table represents the value of the width of the top of the diamond relative to its widest point. In the data, this variable takes values between 43 to 95, with the highest frequency of 52% of the observations being in the values between 55.2 to 58.3. The range of 58.3 - 61.4, with 24.5% data concentration. Outside of the two previous ranges, there is a significant drop in frequency. After a depth of 70, observations become more rare and very few diamonds fall into the extreme low 43 and high 95 width ranges. The data suggest that diamonds with widths between 55 and 60 are the most prevalent or common between the market supply and demand given by the data.

Price (US dollars) variable

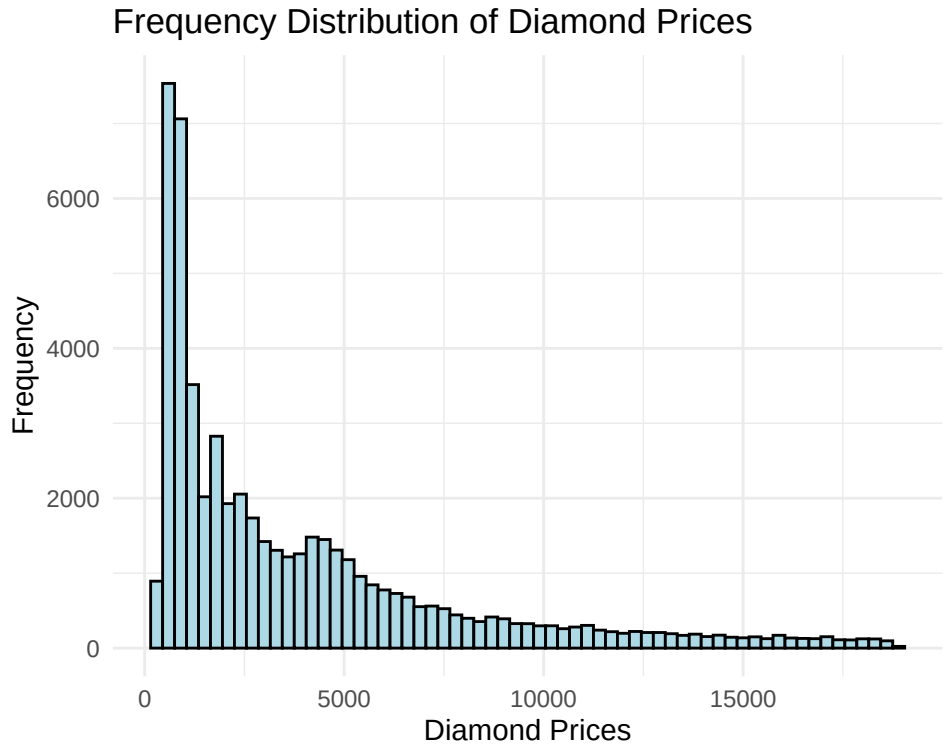
```
freq_table = table(cut(num_vars$price, breaks=10))
knitr::kable(freq_table, caption = "Frequency Table for Price")
```

Table 4: Frequency Table for Price

Var1	Freq
(308,2.18e+03]	25333
(2.18e+03,4.03e+03]	9328
(4.03e+03,5.88e+03]	7393
(5.88e+03,7.72e+03]	3878

Var1	Freq
(7.72e+03,9.57e+03]	2364
(9.57e+03,1.14e+04]	1745
(1.14e+04,1.33e+04]	1305
(1.33e+04,1.51e+04]	1002
(1.51e+04,1.7e+04]	863
(1.7e+04,1.88e+04]	726

```
ggplot(new_diamonds, aes(x = price)) +
  geom_histogram(binwidth = 300, fill = "lightblue", color = "black") +
  ggtitle("Frequency Distribution of Diamond Prices") +
  xlab("Diamond Prices") +
  ylab("Frequency") +
  theme_minimal()
```



The price of the diamonds are given in US dollars with a minimum price found at $326 and maximum value of $18,823. The frequency chart shows a wide range of the prices, however it is skewed to the left with 36% out of the total data set between $326 and $1,410. As the price increases, the number of diamonds sold decreases, which satisfies elasticity of demand properties. After the initial peak, there is a notable drop in frequency and a wide distribution between values above $1,410 to $5,770. Frequencies for diamonds priced above $10,000 become quite low with just 7.5% of the total data. The data shows a preference for prices below $1,410 as the most affordable price for the customers.

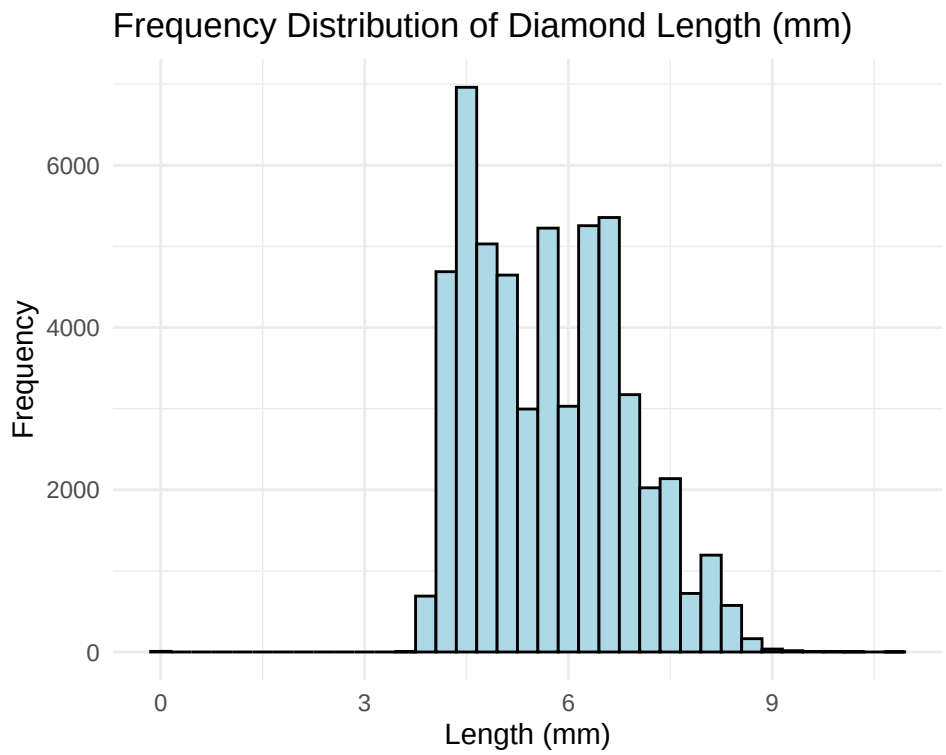
X (Length of the diamond in mm) variable

```
freq_table = table(cut(num_vars$x, breaks=10))
knitr::kable(freq_table, caption = "Frequency Table for Length")
```

Table 5: Frequency Table for Length

Var1	Freq
(-0.0107,1.07]	8
(1.07,2.15]	0
(2.15,3.22]	0
(3.22,4.3]	2934
(4.3,5.37]	20893
(5.37,6.44]	14440
(6.44,7.52]	12205
(7.52,8.59]	3259
(8.59,9.67]	191
(9.67,10.8]	7

```
ggplot(new_diamonds, aes(x = x)) +
  geom_histogram(binwidth = 0.3, fill = "lightblue", color = "black") +
  ggtitle("Frequency Distribution of Diamond Length (mm)") +
  xlab("Length (mm)") +
  ylab("Frequency") +
  theme_minimal()
```



The variable X denotes the length of the diamond in mm, and takes values between the ranges of 3.79 mm to 6.95 mm. Outside of this range, the histogram shows a significant drop in frequencies, indicating that larger diamonds and smaller ones are less common in the data set. The tallest bar of the histogram contains 20% of the observations and represents those that have a length between 6.32 mm and 6.95 mm.

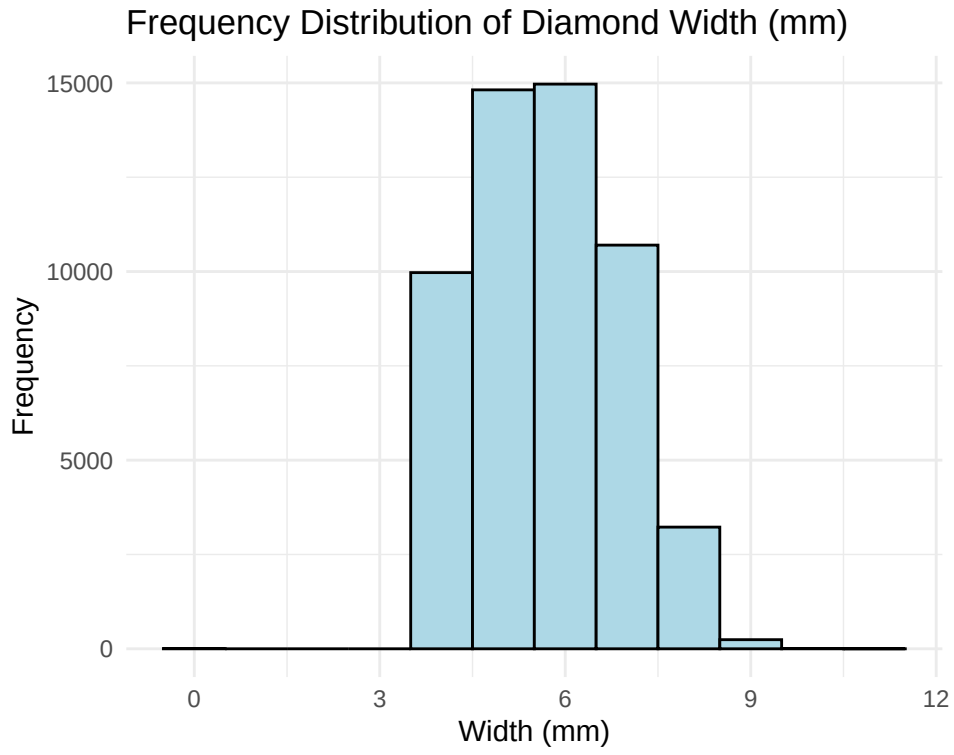
Y (Width of diamonds in mm) variable

```r
freq_table = table(cut(num_vars$y, breaks=10))
knitr::kable(freq_table, caption = "Frequency Table for Width")
```

Table 6: Frequency Table for Width

Var1	Freq
(-0.0105,1.05]	7
(1.05,2.11]	0
(2.11,3.16]	0
(3.16,4.22]	1323
(4.22,5.27]	20885
(5.27,6.32]	13330
(6.32,7.38]	13796
(7.38,8.43]	4257
(8.43,9.49]	330
(9.49,10.6]	9

```r
ggplot(new_diamonds, aes(x = y)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  ggtitle("Frequency Distribution of Diamond Width (mm)") +
  xlab("Width (mm)") +
  ylab("Frequency") +
  theme_minimal()
```



In this dataset, the variable Y represents the width of diamonds measured in millimeters (mm). The observations have widths ranging from 4.34 mm to 6.82 mm, with the highest frequency observed within the interval [4.34 mm, 4.96 mm], which includes 13,155 diamonds. Smaller diamonds are relatively uncommon

in this dataset. Additionally, there is a significant decrease in frequency for diamonds wider than 8.06 mm, suggesting that larger diamonds are more rare.

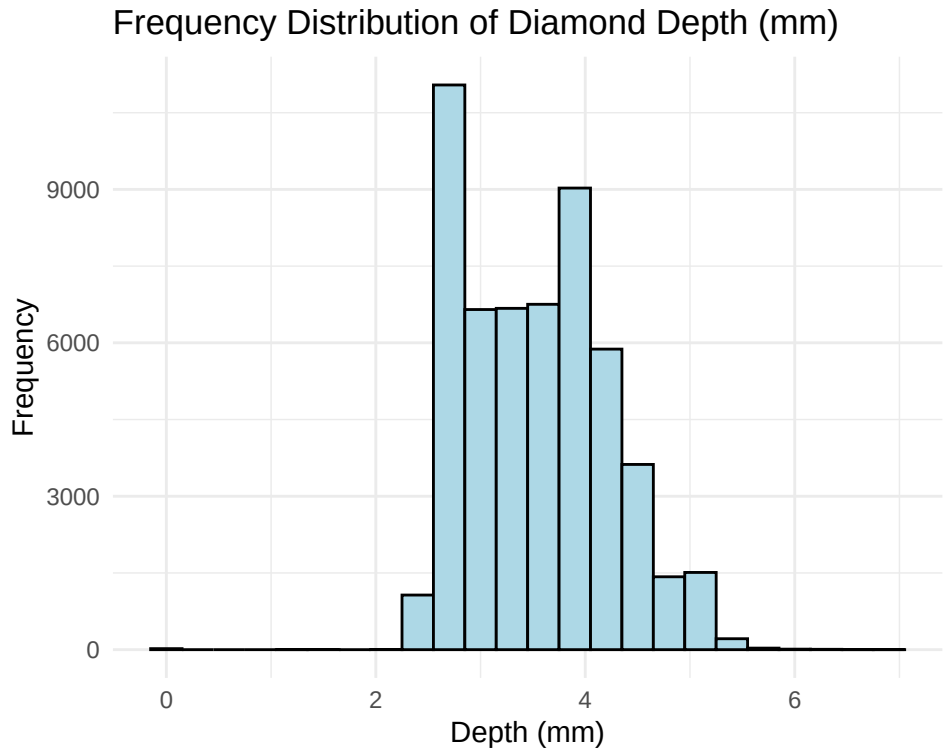
Z (Depth of the diamond in mm) variable

```
freq_table = table(cut(num_vars$z, breaks=10))
knitr::kable(freq_table, caption = "Frequency Table for Depth")
```

Table 7: Frequency Table for Depth

Var1	Freq
(-0.00698,0.698]	20
(0.698,1.4]	1
(1.4,2.09]	3
(2.09,2.79]	10675
(2.79,3.49]	15417
(3.49,4.19]	18098
(4.19,4.89]	7661
(4.89,5.58]	2022
(5.58,6.28]	35
(6.28,6.99]	5

```
ggplot(new_diamonds, aes(x = z)) +
  geom_histogram(binwidth = 0.30, fill = "lightblue", color = "black") +
  ggtitle("Frequency Distribution of Diamond Depth (mm)") +
  xlab("Depth (mm)") +
  ylab("Frequency") +
  theme_minimal()
```



The histogram for the depth measurement variable given by the variable z has a frequency distribution that takes values between 1.9mm to 31.8mm. As seen in the frequency plot, there is no concentration of diamonds after 6.16mm, being the maximum value 31.8 an outlier that expands the ranges of the frequency plot. To give a better understanding of how the diamonds are distributed, the plot is given with a binwidth of 0.30 that gives a more comprehensive view of the concentration of diamonds, which will be given by the 85% concentration around 2.37mm to 4.27mm. After the last value is observed, there is a significant decrease in the frequency of data points for greater depths.

ii) Measures of Centrality

Descriptive Statistics

```
sapply(num_vars, mean)
```

```
##        carat        depth        table        price            x            y
##    0.7979281   61.7494558   57.4572891 3932.7170959    5.7311356    5.7330678
##            z
##    3.5380967
```

```
sapply(num_vars, max)
```

```
##    carat    depth    table    price        x        y        z
##     5.01    79.00    95.00 18823.00    10.74    10.54     6.98
```

```
sapply(num_vars, min)
```

```
## carat depth table price     x     y     z
##   0.2  43.0  43.0 326.0   0.0   0.0   0.0
```

```
sapply(num_vars, var)
```

```
##        carat        depth        table        price            x            y
## 2.246693e-01 2.052367e+00 4.992969e+00 1.591511e+07 1.258301e+00 1.239533e+00
##            z
## 4.828045e-01
```

```
sqrt(sapply(num_vars, var))
```

```
##        carat        depth        table        price            x            y
##    0.4739929    1.4326086    2.2344952 3989.3745126    1.1217401    1.1133432
##            z
##    0.6948414
```

```
library(e1071)
sapply(num_vars, kurtosis)
```

```
##      carat      depth      table      price          x          y          z
##  1.2564821  5.7389655  2.8013247  2.1776058 -0.6183312 -0.6393917 -0.4833309
```

```r
sapply(num_vars, skewness)
```

```
##       carat       depth       table       price           x           y
##  1.11655479 -0.08225704  0.79680931  1.61837529  0.37858829  0.37505246
##           z
##  0.34486256
```

```r
find_mode <- function(x) {
  freq_table <- table(x)
  mode_value <- names(which.max(freq_table))
  return(mode_value)
}
modes <- sapply(categ_vars, find_mode)
print(modes)
```

```
##     cut   color clarity
## "Ideal"     "G"   "SI1"
```

Carat (The weight of the diamond)

The carat variable represents the weight of the diamonds in the dataset. On average, diamonds have a weight of approximately 0.798 carats, with the smallest diamond weighing 0.2 carats and the largest weighing 5.01 carats. The variability in diamond weights is moderate, with a variance of 0.225 and a standard deviation of 0.474, indicating that most diamonds are clustered around the mean but can vary by about half a carat. The distribution is positively skewed, with a skewness of 1.117, meaning there are more smaller diamonds, with a few larger diamonds extending the distribution to the right. The kurtosis of 1.256 suggests that the distribution has relatively moderate tails, meaning there are some outliers, but the overall distribution is not heavily concentrated around the mean. The removal of outliers did not significantly affect any descriptive statistics for this variable.

Cut (Quality of the diamond's cut)

The cut variable represents the quality of a diamond's cut, categorized as Fair, Good, Very Good, Premium, and Ideal. The most common cut quality in the dataset is Ideal, indicating that the majority of diamonds have the highest quality cut.

Color (Diamond color)

The color variable in the dataset indicates the quality of diamond color, ranging from D (the best) to J (the worst). The most common color among the diamonds is G, which represents a near-colorless grade, suggesting that many diamonds in the dataset have desirable color characteristics that fall within the high-quality range. The full scale of diamond colors is as follows: D, E, F (colorless), G, H, I (near-colorless), and J (light yellow), with the color quality declining as you move from D to J.

Clarity (Diamond clarity)

The clarity of diamonds, categorized from I1 (the worst quality) to IF (the best quality), indicates the presence of internal or external flaws, which can affect a diamond's overall appearance and value. In this dataset, the most common clarity rating is SI1 (Slightly Included), suggesting that many diamonds have some minor inclusions that are visible under magnification but generally do not affect the stone's beauty significantly. The clarity scale includes the following categories: I1 (Included), I2, I3, SI1 (Slightly Included), SI2, VS1 (Very Slightly Included), VS2, VVS1 (Very Very Slightly Included), VVS2, and IF (Internally Flawless), with higher categories indicating fewer and less noticeable flaws.

Depth (Total depth percentage)

The depth percentage of a diamond is the ratio of its total height to its average diameter, expressed as a percentage, and it affects the diamond's ability to reflect light and its overall brilliance. The diamonds in this dataset have a mean depth value of 61.75%, with a range spanning from 43.0% to 79.0%. The variance is relatively low at 2.05, and the standard deviation is 1.43, indicating that most of the depth measurements are concentrated close to the mean. The distribution is slightly left-skewed, with a skewness value of -0.08, suggesting a small tendency for depth percentages to lean toward lower values. Additionally, the kurtosis is 5.74, meaning the distribution has heavier tails and a higher likelihood of extreme values compared to a normal distribution, indicating the presence of some outliers in the data.

Table (Width of the top of the diamond relative to the widest point)

The table of a diamond, defined as the width of the top surface relative to its widest point, plays a crucial role in determining the diamond's overall appearance and brilliance. In the diamonds dataset, the mean table percentage is approximately 57.46%, suggesting that, on average, the diamonds have a balanced table size, which contributes positively to their visual appeal. The minimum and maximum values, 43.0% and 95.0%, respectively, indicate a notable range of table sizes among the diamonds, with some exhibiting relatively narrow tables while others possess wider proportions. The variance of 4.99 and standard deviation of approximately 2.23 reflect the degree of variability in table sizes across the dataset. Furthermore, the skewness of 0.80 suggests a slight positive asymmetry in the distribution, indicating that a few diamonds have wider tables than the average, while most tend to cluster around the mean. The kurtosis value of 2.80, which is close to the normal distribution's kurtosis of 3, indicates a distribution that is relatively flat with lighter tails, suggesting fewer extreme values in table percentages. Overall, these statistics provide valuable insights into the characteristics of diamond tables, which can influence their quality and desirability in the market. The removal of outliers did not significantly affect any descriptive statistics for this variable.

Price (US Dollars)

The mean price of the diamonds is approximately $3,932.80, suggesting that most diamonds fall within a moderate price range. However, the minimum price of $326.00 indicates the presence of significantly less expensive diamonds, while the maximum price of $18,823.00 highlights the existence of high-value diamonds in the market. The variance of approximately 15,915,630 and a standard deviation of about $3,989.44 indicate substantial variability in diamond prices, meaning that prices can vary widely from the mean. Furthermore, a skewness of 1.62 suggests a positive skew, indicating that there are more diamonds priced below the mean, but a few extremely high-priced diamonds are pulling the average up. Lastly, the kurtosis value of 2.18 indicates a relatively flat distribution compared to a normal distribution, implying that the data have lighter tails and fewer extreme outliers than a typical bell-shaped curve. Overall, this analysis suggests a market characterized by a mix of both affordable and luxurious diamonds, with the potential for extreme pricing at the high end. The removal of outliers did not significantly affect any descriptive statistics for this variable.

X (Length of the diamond in mm)

The mean length is approximately 5.73 mm, indicating that, on average, diamonds in the dataset tend to fall around this size. However, the minimum length recorded is 0.0 mm, which suggests the presence of some unusually small or possibly faulty entries, while the maximum length reaches 10.74 mm, highlighting the potential for larger diamonds in the market. The variance of 1.26 and standard deviation of about 1.12 mm indicate that there is a moderate degree of variability in diamond lengths, reflecting diversity in sizes. The positive skewness value of 0.38 suggests a slight rightward asymmetry in the distribution, implying that there are more diamonds with lengths less than the mean, but a few larger diamonds stretch the distribution to the right. Conversely, the kurtosis value of -0.62 indicates that the distribution is platykurtic, meaning it has lighter tails than a normal distribution, which suggests fewer extreme values or outliers in terms of diamond length. Overall, this analysis provides a comprehensive overview of the diamond lengths, indicating a central tendency around 5.73 mm with a slight rightward skew and a relatively uniform spread of sizes. The removal of outliers did not significantly affect any descriptive statistics for this variable.

Y (Width of the diamond in mm)

The mean width is approximately 5.73 mm and the values take a range with a minimum width of 0 mm and a maximum width of 10.54 mm, following the removal of outliers. Observations 24,068 and 49,190 were

found to be outliers in the y-variable, with values of 58.9 and 31.8 respectively. After removing these values, the maximum value was 10.54, which shows there was a great effect of these values on certain statistics. The variance of 1.23 indicates some degree of variability in diamond widths, while the standard deviation of about 1.11 mm further highlights the spread of the widths around the mean. Prior to the removal of outliers, the y-variable obtained a skewness value of 2.43 and a kurtosis of 91.2, which indicates significant rightward asymmetry in the data. The kurtosis also suggests a highly peaked distribution with heavy tails. After removing the outliers, the skewness dropped to 0.375, indicating a slight right skew, and the kurtosis became 2.434, maintaining the data has a strong peak, but not as high as it was prior. The analysis of diamond widths shows that there is notable variability and a wide range influenced by outliers, with initial rightward skewness and high kurtosis indicating extreme asymmetry and a peaked distribution, which became more moderate after removing the outliers.

Z (Depth of the diamond in mm)

The average depth measurement is approximately 3.54 mm and the variable takes values that range from a minimum at 0 mm to a maximum depth of 6.98 mm, after removing the outliers. The observation 48,411 was found to be an outlier with a depth value of 31.8, a value almost 40 standard deviations away from the mean. The variance of 0.483 suggests that there is a moderate spread in the depth values, while the standard deviation of approximately 0.695 mm indicates how much individual diamond depths typically deviate from the mean. The depth of diamonds, measured in millimeters, originally displayed a skewness of 1.52. This value indicates a moderate to strong positive skew in the distribution of diamond depths. In practical terms, this suggests that the majority of diamonds in the dataset had depths that were below the mean, with a notable number of diamonds exhibiting significantly higher depths. The presence of these higher depth values likely contributed to the asymmetrical shape of the distribution, pushing the mean to the right of the median and indicating potential outliers in the dataset. After removing the outlier, the skewness of the diamond depths decreased to 0.345. This change signifies a notable shift in the distribution's symmetry. With a skewness of 0.345, the distribution of diamond depths is now only slightly positively skewed. This indicates that while there are still some higher depth values, their influence on the overall distribution has diminished considerably. Furthermore, the extremely high original kurtosis value of 47.08 signifies a distribution with heavy tails, implying that there are numerous outliers with significantly deeper depths compared to the rest of the dataset. After removing the outliers, the kurtosis value changed significantly to -0.483, indicating the distribution has slightly less "peakedness" than a normal distribution. Visually, a small, negative kurtosis describes a distribution that is slightly flat with thinner tails than a normal distribution. These results reflect a market where the majority of diamonds are of standard proportions, but outliers can impact overall pricing and selection based on depth. Additionally, the slight positive skew indicates that while higher depth values are less common, they do exist and may represent premium or unique diamond offerings in the market.

iii) Categorical Data Grouping

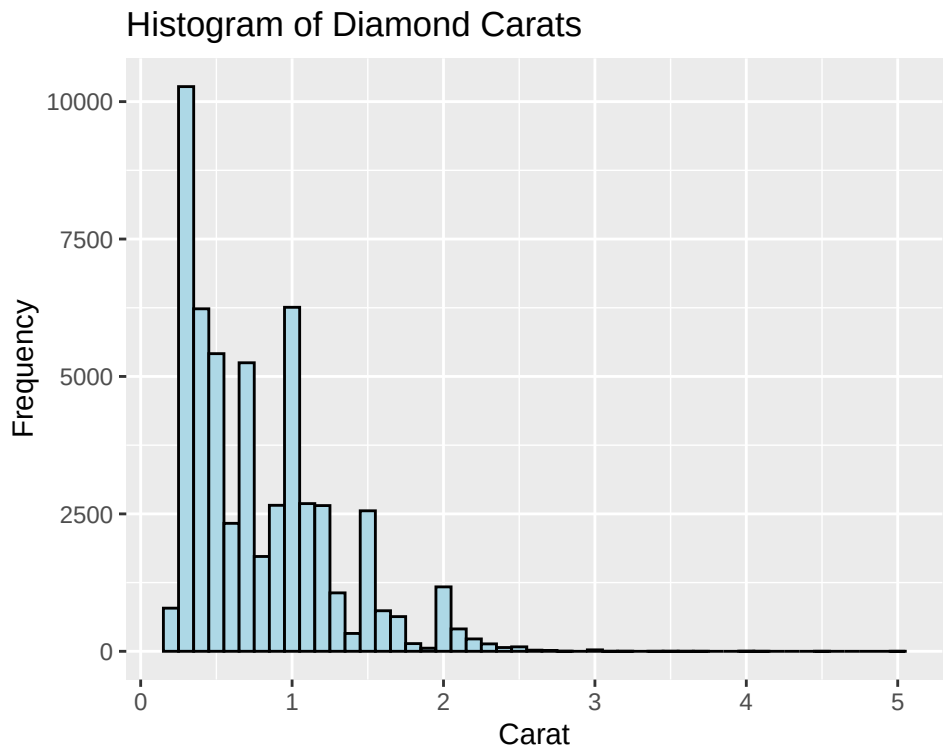
In the diamonds dataset, we are particularly interested in understanding how the categorical variable color relates to various characteristics of the diamonds. First, we will analyze how the price of diamonds varies across different colors, looking for any significant differences in average prices. Next, we will investigate whether there is a correlation between carat size and color, as this may reveal trends in the size distribution of diamonds within each color category.

We will also explore how the distribution of cuts—such as Fair, Good, Very Good, Premium, and Ideal—varies by color, as this can provide insights into the quality of diamonds associated with different colors. Additionally, we will assess whether the average table size differs by color, which might help identify patterns in how table size correlates with diamond color. Finally, we will examine how the clarity ratings differ across color groups, allowing us to understand if certain colors tend to have higher or lower clarity ratings. This comprehensive analysis will provide a clearer picture of the relationships between diamond color and these key characteristics.

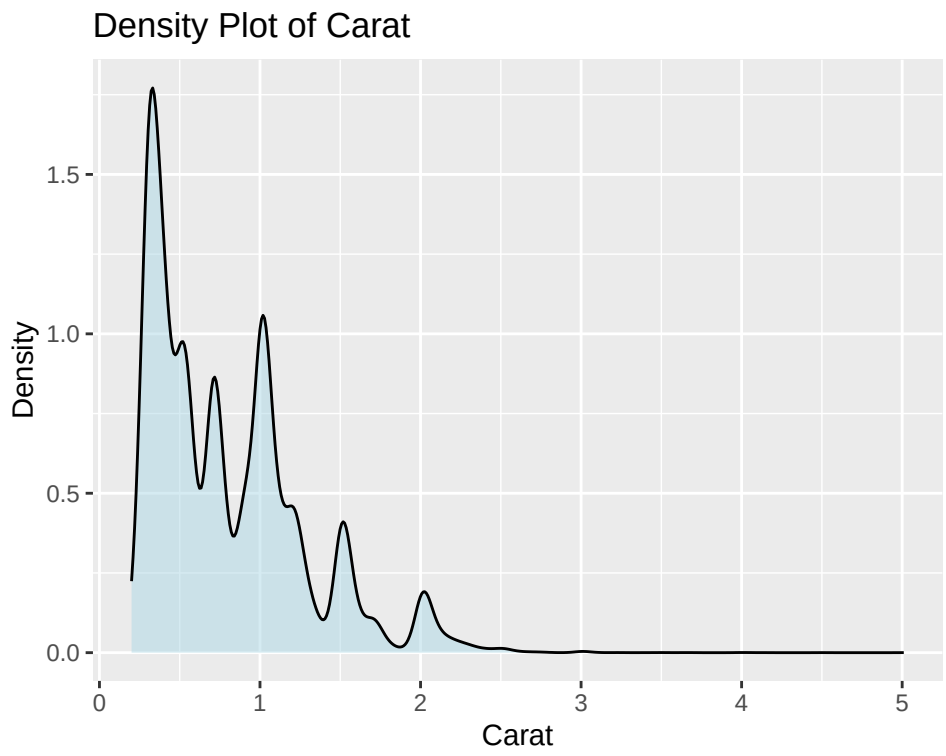
iv) Normality Assessment of the Continuous Variables

Carat (The weight of the diamond)

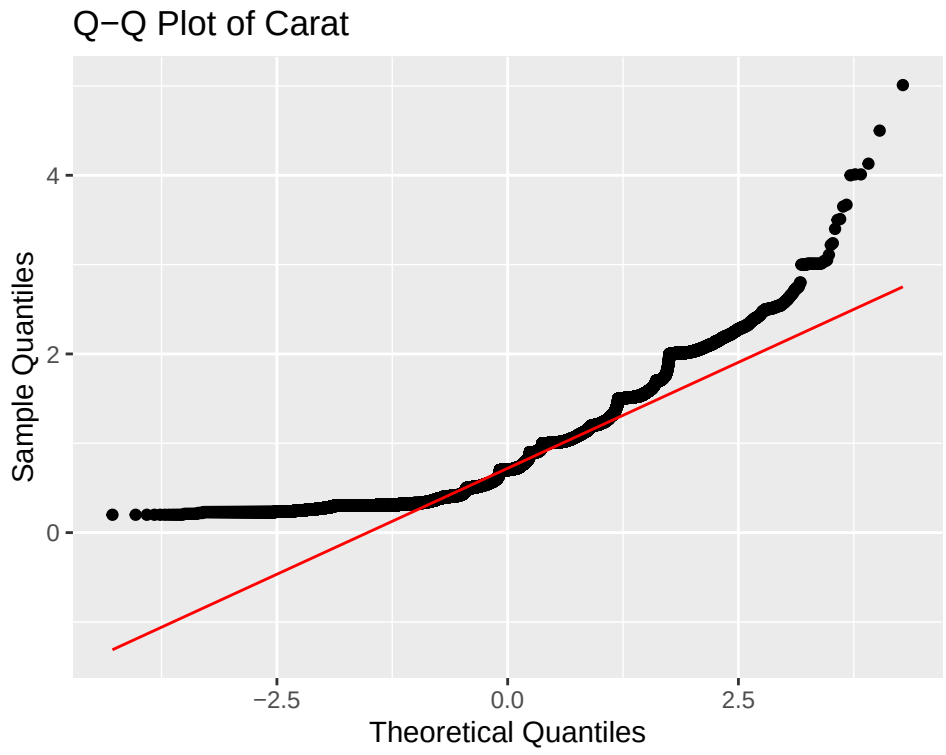
```
hist_plot = ggplot(data.frame(num_vars$carat), aes(x = num_vars$carat)) + geom_histogram(binwidth = 0.1

print(hist_plot)
```



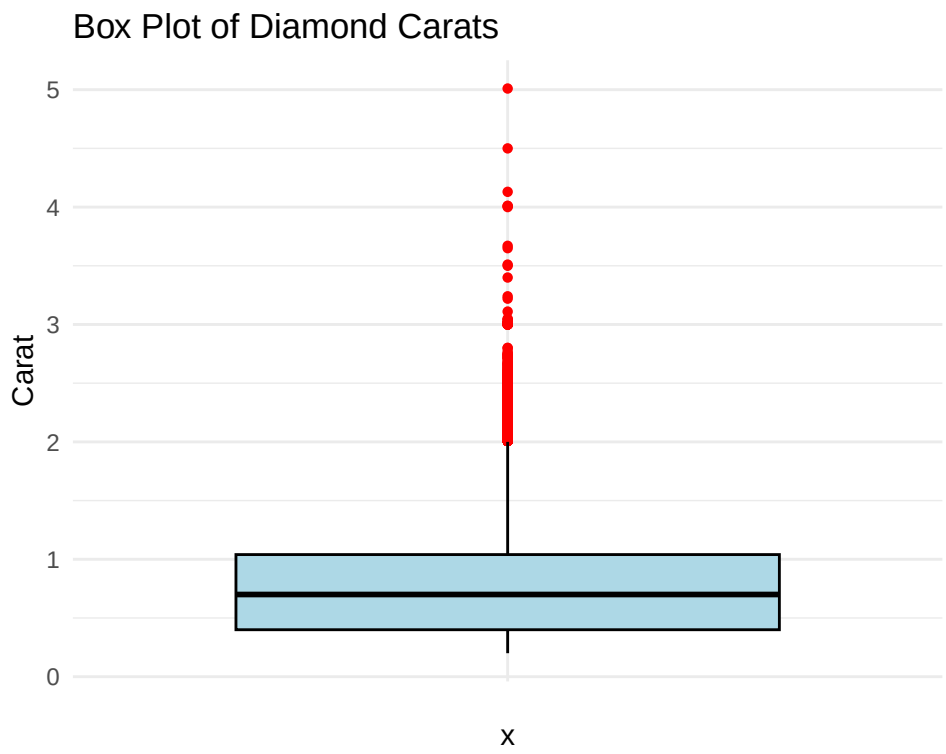

```
density_plot = ggplot(data = data.frame(num_vars$carat), aes(x = num_vars$carat)) + geom_density(fill =

print(density_plot)
```




```
qq_plot = ggplot(data = data.frame(num_vars$carat), aes(sample = num_vars$carat)) + stat_qq() + stat_qq_

print(qq_plot)
```



```r
box_plot = ggplot(data.frame(num_vars$carat), aes(x='',y = num_vars$carat)) + geom_boxplot(fill = "ligh

print(box_plot)
```



Based on the four visualizations—histogram, density plot, Q-Q plot, and box plot—the variable carat in the diamonds dataset does not follow a normal distribution. The histogram shows a clear right-skewed distribution, with a large concentration of diamonds having smaller carat sizes, particularly between 0 and 1 carat, while the frequency of larger carats drops off sharply.

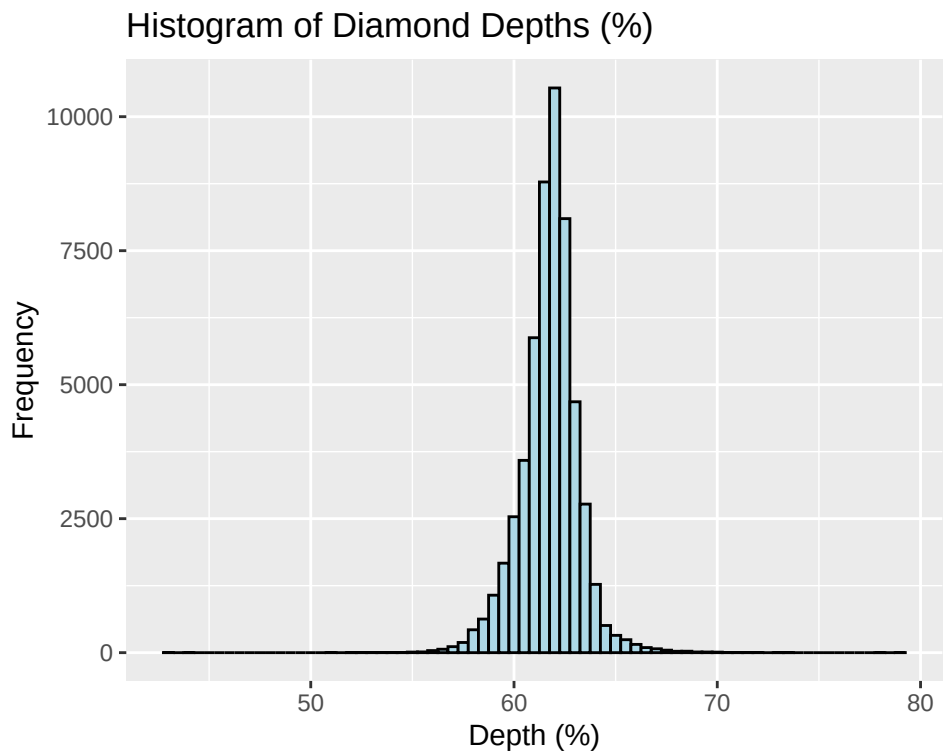
The density plot further emphasizes this skewness, with a multi-modal pattern and a long tail extending towards larger carat values, which is inconsistent with the symmetric bell-shaped curve of a normal distribution. Likewise, the density plot shows multiple peaks, which can suggest subgroups or modes within the data, furthering the deviation from the normal single-peak structure.

The Q-Q plot reveals a significant deviation from the diagonal line that would indicate normality, particularly in the tails. The right side of the plot shows carat sizes far above what would be expected from a normal distribution, suggesting the presence of large outliers and confirming the right-skewness. The left side of the plot shows values also pulling away from the normal line, and, when combined with the right tail pulling, demonstrates the data has a narrow peak.

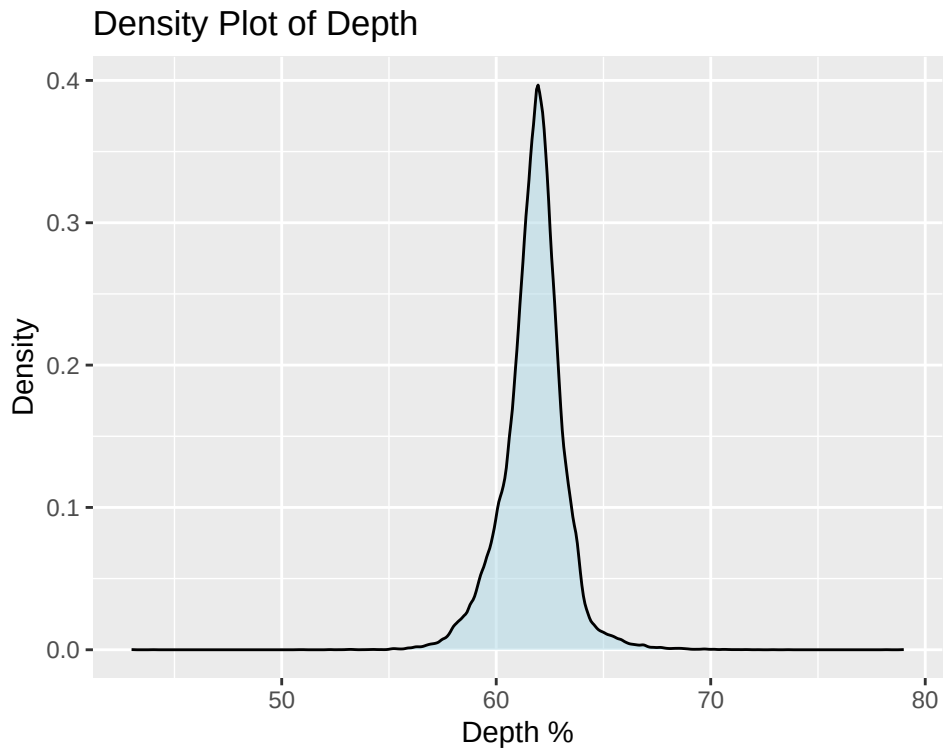
Lastly, the box plot highlights a significant number of outliers for larger carat sizes, with the upper whisker extending far beyond the interquartile range, again indicating a right-skewed distribution. Taken together, these plots strongly suggest that the carat variable is not normally distributed and exhibits significant skewness and outliers.

Depth (Total depth percentage)

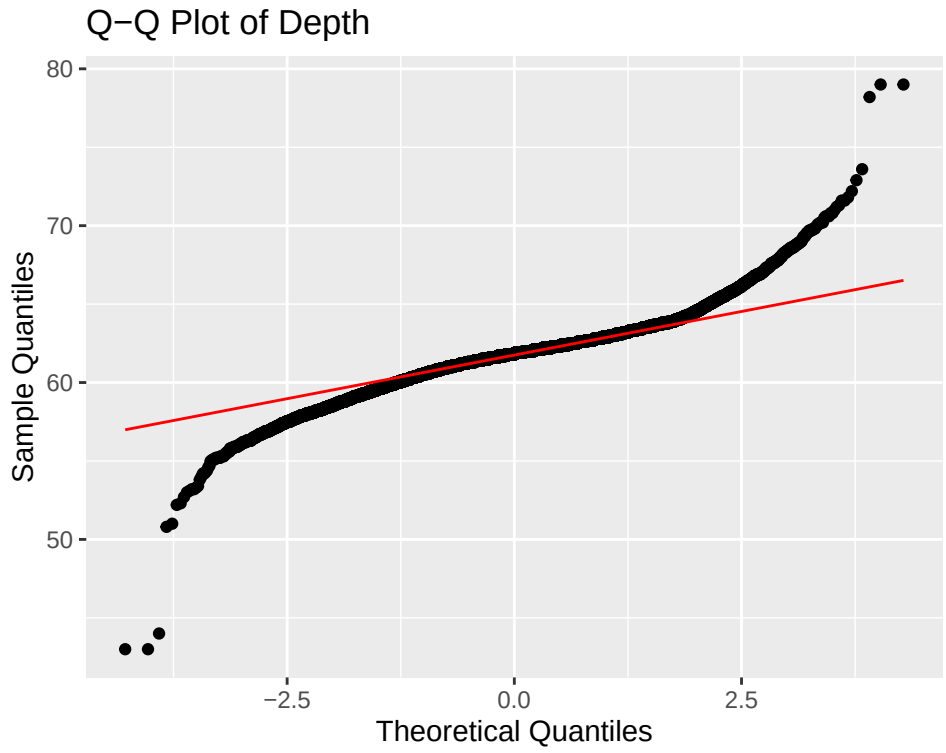
```r
hist_plot = ggplot(data.frame(num_vars$depth), aes(x = num_vars$depth)) + geom_histogram(binwidth = 0.5

print(hist_plot)
```



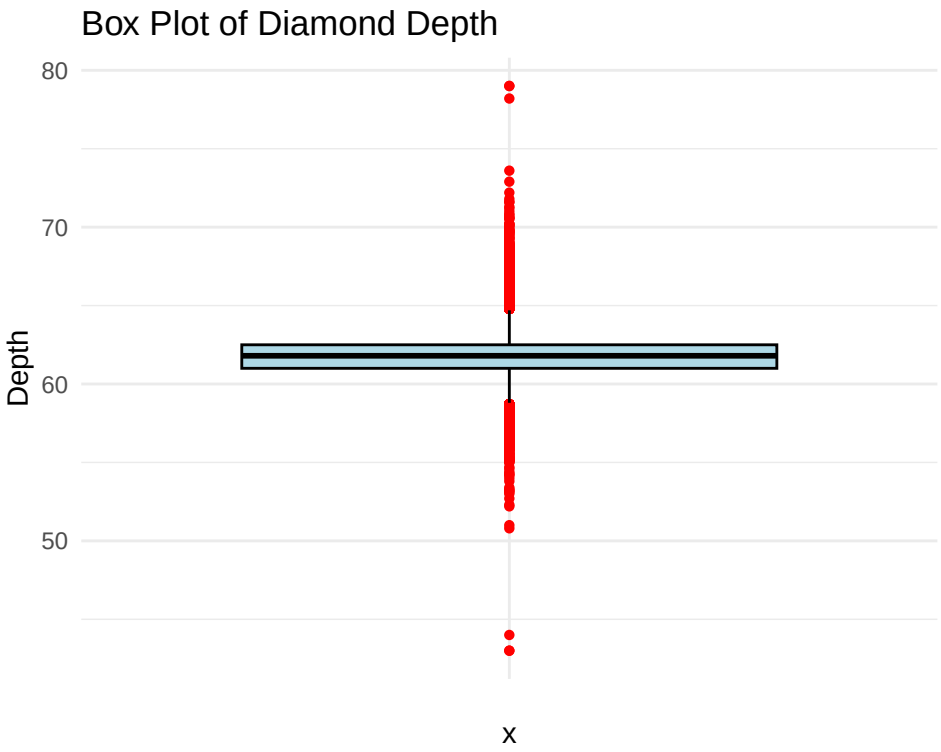

```
density_plot = ggplot(data = data.frame(num_vars$depth), aes(x = num_vars$depth)) + geom_density(fill =

print(density_plot)
```



```
qq_plot = ggplot(data = data.frame(num_vars$depth), aes(sample = num_vars$depth)) + stat_qq() + stat_qq

print(qq_plot)
```




```
box_plot = ggplot(data.frame(num_vars$depth), aes(x='',y = num_vars$depth)) + geom_boxplot(fill = "ligh

print(box_plot)
```



The histogram of diamond depth shows a roughly symmetric distribution, with a strong peak and a relatively narrow spread. Although the shape is somewhat bell-shaped, there are some deviations from a perfect normal distribution, with slight tapering at both tails. However, compared to the carat variable, the depth distribution appears to be more aligned with normality, especially around its central values.

The density plot reinforces the impression of symmetry seen in the histogram. The distribution peaks sharply and tapers off evenly on both sides. However, the tails extend farther than what would be typical in a normal distribution, indicating that while the data is centered in a normal-like manner, the spread in the tails is a bit more pronounced. This suggests a distribution that is close to normal but with slightly heavier tails.

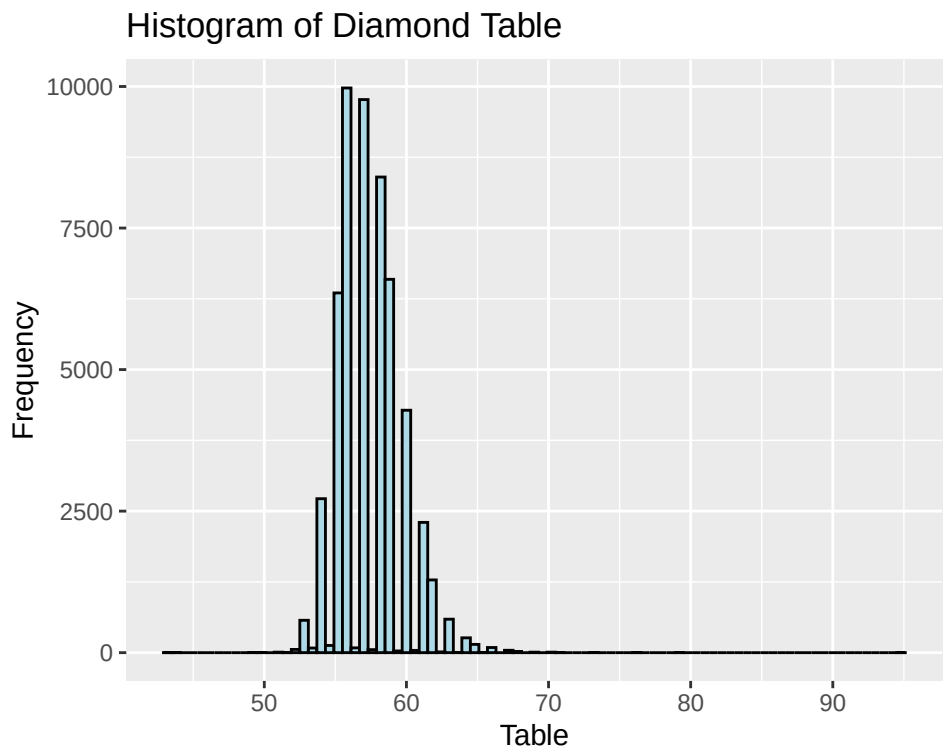
The Q-Q plot shows significant deviations from normality in the tails. While the central quantiles closely follow the diagonal red line, which suggests normality in the central portion of the distribution, the data points at both extremes deviate markedly from the line. The upward curve on the right side and the downward bend on the left indicate that the extreme values are more spread out than a normal distribution would predict. This confirms that the distribution has heavier tails than expected for a normal distribution.

The box plot of diamond depth shows several outliers, both above and below the interquartile range, which corresponds to the longer tails seen in the other plots. While the bulk of the data is tightly packed around the median, the whiskers extend relatively far, and there are numerous outliers beyond the whiskers. These outliers suggest that while the central portion of the data is reasonably normal and evenly distributed, the presence of outliers, particularly in the higher depth values, points to deviations from strict normality.

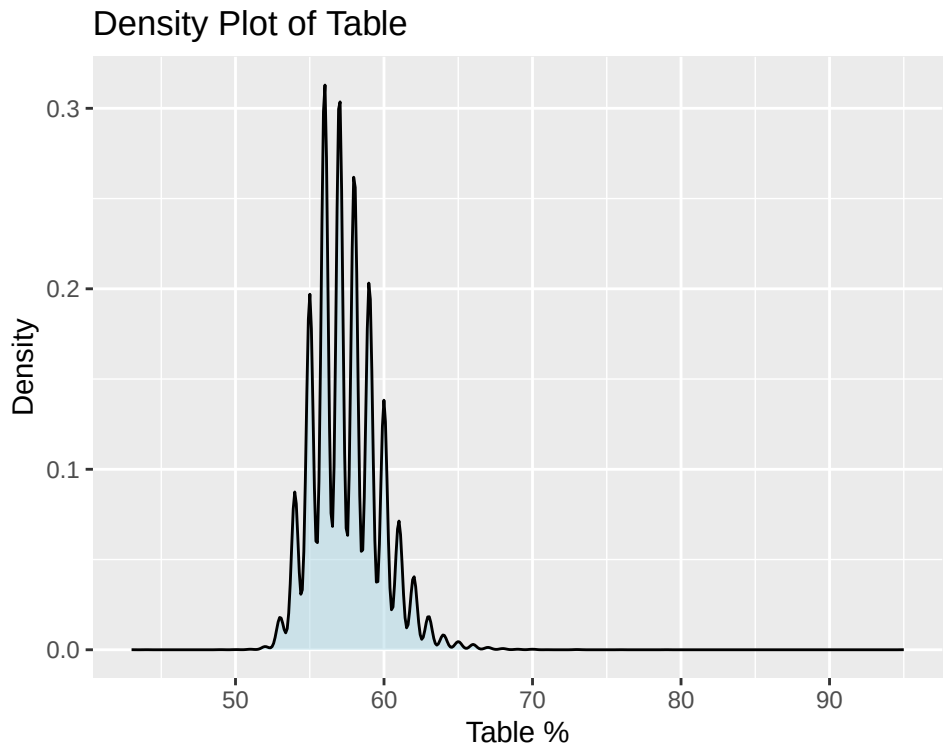
In conclusion, the depth variable shows a distribution that is approximately normal in the central region but has heavier tails and a notable number of outliers, particularly for extreme depth values. While the central mass of the data aligns well with a normal distribution, the presence of deviations in the tails indicates that the depth variable does not perfectly follow a normal distribution.

Table (Width of the top of the diamond relative to the widest point)

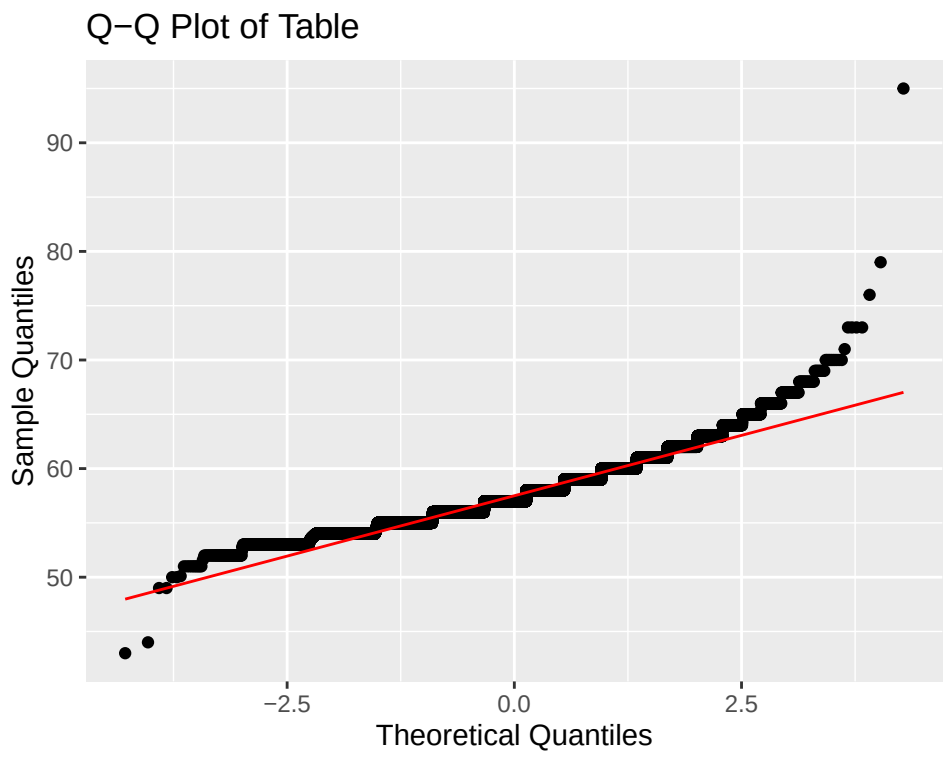
```
hist_plot = ggplot(data.frame(num_vars$table), aes(x = num_vars$table)) + geom_histogram(binwidth = 0.6

print(hist_plot)
```



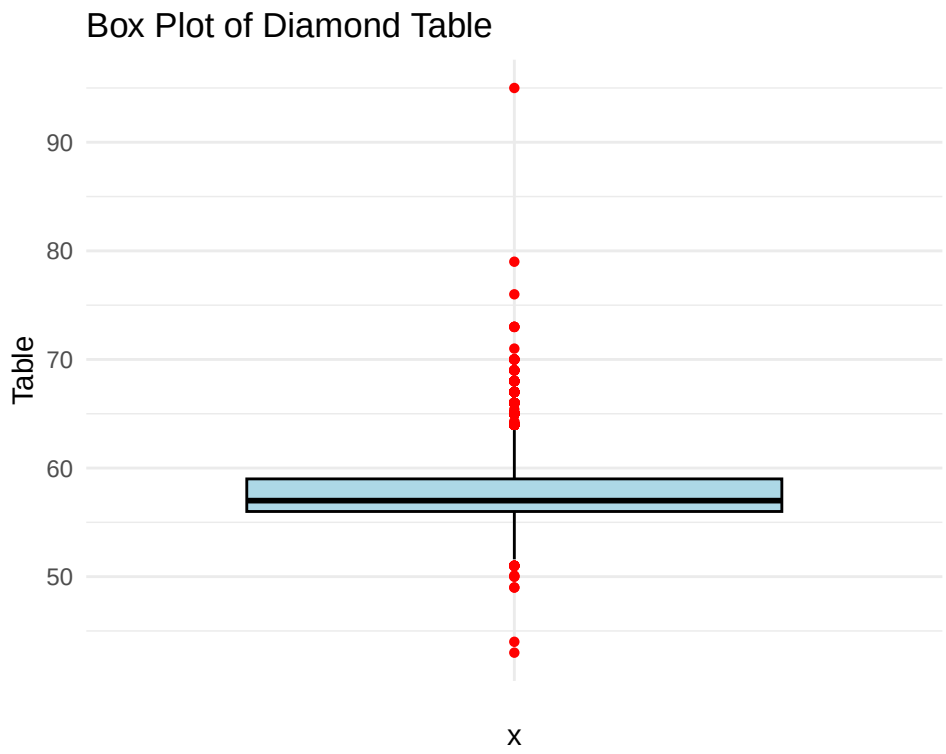

```
density_plot = ggplot(data = data.frame(num_vars$table), aes(x = num_vars$table)) + geom_density(fill =

print(density_plot)
```



```r
qq_plot = ggplot(data = data.frame(num_vars$table), aes(sample = num_vars$table)) + stat_qq() + stat_qq_

print(qq_plot)
```




```r
box_plot = ggplot(data.frame(num_vars$table), aes(x='',y = num_vars$table)) + geom_boxplot(fill = "ligh

print(box_plot)
```



The histogram for the variable table, which is the percentage calculated when dividing the size of the table by the average girdle diameter of the diamond, shows a slightly right skewed distribution. There is a peak around 55 to 60 and significant drop offs on either side, towards the lower and higher values. While the majority of the distribution in the center appears to be symmetric, the right tail, indicating the larger table values, extends further, skewing the distribution.

The density plot also demonstrates the right-skewness shape of the distribution. Similarly to the histogram, the data appears to have a central peak, with a slow taper to the right. This pattern suggests that, while the central portion of the data may appear to have a normal shape, the heavier right tail deviates from the expected shape of a normal distribution.

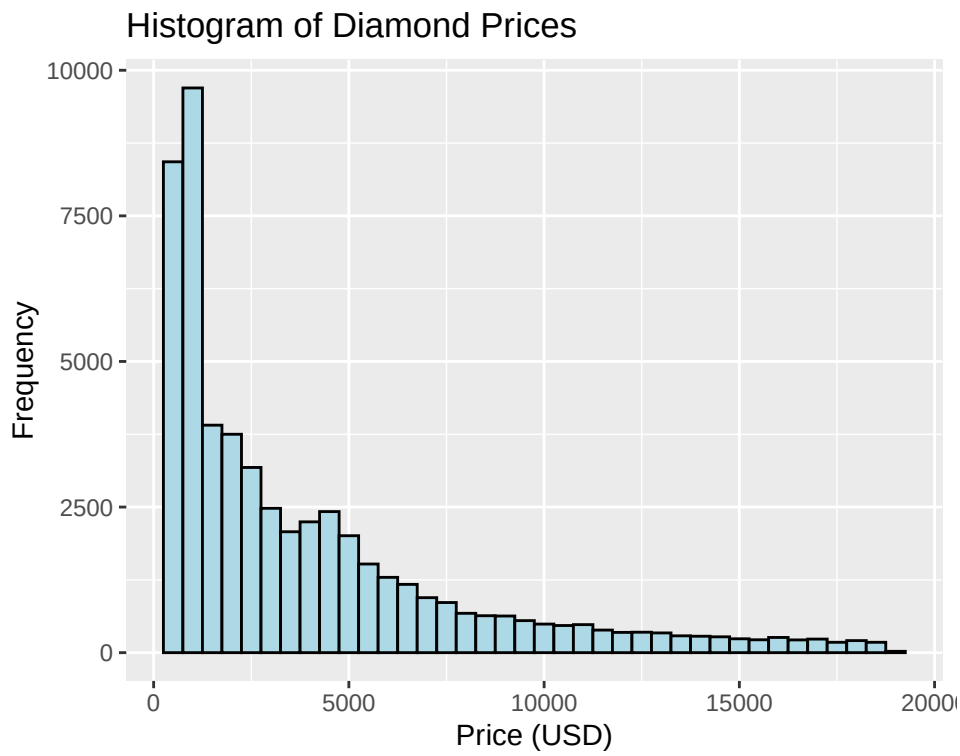
The Q-Q plot further provides us with a detailed view of how the table variable compares to a theoretical normal distribution. While the points in the middle quartiles lie close to the red line, indicating normality in the central portion, the points at both extremes show significant deviation. The upward curvature on the right side suggests that the larger table values are more extreme than a normal distribution would predict.

Finally, the boxplot reflects a similar message as the other plots. The number of outliers at the higher end corresponds to the right tail seen before. The central box, representing the interquartile range, is fairly narrow, but reveals a slight spread towards the right, the larger values, highlighting the right tail. This suggests that, while the bulk of the data is centered around the median, the extreme values and skewness deviate this data from a normal distribution.

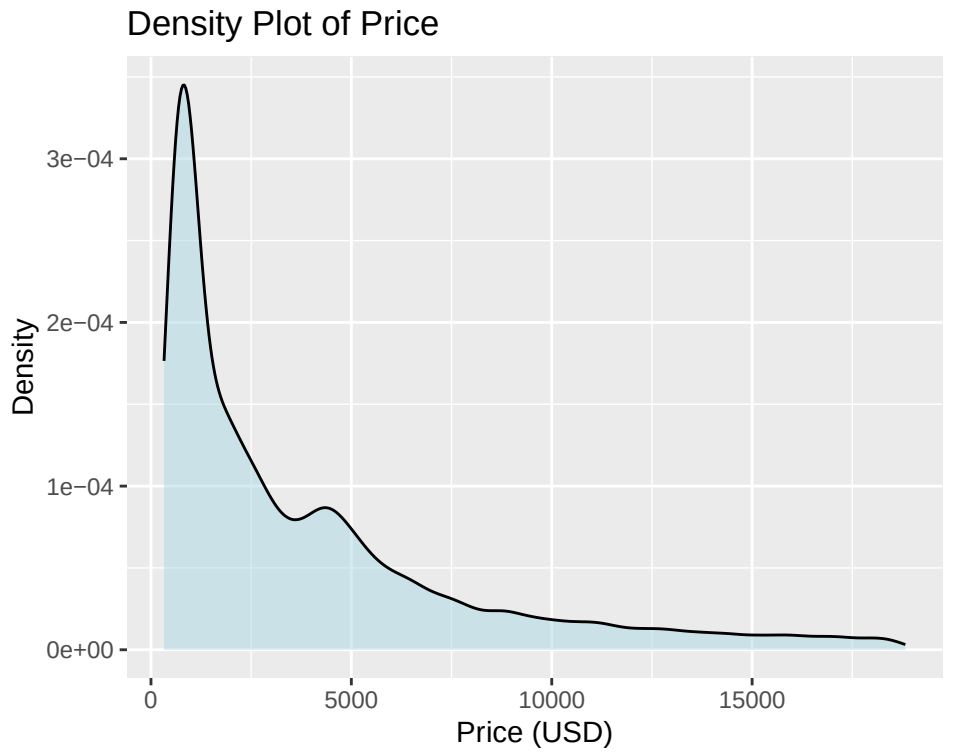
Gathering all of the information from the four plots, the table variable exhibits features of normality, particularly in the central region, but due to the heavy right tails and large outliers, the data deviates from a normal distribution.

Price (US Dollars)

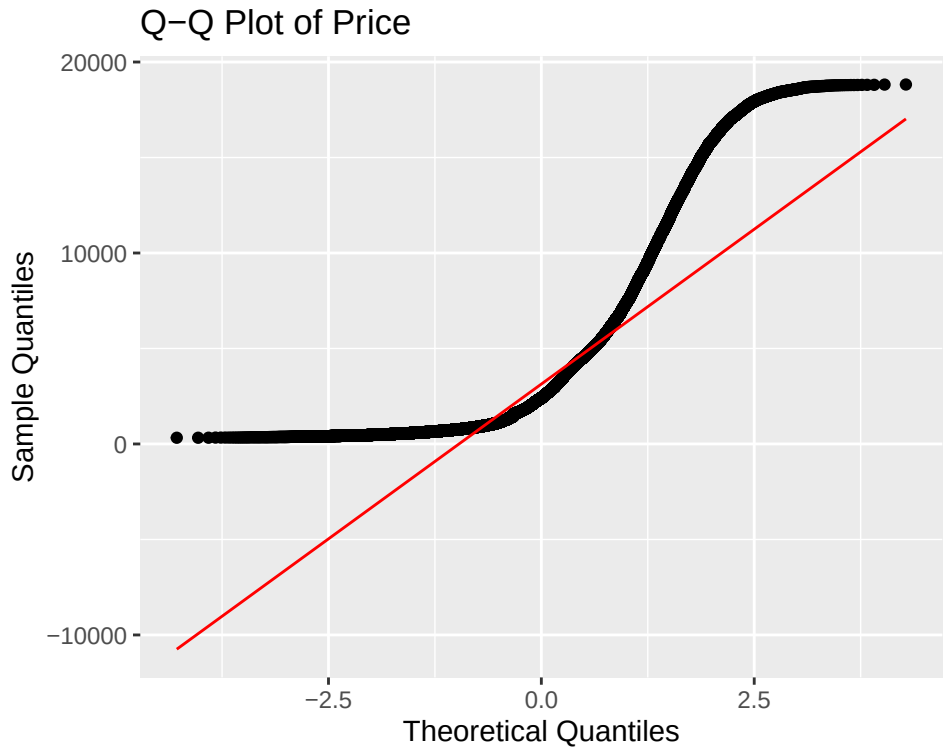
```
hist_plot = ggplot(data.frame(num_vars$price), aes(x = num_vars$price)) + geom_histogram(binwidth = 500

print(hist_plot)
```




```
density_plot = ggplot(data = data.frame(num_vars$price), aes(x = num_vars$price)) + geom_density(fill =

print(density_plot)
```



```r
qq_plot = ggplot(data = data.frame(num_vars$price), aes(sample = num_vars$price)) + stat_qq() + stat_qq_

print(qq_plot)
```




```r
box_plot = ggplot(data.frame(num_vars$price), aes(x='',y = num_vars$price)) + geom_boxplot(fill = "ligh

print(box_plot)
```

The histogram for the price of diamonds (in USD) shows an extremely right-skewed distribution. A higher amount of observations are concentrated in the lower range, and the frequency drops quickly as price increases, however there is a heavy right tail.

The density plot further confirms this skewness, illustrating a strong concentration of data at the lower price values and a long tail extending towards the higher priced values. In a normal distribution, the density plot would appear to be symmetrical about the center, the mean, and have a bell shape, with even tails. Furthermore, the area under the curve in the higher price values indicates the presence of potential high-value outliers.

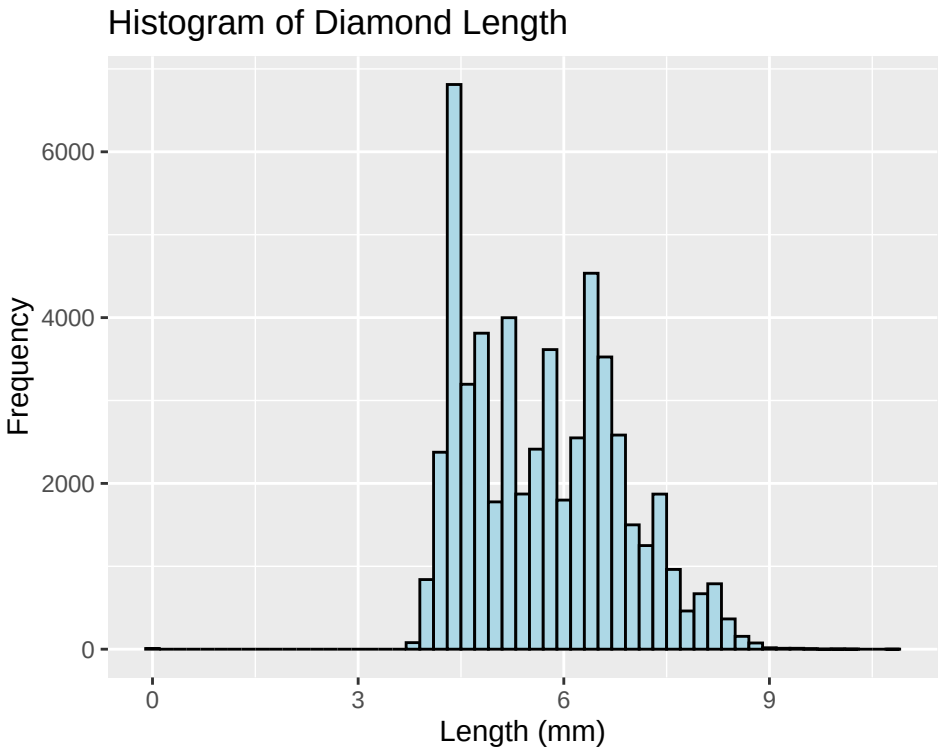
In the Q-Q plot, if the data were normally distributed, the points would appear to follow very closely to the red line, which is the theoretical normal distribution. However, it is clear that the data points deviate significantly from this line, furthering the idea that this variable does not follow a normal distribution. The upper tail, representing higher prices, shows extreme upward deviation, indicating there are observed prices that are greater than what would be expected in a normal distribution. The lower tail of the data points also appears to be compressed, which mirrors the strong concentration of data points in the lower price range.

Lastly, analyzing the boxplots gives a similar idea to the other plots. If the data were to follow a normal distribution, the box plot would appear to be symmetric, with whiskers that extend evenly. In the boxplot for variable price, the box, representing the interquartile range, appears to be compressed, mirroring the concentration of data in the lower price ranges. Similarly, the lower whisker is much smaller than the upper whisker, showing the differences in the spread of each quartile. The red points above the upper whisker represent the prices that lie far beyond the interquartile range, additionally illustrating the heavy right tail.

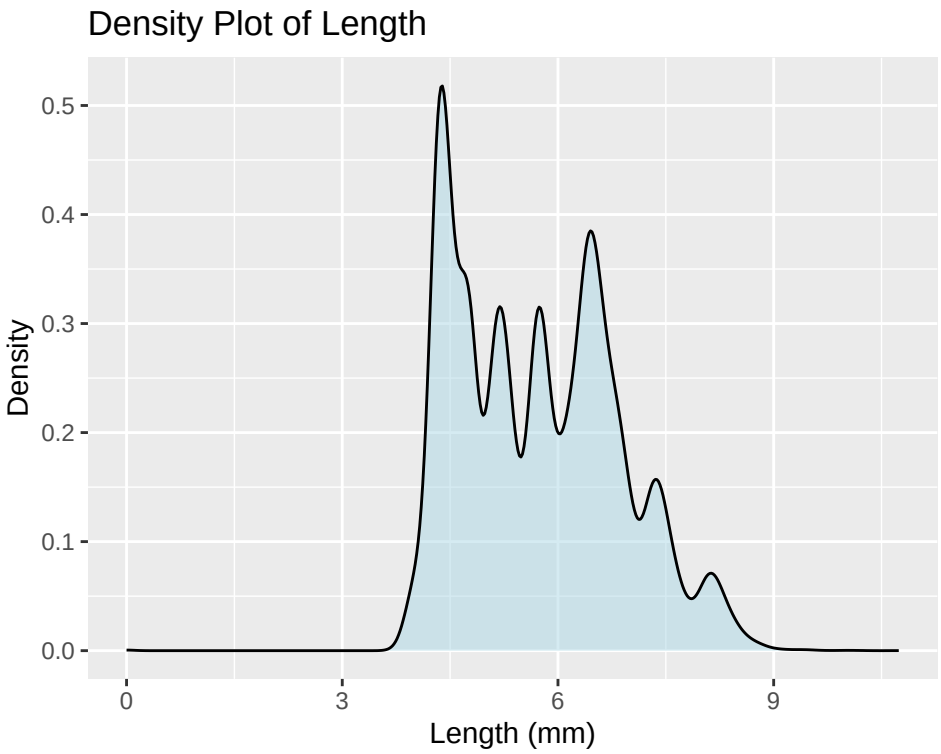
Across all four plots, there are consistent results that the variable price does not follow a normal distribution. The data exhibit a strong right skew, with a heavy concentration of prices in the lower price range and a long tail that extends towards higher prices.

X (Length of the diamond in mm)

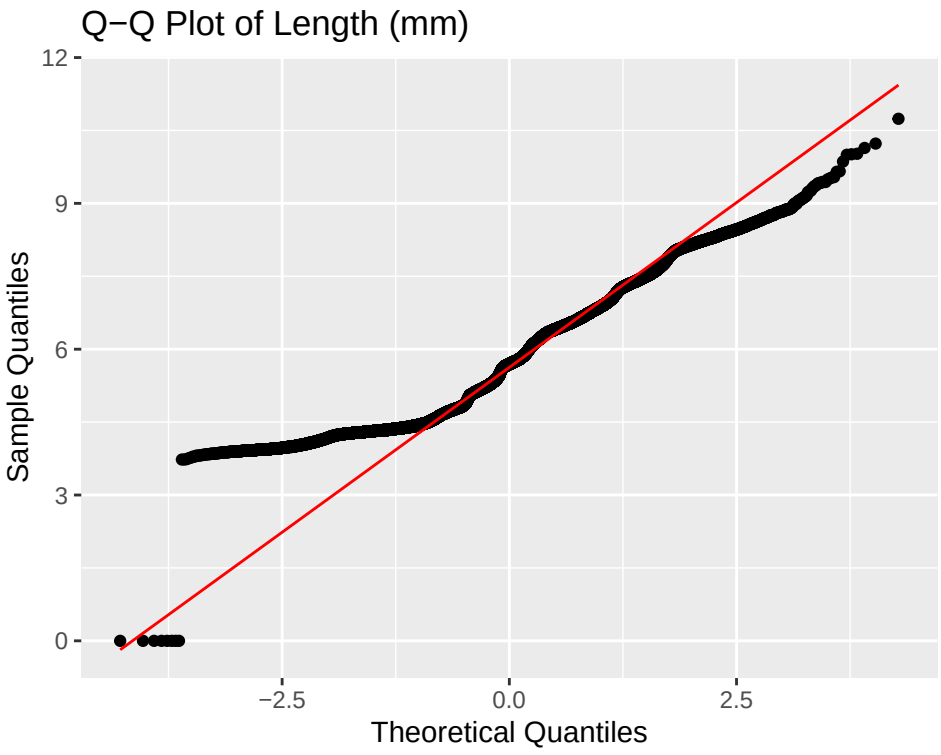
```r
hist_plot = ggplot(data.frame(num_vars$x), aes(x = num_vars$x)) + geom_histogram(binwidth = 0.2, fill =

print(hist_plot)
```



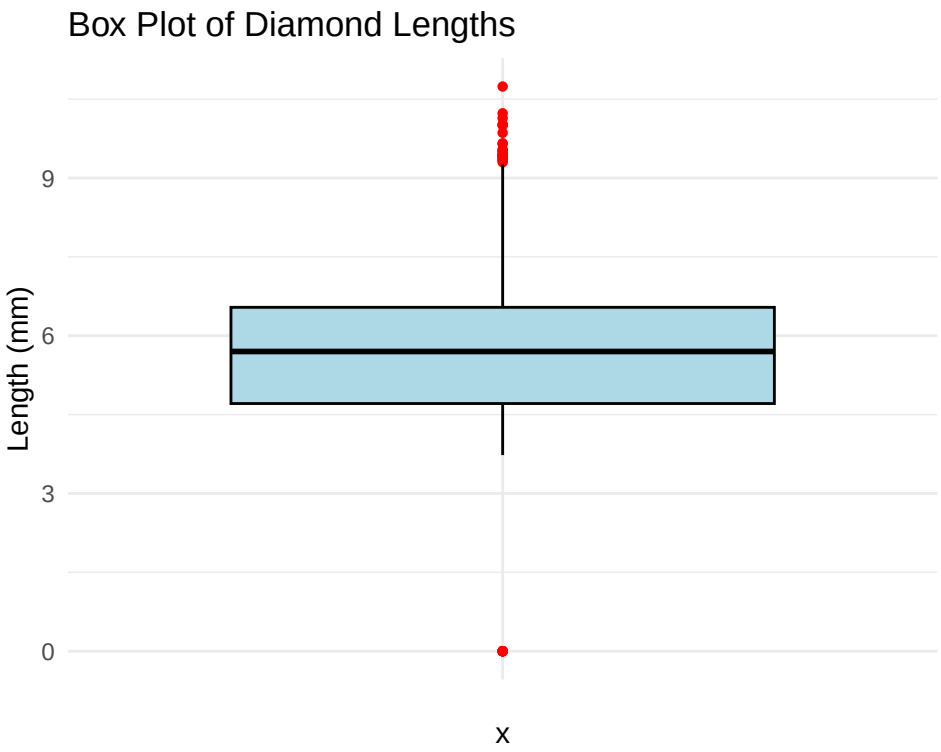

```r
density_plot = ggplot(data = data.frame(data <- num_vars$x ), aes(x = data <- num_vars$x )) + geom_dens

print(density_plot)
```




```
qq_plot = ggplot(data = data.frame(num_vars$x), aes(sample = num_vars$x)) + stat_qq() + stat_qq_line(co

print(qq_plot)
```



```
box_plot = ggplot(data.frame(num_vars$x), aes(x='',y = num_vars$x)) + geom_boxplot(fill = "lightblue",

print(box_plot)
```



The histogram illustrates a frequency distribution of diamond lengths. At first glance, it may appear to be normal, but the data have a multimodal tendency, with multiple peaks within the histogram. A histogram of a normal distribution would have a smooth, singular central peak with symmetric tails. Although there is a "central" peak around 5 to 6 mm, the strong peaks around 4.5 mm and 6.5 to 7 mm incite the idea of multimodality. Additionally, the left tail, indicating shorter lengths, appears shorter than the right tail, longer lengths, which could indicate mild right skewness and further deviating this data from a normal distribution.

The density plot, which smooths the histogram into a continuous probability distribution, highlights the multiple peaks and further confirms the multimodality. The peaks suggest that there are distinct groups or clusters of diamond lengths within the dataset. Furthermore, the density plot shows the tail on the right-hand side which stretches the data and indicates right skewness.

The Q-Q plot compares the quantiles of the sample data to that of a theoretical normal distribution. In this Q-Q plot, there are noticeable deviations from the line, particularly at the lower and higher tails. The points in the lower quartiles, indicating the shorter lengths, fall above the line, while the points in the upper quantiles, indicating the longer lengths, deviate below the line. This pattern indicates a deviation from normality, particularly that there is a stronger lower cluster than expected, and higher values that pull away from the central portion of the data. All of which are characteristics that are not consistent with a normal distribution.

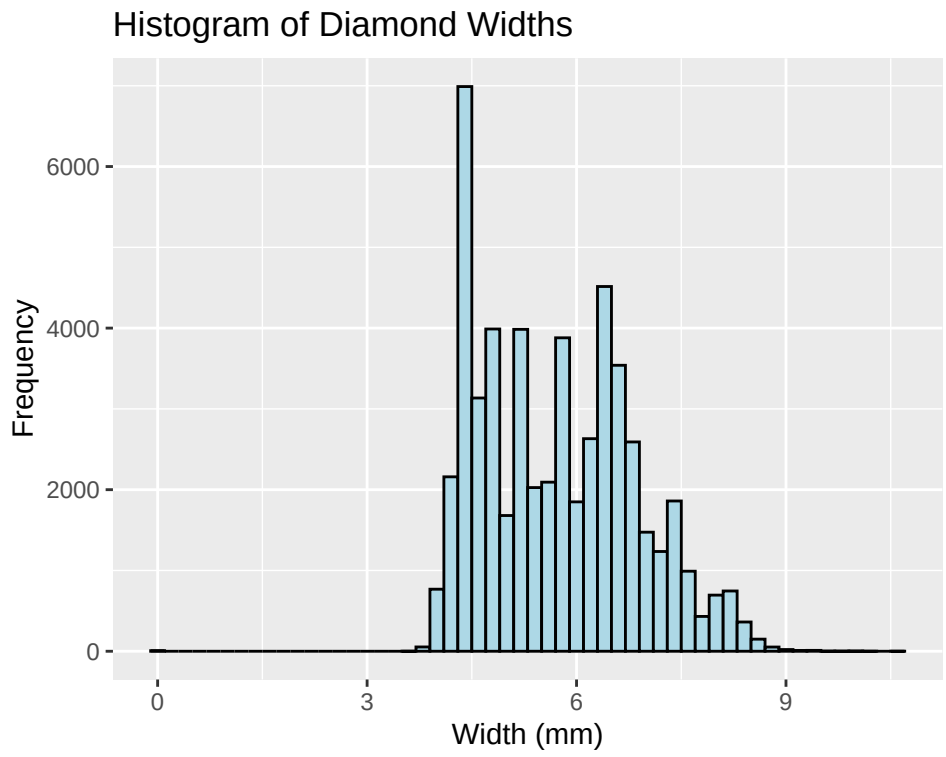
Finally, the box plot gives a visual summary of the distribution of diamond lengths within the quartiles. Although the boxplot appears to be symmetric, it is evident that the lower tail, indicating the shorter lengths, is significantly shorter than the upper tail, the longer lengths, furthering the conclusion that this data is skewed to the right. There is also clustering of outliers in both extremes, which does not align with the normal distribution, either.

To solidify the hypotheses of multimodality and right skewness, a violin plot can show both the central tendency and spread of the data. In the violin plot for variable x, the length of diamonds, there is a clear pattern that is not of a normal distribution. First, the data is the widest at the base and becomes narrower towards the top, except for a second, less wide, peak around 6.25. This plot is just another visualization to show the non-normal patterns of this variable.

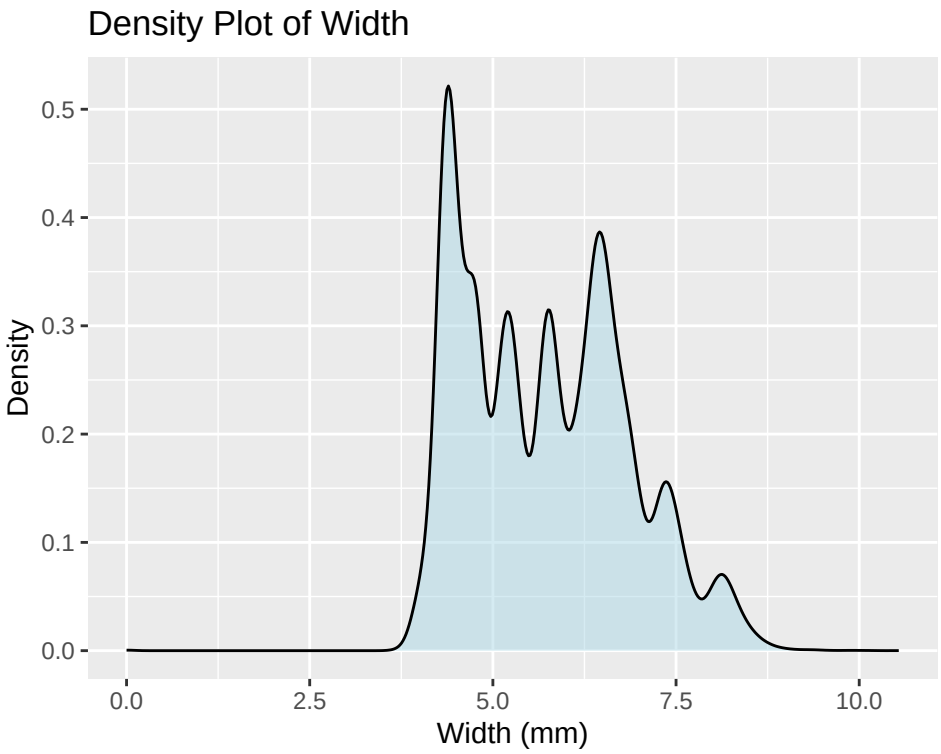
In conclusion, based on the visualizations provided, the variable length has slight multimodality and is skewed to the right. Although the variable length seems to have strong centrality, these results do not allow this variable to be assumed to be normally distributed.

Y: (Width of the diamond in mm)

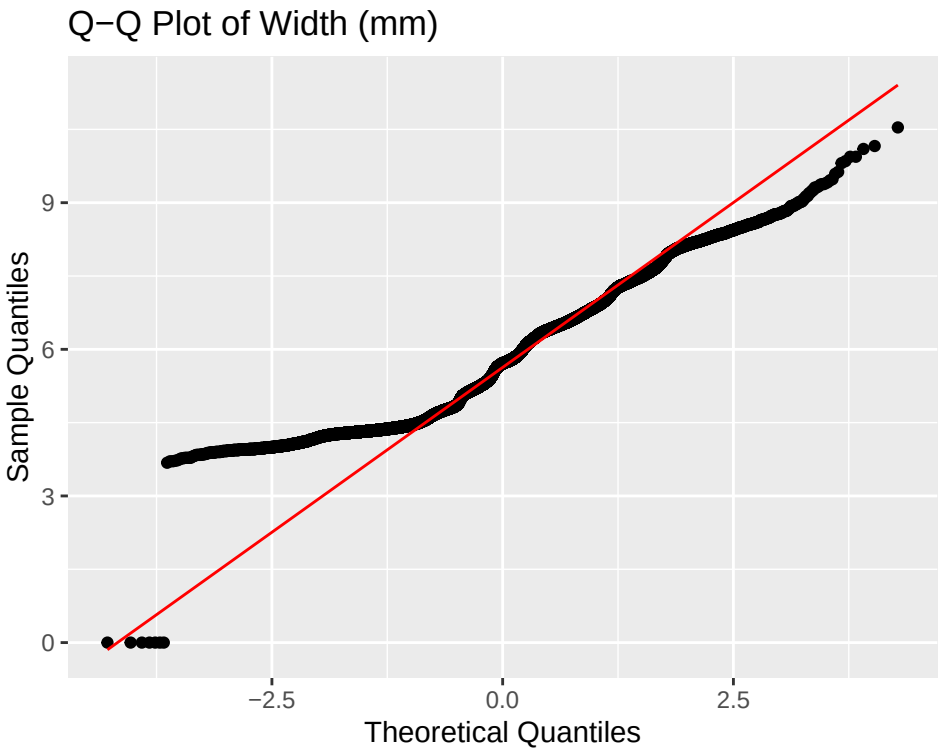
```
hist_plot = ggplot(data.frame(num_vars$y), aes(x = num_vars$y)) + geom_histogram(binwidth = 0.2, fill =

print(hist_plot)
```



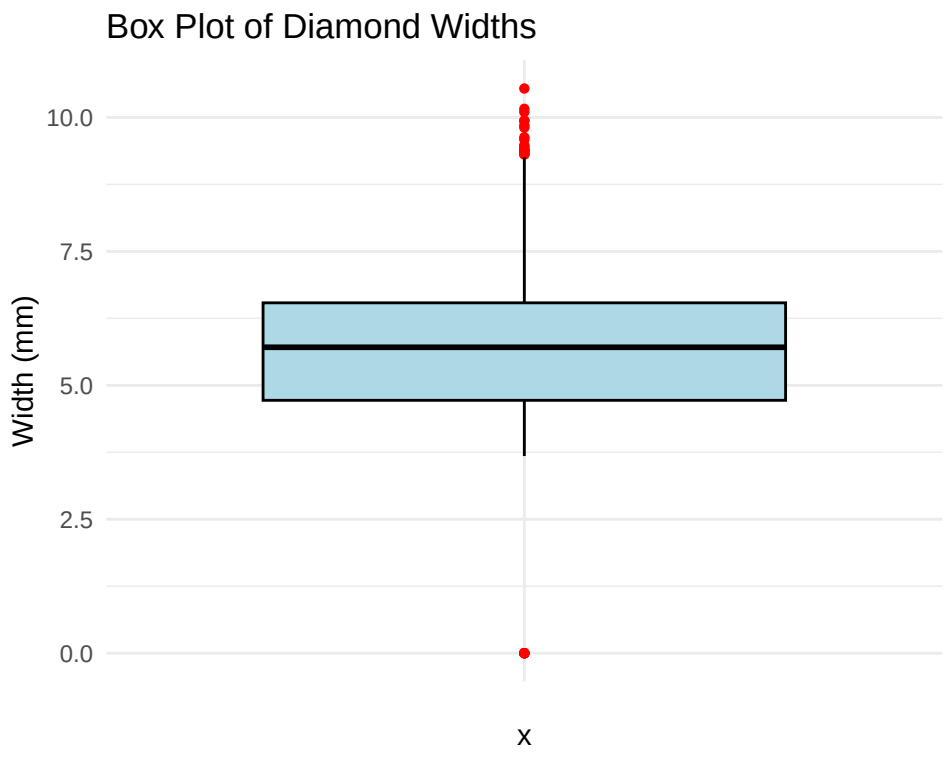

```
density_plot = ggplot(data = data.frame(num_vars$y ), aes(x = num_vars$y)) + geom_density(fill = "light

print(density_plot)
```



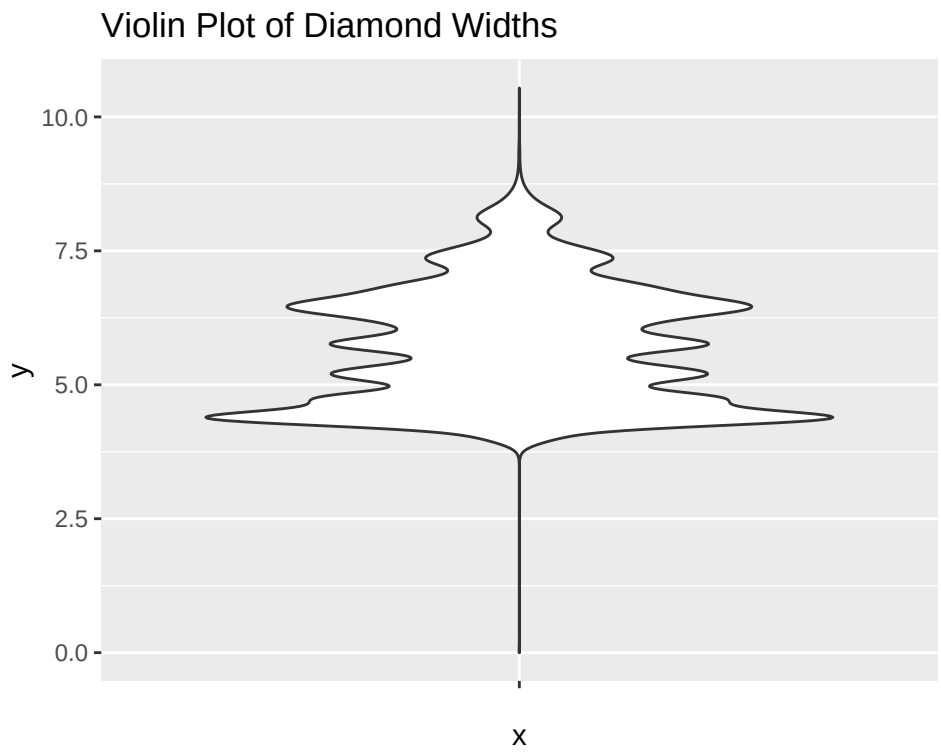

```
qq_plot = ggplot(data = data.frame(num_vars$y), aes(sample = num_vars$y)) + stat_qq() + stat_qq_line(co

print(qq_plot)
```



```r
box_plot = ggplot(data.frame(num_vars$y), aes(x='',y = num_vars$y)) + geom_boxplot(fill = "lightblue", 

print(box_plot)
```




```r
violin_plot =  ggplot(new_diamonds, aes(x = "", y=y))+    geom_violin() +
labs(title = "Violin Plot of Diamond Widths")

print(violin_plot)
```



The histogram representing the frequency distribution of diamond widths could initially give the impression of a normal distribution; however, a closer analysis reveals a multimodal structure with distinct peaks. Unlike a typical normal distribution, which would feature a smooth, singular central peak with symmetrical tails, this data exhibits strong peaks around 4.5 mm and 7 mm, indicating multiple modes. Furthermore, the asymmetry in the tails, with the left tail, shorter widths, being noticeably shorter than the right tail, longer widths, points to moderate right skewness. These results combined further underscore the data deviates from a normal distribution.

The density plot effectively highlights the multiple peaks and multimodal characteristics of the data. These peaks suggest the presence of distinct clusters of diamond widths within the dataset, indicating potential groupings. Additionally, the rightward-stretching tail observed in the density plot further insinuates the indication of right skewness, suggesting an asymmetrical distribution that may influence interpretations of the data.

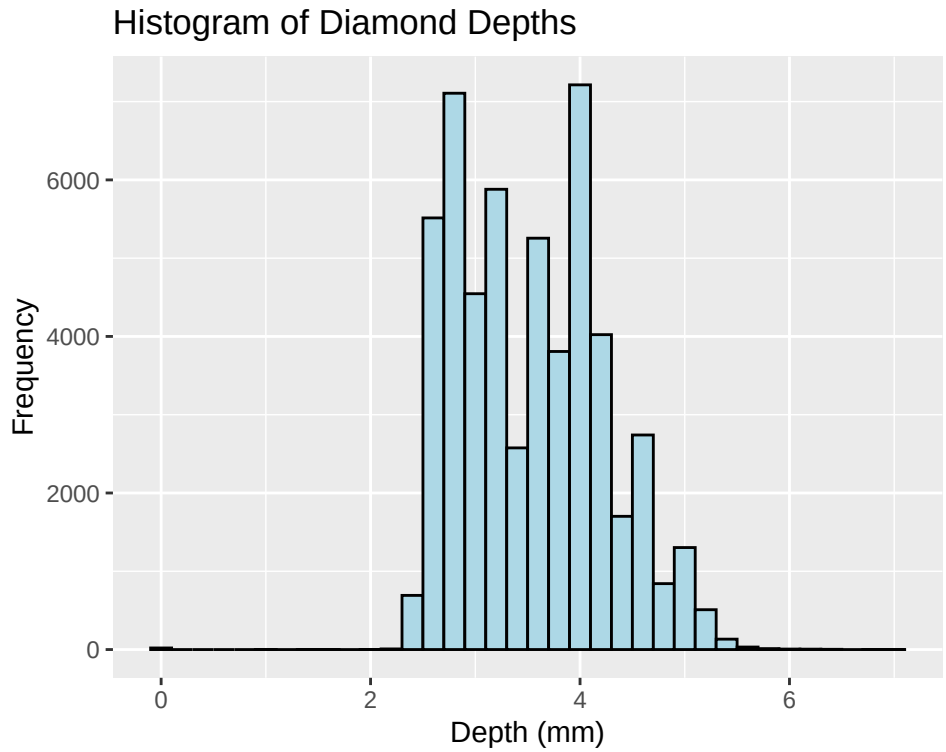
The Q-Q plot provides a comparative visualization of the quantiles of the dataset against those of a theoretical normal distribution. The Q-Q plot for variable width illustrates pronounced deviations from the expected line, particularly in the distribution's tails. In the lower quartiles, which correspond to shorter widths, the plotted points consistently fall above the diagonal line, indicating an abundance in lower values relative to what would be anticipated under normality. This could correspond with the strong peak in the lower region around 4.5 mm. Conversely, in the upper quartiles, representing longer widths, the points are positioned below the line, suggesting a deficiency of higher values. This observed pattern indicates a departure from normality, specifically indicating a large amount of data in lower quartiles while also a more pronounced right tail than would typically be expected. Again, these results combined reflect this variable is not normally distributed.

Lastly, the box plot provides an image of the spread of data within the quartiles. Similarly to the results of the length variable, the boxplot appears to be symmetrical in the central boxes, but it is clear the lower tail, shorter widths, is significantly smaller than the upper tail, greater widths. This illustrates that there is an uneven spread when comparing the lengths of quartiles. If the data were to be normally distributed, the quartiles would appear even in length. The strong amount of outliers in the extreme tails is another characteristic that does not align with normality.

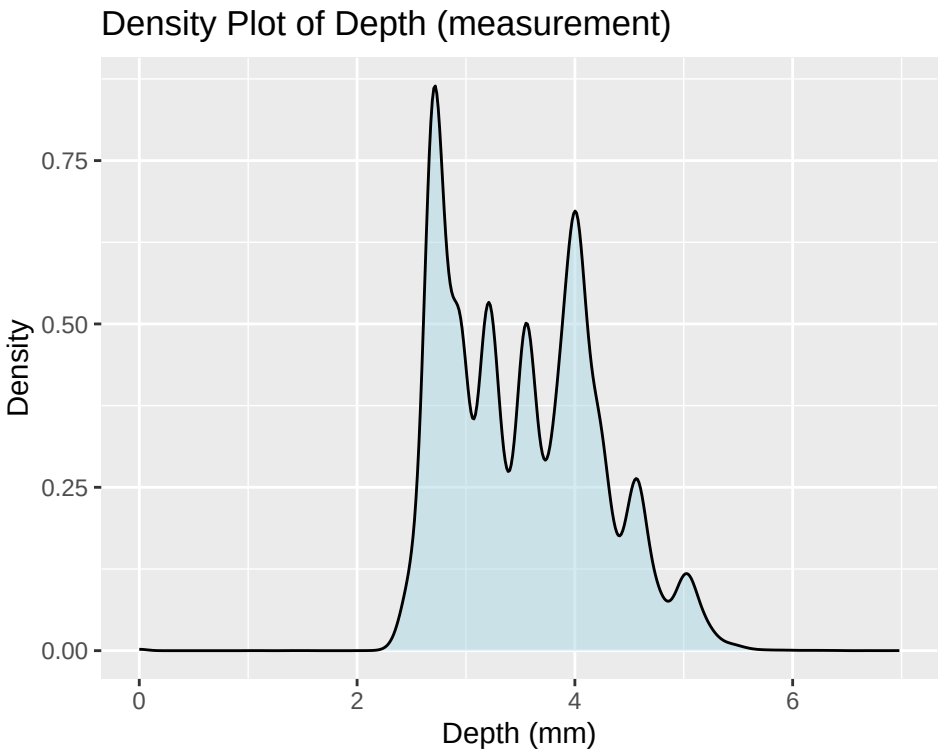
To further analyze multimodality and symmetry, a violin plot can be created. The violin plot for variable y, length of diamonds, clearly shows a non symmetric or unimodal shape. The first strong peak is shown around 4.75, followed by the distribution going closer to the center line, followed by another, yet shorter, peak at 6.25. The violin plot further confirms the hypothesis that this variable is not unimodal and right skewed. Compiling these results, there are several characteristics in the width variable that do not mimic that of a normally distributed variable.

Z (Depth of the diamond in mm)

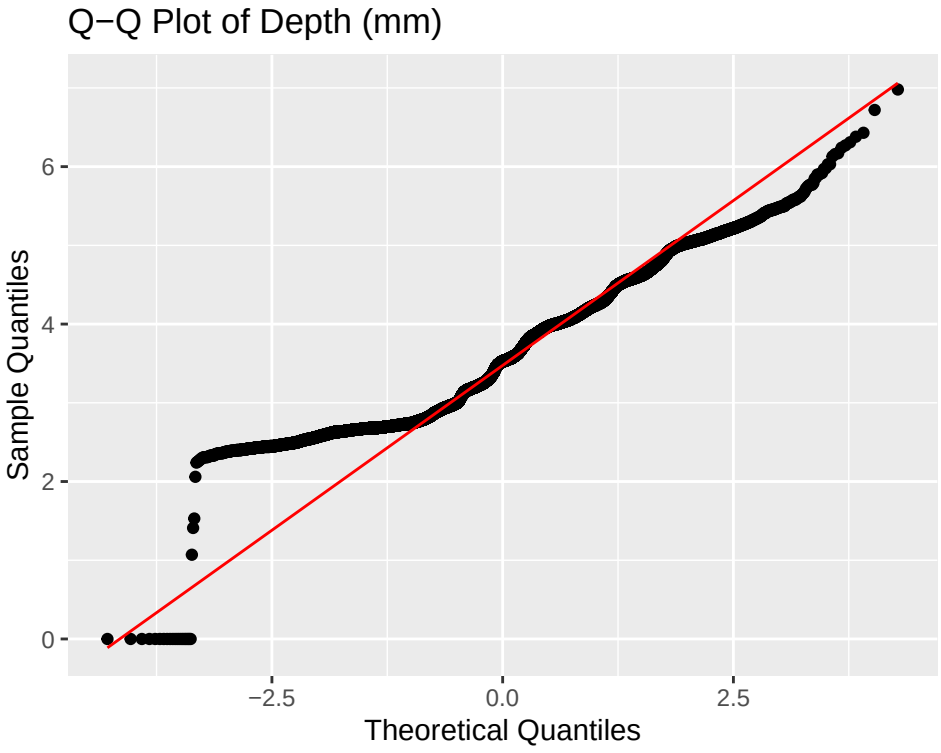
```
hist_plot = ggplot(data.frame(num_vars$z), aes(x = num_vars$z)) + geom_histogram(binwidth = 0.2, fill =

print(hist_plot)
```



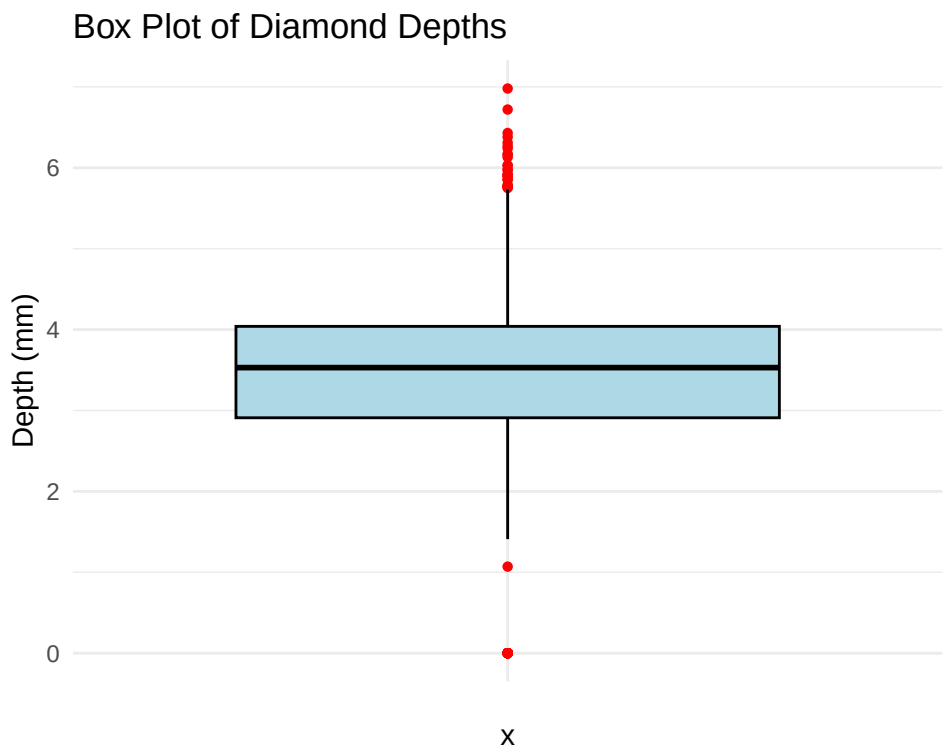

```
density_plot = ggplot(data = data.frame(num_vars$z), aes(x = num_vars$z)) + geom_density(fill = "lightb

print(density_plot)
```




```
qq_plot = ggplot(data = data.frame(num_vars$z), aes(sample = num_vars$z)) + stat_qq() + stat_qq_line(co

print(qq_plot)
```



```
box_plot = ggplot(data.frame(num_vars$z), aes(x='',y = num_vars$z)) + geom_boxplot(fill = "lightblue",

print(box_plot)
```



The histogram of the variable diamond depth shows a general bell shaped curve, which could suggest underlying normality. However, there are slight deviations from a perfectly symmetric distribution. First, there appears to be strong centrality with quick "drop-offs" in the extreme ends. The right tail also appears to have more spread than the left tail, giving the slight assumption that this distribution could be right skewed. This result indicates that, while the majority of depths are clustered around the mean, there are extreme depths that occur less frequently.

The density plot reinforces findings from the histogram, but with some newer insights. The density plot highlights a potential multi-modality, with a strong peak around 2.5 to 3 mm and another around 4 mm. A normal distribution is unimodal, having only one distinct peak. Likewise, the strong peak on the lower range of the data raises concerns that the data is right skewed. Furthermore, the density plot shows the right tail has a more prominent shape than the left, which also goes against characteristics of a symmetric distribution.

The Q-Q plot compares the quantiles of the observed depth against the quantiles of a theoretical normal distribution. While the central points appear to follow the normal line, there are strong deviations at either end. Starting in the lower quantile, the concentration of points above the line indicates there is an abundance of lower values that would not be expected in a normal distribution. This appears to imitate the behavior of the data at the strong left-most peak around 2.5 to 3 mm. On the other hand, the concentration of points that fall below the line in the upper quantile represents the slight lack of higher values, which would make the distribution symmetric, most likely attributing this to the strong peak as discussed earlier. The group of points at the left-most quartile represent the deficiency of lower values which would balance out the distribution with the extreme right tail values.

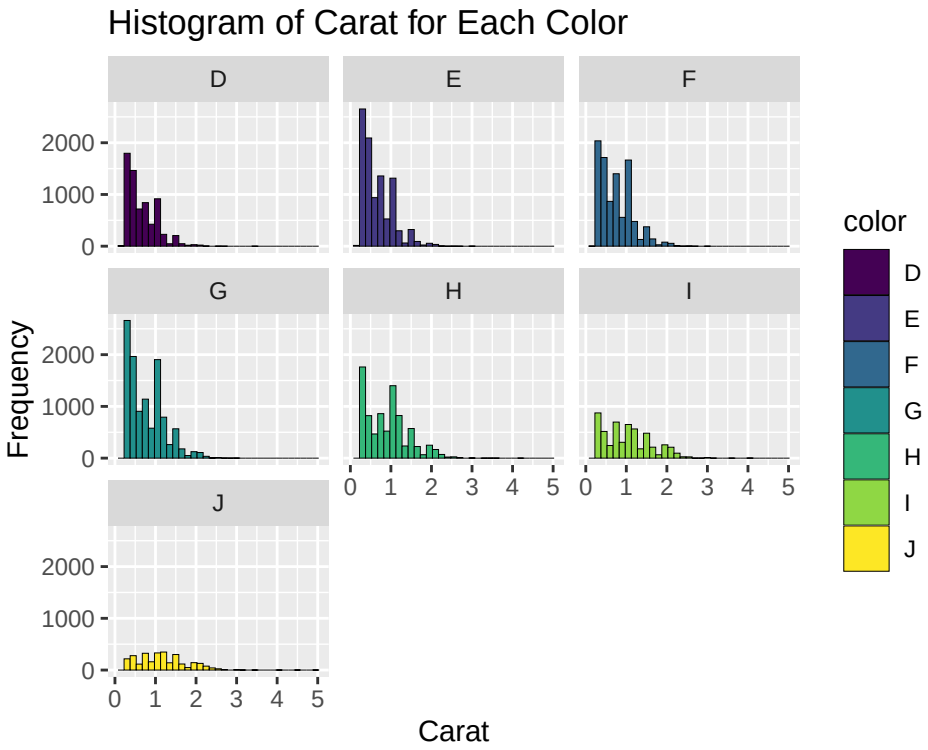
The box plot of variable depth shows an almost symmetric spread within the central quartiles. However, the tails show the large spread in the first and fourth quartiles towards both extremes, highlighting the uneven

distribution and lack of symmetry. The strong concentration of outliers on the upper end indicates the extreme large depth measurements, which further result in a right skewed distribution. With these results combined, the plots demonstrate that the variable depth does not follow a normal distribution, but rather a slightly multi-modal, right skewed distribution.

v) Assessment of Normality of the Continuous Variables Over Each Color Group

Carat (The weight of the diamond)

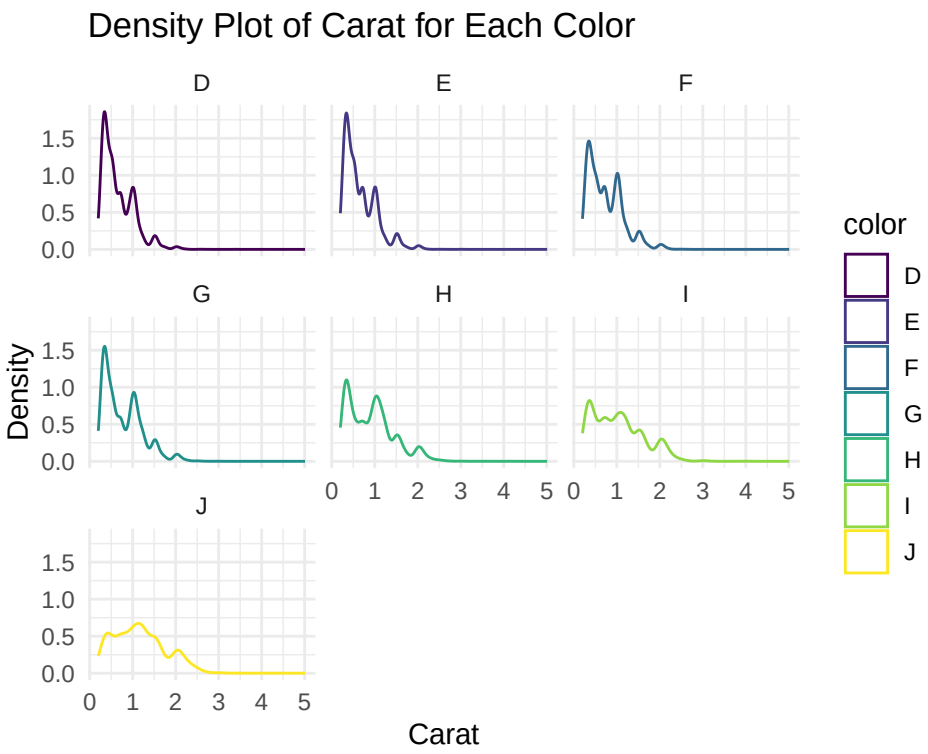
```
#Histogram
hist_plot = ggplot(new_diamonds, aes(x = carat, fill=color)) +
  geom_histogram(binwidth = 0.15, color = "black",linewidth=0.1) +
  facet_wrap(~ color) +
  ggtitle("Histogram of Carat for Each Color") +
  xlab("Carat") +
  ylab("Frequency")

print(hist_plot)
```



The histogram depicting the distribution of diamond carat weights across different color grades reveals intriguing insights into the frequency of diamonds available in each category. For color D, there is a notable concentration of diamonds with carat weights below 0.5. This suggests that lighter diamonds are more prevalent in this color grade. When comparing color grades D, E, and F, a consistent left-shifted distribution emerges, with pronounced peaks in carat weights ranging from 0.2 to 0.6. In contrast, for colors G and H, although the distribution also shifts to the left between 0.2 and 1 carat, the peak frequency is observed at the lower carat weights of 0.2 and 0.4. Color I exhibits a broader range of distribution, indicating a less pronounced left peak. Here, frequencies are more evenly distributed across carats from 0.2 to 1.6, with distinct peaks observed in the 0.2–0.4 and 1.0–1.2 carat ranges. This suggests that color I is perceived as a

more luxurious option, showcasing a diversity of carat weights despite a smaller overall inventory. Diamonds with color J are more evenly spread across the carat range, with a concentration between 0.8 and 1.5 carats. This results in a more uniform distribution of data points. Overall, the analysis reveals that the majority of diamonds fall within the 0.2 to 0.8 carat range, underscoring this segment's prominence in the data given. In the context of this dataset, when examining the normality through the charts provided, it becomes evident that the data exhibits an asymmetric, or skewed, distribution. Specifically, the data is right-skewed, meaning that there is a long tail extending to the right while the majority of the values cluster towards the left side. This indicates that there are fewer occurrences of higher carat diamonds, but these rare, larger diamonds have an influence on the distribution and have a greater spread. The presence of peaks in lower ranges indicates these values are more prevalent in the market.

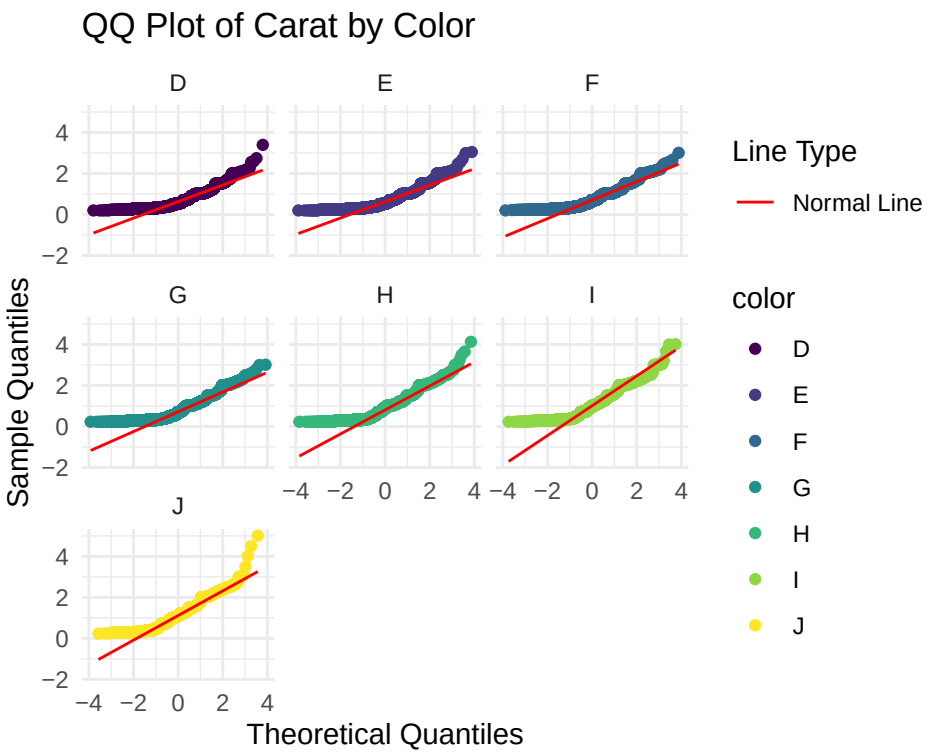
```
#Density plot
density_plot = ggplot(new_diamonds, aes(x = carat)) +
geom_density(aes(color = color), alpha = 0, adjust = 1.2) + # Set alpha to 0 for no fill
facet_wrap(~ color) +
ggtitle("Density Plot of Carat for Each Color") +
xlab("Carat") +
ylab("Density") +
theme_minimal()

print(density_plot)
```



The density plots provide valuable insights into the distribution of carat weights for diamonds, categorized by color, and allow for a clear comparison across different color grades. For colors D through H, two prominent peaks are observed: one between 0 to 0.5 carats and another at exactly 1 carat, with a common mode around 0.3 carats, indicating this is the most frequently occurring weight for these colors. In contrast, the density plots for colors I and J display more dispersed distributions. Specifically, the plot for color J shows a flatter curve, indicating a wider range of carat weights and less concentration around specific values. This suggests that diamonds in these colors are more varied in size, while colors D through H exhibit a preference for

lighter diamonds. The asymmetry in the plots across all color categories confirms the patterns noted in the histograms, showing a rightward skew with long tails, particularly for higher carat values. This suggests the presence of larger, rarer diamonds. Additionally, the peaks for colors D to G highlight popular weight ranges, but the multiple peaks observed across the data indicate that the distributions are not unimodal, suggesting the existence of subgroups within the color categories. Colors H, I, and J show flatter distributions with lower densities as carat weights increase, indicating fewer heavy diamonds in these groups. These findings reinforce the idea that the data is heavily skewed and non-normal.

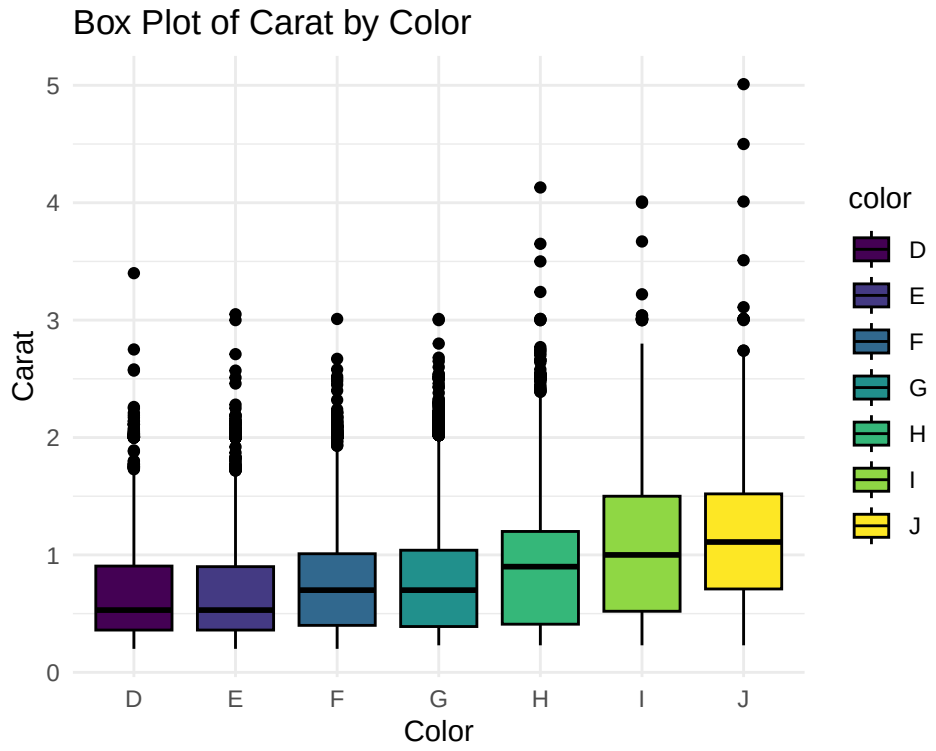
```
# QQ plot
qq_plot = ggplot(new_diamonds, aes(sample = carat, color = color)) +
  geom_qq(distribution = qnorm) +
  geom_qq_line(aes(linetype = "Normal Line"), color = "red") +
  facet_wrap(~ color) +
  labs(title = "QQ Plot of Carat by Color",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  scale_linetype_manual(values = "solid", name = "Line Type", labels = c("Normal Line")) +
  theme_minimal()

print(qq_plot)
```



For a normal distribution, the expected data points will closely align along the straight line. However, the data points for all color categories deviate substantially from the diagonal line that would represent a perfectly normal distribution. In particular, the points at the extreme upper and lower quartiles show the greatest deviation, suggesting that the tails of the distribution are heavier than what would be expected in a normal distribution. This indicates that the carat sizes for diamonds in all color categories exhibit significant non-normality. The separation of the left tails from the normal line also implies that there are more lower carat weights in the dataset than would be anticipated if the data followed a normal distribution. For color J, as previously observed, there is a notable presence of outliers at the right tail of the distribution. The

existence of these outliers may indicate the presence of particularly rare or high-quality diamonds between the color J. The patterns seen in the QQ plot validate the conclusion that the distribution of carat sizes across all color categories is non-normal.

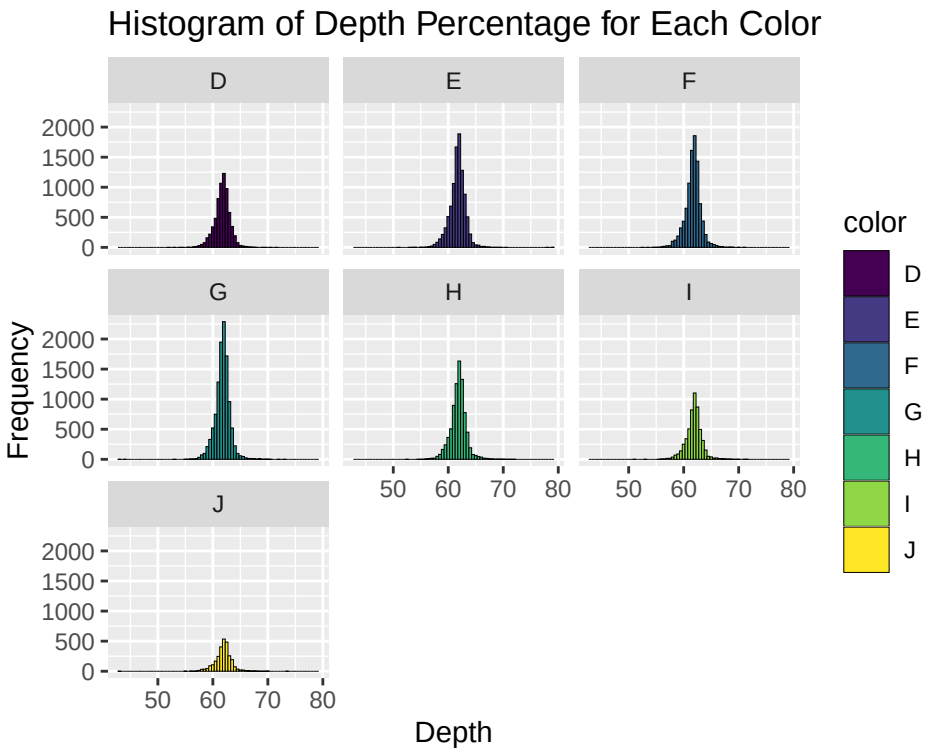
```
#Boxplot
box_plot = ggplot(new_diamonds, aes(x = color, y = carat, fill=color)) +
  geom_boxplot(color = "black") +
  ggtitle("Box Plot of Carat by Color") +
  xlab("Color") +
  ylab("Carat") +
  theme_minimal()

print(box_plot)
```



In the boxplots for the carat variable, Colors D and E exhibit similar wide frequency distributions, with the interquartile range (IQR) showing that 50% of diamonds in these categories have carat weights between 0.4 and 0.9. The medians are closer to the first quartile, indicating right skewness. Colors F and G show a trend towards higher carat weights, with the central 50% ranging from 0.5 to 1.2 carats. The longer upper whiskers further suggest a right-skewed distribution. Color H's median is closer to the third quartile at about 1.3 carats, with more pronounced outliers, indicating a greater spread in this category and a tendency for higher-than-average carat weights. Color J stands out with the highest carat values, a median of 1.11 carats, and 50% of diamonds falling between 0.8 and 1.5 carats. Significant outliers in this category, including diamonds up to 5 carats, reflect its prestige and rarity. Across all color categories, the upper 25% of values display considerable variability, with outliers that can triple the third quartile values. This highlights the asymmetry in carat distributions, with wider boxes showing a large dispersion between the first and third quartiles. These findings indicate a skewed, non-normal distribution of carat weights across all categories.

Depth (Total depth percentage)

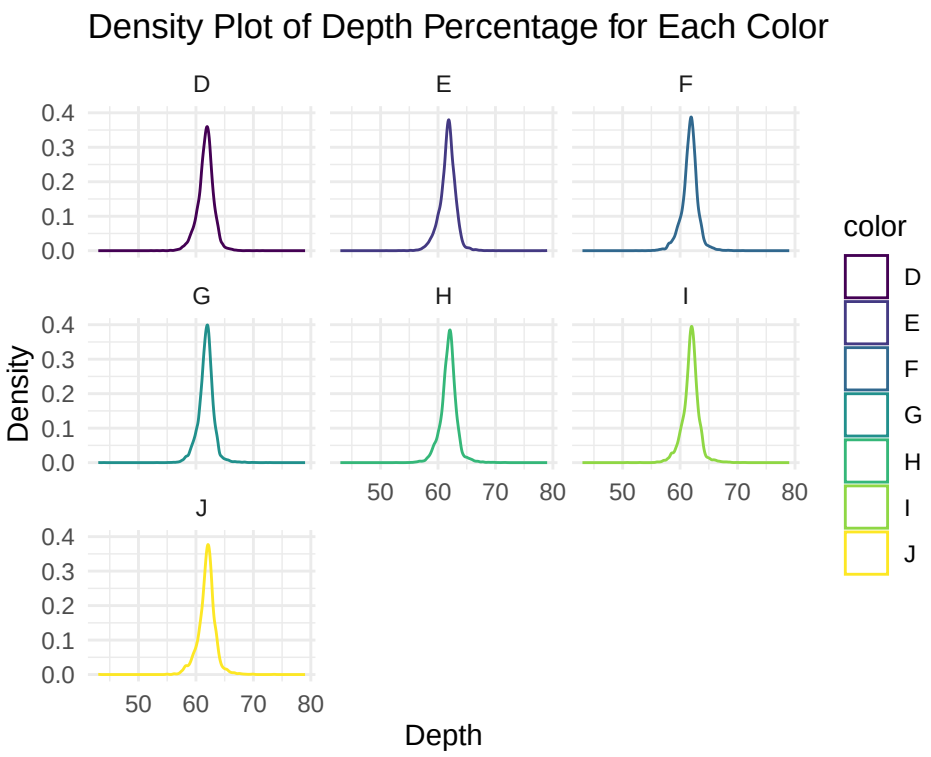
```r
#Histogram
hist_plot = ggplot(new_diamonds, aes(x = depth, fill = color)) +
  geom_histogram(binwidth = 0.5, color = "black", linewidth = 0.1) +
  facet_wrap(~ color) +
  ggtitle("Histogram of Depth Percentage for Each Color") +
  xlab("Depth") +
  ylab("Frequency")

print(hist_plot)
```



Histograms created for the different diamond colors provide insight into the distribution of the depth variable. For all colors, the concentration of values begins to increase around a depth of 60 and decreases smoothly, without sharp drops. When using narrower bin widths in the histograms, the normal distribution tendency becomes clearer across the depth of the diamonds. As seen in the histograms, the peaks of the depth values show an approximately normal distribution for all colors, meaning the values are symmetrically distributed. Even though color J has fewer diamonds, its values still concentrate around the middle range, between 60 and 65. The depth values for diamonds of all colors cluster in the center and gradually taper off at the tails, creating a symmetric curve. The thin tails indicate that extreme high or low values are less frequent within this range. Overall, the depth distribution does not vary significantly between colors, showing a normal distribution with a shared mean between 61.7 and 61.9. We can conclude that depth is not an indicator of diamond color.

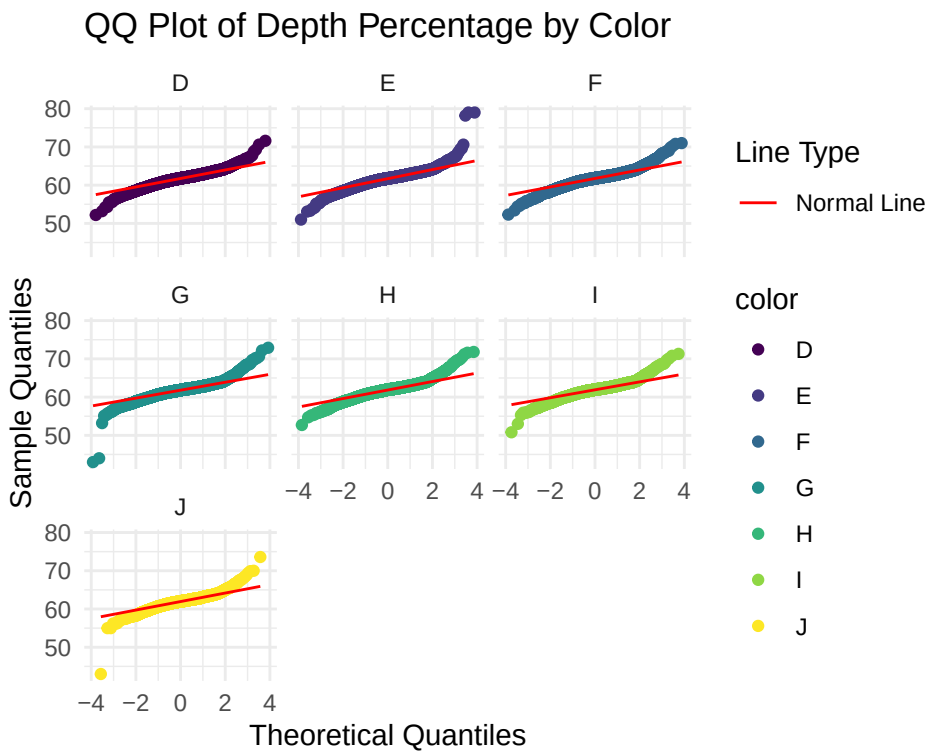
```r
# Density Plot
density_plot = ggplot(new_diamonds, aes(x = depth)) +
  geom_density(aes(color = color), alpha = 0, adjust = 1.2) + # Set alpha to 0 for no fill
  facet_wrap(~ color) +
  ggtitle("Density Plot of Depth Percentage for Each Color") +
  xlab("Depth") +
  ylab("Density") +
```

```
  theme_minimal()

print(density_plot)
```



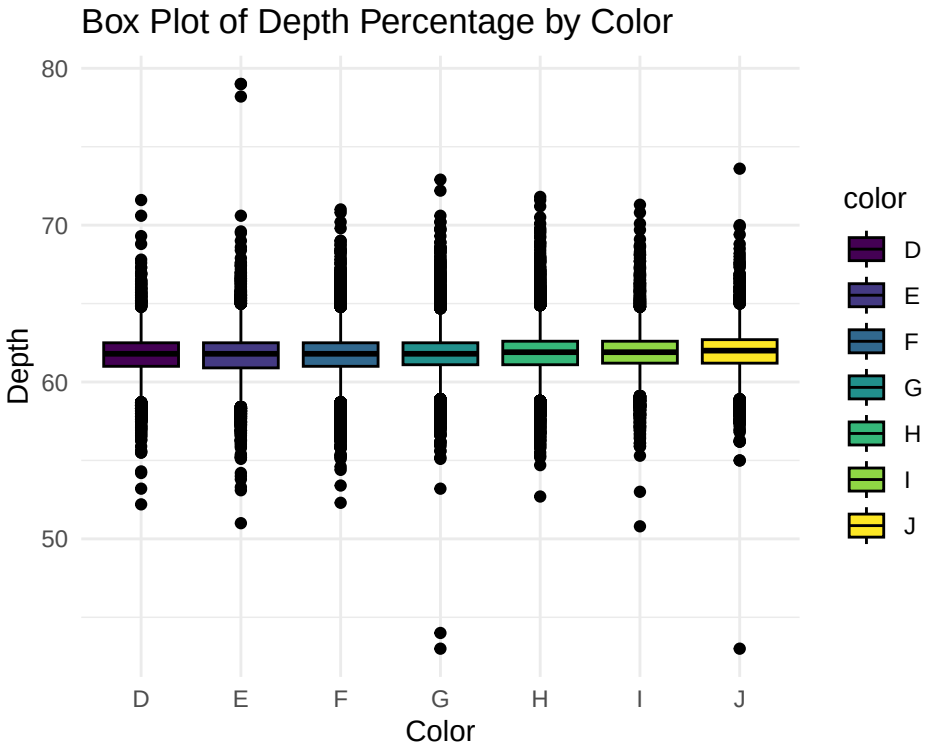
The density plots for the variable depth further support the normal distribution of depth for diamonds categorized by color. The symmetrical bell shape in the plots, with a single peak around the median values, reinforces this observation. Most diamonds have a depth between 60 and 64, with 87% of the diamonds falling within the range of 59.9 to 64.2. When splitting the data by color, there are no significant differences. As shown in the histograms, depth does not vary across the different colors, and the normal distribution remains, with thin tails and no signs of skewness.

```
#QQ Plot
qq_plot = ggplot(new_diamonds, aes(sample = depth, color = color)) +
  geom_qq(distribution = qnorm) +
  geom_qq_line(aes(linetype = "Normal Line"), color = "red") +
  facet_wrap(~ color) +
  labs(title = "QQ Plot of Depth Percentage by Color",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  scale_linetype_manual(values = "solid", name = "Line Type", labels = c("Normal Line")) +
  theme_minimal()

print(qq_plot)
```



As previously discussed, data points from a normal distribution will fall along the normal QQ line with little to no pattern or deviation. The depth values near the center of the distribution fall along this line, reinforcing a central normality. Across the depth variables, since the mean depth does not differ by color, the distribution appears to be normal for each color. However, in the QQ plot for each color, the right tails deviate slightly from the straight line, suggesting some probability of extreme values that do not follow a normal distribution. Despite these outliers, they do not alter the overall normality of the depth variable in the dataset. Outliers for each color category (from D to J) appear above the median in the depth range of 68 to 74, which could challenge the assumption of normal distribution. However, since other plots support the normality, these outliers, common in large datasets, do not change our conclusion.

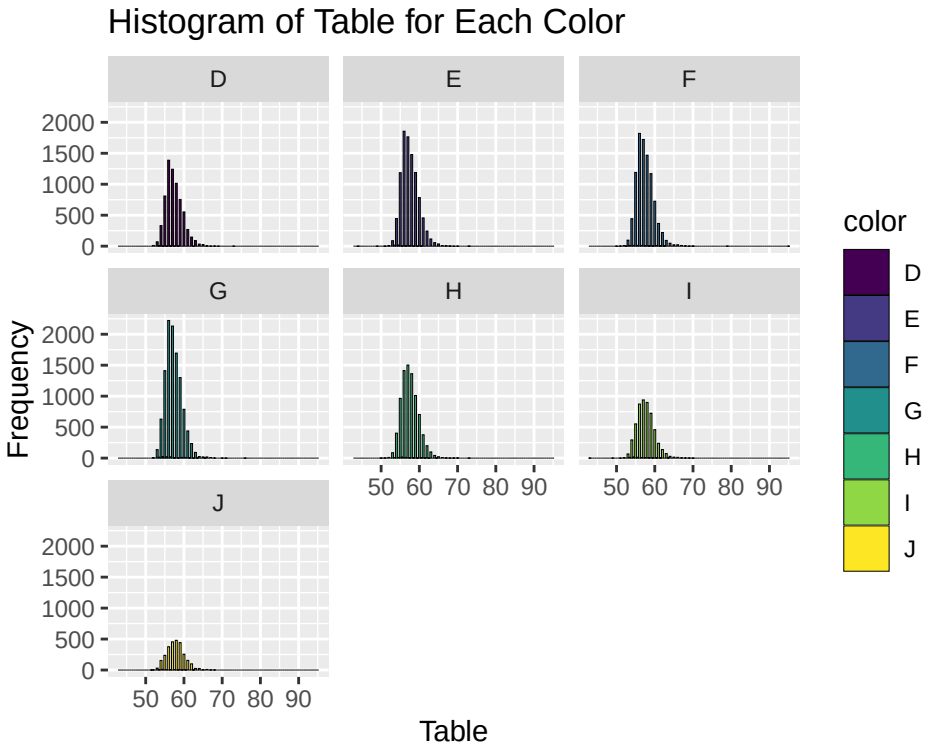
```
#Boxplot
box_plot = ggplot(new_diamonds, aes(x = color, y = depth, fill=color)) +
  geom_boxplot(color = "black") +
  ggtitle("Box Plot of Depth Percentage by Color") +
  xlab("Color") +
  ylab("Depth") +
  theme_minimal()

print(box_plot)
```



The boxplot for depth values reinforces the distribution indicated by the frequency analysis, showing that 50% of the data for all color categories falls within the range of 61 to 63. On the lower side of the median, colors D, E, F, and H have outliers between depths of 50 and 54, while colors G to J show more widely dispersed outliers ranging from 42 to 56. This consistency suggests a uniformity in depth characteristics, meaning that diamond cut proportions are relatively similar regardless of color. However, it's important to note that colors E, G, and J show more outliers, which may represent diamonds with significantly deeper or shallower proportions than the majority. These outliers could indicate unique cutting styles or qualities that deviate from the norm. While the boxplots for depth percentage show early signs of a normal distribution, the presence of outliers—common in large datasets—does not significantly change the overall distribution for this variable.

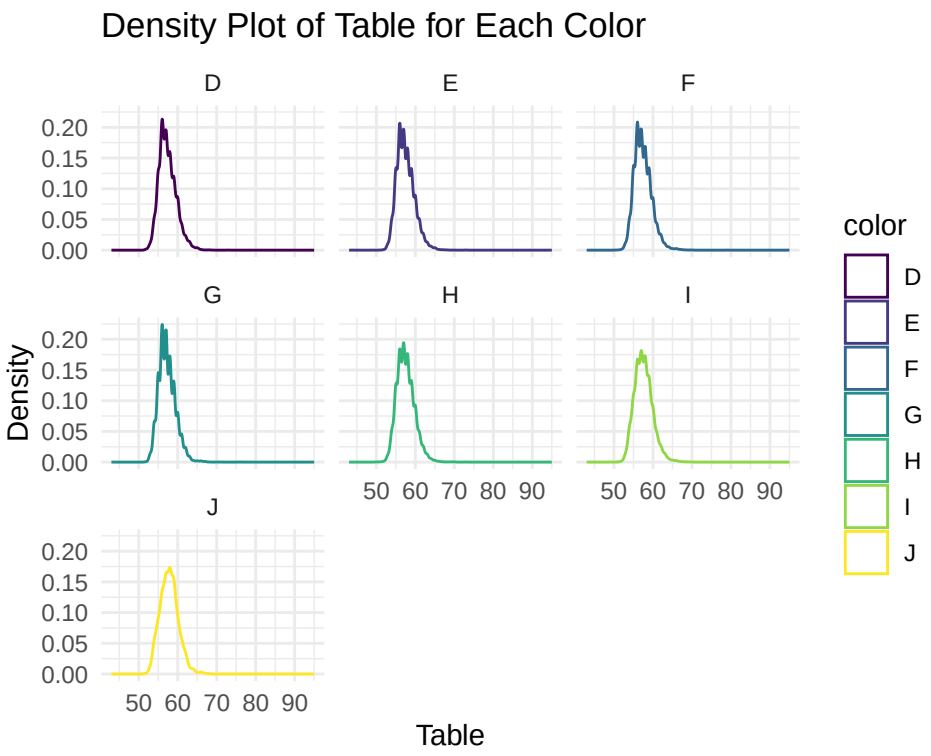
Table (Width of the top of the diamond relative to the widest point)

```
# Histogram
hist_plot = ggplot(new_diamonds, aes(x = table, fill = color)) +
  geom_histogram(binwidth = 0.5, color = "black", linewidth = 0.1) +
  facet_wrap(~ color) +
  ggtitle("Histogram of Table for Each Color") +
  xlab("Table") +
  ylab("Frequency")

print(hist_plot)
```



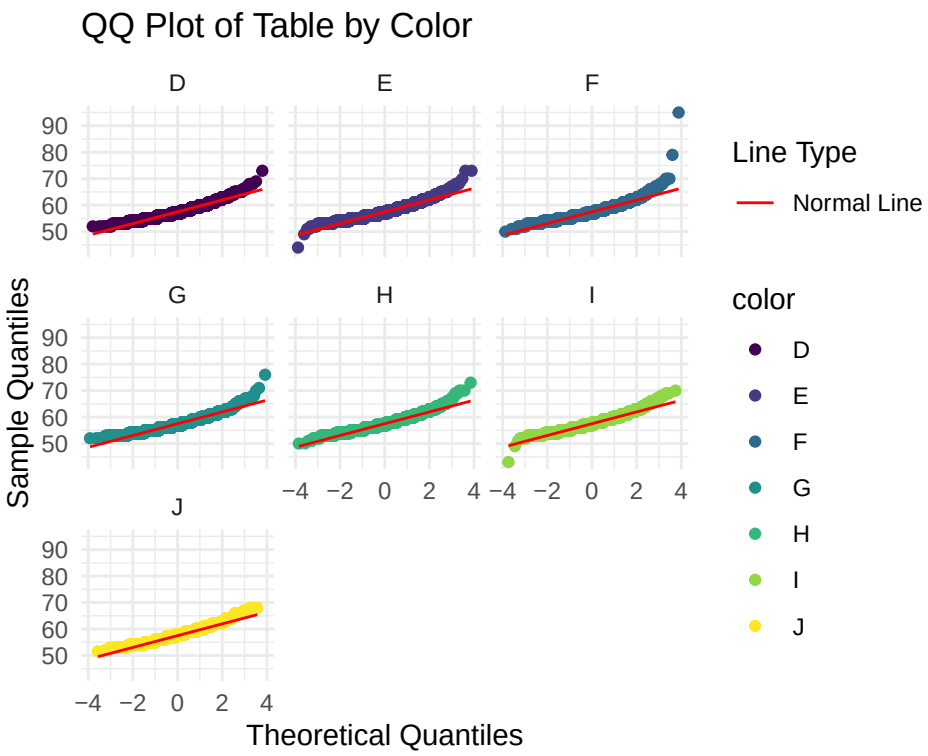
The table variable, which represents the width of the top of the diamond relative to its widest point, offers critical insights into the diamond's overall proportions. The histogram of the distribution of table measurements across different colors validates the concentration in the range between 55 and 60. For each color, starting with D and J, we see the most significant peaks at 54-57, with 30 and 49 diamonds, respectively, in those ranges. From colors G to I, the peak values tend to concentrate between 57-60, as previously observed. For colors E and F, which have the largest number of diamonds (18% and 17% of the total data, respectively), there is little variation, with 45% of their values falling within the 57-60 range. While the histograms for the "table" variable might initially suggest a more normal, bell-shaped distribution, this can be misleading when analyzing normality. The histograms show a concentration of values around a central point, which may give the impression of a Gaussian-like curve, with values decreasing symmetrically after the peak. However, closer inspection reveals a slight rightward shift in all color categories, indicating a subtle right skew in the distribution. This shift points to a longer right tail, suggesting that higher table values occur more frequently than in a true normal distribution. The apparent cohesiveness of the peaks in the histograms may obscure the underlying non-normality observed in other visualizations, such as the box plots. In the box plots, clear signs of asymmetry and skewness were noted. To rule out a normal distribution, further analysis of the other plots for the "table" variable is necessary.

```
#Density plot
density_plot = ggplot(new_diamonds, aes(x = table)) +
geom_density(aes(color = color), alpha = 0, adjust = 1.2) + # Set alpha to 0 for no fill
facet_wrap(~ color) +
ggtitle("Density Plot of Table for Each Color") +
xlab("Table") +
ylab("Density") +
theme_minimal()

print(density_plot)
```



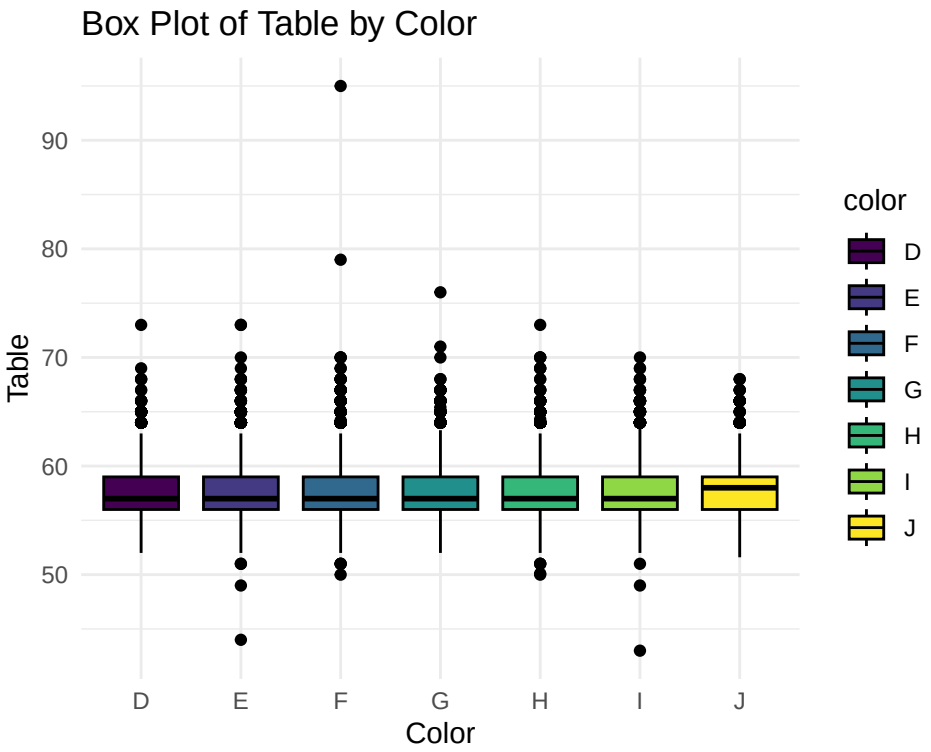
The density plots for each color category reveal the presence of multiple underlying distributions within the table variable, showing complexities not immediately apparent in the histograms, which capture broader frequency ranges. However, in the density plot, we can observe how concentrations may vary by as little as 0.5% in the table variable. The observed fluctuations in the density plots can be attributed to the choice of the smoothing parameter (bandwidth), which affects data visualization. A smaller bandwidth can reveal more intricate patterns and variations. Given this information, fluctuations become noticeable when adjusting the smoothing parameter by bandwidth. Since the table variable is a percentage of the width of the top of the diamond relative to its widest point, more fluctuations appear when viewed in detail. This makes defining normality for the "table" variable more challenging.

```
# QQ plot
qq_plot = ggplot(new_diamonds, aes(sample = table, color = color)) +
  geom_qq(distribution = qnorm) +
  geom_qq_line(aes(linetype = "Normal Line"), color = "red") +
  facet_wrap(~ color) +
  labs(title = "QQ Plot of Table by Color",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  scale_linetype_manual(values = "solid", name = "Line Type", labels = c("Normal Line")) +
  theme_minimal()

print(qq_plot)
```



The straight red line in the QQ plot represents the theoretical normal distribution of the table variable across different color categories. This visualization suggests that the overall distribution of the table variable aligns closely with a normal distribution, as the data points tend to follow the line. However, notable outliers appear, particularly in color categories E, F, and I, where values above 67 deviate from the normal distribution, forming a right tail. These outliers, as previously noted in the box plot analysis, suggest greater variance in these categories. Additionally, the QQ plot does not show significant separation in the tails beyond these outliers, further supporting the conclusion that the table variable is largely characterized by a normal distribution. In conclusion, the "table" variable presents mixed evidence of normality depending on the visualization method. However, the QQ plot suggests no major deviations from normality, as the points follow the theoretical normal line without significant departures. While other plots indicate some variability, the absence of marked curvature in the QQ plot leads to a reasonable inference that the "table" variable is approximately normally distributed.

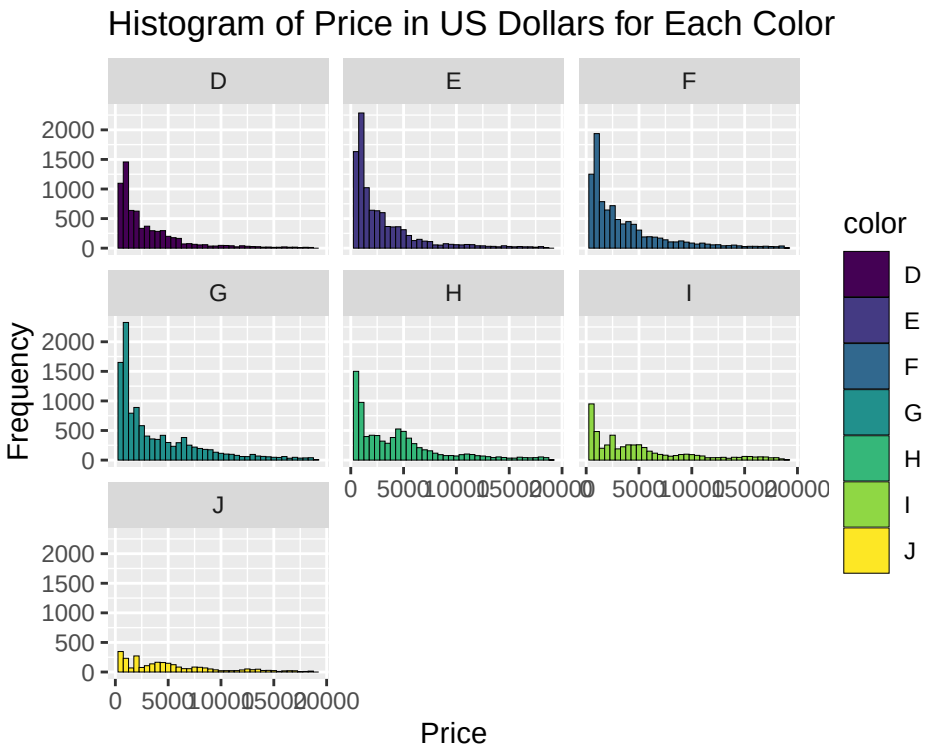
```
#Boxplot
box_plot = ggplot(new_diamonds, aes(x = color, y = table, fill=color)) +
  geom_boxplot(color = "black") +
  ggtitle("Box Plot of Table by Color") +
  xlab("Color") +
  ylab("Table") +
  theme_minimal()

print(box_plot)
```



The box plot, categorized by color grades, reinforces trends observed in the frequency plot, with widths between 55% and 60% being the most common among diamonds. Upon further analysis, it becomes clear that there is no significant difference in table width across the various color grades. All colors exhibit first and third quartiles that closely align with these preferred width ranges. However, colors F, G, and I display outliers that deviate substantially from their means of 57.4%, 57.3%, and 57.6%, respectively, indicating more extreme values in these categories, which contributes to the non-normal distribution. When examining the normality of the "table" variable across different color categories using box plots, signs of asymmetry become evident. This asymmetry is reflected in the median's position, which, in most cases, is closer to the first quartile (25th percentile) rather than centrally located within the interquartile range (IQR). There is a tendency for the data to concentrate toward the lower values of the variable. Furthermore, the whiskers of the box plots, representing the tails of the distribution, show that, in all color categories, the data is more concentrated in the right tail (upper end). This results in a wider distribution, as the right-side whiskers extend further, capturing more variability in higher table values. An interesting observation is in color J, where the median is closer to the third quartile (75th percentile), indicating that most data is concentrated at higher table values. Despite this upward shift in the median, the whiskers in color J are more balanced, with relatively equal lengths between the lower and upper tails. In summary, across all color categories, the "table" variable shows signs of non-normality and a concentration of values in the lower range.

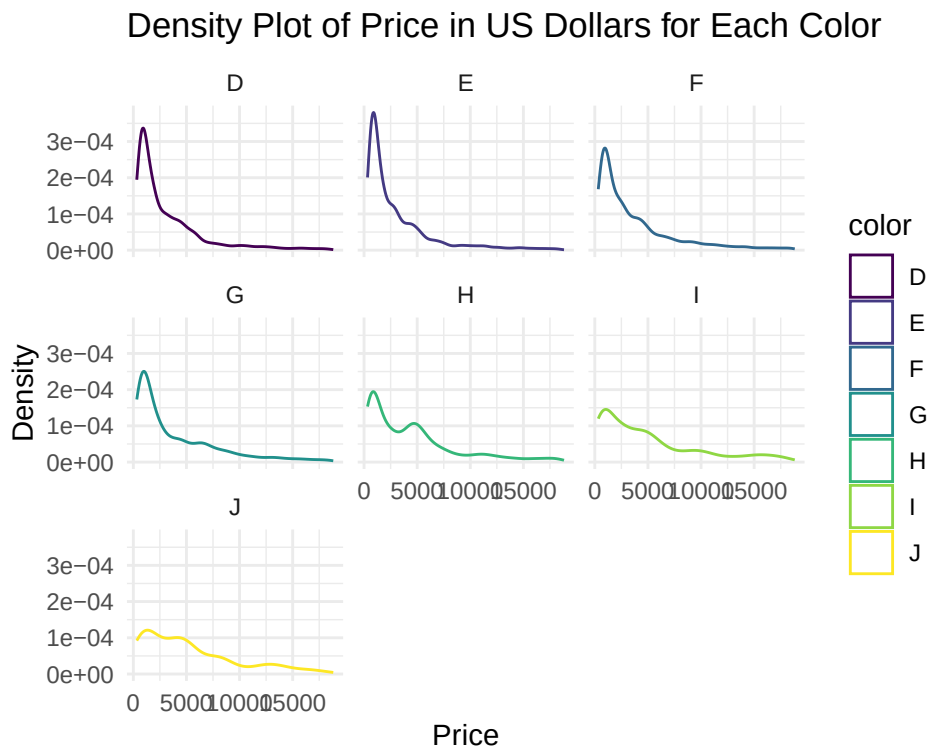
**Price (US Dollars) *

```
#Histogram
hist_plot = ggplot(new_diamonds, aes(x = price, fill=color)) +
  geom_histogram(binwidth = 500, color = "black",linewidth=0.1) +
  facet_wrap(~ color) +
  ggtitle("Histogram of Price in US Dollars for Each Color") +
  xlab("Price") +
  ylab("Frequency")

print(hist_plot)
```



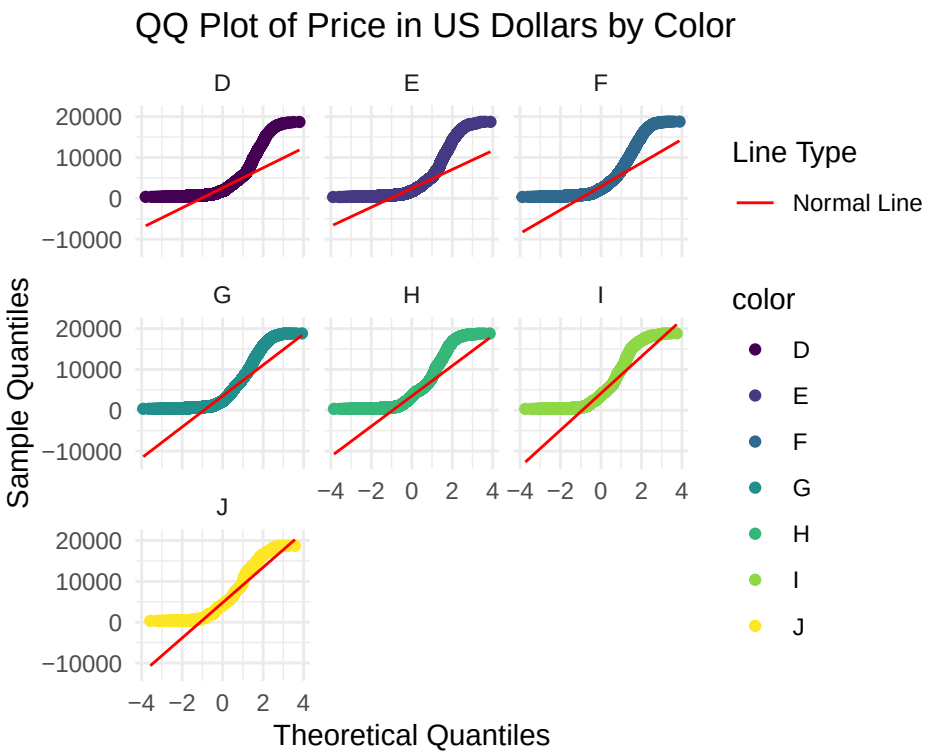
The histogram of the process categorized by color shows a wide distribution across ranges. However, for colors D to I, there is a noticeable left skew, indicating a higher concentration of diamonds at lower prices in most of these colors, with an overall peak between approximately $1,200 and $1,600. In contrast, for color J, the distribution expands further, as these diamonds often have more carats and thus can command higher prices, with a more diverse price range. Based on the histograms showing price distribution for different colors, the data is concentrated towards the left, featuring a notable peak followed by a long right tail. This type of distribution is typically described as right-skewed or positively skewed, meaning most data points cluster around lower price values, but a few higher price values stretch the distribution towards the right. This skewness, especially with a long tail, suggests that the data does not follow a normal distribution, which typically has a symmetrical, bell-shaped curve. Previous analyses, such as the identification of outliers in box plots, further support the conclusion of non-normality. The skewed nature of the data and the presence of outliers indicate that extreme values are pulling the distribution away from normality.

```
#Density plot
density_plot = ggplot(new_diamonds, aes(x = price)) +
geom_density(aes(color = color), alpha = 0, adjust = 1.2) + # Set alpha to 0 for no fill
facet_wrap(~ color) +
ggtitle("Density Plot of Price in US Dollars for Each Color") +
xlab("Price") +
ylab("Density") +
theme_minimal()

print(density_plot)
```



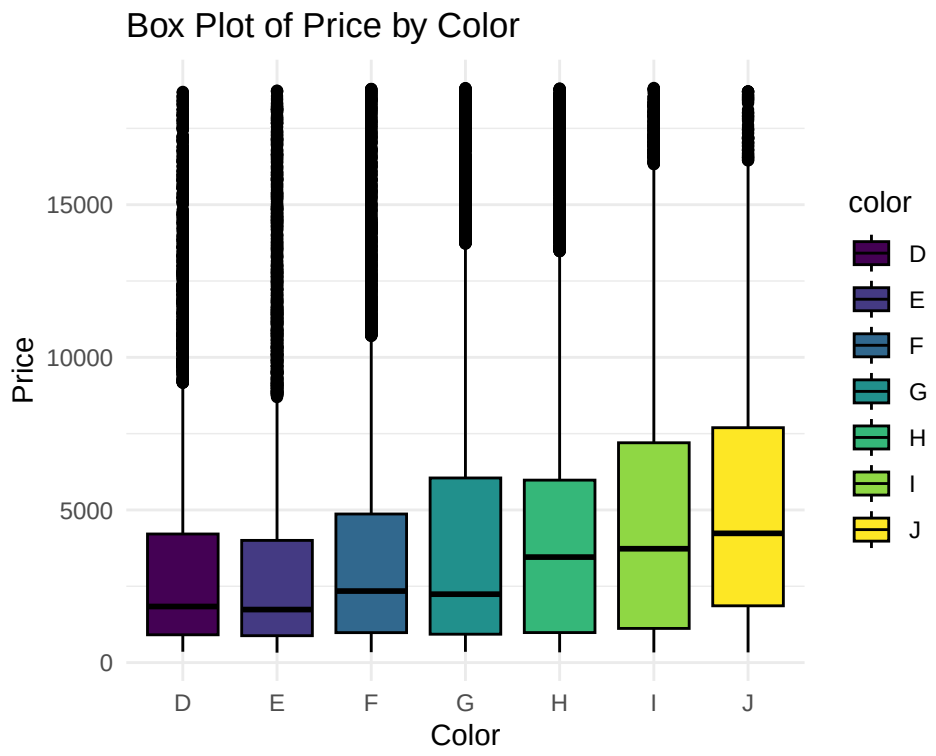
A review of the density plots confirms the data's non-normality. The curves show a clear concentration towards the left side for most colors, especially in the lower price ranges. As the curve progresses, it broadens, indicating a long tail to the right. For colors H, I, and J, the curves display wide peaks with pronounced flattening and elongation, suggesting the distribution for these colors is less concentrated around a specific value and spread over a wider range. This highlights the higher variability in price distributions for these colors. These flattened curves and long tails suggest a substantial portion of the data exists far from the central peak, reinforcing the right-skewed distribution and departure from normality. The broad peaks for these colors may indicate the absence of a single dominant price range, suggesting that the distribution is multimodal or heavily dispersed. This contrasts with a normal distribution, where data points typically cluster tightly around a central value. Therefore, based on both the histograms and density plots, the data clearly does not follow a normal distribution, particularly for colors H, I, and J. The skewed shape, presence of outliers, and the flattened, elongated density curves all emphasize the non-normality of the dataset.

```
# QQ plot
qq_plot = ggplot(new_diamonds, aes(sample = price, color = color)) +
  geom_qq(distribution = qnorm) +
  geom_qq_line(aes(linetype = "Normal Line"), color = "red") +
  facet_wrap(~ color) +
  labs(title = "QQ Plot of Price in US Dollars by Color",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  scale_linetype_manual(values = "solid", name = "Line Type", labels = c("Normal Line")) +
  theme_minimal()

print(qq_plot)
```

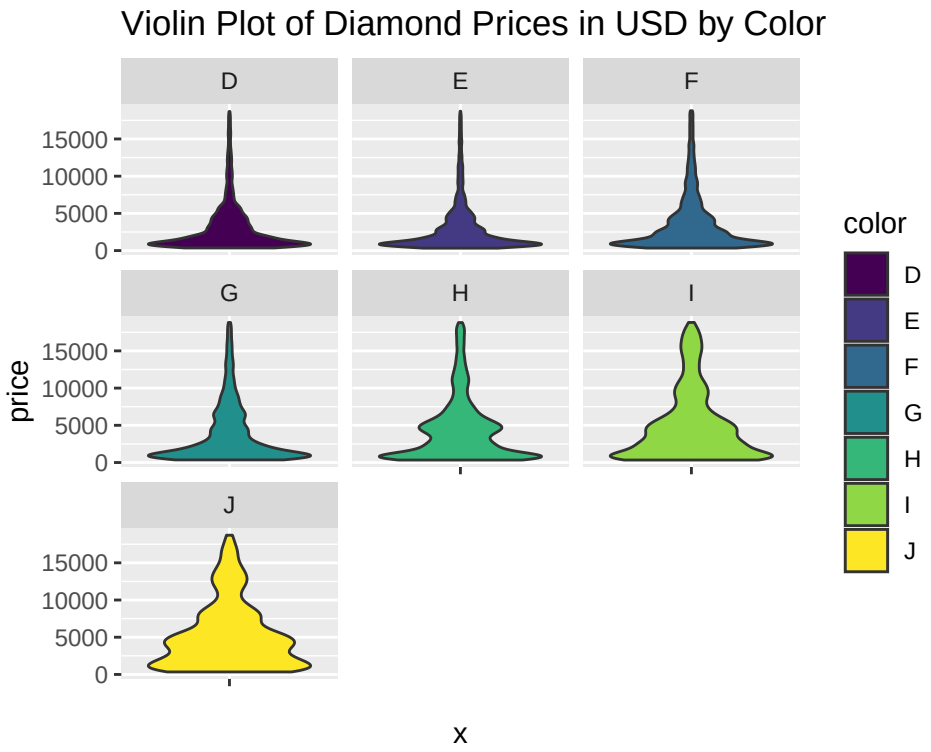


The QQ plots further indicate significant deviations from normality. In these plots, both the right and left tails show notable departures from the diagonal line representing a normal distribution, suggesting that the data points do not align closely with a normal distribution, where points should fall along this line. Moreover, the extreme curvature in the QQ plots emphasizes the data's lack of normality. The pronounced bends away from the diagonal line indicate thicker tails than in a normal distribution, consistent with the right skewness and long tails observed in the histograms and density plots. The price of diamonds, in US dollars, ranges from a minimum of $326 to a maximum of $18,823. The boxplot reveals interesting insights, showing that for each color category, 50% of the diamonds fall within relatively wide interquartile ranges. For example, in color D, 50% of the diamonds are priced between $911 and $4,213. As the color grade increases, the price range tends to widen; for color H, the range extends from $984 to $5,980. Colors I and J exhibit the greatest price variety, with 50% of the diamonds priced between $1,200 and $7,600. As mentioned, carat weight is a significant indicator of a diamond's price, so it's not surprising that colors I and J, which include larger diamonds, have a broader distribution of higher prices. Additionally, all color categories show outliers reaching prices of up to $18,000, highlighting the presence of exceptionally rare diamonds.

```
#Boxplot
box_plot = ggplot(new_diamonds, aes(x = color, y = price, fill=color)) +
  geom_boxplot(color = "black") +
  ggtitle("Box Plot of Price by Color") +
  xlab("Color") +
  ylab("Price") +
  theme_minimal()

print(box_plot)
```



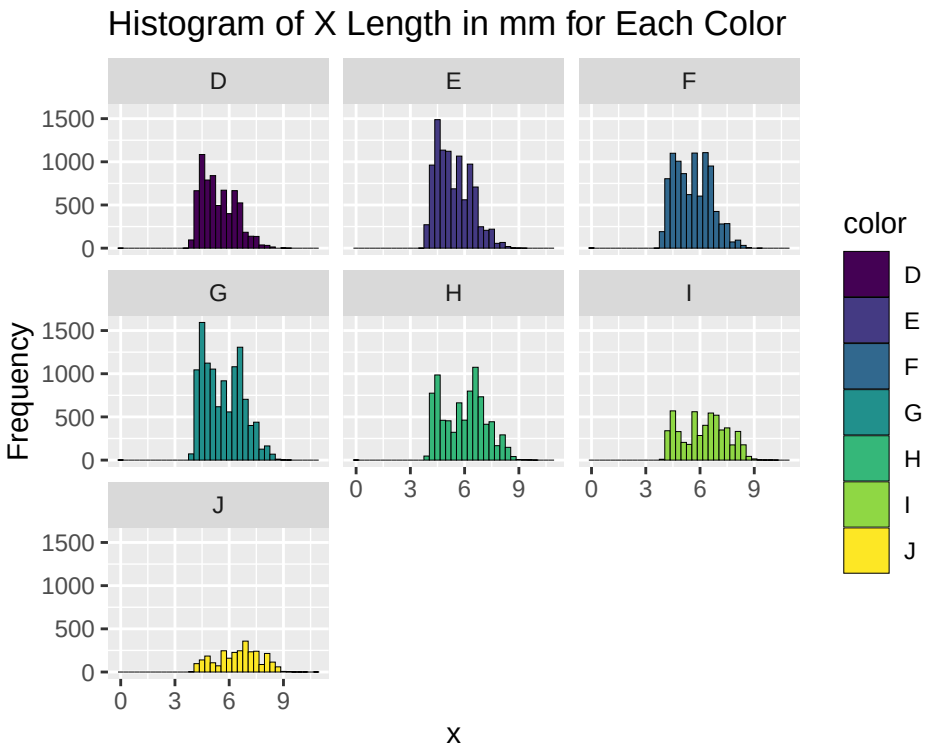

```
#Violin plot
violin_plot =  ggplot(diamonds, aes(x = "", y = price, fill = color)) +
   geom_violin() +
   labs(title = "Violin Plot of Diamond Prices in USD by Color") +
   facet_wrap(~color)

print(violin_plot)
```



The boxplot for price by color reveals considerable variability in diamond prices across all color categories. This variability is evident from the wide interquartile ranges (IQR), showing that the middle 50% of prices span a broad range. This wide spread indicates substantial price variation within each color category, further reinforcing the data's non-normality. Additionally, the long whiskers on the right side of the boxplots indicate strong right-skewness in the data. This suggests that higher diamond prices, though less frequent, extend far beyond the bulk of the data. The extended right tails confirm that a significant number of diamonds are priced much higher than most, making the distribution even more asymmetrical. These factors strongly support the conclusion that the distribution of diamond prices, categorized by color, is far from normal and exhibits substantial skewness and variability.

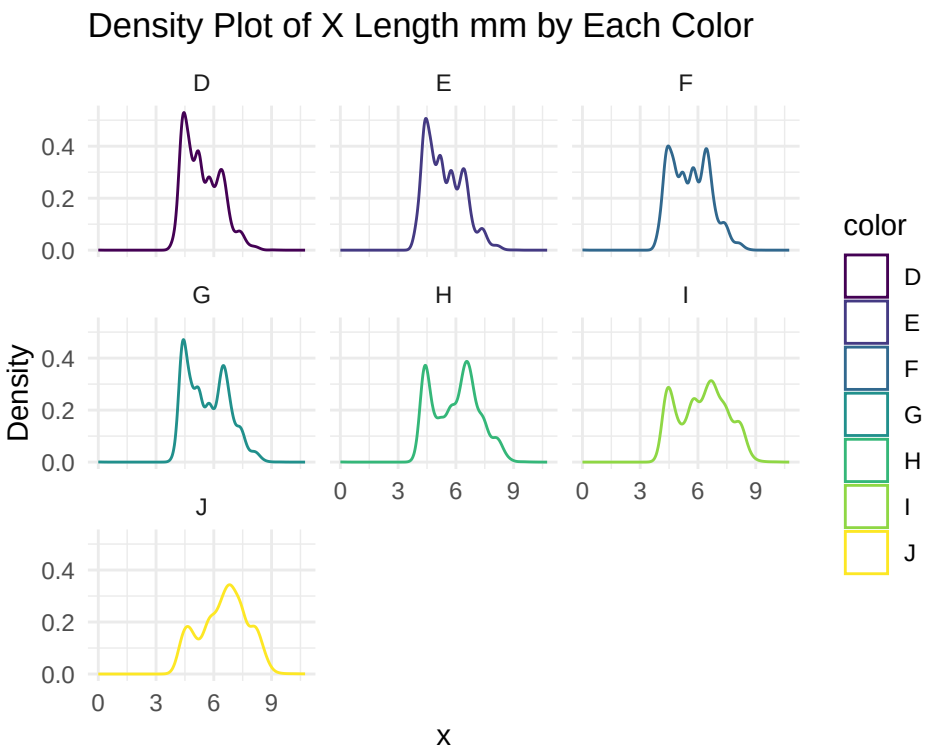
X (Length of the diamond in mm)

```
#Histogram
hist_plot = ggplot(new_diamonds, aes(x = x, fill=color)) +
  geom_histogram(binwidth = 0.30, color = "black",linewidth=0.1) +
  facet_wrap(~ color) +
  ggtitle("Histogram of X Length in mm for Each Color") +
  xlab("x") +
  ylab("Frequency")

print(hist_plot)
```



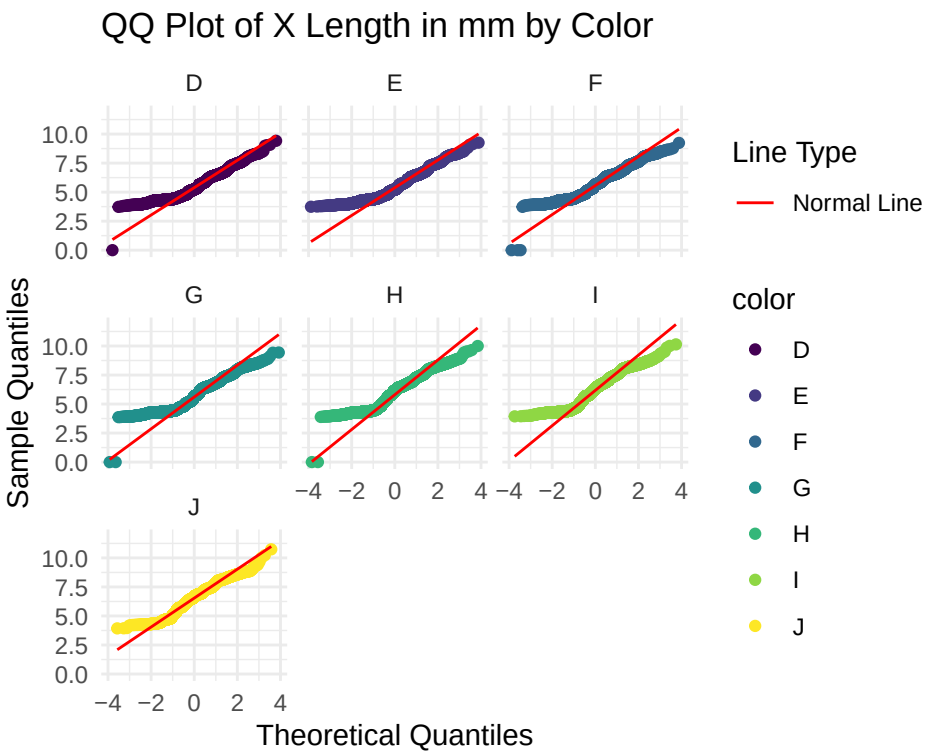
Looking at the histogram graphs for the variable x, which represents the length of the diamonds in mm, the distributions appear slightly different across the seven colors. For the D, E, F, and G variable, the histogram appears to have a strong peak on the lower end, representing more diamonds with shorter lengths, and they taper off towards higher values, indicating diamonds with longer lengths. Colors F and G not only show a right skewed distribution, but along with H, these colors also appear multimodal, with more than one peak in the data. This is a pattern not seen in normally distributed variables. Color I appears to have a distribution closer to a uniform, but the strong peak on the lower end deviates this variable from a normal distribution. Color J appears to be a unimodal and almost symmetric distribution. Although the frequencies are not as high in J as in the previously discussed colors, the data appears to be centered around the mean. However, there may be a slight left skew, which can be revealed in other graphics.

```
#Density plot
density_plot = ggplot(new_diamonds, aes(x = x)) +
geom_density(aes(color = color), alpha = 0, adjust = 1.2) + # Set alpha to 0 for no fill
facet_wrap(~ color) +
ggtitle("Density Plot of X Length mm by Each Color") +
xlab("x") +
ylab("Density") +
theme_minimal()

print(density_plot)
```



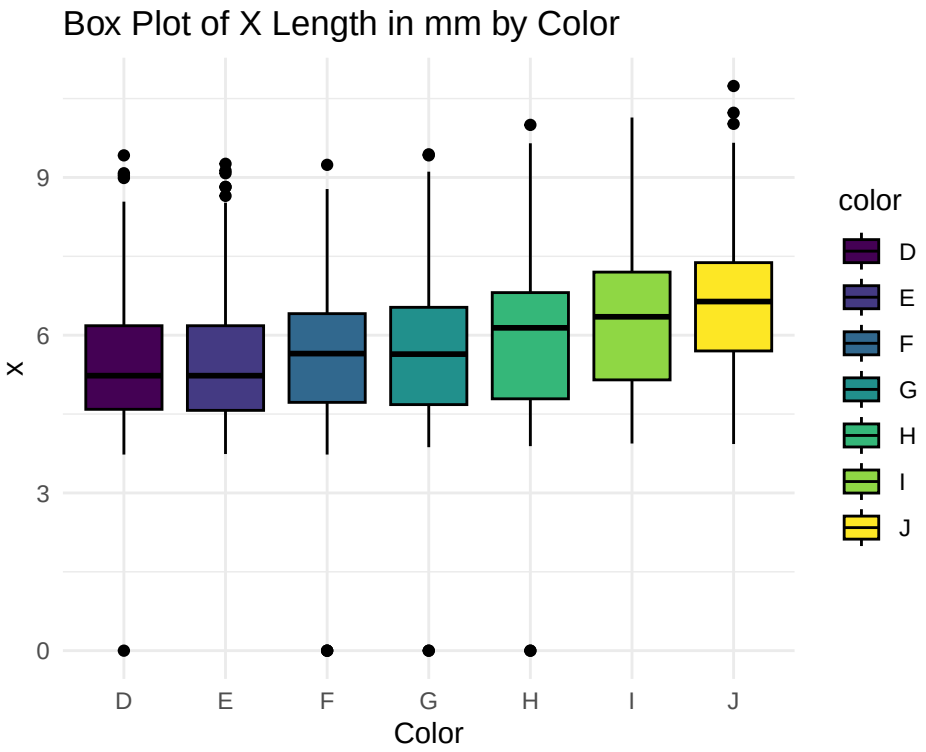
Density plots can further reveal patterns predicted in the histograms. As expected, colors D, E, F, G, and H show strong multimodality, all with a strong peak on the lower end, for the diamonds with shorter lengths, as well as on the high end, for diamonds with longer lengths. The density plot also uncovers the variables that have a stronger skewness than others, that is, colors D, E, and G are right-skewed. Although the histogram for F showed potential skewness, the density plot illustrates the peaks at either extreme are of the same length, with only a slight tail. Color I remains to have a central pattern, but the strong peak on the lower end moves this distribution from a normal. Finally, the density plot for color J appears to be the closest to a symmetric of all seven colors, but there is still a moderate left-hand tail.

```
# QQ plot
qq_plot = ggplot(new_diamonds, aes(sample = x, color = color)) +
  geom_qq(distribution = qnorm) +
  geom_qq_line(aes(linetype = "Normal Line"), color = "red") +
  facet_wrap(~ color) +
  labs(title = "QQ Plot of X Length in mm by Color",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  scale_linetype_manual(values = "solid", name = "Line Type", labels = c("Normal Line")) +
  theme_minimal()

print(qq_plot)
```



The QQ plots of variable x, diamond length, by color can show the actual data points plotted against the theoretical line of what a normal distribution would expect. Interestingly, all seven colors show a similar pattern. In the lower quartiles on the left hand side of each plot, the data pulls away from the line upwards. This indicates that our data views an abundance of observations in the lower values that what would be expected in a normal distribution. This, however, is not unexpected, considering all of our plots had strong peaks in the lower values, which skewed the shape and deviated from a normal distribution. In the central quantiles, the data seems to follow along the normal line closely, indicating the central parts of the data, indicating the middle lengths, followed an almost normal distribution. The points in the upper quartiles to the right of each plot that fall below the line demonstrate how the data lacks observations in those values, most likely due to the abundance of values in the lower ranges. The points in the very bottom left corners of the D, F, G, and H plots are most likely very small outliers, again, that a normal distribution would not expect. As discussed previously, the J color appeared to have the closest distribution to that of a normal, and this is clear in the QQ-plot as well. Compared to the other six colors, the left hand tail pulls away the least amount from the normal line.

```
#Boxplot
box_plot = ggplot(new_diamonds, aes(x = color, y = x, fill=color)) +
  geom_boxplot(color = "black") +
  ggtitle("Box Plot of X Length in mm by Color") +
  xlab("Color") +
  ylab("x") +
  theme_minimal()

print(box_plot)
```

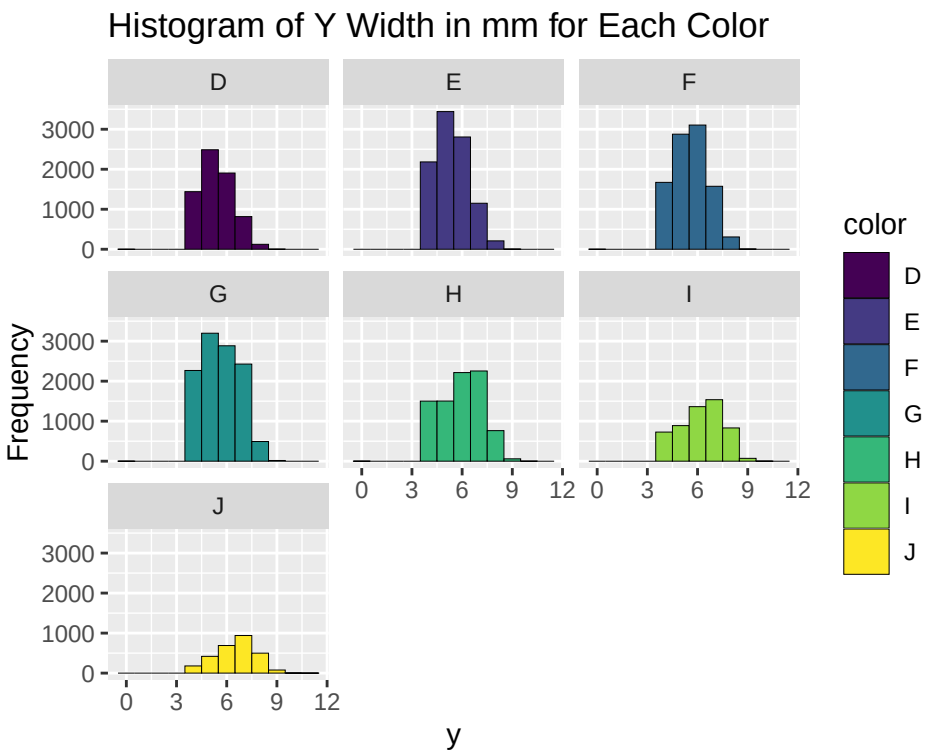


Boxplots give a visual representation of the quartiles of the distribution, particularly that of the spread. A boxplot of a normal distribution would appear to have symmetry over all four quartiles, the median line in the center of the box, and very few, if any, outliers. Looking at the first two boxes, D and E, it is clear the median line does not fall in the center, rather it falls towards the left. This is most likely due to the strong peaks in the data in the lower values. The left tail is also extremely shorter than the right tail, which indicates that there is a right skewness, as both tails should represent the same amount of data, 25% each, but with one being longer, this quartile is much more spread out. These box plots resemble a right skewed distribution. Moving on to the boxplots for F and G, The median appears to be more in the center, indicating the middle 50% is almost symmetric. However, the same characteristic occurs in the tails that the left tail is much shorter than the right one, indicating right skewness. The boxplot for color H and I show the median closer to the right side, indicating an abundance of values in the middle 50% that are higher. Even though the medians are pulled towards the higher values, these variables appear to be right skewed, as the left tails are significantly shorter than the right tails. Finally, color J continues to have characteristics closer to a normal, symmetric distribution compared to the others. The median appears slightly to the right of the center box, indicating more values in the medium-high ranges, but the tails appear to be more similar in length. This could align with a slight left skew, and after obtaining the skewness value of -0.1508 for variable x in group J, it is clear there is a negative skew in this group. In conclusion, the x variable, representing the length of diamonds, does not tend to follow a normal distribution when assessed over each color grouping. Instead, the data appears to be multimodal and right skewed. Although one color grouping, J, was close to a symmetric, unimodal distribution, there was still skewness in the data.

Y (Width of the diamond in mm)

```
#Histogram
hist_plot = ggplot(new_diamonds, aes(x = y, fill=color)) +
  geom_histogram(binwidth = 1, color = "black",linewidth=0.1) +
  facet_wrap(~ color) +
  ggtitle("Histogram of Y Width in mm for Each Color") +
  xlab("y") +
  ylab("Frequency")
```

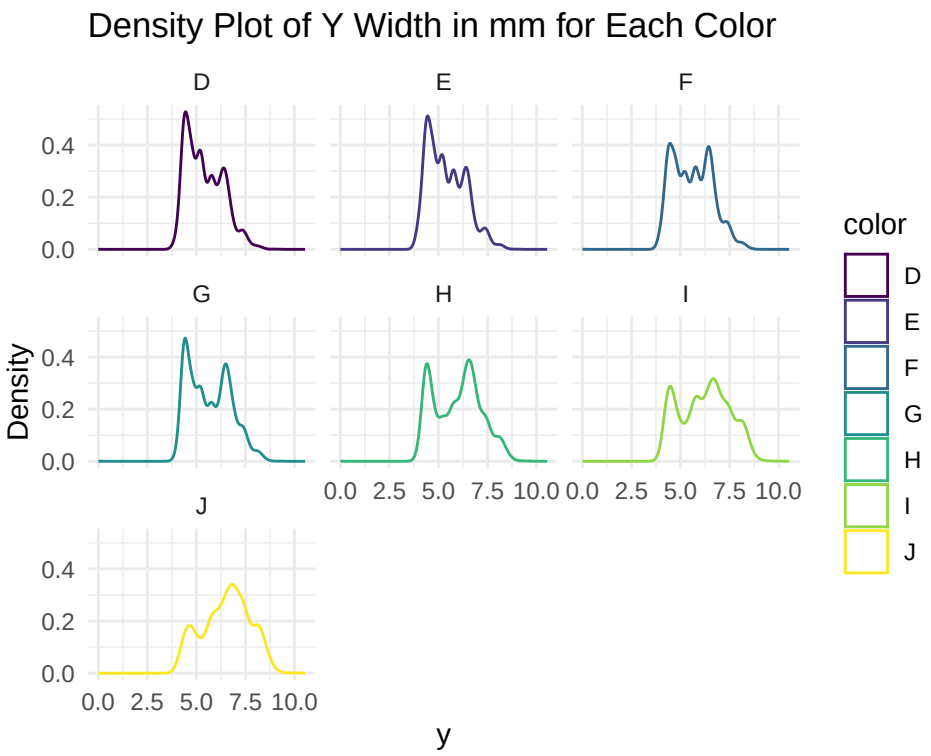
```
print(hist_plot)
```



The histogram plots of the variable y, which is the diamond widths measured in millimeters, show the frequency distribution of this variable over the seven groups of colors. A normal distribution would have a histogram plot that appears to be symmetrical about the center, and unimodal, meaning it has one peak. The histograms for colors D and E show the strongest peak towards the left of the center, meaning there is a concentration of data values in the lower measurements, and the median is less than the mean as the bulk of the data is on the lower range. Both plots appear to taper off towards the right, indicating a right skewed distribution. This aligns with the result of the y variable having a positive skewness value, found to be ***. The plots of F, G, and H appear to not only have a strong peak towards the lower values, but a second strong peak towards the higher values. This result can be explained as a multimodal distribution, where there are more than one peaks in the data. Similarly, in all three, the stronger peak is in the lower range, indicating more diamonds possess smaller widths. Plots F and G also appear to have stronger right tails compared to their left tails, further confirming the result of a positively skewed distribution. The color I appears to show a less skewed distribution, however the shape appears to be multimodal, with a strong peak before 5 mm and another peak around 7mm. Finally, the J color, although having the least amount of observations, appears to be the most symmetrical. Although there is a small peak before 5 mm, it is shorter than the central peak at 8 mm, a characteristic color I did not have. Following initial observations, color J could be assumed to be normally distributed, however looking at other plots is crucial before confirming this result.

```
#Density plot
density_plot = ggplot(new_diamonds, aes(x = y)) +
geom_density(aes(color = color), alpha = 0, adjust = 1.2) + # Set alpha to 0 for no fill
facet_wrap(~ color) +
ggtitle("Density Plot of Y Width in mm for Each Color") +
xlab("y") +
ylab("Density") +
```

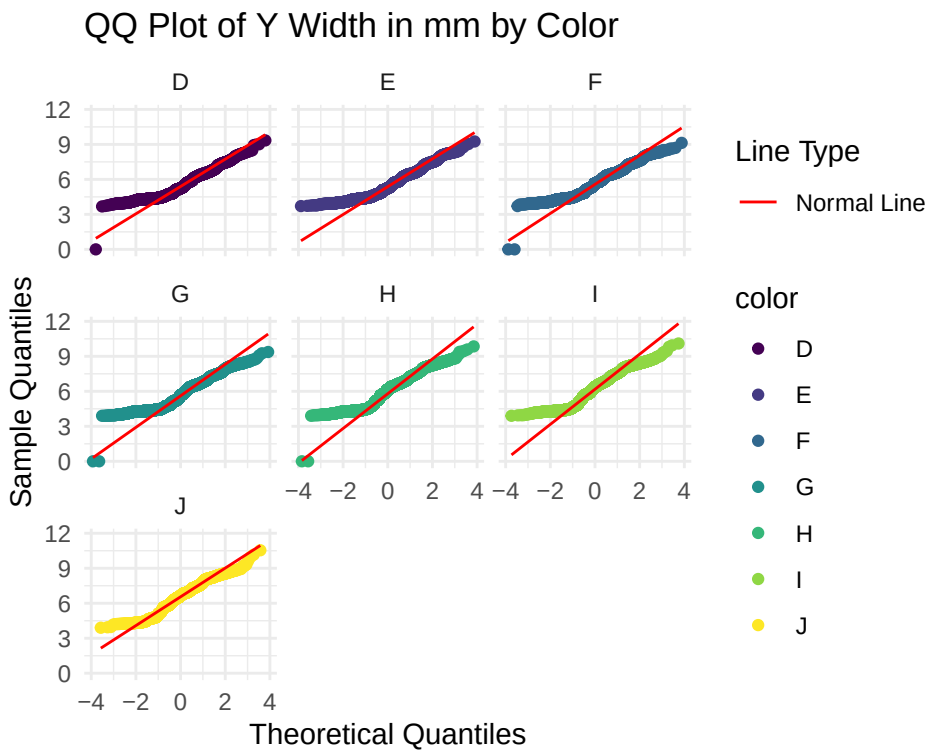
```
theme_minimal()

print(density_plot)
```



The density plots show a curved line that follows the distribution of variables, not necessarily the frequency. The density plots for colors D and E align with the results obtained from observing the histograms. Both distributions have a strong peak on the left, where the lower measurements are, and taper off with several, weaker peaks, towards the greater values. This directly aligns with a positively skewed distribution. The density plots for F, G, and H also align with the multimodality results found in the histograms, with a strong left hand peak, followed by the graph making a valley, then another strong peak in the higher values. This may be evidence enough to justify that these groups follow a bimodal distribution. In the density plot for color I, the left hand peak is still too strong to justify this group is symmetric, as the central peak and the left hand peak are at almost the same height. Although the peaks are not as distant when compared to groups like G and H, the drop off after the first peak may be too large and therefore color group I could also be considered bimodal. Previously, the histogram showed color group J as being almost symmetric. However, the density plot shows that it may actually be left skewed, with a strong peak towards the right of the center, and a longer left tail. Calculating the skewness of variable y in color group J, the obtained result is -0.1546, which aligns with a very slight left skew. When compared to the other groupings, this group appears to be the most symmetric and unimodal, despite the slight skewness it contains.

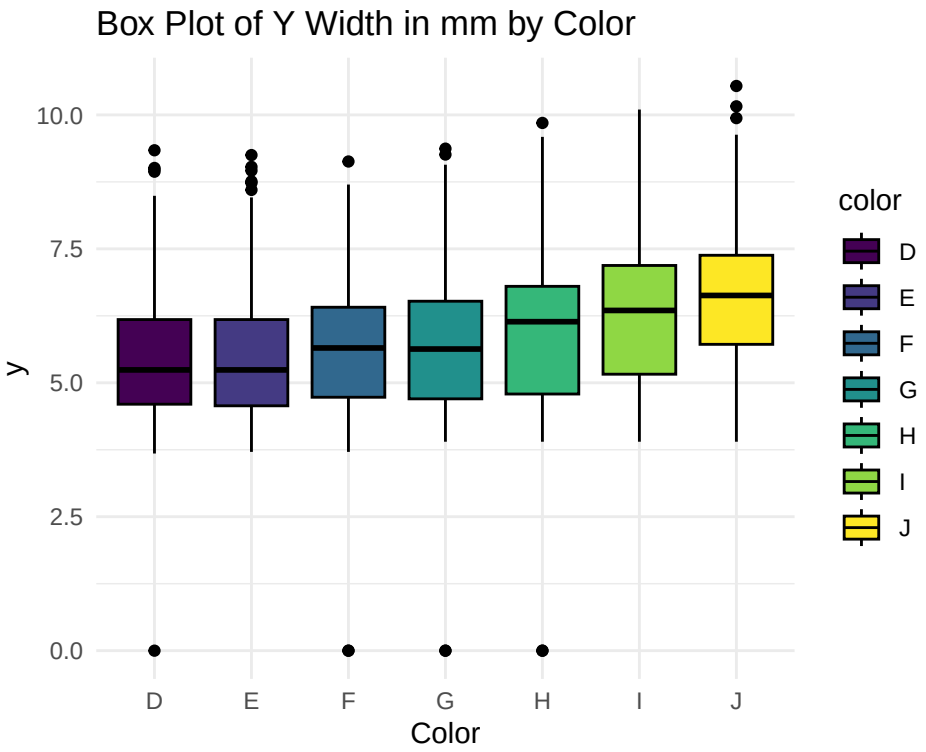
```
# QQ plot
qq_plot = ggplot(new_diamonds, aes(sample = y, color = color)) +
  geom_qq(distribution = qnorm) +
  geom_qq_line(aes(linetype = "Normal Line"), color = "red") +
  facet_wrap(~ color) +
  labs(title = "QQ Plot of Y Width in mm by Color",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  scale_linetype_manual(values = "solid", name = "Line Type", labels = c("Normal Line")) +
```

```
  theme_minimal()

print(qq_plot)
```



The QQ plots show the actual observed points along a normal line, which represents what a theoretical distribution would expect from our data. A normally distributed variable would follow the line almost perfectly, with no deviations or far off the line. In all of the plots, there is light pulling upwards, away from the line, in the lower quartiles. These points represent the abundance of data in the lower values, which is not a characteristic of normal distributions. This abnormality was previously observed in the histograms and density plots. QQ plots for colors D, F, G, and H show points in the lowest quartile below the normal line. These points represent width measurements that are much smaller than a normal distribution would predict. In several of the plots, there is a recurring pattern of the points in the higher quartiles falling below the line. This represents how, when considering the amount of observations we have in the lower quartiles, there is a lack of measurements in the higher values. A normal distribution is symmetric, so it is expected that there are a similar amount of values in either extreme. From the histogram and density plots, all values showed a right skewed distribution, where the distribution peaked on the left, and tapered to the right. As discussed before, color J appeared to be the color that was closest to a normal distribution. When comparing the QQ plot of J to the other colors, the tails in the extreme quartiles pull away less from the normal line, indicating the data is very close to a normal distribution.

```
#Boxplot
box_plot = ggplot(new_diamonds, aes(x = color, y = y, fill=color)) +
  geom_boxplot(color = "black") +
  ggtitle("Box Plot of Y Width in mm by Color") +
  xlab("Color") +
  ylab("y") +
  theme_minimal()

print(box_plot)
```



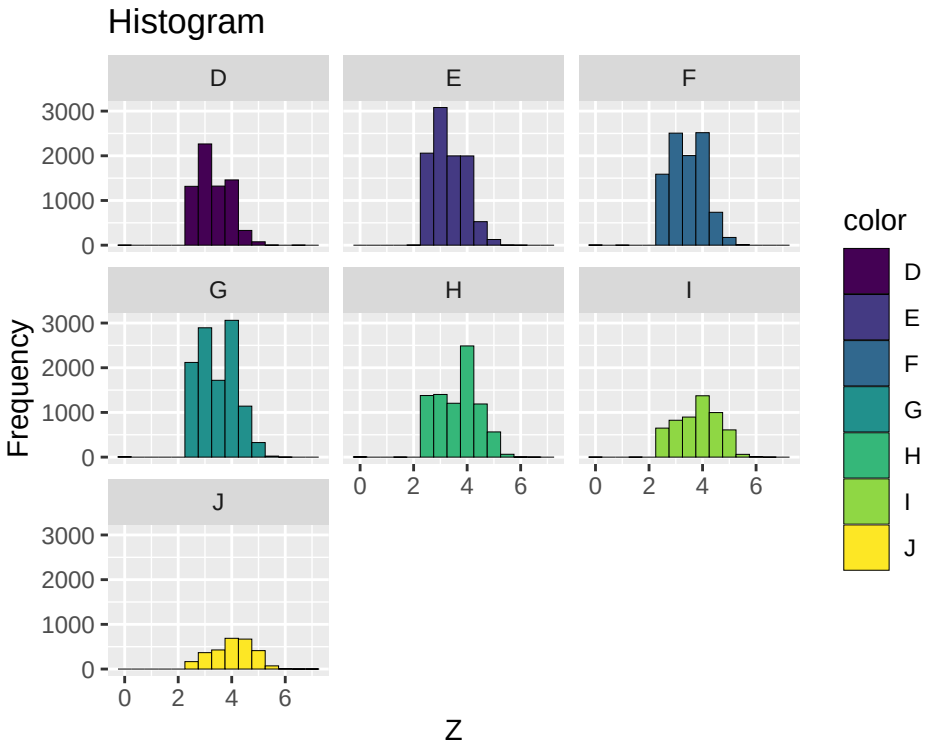
Boxplots are a helpful visualization to see how the data is spread through the four quartiles. A normally distributed variable would have a symmetric boxplot, where all four segments have the same length and the median falls in the center of the box. The first four box plots, colors D, E, F, and G, do not show a normally distributed variable. For colors D and E, the box plot shows the median closer to the left of the box, indicating the median is being pulled by a larger amount of data in the lower values. This is expected as the same result was observed in the previous plots. Additionally, the left tail for plots D, E, F, and G are significantly shorter than the right tail, indicating the longer widths have a greater spread than the shorter widths. These results combined support the prediction that these colors are skewed right. Boxplots for colors H and I also show a similar pattern in the tails, but differ in the box. These plots show the median closer to the larger values than the shorter values, indicating there is something pulling the median that way. In the previous plots, it was hypothesized that H and I follow a bimodal distribution. Finally, the boxplot for color J appears almost symmetrical, however the median is slightly towards the larger widths. This result was expected as in the previous plots, J appeared to be slightly left skewed, and has a small skewness value of -0.1546, which is a value corresponding to a slight left skew. Gathering the results, the variable y, containing the widths of diamonds in millimeters, when distributed over the color groups shows varying results. The majority of the groups, D, E, F, and G, show a strong right skew distribution, with a greater amount of value in the lower measurements than in the high measurements. Color groups H and I show a bimodal distribution of the diamond widths, such that there is a moderate amount of shorter widths, a small amount of middle widths, and a moderate amount of longer widths. Color group J shows the most normal, symmetric, and unimodal distribution of variable y. Most observations for this color fall near the average value, and taper off at both extremes. **Z (Depth of the diamond in mm)**

```
install.packages("ggpubr", repos = "https://cloud.r-project.org/")
```

```
## Installing package into 'C:/Users/cream/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'ggpubr' successfully unpacked and MD5 sums checked
##
```

```
## The downloaded binary packages are in
##  C:\Users\cream\AppData\Local\Temp\Rtmpuiqghc\downloaded_packages
```

```
library(ggpubr)
#Histogram
hist_plot = ggplot(new_diamonds, aes(x = z, fill=color)) +
  geom_histogram(binwidth = 0.5, color = "black",linewidth=0.1) +
  facet_wrap(~ color) +
  ggtitle("Histogram") +
  xlab("Z") +
  ylab("Frequency")

print(hist_plot)
```



The variable z represents the depth of the diamond, measured in millimeters. The histogram separates the data into several bins, or ranges, and then shows the frequency distribution of that data. A normally distributed variable would have a bell shaped histogram, with one peak in the middle, defined as unimodal, and symmetrical spreads as the frequency approaches either extreme. In several of the histograms, this pattern is not seen. Color groups D, E, F, G, H, and I show a strong peak on the left of the plot, where the shorter depths are. Graphs F, G, H, and I also show an interesting pattern of multimodality, where there is a second peak towards the higher values. This is another pattern that does not align with a normal distribution. As, for these six colors, the left peak is larger than the right peak, it could lead to as assumption that the data is also right skewed, which describes a data set where a majority of the observations are in the lower range, but the values in the higher range have a greater spread, and thus the data is "pulled" towards the right. The histogram for color J shows an almost symmetric distribution. Although the frequency for color J is lower than the other groups, it appears there could be a slight left skew in this variable, which is a result other plots could confirm.

```r
#Density
density_plot = ggplot(new_diamonds, aes(x = z)) +
geom_density(aes(color = color), alpha = 0, adjust = 1.2) + # Set alpha to 0 for no fill
facet_wrap(~ color) +
ggtitle("Density") +
xlab("Z") +
ylab("Density") +
theme_minimal()

library(diptest)

#Dip Test
library(diptest)
diamonds_F <- subset(new_diamonds, color == "F")
dip.test(diamonds_F$z)
```

```
##
##  Hartigans' dip test for unimodality / multimodality
##
## data:  diamonds_F$z
## D = 0.034532, p-value < 2.2e-16
## alternative hypothesis: non-unimodal, i.e., at least bimodal
```

```r
diamonds_H <- subset(new_diamonds, color == "H")
dip.test(diamonds_H$z)
```

```
##
##  Hartigans' dip test for unimodality / multimodality
##
## data:  diamonds_H$z
## D = 0.052313, p-value < 2.2e-16
## alternative hypothesis: non-unimodal, i.e., at least bimodal
```

The density plot shows a smoothed version of the histogram, but instead of representing a frequency count, it represents a relative frequency in the form of area. The patterns previously observed in the histogram are apparent in the density plots for this variable. Color groups D, E, and G show a strong left peak, indicating the dataset had many diamonds of shorter depths in these color groups. Following the peak, the data begins to taper towards the higher values. The concentration of data in the lower values followed by a slight decrease in distribution towards higher values is a strong indication of a right skewed distribution. Calculating the skewness measurement for variable z over colors D, E, and G are 0.5497, 0.5592, and 0.2488, respectively. Positive skewness values indicate right skewness, confirming this hypothesis. Density plots for colors F and H show a similar result as the histogram with a possible bimodal distribution. Hartigans' Dip Test For Unimodality / Multimodality can test For this hypothesis numerically, not just visually. This test has a null hypothesis that the data is unimodal, and an alternative hypothesis that it is not. For both color groups F and H, Hartigans' Dip Test returns very small p-values, less than 2.2e-16, which leads to rejecting the null hypothesis and that there is evidence these distributions are not unimodal. Color I also appears to have potential multimodality, although, compared to F and H the peaks are smoother. Running the Hartigans' Dip Test for color I and variable J returns the same result as before, a p-value less than 2.2e-16, leading to the rejection that this distribution follows a unimodal pattern. Lastly, color J in the histogram appeared to be the closest distribution to that of a symmetric bell-shaped graph. Although there is a strong peak central after 4 mm, and it appears symmetric immediately after, the left tail is stronger than the right tail. This could indicate this distribution is left skewed, and with a skewness value of -0.1314, this distribution has a very slight negative skew.

```
#QQ
qq_plot = ggplot(new_diamonds, aes(sample = z, color = color)) +
  geom_qq(distribution = qnorm) +
  geom_qq_line(aes(linetype = "Normal Line"), color = "red") +
  facet_wrap(~ color) +
  labs(title = "QQ Plot",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  scale_linetype_manual(values = "solid", name = "Line Type", labels = c("Normal Line")) +
  theme_minimal()
```

The Q-Q plots graph the actual distribution points along a normal line, which represents what a theoretical normal distribution of this variable would look like. In all of the plots, the groups show significant pulling in the lower quartiles away from the line. This indicates that the observed distribution has a greater amount of values in the lower ranges than a normal distribution would expect. The points that pull away below the line in the higher quartiles indicate that, when compared to the amount of data in the lower ranges, there is not the same amount in the higher ranges. This result is expected as the distributions showed a concentration of observations on the left, with a smaller amount on the right. This supports the hypothesis that there exists a right skewness in the variable. The points that fall below the normal line in colors G, F, G, H, and I indicate values that are smaller than what the normal distribution would expect. As anticipated, the color J appears to show the lower tail having a shorter distance away from the normal line, as this color previously appeared to have a weaker concentration of values to the left compared to the other color groups.

```
#Boxplot
box_plot = ggplot(new_diamonds, aes(x = color, y = z, fill=color)) +
  geom_boxplot(color = "black") +
  ggtitle("Box Plot") +
  xlab("Color") +
  ylab("Z") +
  theme_minimal()
```

The boxplot gives a representation of the spread of the variable over the four quartiles. The boxplots for colors D and E show similar patterns. First, the median line is not in the center, but more towards the left. This indicates that there is a strong concentration of data points in the second quartiles that pull the median away from the actual "center". The short left tail accompanied with a longer right tail indicate a right skewed distribution, as the whiskers represent the same amount of data (25% each), but the spread is significantly different. The boxplots for colors F and G show the median in the middle of the box, indicating there is no significant imbalance of concentration that pulls the median to either extreme. However, there is a significant difference in the length of the whiskers, similarly to that of D and E, which indicates more positive skewness. Interestingly, the boxplots for H and I show the center line pulled towards the greater values, indicating more observations in the larger depths that pull the median. In comparison to the other groups, these groups also have the greatest spread in both range and interquartile range, as the length of the whiskers and of the boxes are much greater than those of the other colors. The lengths of the whiskers do not appear significantly different, therefore, from the boxplot alone there is not enough evidence to support a skew in either direction. Finally, the boxplot for color J has several interesting characteristics. First, the center line is pulled slightly towards the right. As seen in the histogram and density plot for this variable, the data appeared to be slightly skewed left, as the highest peak was towards the right of the graph. This also aligns with the previous result of this variable having a skewness value of -0.13 in color group J. The normality analysis of the depth variable z reveals a diverse distribution across color groups, with many exhibiting right skewness characterized by a concentration of diamonds with shorter depths and a notable spread in deeper values. This pattern suggests that while most diamonds tend to have shorter depths, there are premium options available, influencing market dynamics by potentially increasing the value of diamonds with greater depth due to their rarity, thus impacting pricing strategies and consumer preferences in the diamond market.

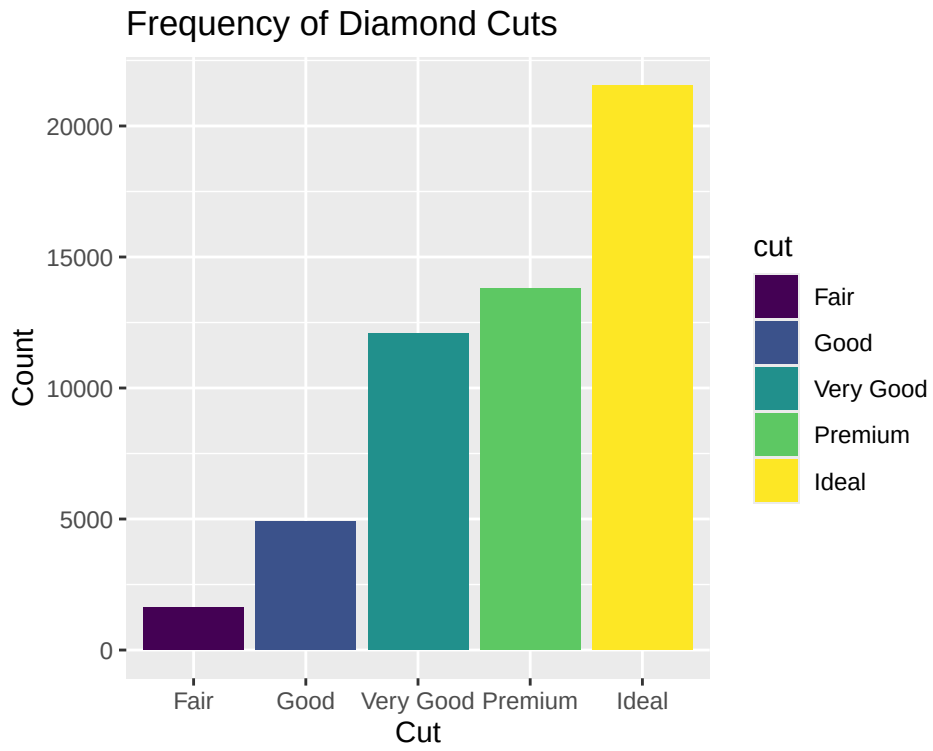
vi) Frequencies and Relationships of the Categorical Variables

```
freq_table = table(diamonds$cut)
knitr::kable(freq_table, caption = "Frequency Table for Cut")
```

Table 8: Frequency Table for Cut

Var1	Freq
Fair	1610
Good	4906
Very Good	12082
Premium	13791
Ideal	21551

```
ggplot(diamonds, aes(x=cut)) + geom_bar(aes(fill=cut)) +ggtitle("Frequency of Diamond Cuts")+xlab("Cut")
```



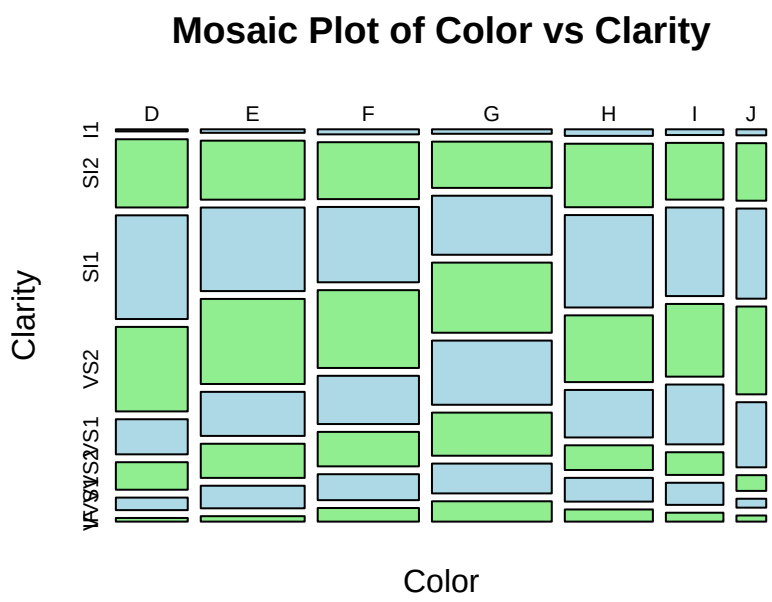
Cut (Quality of the diamond's cut) The variable cut describes the quality of the cut of the diamond which affects how well it reflects light. Diamonds can have different quality cuts, from Fair, the lowest quality, to Ideal, the highest quality, with categories Good, Very Good, and Premium in between. The frequency distribution of the dataset in this variable shows an interesting result. The frequency spread of the data shows a right skewed distribution, with the amount of observations increasing as we move from each category, with the lowest amount in Fair and the highest amount in Ideal. This gives insight into the diamond market that most diamonds have an ideal cut, most likely for retailers to have as many diamonds with the better cuts. On the other hand, this indicates a strong preference consumers have for higher-quality cuts.

```r
#Contingency Table and Mosaic Plot
contingency_table = table(new_diamonds$color,new_diamonds$clarity)
knitr::kable(contingency_table, caption = "Contingency Table for Color and Cut")
```

Table 9: Contingency Table for Color and Cut

	I1	SI2	SI1	VS2	VS1	VVS2	VVS1	IF
D	42	1370	2083	1697	705	553	252	73
E	102	1713	2426	2470	1279	991	656	158
F	143	1609	2131	2201	1364	975	734	385
G	150	1548	1976	2347	2148	1443	999	681
H	162	1562	2275	1643	1169	608	585	299
I	92	912	1424	1169	962	365	355	143
J	50	479	750	731	542	131	74	51

```r
mosaicplot(contingency_table,
           main = "Mosaic Plot of Color vs Clarity",
           xlab = "Color",
           ylab = "Clarity",
           color = c("lightblue",'lightgreen'),
           shade = FALSE)
```



Color and Clarity: The contingency table summarizes the frequency of observations across different categories of color and clarity. Each cell in the table represents the count of diamonds that fall into a specific combination of color and clarity. This can be presented visually with a mosaic plot that uses tiles with specific widths and heights to represent the count of observations with a particular Clarity and Color combination. The color variable ranges alphabetically from D to J, with D being colorless and higher grade

to J being the lowest grade with a slightly yellow hue. The clarity variable, which describes the amount of internal and external flaws, ranges from IF, which stands for internally flawless, having no flaws under strong magnification, to I1, which means the diamond has noticeable inclusions.

From the mosaic plot, the largest tiles appear in the center of each variable, most likely attributing that most diamonds have an average quality. The smaller tiles appear in either extreme, most likely indicating that there are few diamonds of extremely poor quality and undesirable features, and vice versa, there are few diamonds of extremely high quality with highly desirable features. The most common combination of clarity and color are diamonds that have an E grade color and a VS2 clarity, meaning the color is fairly good and there are very few inclusions that can most likely not be seen with the naked eye. The average price for diamonds with this color and cut combination is $2750.94, which falls below the entire dataset's average of $3932.72, indicating this diamond not only provides the consumer with a fairly good product, but it is also relatively affordable, compared to the average cost. The least amount of observations have both the best color, D, and the best clarity, IF, with only 42 observations. The average price for diamonds with this color combination is $8307.37, which is significantly higher than the average of $3932.72. This result is expected as these diamonds possess the greatest quality and are very rare. Individually, the greatest amounts of diamonds in this data set have either a VS2 or SI1 clarity, and either a color of E, F, or G. In summary, the diamond market appears to be structured around a balance between quality and affordability. While high-quality diamonds are rare and expensive, the bulk of the market caters to average-quality diamonds, which offer good value for the price. This reflects consumer preferences for diamonds that look good but aren't extremely costly.

Conclusion

This project sought to analyze various characteristics about the Diamonds data set, including centrality, normality, and how this sample can give information about real market patterns. For most of the continuous variables, the observations did not follow a normal distribution, but more a right skewed distribution, with a concentration of values in the lower ranges and less in the higher, but a greater spread for the higher ranges. For some, there were even bimodal patterns, with a strong peak in the lower ranges and a second, shorter, peak higher up. These patterns were also seen when splitting the continuous variables among the seven color groups. Finally, when analyzing the relationships and frequencies of categorical variables with the others, it was observed that, for these qualitative measurements, many diamonds fell in the middle-high ranges, with most being in the second best color and the middle clarity range. In the diamond market, these patterns can suggest that the majority of consumers are purchasing diamonds that are more affordable with moderate qualities, while a smaller number of high-end diamonds cater to buyers who are willing to pay significantly more for better characteristics. The bimodal patterns can be due to the nature of the diamond market being potentially split into a mass-market segment, which targets everyday consumers, and a luxury segment, which targets wealthier buyers. The patterns seen in the relationship of color and clarity shows that the diamond market tends to have more diamonds in these desirable but not top-tier categories, which reflects consumer preferences for high quality without paying the premium prices. Overall, these results suggest that, while the majority of diamonds sold are aimed at a broader consumer market, the high-end diamond market shows significant variation, with prices driven by scarcity, size, and quality, creating a separation between the "mass-market" and "luxury" segments.