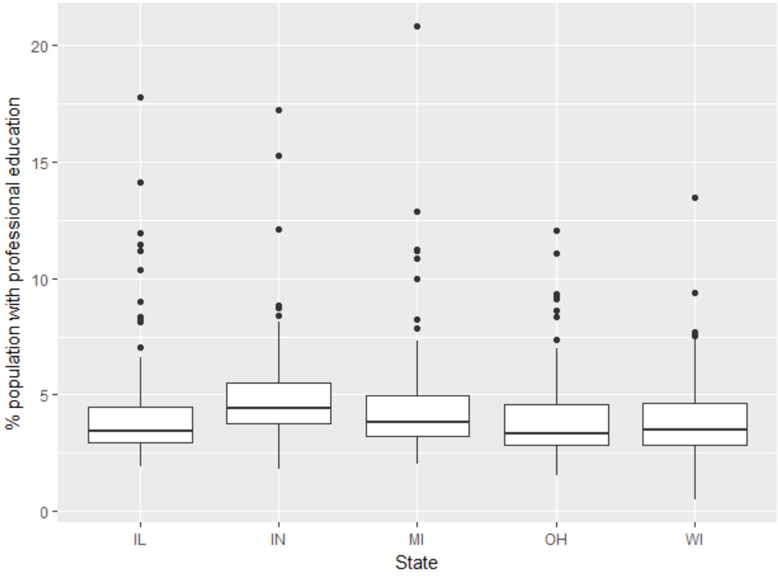
Matthew L. Miller
mmiller319

1. Professional Education by State

To visualize the distribution of professional education levels by state, I chose to use a box-plot because it most effectively shows the median values for each state, along with the size of the Interquartile Range (values between the 25th and 75th percentiles), and counties that have outlying values. To create the box plot, I used the following code:

```
ggplot(midwest, aes(state, percprof)) + geom_boxplot() + xlab(“State”) +
ylab(“% population with professional education”)
```

The following box plot was the result:



Based on the distributions, Indiana clearly has the highest median percentage of population with a professional education, while Illinois has the lowest. All the states have medians of less than 5%, but many have counties with outlying values between 10-20%. When examining which counties have outlying values, most of them are counties that contain a large university. The highest outlier is Washtenaw County, MI with 20.79% and contains the city of Ann Arbor where the University of Michigan is located. The highest outliers for the other states are also counties that include large universities such as the University of Illinois in Champaign

County, IL, Indiana University in Monroe County, IN, the University of Ohio in Athens County, OH, and the University of Wisconsin in Dane County, WI.

Another interesting property is that the outliers tend to be counties with large urban areas. Many of the mid-range outliers (those in the 7.5 - 12.5% range) contain big cities or their suburban communities. These include counties that either contain or are just outside of cities like Chicago, Indianapolis, Columbus, Cincinnati and Cleveland.

While the box plot clearly shows that Indiana has the highest percentage of people with a professional education, the determination of the state with the lowest percentage was less clear because the bars representing the median values are so close together. In fact, it was difficult for me to determine at first whether Illinois, Ohio, or Wisconsin had the lowest median percentage using the box plot. Therefore, I created some histograms based on the median value for each state, as well as the total percentage of population.

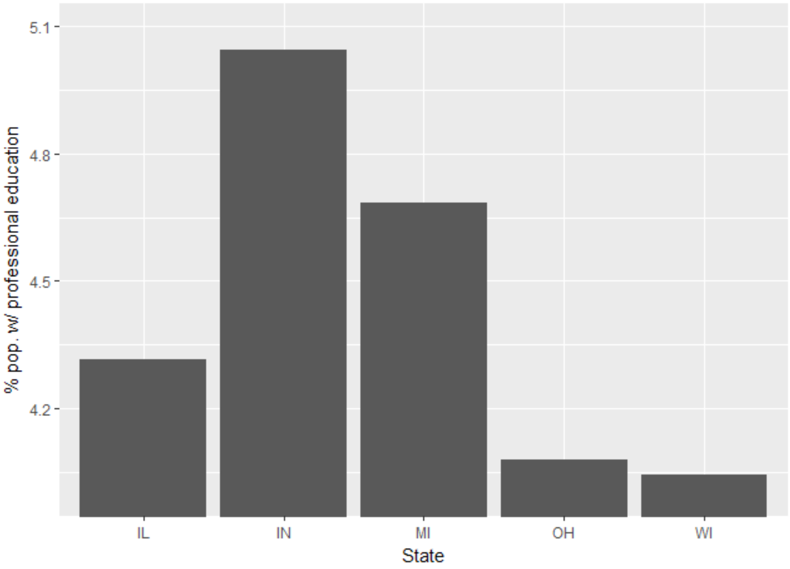
I first created an aggregate of the "percprof" values based on "state" to determine the aggregate mean, using the code:

```
midwest_percprof_ag <- aggregate(percprof ~ state, midwest_percprof_state, me
an)
```

I then created a bar chart representing the combined average for the states with the code:

```
ggplot(midwest_percprof_ag, aes(x=state, y=percprof)) + geom_bar(stat="identi
ty") + coord_cartesian(ylim=c(4,5.1)) + xlab("State") + ylab("% pop. w/ profe
ssional education")
```

The resulting chart gives a clearer picture of which states have the highest and lowest combined mean of people with a professional education:



Because all the states have a combined mean of between 4.0-5.05%, the scale values for the y-axis are set to 4.0 - 5.1. This creates a clearer picture of the degree of difference between the combined means for the states and shows that Wisconsin has the lowest combined mean at 4.04%, just slightly lower than Ohio at 4.08%. The differences in the combined means for Wisconsin, Ohio, and Illinois are readily visible in the bar chart as opposed to the box plot.

2. School and College Education by State

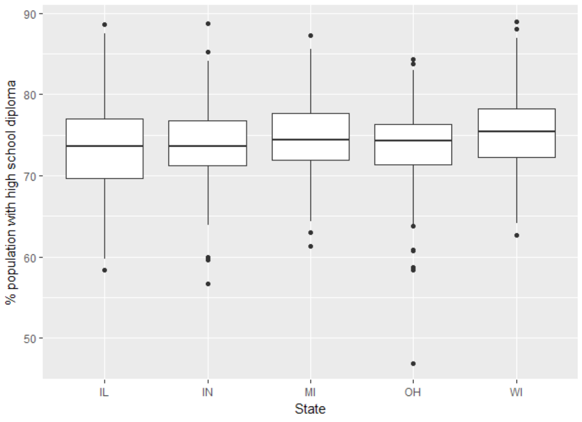
To analyze the relationship between high school and college education levels, I began by creating a simplified data frame from the midwest dataset called midwest_edu which contained only the columns “PID”, “county”, “state”, “perchsd”, and “percollege” using the following code:

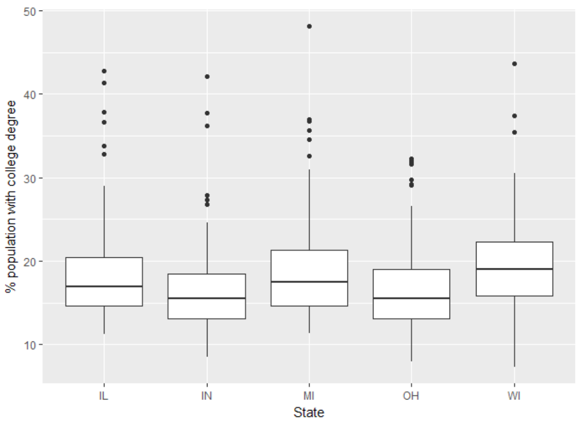
```
mycols <- c("PID", "county", "state", "perchsd", "percollege")
midwest_edu <- midwest[mycols]
```

I then created two box plots illustrating perchsd by state and percollege by state using the code:

```
ggplot(midwest_edu, aes(state, perchsd)) + geom_boxplot() + xlab("State") + y
lab("% population with high school diploma")

ggplot(midwest_edu, aes(state, percollege)) + geom_boxplot() + xlab("State")
+ ylab("% population with college degree")
```





As can be seen from the box plots, the percentage of the population with a college education is substantially lower than the percentage with a high school diploma. Based on the median bars, all states have an average of 70-80% of people with a high school diploma and 15-

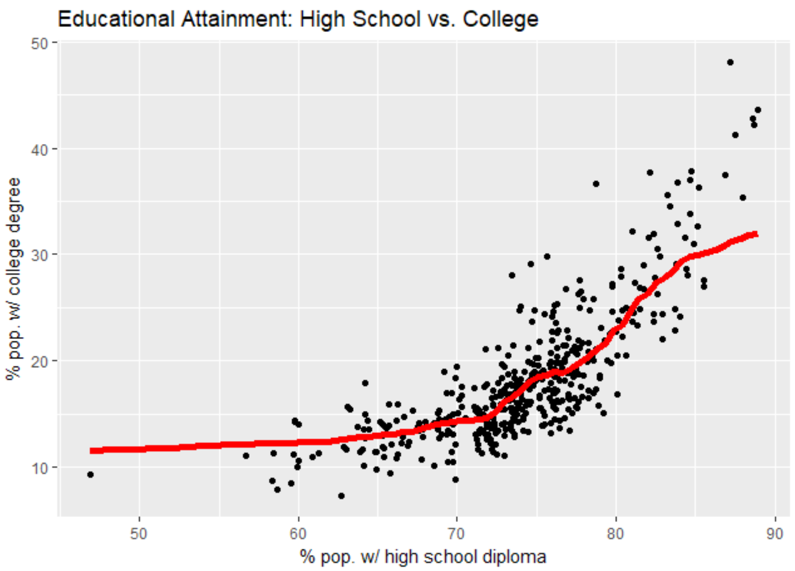
20% with a college degree. The state of Wisconsin has the highest mean for both categories with slightly more than 75% with a high school diploma and slightly less than 20% with a college degree.

However, there are a substantial number of outliers, with several counties having close to 90% of adults with a high school diploma and 30% or more with a college degree. For the outliers in the percollege category, the counties with the highest percentages generally correspond to the counties with the highest percentages of adults with a professional education in Part 1, namely Washtenaw County, MI, Dane County, WI, Champaign County, IL, etc. These are the same counties that contain large universities.

For the percentage of adults with a high school diploma, there are also some outliers that are well below the mean. The lowest is Holmes County, OH with only 46.9% of adults having finished high school. Others include Lagrange County, IN, with only 56.6%, Gallatin County, IL, and Adams County, OH, with 58.4% each. These counties tend to be in rural areas that do not include large cities or universities.

Next, I created a scatterplot to analyze the relationship between the percentage of adults with a high school education and a college education using the code:

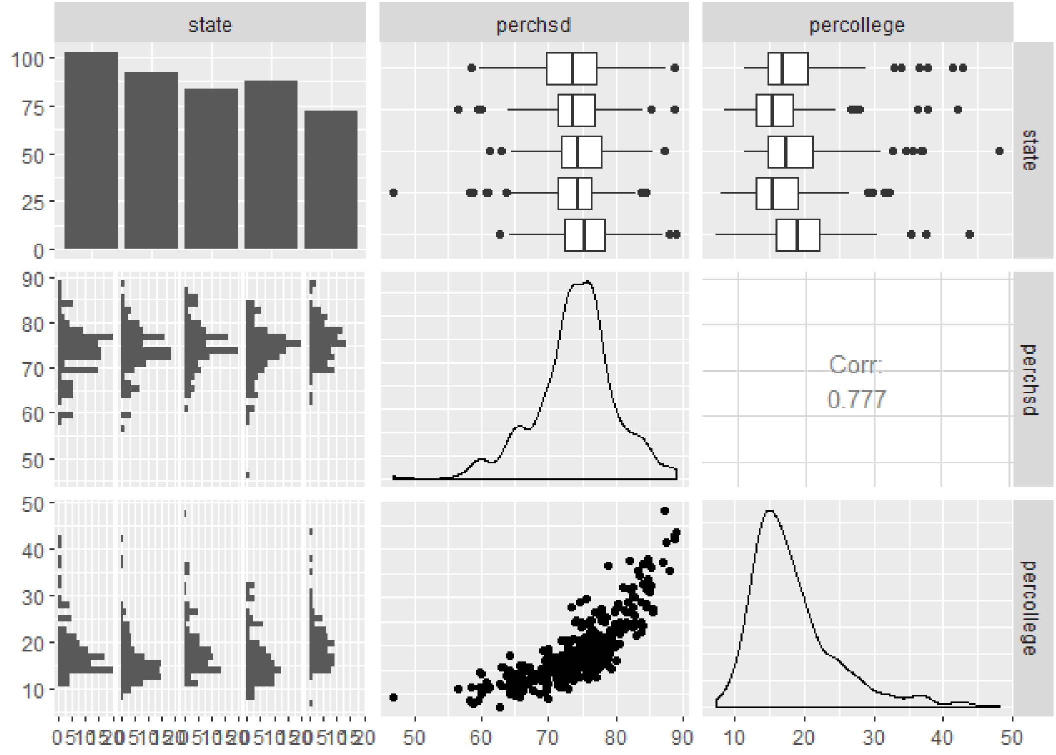
```
qplot(perchsd, percollege, data=midwest_edu, main = "Educational Attainment:
High School vs. College", xlab="% pop. w/ high school diploma", ylab="% pop.
w/ college degree") + stat_smooth(method="loess", method.args = list(degree=0
), span=0.2, se=FALSE, color="red", size=2)
```



As can be expected, the scatterplot shows a correlation between the percentage of the population with a high school diploma and the percentage with a college degree. The plot shows a red smoothed line curve which is a weighted average of the data points. The line used the default "loess" method with a span of 0.2, which proved to be optimal for illustrating the relationship (a span of 0.1 created a noisier line while a span of 0.3 or higher created a flatter line that did not adequately represent the upward correlation).

Finally, I experimented with a creating a combined pair-wise plot using ggpairs in the following code:

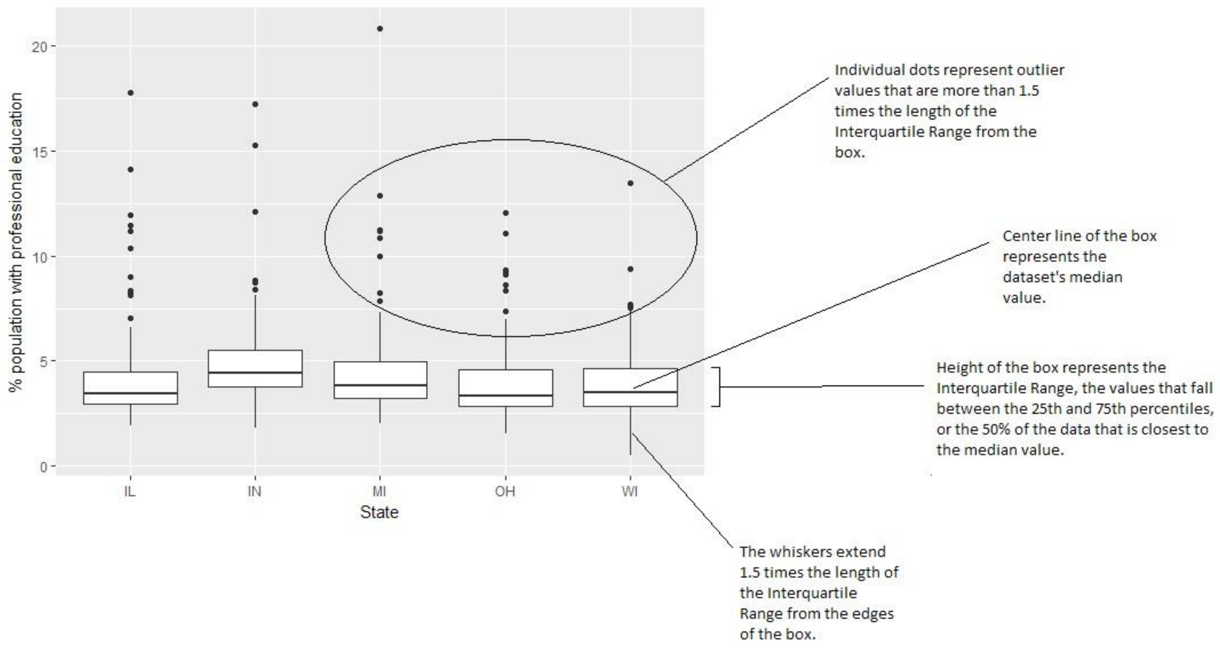
```
ggpairs(midwest_edu, columns = c("state", "perchsd", "percollege"))
```



The pair-wise plot illustrates the relationships between all three variables in one combined graphic. The box plots for state vs. perchsd and state vs. percollege show the same relationship as the stand-alone box plots above. Also, the relationship between percollege and perchsd is the same as the scatterplot created above. However, the pair-wise plot includes some additional information such as two histogram charts (perchsd vs. state and percollege vs. state) showing the numbers of people with high school diplomas and college degrees by county, and two line graphs (center and lower right boxes) showing the frequency distributions of the percentages of people with high school diplomas and college degrees. Also, because the pair-wise plot contains two boxes for perchsd vs percollege (bottom row, center column and center row, right column), only the bottom center box shows the scatter plot while the center right column shows the correlation. Not surprisingly, the correlation between the two variables is reasonably high at 0.777.

3. Comparison of Visualization Techniques

A box plot's main feature is a box with an inner line and a pair of "whiskers" that extend in either direction from the box. The line inside the box represents the median value of the data being represented. It may lie towards one end of the box depending on the skewness of the dataset. The width of the box represents the Interquartile Range (IQR), which is the interval between the first and third quartiles (i.e. values above the 25th percentile but below the 75th percentile). Therefore, the width of the box represents the central 50% of the data that is closest to the median value. The whiskers extend 1.5 times the length of the IQR from either edge of the box. Any outliers that fall outside of the range of the whiskers are shown as individual dots either above or below the extent of the whiskers.



Individual dots represent outlier values that are more than 1.5 times the length of the Interquartile Range from the box.

Center line of the box represents the dataset's median value.

Height of the box represents the Interquartile Range, the values that fall between the 25th and 75th percentiles, or the 50% of the data that is closest to the median value.

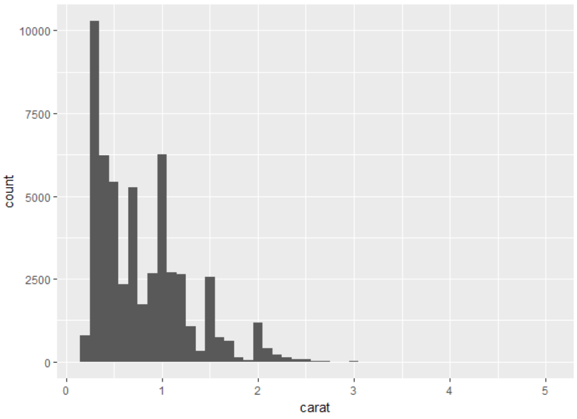
The whiskers extend 1.5 times the length of the Interquartile Range from the edges of the box.

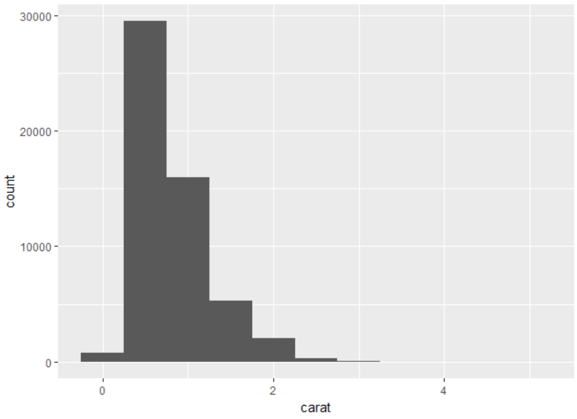
A box plot is especially useful when illustrating the distribution of the data, namely the median value, the values that fall between the 25th and 75th percentiles, and the extent of outlying values. The variation in the size of the box, the location of the median line, and the length of the whiskers illustrate different aspects about the dataset being represented. A larger box and longer whiskers, for example, will show that the Interquartile Range of the data is more broadly distributed. A smaller box with shorter whiskers will indicate that the dataset is more tightly clustered around the median value.

If the median line is not centered within the box, it shows that the data is skewed to one side of the distribution. In the diagram above, for example, the median line in the box for Ohio falls closer to the bottom of the box, meaning that the data for Ohio is skewed towards lower values. Additionally, the individual dots represent outliers that fall more than 1.5 times the length of the Interquartile Range from the median. In the diagram above, there are a significant number of outlying values, with the highest outlier being for the state of Michigan and having a value of over 20%.

Histograms are useful when one wishes to show the count of a particular variable. Histograms are one-dimensional plots that show a continuous variable along the x-axis and the count of how often that variable occurs on the y-axis. Because there may be a very large number of observations of a variable, histogram data is often divided into bins that represent a certain range of values for the variable. For example, if a variable has values that range from 0 to 100, a histogram may divide them into bins with a length of 10 in order to group observations that have values of 0-10, 10-20, 20-30, etc. Choosing the right size of a bin is important in creating a good histogram since a bin size that is too large may reduce the detail of the histogram and not represent trends in the data. A bin size that is too small, however, may result in a histogram that is too "noisy" and crowded with too many observations, thus making it difficult to discern trends in the data.

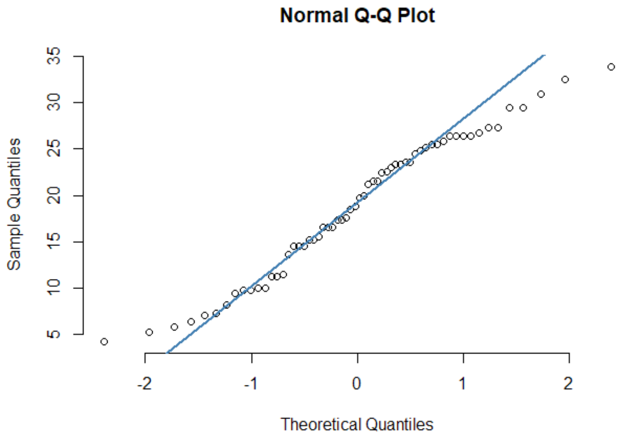
An example of a histogram with different binwidths is shown below. It is derived from the "diamonds" dataset and shows the count of diamonds by carat. Having a larger binwidth (the right histogram has a binwidth of 0.5 compared to the left with a binwidth of 0.1) results in a coarser histogram that does not reveal as much detail in the data distribution.





A QQPlot is known as a quantile-quantile plot and is used to compare to datasets based on their quantiles. They are most useful for comparing sampled datasets and are basically scatter plots showing the relationship between the sampled distributions. They consist of several common shapes illustrating how the datasets relate to each other. If the scatterplot has a slope of 1 and passes through the origin, then the two distributions are likely sampled from the same dataset, while a slope of 1 but not passing through the origin indicates that one distribution is shifted relative to the other. If the slope of the scatter plot is not 1, then the distributions are likely translated and scaled relative to each other. In addition, if the scatterplot has an S-shape along either the x or y-axis, then one dataset has a longer tail in its distribution along the variable represented by the axis.

An example of a QQPlot based on the "ToothGrowth" dataset is shown below.[1] It shows a normal QQ Plot of the variable "len" along with a regression line. Since the slope of the line is approximately one, the quantiles of the data are distributions from the same dataset and therefore have a similar shape and spread.

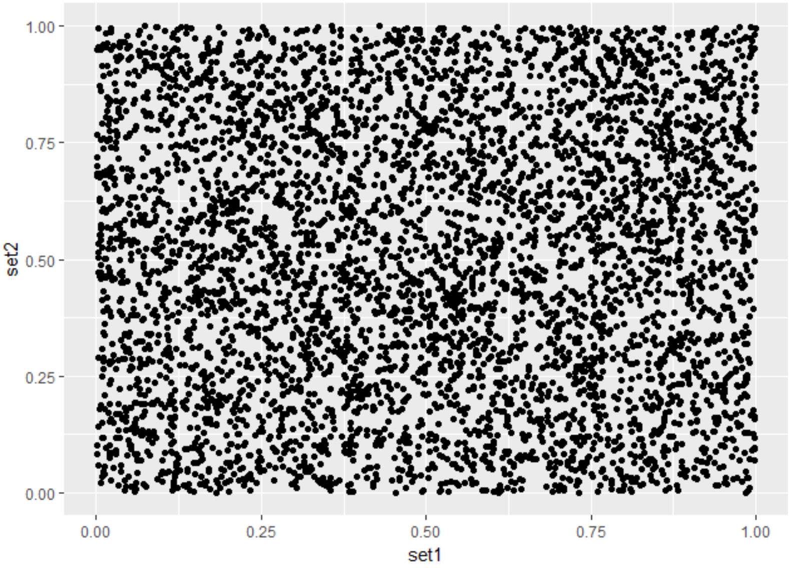


4. Random Scatterplots

To generate scatterplots I used the runif() function with a series of parameters starting at 10 and increasing to 5000 at intervals of 500 with the following code:

```
set1 <- runif(5000)
set2 <- runif(5000)
qplot(set1, set2)
```

An example of the scatterplot with 5000 points in both set1 and set2:



I exported each of the scatterplots into four formats: pdf, jpg, png, and eps. I then created a data frame using the size of the resulting files in kilobytes:

```
n <- c(10, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000)
pdf <- c(5, 34, 63, 92, 121, 150, 179, 207, 236, 265, 294)
jpg <- c(35, 122, 181, 219, 251, 273, 290, 305, 314,320, 325)
png <- c(5, 9, 13, 16, 18, 20, 21, 22, 24, 24, 25)
eps <- c(9, 21, 34, 46, 58, 70, 82, 94, 106, 118, 130)
scatterplot_size <- data.frame(n, pdf, png, jpg, eps)
```

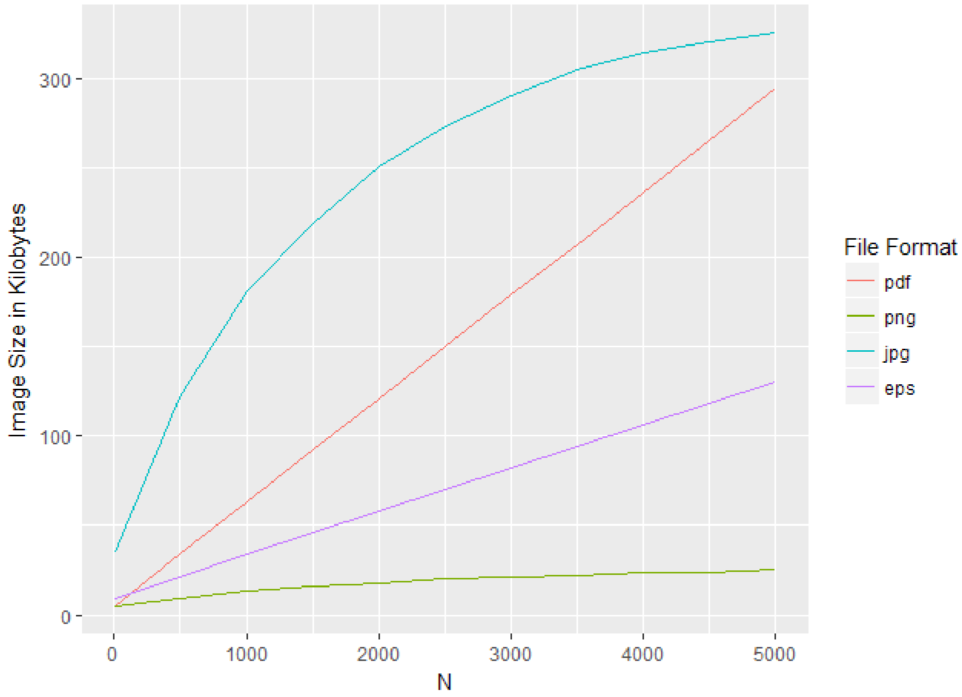
To make the analysis simpler, I used the melt() function on the scatterplot_size data frame to combine the file types into the “variable” column and their respective file sizes into the “value” column:

```
scatterplot_size_melt <- melt(scatterplot_size, id="n")
```

The scatterplot_size_melt data frame was then used to create a graph with the following code:

```
ggplot(scatterplot_size_melt, aes(x=n, y=value, color=variable)) + geom_line(
) + xlab("N") + ylab("Image Size in Kilobytes") + labs(color="File Format")
```

The resulting graph shows the relationship between the file size for each of the four image formats based on the number of points graphed in the scatter plots:

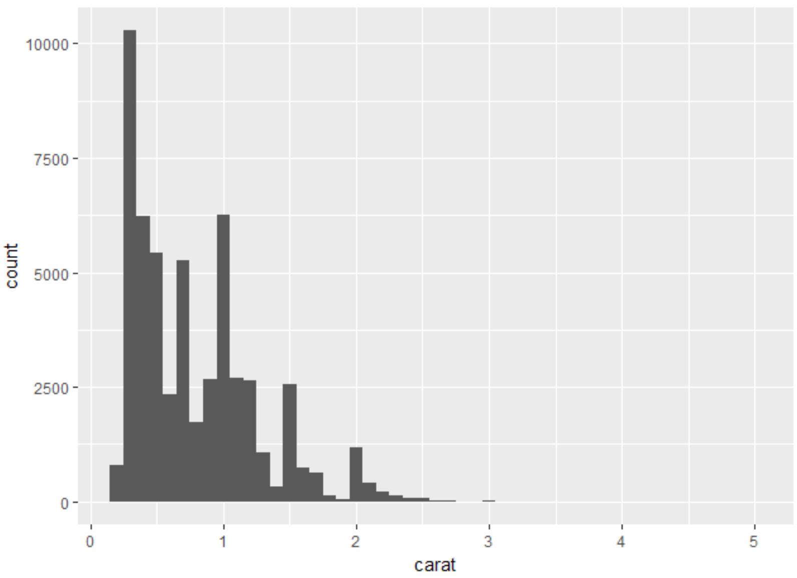


The graph shows that the .jpg format has the largest file size for all values of N. However, the size of the .jpg's increase more rapidly for values of N below 2000 and then the growth of the file size decreases as it grows to 5000. The growth rates for the .pdf and .eps formats appear constant. The .png format has the smallest file size, but it also shows a faster rate of growth for values of N below 2000 and then its rate of growth appears to slow as N grows from 2000 to 5000.

5. Diamonds

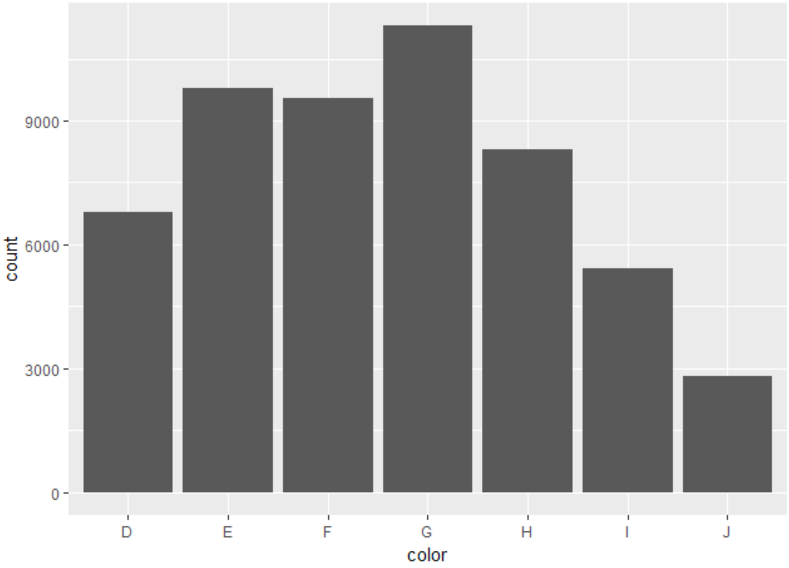
For the diamonds dataset, I created a series of histograms based on carat, price, and color using the code:

```
ggplot(diamonds, aes(x=color)) + geom_histogram(binwidth=1, stat="count")
ggplot(diamonds, aes(x=carat)) + geom_histogram(binwidth=0.1)
ggplot(diamonds, aes(x=price)) + geom_histogram(binwidth=10)
```

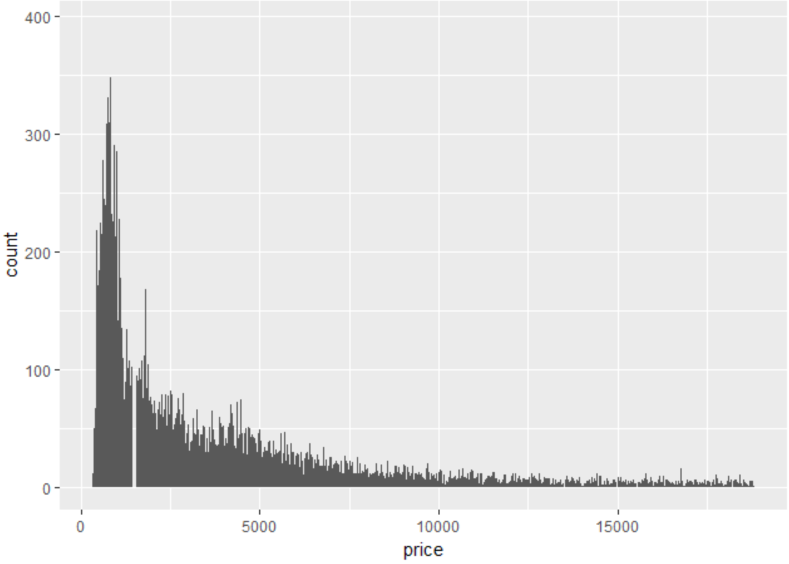


For the histogram by carat, the most commonly occurring carat is less than 0.5. However, the histogram also shows a high concentration of diamonds at even-numbered carats such as 1, 1.5, and 2. This indicates that the carat of many diamonds fall on even values with fractional values being less common. There are also very few diamonds with carats greater than 3, although the highest carat value in the dataset is 5.1. Therefore, the dataset has a tail that extends from 3 to 5.1, but the number of diamonds with those carats are too small to be visible in the histogram.

I used a binwidth of 0.1 which seemed to be the optimal binwidth to show the pattern of the number of diamonds by carat. A smaller binwidth made the histogram rather noisy and the pattern of larger numbers of diamonds falling on carats of 1, 1.5, and 2 was less apparent. Increasing the binwidth above 0.1, however, made the histogram coarser and also made the pattern less clear.

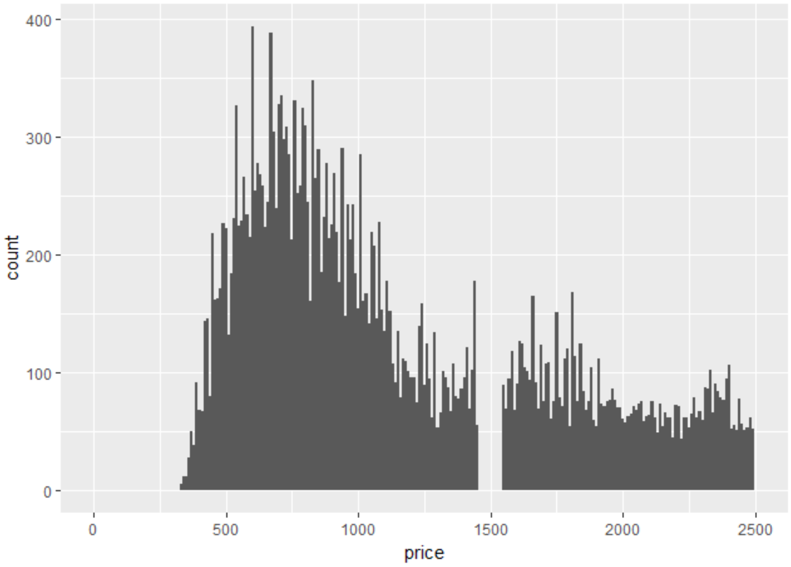


The histogram showing the number of diamonds by color shows that J is the least common color while G is the most common. The colors surrounding G such as E, F, and H are the next most commonly occurring diamonds in the dataset, indicating that these four colors account for the bulk of the diamonds. However, the dataset clearly shows a skew towards the lower end of the distribution since the counts for colors D and E are much higher than the counts for the colors I and J at the opposite end of the distribution.



The histogram for price shows a strong skew towards the lower end of the distribution, with the highest counts falling below 2500. It also shows a relatively steady decrease in the count as the price increases. However, the dataset shows a long tail in which a fair number of diamonds have prices over 15,000, with the highest price in the dataset being 18,823.

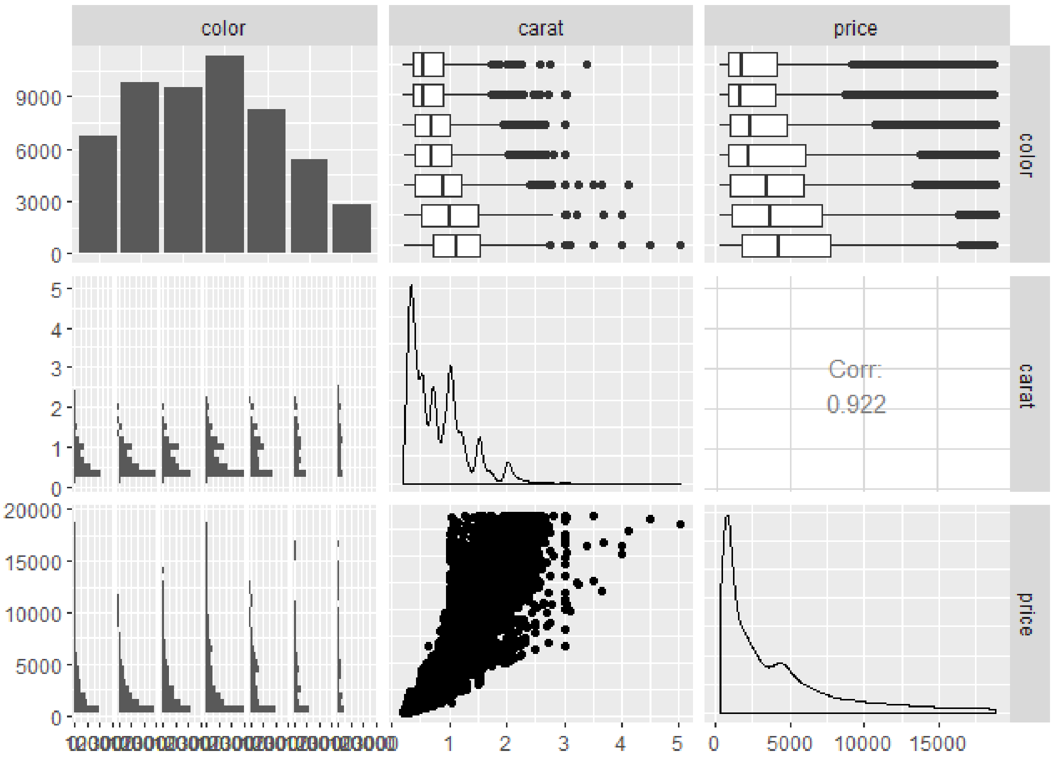
One interesting observation is the lack of diamonds around the 1500 price point. To investigate this further, I added an xlim(0,2500) parameter to limit the histogram to prices below 2500:



As the histogram shows, there are no diamonds around the 1500 price point. Examining the diamonds dataset shows that prices generally increase in increments of 1 until the price of 1454 and then jumps to a price of 1546. Why no diamonds have prices that fall within the range of 1454-1546 is unknown, but the most likely explanation is that data is simply missing for diamonds in that price range.

Finally, I created a pair-wise plot showing the relationships between color, carat and price using the code:

```
ggpairs(diamonds, columns = c("color", "carat", "price"))
```



In the pair-wise plot, the histograms for color, carat, and price are very similar to the individual histograms above. However, the pairwise plot includes box plots comparing color vs. price and carat vs. color in the top row. From the color vs. price box plot, it is clear that the median price increases as color increases, which explains why the color histogram is skewed to the lower end and the I and J colors had the lowest counts, since the less common colors are also more expensive. The carat vs. color box plot shows a similar pattern, with the I and J colors having higher median carat values.

The plot also shows histograms comparing carat vs. color and price vs. color. These also show a similar relationship as the box plots, namely that the higher carat diamonds tend to be found in the rarer colors and that price increases with those same colors. The bottom row, center column shows a scatter plot relating price and carat. As can be expected, price increases as carat increases. The middle row, right column shows the correlation between price and carat at a very strong 0.922.

References:

1. QQPlot based on the ToothGrowth dataset derived from:
   QQ-Plots: Quantile-Quantile plots - R Base Graphs. *Statistical tools for high-throughput data analysis.* Retrieved June 9, 2018, from: http://www.sthda.com/english/wiki/qq-plots-quantile-quantile-plots-r-base-graphs