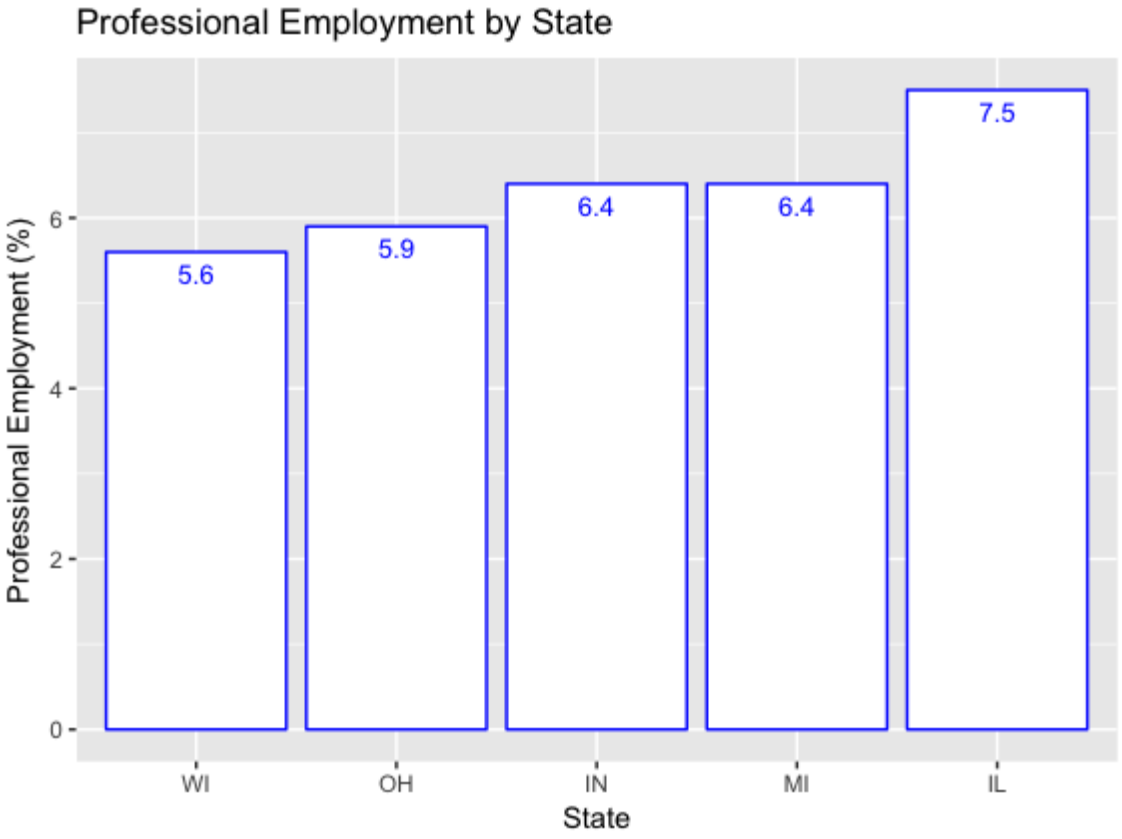
dgiron3

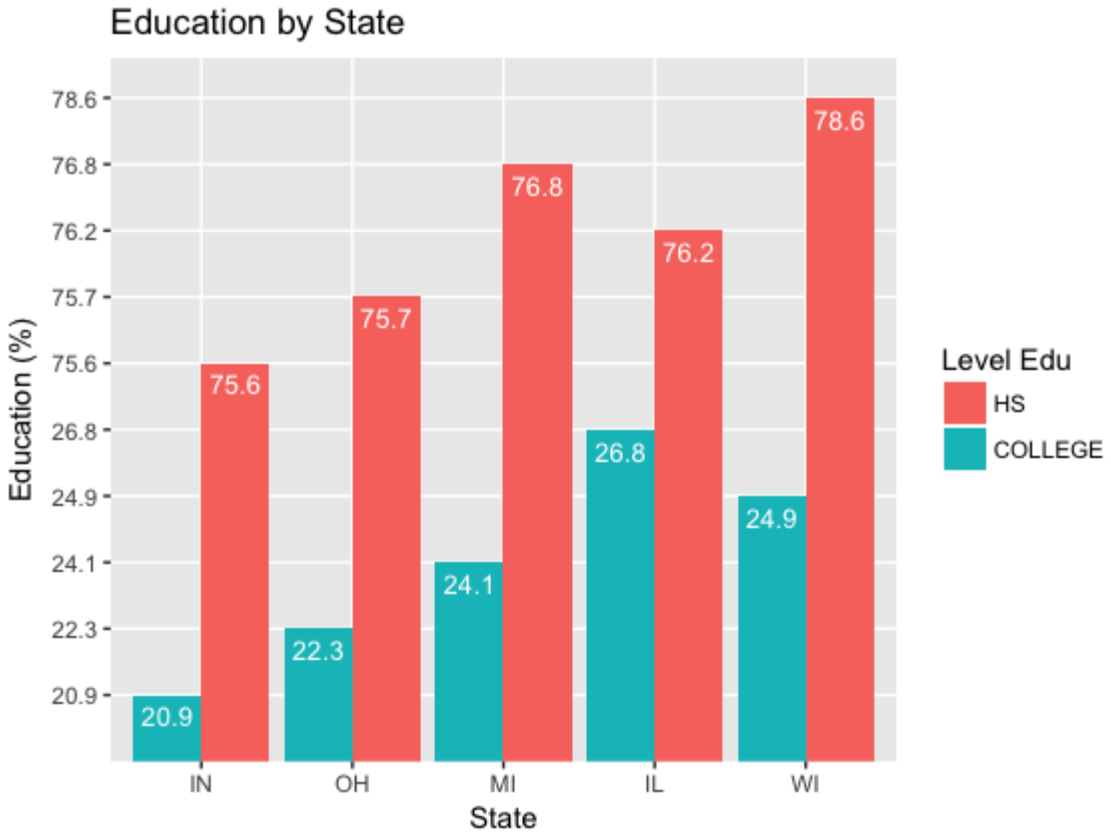
HW 2 Report

1. Interpretation A: Through data manipulation, I was able to group by state to include pertinent professional employment information (total number of people with professional employment in each county and then in each state). In doing so, I was able to determine the total percentage of professional employment in each state. Once I had this data, I used ggplot to give a graphical representation.



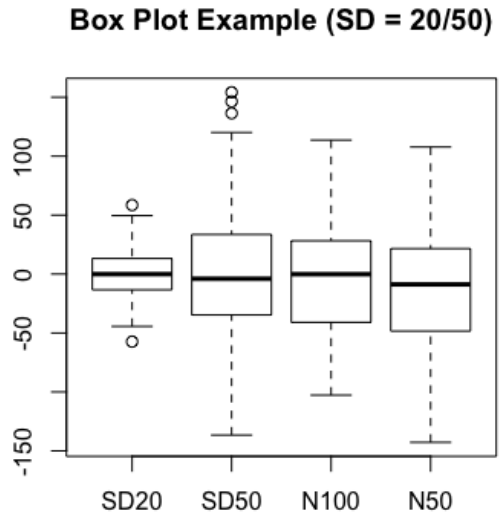
The above graph clearly shows that the states in the Midwest region have relatively similar professional employment percentages (all within the range of 2% of each other). Also, it can easily be concluded that Wisconsin has the lowest professional employment percentage (5.6%) and Illinois has the highest professional employment percentage (7.5%). These percentages were of the total adults population in each respective state.

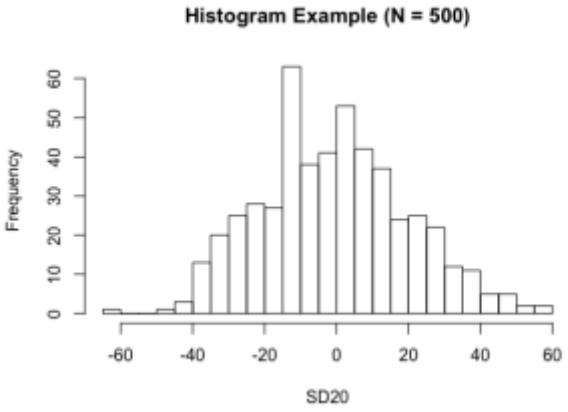
2. Interpretation A: Similar to problem 1, through data manipulation, I was able to group by state to include pertinent education information (total number of people with high school and college educations in each county and then in each state). In doing so, I was able to determine the total percentage of people with high school and college degrees in each state. Once I had this data, I used the melt() function in addition to ggplot to make a multi-bar graph displaying percentages for both high school and college degrees in each state.

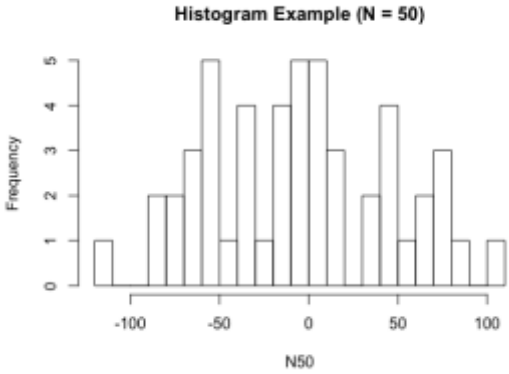


While the graph is not to scale, it does a great job at showing the relationship between the percentage of people with a High School diploma and the percentage of people with college education for each of the 5 states in the region. For the most part, there is a positive correlation between the percentage of people with high school diplomas and the percentage of people with college education. The states with the two lowest percentages of high school diplomas also have the two lowest percentages of college educations. Wisconsin has the highest percentage of high school degrees and the 2nd highest percentage of college degrees of the 5 states. Illinois has the highest percentage of college degrees and is right behind Michigan with the 3rd highest percentage of high school degrees. In all states, the difference between the percentage of college degrees and the percentage of high school degrees is roughly within the range of 50% and 55%.

3. A box plot is used to display the distribution of data. The box plot shows the minimum and maximum values of the data set. It also shows the first quartile, the median and the third quartile. A box plot is best used for summarizing large quantities of data. As shown in the Box Plot Example to the right, there are 4 different box plots. The first 2 (SD20 and SD50) have N = 500 data points and mean = 0. The only difference is the standard deviation (20 and 50 respectively). This box plot does a good job at showing SD20 has a narrower average range while the SD50 has a wider average range. This makes sense due to their known standard deviations. The last three box plots all have the same mean = 0 and standard deviation = 50, but their N values are 500, 100, and 50 respectively. Based on these box plots it is difficult to really see their differences. Also, a box plot does not do a good job of displaying the mean or mode of the data set. These are cons of the box plot. As also shown to the right, the histogram of when N = 500 is a very different picture to the histogram of when N = 50. The historgram shows the frequency distribution of a data within equal intervals. Same with box plots, histograms are only used for numerical data. Also, they can do a better job at giving and idea where the mode range will be between (given the interval range).

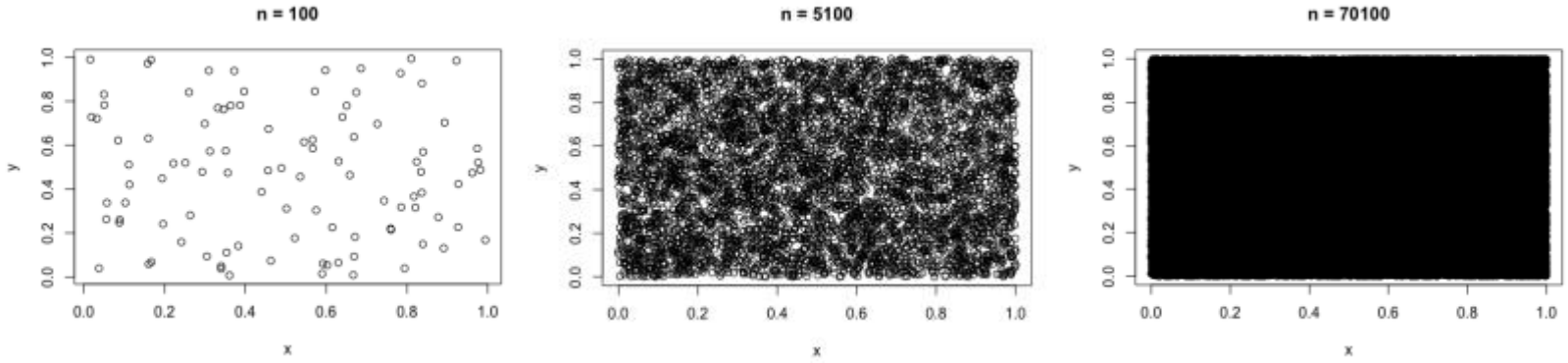




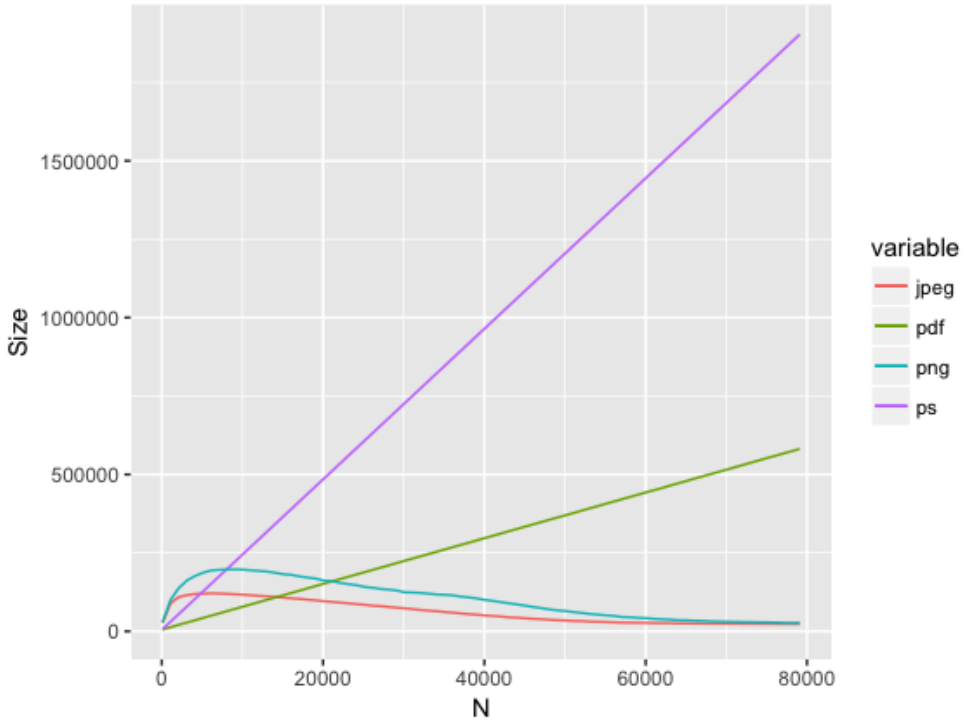


When deciding when to use a box plot, histogram, or QQPlot, it is important to understand what metrics you are searching for to see visually given the sample of numbers. If you are given large data sets, and you wish to know where the median will fall and where the max, min, and quartiles will fall, choose the Box plot. If you wish to know how the data is distributed based on interval frequency, choose the histogram. If you wish to see if the distribution of data has a normal or exponential distribution, choose the qqplot.

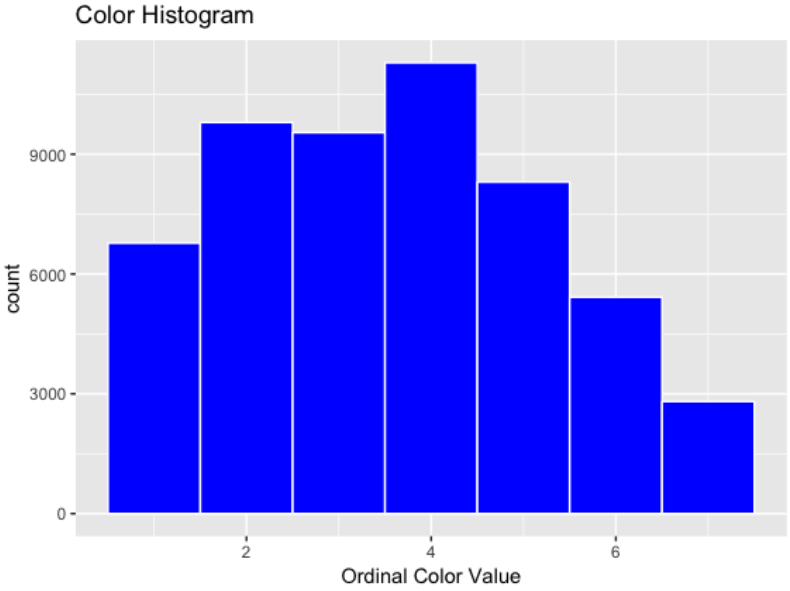
4. After testing the different file types and the sizes for each file type given the n value, I was able to find very interesting results. Here are examples of 2 scatter plots generated:

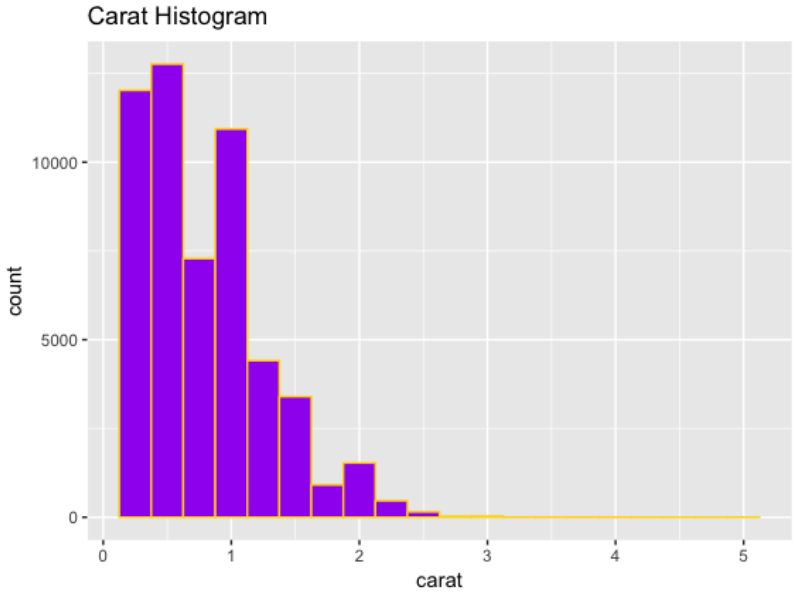


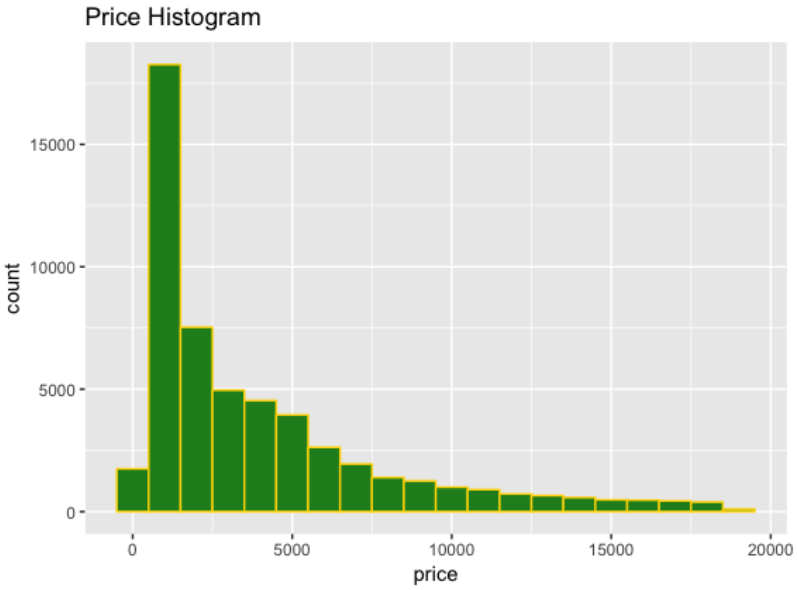
As shown above, as n increases, the scatter plot gets more and more filled in until it is just a seemingly blacked out rectangle. At first, I did not account for the higher values of n (n > 10,000) and I assumed that the file size would get bigger and bigger for each file type. Then when I ran the final test taking n over 70,000, I was able to see that the file size actually shrunk for the picture files (png and jpeg). This started to make sense the more I thought about it. Another interesting take away was that the pdf and ps file sizes never shrunk, no matter how large the n value got. The post script file was significantly the largest once n surpased 10,000. The png was the largest for most of n < 10,000. Also, another intersting observation was that the pdf and ps lines had a linear relationship while the other two obviously did not.



5.







From looking at the 3 histograms, there are few that first come to mind. The Color histogram is the most stochastic out of all them, because it is not highly front loaded like the other two. This tells us that the majority of diamonds are of average color. In fact, there are somewhat more higher leveled colors than lower leveled colors (lower number means higher level). The carat and price histograms have a trend of decreasing counts as the number of carats increases and as the price increases. This means that the majority of diamonds have a small number of carats (1 or less) and are around the same price of $1000-$2000. Knowing that the colors are ordinal variables, it will be interesting to see the three way relationship between these variables.

As shown below, I plotted scatter plots, based on the color of the diamonds with the number of carats along the x-axis and the price along the y-axis. As shown below, as the level of color decreases (D>E>F>G>H>I>J) the price per carat also decreases (the positive correlation slowly flattens out for the most part). Another observation, is that higher leveled colors do not have as many higher numbered carats. Most D diamonds are well less than 2 carats. On the other hand, 2 carats seem to be around the average for J-colored diamonds.