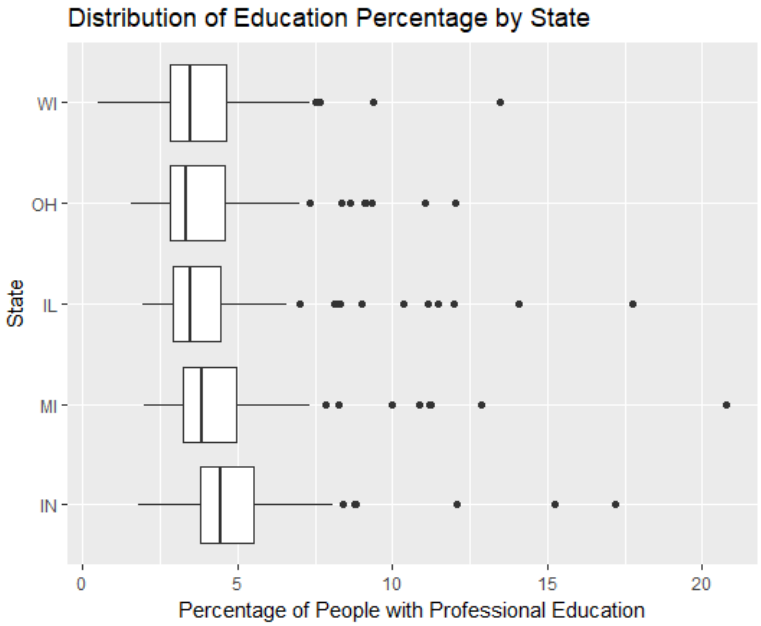
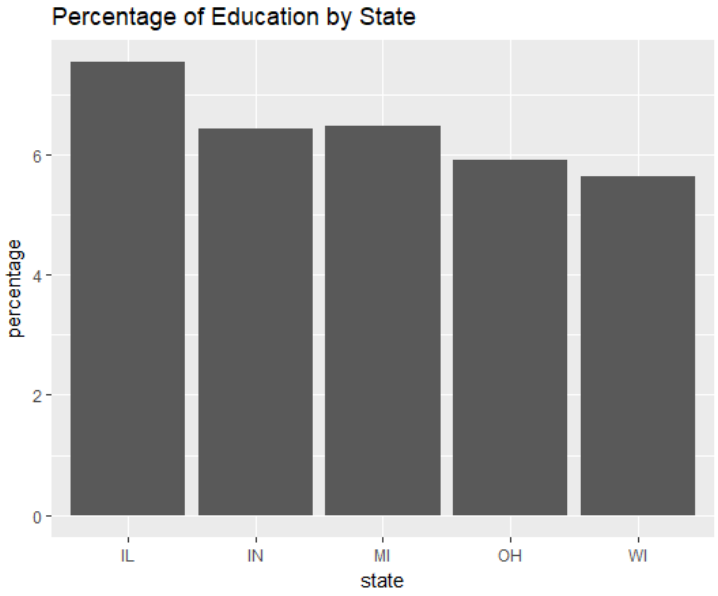
hyang390
903320189

Homework 2: Data Visualization

1. Professional Education by State [20 points]



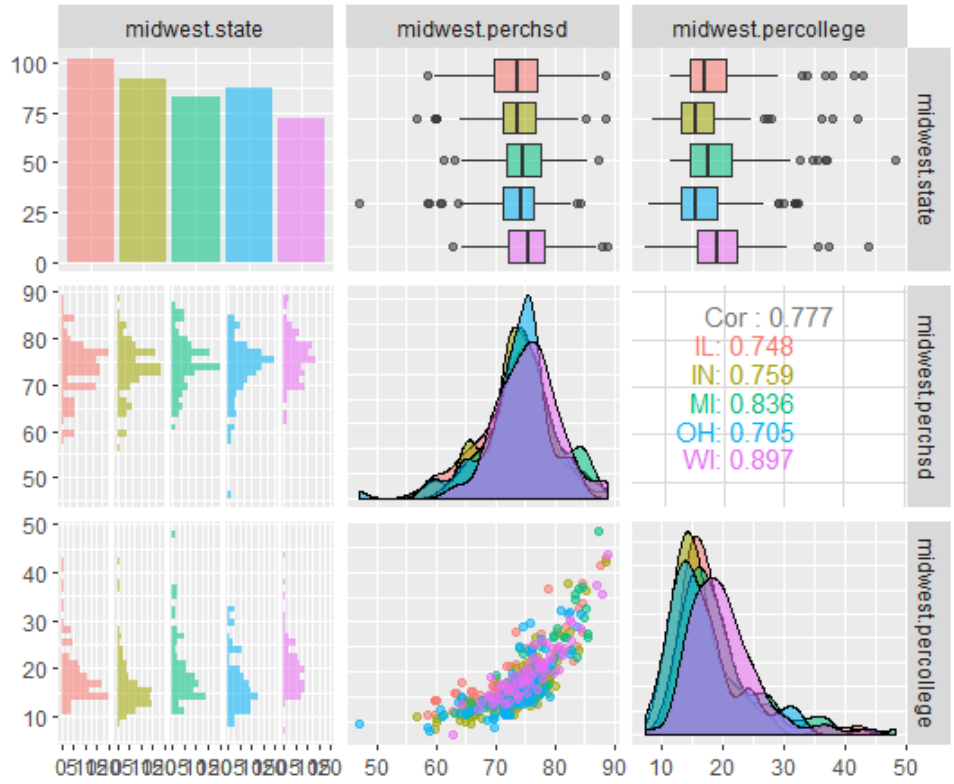
The above box plots illustrate the distribution of the percentage of people with professional education for each state as ordered by the mean. As shown from the lecturers, the box plots show the IQR as well as the outlier counties for each state, and with the middle line of each box plot indicating the median of each state. As we can see from the figure above, Indiana (IN) has the highest average of percentage of people with professional education, and Wisconsin (WI), Ohio (OH) and Illinois (IL) have around the same median of the percentage with professional education, with OH having a slightly lower median. OH does have more outlier points, which probably boosted its mean value. Also, it's interesting to point out that MI has the largest spread and IL has the most number of outliers.



hyang390
903320189

I also wrote some code to aggregate the total educational population of each state and this is then divided by the total population of the respective states to retrieve the percentage population of each state with professional education, the bar graph is shown above. As clearly shown, WI has the lowest population with professional education and IL has the highest population with professional education. Note this is a slightly different result than just looking at the mean (ordering) and median (line in box plots) from the box plots.

2. School and College Education by State [20 points]



For this question, I used the library *GGally* to compare the three-way relationships between state, perchsd (percentage of high school educated population per county) and percollege (percentage of college educated population per county). By observing the plots above, we can see that there is positive correlation between perchsd and percollege, in fact this is calculated to be about 0.777.

Then, looking at the relationship between state and percollege, we can find that WI has the highest median and IQR of population with college education, and IN has the lowest median value and IQR with college education. MI has the largest spread among the five states.

Finally, looking at state and perchsd, we can see that again WI has the highest median of people with high school education and IN and IL are tied for the lowest median for percentage of people with high school diplomas. However, IL has a far larger IQR compared to IN. It's interesting to note that OH has one extreme county on the left side of the IQR, which signifies it is the county with the lowest percentage of people with high school diplomas out of the five states.

hyang390
903320189

3. Comparison of Visualization Techniques [20 points]

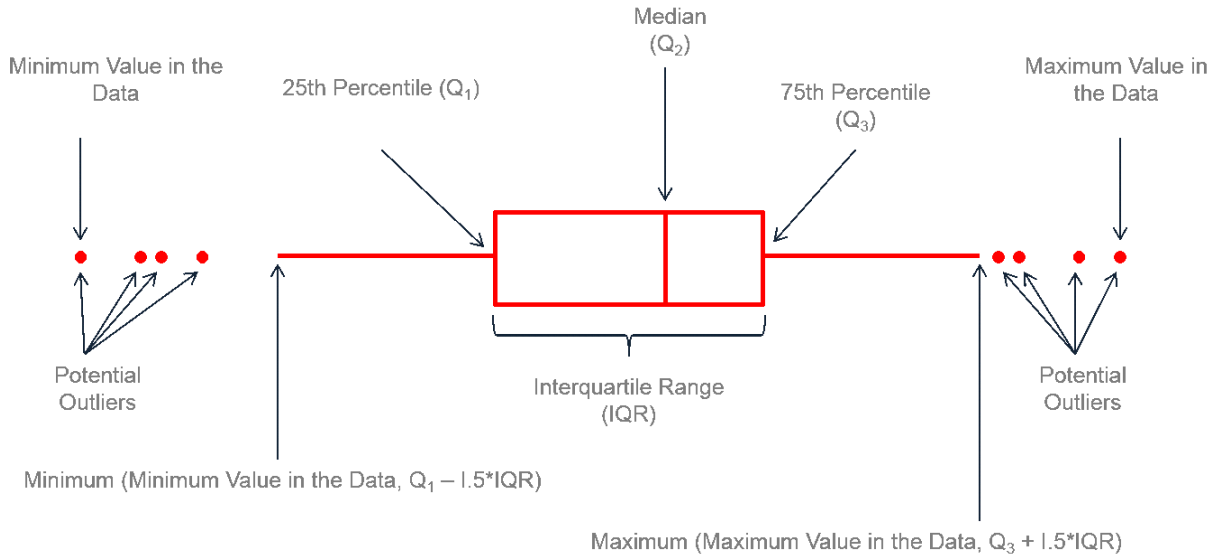


Figure Reference: https://www.leansigmacorporation.com/box-plot-with-minitab/

As we learned in the lecturers, a box plot shows the distribution of a dataset. Specifically, it shows a box that includes the difference between the 25 percentiles to 75 percentiles of the data in a box, with a line within the box that denotes the median of the dataset. This box is known as the interquartile range (IQR), since it includes the middle 50% of the data set. The ends of the whiskers denote the maximum value in the data on the right and the minimum data on the left, with the right being Q3+1.5*IQR and the left being Q1-1.5*IQR. Any values that lie beyond these whiskers are known as outliers. The spread of the data is also shown in the box plot by the range from the left most outlier to the right most outlier, or the whiskers if there are no outliers.

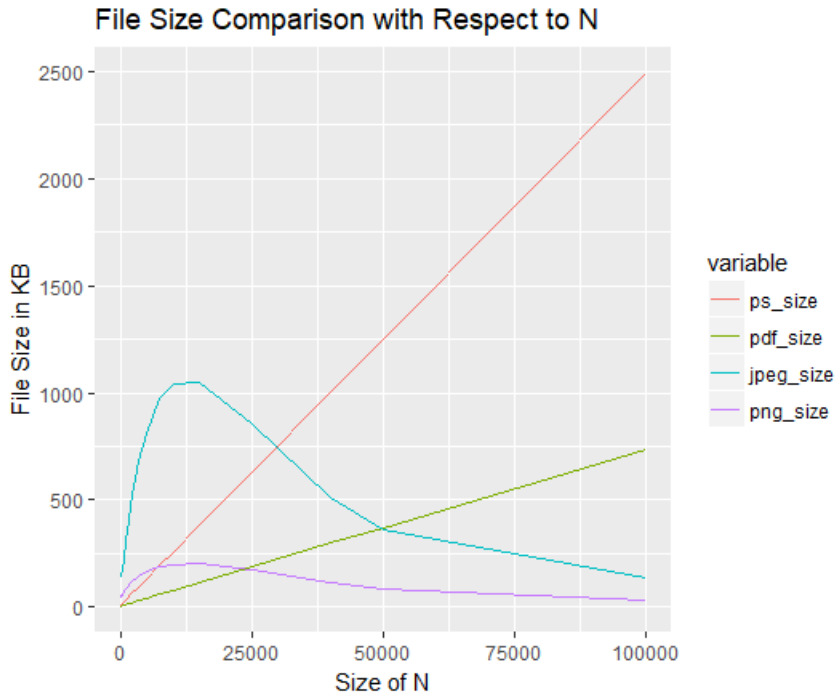
A histogram is an easy way to view the distributions of a numeric data set. Histograms are very easy to read and if the correct bin width is used, it is very effective in presenting the specific distributions of the dataset. If, however, the bin width is too big, then the distributions shown will lose accuracy. Similarly, if the bin width is too small, the histogram will be smoother, but it will be harder to deduce patterns. Another con of using histograms is that it is difficult to compare multiple variables side by side as they are not visually effective. A good use case of histograms will be to see the distribution of a country's population with each decade of life being a bin. This is a good distribution of one variable and the bin width will clearly show the distribution of each age group.

A box plot is also effective in showing the distribution of a dataset. Compared to a histogram, it provides more details, such as the clear median of the dataset, the box showing the middle 50% of the data set (IQR), the spread as well as the outliers of the dataset. The box plot is also very effective in comparing different variables. On the other hand, it is not particularly useful with a small dataset since the quartiles might be meaningful enough, it is also not effective in seeing some detailed patterns in the distributions, such as clustered distributions we saw in the lecturers regarding eruption times. A good example is what we saw above, comparing the distributions of populations that are educated by state, these are multiple large datasets that are comparable, and plotting box plots allow us to compare them side by side and look at the individual characteristics of the datasets, such as the medians.

hyang390
903320189

Lastly, the QQ plot is an excellent tool to compare the distributions between two datasets, it is a scatter plot of quantiles of one data set on the x axis versus the quantiles of the other dataset on the y axis. One of the datasets can be a sample dataset of a theoretical distribution. The slope of the dataset denotes the relationship of the two distributions. One example usage of a QQ plot might be comparing the distribution of a country's population with a theoretical normal distribution to if the country's population distribution is indeed close to a normal distribution.

4. Random Scatterplots [20 points]

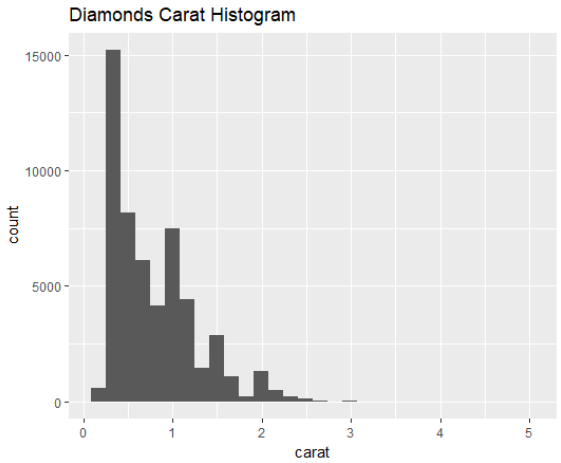


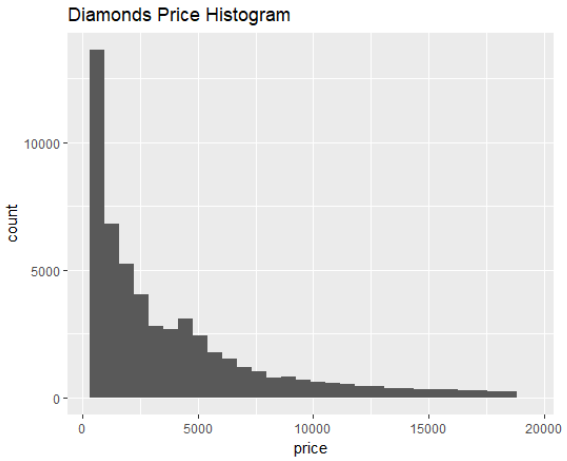
For this problem, two different datasets are generated using the *runif* function and plotted in a scatterplot with *ggplot2.* I then saved the figures using the four different file types with *ggsave* and calculated their sizes with *file.size*.

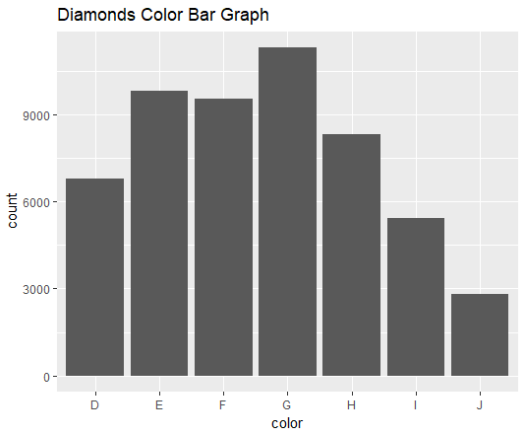
Then, I plotted the file sizes in kilobytes of the different types of files with N ranging from 50 to 100000. As seen above, the size of pdf and ps files increase linearly with respect to the size of N, with the size of ps files increasing much faster than the size of pdf files. Jpeg and png files, on the other hand, increase for smaller values of N, then gradually decreases as N increases beyond 10000, and stabilizes as N gets bigger. For smaller values, the size of jpeg files is the largest, followed by png, then ps and pdf files. However, as the value of N gets larger, the order now becomes ps being the largest, followed by pdf, then jpeg and lastly png files being the smallest.

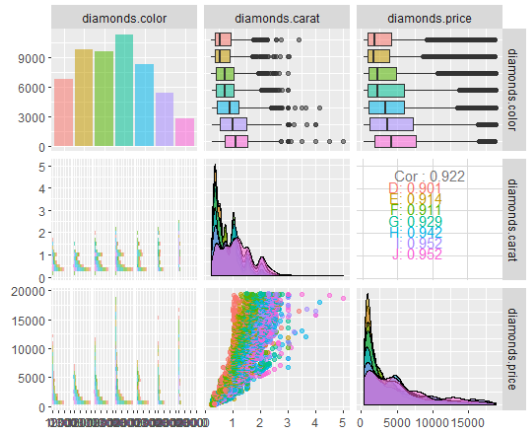
hyang390
903320189

5. Diamonds [20 points]









Looking at the carat and price histograms, we can see that the distributions is skewed to the right, which signals the mean is greater than the median, and the mode lies far towards the left. For example, for carat, we can see the mode of the data is around 0.25 carats; and for price, the mode is around 1500. As for the color of the diamonds, we can see that color G is the most often appearing, and color J appears the least.

As for the three-way relationship between carats, price and color, we can first see there is a positive correlation between carat and price, the higher the carat, the higher the price. In the box plot above, it is also noticeable there is also a positive relationship between carat and color. For example, color J has the largest median of carats, it is also the rarest, this applies for the other colors as well. As rarity of the color increases, the higher the carat of the diamond. There is also a positive correlation between price and color, with the rare colors being more expensive. For example, we know from our color histogram that I and J occur more rarely, and in the box plot, we can see they have larger medians compared to other colors.