Marina E. Johnson - mjohnson393

Homework 2

1. Professional Employment by State

According to Figure 1, we observe that in all of the states, some counties have much higher percentages of professionals as compared to the other counties in the same state (outliers). For example, 1st and 3rd quartiles of IL are less than 5. However, there is a county in IL whose percentage of professionals is above 17.5%. Probably, these outlier counties with a lot of professionals are urban areas, such as Madison County (IL) where Chicago is or Franklin County (OH) where Columbus is. According to the graph, 1st and 3rd quartiles of IN are greater than other states. I believe, if we conduct a t-test, we will see that the mean percentages of professionals in these states do not differ significantly.

Figure 1.1

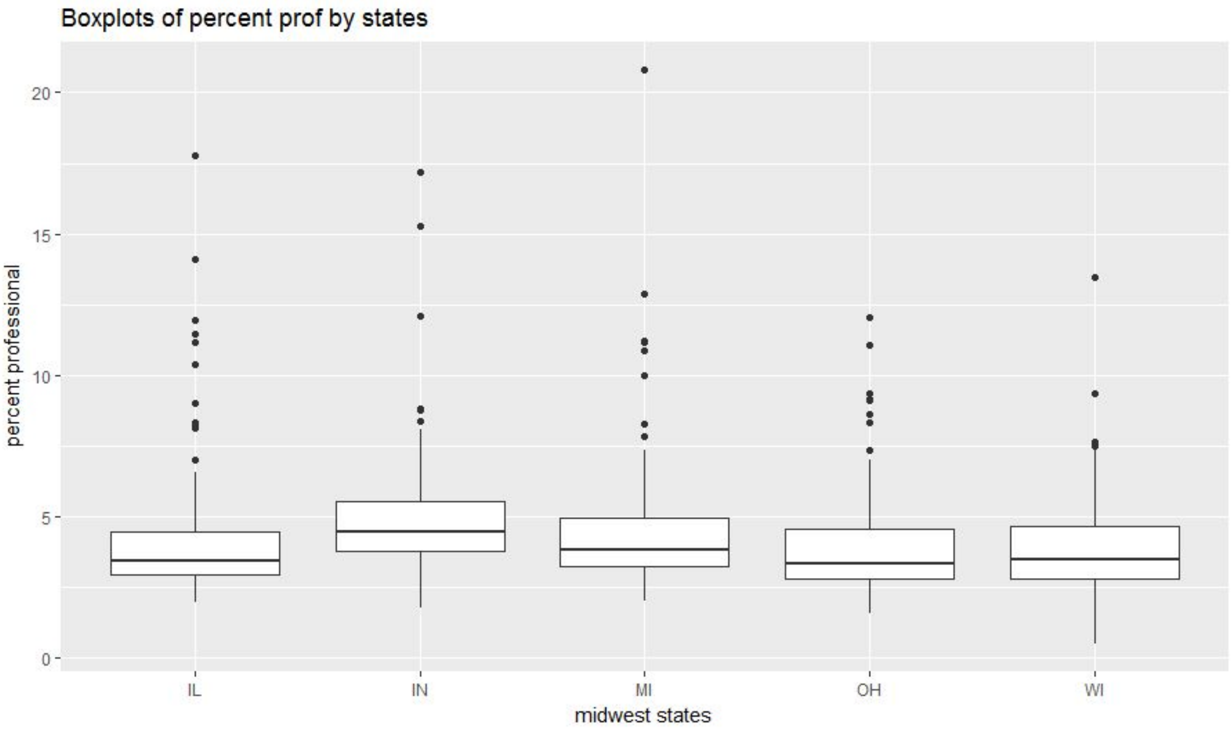
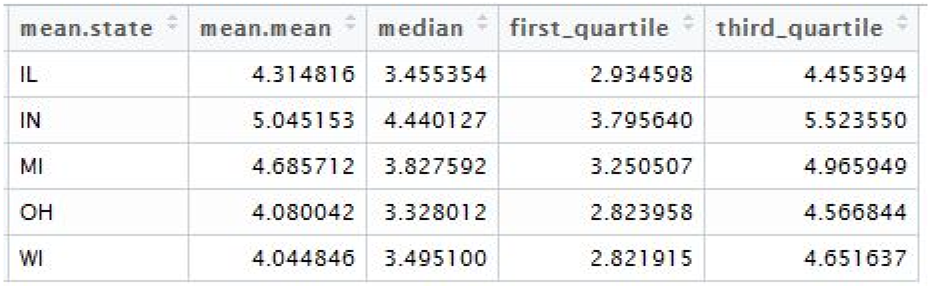


Table 1.1 below illustrates the descriptive statistics of the percent professional in counties by Midwest states. This is actually the representation of the boxplots above. The bottom sections of the boxes are the 1st quartile, the lines in the middle of the boxes are the means, and the top section of the boxes are the third quartile. The total length of the box plot including the vertical lines cannot exceed 1.5 times of the box size (3rd quartile - 1st quartile). The table below proves that in average counties in IN has higher percentage of professionals as the first and third quartile of IN are greater than of other states.

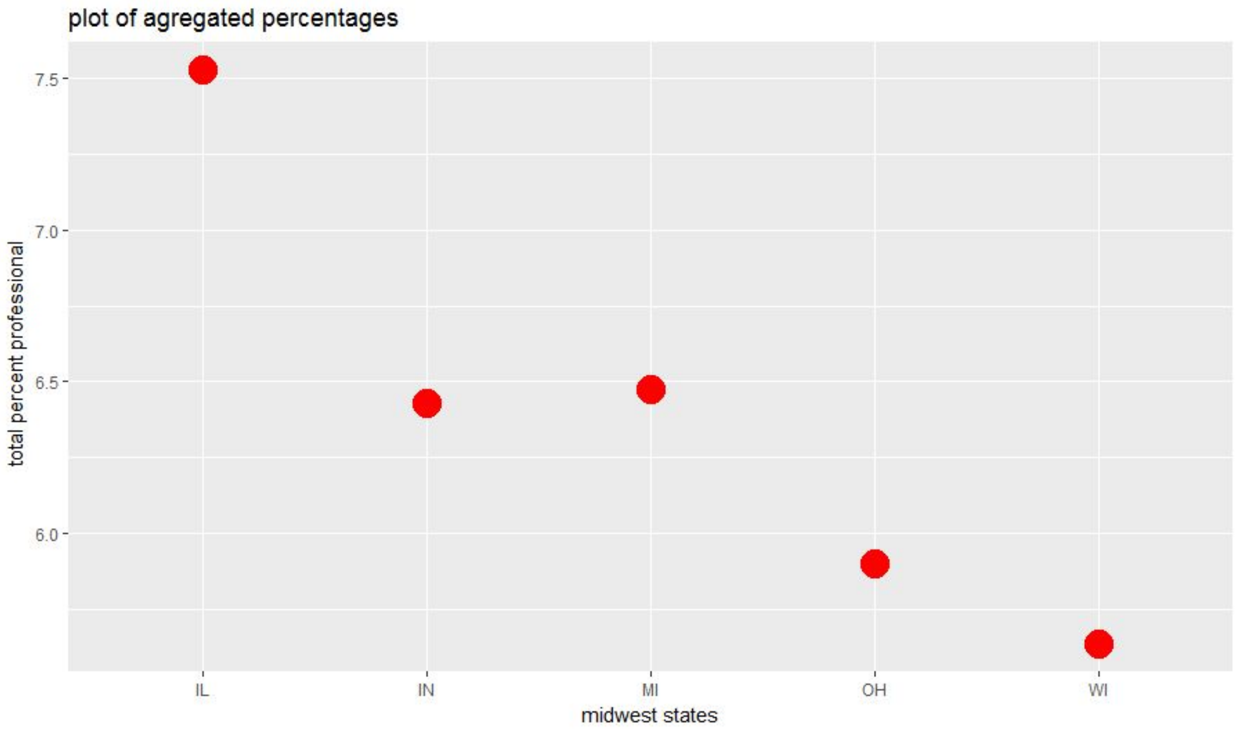
Table 1.1



mean.state	mean.mean	median	first_quartile	third_quartile
IL	4.314816	3.455354	2.934598	4.455394
IN	5.045153	4.440127	3.795640	5.523550
MI	4.685712	3.827592	3.250507	4.965949
OH	4.080042	3.328012	2.823958	4.566844
WI	4.044846	3.495100	2.821915	4.651637

When we aggregate the percentages using weighted average based on population, we see that IL has the highest percentage of professionals. This is not surprising given that Chicago has a lot of opportunities for professionals. WI has the lowest percentage of professionals. The total number of professionals in WI is 275,714 (less than 6%) whereas IL has 860,467(more than 7.5%) professionals.

Figure 1.2



When we aggregate the percentages using the weighted average based on population, we see that IL has the highest percentage of professionals. This is not surprising given that Chicago has a lot of opportunities for professionals. WI has the lowest percentage of professionals. The total number of professionals in WI is 275,714 (less than 6%) whereas IL has 860,467(more than 7.5%) professionals.

2. School and College Education by State

As seen in Figure 2.1, regardless of the states, there is a positive correlation between the high school and college percentages. If a county in a state has greater high school graduate percentage, it is likely that they have higher college graduate percentage. It also seems that the dots on the WI scatter plot forms a narrower cluster; this means that the correlations between the high school graduate % and college graduate percentage in WI and MI are greater than of the other states. It also seems that the average college graduate % is about 20% regardless of the states. However, there are some counties in each state that they have substantially more high school and college graduates. Probably, these counties are urban areas, such as Indianapolis, Chicago, Cincinnati, etc.

Figure 2.1

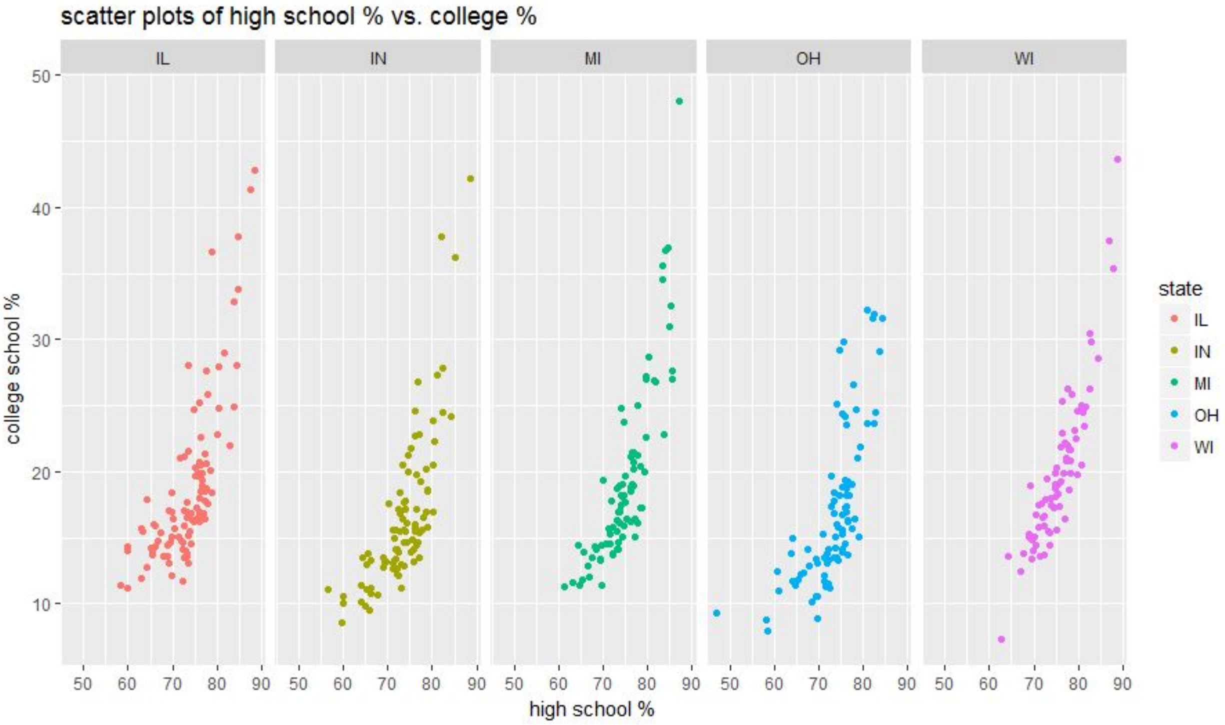


Figure 2.2 illustrates the boxplots of college and high school graduate percentages by state. It appears that IL has the highest variation in high school graduate percent because Illinois's boxplot is the tallest boxplot. The high school graduate percentage means of each state are almost the same (the line in the boxes). There are not many counties with a significantly more high school graduates (outliers). However, there are numerous counties with much more college graduates within the same state (dots above the lines). It also appears that IN and OH have counties with fewer college graduate percentage as compared to the counties in other states.

Figure 2.2

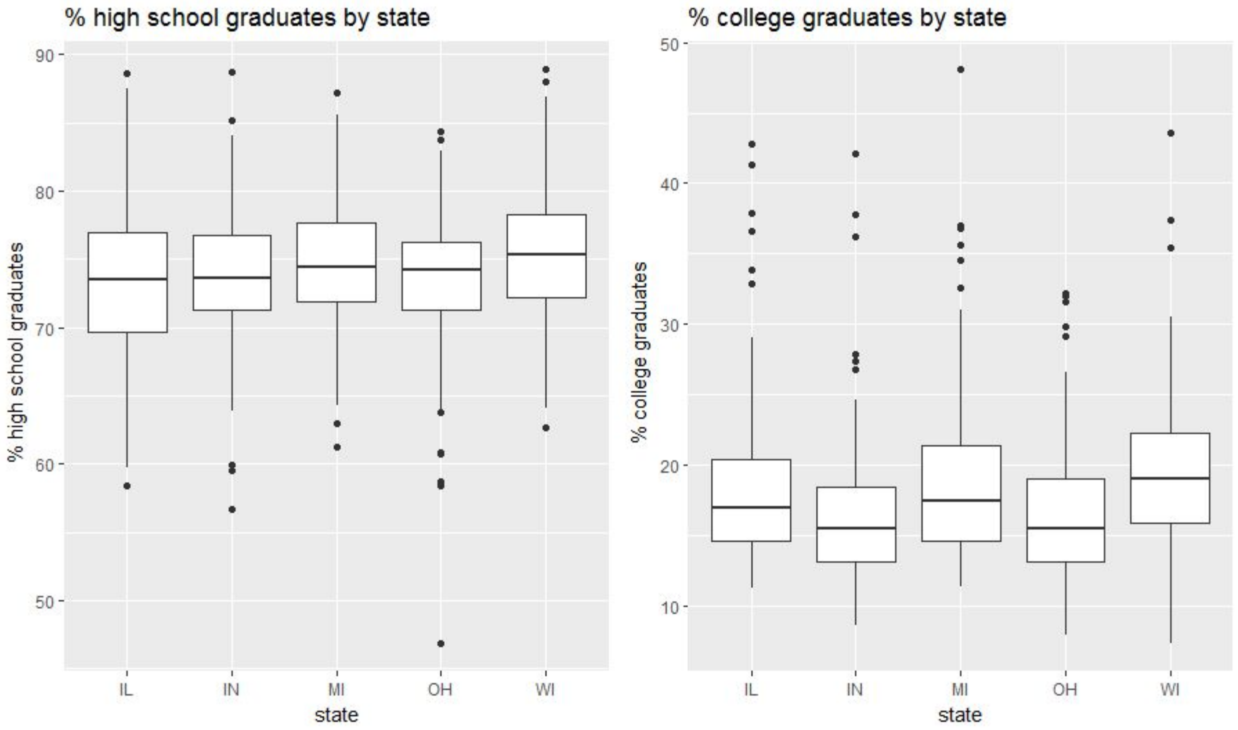


Figure 2.3 shows the aggregated college and high school percentages by state. As indicated before, almost all states have similar high school percentages (around 80%). However, we can't say the same thing for the college graduate percentages. It appears that IL, MI, and WI have greater college graduate percentages as compared to OH and WI.

Figure 2.3

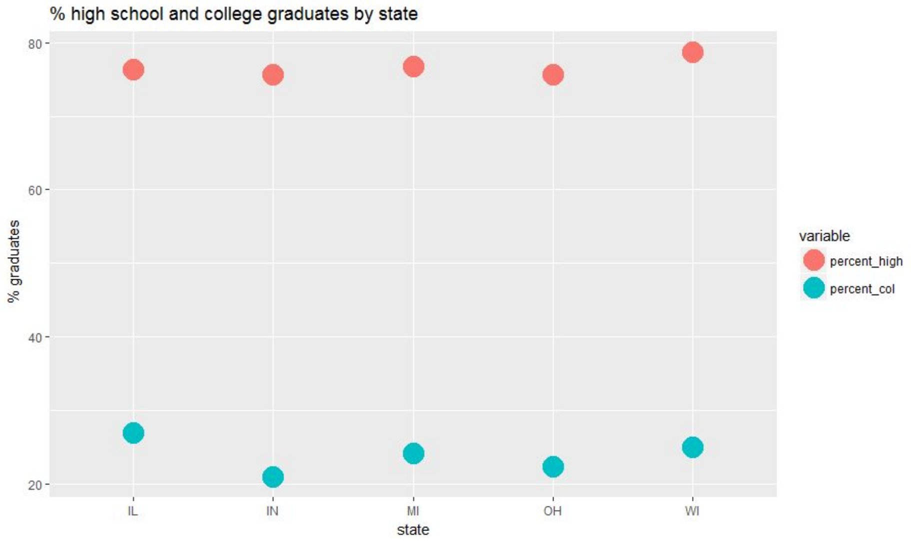
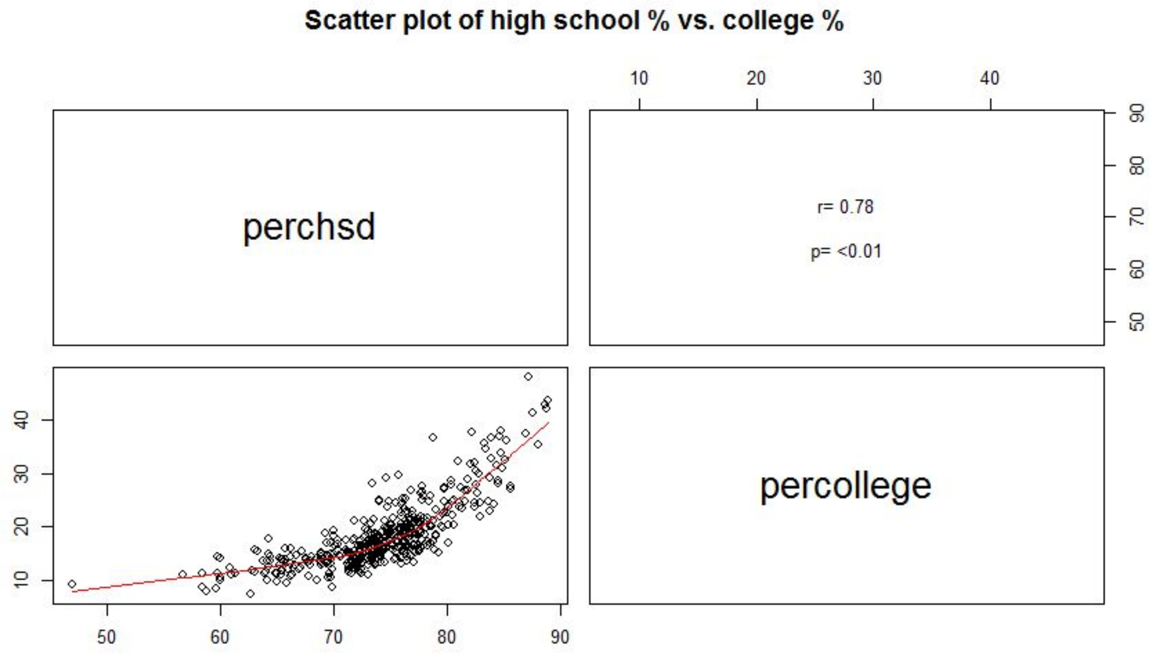


Figure 2.4 just analysis the 2-way relationship between the two numerical values (college % and high school %). The graph below proves that there is a very strong correlation (r = 0.78) between the high school graduate % and college graduate %. We can say that counties that have more high school graduates tend to have more college graduates.

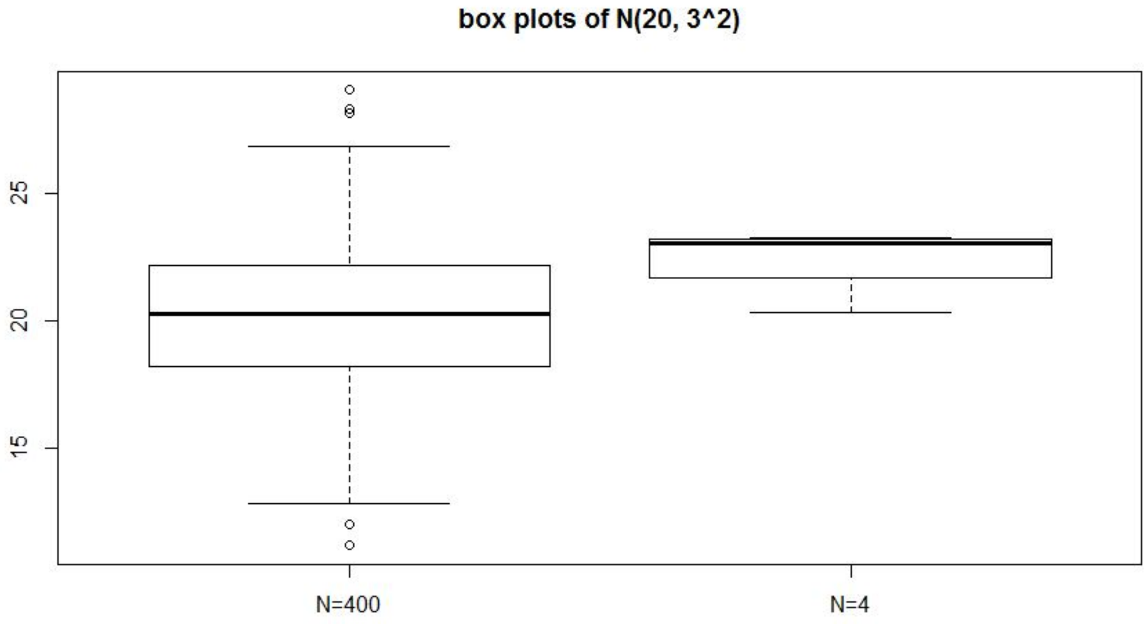
Figure 2.4



3. Comparison of Visualization Techniques

Box plots have three important statistics on them: first quartile, mean, and the third quartile. With the help of these statistics, we can observe the variation of a dataset and whether or not the data is skewed. Small sample sizes will cause misinterpretation. I will explain this phenomenon with an example. Checking Figure 3.1, we would think that N=400 has much greater variability than N=4. But, both data sets were generated by randomly sampling from a normal distribution with a mean of 20 and a standard deviation of 3 ~ N(20, 32). That is, the data for both plots come from the same population. The reason why boxplots are so different is because N=400 contains 400 data points whereas N=4 contains only 4 data points. The small sample size shrinks the whiskers and gives the boxplot the illusion of decreased variability as seen in the figure below. Another issue with using a boxplot with small samples is that the first and third quartiles will become meaningless. For example, if you have only 4 data points, it makes no sense to display an interquartile range that shows the "middle 50%" of the data (interquartile range is also called the fourth spread and calculated by subtracting the first quartile from the third quartile). This issue is also seen in the figure below. When N=4, the quartiles serve no purpose.

Figure 3.1



Boxplots and histograms have a lot of similarities. They both try to convey the underlying distribution of the data. The histogram gives a more detailed picture of the distribution of the data. However, we sometimes are not interested in the exact distribution or overall shape of the data. Instead, we want to compare several data sets. In this case, plotting different datasets on a boxplot side by side will help us to compare the quartiles of the datasets and see their fourth spread values. We also need to note that the histogram does not contain several statistics that the boxplot has, such as the median, first and third quartile. I would say the histogram is useful if we are interested in the distribution of one single dataset. The boxplot is useful if we want to compare several data sets. The Q-Q plot will be a great option if we want to make sure that two datasets are coming from the same distribution. If it is the case, the Q-Q plot will follow a 45-degree line. So the Q-Q plot will not tell us about the median, quartiles, or other things. The Q-Q plot can only compare two data sets and indicate if their probability distributions are the same.

4. Random Scatter Plots

Figure 4.1 illustrates two scatter plots: (1) two uniform distributions with 100000 data points and (2) two uniform distributions with 100 data points.

Figure 4.1

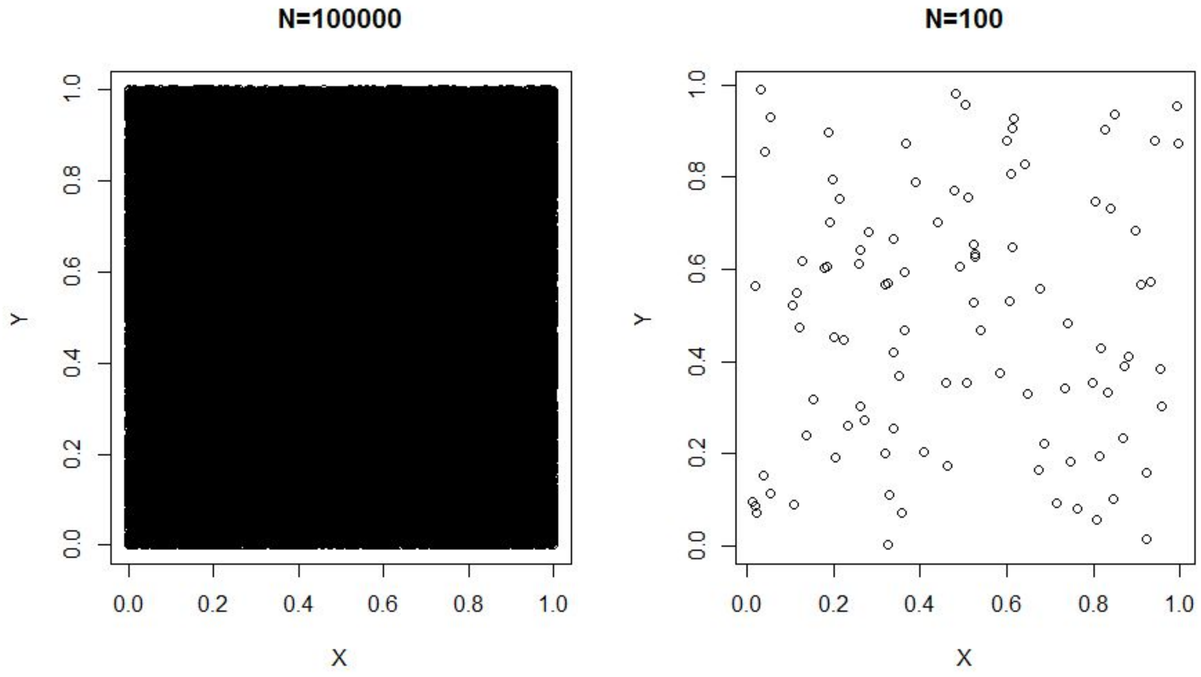
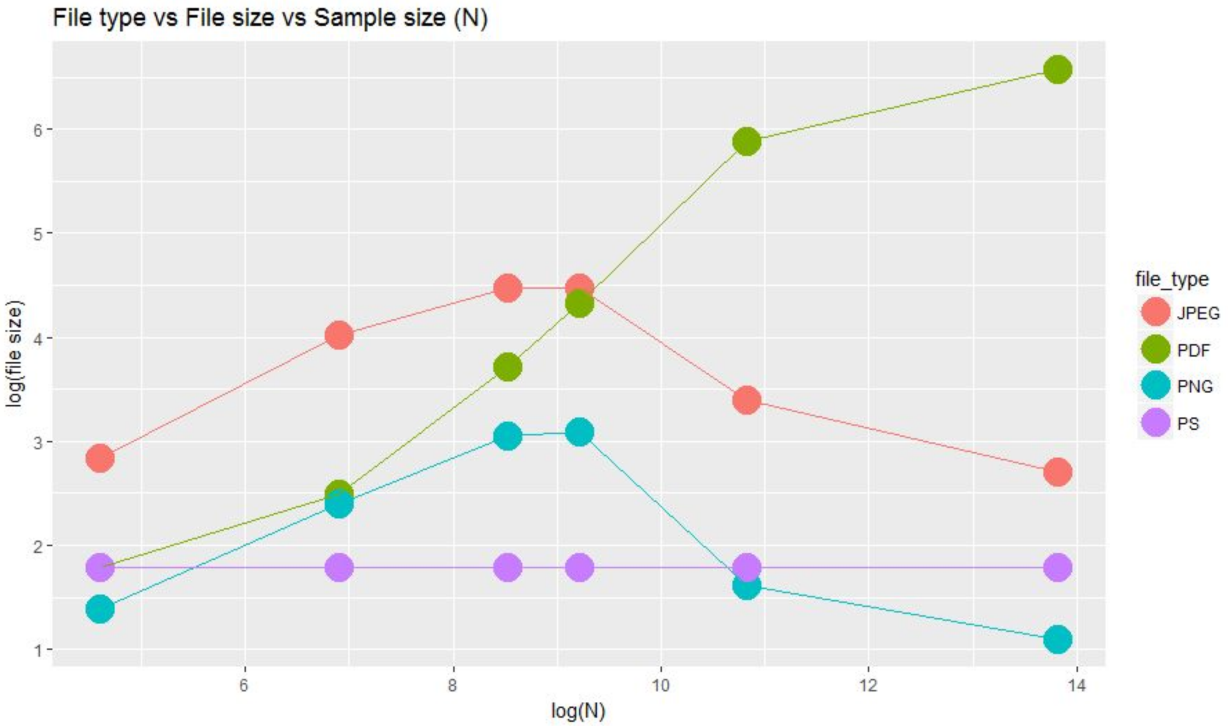


Figure 4.2 below delineates how the file size changes with respect to the sample size for each file format.

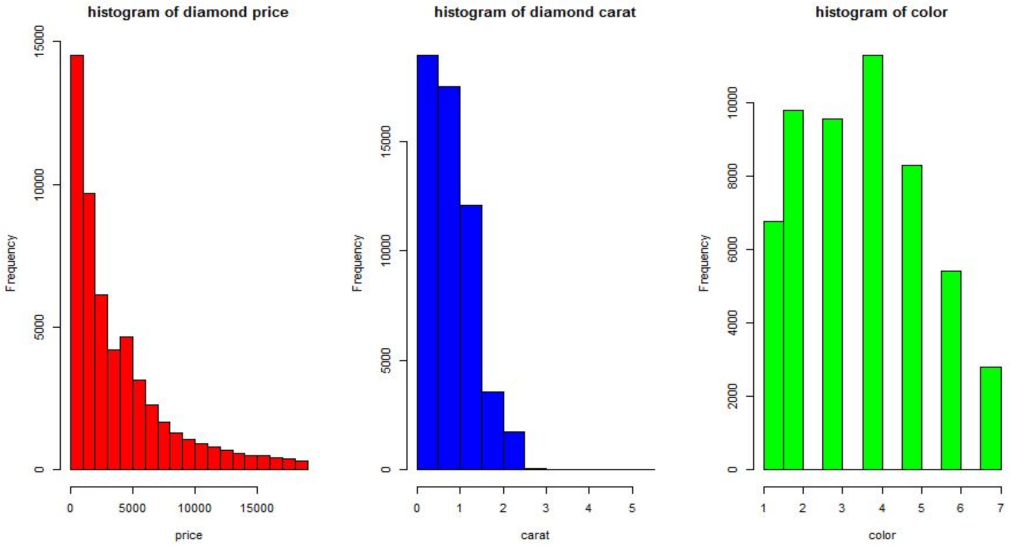
- It appears that the PDF format has the highest file sizes. The file size for PDF increases with respect to the sample size. For PDF files, there is almost a linear relationship ship and positive correlation between the sample size and the file size.
- We cannot say the same thing for JPEG and PNG. It appears that the file size increases up to a certain point and then starts decreasing (after log(N) = 9.1, where N is 10000). Thus, for JPEG and PNG, the file size when N=100 is almost the same with the file size when N=1000000. So, this change is concave.
- For PS format, the file size stands still no matter what the sample size is. I am not sure if this is happening because my computer cannot read the postscript files.

Figure 4.2



5. Diamonds

Figure 5.1



As seen in Figure 5.1, the distributions of price and carat are positively skewed, whereas the distribution of color is almost uniform. It is not a surprise to see such a distribution for because very few people can afford diamonds above 10k. We see a similar distribution for the carat as well. This is, of course, because the carat and price variables are positively correlated. For the color variable, good (D and E) and poor colors (I and J) are not as frequent as the average colors (F, G, and H). This is probably because the average colors (F, G and H) are more affordable than the good colors (D and E).

Figure 5.2

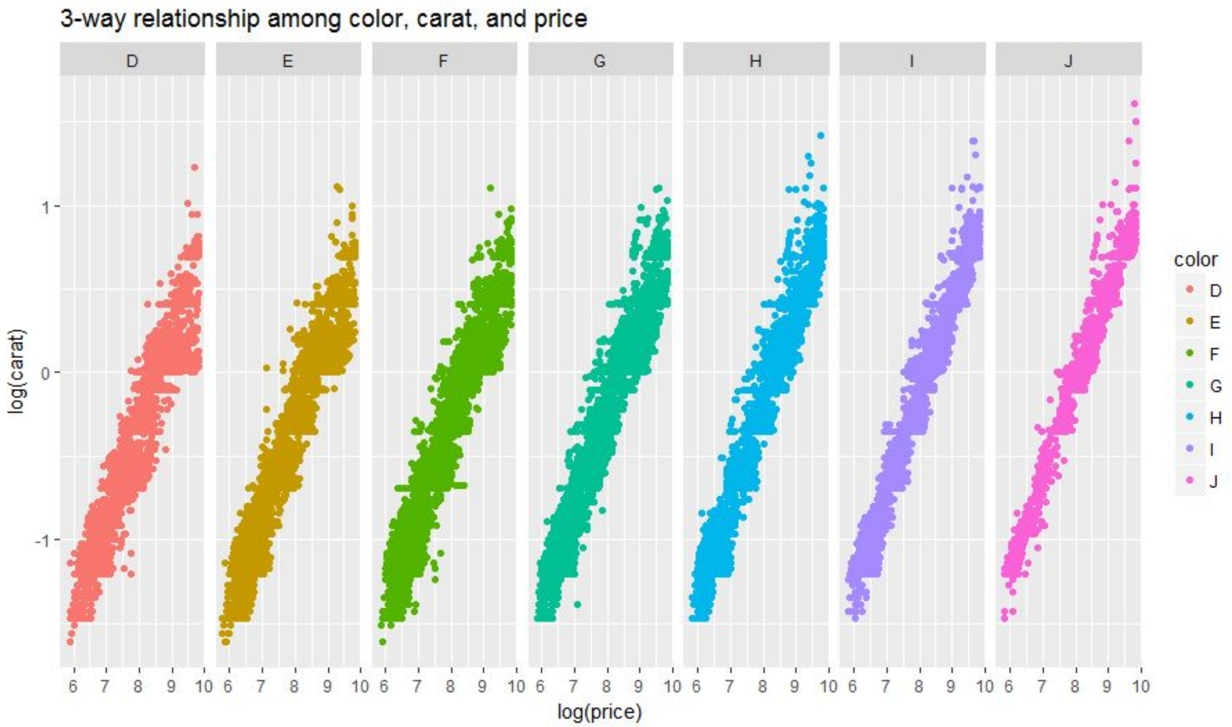
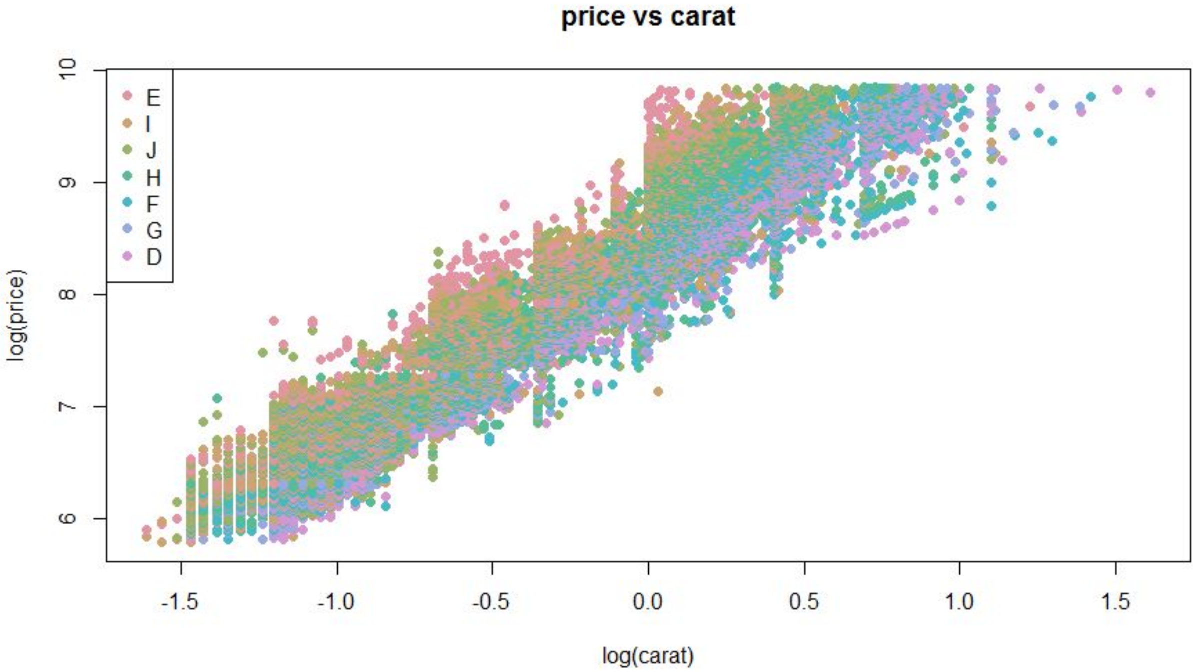


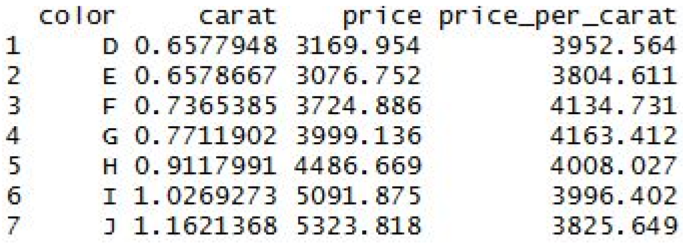
Figure 5.2 above shows how price and carat are correlated given the color. For each color, we observe the same trend. There is a positive correlation between the carat and price variables. However, it seems that this relationship is getting stronger when colors go from D to J.
For instance, you can see that the points in color J are not as sparse as the points in color D. The points on the graph almost form a 45-degree straight line in color D. In color D and E, you see that the same carat might have different prices. This means, for color D and E, besides the carat variable, there might be another factor driving the price up or down. When you check color J, the same carat values almost have the same price (the dots almost build a narrow line).

Figure 5.3



I plotted this graph to see of there are any clusters formed based on the color variable. It seems that there are no apparent clusters, D, E, F, etc. are all overlap one another. This proves that we cannot say color D is significantly more expensive than color F. The last column in Table 5.1 shows the price per carat for each color. As seen in the table, for different colors, there is not a significant difference in the means of 'price_per_carat'.

Table 5.1



```
  color     carat    price price_per_carat
1     D 0.6577948 3169.954        3952.564
2     E 0.6578667 3076.752        3804.611
3     F 0.7365385 3724.886        4134.731
4     G 0.7711902 3999.136        4163.412
5     H 0.9117991 4486.669        4008.027
6     I 1.0269273 5091.875        3996.402
7     J 1.1621368 5323.818        3825.649
```