Homework 2 Report

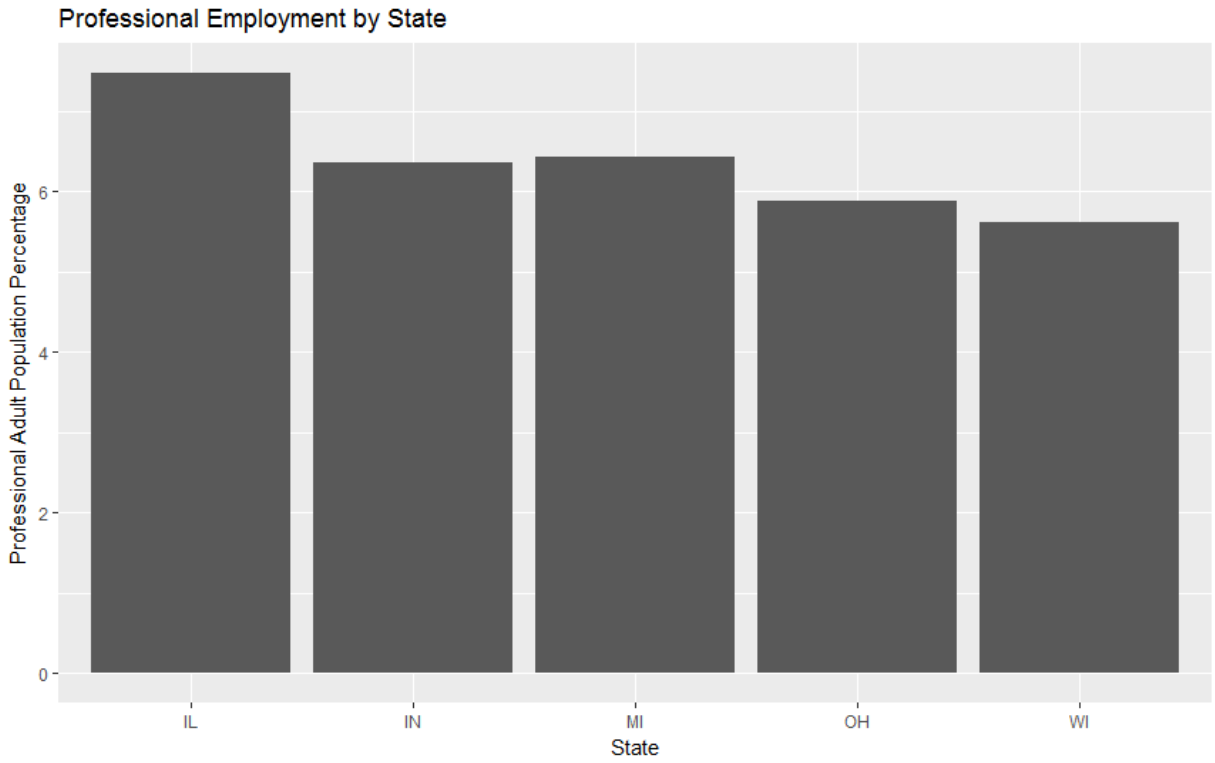
Mohammad Hassan Salim

Msalim7

Part 1: Professional Education by State

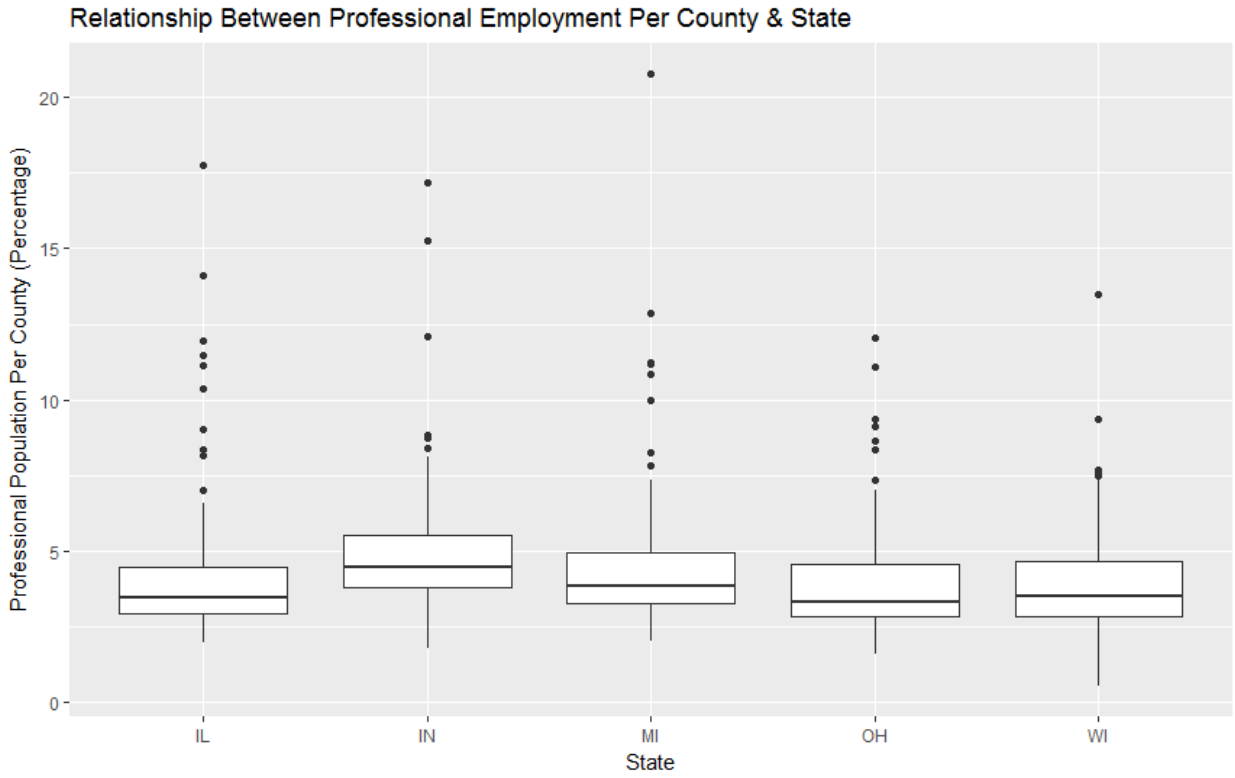
For this question, I went with interpretation A. The state with the highest professional adult population percentage is Illinois. The state with the lowest professional adult population percentage is Wisconsin.

It's a bit difficult to learn anything from interpreting this data. We know that Wisconsin has the least professional adult population. This could be that there aren't many professional opportunities there. Maybe there is a source of poverty preventing people to get an education required at a professional job. It's also possible the culture there does not promote working a professional career. The reciprocal is true about Illinois. Chicago is a heavily populated and professional area that probably gives Illinois the edge it needs to be populated with the most professional adult by percentage.



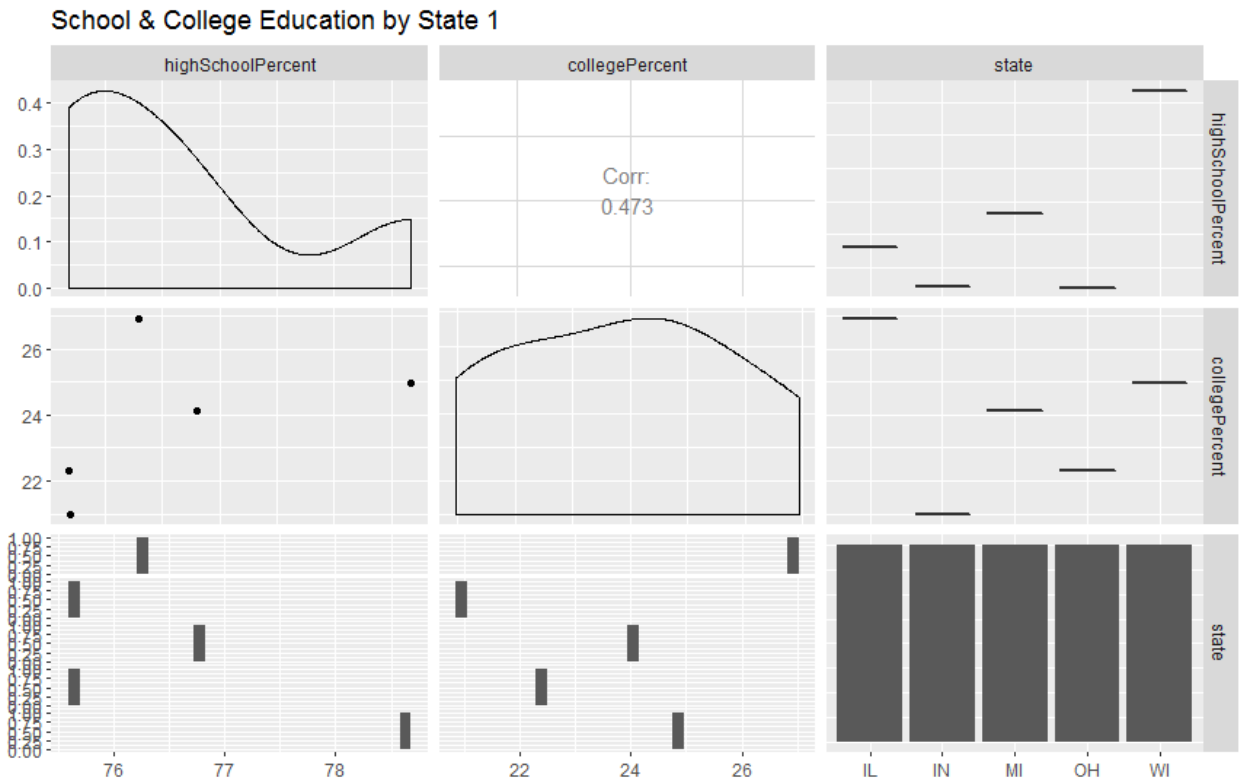
Next, we'll look at the relationship between states in the Midwest region and the percentage of people that have professional education per county. All states have a median below 5%. All states except Indiana have 5% or less third quartile. Although Indiana seems to have a higher first quartile, third quartile, and media than Illinois, it still doesn't have the highest adult population percentage.

Michigan has the highest outlier. Illinois has the most outliers (which probably plays a factor in it having the highest adult population).



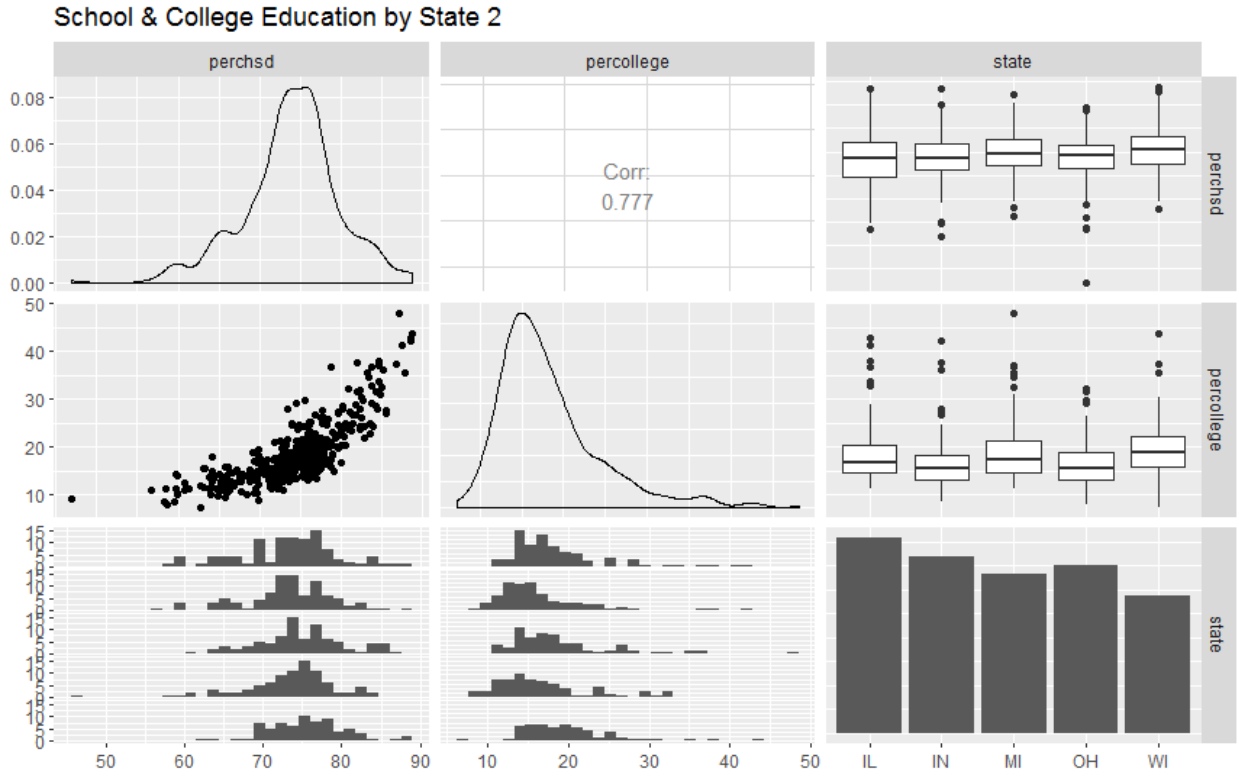
Part 2: School and College Education by State

Initially when writing my code, I went with interpretation A. I aggregated the percentage with high school diploma and the percentage with college degrees. I created a GG pairs plot, but the information revealed is very little. In fact, I have no idea how to read some of these plots. State by state is useless. State by anything else is not useful. Since we aggregated the percentages by state, we only have 5 entries for high school percent and 5 entries for college percent. There won't be a lot of data points for high school percent by college percent. There is .0473 correlation between high school population percent and college population percent. This could be because whoever has a college degree most likely has a high school diploma of some sort.



I was going to end it at that, but I was curious as to what interpretation B would tell me. This is interesting because we have a higher correlation between the raw percent values of high school diplomas and college degrees. Far more people finish high school in these 5 states. In fact, there are more outliers with counties with smaller high school diploma percentages. The opposite is true about college. Far less people finish college and those with high percentages are considered outliers. This is another indication that people just don't have college degrees, and that is needed for a professional job (referring to problem 1). As the percent of high school diploma increases, the percent of college degree increases. This could be because counties that have people completing high school probably endorse education beyond a high school diploma. Looking at the histograms of state by high school percent and state by college percent tell us more of what we can already infer. These states have more counties who have populations that complete high school. These same states have less counties who have populations

that complete college.



This is interesting because we learn more by no aggregating our data. This could be because we produce far less data points once we aggregate it, so not much can be learned with pair-wise plots.

Part 3: Comparison of Visualization Techniques

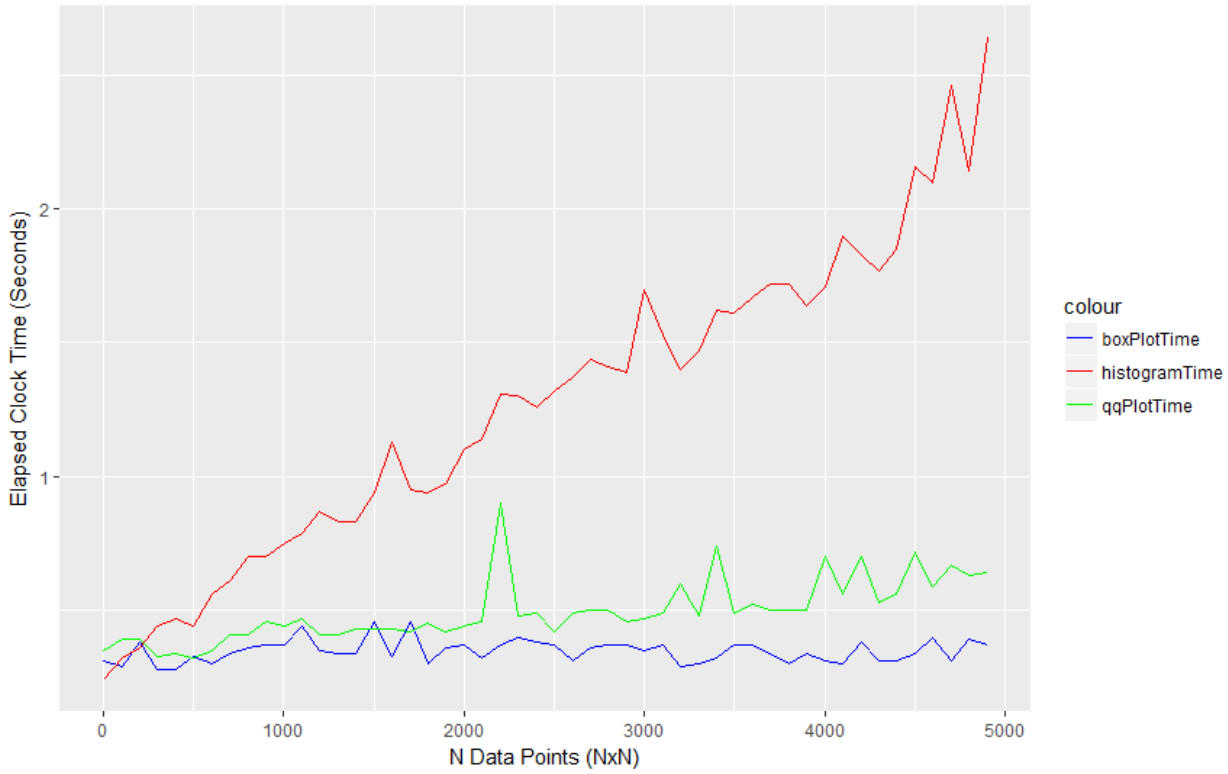
A simple box plot displays a handful of features: min, 25th percentile, median, 75th percentile, max, and outliers (beyond min and beyond max). Depending on how the sample of numbers is distribute, it could easily shift the entire interquartile range. Visually, the whiskers with min and max seem fixed. You don't know what values are aggregated near min and max.

Box plots are useful when you need to know the exact value of a median or a certain percentile. It tells you the calculated values as needed. However, you cannot how many of what values are distributed where. If you want a more detailed visual of the distribution, a histogram is a better choice in my opinion. The issue with histogram is that it doesn't calculate the exact values you want (such as median) which is exactly what box plots do.

I'd use a box plot if I want to know what the IQR, median, or outliers are. If I want to know a specific value. I'd use a histogram if I wanted to view the distribution of the data. I'd use a QQ plot if I wanted to compare the relationship and shape between 2 distributions.

I want to mention a final pro/con comparison. Say you just want to learn something from a very large data set. It would be beneficial to use box plot or QQ plot over histogram because histogram is far slower. I checked wall times of rendering each plot with NxN data sets where each N goes up 5000

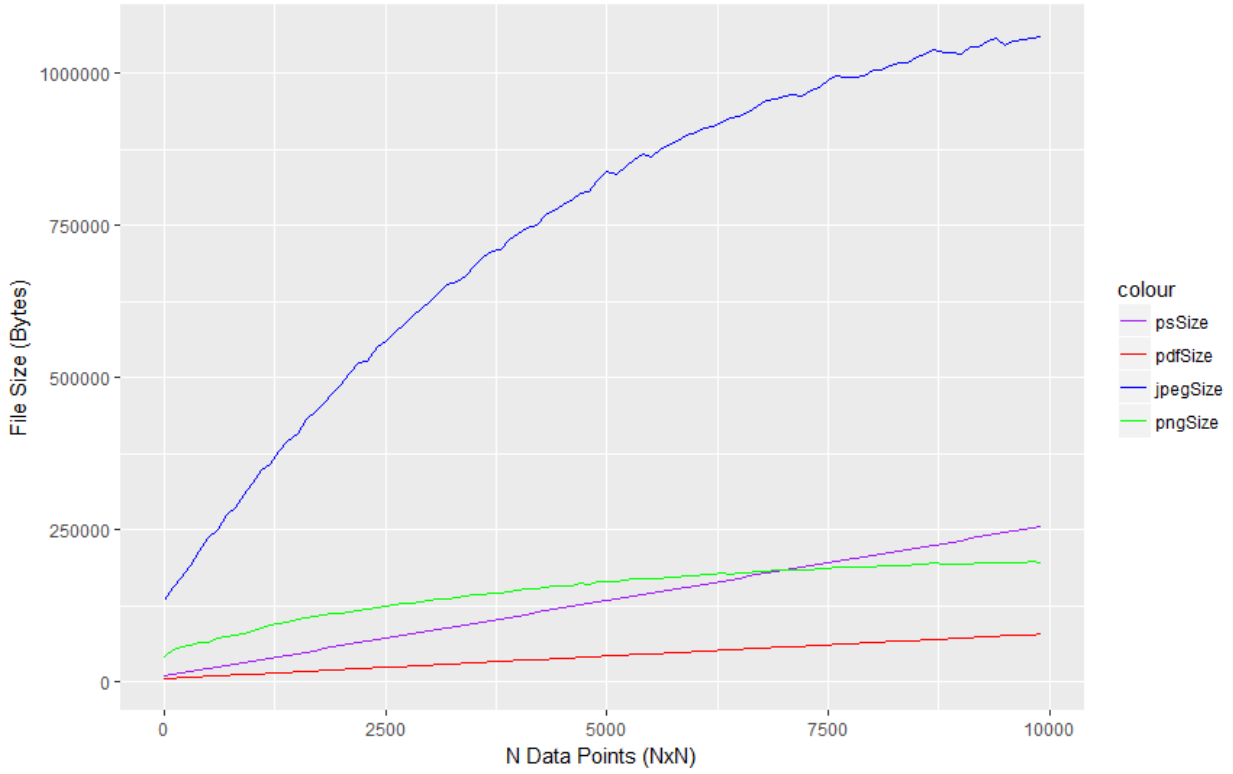
randomly (uniform distribution) generated data points. Histogram is probably slowest because it must group each entry then render it proportionally.



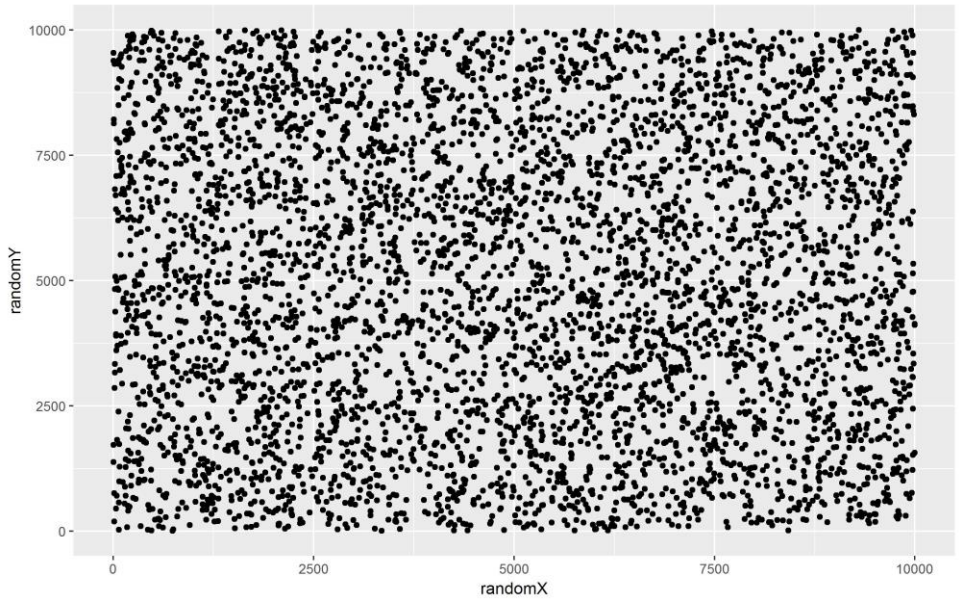
Part 4: Random Scatterplots

Like problem 3, I saved scatter plots produced from NxN data sets where each N is 5000 randomly (uniform distribution) generated data points. PDF saves the most space and scales very well. PS files seem to be the second most space efficient about when N is 7000. PS does not scale well. PNG takes the second spot for large sizes of N. The worse in space efficiency is JPEG. For all ranges of N, it just easily takes about a lot of memory.

Both PDF and PS display a linear relationship. PNG has a positive non-linear relationship, but seems like it starts to flat line near the end. JPEG has a terribly rapid positive non-linear relationship.



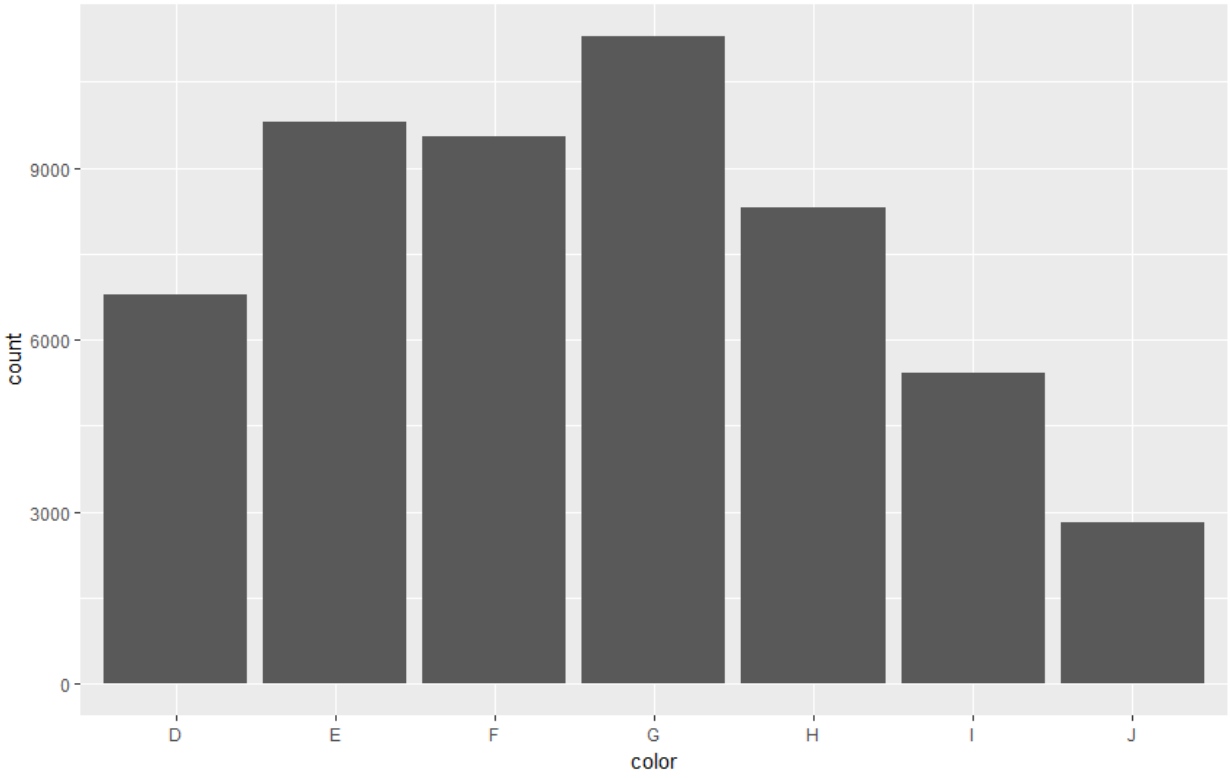
The question is now: why do all these file types give these results? PDF probably compresses the image as much as possible and loses all pixel information. PNG and JPEG both probably contain pixel info, but PNG probably compresses it better. PS probably compresses the image like PNG, but the compression algorithm for PS probably doesn't scale well. Below is an example of how one of these scatter plots look like. It would be interesting to try different types of plots and adding color to images. It would also be interesting to check the speed of generating and saving these files.



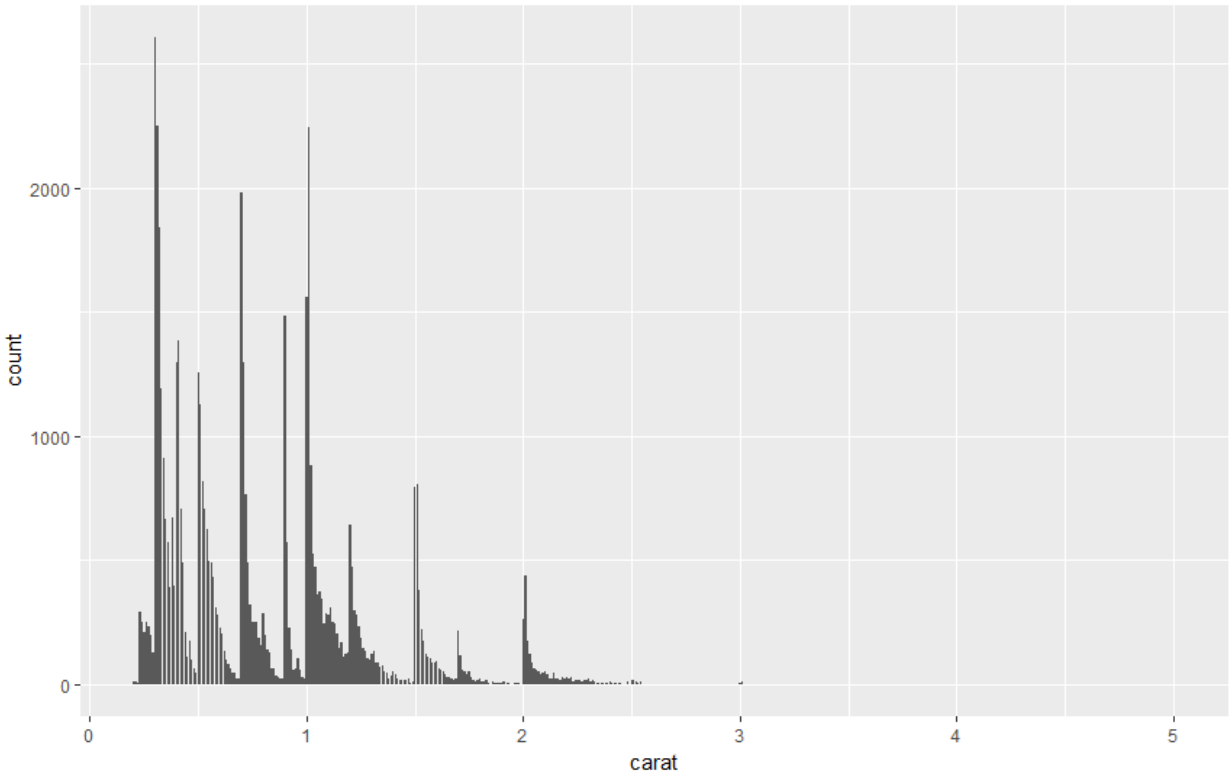
Part 5: Diamonds

The diamond dataset has many features. We'll be looking at color, carat, and price. If we look at the distribution shape of each, we'll understand the kind of data we're dealing with.

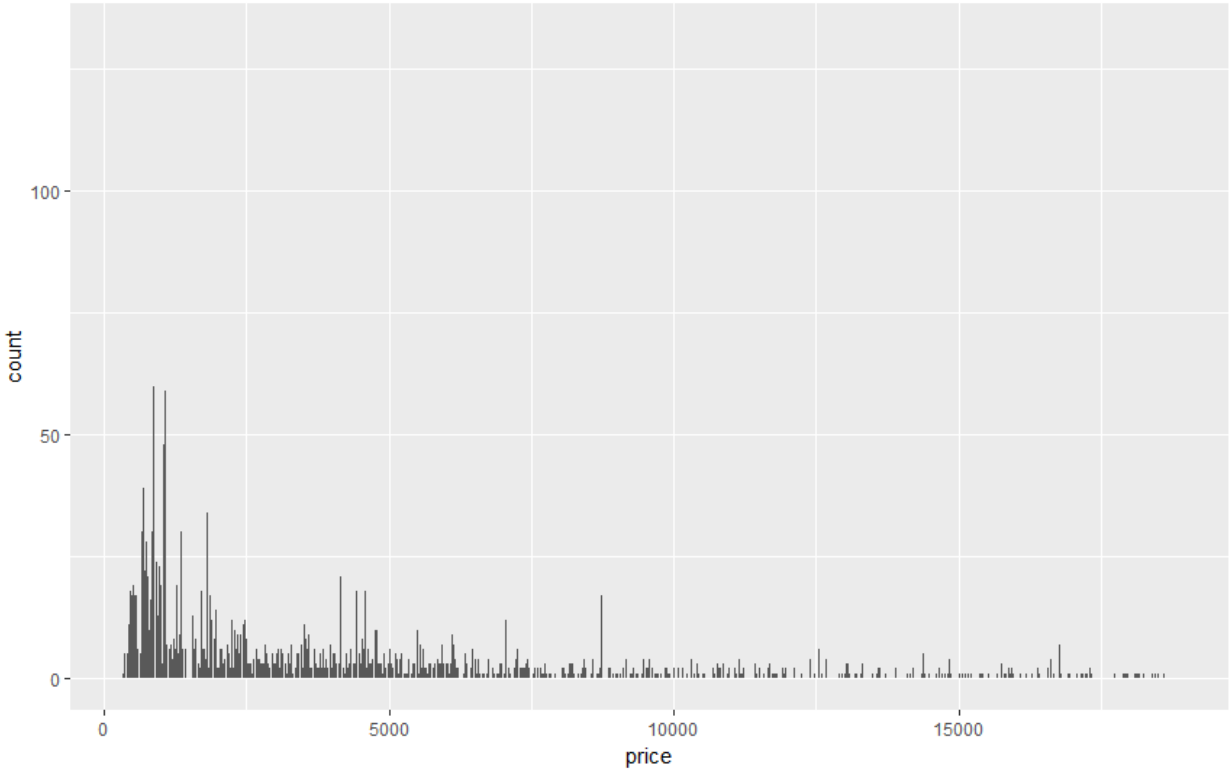
The color histogram seems slightly right skewed, but closer to being symmetric than the other histograms we'll look at. This probably means it's easier to get diamonds with average color quality. Of course, it's still difficult to get high quality colored diamonds.



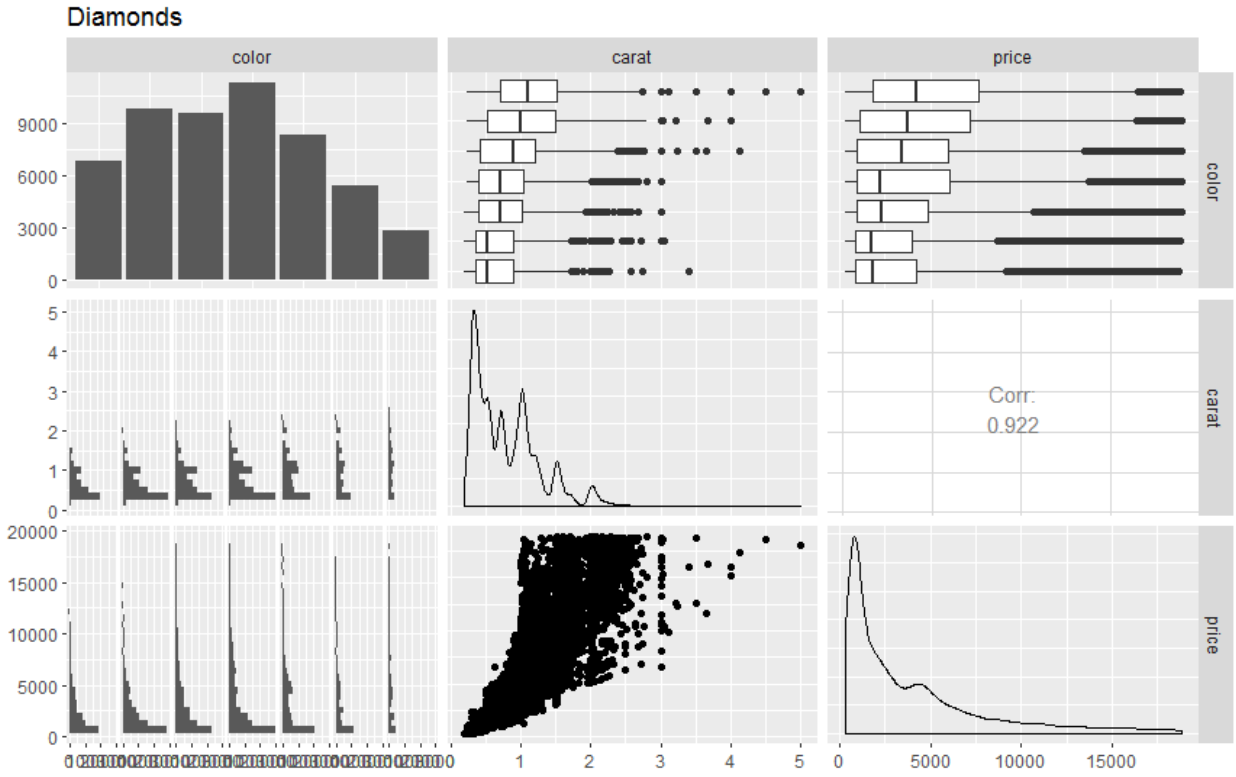
The carat histogram has an interesting distribution. The overall distribution is right skewed. However, if you zoom in, you'll seen there are multiple right skewed distributions within the overarching distribution. This could be because the companies that produce diamonds using their own distribution. This would explain all the mini distributions through the overall distribution. They need to keep diamonds with higher carat less common. If high carat diamonds are more valuable, people probably won't by as much hence the overall right skewed distribution.



The price histogram is the least interesting. It's strictly right skewed. This is expected. Expensive diamonds probably don't sell as much and therefore aren't in demand as much as cheaper diamonds.



Now let's look at a 3x3 pair-wise plot. This results shouldn't be too surprising with the analysis we've discussed so far. Carat and price have a strong correlation of 0.922. The higher the carat, the more the diamond costs. You can see this graphically with the bottom-middle graph. If we look at the price and color distributions on the bottom-left, each color-to-price distribution has a similar shape of skewed top. For each color, they have many diamonds that cost less and a handful that cost a lot (except for color G where the distribution is closer to being flat). You'll see the same with middle-left graph displaying color-to-carat. Finally, the last 2 graphs worth discussing are the box plots. The price-to-color box plot reveals there are a lot of outliers. This must mean each color group just has a handful of diamonds that are just too expensive compared to the rest of the diamonds in that color group. The carat-to-color box plot reveals something similar. This could just be how the diamond market structures its diamonds its selling.



I'd be curious to know where these diamonds come from and who buys them, then try to find a correlation there. Most likely we'll see couples buying each other jewelry, but I'd like to see which areas have been tapped and which haven't been. This way a company could decide to open up shop in a certain area with many young couples who are willing to buy diamonds.