Jeremy Kang
GTID: 903136703
GT Name: jkang315

1. Interpretation A:

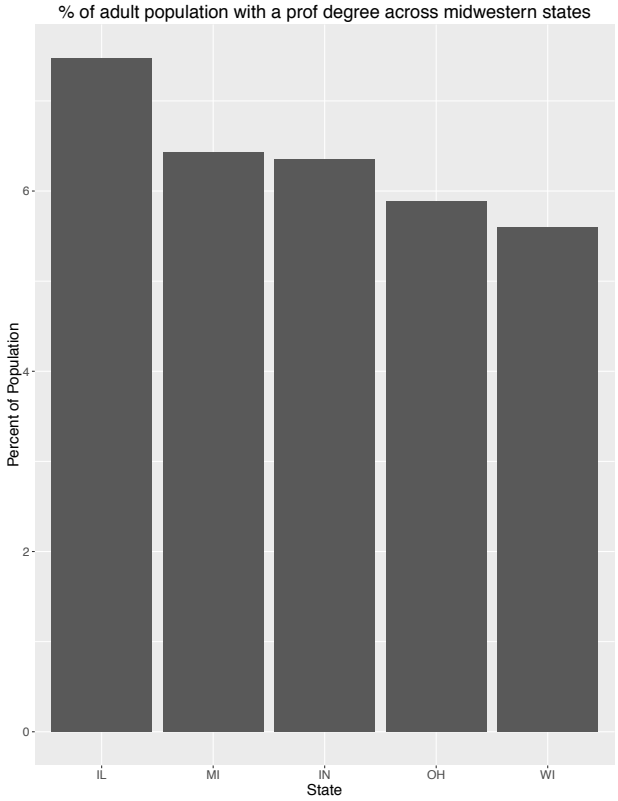

Figure 1. Percent of adult population with a professional degree ranked from highest to lowest across five Midwestern states.

A subset for each state from the Midwest data was created to make the analysis easier. Using the interpretation A and the formula provided (where the sum of the percent of professionals is multiplied by the population of adults and the divided by the sum of the population adults to a new column representing the percentage of adult population with a professional education. Figure 1 shows that **Illinois has the highest percentage of adult population with a professional education at 7.47%. Wisconsin has the lowest at 5.60% of the adult population with a professional education.**

Note that since interpretation A is averaged across the state, the interpretation doesn't give much insight into the counties within the state. Thus, I included interpretation B for further analysis.

Interpretation B

For interpretation B, I still utilized the subsets created in interpretation A. However, by just interpreting the raw percent of professionals, Indiana appears to have the

highest median at 4.44% **and** average at 5.05% compared the five states. The lowest median percent of adults with a professional education appears to be in Ohio at 3.33%, but if we look at the mean, Wisconsin still comes to be the lowest at 4.05%.

Note, these conclusions differ from interpretation A because the adult population for interpretation A was averaged across the whole state, whereas the raw percentages of high school and college educated was based on the average at the county adult population level.

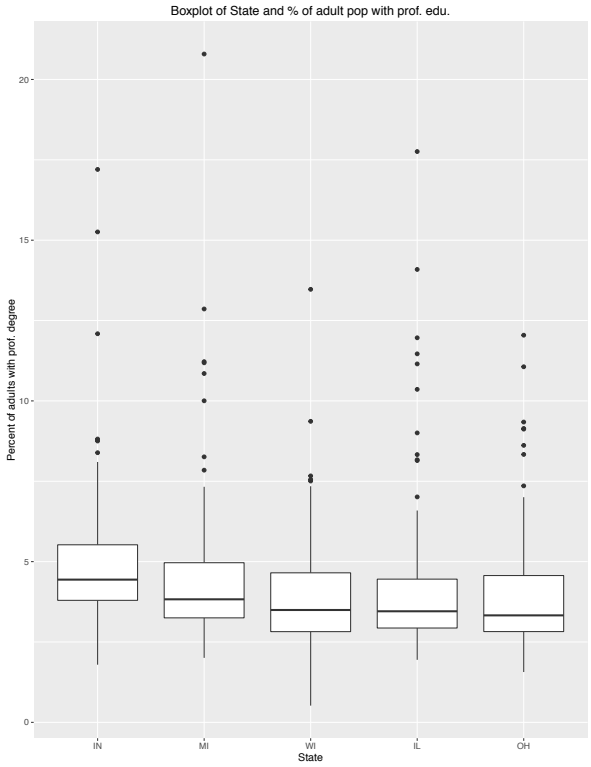

Figure 2. The boxplots are graphed from highest (left) to lowest (right) using the median. Each point represents a single county within that state.

In figure 2, it's difficult to tell in the shape of the distribution using a boxplot across the states. Let's take a look at the subset of states to get a better picture at the county level. Figure 3-7 shows percent of adult population with a professional degree within each county in each state ranked from highest to lowest percentage.

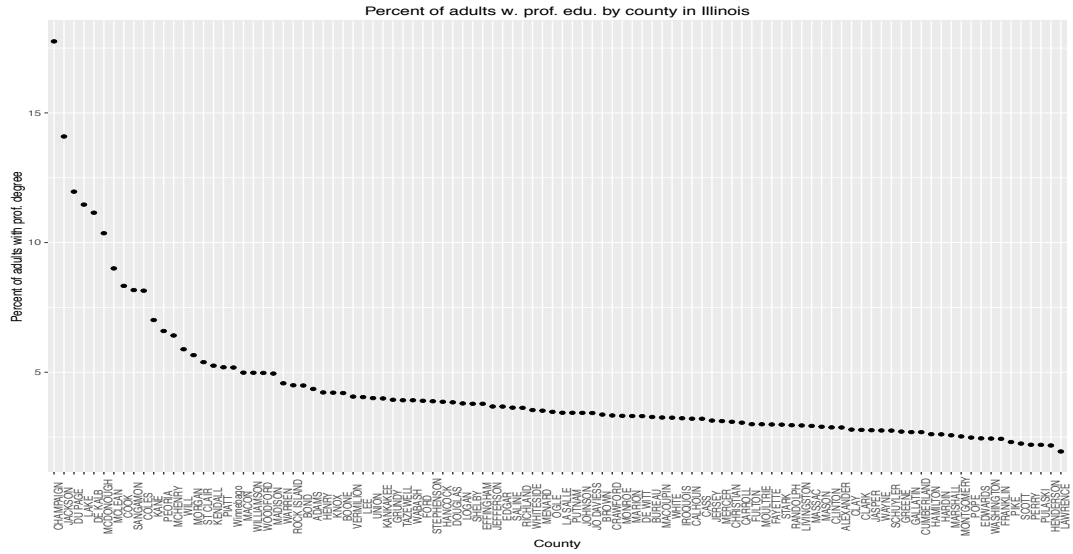

Figure 3. Percent of adult population with professional education by county in IL.

The distribution of adult population with a professional education seems to be disproportionally skewed to certain counties. The county with the highest percent of adult population with professional education is Champaign, IL at 17.76% and the lowest in Lawrence, IL at 1.95% - which is a 15.81% difference. 27 counties (26.47% of the state) have an adult population with a professional education above the average.

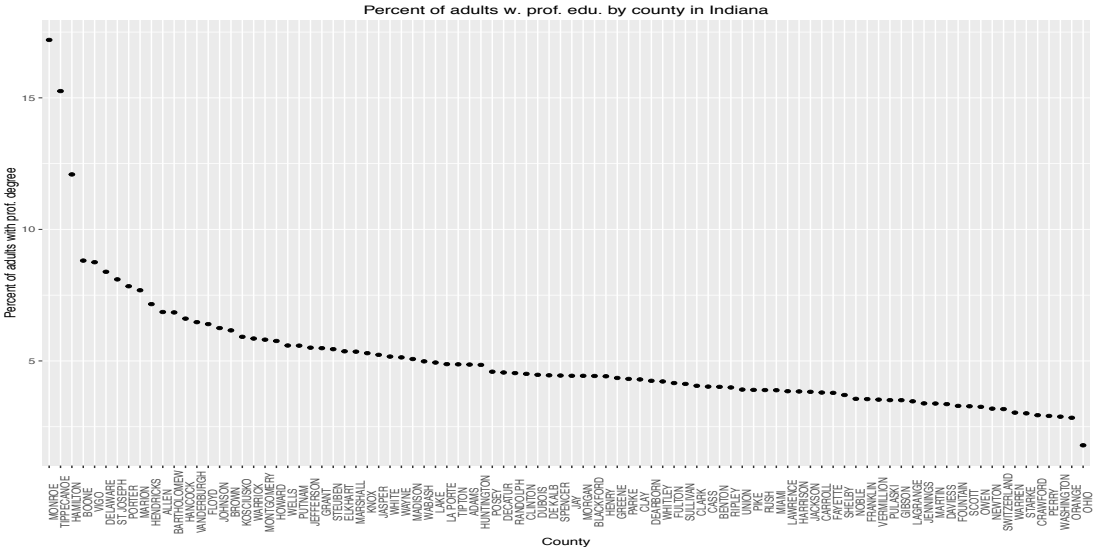

Figure 4. Percent of adult population with professional education by county in IN.

The distribution of the adult population with a professional education is more spread out in Indiana with about 35.87% of the counties in the state that are higher than the average. Still, the disparity between the highest county – Monroe, IN – and the lowest county – Ohio, IN – is 15.41%.

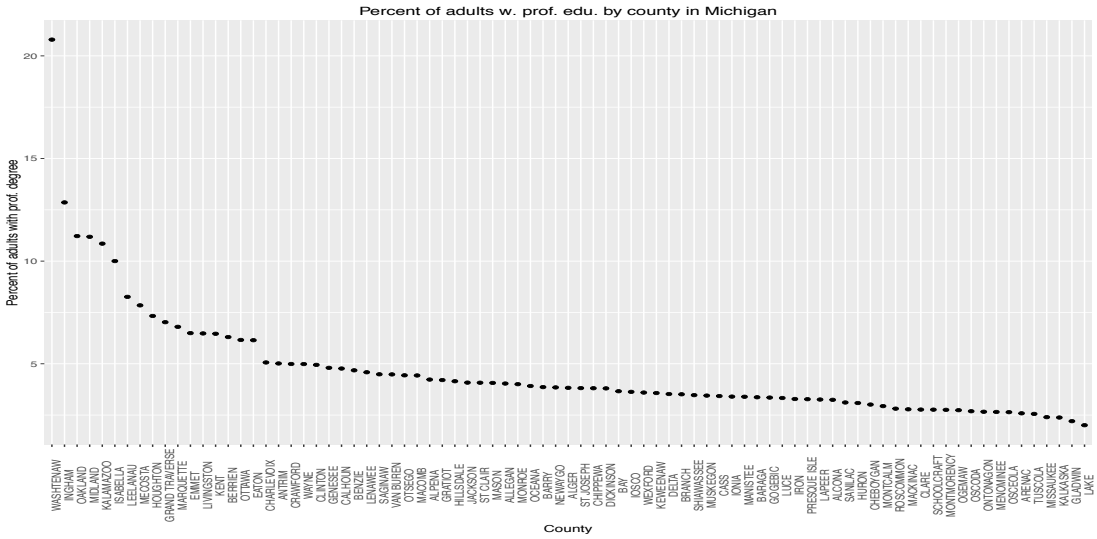

Figure 5. Percent of adult population with professional education by county in MI.

Washtenaw, MI has, not only the highest percent of adult population with a professional education within the state of Michigan, but also in *any* state at 20.79%. This disparity creates a difference between Washtenaw, MI and Lake, MI at 18.78%, - the largest range in distribution between the five states. This explains why there are 28.92% of counties that are above the mean of percent of adult population with a professional education.

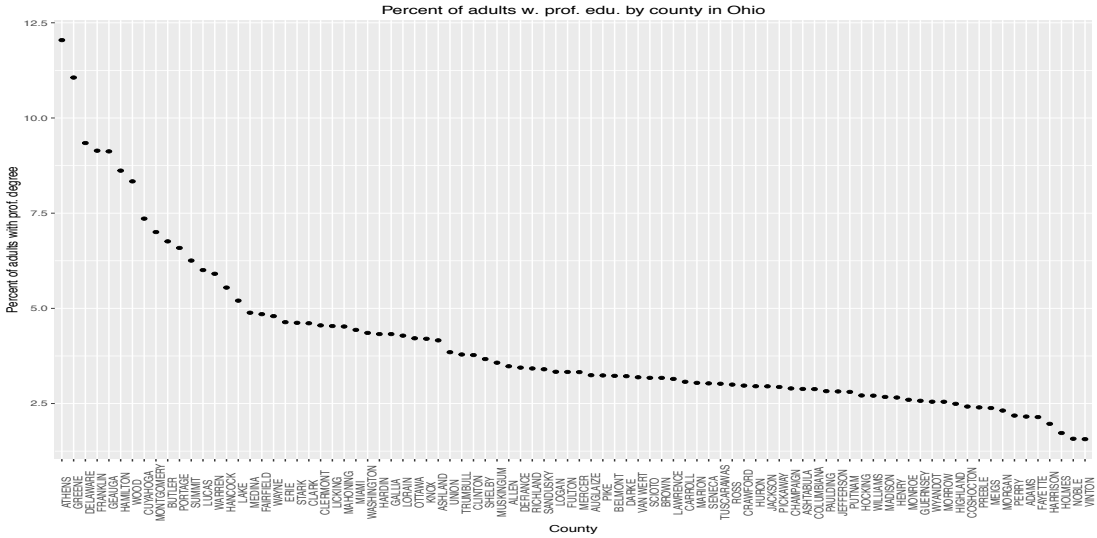

Figure 6. Percent of adult population with professional education by county in OH.

Ohio tied with Wisconsin to have the largest distribution of percent of adult population with a professional education at 37.5% of the counties above the average. However, Athens, OH is the lowest highest percent of adult population with a professional education compared to the other four states at 12.04%.

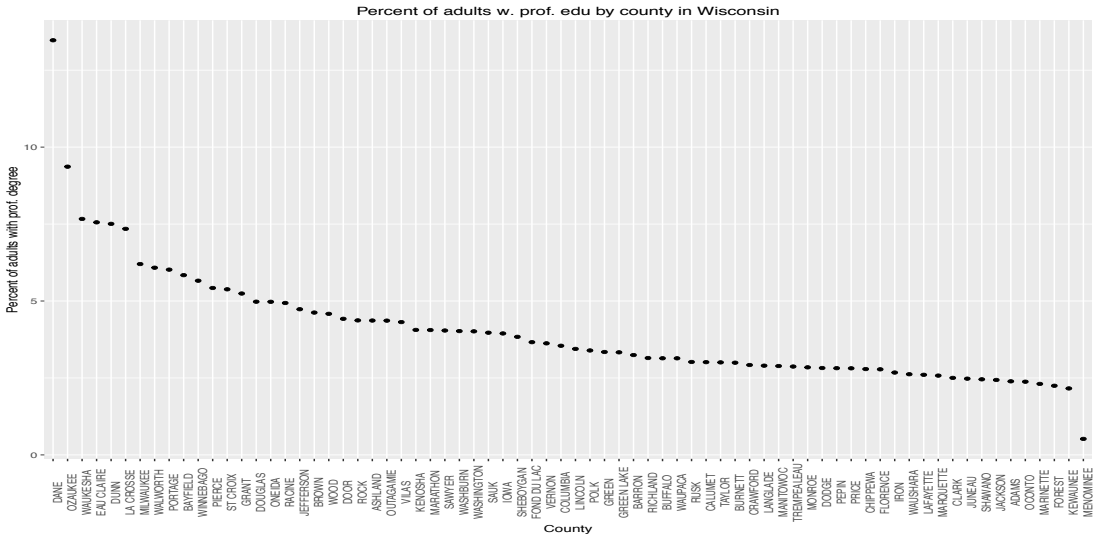

Figure 7. Percent of adult population with professional education by county in WI.

Wisconsin also had a better distribution of percent of adult population with a professional education compared to Illinois, Indiana, and Michigan at 37.5% of the counties being above average. **However, Menominee, WI had the lowest percent of adult population with a professional education of any counties in any state at 0.52%.**

2. Interpretation A

Using the same formula and subsets by state, I aggregated the percent of adult population with a high school education and college education as two new columns into each state data frame.

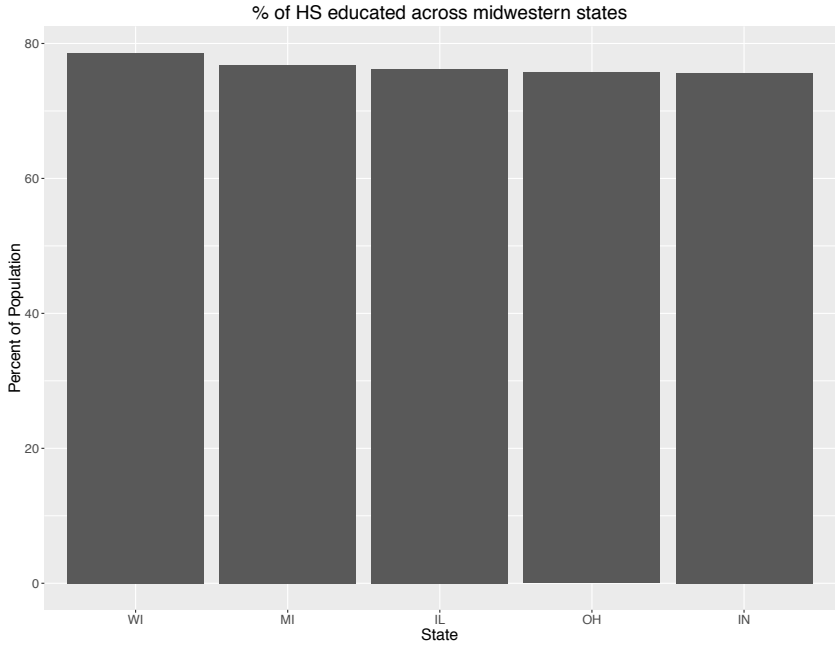

Figure 8. Bar chart of percent of high school educated by state.

78.60% of the adult population has a high school education, making it the highest amongst the five states. Indiana's adult population has the lowest high school education at 75.64%, which is close to Ohio's 75.67%.

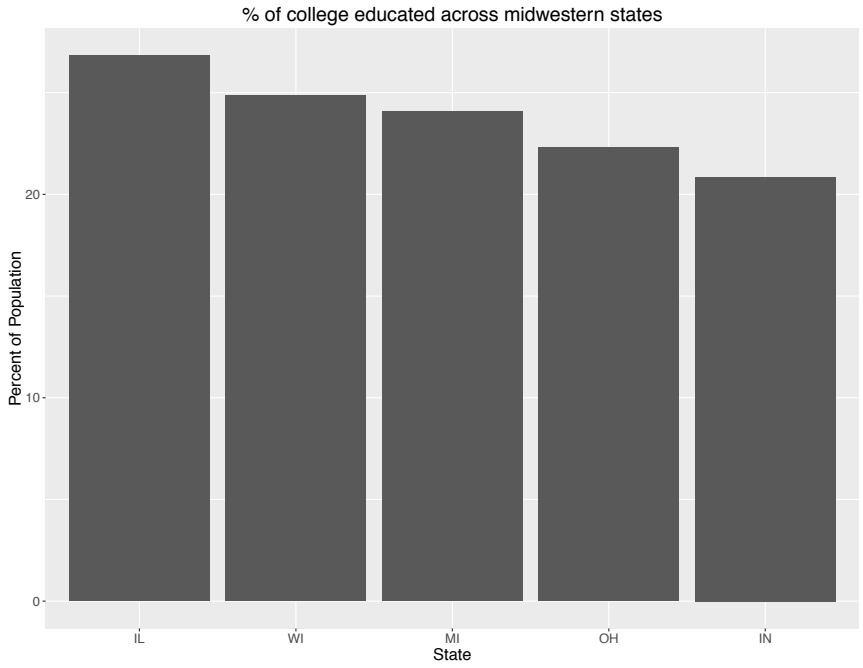

Figure 9. Bar chart of percent of college educated by state.

Illinois has the highest percent of the adult population with a college education at 26.82%. On the other hand, only 20.85% of the adult population in Indiana has a college education.

Since you cannot visualize a three-way relationship based off of one variable being an aggregate percentage (i.e., percent of adult population with a high school education and percent of adult population with a college education), interpretation B was added.

Interpretation B

Boxplots were first created to explore the relationship between percent of adult population with a high school education vs. the states (Figure 10) and percent of adult population with a college education vs. the states (Figure 11).

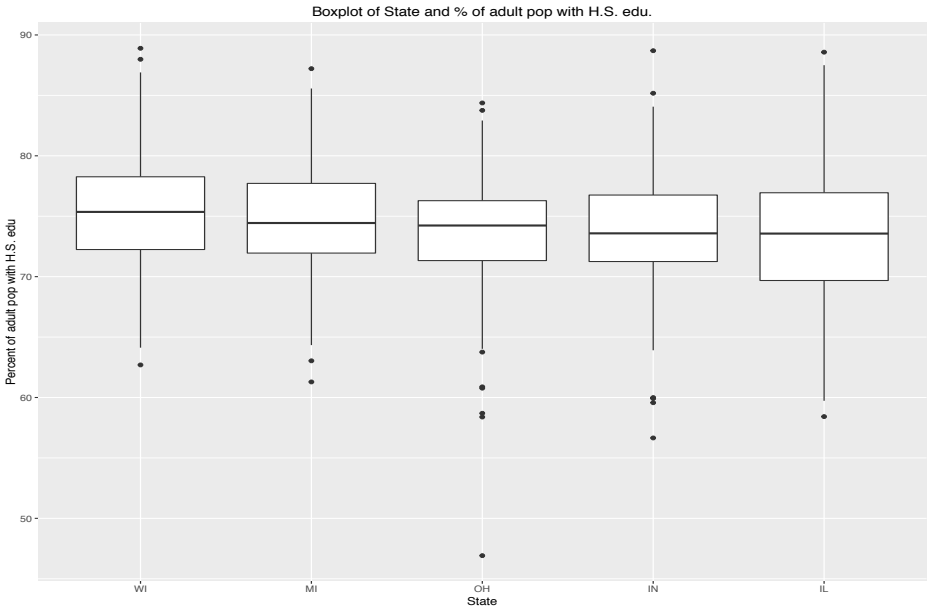

Figure 10. Box plot of percent of adult population with a high school education by state.

As seen in figure 10 in Ohio, one county in that state has the lowest minimum percent of adults with a high school education at 46.91% while a county in Wisconsin actually had the highest minimum percent at 62.70%. The medians and averages of all the counties in each state were quite similar, but **a county in Wisconsin had the highest median percent of adults with a high school education at 75.36% versus the lowest median being a county in Illinois at 73.56%. The highest average percent of adults with a high school education was, again, a county in Wisconsin at 75.52% while a county in Ohio had the lowest average percent at 73.21%.** As shown as the top points in figure 10, the highest maximum percent of adults with a high school education was also very similar across the counties in the five states with a county in Wisconsin at 88.90% and a county in Ohio at 84.37%.

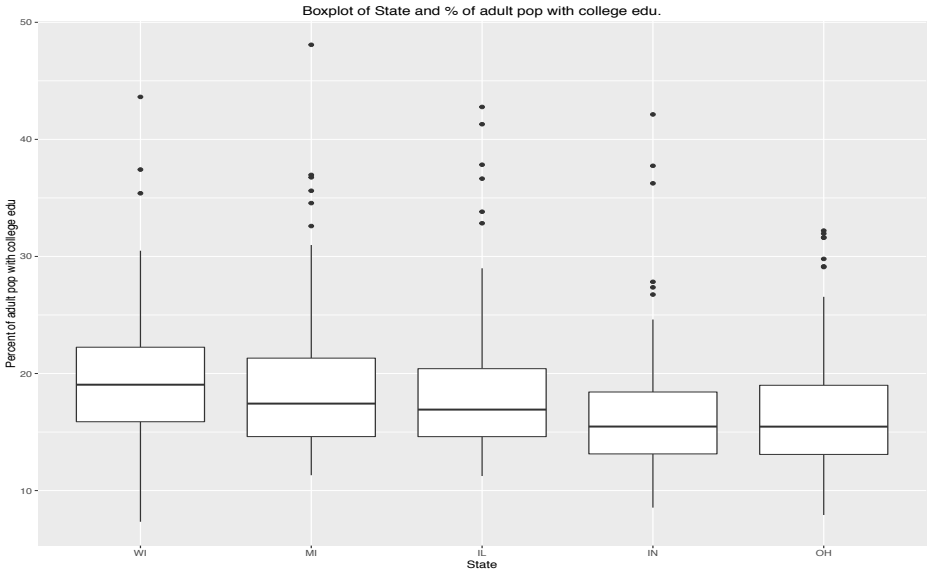

Figure 11. Box plot of percent of adult population with a college education by state.

The percent of the adult population with a college education shows a more interesting relationship across the counties in each of the states. While counties in Wisconsin had the highest percent of the adult population with a high school education, a county in Wisconsin also had lowest minimum of any county in all the states at 7.34%. A county in Michigan had the highest minimum at 11.31%. **Counties in Wisconsin also had the highest median *and* mean percent of the adult population with a college education at 19.04% and 20.02%, respectively.** This suggests the distribution is robust with high and low percentages in the Wisconsin area and can be seen in figure 11. **A county in Ohio had the lowest median percent at 15.46%. A county in Indiana had the lowest average percent at 16.62%.**

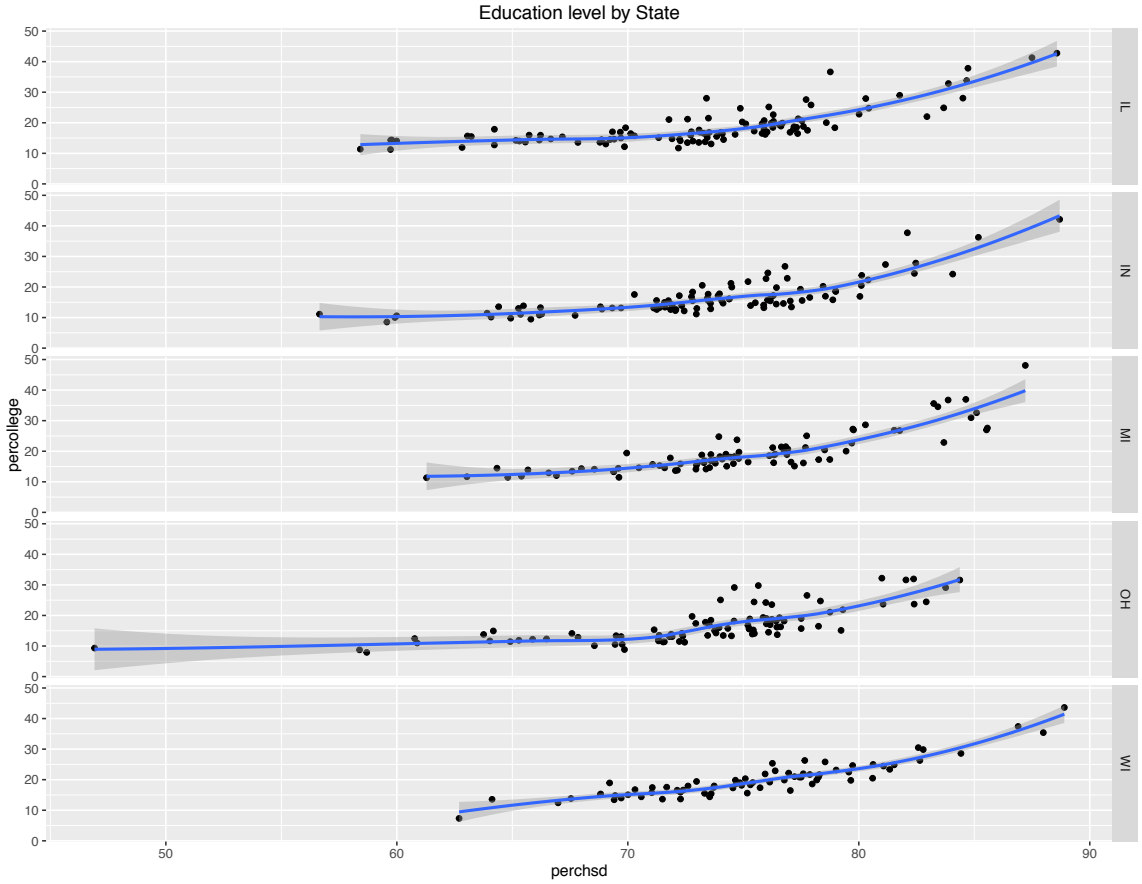

Figure 12. Three way relationship between perchsd, precollege, and states.

Although figure 12 displays a nice comparison between all of the states and the relationship between percent of the adult population with a high school education and percent of adult population with a college education, it's difficult to interpret without solid numbers. The intercepts and the slopes for each state were calculated to get a better understanding of what this three-way relationship actually reveals. The table is below in figure 13.

	q1labels	q2intercept	q2slope
1	IL	-36.44086	0.7532536
2	IN	-39.83337	0.7694763
3	MI	-57.64150	1.0291892
4	OH	-32.70033	0.6774203
5	WI	-57.69921	1.0290854

Figure 13. Fitted line for all states.

Supporting our findings from interpretation A and B, we see that Ohio has the lowest slope at 0.6774, which suggest that a 1% increase in the percent of high school educated in Ohio increases by 1, the percent of college educated in the state only increases by 0.6774%. Compared to the other four states, the change in the population with a college education in Ohio is not *as* dependent on the change of the population with a high school education. This could mean that other variables such as vocational choices made by the adult population that do not require a college degree is chosen more in the state of Ohio could be influencing the population of college educated more so than the percent of the population with a high school education (a hypothetical thought, not fact).

The largest slope for percent of high school educated relating to the percent of college educated is Michigan at 1.0292. This suggest as the percent of high school educated increases in the state of Michigan, the percent of college educated increases by 1.03%. This could be interpreted as a promising achievement for the state as a 1:1 relationship suggests a promising high school education system (depending on other analyses of high school and college population in state) – again, this is purely speculative, rather than what is drawn from the data/illustrations).

3. Histograms

Histograms offer a clear visualization of one-dimensional data that divide the ranges into bins and counting the frequencies in each bin as a measure. This is an easy way to see the density (frequency/**mode**) and variance of the distribution in your data, as well as the shape of your data (e.g., if it is normally distributed, negatively/positively skewed, etc.). However, you must set an appropriate bin sizes (the ranges of values in which the bars divided) so that you can see important areas in the data, but also not get distracted by noise/outliers in your data. As the sample fluctuates in size, you want to adjust your bins accordingly so that each of the bars gives insight into each range as opposed to having too wide of a bin which could drastically change your histogram shape, leading to false assumptions about your data. **It's also difficult to infer other statistical properties such as the minimum and maximum, median, and quartile values from the data** as histograms usually focus on frequencies more so than the aforementioned statistics.

The greatest strength of histograms is measuring frequencies in a sample of numbers. **As mentioned, histograms show a clear occurrence of a data point in your data so it would be the most helpful in problems with making sure your sample size is normally distributed (e.g., you want to make sure the sample you're collecting is representative of the population you want to observe).**

For example, say you had the ability to collect a sample of the student population and wanted to know students' choice of college majors and you also gather demographic information such as age, gender, and if they are out-of-state students. The school reports that the average undergraduate student age is 23 so you expect the mean (the middle of your histogram) to be about 23. A histogram plot of this data would not only be useful in seeing if this information is accurate, but also seeing what other ages are common at the college. Below is a normally distributed histogram of a fictional college matching our example.

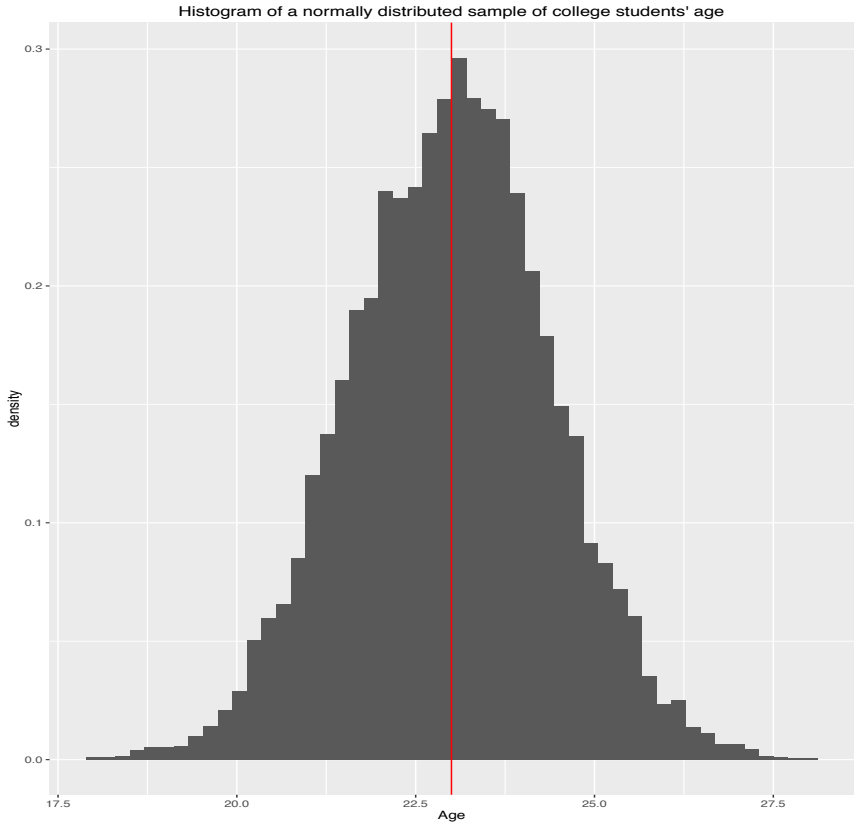

Figure 14. Histogram example of college student age (artificial data).

Box Plot

Box plots are also a type of visualization of the distribution in your data, but in regards to quartiles rather than just a frequency. The box refers to the square that is drawn using the **first quartile**, **median**, and **third quartile**. The **minimum** and **maximum** are clear to see in box plots. The two lines (whiskers) represent points outside of the first and third quartiles, before being flagged as potential outliers. Lastly, the points outside of the whiskers, called the fence, are potential outliers –

which is 1.5 x below or above the IQR. The advantage of using box plots are that you are able to visualize the range at which your data is distributed more clearly and don't have to be concerned about bins or missing potential important outliers.

You're also able to see where the majority of your data is distributed (including clear outlier points and skewness) and where the median and inter-quartile range (IQR) are in your data. A major disadvantage to using box plots is you have less information to see individual points in the data such as **not being able to see frequencies** of your sample, **not being able to see what your average** is (unless you specify the line to be mean in which case it could cause confusion because the mean could lie outside of the first quartile, median, and third quartile), and **also the requirement that at least one of the dimensions in your data be continuous to utilize box plots**.

Regardless, box plots excel at displaying the range of your data and should be used in cases where you want to confirmed that the frequencies in your data seem normally distributed, but want to also make sure that your data has robust enough variability to make your data representative of the population you're sampling from.

Going back to the example of the college students, you may want to see if female and male respondents show the same age range and similar variability distribution. Note that the box plot example is using a different set of data.

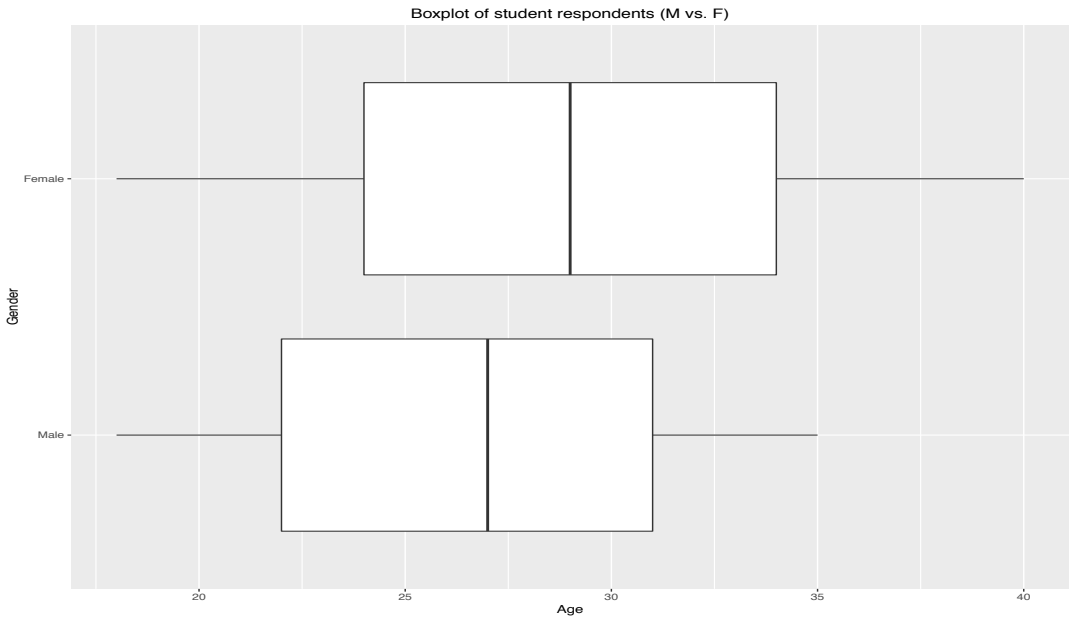

Figure 15. Boxplot example of Male vs. Female respondents age (different simulated data).

QQ Plots

QQ plots are highly sensitive to minimum and maximum values that skew the data towards either direction (i.e., outliers). Similar to histograms, QQ plots generate a graphical representation of the distribution of your data. What makes QQ plots unique is that they focus on quantiles (thus the quantile-quantile in *QQ* plots) of the values against another quantile distribution (typically a normal distribution, but can also be the distribution of another data set) of data. **This is very powerful in that you are able to test your sample distribution for normality against a theoretical normal distribution in cases where you're not sure if your sample is representative of the population or not.** It is very easy to spot if something is wrong with the distribution of your data using QQ plots because of its sensitivity to outliers.

The qq plots below use the same data used for the box plot example.

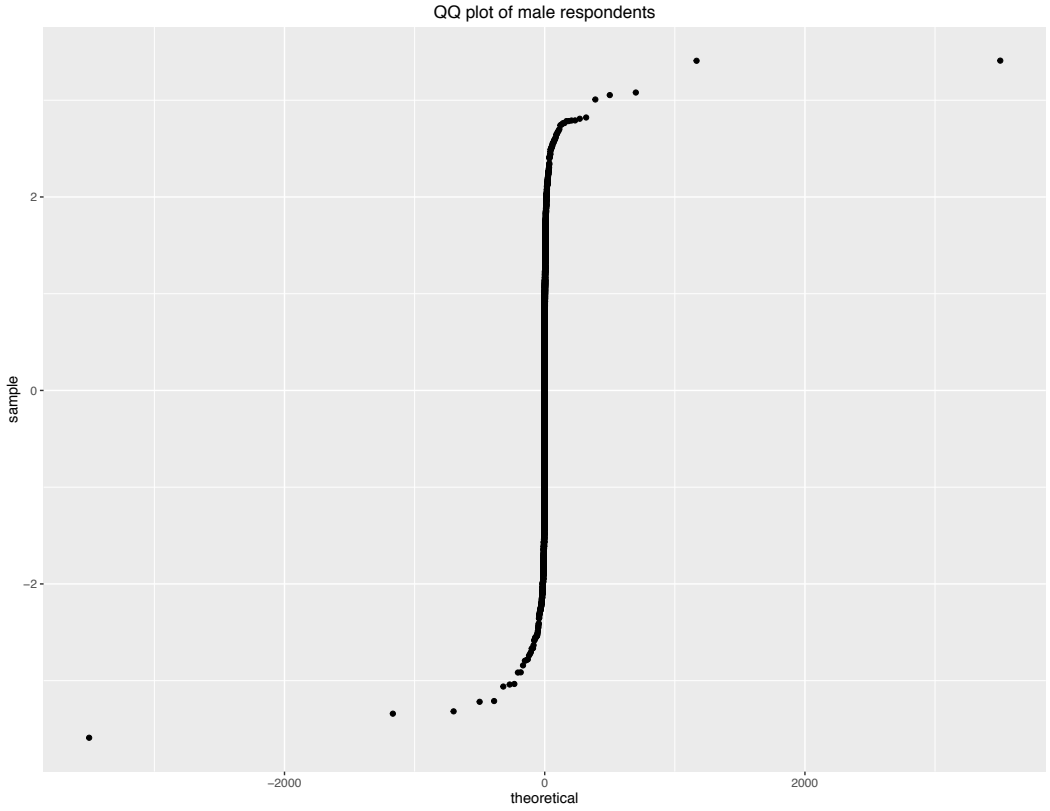

Figure 16. QQ plot of male respondents.

The qq plot suggests a heavy distribution towards the median of the sample. This is not surprising since the age range (18-35) we sampled may not have been representative of all colleges (the population in which you are wanting to have a similar distribution for). The upper tail also suggests there were not many male students towards the end of the distribution in the sample than to be theoretically expected.

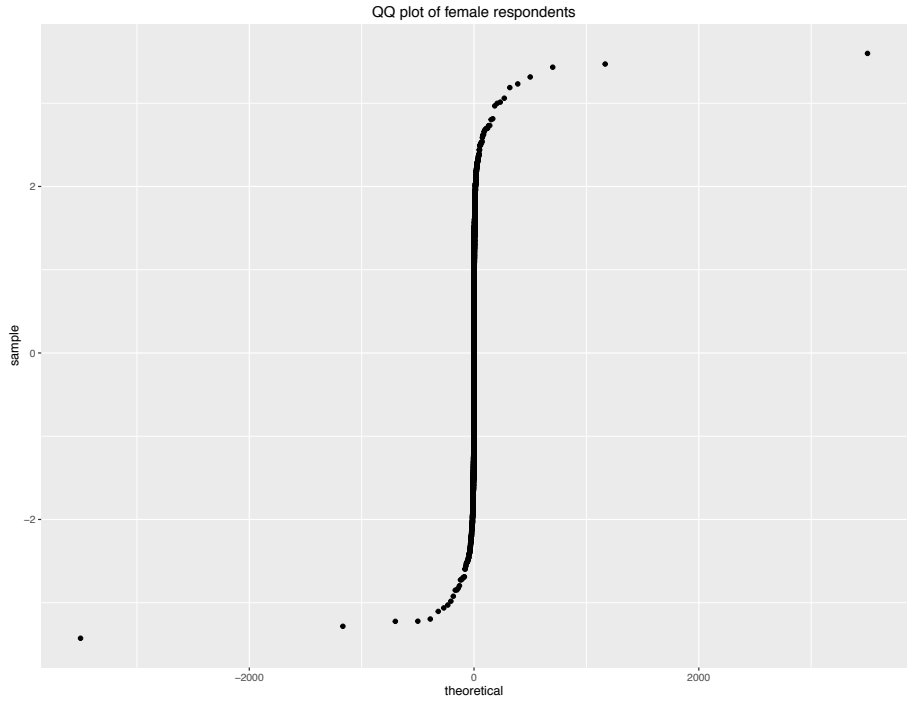

Figure 17. QQ plot of female respondents.

The qq plot for female respondents is very similar in which the distribution is heavily towards the median. What's more distinctive is that there appears to be a "more"/expected distribution towards the upper tail than the male respondents which suggest the sample for females could be more representative of the population in which you are collecting data from. However, further investigation is needed to confirm if this speculation is statistically significant.

4. Random values were generated for X any Y with a uniform distribution using the runif() function. Samples were collected at 10, 50, 100, and 200. N was then increased by 200, until 2,000, increased again by 2,000, until 20,000, increased again by 20,000 all the way to 100,000 samples. Admittedly, the reasoning for this was because of my lack of familiarity with R and not knowing how to programmatically vary the N's. However, I believe I was able to obtain enough samples to allow me to see a pattern with the different file types. Below are two examples of the plots that were being saved.

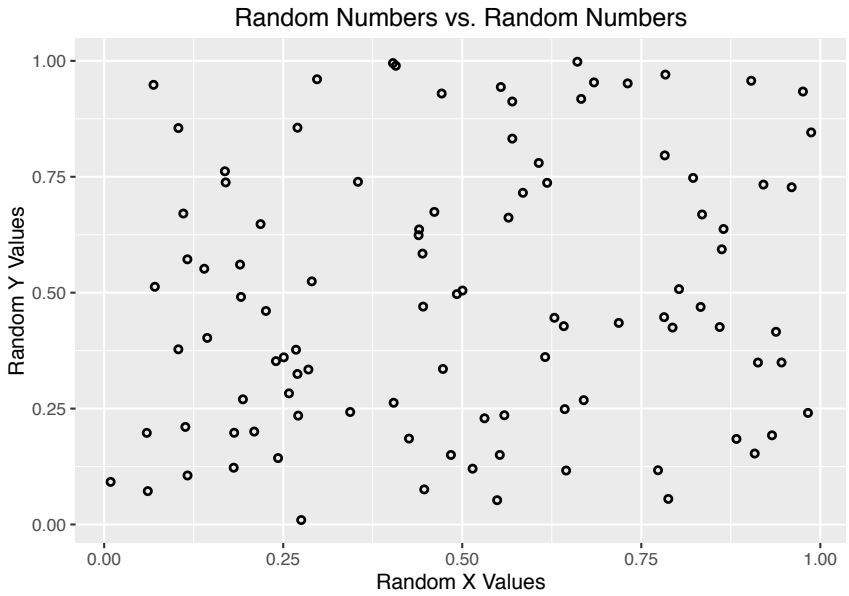

Figure 18. N = 100. Notice, there doesn't appear to be a relationship between X and Y (which is to be expected).

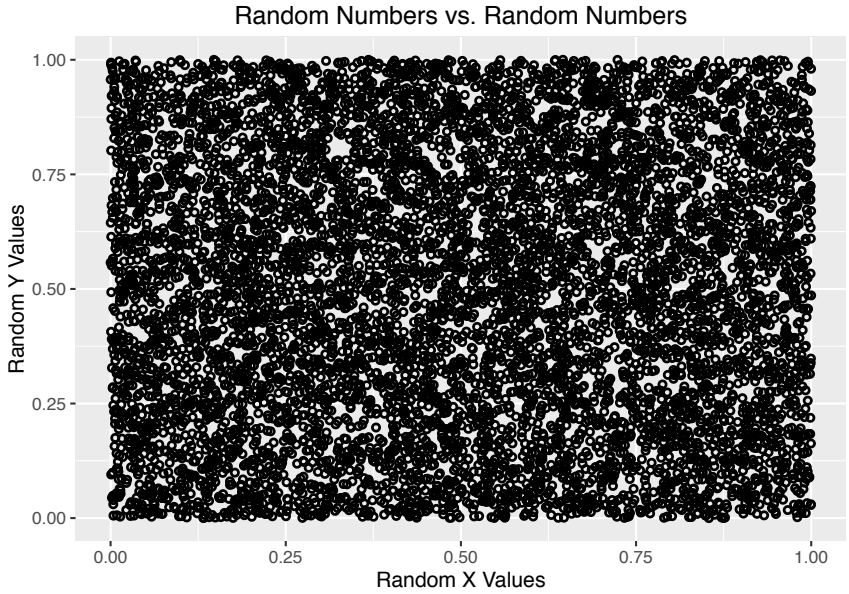

Figure 19. N = 8000. Again, no relationship is apparent in this plot.

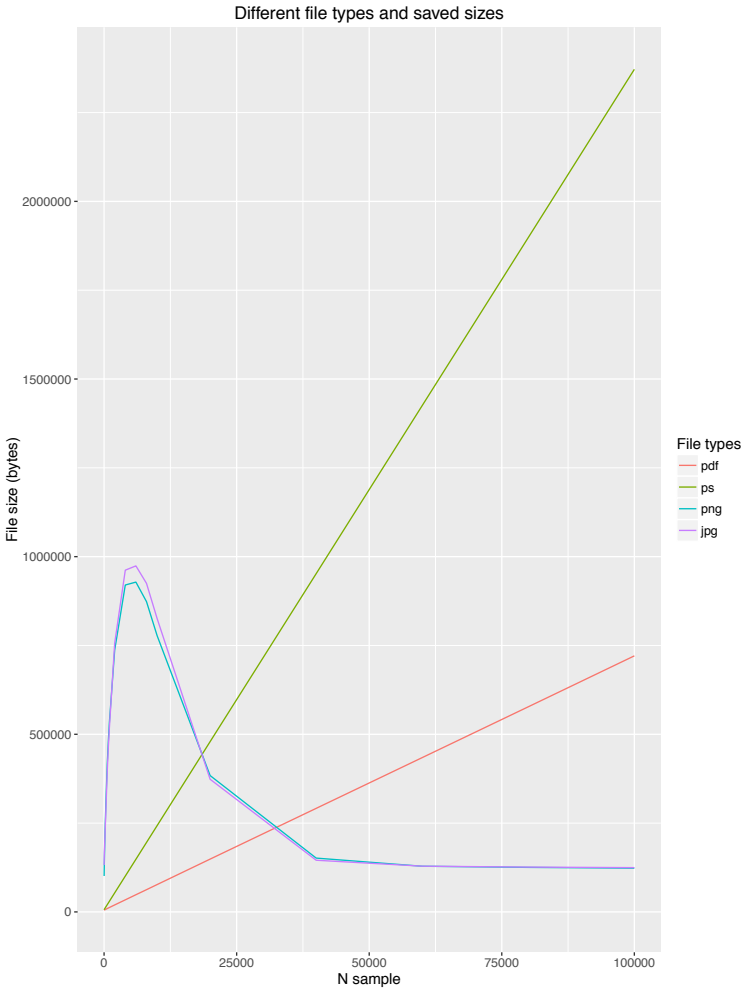

Figure 20. Line plot of the different file types varying N from 10 to 100,000.

Other than the png and jpg files, the other file types are quiet distinct from one another. Pdf files showed a linear growth and were the smallest file type compared to the others until about N=2600 where png and jpg files begin to become smaller. Ps files also showed a linear growth and were smaller than png and jpg files until about N=12,000. As per described in the reading of Data Visualization Chapter 10, page 341-342, pdf and ps files use vector graphics which allow for zooming "in to arbitrary precision" (i.e., a higher resolution is produced in these file types). Note, the slope for ps files are significantly higher than the slope of pdf, suggesting pdf files are more efficient with file compression.

The png and jpg files also showed an interesting and near-identical pattern to one another throughout the samples gathered with a logarithmic growth and then significant decrease in size. This is due to the raster graphics used by png and jpg filesThe jpg files were slightly larger compared to png files at samples around 4,000 to 10,000, however it doesn't seem too significant and both eventually are practically identical at N=10,000. What was more surprising was that the file sizes for png and jpg files eventually decreased to about 120 KB from being 900 KB at

samples from 4,000 to 10,000 and stayed at about 120KB all the way to N=100,000. This leads me to believe that the compression algorithms for png and jpg files are highly adept at handling images with large pixel counts, while underperforming, comparatively, to pdf and ps files at lower pixel counts.

5.

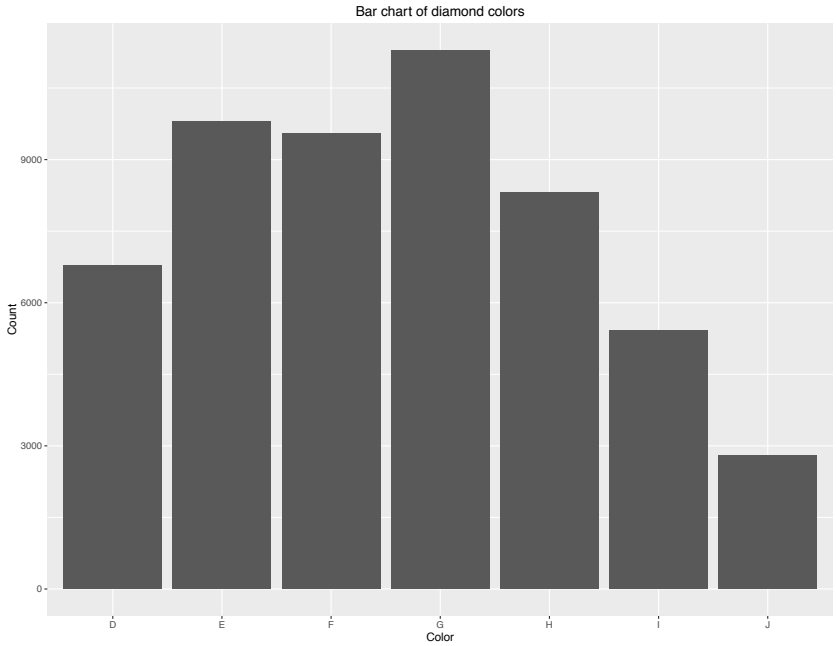

Figure 21. Bar chart of diamond colors.

G is the most frequent diamond color with a count of 11,292 diamonds. That is reasonable to expect, but the second most frequent is E – the second from the worst color. Although the difference is probably insignificant between E and F, it's interesting to see that E is the second most frequent in this data set. Regardless, the distribution seems almost normal (with a slight left skew that indicates lower popularity with J colored diamonds).

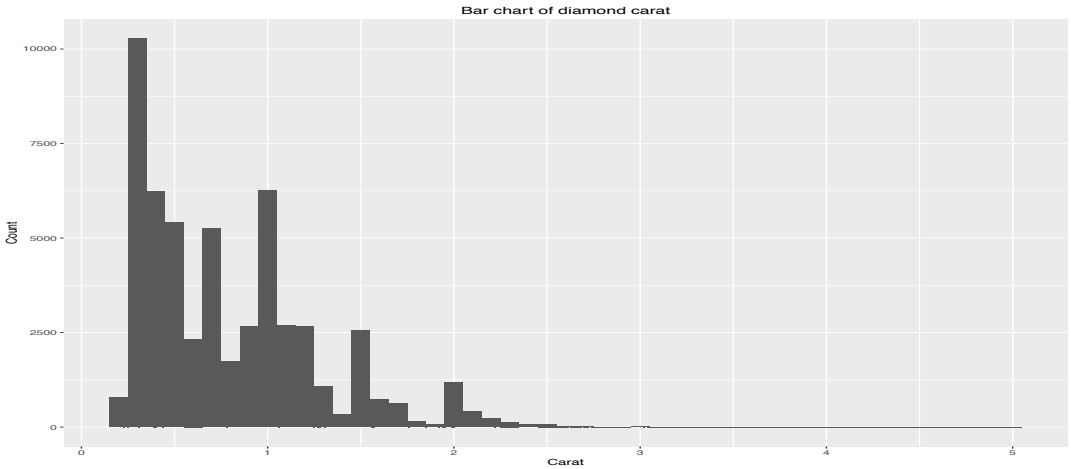

Figure 22. Bar chart of diamond carats.

The bar chart for carat is a little more interesting. Figure 23 shows a summary statistic table of the carat variable. There is a left skew in the graph that suggest diamonds are typically smaller than 2 carats. This is supported with the median being 0.7 and average diamond size at 0.7979. There is a huge spike in the frequency of 1 carat diamonds. We have a wide range of carat sizes from 0.2 to 5.01 carats.

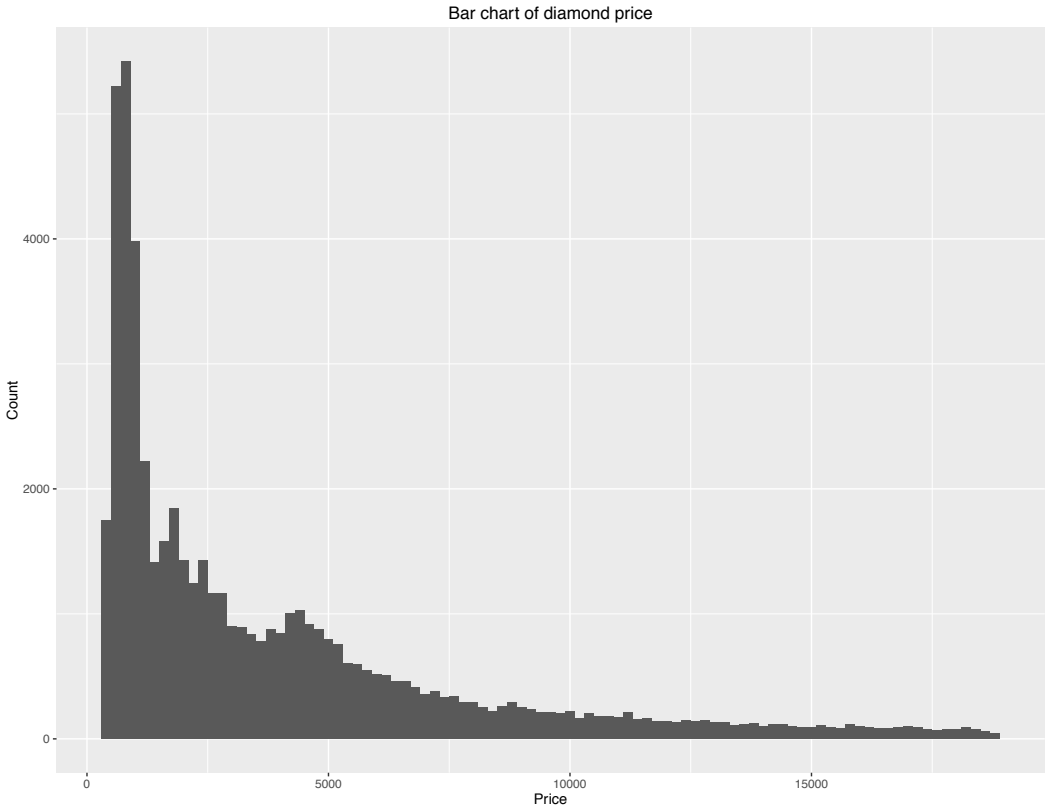

Figure 23. Histogram of diamond prices.

Again, we see a left skewed histogram for the price of diamonds with a median of $2,401.00 and average of $3,932.80. 36.44% of the diamonds in the data set have an above average price. What's the most interesting about this data is the range of the diamond prices from $326.00 to $18,823.00, suggesting a very robust sample of data.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Carat	0.2	0.4	0.7	0.7979397	1.04	5.01
Price	326.0	950.0	2401.0	3932.7997219	5324.25	18823.00

Figure 24. Stats for diamond carats and prices.

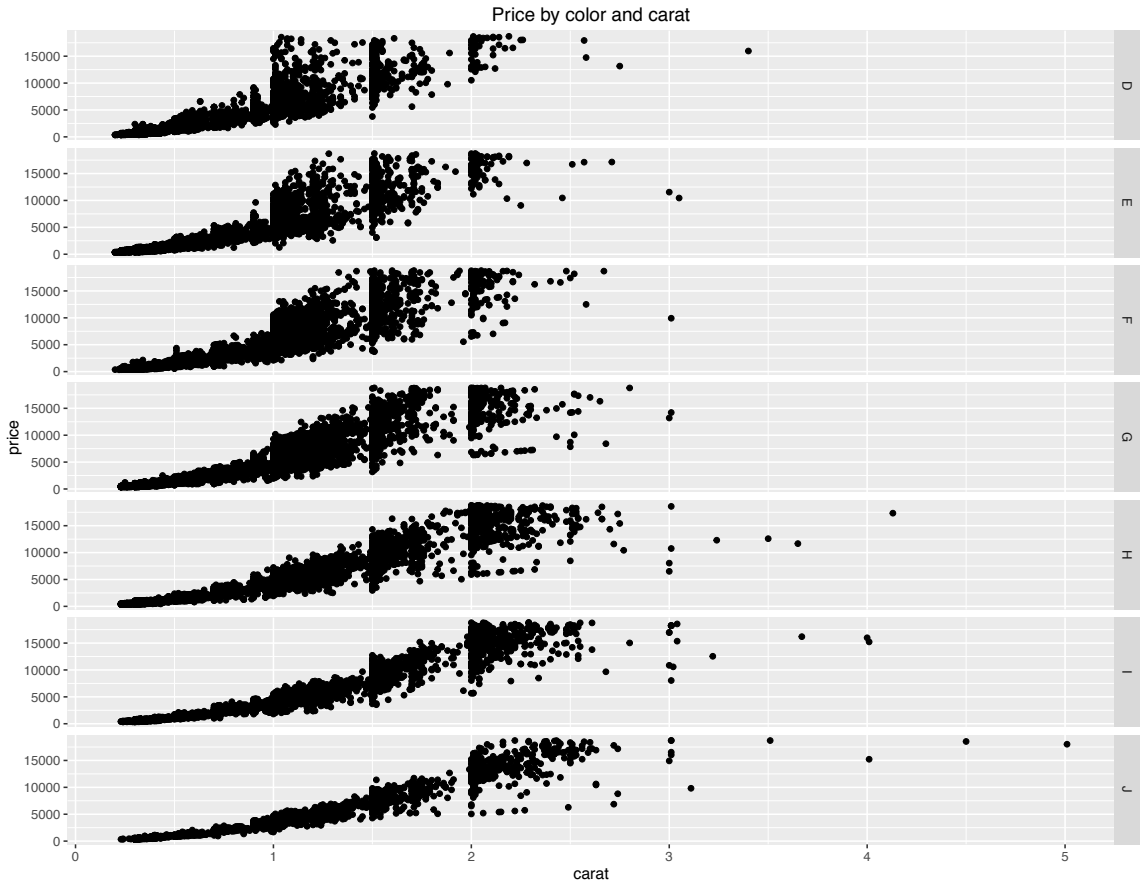

Figure 25. Three-way relationship between carat, price, and price.

The price of the diamond appears to increase exponentially in all colors and carat sizes. There are three distinct price jumps/clusters across diamond colors: D. E. F. G The first one is at 1 carat, the next at 1.5, and the next at 2 carats. Diamond colors: H, I, J, also has two distinct clusters at 1.5 and 2 carats, but are missing the clear clusters seen in the other colors. A speculative reason might be that diamond cutters purposefully sell only at these carat sizes and/or it could be that diamond buyers seek out certain size brackets. There appears to be more outliers as you look at worse color. All colors appear to have similar ranges (i.e., you are able to choose a J colored diamond at a cheaper price, but may have to sacrifice other attributes such as color or size (carat).