

hw2_report

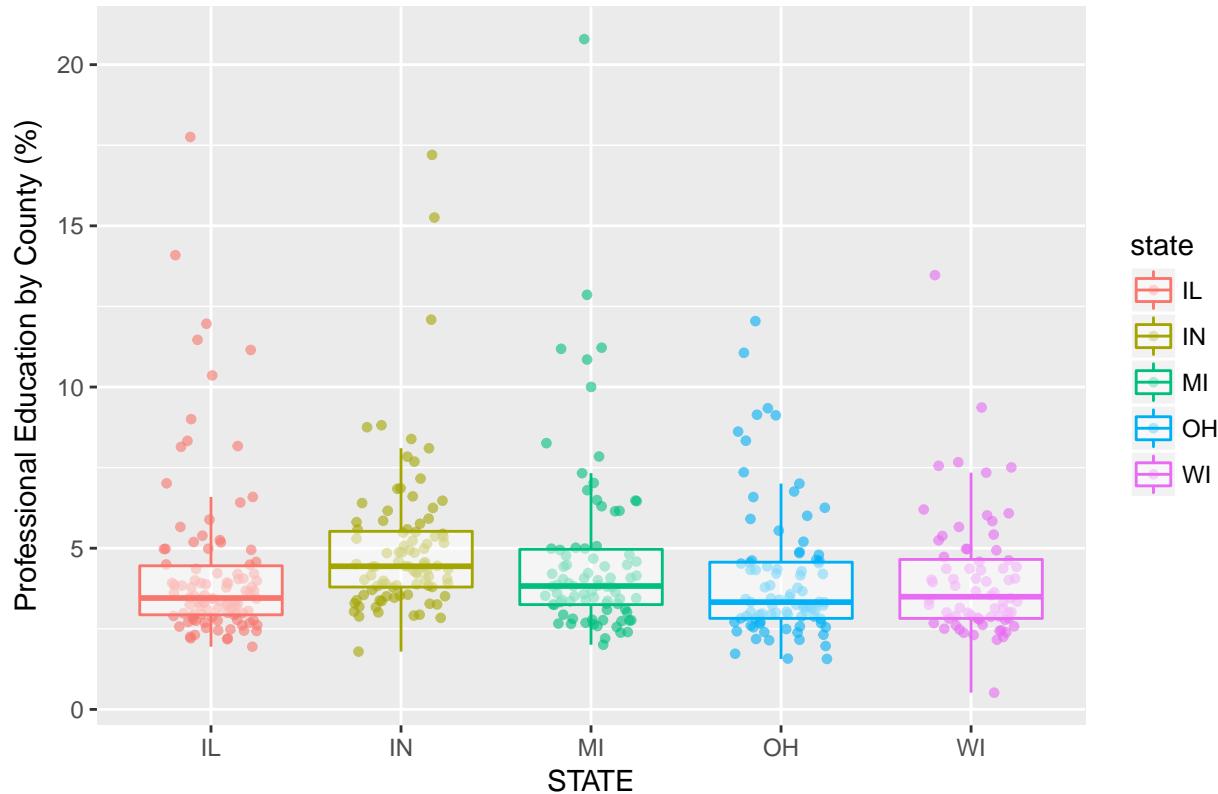
jlonai3

September 24, 2017

Problem 1

A) Visualize the % of people that have a professional education for each county, grouped by state.

Percent County Population with Professional Education by State



```
##   state median percprof
## 1    IL    3.455354
## 2    IN    4.440127
## 3    MI    3.827592
## 4    OH    3.328012
## 5    WI    3.495100
```

B) Describe Distributions by pointing out 2 relevant/interesting properties from your plot.

My plot shows that OH has the lowest median percentage of people with professional education (3.33%) and IN has the highest median percentage (4.4%). It is also interesting to Note some of the outliers. For instance MI has the county with the highest professional education percentage (~20.8%) but WI has the county with the lowest percentage of only 0.52%

C) Can you point out which state has the lowest and highest percentage of population with a professional education?

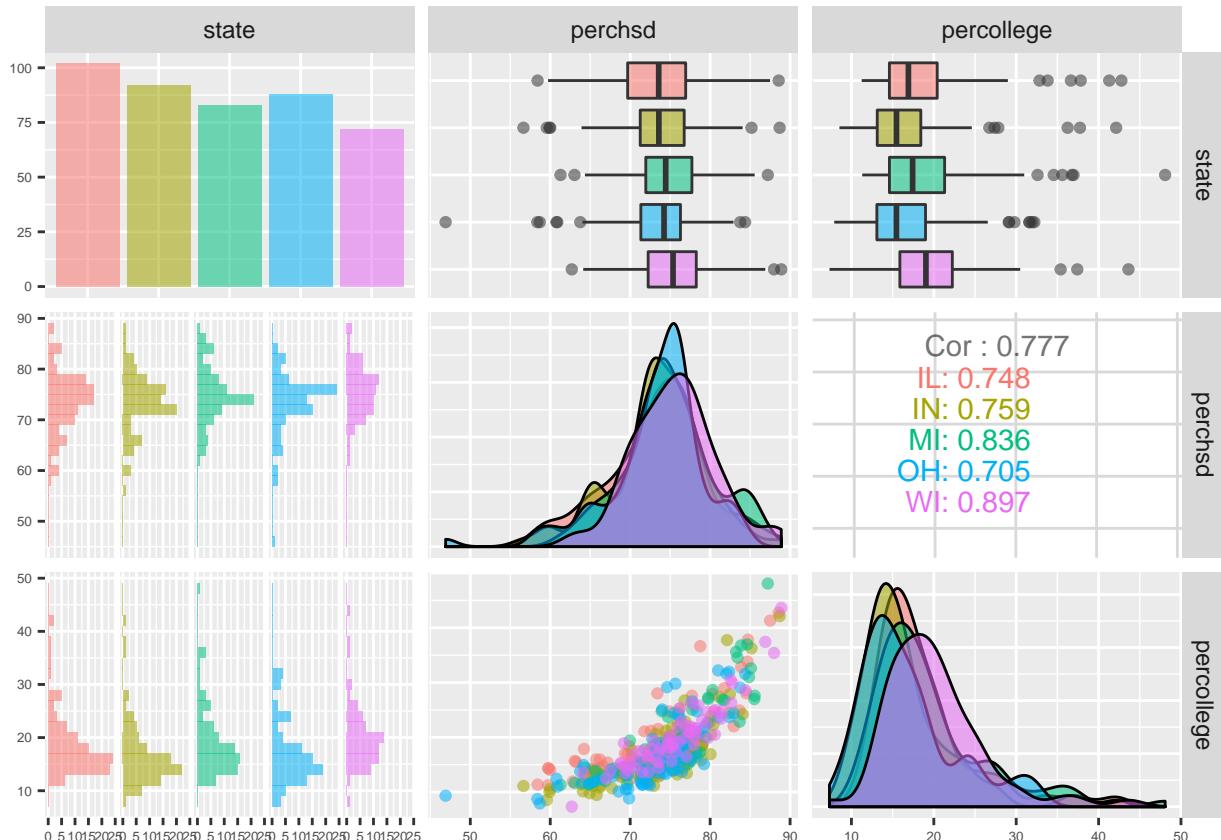
Simply from the graph you cannot tell which state has the highest percent population with professional education because we are not showing a metric that represents the population for each county. For example, MI has the county with the highest percentage of professional education but what if this is the smallest county in MI? We need more analysis to determine the state total percent education. Below I have calculated the number of people with professional education for each state as well as the total population for each state as provided in the midwest dataset. **The results are illustrated in the table below and show that actually IL has the highest overall percentage of the population with a professional education!(7.53%)**

```
##   state poptotal pop.prof.edu percentEducate
## 1    IL 11430602     860466      7.53
## 2    IN  5544159     356366      6.43
## 3    MI  9295297     601641      6.47
## 4    OH 10847115     639716      5.90
## 5    WI  4891769     275716      5.64
```

Problem 2

A) Explore the three-way relationship between the percentage of people with a high school diploma in each county (perchsd), the percentage of college educated population in each county (percollege), and the state. Illustrate these relationships using 3 separate plots (perchsd vs. state, percollege vs. state, perchsd vs. percollege), or a combined pair-wise plot

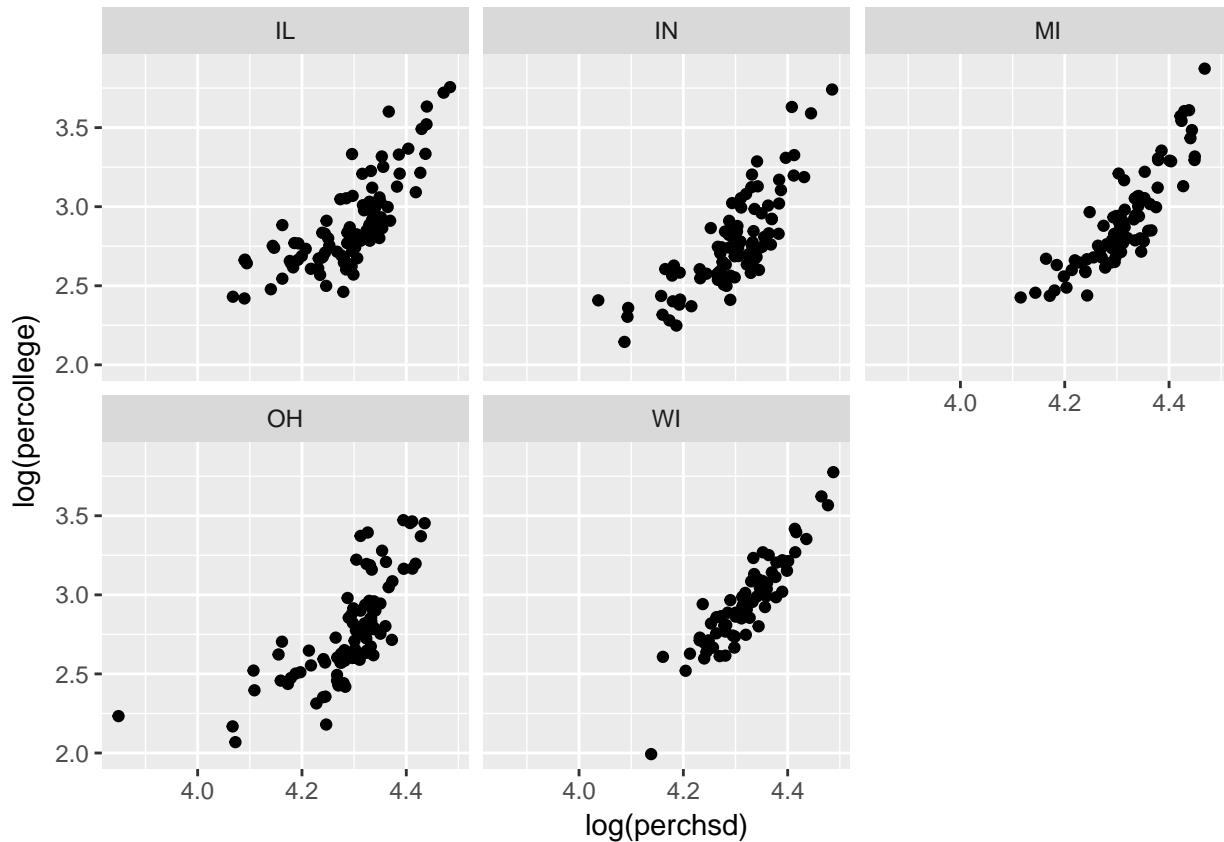
We first plotted a ggpairs plot for the data. The top middle box plot and the left-middle histogram set shows the relation between percent county with high school diploma and state, and the top right box plot and left-bottom histogram set show the relation between percent of county with college diploma by state. The bottom middle plot is a scatter plot that shows the relationship between high school diploma and college diploma percentage for each county, colored by state.



B) For at least 2 of the 3 pairs, describe the relationship you observe (if inconclusive, explain why).

The box plots/histograms make it evident that almost all states have a similar percentage of students graduating with a high school diploma. The median high school diploma percentage across all counties for each state is between 70 and 80 percent. The same can be said for college diploma as, but with a much smaller median percentage between 15 and 20 percent for all states.

A more interesting relationship can be seen when we examine that between the percent of high school diplomas and the percent of college diplomas in each county. Based on the bottom-middle scatter plot in the ggpairs plot above we can see a somewhat exponential relationship between the percent of a counties population that graduated high school and the percent that graduated college. I decided to take a more in depth look and plotted each state out separately with a log-log scale (see images below). These graphs shows that all states show an somewhat linear relationship between percent with highschool diploma and percent with college diploma on a log-log scale. The best example of this being WI, showing a very linear relationship in log-log scale.



Problem 3

A) Describe the different elements of a Box Plot and how they illustrate different statistical properties of a sample. Show a Box Plot diagram with these elements labeled (need not be based on actual data).

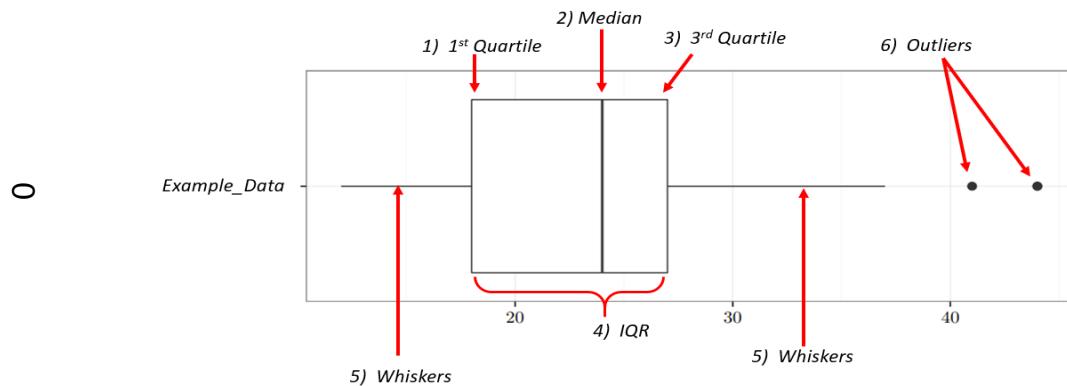
- 1) 1st Quartile (or 25-percentile) is the point at which 25% of the data is below, and 75% is above.
- 2) Median is the “middle-point” in the data set. It is the location at which 50% of the data falls below and 50% falls above. It is shown as the solid line which bisects the box in a box-plot.
- 3) 3rd Quartile (or 75-percentile) is the point at which 75% of the data is below and only 25% is above.
- 4) The Inner Quartile Range (IQR) is shown as the box in a box plot and represents the center 50% of the

data. It stretches from the 1st Quartile to the 3rd Quartile.

5) Whiskers are the lines that extend in either direction from the box. These lines run to the most extreme high and low data points in the sample that are less than or equal to 1.5 times the IQR window width away from the edge of the IQR box in the figure.

6) Outliers are any data point outside the reach of the whiskers in a box plot. This means points that are farther than 1.5 times the IQR window's width away from the edges of the box (IQR).

—All Items are numbered in the figure below—



B) When would you use a Box Plot, a Histogram, vs. a QQPlot to graphically summarize data? Describe the primary purpose and mention 1 example use case for each type of plot.

Box Plot - Primary purpose is to summarize a numeric data set into several easy-to-read values. Box plots makes it easy to see the spread and median of data sets and compare them to one another. For example, if you had the data for the average distance people drove to work for every county in every state you could create a box plot to show the distribution of miles driven for each state. This would allow you to easily compare the spread(IQR)and median distance driven in each state to all other states.

Histogram - Primary purpose is to summarize single dimensional data and see the rough distribution of values across a data set. Histogram allows you to see what values occur most frequently in your data set. For example, if we had dataset that contained the time of day that people most enjoyed running we could create a histogram and find out what times were most popular. My guess would be that such a histogram would show two peaks with one in the morning and one in the evening (aka cool times of day, and before/after work).

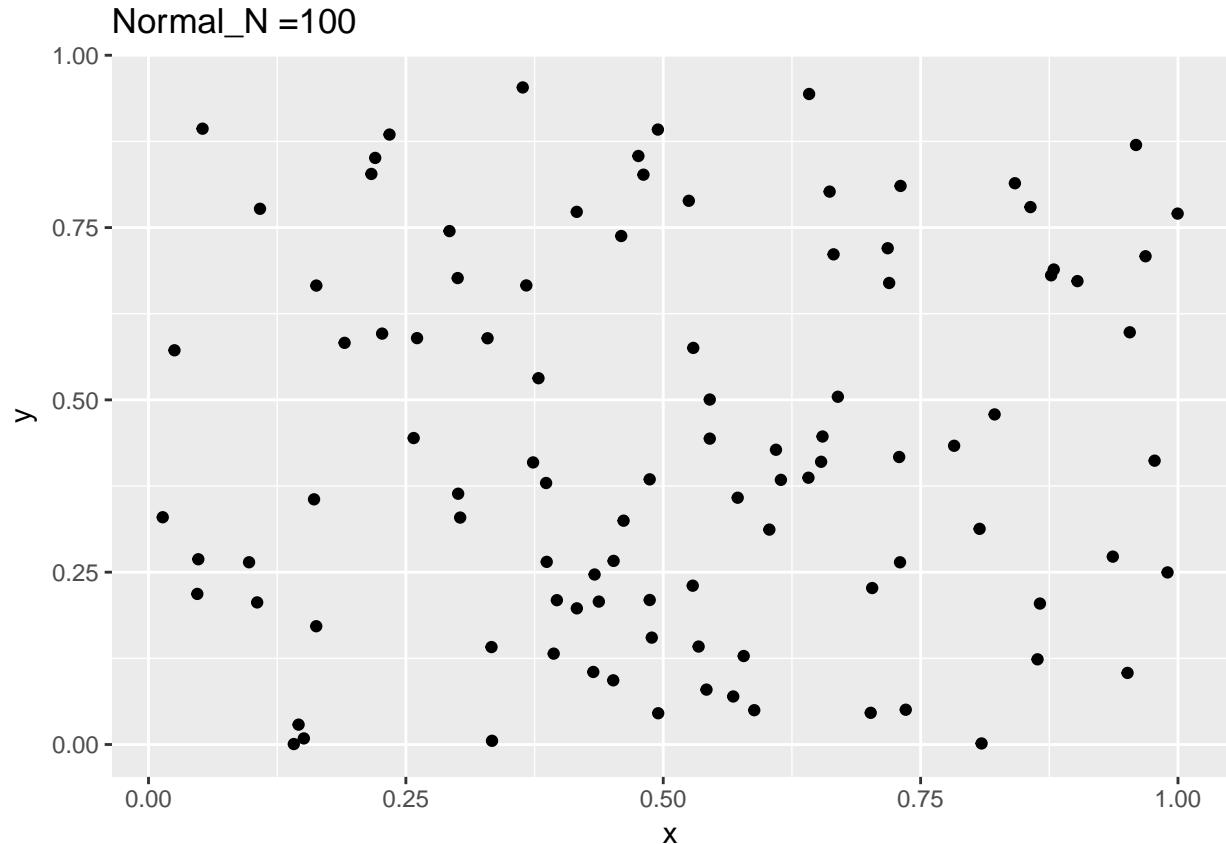
QQplot - Primary purpose of a QQplot is for comparing two different data sets. Often one of the data sets will be drawn from a certain distribution (normal, Poisson, etc). For example if you have a set of data from a survey and want to know if your data is normally distributed you can use a QQplot and set

the second dataset to be a set of data points from a known normal distribution with a certain center and width. The resulting graph will allow you to tell if you have a similar origin, spread, and or tail distribution.

Problem 4

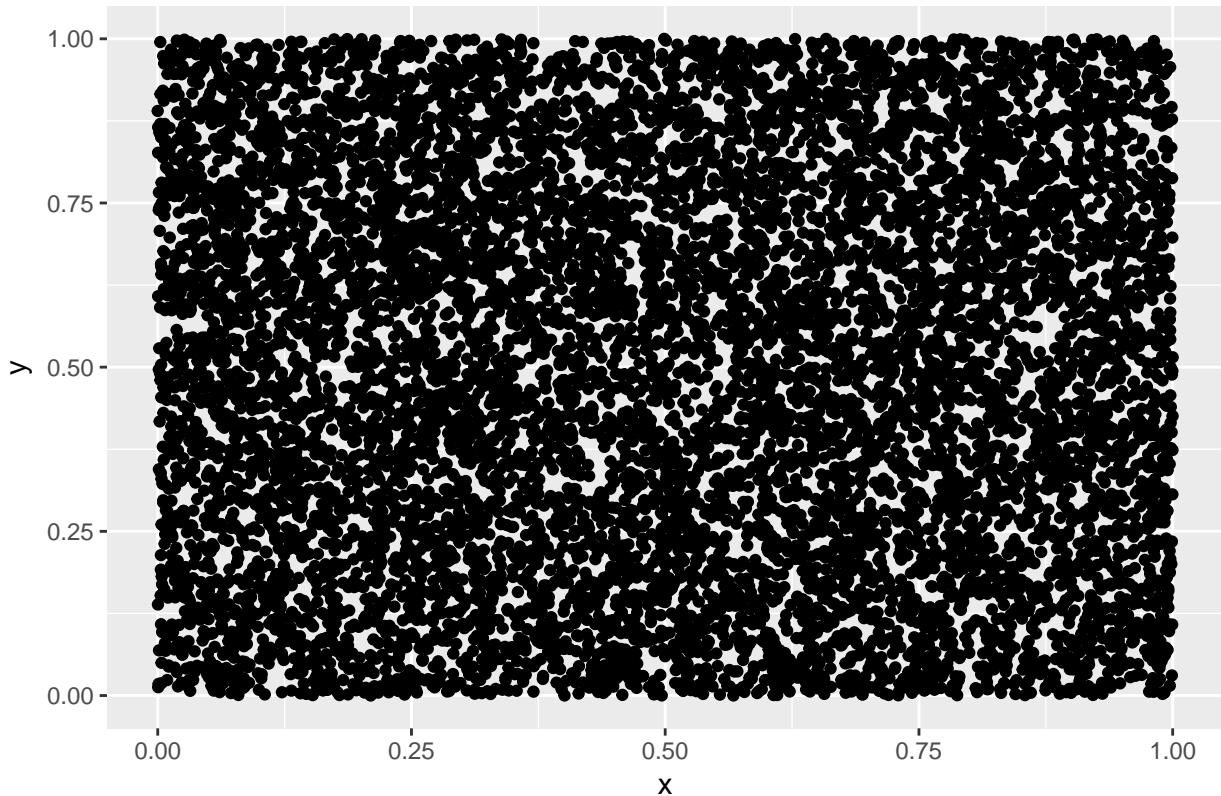
A) Generate two sets of N random uniformly-distributed values using the function runif() and display a corresponding scatterplot using one set as X-values and the other as Y-values.

For this section I plotted graphs for values of N_sizes:1E2,1E3,1E4,1E5,1E6. Lets plot a couple example graphs for this report. The first graph will have a low value for N (N=100)



For the second example graph we will use larger value of N=10,000.

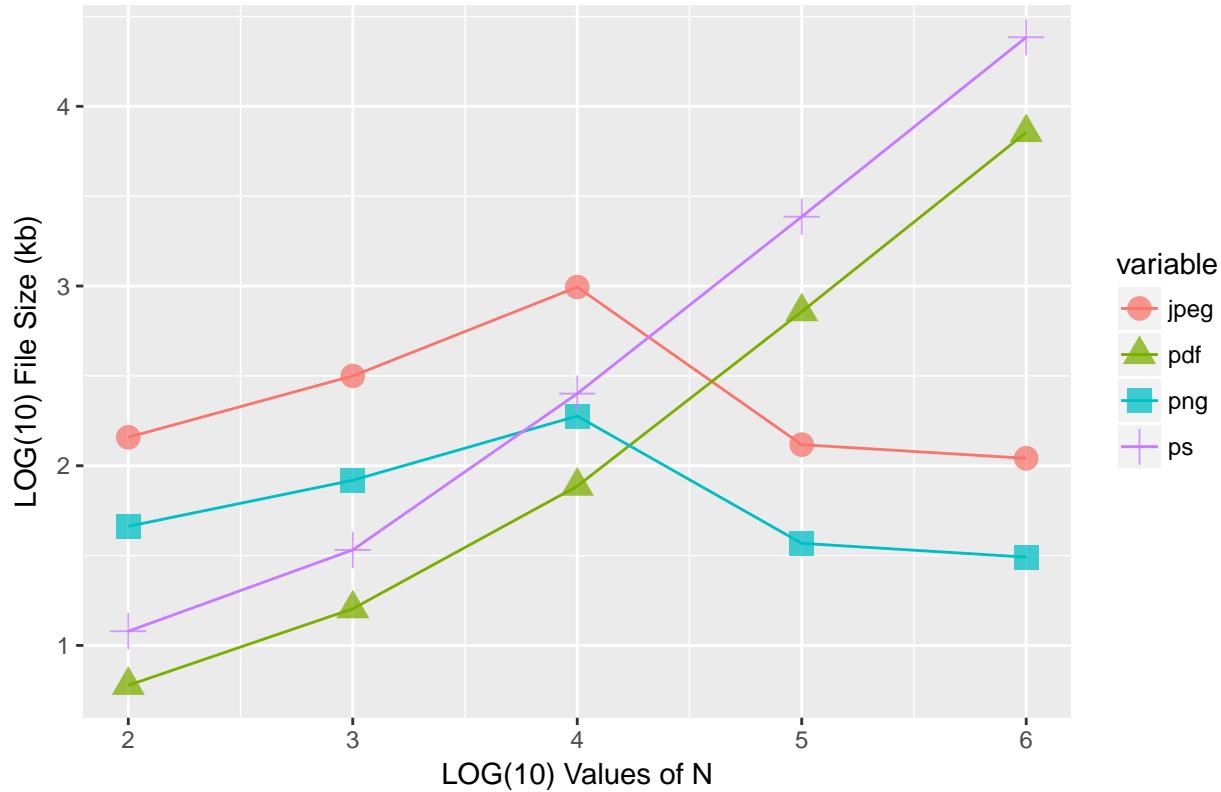
Normal_N =10000



I created a FOR LOOP that iterates through my five different sizes for N, creates a ggplot for each just like those shown above, and then saves them in the following formats(jpeg, png, ps, pdf).I then opened the folder I saved all of these files to and recorded the file sizes (kb) in a data frame(shown below).

```
##      N jpeg  pdf png    ps
## 1 1e+02 144    6  46   12
## 2 1e+03 316   16  83   34
## 3 1e+04 987   77 189  252
## 4 1e+05 131  720  37 2431
## 5 1e+06 110 7154  31 24223
```

File Size vs. File Type



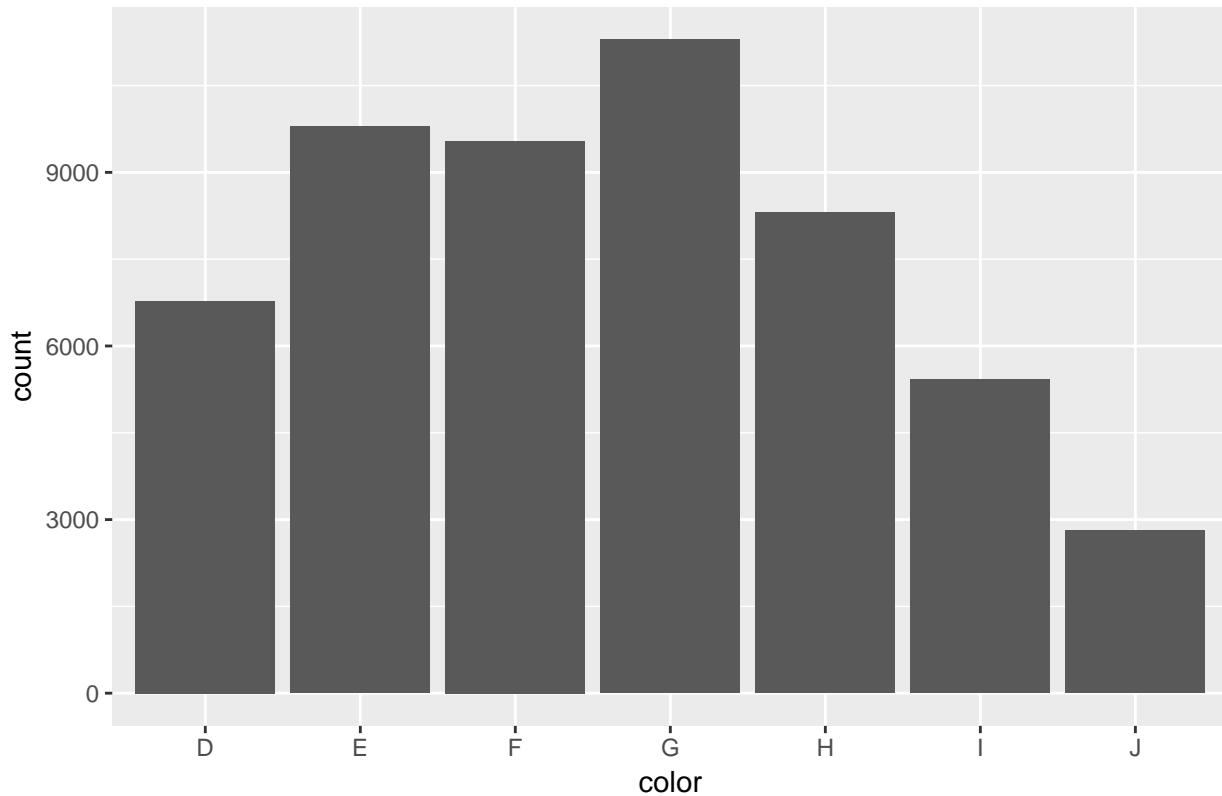
I plotted the file size vs. file type data for my different values of N using a log-log scale. I used base 10 orders of magnitude for my values for N, so the log-log scale in the graph was calculated using base 10 for both the x and y axis.

The pdf and ps file types show the expected increase in file size as our values of N increased. We can easily see that both these file types show exponential increase in file size with increasing N. However, the jpeg and png files show an interesting trend. The file size increases up to N=1E4, but at N=1E5 we see that the file size starts to decrease again. The reason for this is that for values of N>= 1E5 the plot becomes almost completely black. Since jpeg and png use image compression, for these graphs rather than a white graph with +10,000 black points, the compression software sees a black graph with only a few dozen white spots of open graph showing through. These graphs thus require less space to compress and store so we see them drop off in file size until the image is completely black at N=1E6 and the file size is even smaller than the lowest value with N=1E2.

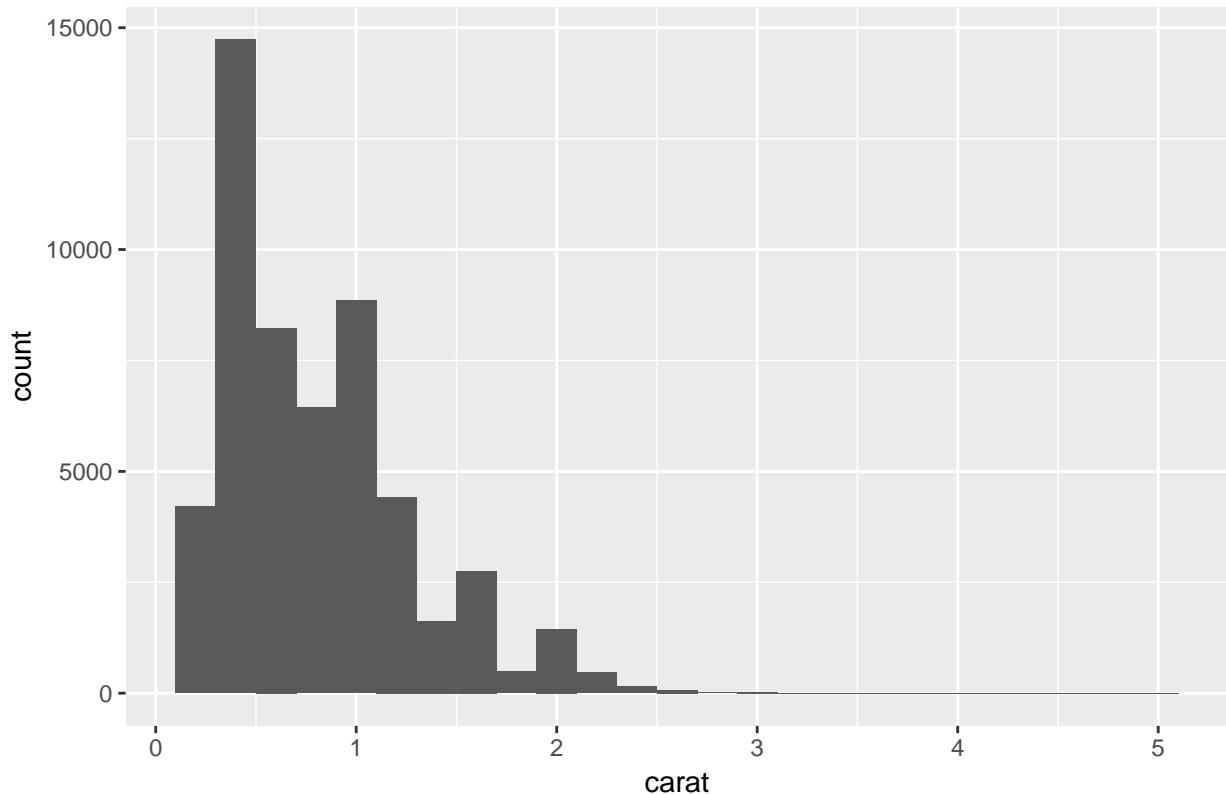
Problem 5

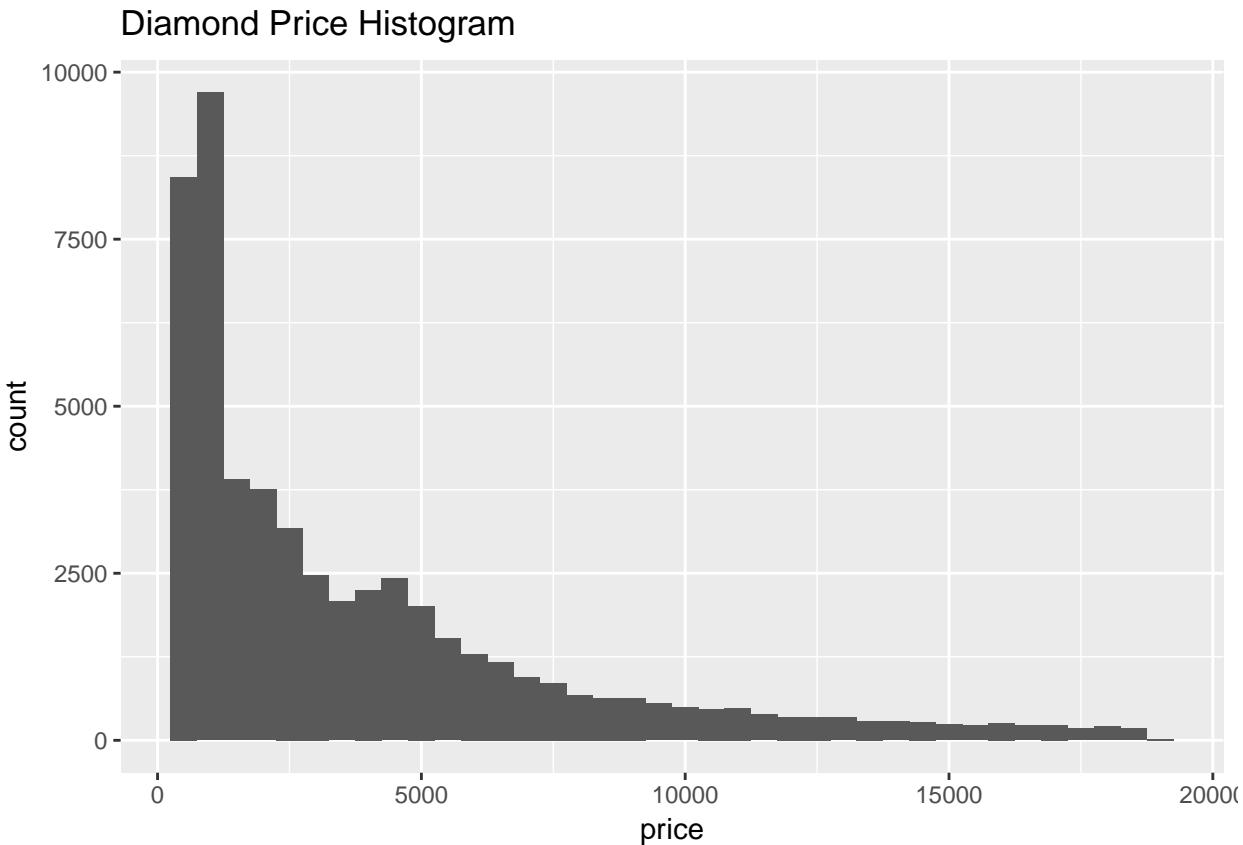
- A) Plot histograms or bar charts for color, carat, and price, illustrating their distributions. What can you infer from these distributions

Diamond Color Histogram



Diamond Carat Histogram





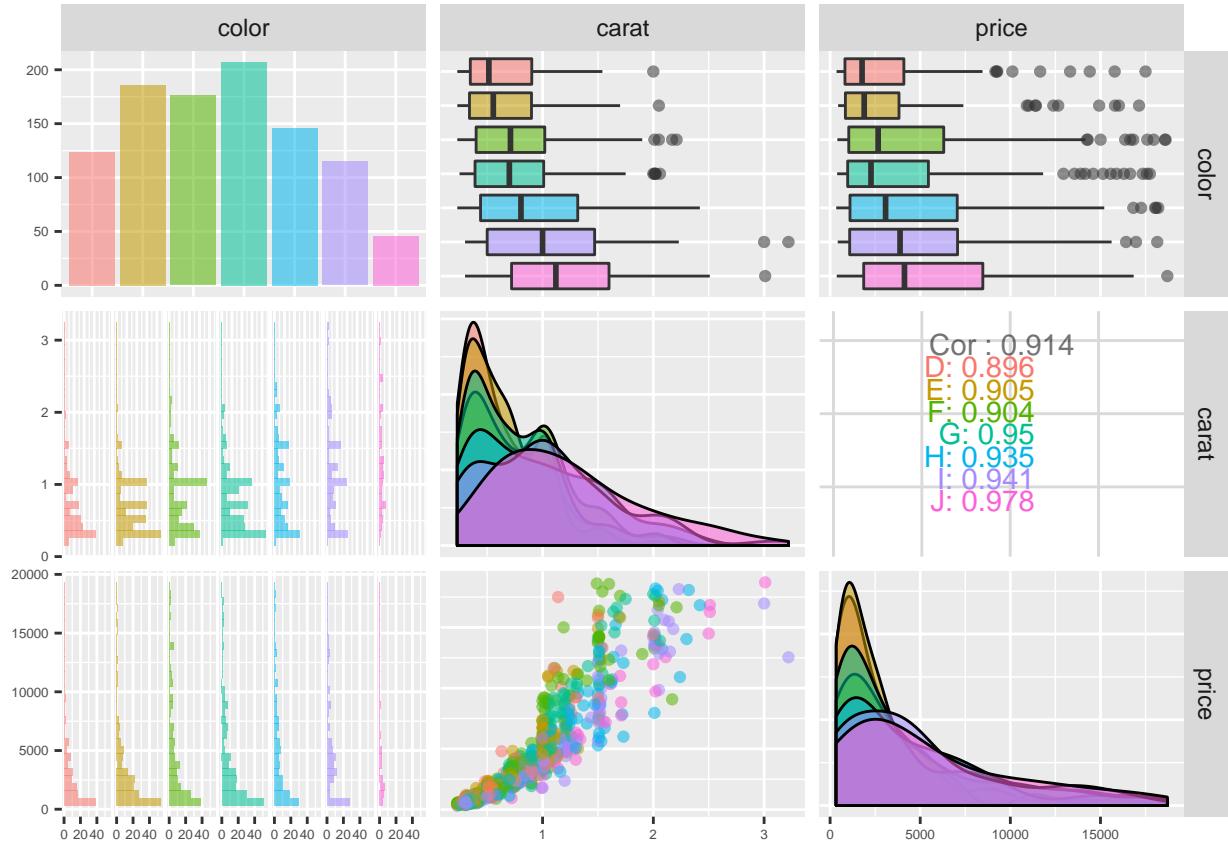
Color Histogram This histogram doesn't tell us too much about the distribution of diamonds. We can tell that 'J' is the least common diamond color in the sample, while 'D' and 'I' are a little more common and 'E-H' are all present in pretty similar amounts. If we can assume the order of diamond color is intentionally alphabetic, we can see a faint normal distribution skewed with the center shifted to the left.

Carat Histogram We can see a normal distribution skewed partially to one side so it is centered between 0.5 and 1.0. The tail declines quickly as we would expect from a normal distribution.

Price Histogram The number of diamonds available for a certain price shows an exponentially decreasing distribution as price increases (like normal distribution skewed completely to one side). We can see many diamonds are available for sale at low prices, but only a few diamonds are available at any of the high to very high prices (long tail).

B) Investigate the three-way relationship between color, carat and price. What are your conclusions? Provide a combined pair-wise plot and/or separate graphs that illustrate these relationships.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the ggpairs graph above we can see a couple things. 1) The color vs. carat histogram shows us that several different color diamonds have a bi-modal distribution of carat sizes. This is especially clear for F,G,H diamonds and is something we couldn't extract from the total carat histogram. 2) The distribution for carat and price as shown by the box plots parallel one another. This means colored diamonds with higher median carat sizes will have a higher median price, and those with a lower median carat size will have a lower median price. 3) This is further exemplified in the carat vs. price scatter plot where we see an exponential increase in price as carat size increases. I wanted to further examine this relationship so I plotted the carat vs. price for each color separately in the facet plot below on a log-log scale. The results show a nice linear relationship on the log-log scale, with G, and H colored diamonds showing a bit of a heavier tail.

```
## `geom_smooth()` using method = 'loess'
```

