CSE6242 HW2: Data Visualization

X. Sheldon Gu (xgu60)

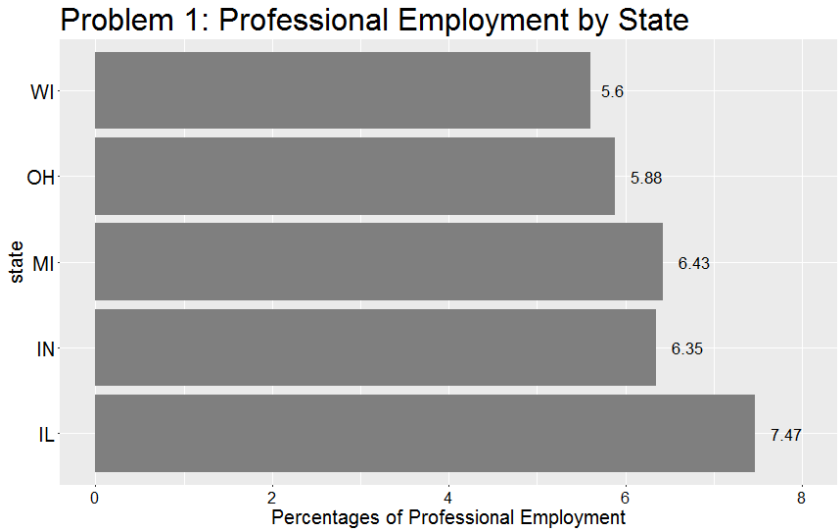
Problem 1.

Answer: (*Interpretation A*) The *midwest* dataset has each county's adults population ('popadults') and percentage of professional employment ('percprof'). I used the following formula to calculate the percentage of adult population with a professional employment for each state:

$$percprof.state = sum(popadults * percprof) / sum(popadults)$$

The final results are plotted in Figure 1. It clearly shows that WI has the lowest percentage of professional employment (5.6%), while IL has the highest percentage of professional employment (7.47).

Figure 1: Percentage of professional employment of each state in *midwest* dataset.



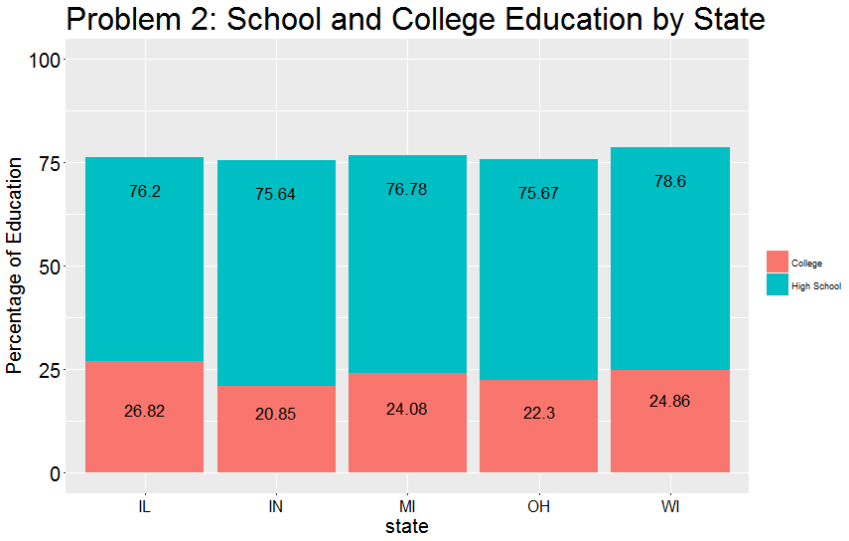
Problem 2.

Answer: (*Interpretation A*) I used similar formulas to calculate the percentage of people with high school diploma and the percentage of people with college education.

$$perchsd.state = sum(popadults* perchsd) / sum(popadults)$$
$$percollege.state = sum(popadults * percollege) / sum(popadults)$$

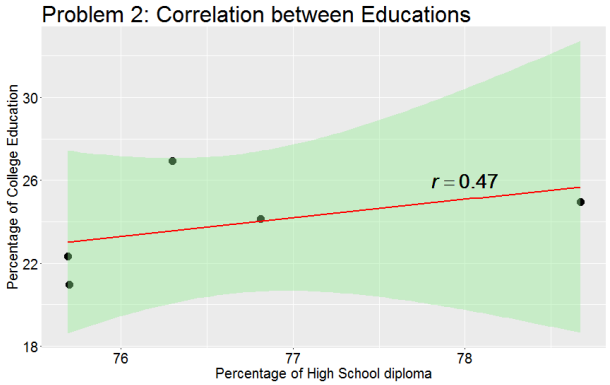
The results are plotted in Figure 2A. It shows WI has highest percentage of high school diploma (78.6%) and IN has the lowest percentage of high school diploma (75.64%), while IL has highest percentage of college education (26.82%), and IN has the lowest percentage of college education (20.85%).

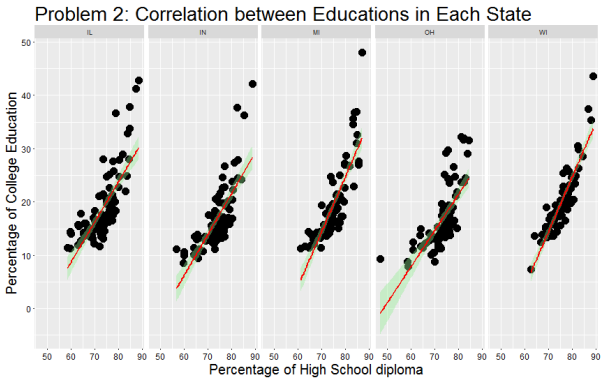
Figure 2A: Percentage of high school diploma (blue) and college education (red) for each state in *midwest* dataset.



It seems a week correlation (r = 0.47) between states' percentages of high school diploma versus percentages of college education (Figure 2B). The weak correlation may due to the lack of data points. (*Interpretation B*) I further investigated their correlations at different states using their county data (Figure 2C). They all show good correlations between high school diploma and college education.

Figure 2B: the correlation between states' percentage of high school diploma and percentage of college education (left). **C:** the correlation between counties percentage of high school diploma and percentage of college education (right).





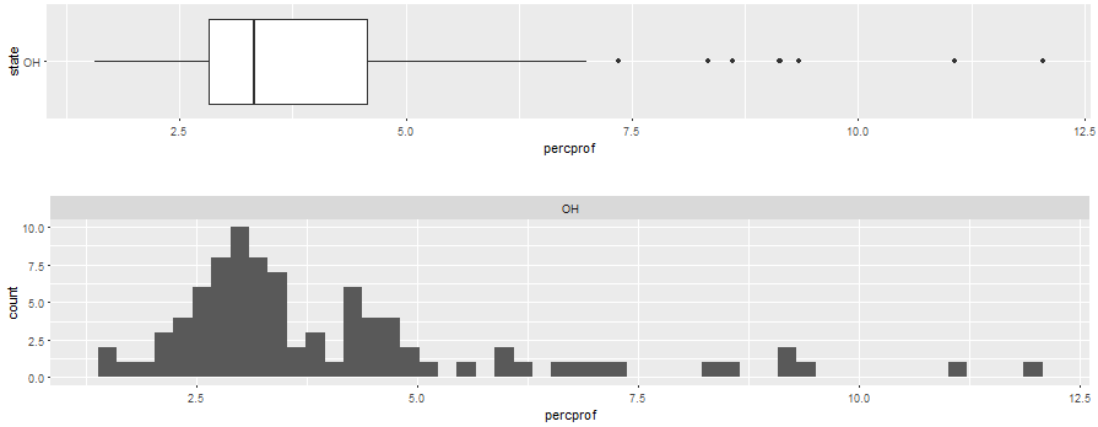
Problem 3.

Answer: Boxplot shows statistic information of a dataset in an elegant way. The box shows 25 and 75 percentile of the dataset. The middle black line in the box shows median (50 percentile) of dataset. The interval between the first and third quartiles is called the inter-quartile range (IQR). The whiskers extend to the most extreme points, but must be less than 1.5 times the length of the IQR. If there are still points not covered in box or whisker, they are called outliers and are plotted separately.

Figure 3A shows the distribution of percprof of Ohio state using Boxplot (upper) or histogram (lower). The boxplot gives lots of important statistic information, e.g. median, first quartile, third quartile, but does not give too much details as histogram (show bi-model). Histogram gives detailed distribution information, but does not clearly show where is the median, first quartile, and third quartile.

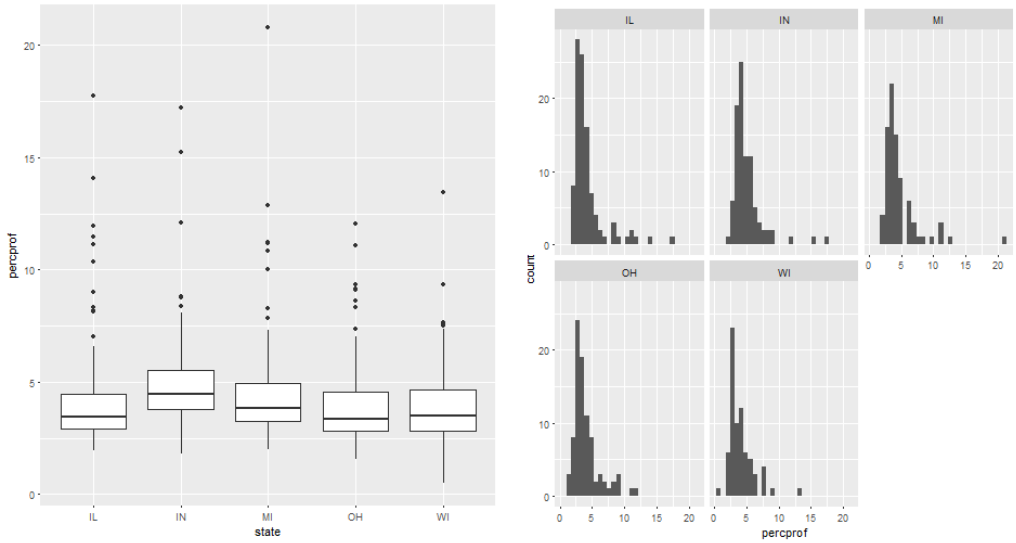
Although boxplot does not give detailed distribution information, it still gives some information about the distribution by check the location of median in the box, and the whiskers. If a dataset is normal distributed, the boxplot should be symmetrical with median line exactly at the center. The percprof data of OH state is skewed as shown in histogram (Figure 3A lower), thus in boxplot (Figure 3A upper) the median line is located at left side of the box. If a dataset is left skewed, then the median line will be located at the right side of the box. The information about the distribution of data provided by boxplot is not as accurate as histogram.

Figure 3A: Boxplot of percprof of Ohio state (upper) and histogram of percprof of Ohio state (lower).



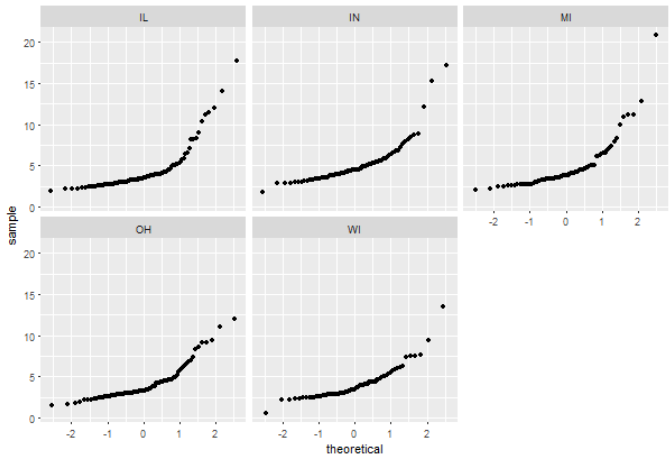
Another advantage of boxplot is that it is very convenient for comparison between different datasets. The plot is much clean and elegant. Figure 3B shows the percprof data of each state in midwest dataset in boxplot (left) and histogram (right). Histogram plot is kind of crowd and hard to compare between state to state.

Figure 3B: Boxplot of percprof of states in midwest dataset (left) and histogram (right)



qq-plot (Quantile-quantile plots) is very useful for comparison the distribution of two datasets. It can also be used to tell whether a dataset is normal distributed by plotting the dataset with a normal distributed dataset. Figure 3C plots the distributions of county percprof of each state in midwest dataset against standard normal distributed dataset. It clearly shows none of them is normal distributed which is consist with histogram shown in Figure 3B.

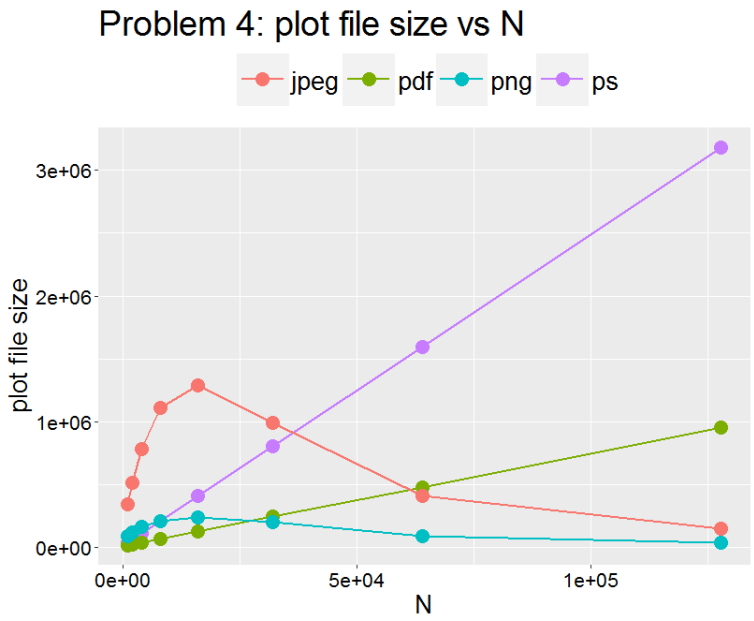
Figure 3C: qq-plot of percprof of each state in midwest dataset.



Problem 4:

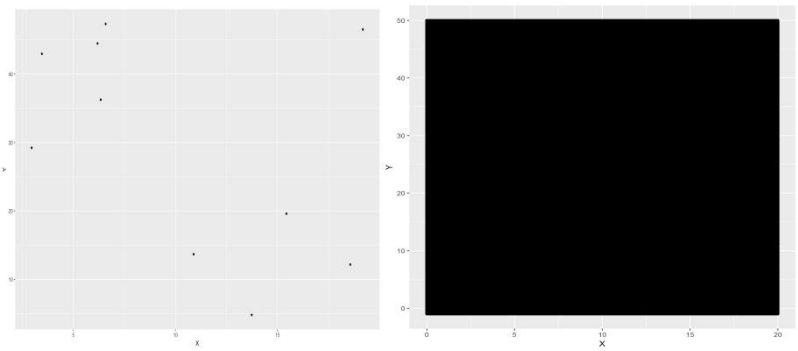
Answer: As shown in Figure 4A, when N is small (< 1000), the saved files' sizes follow the order: jpeg > png > ps > pdf, and it looks like a linear increment (at least for small N). However, when N further increases (> 10000), the files' sizes of jpeg and png decrease, while files' sizes of pdf and ps keep increasing in a linear mode. When N reaches 10,000,000, the new order is ps >> pdf >> jpeg > png.

Figure 4A: File size versus N for jpeg, pdf, png and ps.



There are different techniques to store data points. For pdf or ps files, they saved actual data information into file, thus the file size increase with point numbers. However, jpeg or png store data after compression. When the data points reach a threshold, jpeg or png no longer store the information of actual points but a black block (Figure 4B).

Figure 4B: Scatter plots with 10 points (left) and 10000000 points (right) stored as png files.

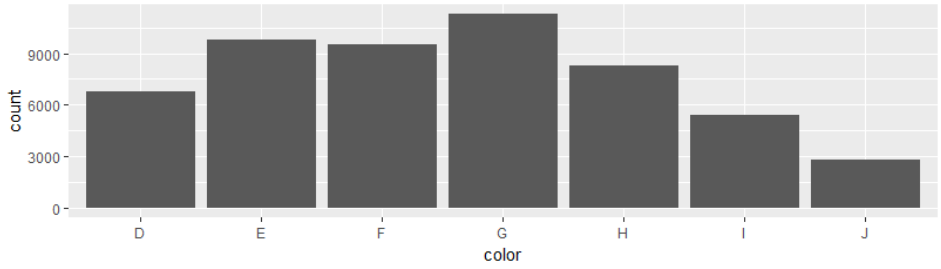


Problem 5:

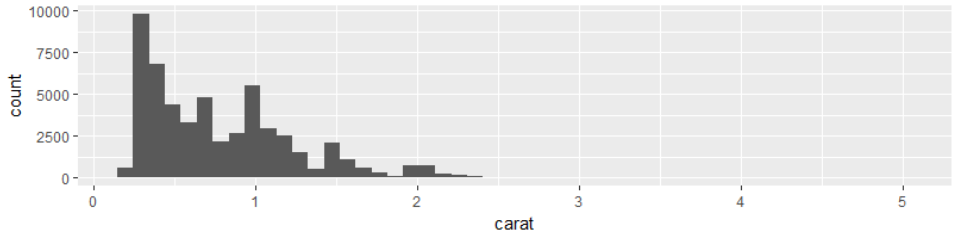
Answer: Figure 5A1 shows the number of diamonds with different color scales follows the order: G > E > F > H > D > I > J. Figure 5A2 shows majority diamonds are small, the number of diamonds decreases with the carat, but it does not monotonically decrease, and it is a relative large number for diamonds that weight 1 carat. Figure 5A3 shows that majority of diamonds have low price. The number of diamonds increases with the price at the very beginning, then quickly decreases with the price, and there are a few diamonds have very high price (tailing for the histogram).

Figure 5A: Bar chart for color (1), and histograms for carat (2) and price (3) for diamonds dataset.

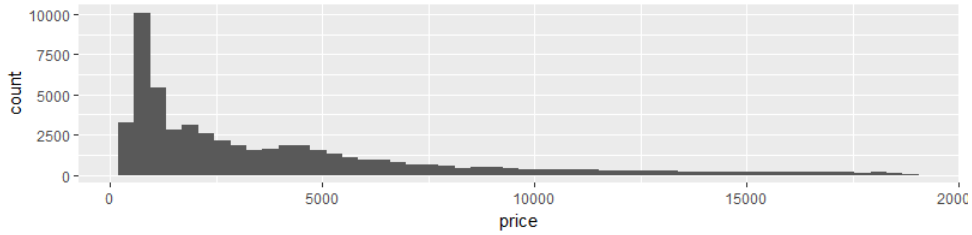
1.



2.



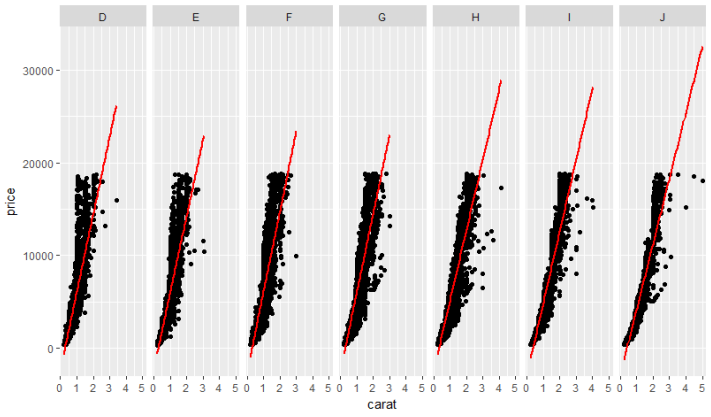
3



Some observations about the relationships between price, carat and color:

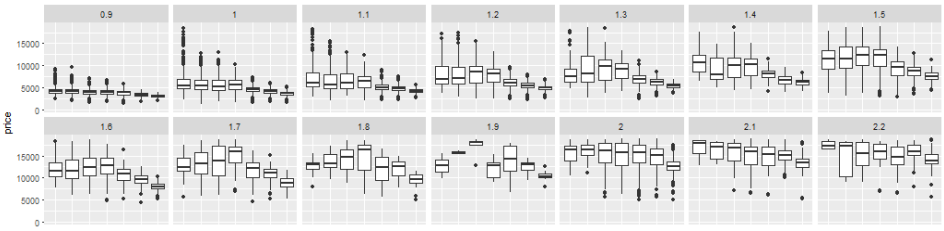
1. Price is highly correlated with carat. Large carat diamonds have high price in all color scales.

Figure 5B: Scatter plot of price vs carat for each color scale diamonds.



2. For diamonds in a small range of carat, diamonds with color scale H,I,J have relative low prices compared with other color scales.

Figure 5C: Box plots the prices of diamonds with different color scales in small range of carats.



3. Diamonds with color scale I, J are larger that diamonds with other color scales. It is difficult for nature to generate big and colorless diamonds.

Figure 5D: Box plots the carats of diamonds with different color scales.