Question 1: Professional Education by State

Figure 1.1: Boxplot of percentage of people having professional education by state

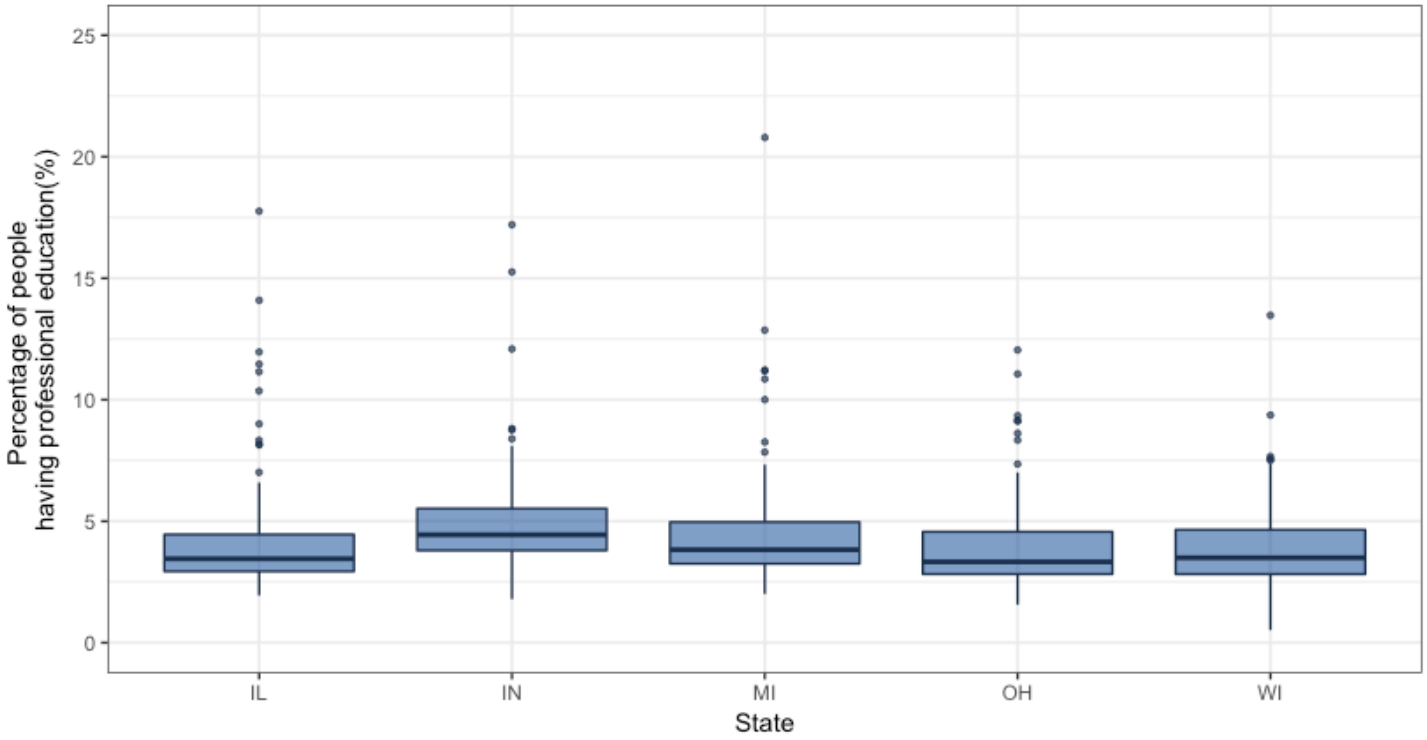


Figure 1.1 is the visualization of the distribution of “percprof” values grouped by state using box plot. We could infer the following interesting properties on the basis of Figure 1.1:

1) State IN has the highest mean value of “percprof” among five states.
2) State MI, OH and WI have similar mean of “percprof” values, but the “percprof” values of state MI are more variable than others.
3) Sate MI has the maximum “percprof” values among five states.

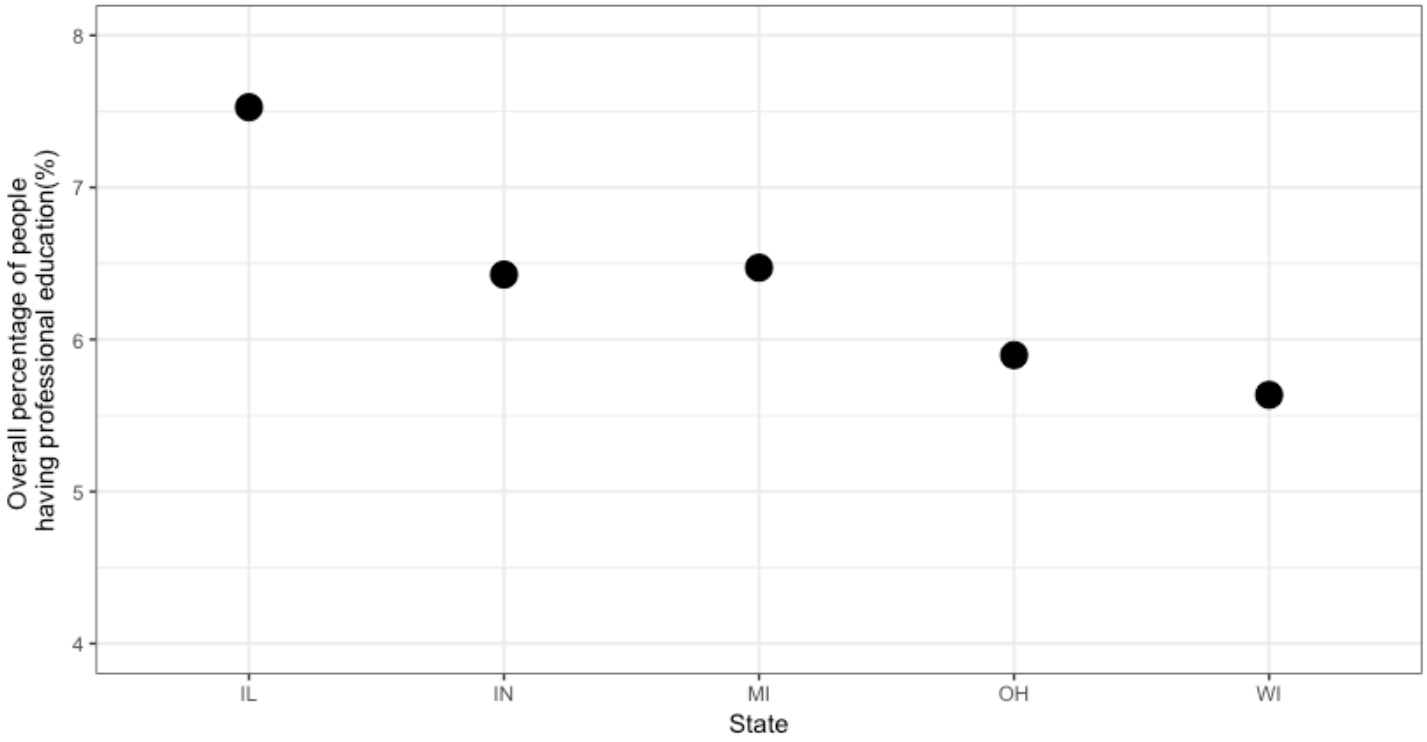
The overall percentage of people that have a professional education for each state could be calculated by the following function:

$$(P_1 * N_1 + P_2 * N_2 + P_3 * N_3 + P_4 * N_4 \ldots\ldots + P_m * N_m) / (N_1 + N_2 + N_3 + N_4 + \ldots\ldots N_m)$$

where P_m is the percentage of people that have a professional education in county m, and N_m is the number of total population in county m.

The results were summarized in Figure 1.2 and it shows that state IL has the highest percentage of population with a professional education, and state WI has the lowest percentage of population with a professional education.

Figure 1.2: Scatter plot of overall percentage of people having professional education by state



Question 2. School and College Education by State

Figure 2.1: Relation between percentage of college educated population in each county and percentage of people with a high school diploma in each county

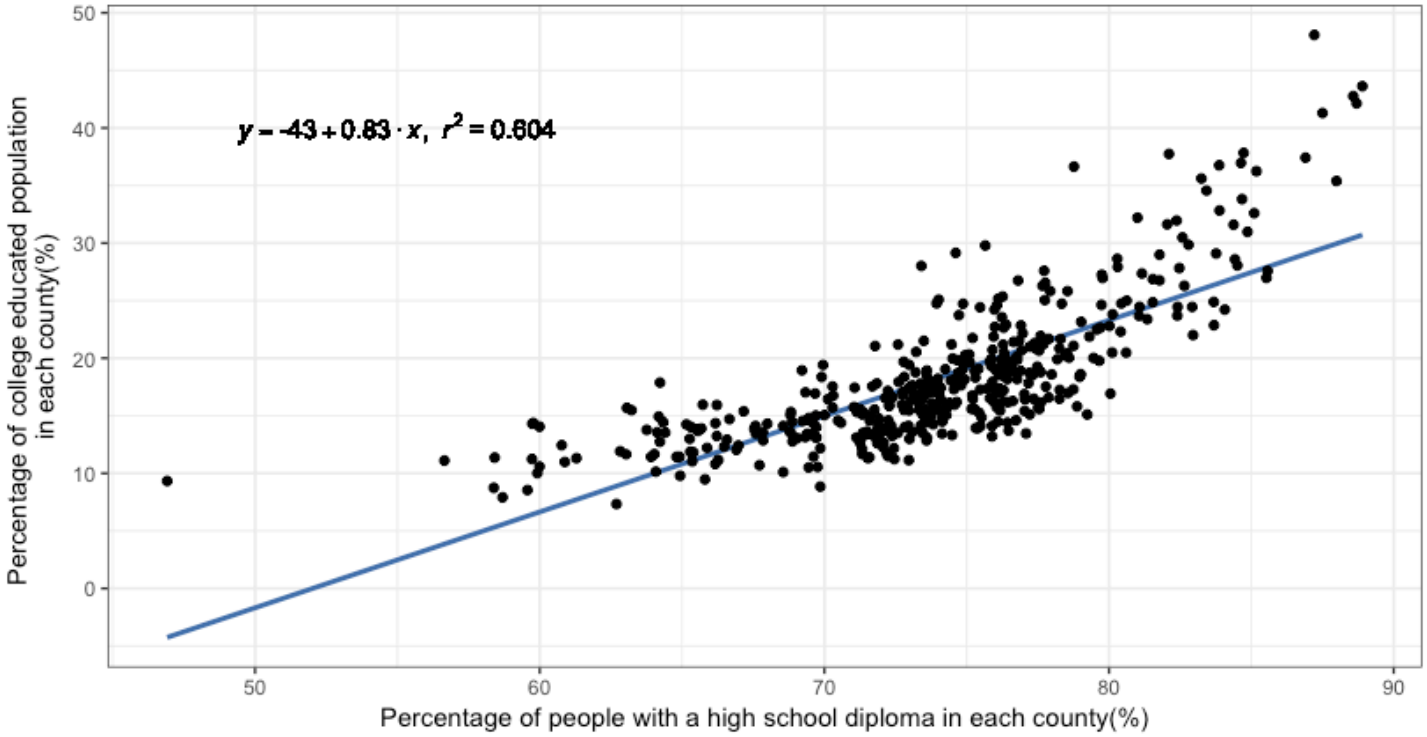


Figure 2.1 illustrated the relationship between the percentage of people with a high school diploma in each county (perchsd), and the percentage of college educated population in each county (percollege). From Figure 2.1 we could see the coefficient of determination (R^2) is 0.604, which indicates “perchsd” has a moderate linear relationship with “precollege”. The Pearson Correlation Coefficients between “perchsd” and “precollege” is 0.777 and it also reveals there is a positive correlation between “perchsd” and “precollege”.

Figure 2.2 shows that the mean of the percentage of people with a high school diploma in each county (perchsd) of the states are similar, but the “perchsd” values of state OH are more variable than others.

Figure 2.3 shows that state WI has the highest mean of the percentage of college educated population in each county (precollege), and the “percollege” values of state OH are less variable than others.

Figure 2.2: Boxplot of percentage of people with a high school diploma in each county by state

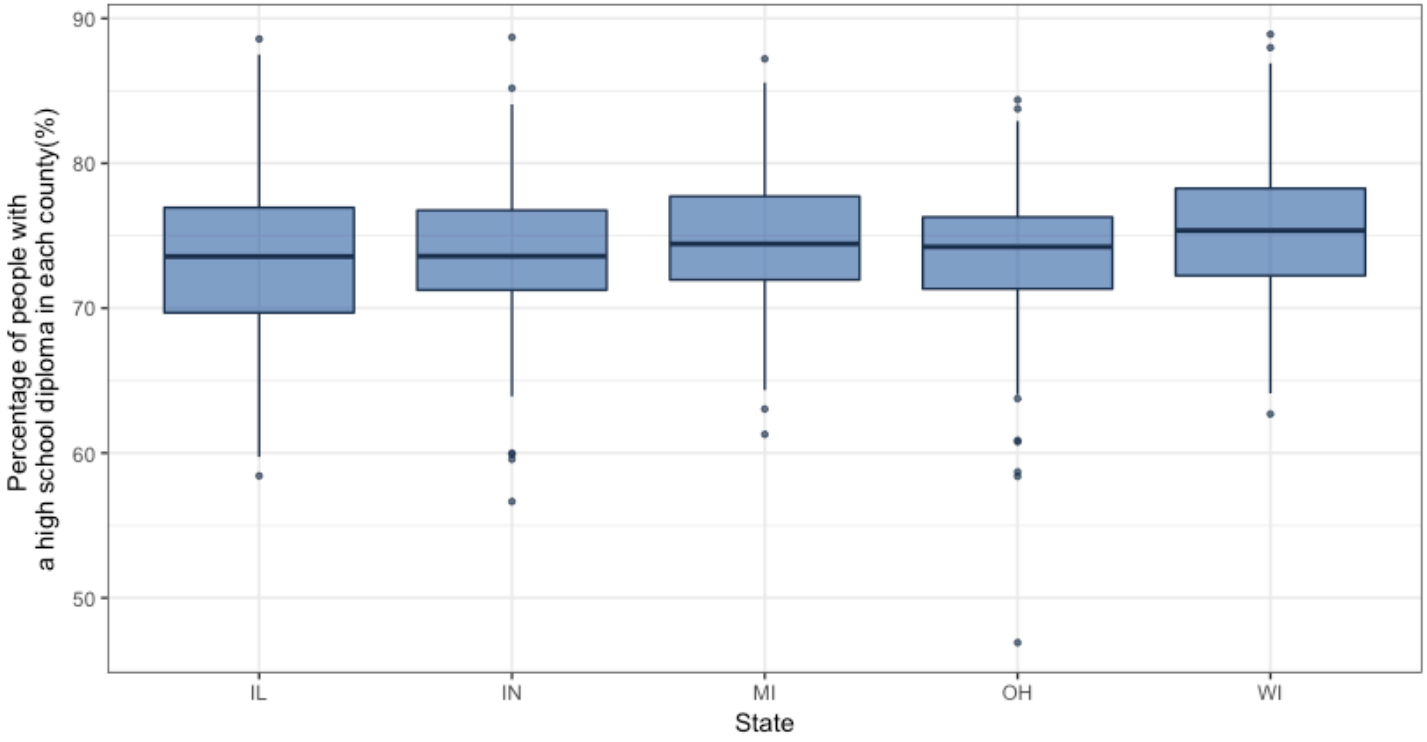
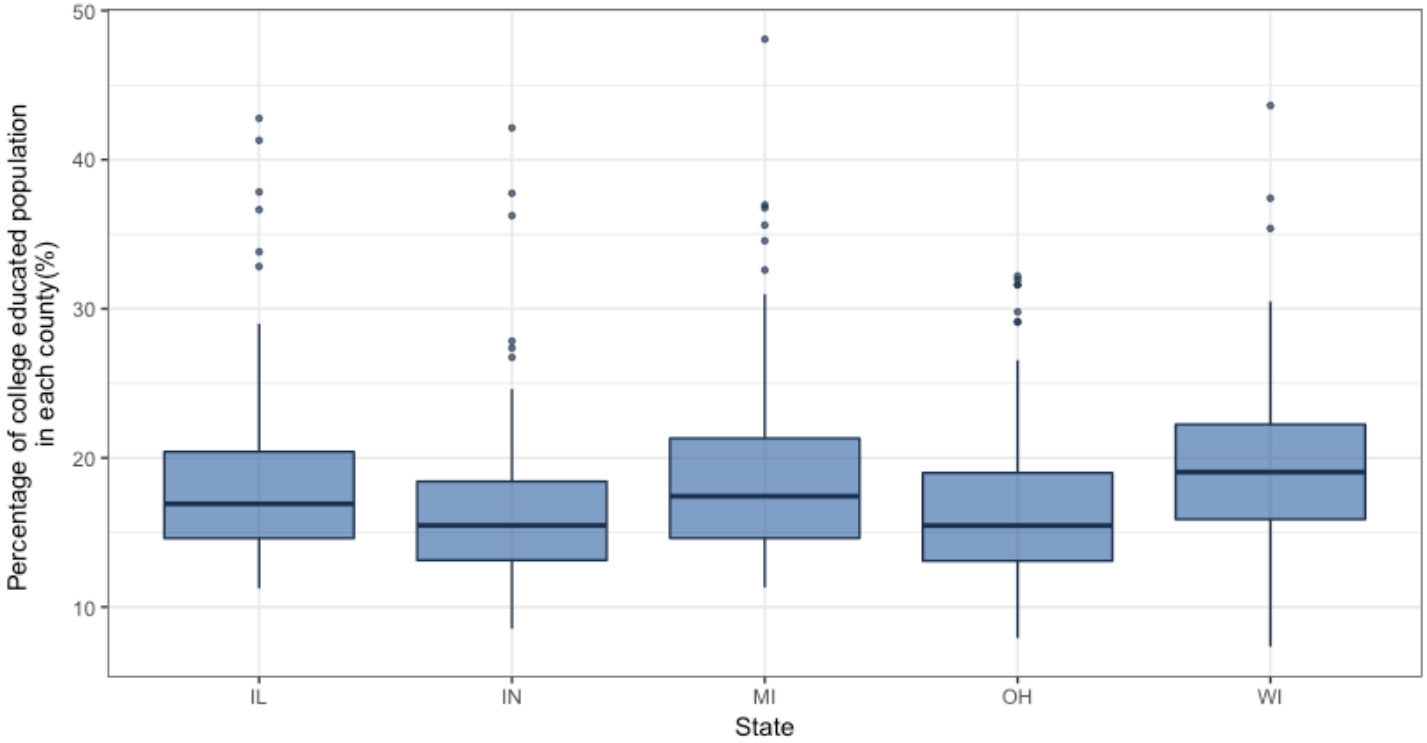


Figure 2.3: Boxplot of percentage of college educated population in each county by state



Question 3. Comparison of Visualization Techniques

Figure 3.1: Boxplot of percentage of college educated population in each county by state

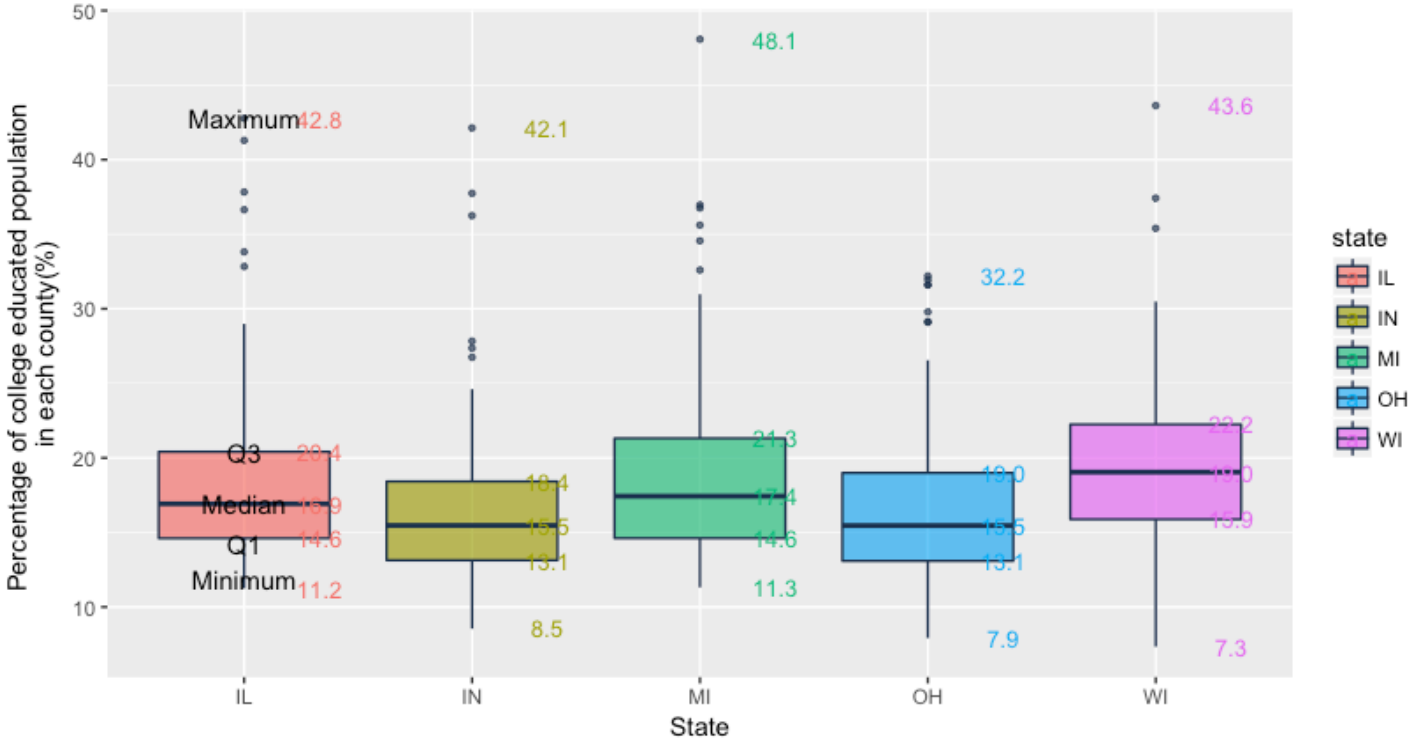


Figure 3.1 is a boxplot of percentage of college educated population in each county by state with labeled elements. We could use state IL as an example to explain how the labeled elements illustrate different statistical properties of a sample:
1) The Maximum value of the sample is 42.8; 2) The third quartile of the sample is 20.4; 3) Median of the sample is 16.9; 4) The first quartile of the sample is 14.6 and 5) The minimum value of the sample is 11.2.

Boxplot describes the distribution of one variable grouped by another variable (Figure 3.1), where as Histogram often exhibits the distribution of one variable and does not take other variables into account, as shown in Figure 3.2.

Q-Q plot is a way to estimate if samples came from some theoretical distribution such as a normal distribution. The line in Figure 3.3 passes the first quartile and the third quartile of the values of the percentage of college educated population in each county, and the hypothesis was the samples come from a normal distribution. As shown in Figure 3.3 the data points fall along this line in the middle of the graph, but curve off in the two ends. Thus we could not expect the dataset truly comes from a normal distribution.

Figure 3.2: Histogram of percentage of college educated population in each county by state

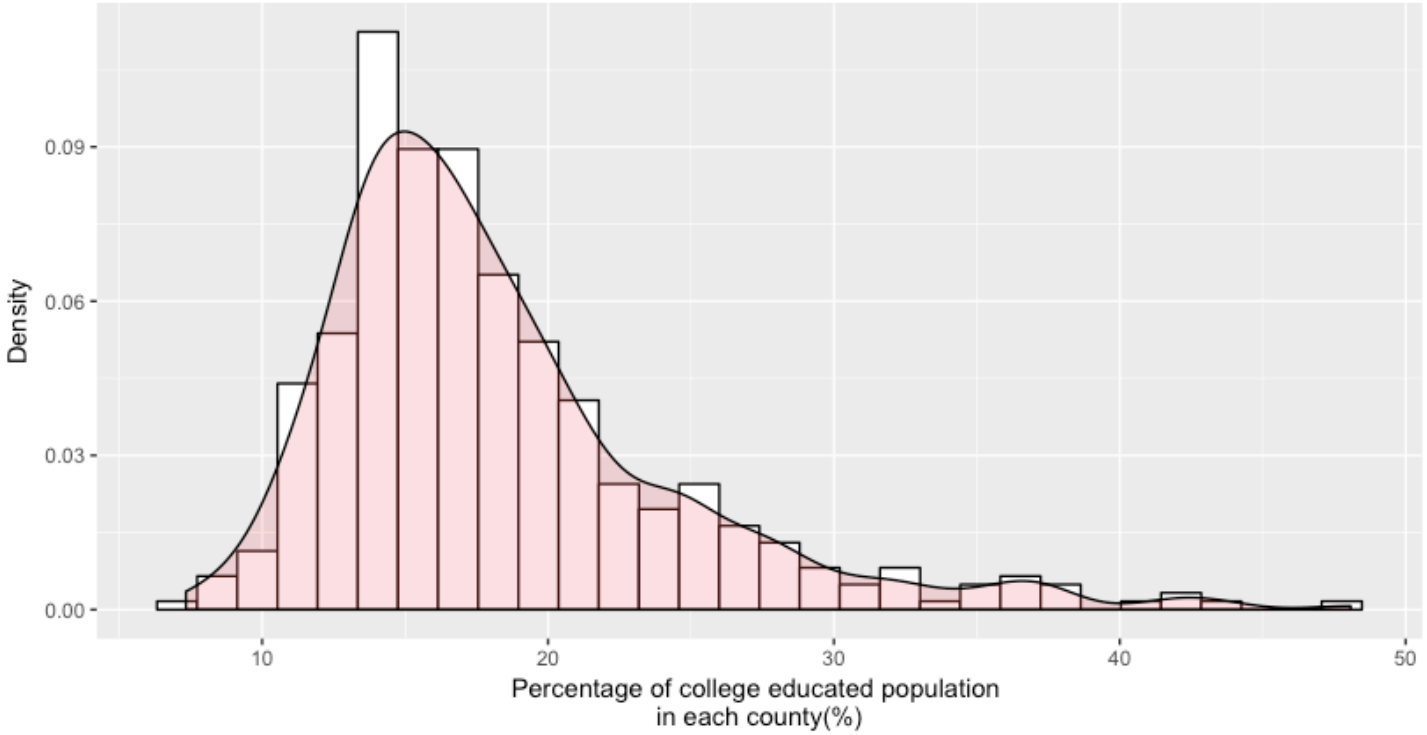
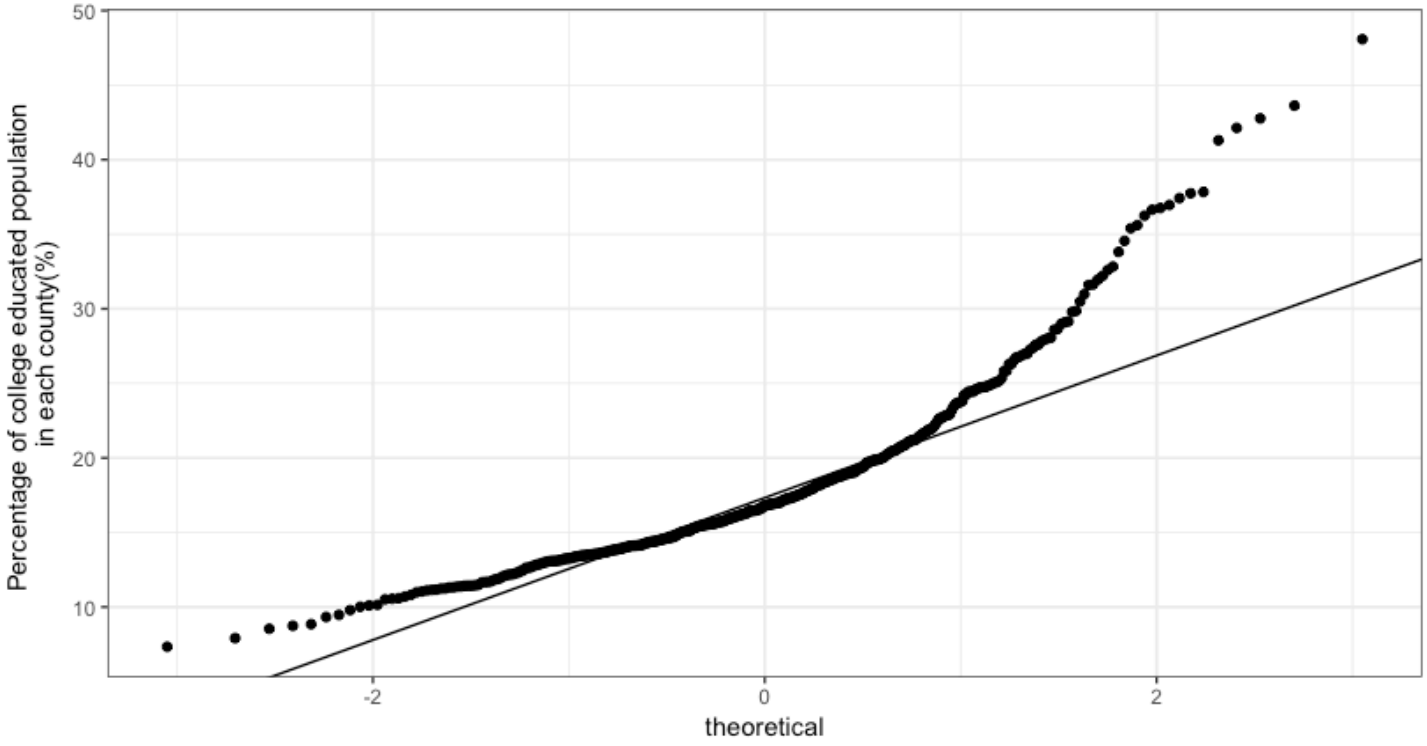


Figure 3.3: QQ plot of percentage of college educated population in each county by state



Question 4. Random Scatterplots

Figure 4.1: Scatter plot of two sets of 10 random uniformly-distributed values

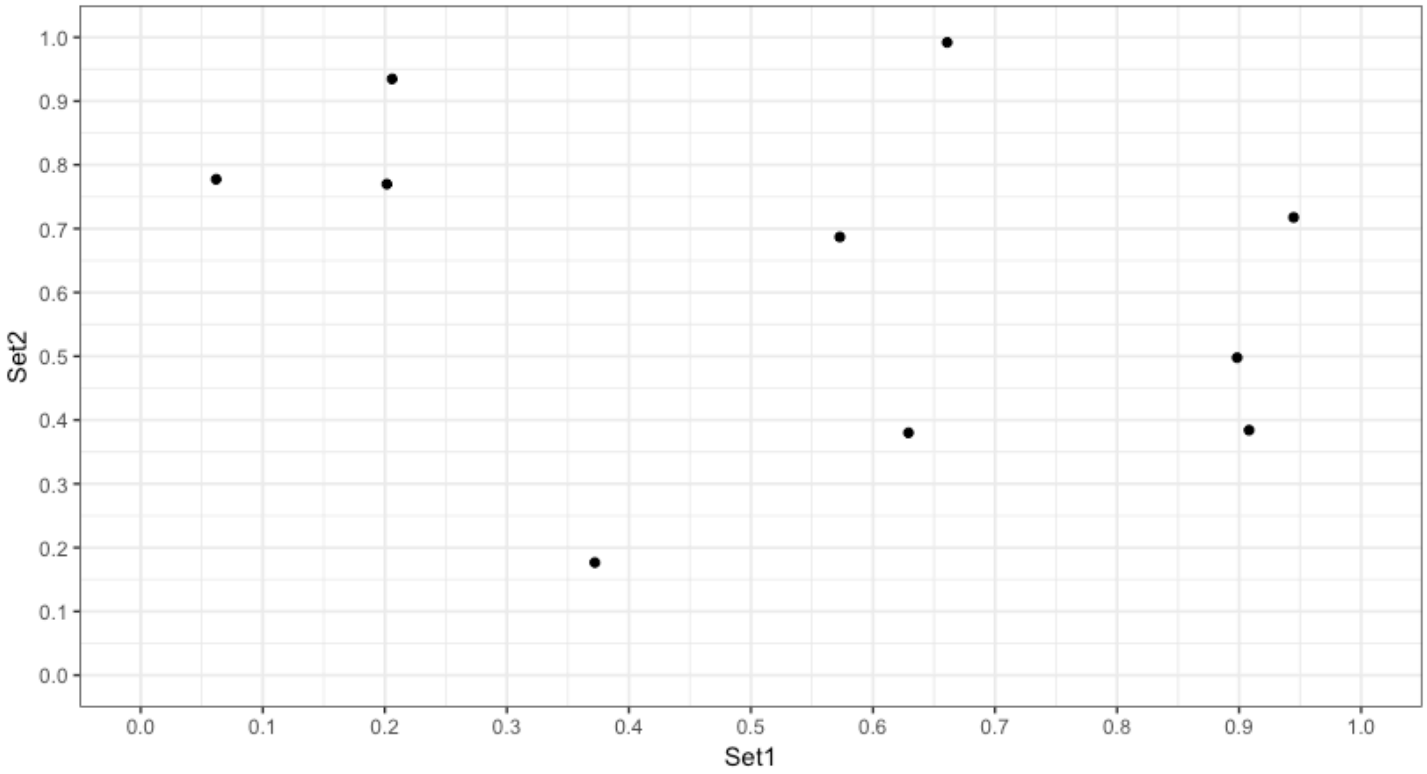


Figure 4.2: Scatter plot of two sets of 1000 random uniformly-distributed values

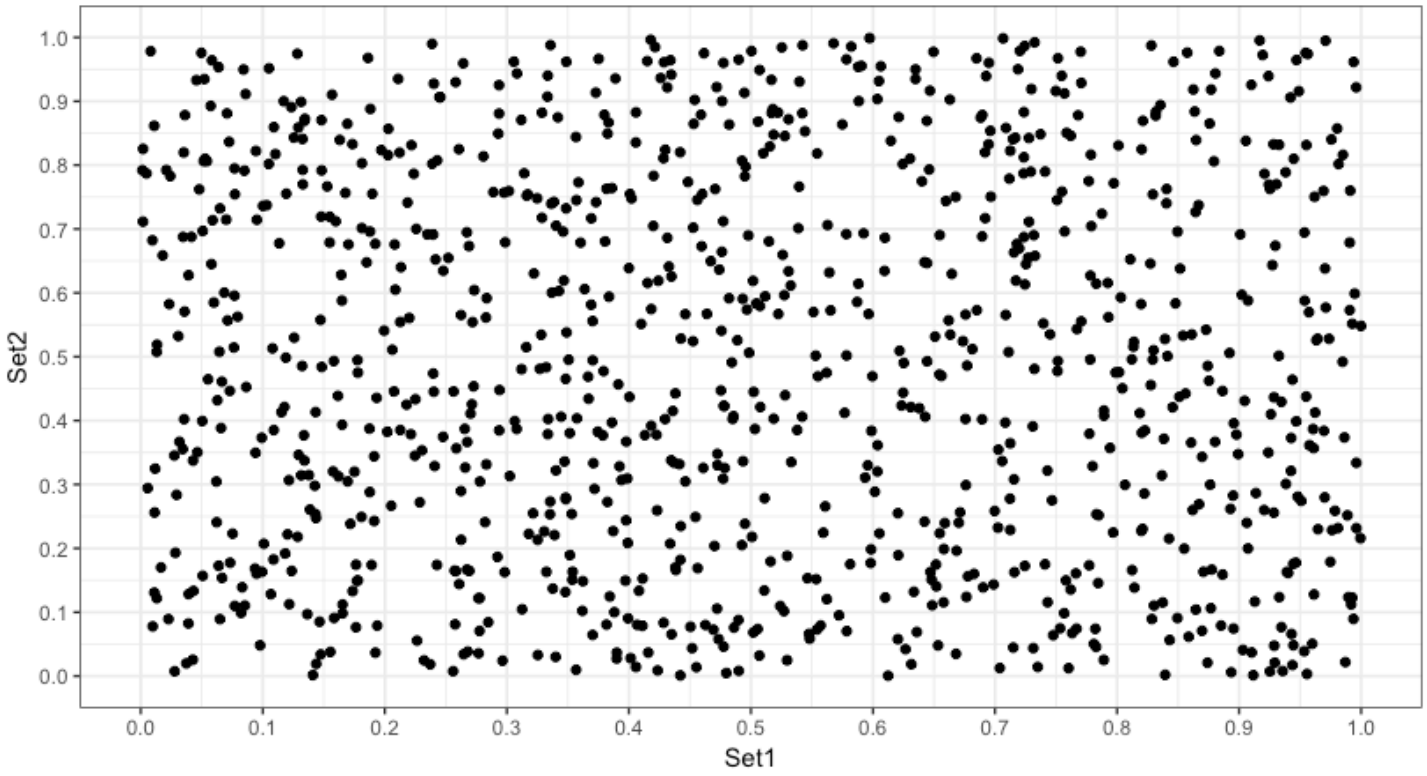
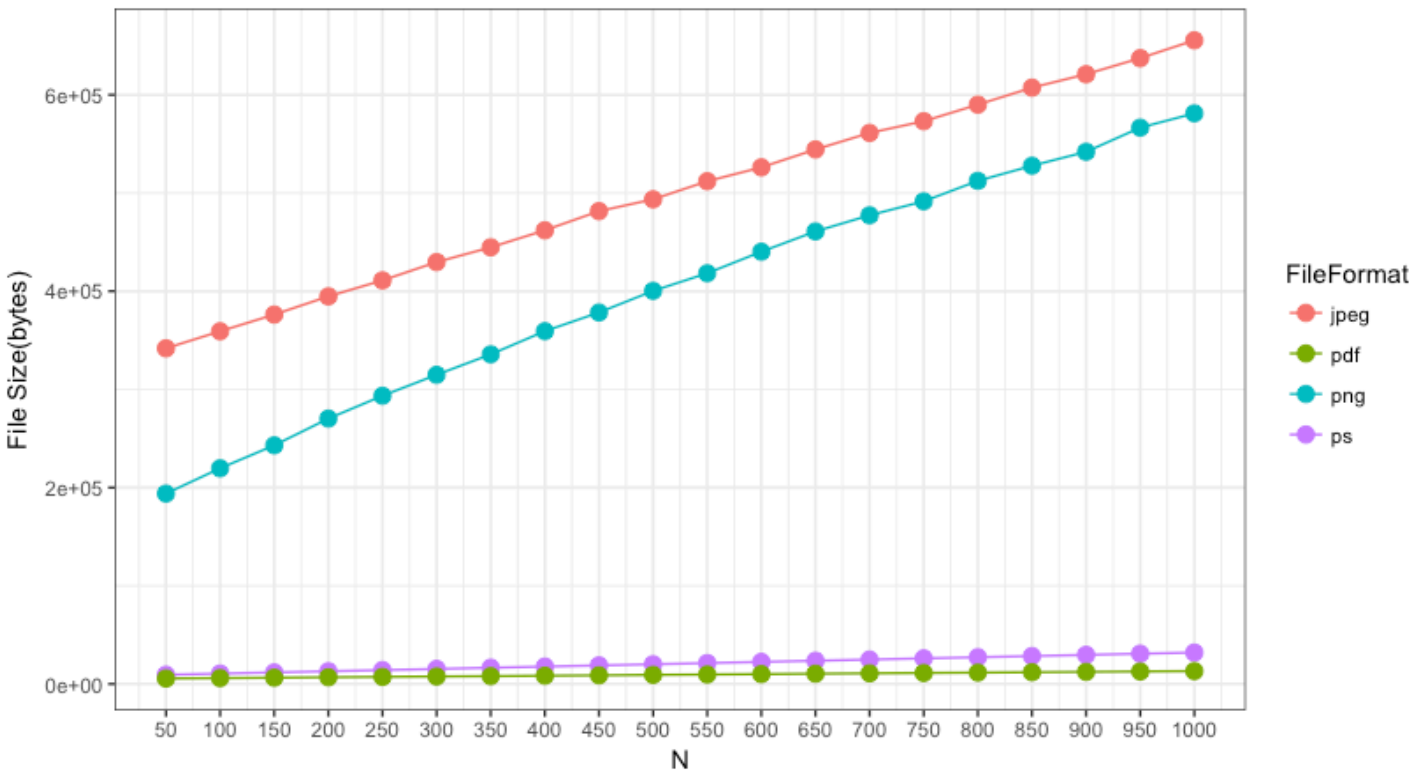


Figure 4.1 and Figure 4.2 show the scatterplots of two sets of N random uniformly-distributed values with N = 10 and N = 1000, respectively. Figure 4.3 reveals the relationship between file size and N for different file formats. The conclusion is if the value of N is fixed, there is a strong correlation between file size and file format, and we could always have the following order about file size:

File size of jpeg format > File size of png format > File size of ps format > File size of pdf format

Figure 4.3: Relationship between file size and N for different file formats



Question 5. Diamonds

Figure 5.1: Number of diamonds in each color

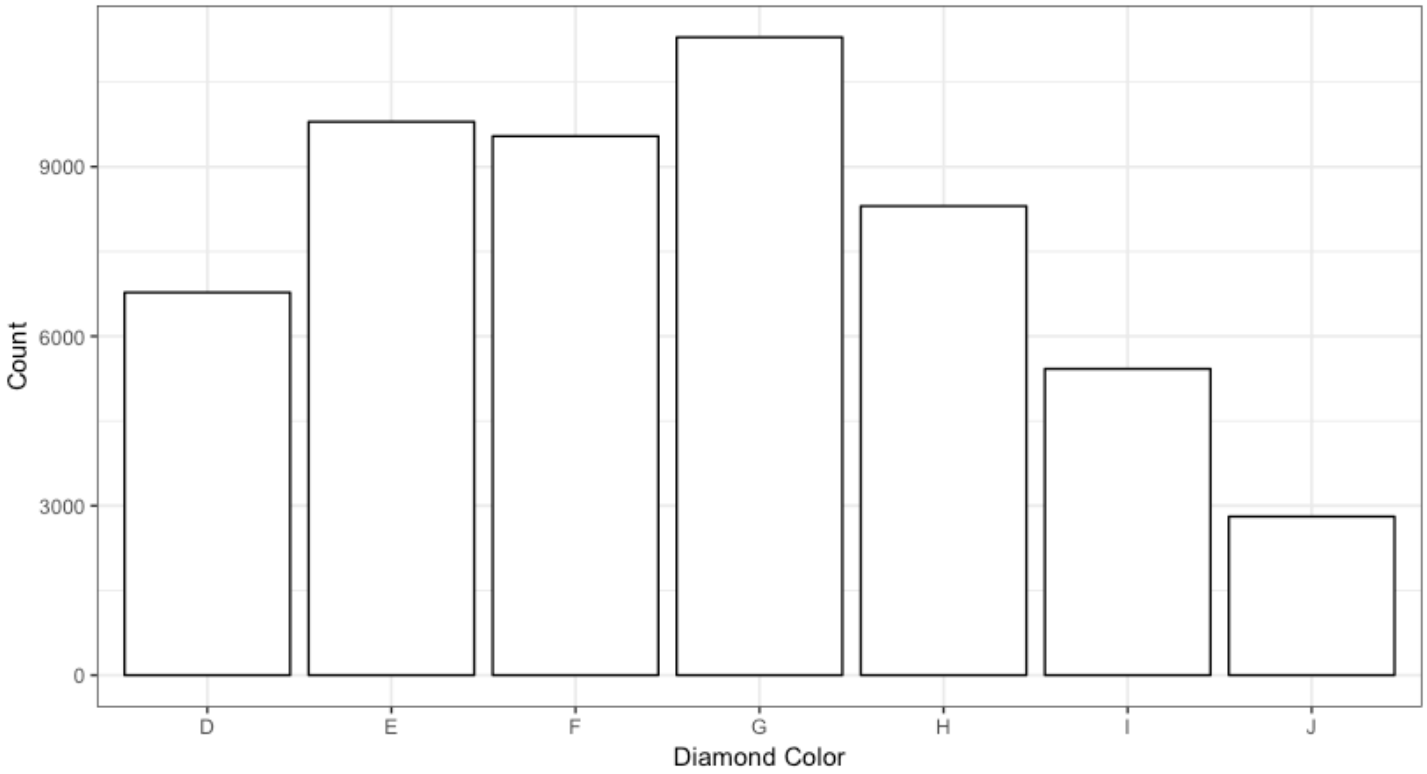


Figure 5.2: Histogram of diamond carat

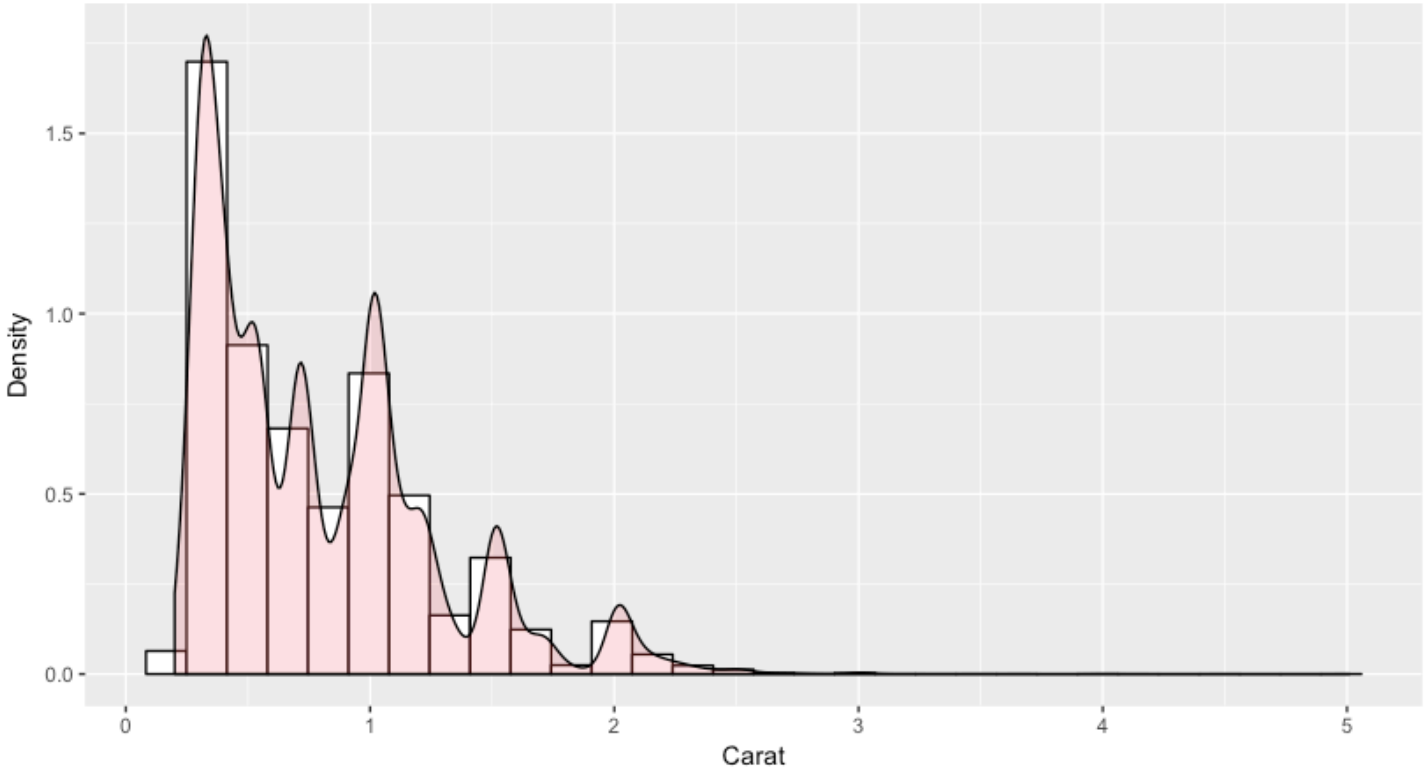


Figure 5.3: Histogram of diamond price

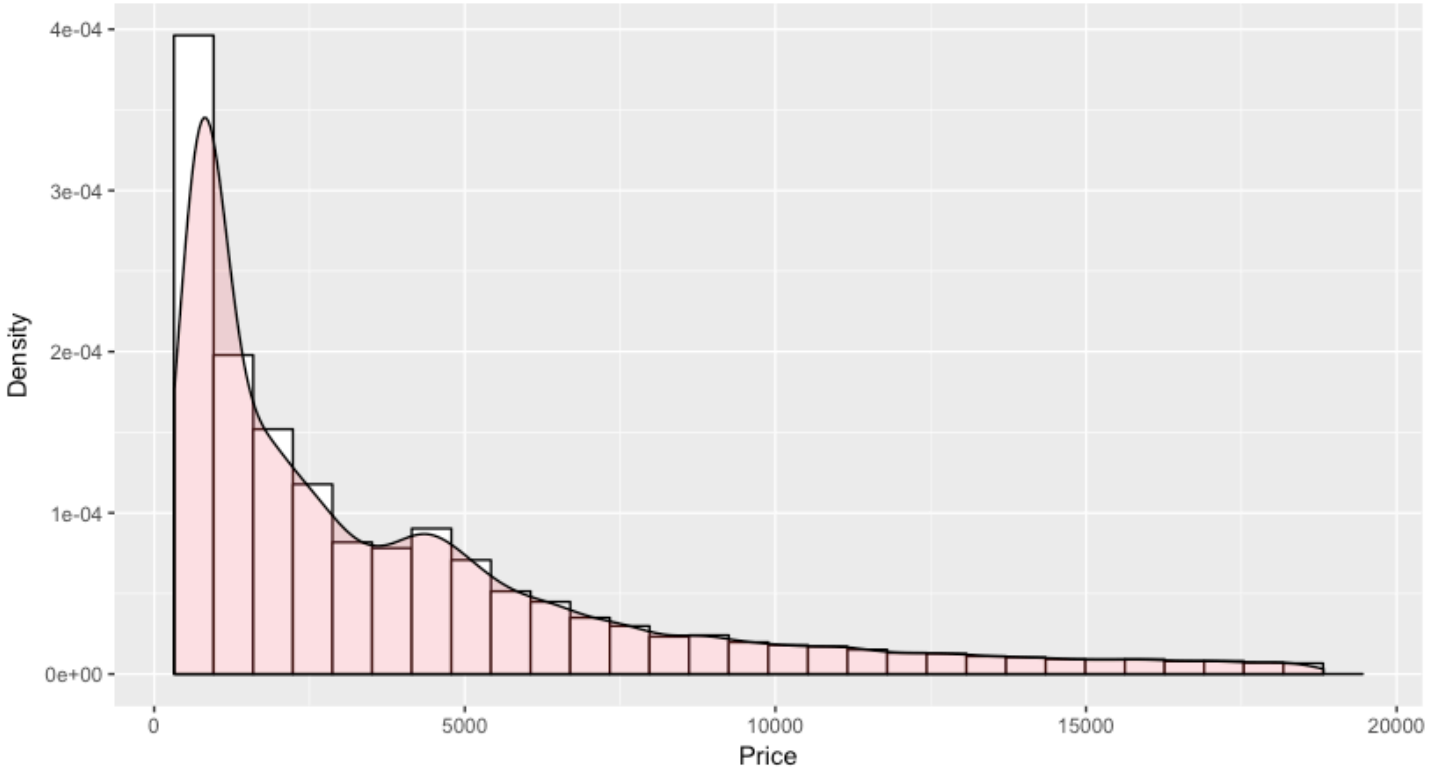


Figure 5.1 is the bar chart of the diamond colors and we could see color “G” has the maximum count of diamonds and color “J” has the minimum count of diamonds. Figure 5.2 and 5.3 describe the distribution of variale “price” and that of variable “carat” using Histogram. The shape of distribution of “price” is single peak whereas that of “carat” is multiple peaks, and both of them are skewed right, especially the distribution of “price” has a long tail on the right of the graph.

Figure 5.4 shows that color “J” has the highest mean of “carat”, and it also has the most variable values of “carat” comparing with other colors. Color “D” has the lowest mean of “carat”.

Figure 5.5 shows that color “J” has the highest mean of “price”, and all colors have the similar maximum value of “price”. Color “E” has the lowest mean of “price”.

Figure 5.6 illustrated the relationship between “carat” and “price”. From Figure 5.6 we could see the coefficient of determination (R^2) is 0.849, which indicates “carat” has a strong linear relationship with “price”. The Pearson Correlation Coefficients between “carat” and “price” is 0.922 and it also reveals there is an extremely positive correlation between “carat” and “price”.

Figure 5.4: Boxplot of carat by color

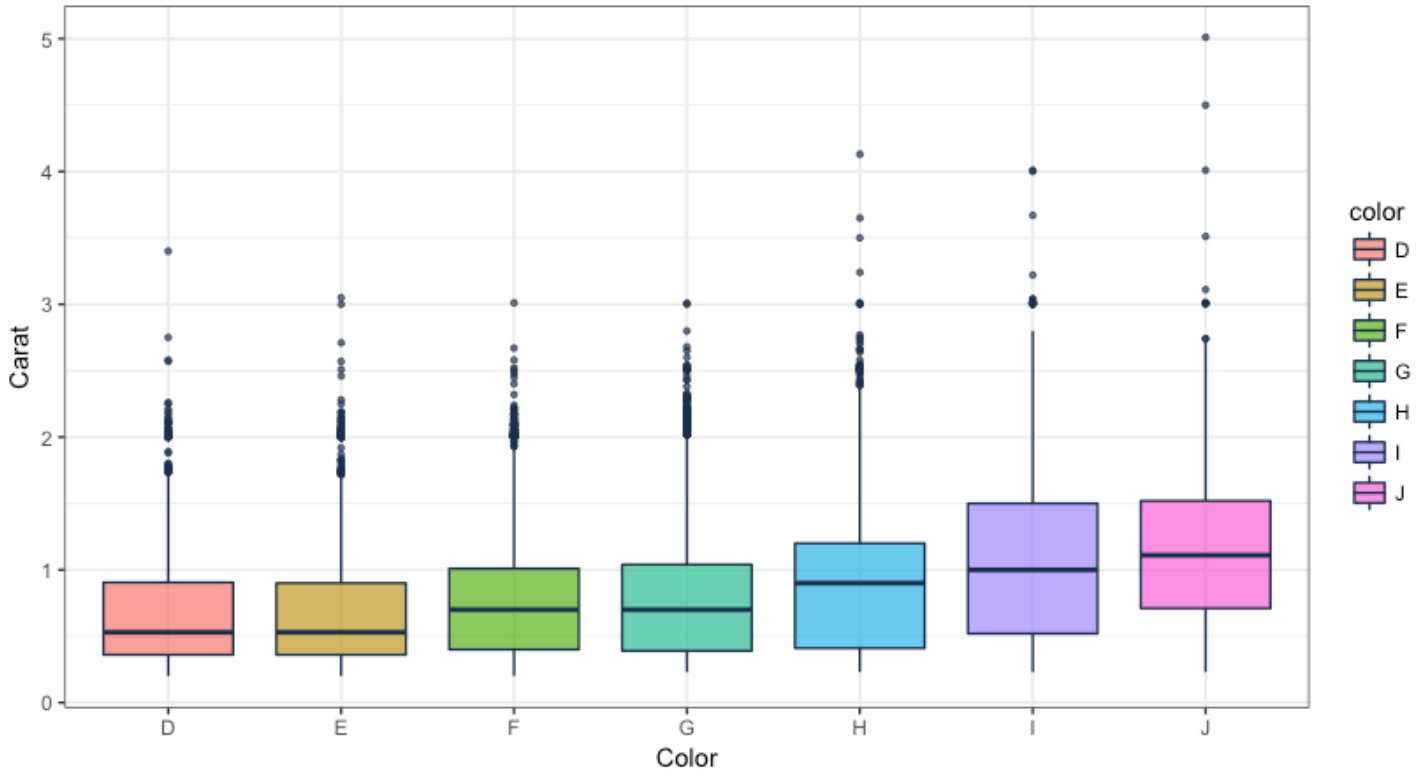


Figure 5.5: Boxplot of price by color

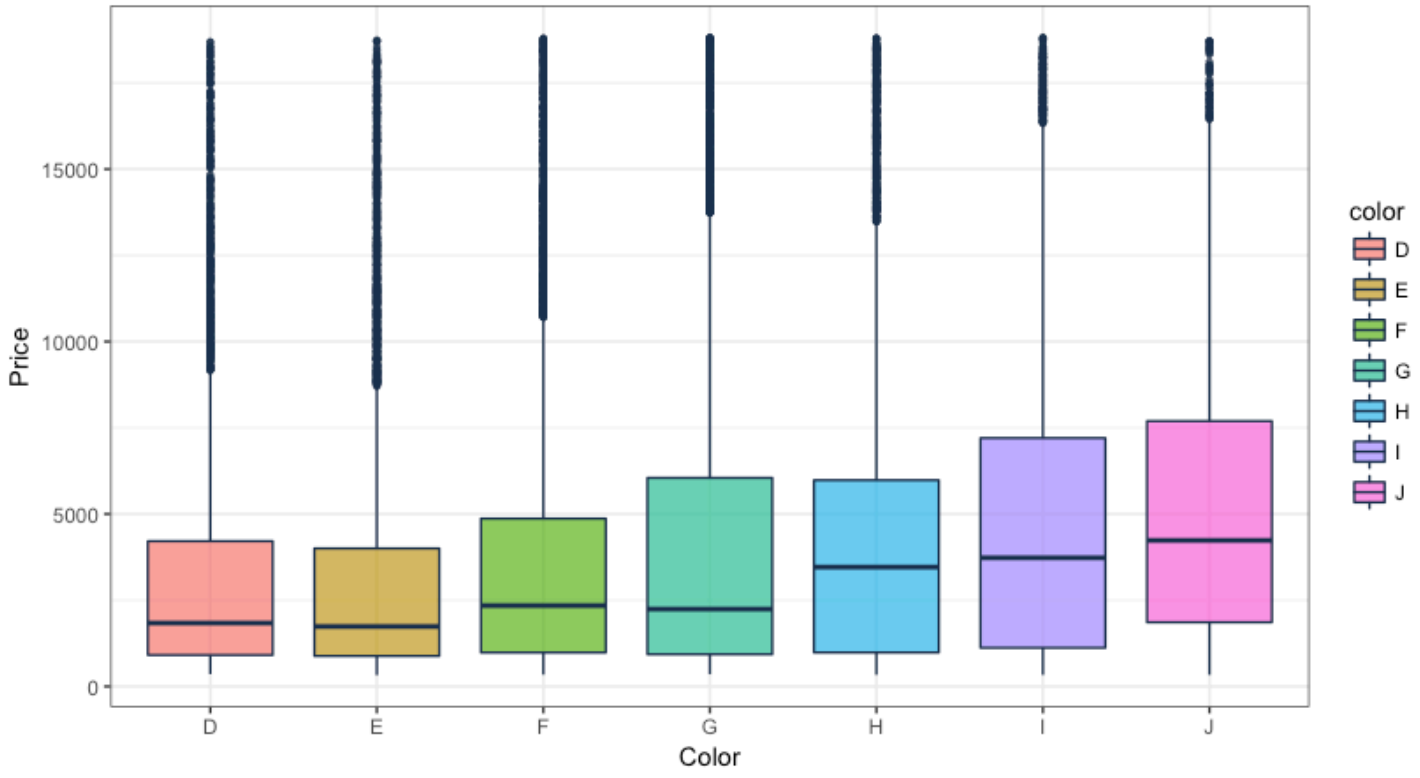


Figure 5.6: Relation between diamonds carat and price