HW2: R Programming

author: JD Plagianis (jplagianis3)

due date: September 24th, 2017, 11:59PM UTC-12

1. Professional Education by State [20 points]:

Visualize the percentage of people that have a professional education (percprof column) for each county, grouped by state, using a suitable type of plot that can help show the properties of the distribution of percprof values found within each state.

Describe the distributions by pointing out at least 2 relevant and interesting properties that your plot illustrates, such as the mean or median value for some states, the spread of values for some states, states that have outlying counties, and so on.

Can you point out which state has the lowest and highest percentage of population with a professional education? If not, explain why. You may interpret this in one of two ways: A) compute a summary statistic for each state (e.g. combined mean) and compare that across states; or, B) directly compare the percprof distributions across states. Support your claim with additional statistics/graph(s) here, if necessary.

Submit:

Report: Visualization of the distribution of percprof values grouped by state; description of at least 2 relevant and interesting properties that can be inferred from the plot; comment on states with lowest and highest percentage (or explanation why you cannot ascertain them), with optional supporting statistics/graph(s).

Observations (for code and additional comments, see attached .Rmd file):

I began this analysis by poking around the dataset and finding corroborating data on the internet to validate that the data was reasonable. Once I was satisfied, I generated a box and whisker plot to visualize the distribution of the percprof values for each county in each state (Figure 1). This highlighted some minimally-insightful information, such as 2-7 percent of the population in every county in the Midwest seems to have a professional education. Obviously if there's no plumber in a county, somebody's going to move in to exploit that market vacuum. This will ensure that there will always be some non-zero percentage of any given profession. The median seems to be consistently in the 3 to 4 percent range across all states, which seems right for any given profession, but this wasn't what caught my attention. What surprised me was the large number of outliers on the high end of the boxplot for every state! This called for a deeper analysis, so I added some factors to the data frame that would allow me to get down into each county's distinct flavor to see what was driving all the high outliers.

As you can see in Figure 2, I've broken the percprof values out by population size and density. Looking at the data this way shows an obvious correlation between population density and the percent of the population with professional educations. The non-metro areas tend to cluster just above and below the median (black line), while the metro and high-density metro areas consistently shift to higher percentages above the median. This makes a great deal of sense to me, as demand for services increases in urban environments. For example, areas with more kitchens per square mile will demand more plumbers per square mile. The point that surprised me the most was Washtenaw, MI as it completely eclipsed all other counties with a percprof over 20 percent! Upon examining a map; however, I suspect its proximity to Detroit (a national manufacturing hub) and its abundance of institutions of higher education simply created a culture where all types of education are valued. This is backed up by the similarly outstanding percollege value for the same county (48%).

As to which states had the highest and lowest percentage of professional educations, this question could not be directly answered by aggregating the per-county percprof and comparing, because as we saw in Figure 2, there are several counties where the huge populations would completely skew the result. Instead, I created some additional fields in the data to aggregate the ***number*** of people per county with professional education. With those values, I was able to aggregate across each state to produce the pareto chart show in Figure 3. This clearly shows that Illinois, buoyed by the 5 million people in Chicago, has the greatest percentage of people with professional educations, with Wisconsin coming in last.

Sample Output:

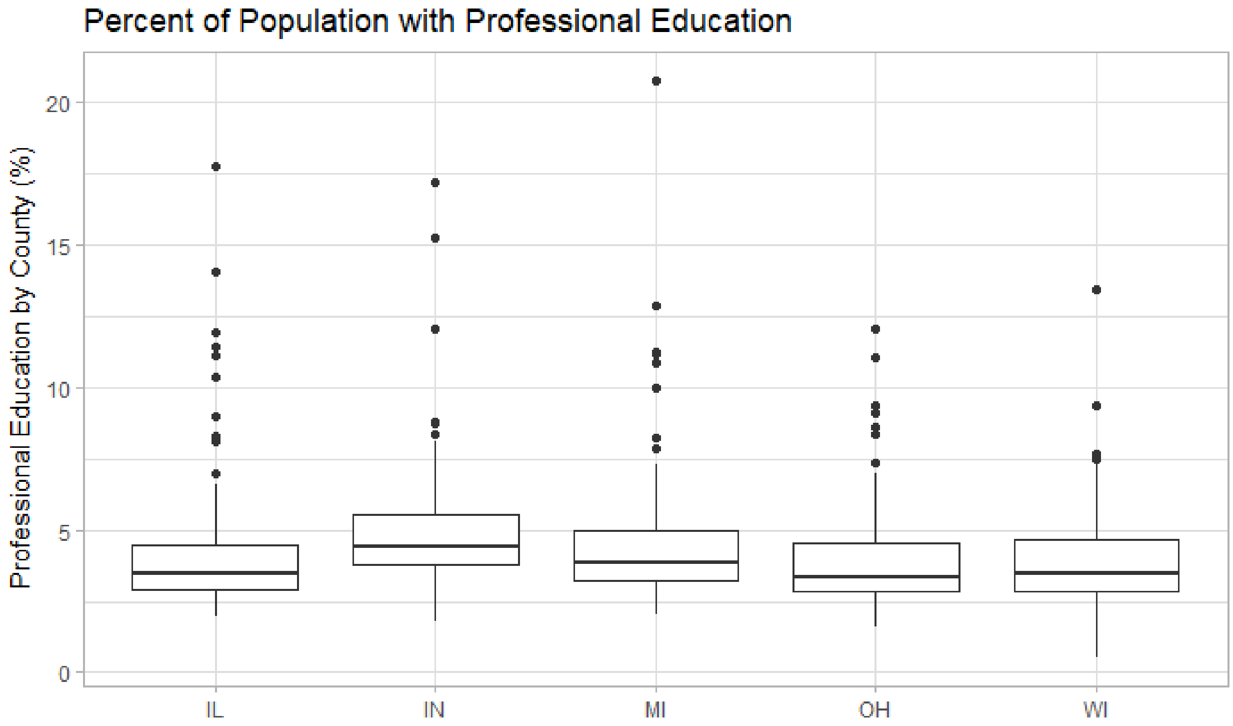


Figure 1: A box and whisker plot displaying the distribution of percprof values for each county in each state

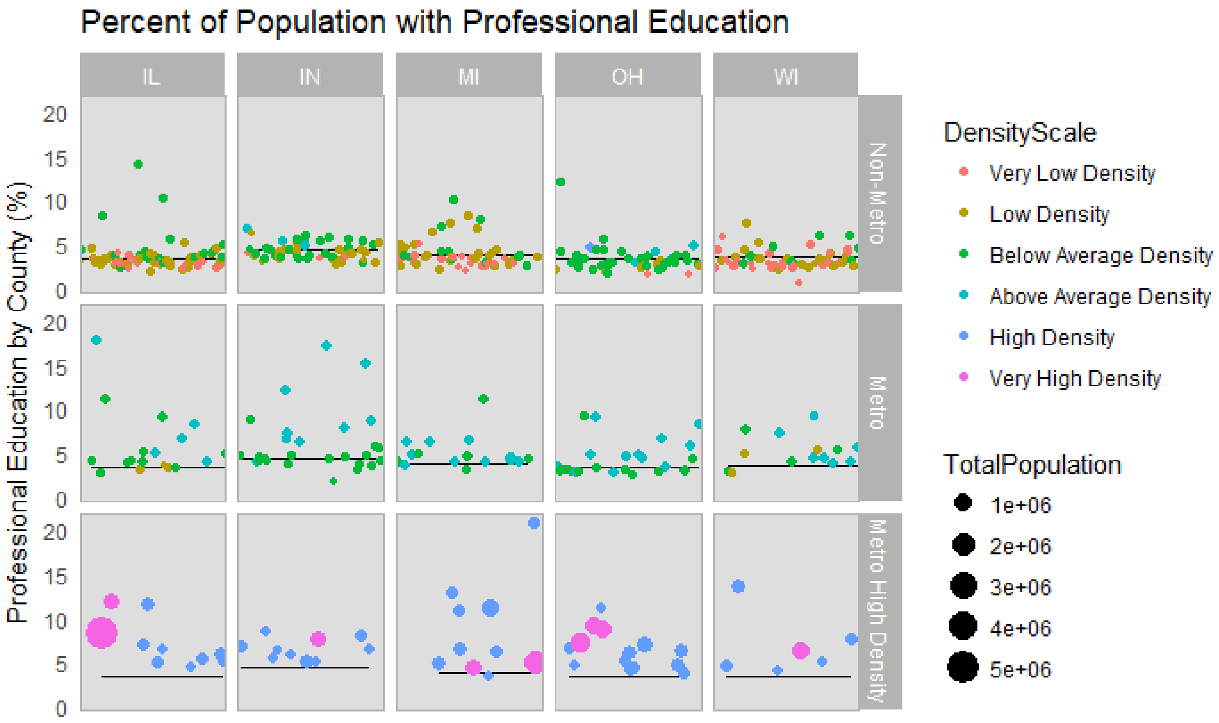


Figure 2: A set of faceted scatterplots displaying percprof as a function of county size/density

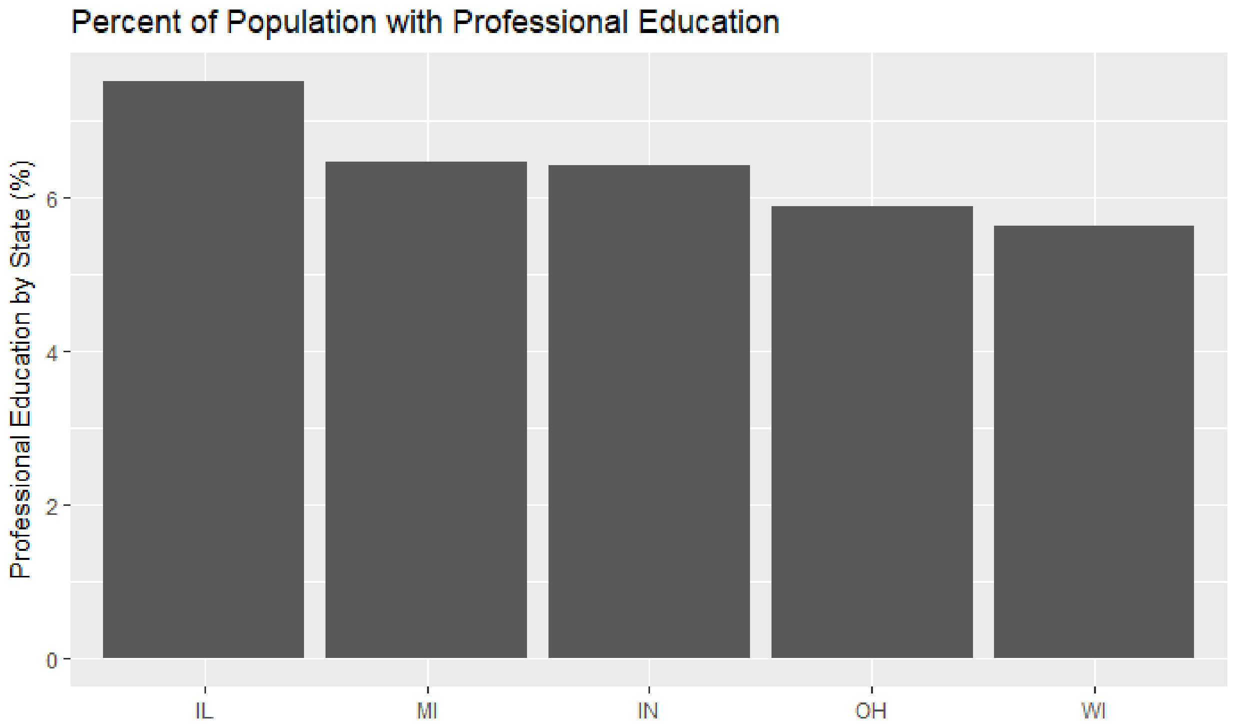


Figure 3: A pareto chart displaying percprof at a state level

2. School and College Education by State [20 points]:

Using the midwest dataset:

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 (e.g. using ggpairs).

For at least 2 of the 3 pairs, describe the relationship you observe (if inconclusive, explain why). You may compute and present additional statistics to support your claim (e.g. correlation), if necessary.

Submit:

Report: Graph(s) illustrating the relationships between perchsd, percollege and state; at least 2 observations that can be inferred from the graph(s), with optional supporting statistics.

Observations (for code and additional comments, see attached .Rmd file):

This analysis was pretty straightforward. After generating a ggpairs (Figure 4) plot, my eyes were immediately drawn to the scatter plot between perchsd and percollege. I generated a scatter plot with a best fit line (Figure 5) to highlight the positive relationship between college graduation and highschool graduation, but there was something more that popped out of this chart. It seemed almost like there was a threshold (~70% highschool graduation) below which the relationship with college graduation seemed muted, and above which seemed much stronger. This lead me to speculate that a certain level of educational attainment among the population of a county can create a feedback loop that either encourages or discourages further educational attainment. To test this hypothesis, I added a factor to the data that would compare each county's highschool graduation rate against the average for that county's state. I then plotted these by state against college graduation rates in a faceted box plot to generate Figure 6. This is a very exciting plot, because it clearly shows that counties with higher highschool graduation rates also had higher college graduation rates. It shows that regardless of which state one lives in, one could predict the likelihood of any given student graduating college just by looking at whether that county invests in creating a culture that produces higher highshcool graduation rates.

Sample Output:

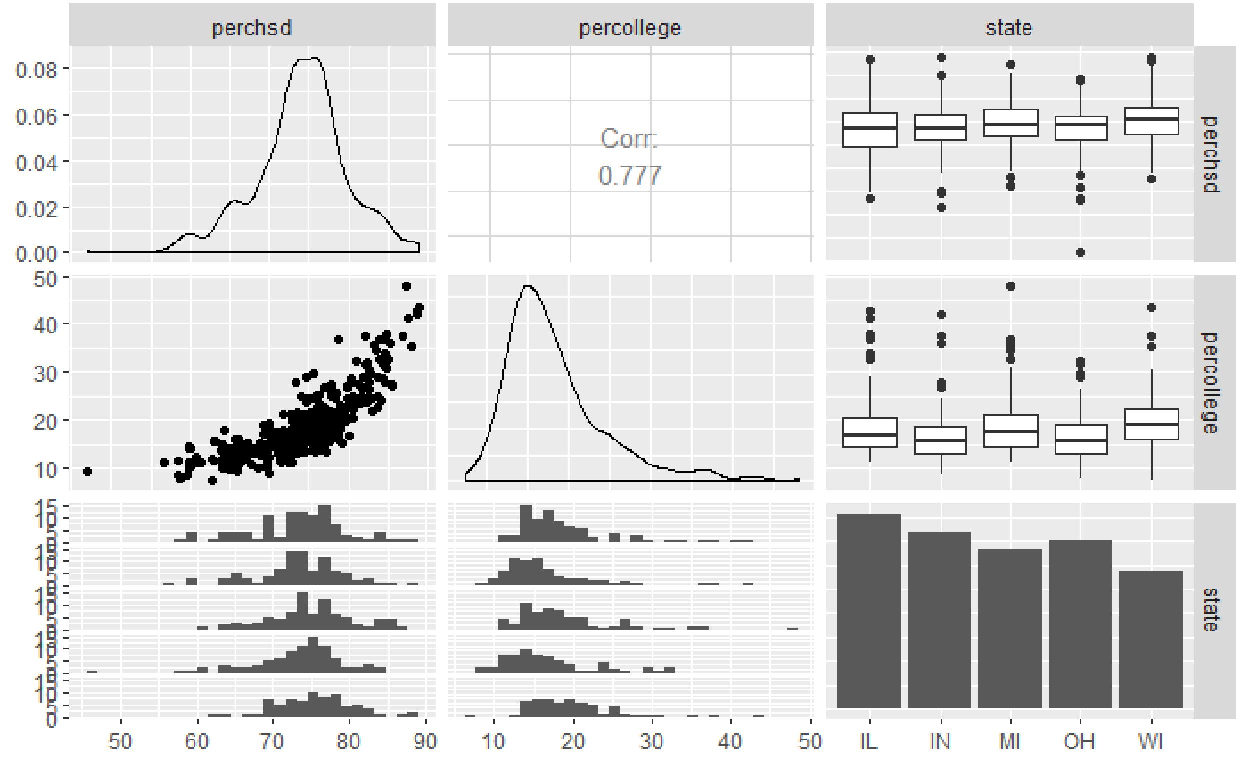


Figure 4: A ggpairs visualization of perchsd, percollege, and state

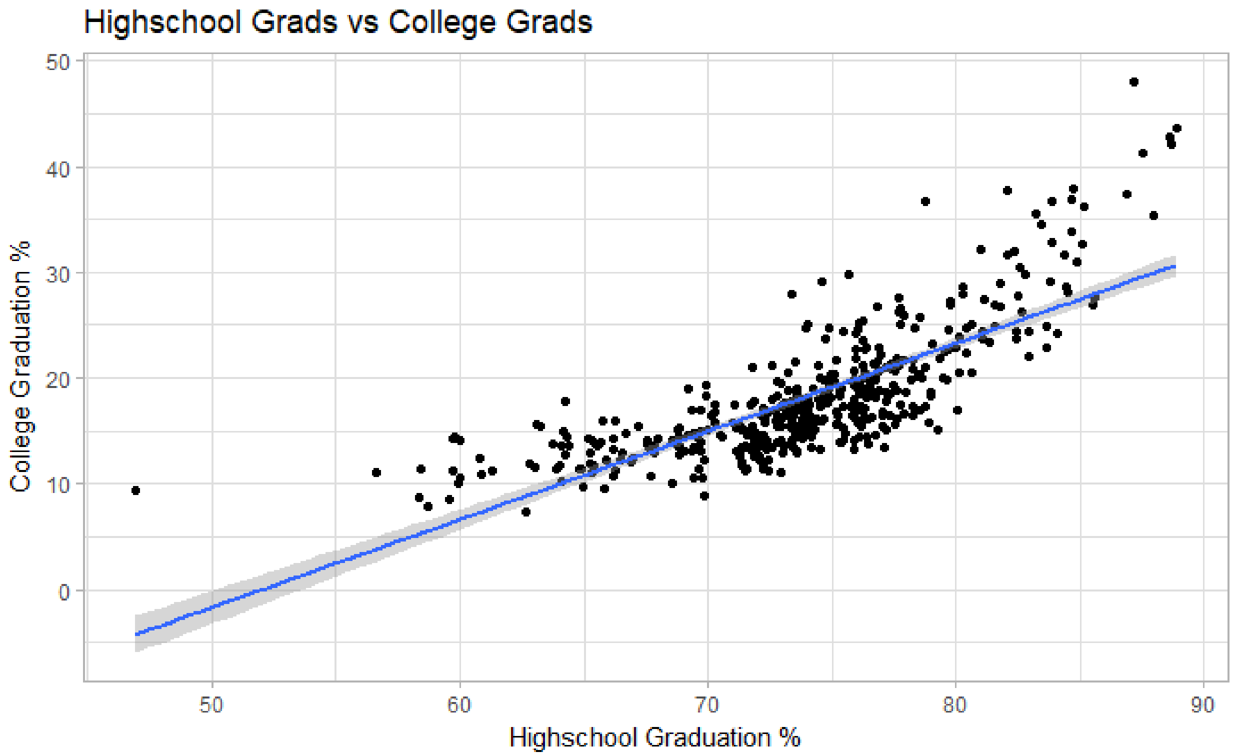


Figure 5: A scatterplot of graduation rates (college vs highschool) showing a best fit line with a positive slope

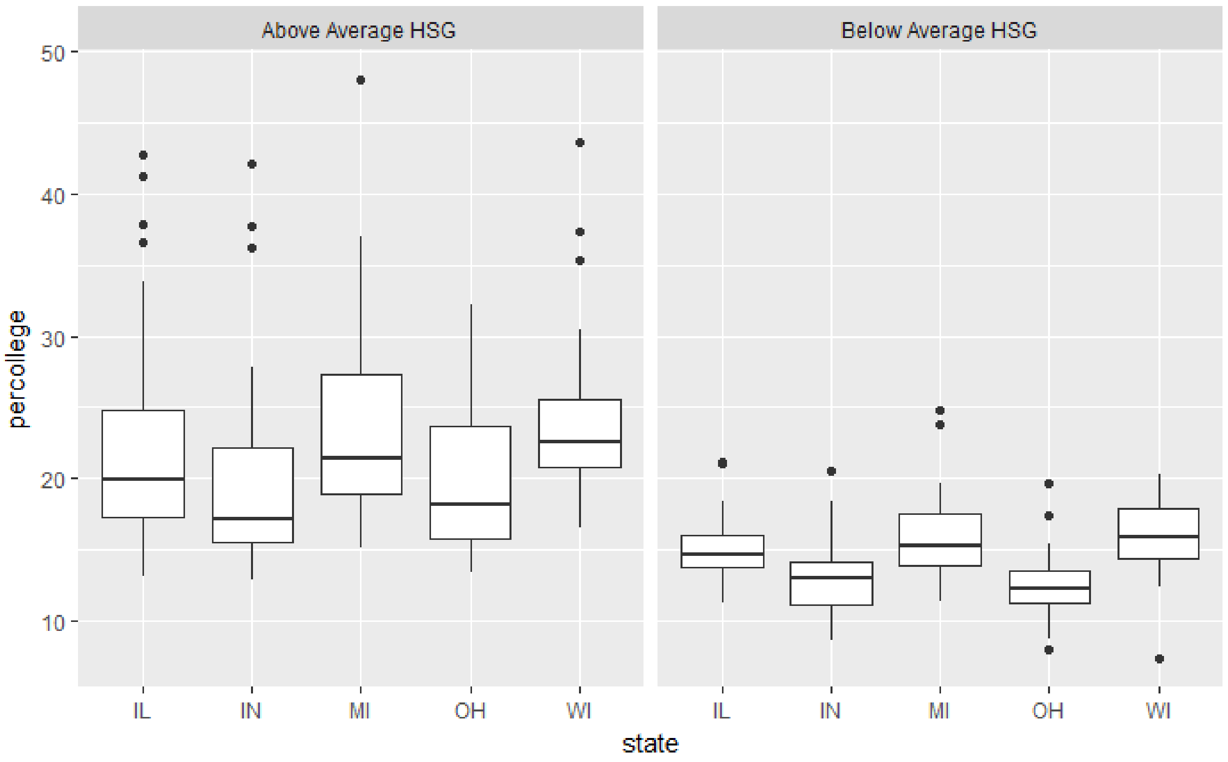


Figure 6: A comparison of the likelihood of a student graduating college if they lived in a county with an above-average highschool graduation (HSG) rate versus a below-average highschool graduation rate

3. Comparison of Visualization Techniques [20 points]:

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).

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.

Submit:

Report: Labeled Box Plot diagram with description of different elements, and how they illustrate sample statistics; a comparison of when you would use each of the following types of plot, with an example each: Histogram, Box Plot, QQPlot.

Observations (for code and additional comments, see attached .Rmd file):

For this activity, I generated a random normal distribution centered around 100 with a standard deviation of 25. I sampled 80% of this distribution and then added some large numbers to skew the dataset away from the theoretical normal distribution.

Figure 7 addresses the first requirement of this activity, by labelling the features of the boxplot. First, the thick horizontal line across the middle of the box is the median of the entire dataset. This represents the point at which 50% of all values are above the line and 50% are below. Next, the 25% Quartile represents the median of the bottom 50% of values (everything below the formerly described median). Same goes for the 75% Quartile, which represents the median of the upper 50% off the values in the dataset. These quartiles represent the top and bottom of the “box” in the boxplot, which is also called the InterQuartile Range (IQR). The bottom whisker represents the lower 25% of the data (everything below the 25% Quartile). The top whisker represents the upper 25% of the data (everything above the 75% Quartile). Because I added some extra numbers to the dataset with much larger values than the normal distribution would typically show, we get to see 5 points that sit beyond the upper whisker. Typically, in a normal distribution, all points will fall within 1.5x the range of the IQR. This arbitrary 1.5 comes from a guy named Tukey, who invented the box plot. The range works well, so it has stuck as a best practice. Points that lie beyond (above or below) the 1.5x range are called outliers. Points that lie significantly beyond the 1.5x range are called extremes and are just a subset of outliers.

The boxplot doesn’t care if the data is normally distributed. Its value comes in getting a good feel for the spread of the middle 50% of the data. It’s also useful for spotting differences in the center and spread of the data when comparing multiple datasets (Figure 6 from problem 2 is a good example of this).

The histogram shows how data is distributed. It’s extremely useful for spotting modes in multi-modal data. Figure 8 shows the histogram for the same dataset used in Figure 7. You can clearly see the normal distribution centered around 100, along with the large numbers that sit to the right. Had I added five or ten more large numbers clustered around 200, a multi-modal distribution would have been very apparent using the histogram, while the boxplot would only have shifted its median line up. The histogram is highly sensitive to bucket size. Over- or under- sizing the buckets can obscure local maxima.

The Q-Q (quantile-quantile) plot is designed to help understand if the data is normally distributed. It plots the values in your dataset against values from a standard normal distribution. If your data is normal, this should create a line with a roughly constant slope. Figure 9, which is created off the same dataset as Figure 7 and 8, clearly shows that the data is heavy on the upper tail, skewing the data away from normal. Like most charts, the assessment is subjective in practice, but is a useful tool when getting to know the data.

Sample Output:

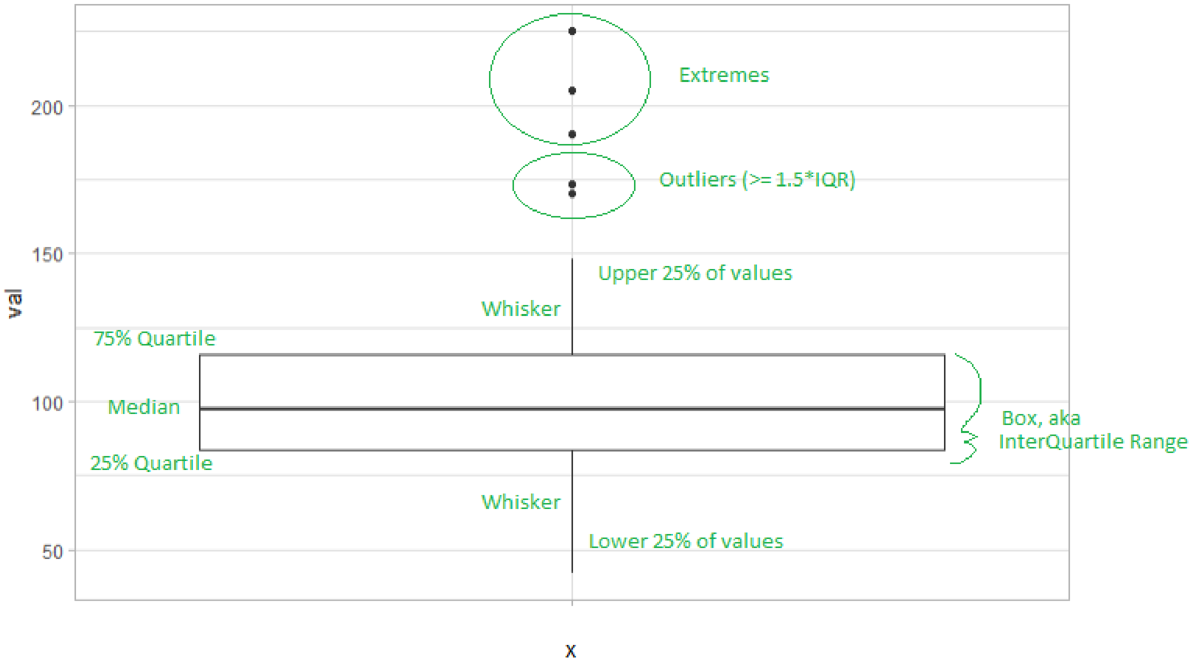


Figure 7: A labelled box and whisker plot with outliers

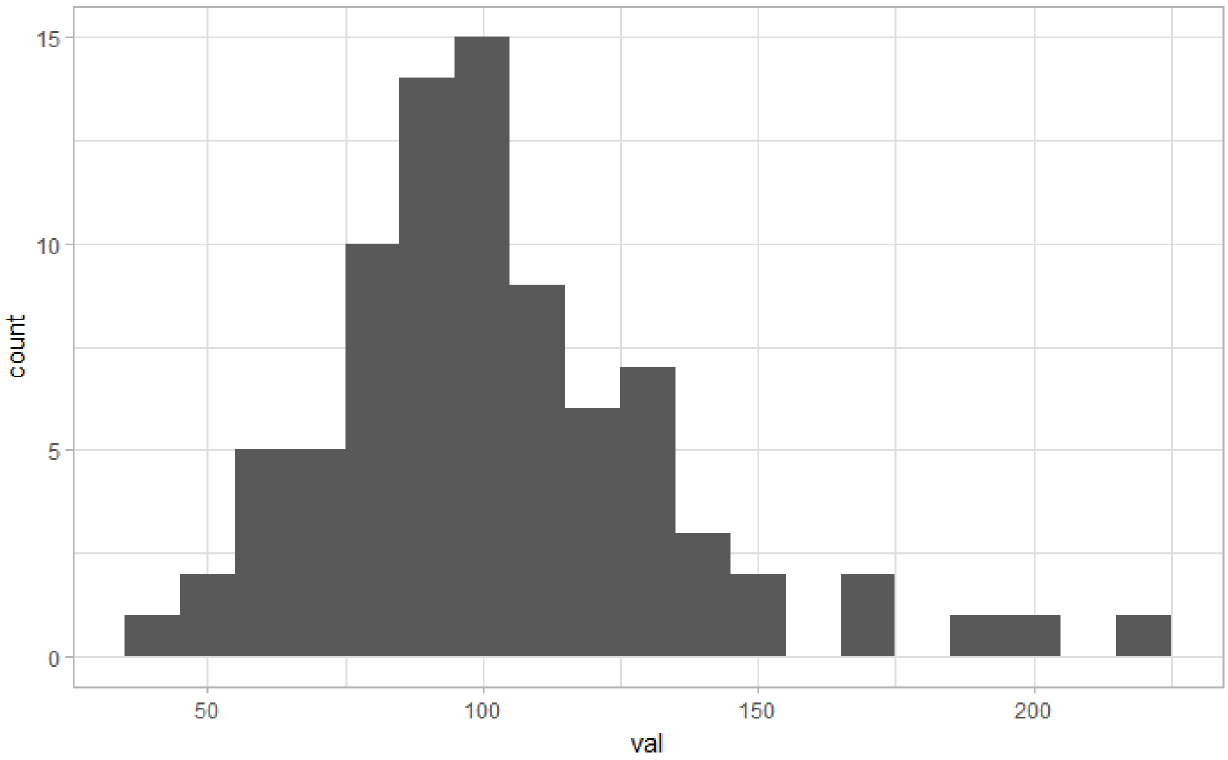


Figure 8: Histogram with a single mode centered around 100

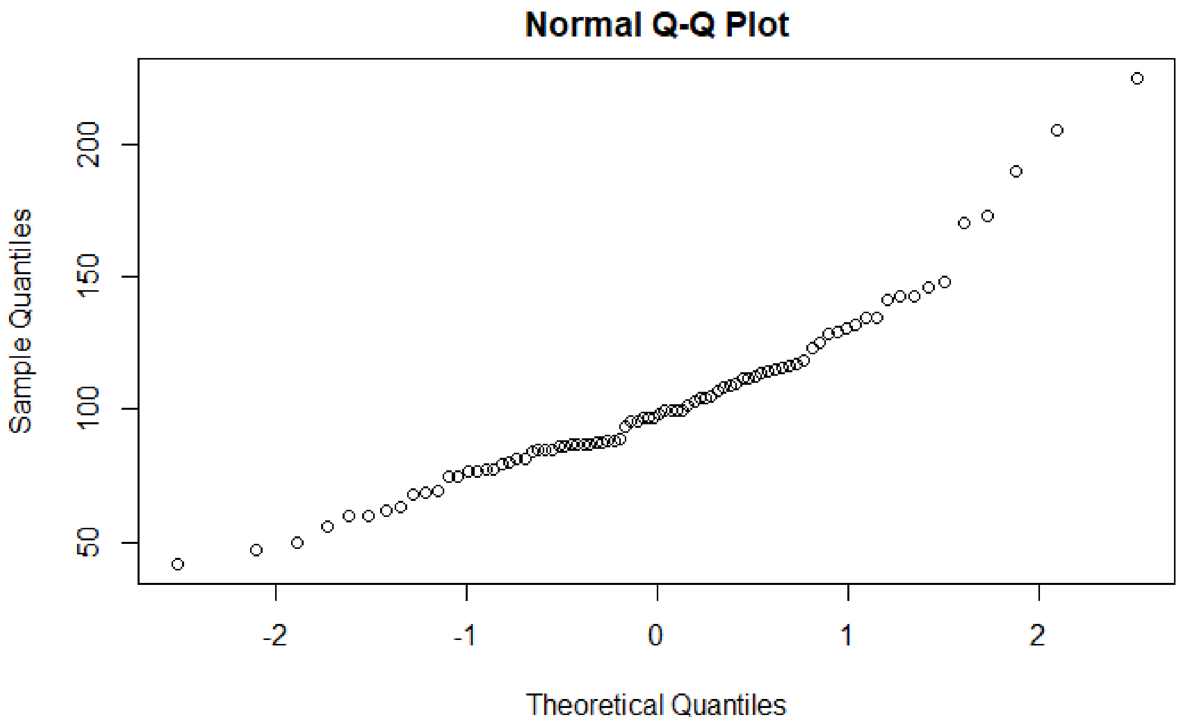


Figure 9: Q-Q plot showing a normal distribution with a suspicious heavy tail on the top end of the data

4. Random Scatterplots [20 points]:

Generate two sets of N random uniformly-distributed values using the function runif() (read "random-uniform" not "run-if"), and display a corresponding scatterplot using one set as X-values and the other as Y-values.

If you save the plot to disk, what is the resulting file size for the following file formats: ps, pdf, jpeg, png? How do the file sizes scale with increasing N? Display your results by plotting file size vs. N, over a suitable range of N, for each format. Your visualization must support your observations.

Note: To save a plot to disk, you may need to open a graphics device before plotting (for base graphics) or use ggsave() (for ggplots). If you have trouble saving the plots in any of these formats, you may pick some alternatives that are available on your system (such as tiff), but ensure that you plot file size vs. N for at least 3 different formats.

Submit:

Report: One or two sample scatterplots using randomly generated values (e.g. one with low N, one with high N); a suitable plot illustrating the relationship between file size and N for each of the chosen file formats (at least 3 different formats); your observations on these relationships (asymptotic space complexity is not compulsory, however, try to be as precise in your description of the relationships as evident from your plot).

Observations (for code and additional comments, see attached .Rmd file):

This was my favorite problem in this homework because I was completely caught off guard by the outcome. As the entire chart becomes covered by black dots, the JPEG and PNG stop increasing in size and actually start decreasing in filesize! My guess is that these two file compression algorithms begin to represent the entire area as a single black shape instead of n overlapping black shapes like the PDF and PS algorithms do. It appears that the PS and PDF files, while both extremely efficient at storing very small numbers of points, will continue to grow linearly as the number of points increases, while the JPEG and PNG files will level off and stop growing. Specifically, it appears that the filesize for the JPEG and PNG is influenced primarily by the saturation of the plot area. Figure 10 shows the relationship between filesize and number of points on the scatterplots. Figures 11 and 12 show the scatterplots at 500 and 500,000 points, respectively on 10 cm x 10 cm plots.

Sample Output:

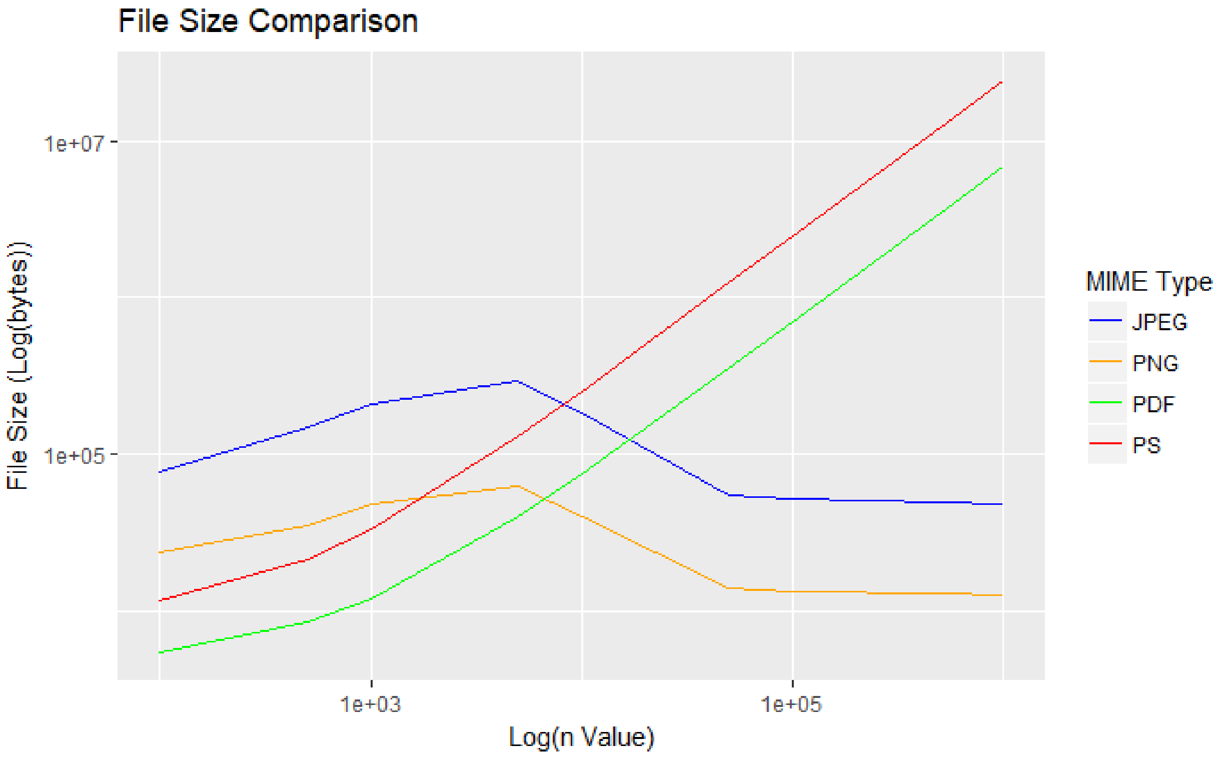


Figure 10: File size versus number of points on a 10 cm x 10 cm image of a scatterplot

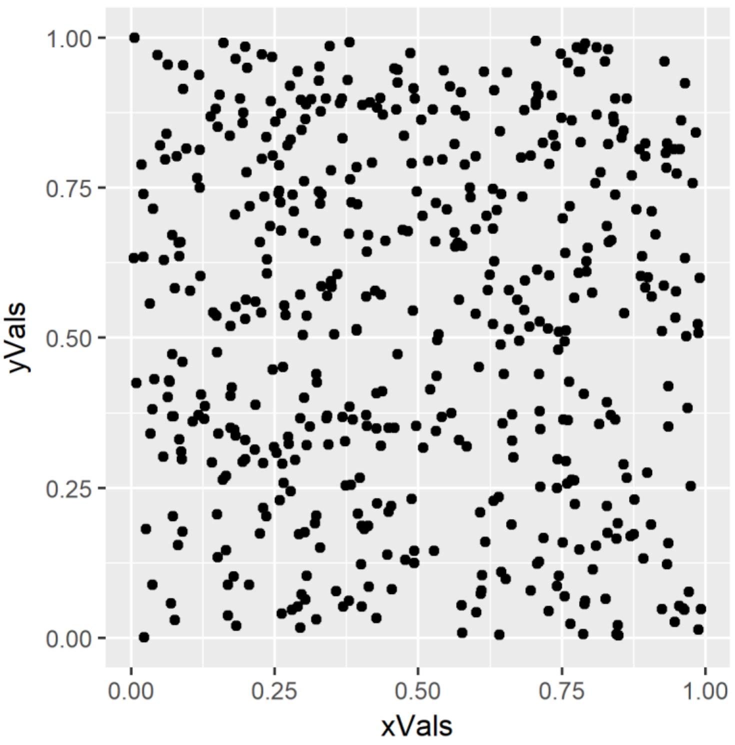


Figure 11: A 10 cm x 10 cm image of a scatterplot with 500 points

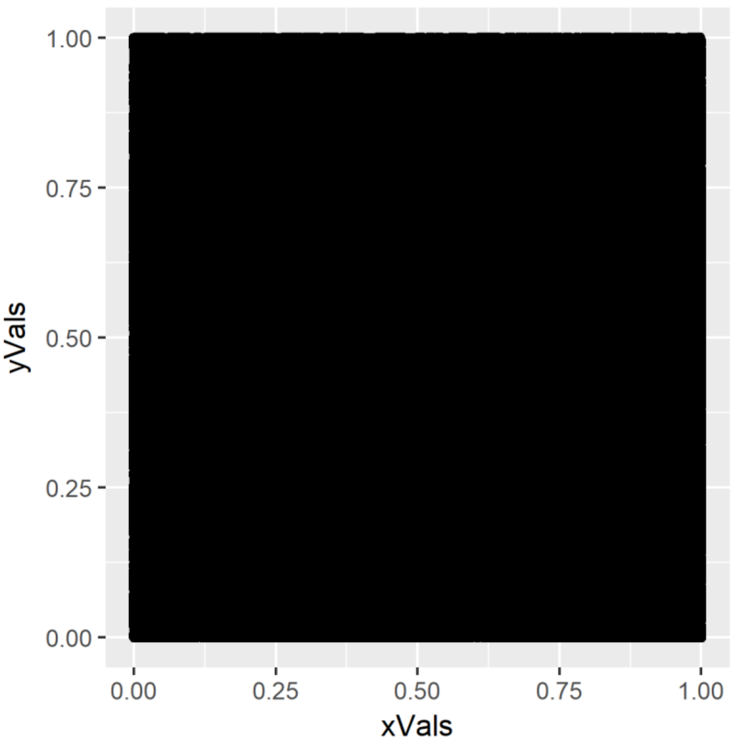


Figure 12: A 10 cm x 10 cm image of a scatterplot with 500,000 points

5. Diamonds [20 points]:

The diamonds dataset within ggplot2 contains 10 columns (price, carat, cut, color, etc.) for 53940 different diamonds. Type help(diamonds) for more information.

Plot histograms or bar charts for color, carat, and price, illustrating their distributions. What can you infer from these distributions (e.g. shape, skew, tail, etc.)?

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.

If you encounter computational difficulties, consider using a smaller dataframe whose rows are sampled from the original diamonds dataframe. Use the function sample() to create a subset of indices that may be used to create the smaller dataframe.

Submit:

Report: Your observations on the distribution of values for color, carat, and price, and a graph for each; a brief writeup explaining the three-way relationship between price, carat, and color, with appropriate graph(s).

Observations (for code and additional comments, see attached .Rmd file):

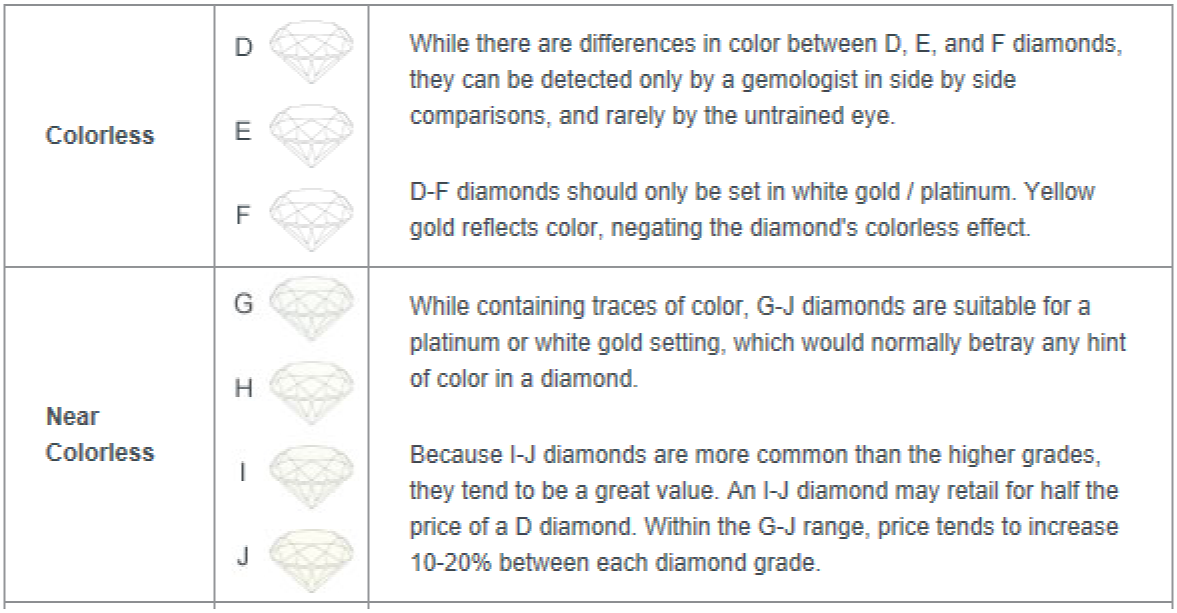
Beginning with color, reference Figure 13, which displays a color chart for diamonds rated D – J. This chart is extremely informative from a market value perspective. Naturally occurring diamonds will almost always have impurities trapped in their crystalline structure and these impurities impart different colors to the diamond. Some of these colors can add value to the gem, but from the perspective of the greater diamond market, a colorless stone is worth the most. Since this dataset only contains cut diamonds, it is safe to assume that the diamonds in the dataset will be skewed towards the colorless side of the spectrum as there are finite resources for cutting and they will tend to favor diamonds that would fetch higher prices (more colorless). This bears out in the chart (Figure 14), with grades D, E, F, & G accounting for 69% of the total stones in the dataset instead of the 57% one would expect if there were similar quantities of every color grade.

The carat aspect of a diamond (which represents weight) has some very interesting properties. As the narrow-bin-width histogram in Figure 15 shows, there are 10 significant modes the diamonds in the dataset tend to cluster around. The most significant local maxima fall at .25, .75, and 1 carat. Not surprisingly, the other local maxima also tend to fall around increments of .25. This is likely because the markets have converged on weights based on cultural norms, manufacturing efficiencies, and jewelry setting standards. Also, unsurprisingly, the overwhelming majority of the diamonds in the dataset are under 2.5 carats as this is what the mass market can afford.

The price of a diamond should technically be whatever the market will bear. Surprisingly, people seem willing to bear quite a bit for their sparkly rocks! The dataset contains prices as low as $326 and as high as $18,823, with half of the diamonds valued at more than $2,401 (median). When displayed as a histogram (Figure 16), three things pop out immediately: 1. There are no stones less than $300 dollars- this seems to be the minimum that sellers in the market will transact for. This means if you can cut a diamond you find on the ground, your time and effort could be worth at least $300! 2. Twenty-five percent of the stones cost less than $1,000. There is then a very long tail representing the prices of the remaining 75% of diamonds in the dataset. 3. There is a very conspicuous absence of diamonds that cost $1,500 ± $40. This could be a gap in the dataset or represent a step function increase in cost at a certain threshold of consumer. I'm leaning towards a gap in the data, because there are so many subjective measures of quality in jewelry that an $80 diamond desert seems unlikely. As a side note, a boxplot would not have caught this gap.

Finally, Figure 17 displays the ggpairs visualization comparing the three facets (pun intended) of the diamond dataset just investigated. The visualization that seems most obvious is the price as a function of carat plot in the bottom center. You could easily draw a line that would act as a minimum price per carat, but the maximum price per carat does not seem to have any upper bound. The ggpairs plot; however, doesn't really indicate a three-way relationship between color, carat, and price. Instead, I created Figure 18, which clearly shows that small, high-grade colorless diamonds are priced the same as large, middle-grade near-colorless diamonds. You can see that the preponderance of sub-2-carat gold dots at any given price in the 1 carat range for D grade diamonds gives way to the larger 2-4 carat green and blue dots in the lower E-J grade diamonds.

Sample Output:



Colorless	D E F	While there are differences in color between D, E, and F diamonds, they can be detected only by a gemologist in side by side comparisons, and rarely by the untrained eye. D-F diamonds should only be set in white gold / platinum. Yellow gold reflects color, negating the diamond's colorless effect.
Near Colorless	G H I J	While containing traces of color, G-J diamonds are suitable for a platinum or white gold setting, which would normally betray any hint of color in a diamond. Because I-J diamonds are more common than the higher grades, they tend to be a great value. An I-J diamond may retail for half the price of a D diamond. Within the G-J range, price tends to increase 10-20% between each diamond grade.

Figure 13: A table explaining the different color grades of diamonds that I pulled from the website: https://www.lumeradiamonds.com/diamond-education/diamond-color

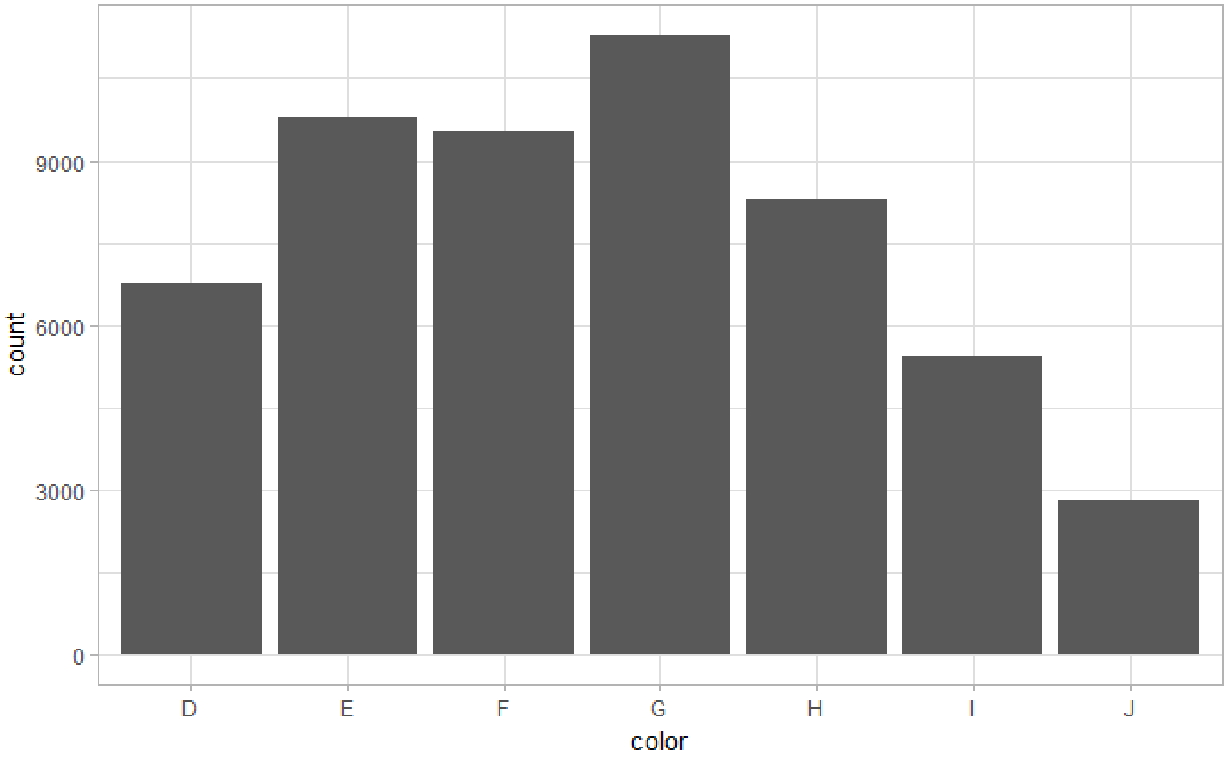


Figure 14: A simple bar chart showing the relative quantities of different color grades in the dataset

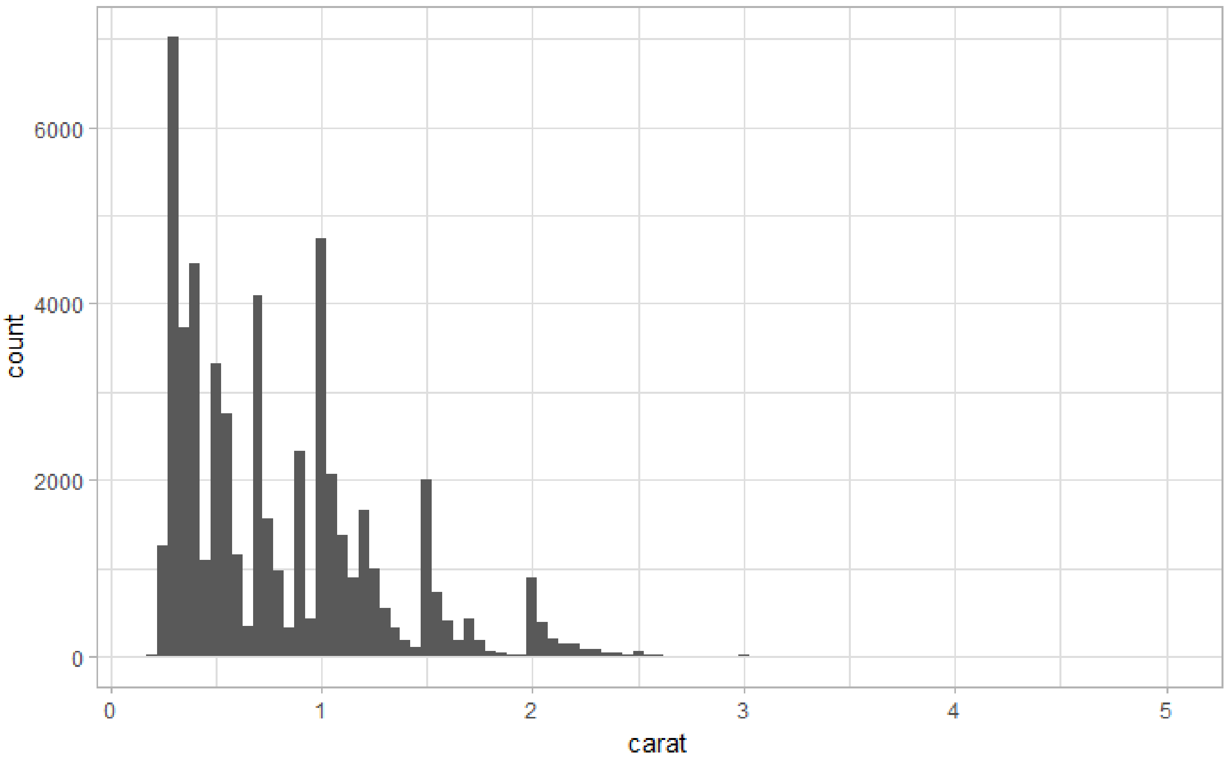


Figure 15: Histogram showing the diamond carat clusters around high demand jewelry standards

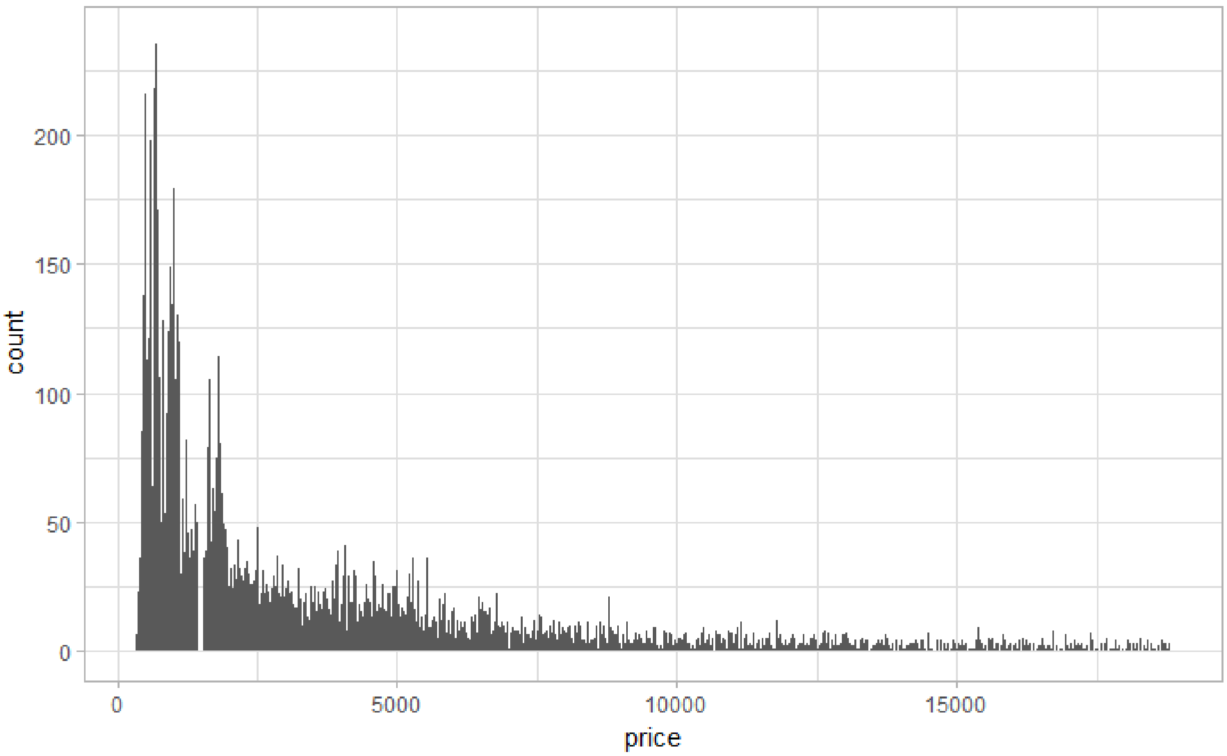


Figure 16: Histogram showing diamond price distribution with a noticeable gap at $1,500

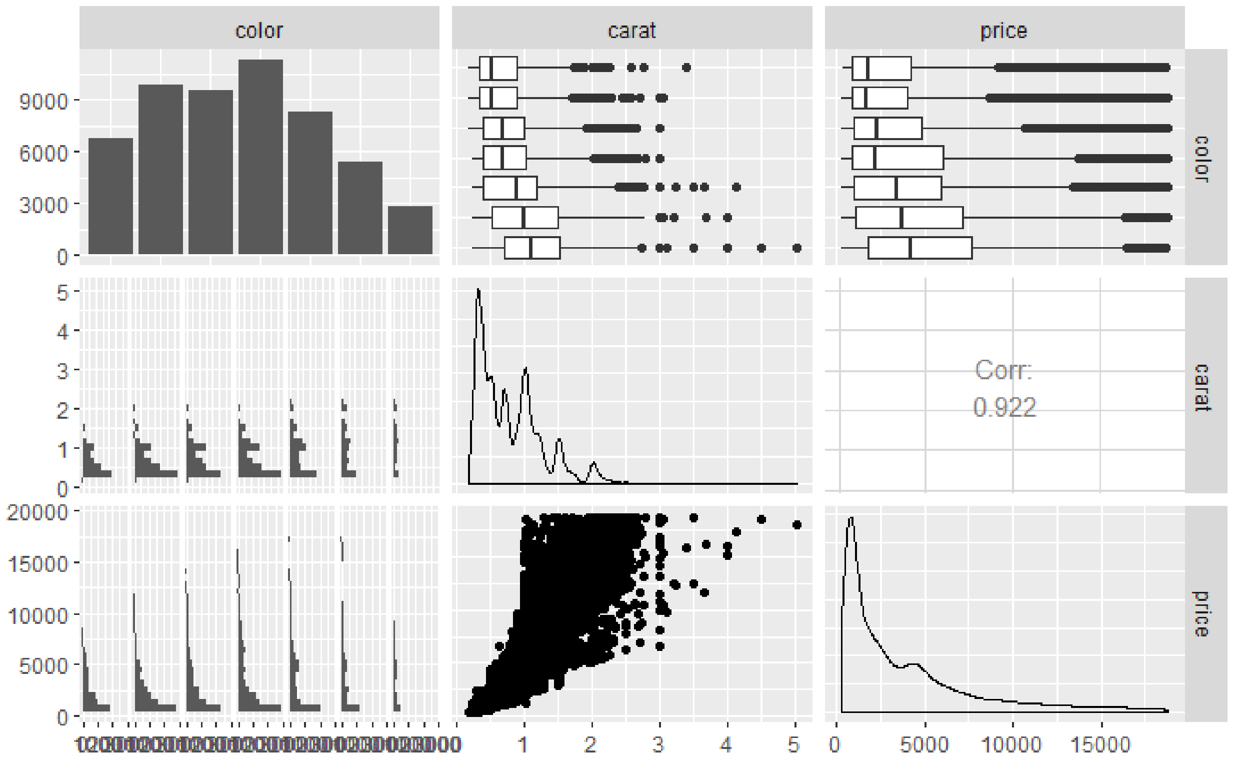


Figure 17: A ggpairs visualization of color, carat, and price

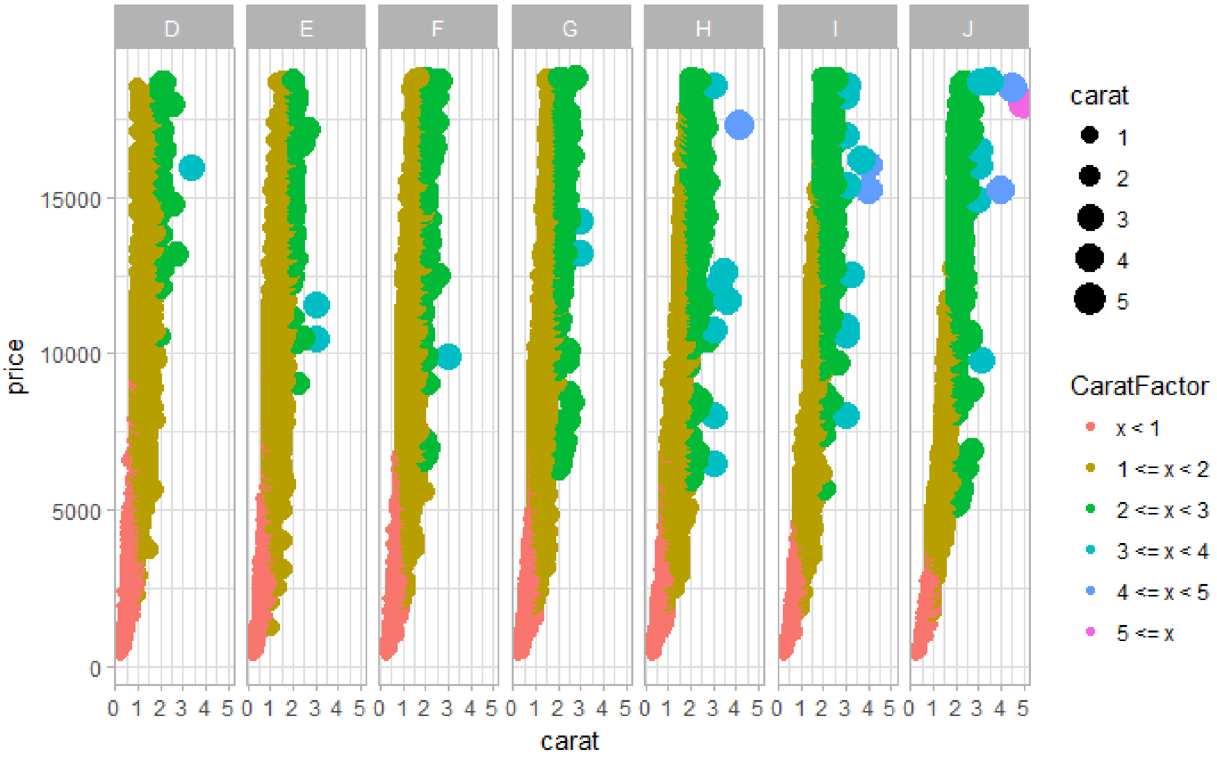


Figure 18: A visualization of the relationship between color, carat, and price