Jeremy Kang

GTID: 903136703

GT Name: jkang315

1.

2.

3. Add pros and cons

Add QQPlot

Add sample of N numbers (example of graphs)

When would you use a Box Plot, a Histogram, or a QQPlot to graphically summarize N points?

**Histograms**

Histograms are a data visualization that explores one-dimensional data by dividing the ranges into bins and counting the frequencies in each bin as a measure. This is an easy way to see the density of the distribution in your data, as well as the shape of your data. However, you must set your bins correctly so that you can see important areas in the data, but also not get distracted by noise/outliers in your data.

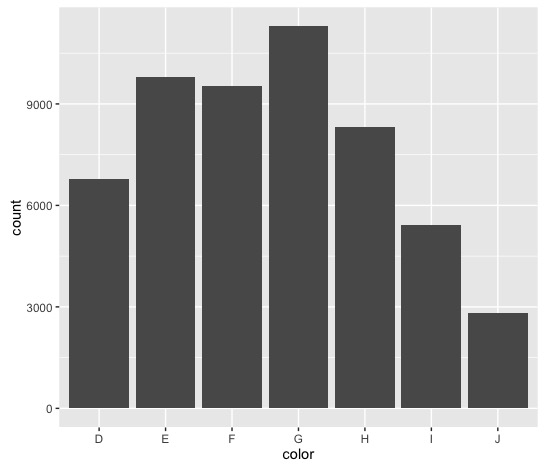
**Box Plot**

Boxplots also offer a visualization of the distribution in your data, but in the form of quartiles rather than a frequency. 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 to lower the chance of being distracted with outliers. In fact, outliers are clearly pointed out by points that are even further from the “inner box” which represents the inter-quartile range. You’re also able to see where the majority of your data is distributed, the median, and IQR. A major disadvantage to using box plots is you have less information to see individual points in the data.

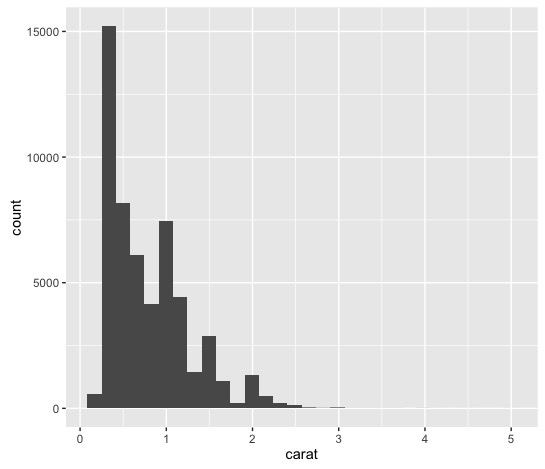
I appreciate what both visualizations have to offer, but histograms are probably my go-to visualization. It is easier to understand the overall distribution of the data, and if the bins I’m setting is still not solving any outlier issues, I can use box plots to get a better understanding of what the issue could be.

4. The file size changes in a logarithmic way in that there is an increase in file size with smaller samples and it actually decreases after a certain point. This is explained by different compressions used in the file types, specifically, as the points (black dots) of the file fill the plot more and more, they become harder to discern visually and in regards to where it is on the pixel.

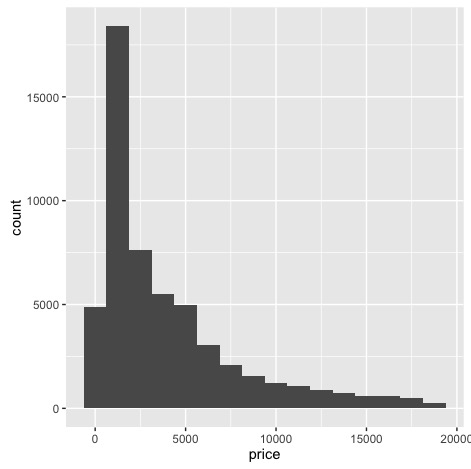
5.



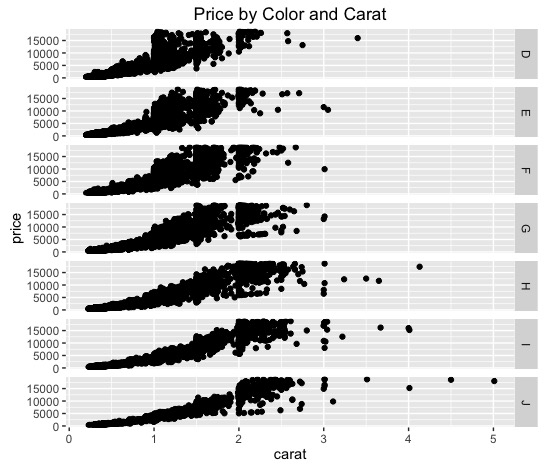
G is the most frequent diamond color with a count of 11,292 diamonds. That is reasonable to expect, but the second most frequent is E, the second from the worst color. Although the difference is probably insignificant between E and F, it’s interesting to see that E is the second most frequent in this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



The histogram for carat is a little more interesting. There is a left skew in the graph that shows that diamonds are typically smaller than 2 carats. We have a wide range of carat sizes from 0.2 to 5.01 carats with the IQR being 0.64 and average size being 0.798.



Again, we see a left skewed histogram for the price of diamonds with an IQR of $4,374.25 and average of $3,932.80. What’s the most interesting about this plot is that it shows there are good portions of diamonds that are above the average price, even though the majority is within the $4,000 range.



The price of the diamond appears to increase exponentially in all colors and carat sizes. As we saw in the histogram of carats before, the majority of carats range from 0.40 to 1.04 carats. What’s interesting is that the majority of the colors show some gap between 1.5 and 2 carats, indicating there could be some clusters for cars at or smaller than 1.5 and carats 2 or larger. This observation isn’t apparent in diamonds with H color. However, further investigation is needed to see if that observation is credible. There appears to be more outliers as you look at better 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 cut or size [carat] or vice versa with diamonds of D color).