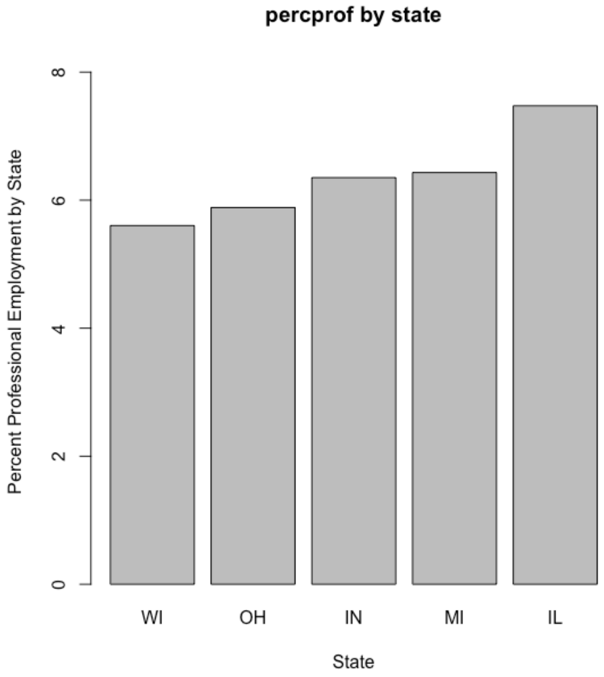
GT Account Name: tbobik3
Tyler Bobik

HW2: Data Visualization

Problem #1

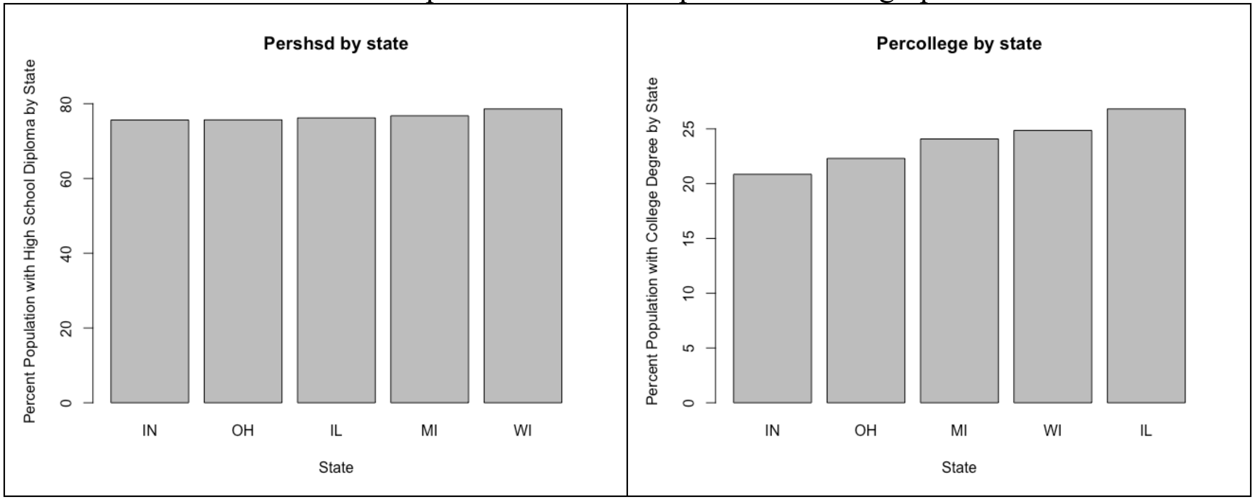


For this problem 1 used Interpretation A:
Aggregate percprof for each state by combining the values from each county in that state, and generate a "percprof by state" plot.
Note that since percprof is a fraction of popadults, when aggregating it, you will have to use (s = state, c = county): $\forall s, percprof_s = (\Sigma_{c \in s} percprof_c \times popadults_c) / \Sigma_{c \in s} popadults_c$

As you can see here the state Illinois (IL) has the highest percentage of professional employment with 7.47%, the lowest percentage of professional employment is the state Wisconsin (WI) with 5.60%.

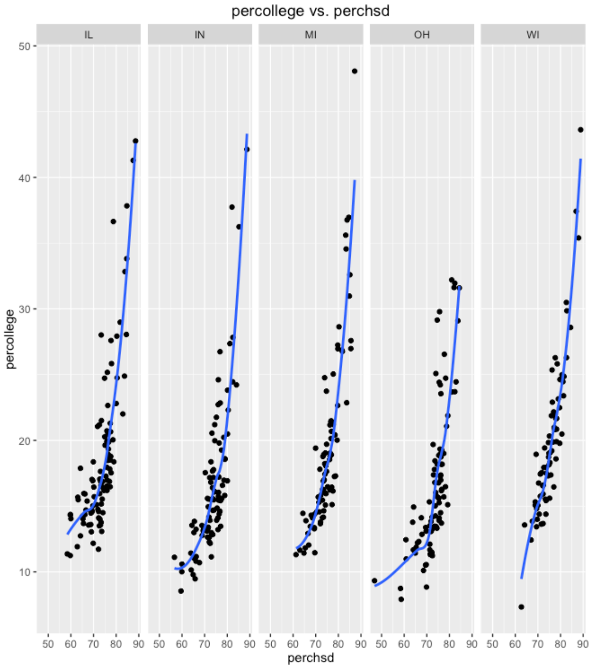
Problem #2

I used interpretations A to complete these two graphs.



In the pershsd by state histogram, you can see that WI has the highest percentage of high school graduates given the adult population and IN has the lowest percentage, although they are all pretty equal. I suspect this is because high school is mandatory in the US. It is interesting that in the Percollege by state histogram IL has the highest percent of college graduates given the adult population because IL was only the third highest state in terms of percent of high school graduates in the pershsd by state histogram.

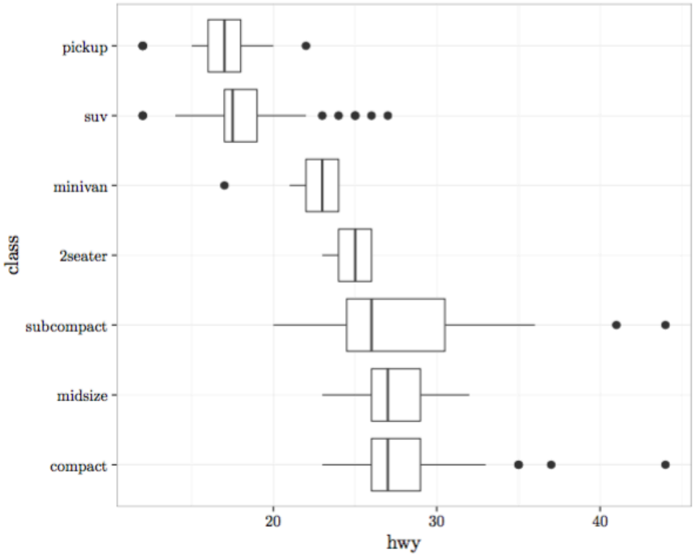
For the next graph I decided to use interpretation B because I feel that looking at perchsd and precollege by county in each state will return a more granular result that may provide more insight.



You can see here that there is a high positive correlation between the percentage of high school graduates (perchsd) and the percentage college graduates(percollege) for each county grouped by state. Specifically, for each county you can see that a higher percentage of high school graduates results in a higher percent of college graduates, which makes intuitively.

Question #3

A boxplot has the following elements: a box denoting the IQR (interquartile range) is the range between the 25th percentile and the 75th percentile which is denoted using a box, an inner line bisecting the box denoting the median or the 50th percentile, whiskers that extend to either side of the box which indicate the range of the data while we exclude outlier to show where the majority of the data reside in. The median may or not be in the center of the box, depending on whether the distribution of values is skewed or not. The whiskers extend to the most extreme point no further that 1.5 times the length of the IQR away from the edges of the box. The outliers, which are marked separately outside of the whiskers are graphed as separate points which is useful for avoiding a misleading viewpoint where there are a few extreme non-representative values.



Differences in size of box: if there is a big difference in a range of numbers in a dataset this would cause the box to increase in size. If the difference between a range of numbers is small in a dataset this would cause the box to decrease in size. (compare subcompact to pickup in chart above).

The data is not symmetric: If the median is skewed to the left or right. (ex. Median skewed right: subcompact in plot above).
The data is symmetric: If the median is in the geometric center of the box. (ex. minivan in plot above).

Data's upper quartile is equal to the maximum: No whiskers to the right of the box. (ex. Minivan in plot above)
Data's lower quartile is equal to the minimum: No whiskers to the left of the box.

Lowest data point is within 1.5 IQR of the lower quartile and the highest point is within 1.5 IQR of the upper quartile: Whiskers on both sides of the plot (ex. Midsize in plot above).

Highest data point is above 1.5 IQR of the upper quartile: Outlier point is present outside the box and whiskers. (ex. Subcompact in plot above)
Lowest data point is below 1.5 IQR of the lower quartile: Outlier point is present outside the box and whiskers. (ex. Minivan in plot above)

It should be noted that small samples can cause the quartiles to become meaningless.

Pros of Box Plots:
Box plots are alternatives to histograms that are usually more lossy because they lose more data but emphasis quantiles and outliers in ways histograms can't. Useful where displaying the whole dataset is not possible. Good at summarizing large amounts of data. Provide some information of data's skewness. Good for side by side comparison of more than one distribution.

Cons of Box Plots:
Not good with small number of observations. Can only be used with numerical data. Box emphases attention on the central half of the data. Does not convey the multimodal nature of distribution like histograms can.

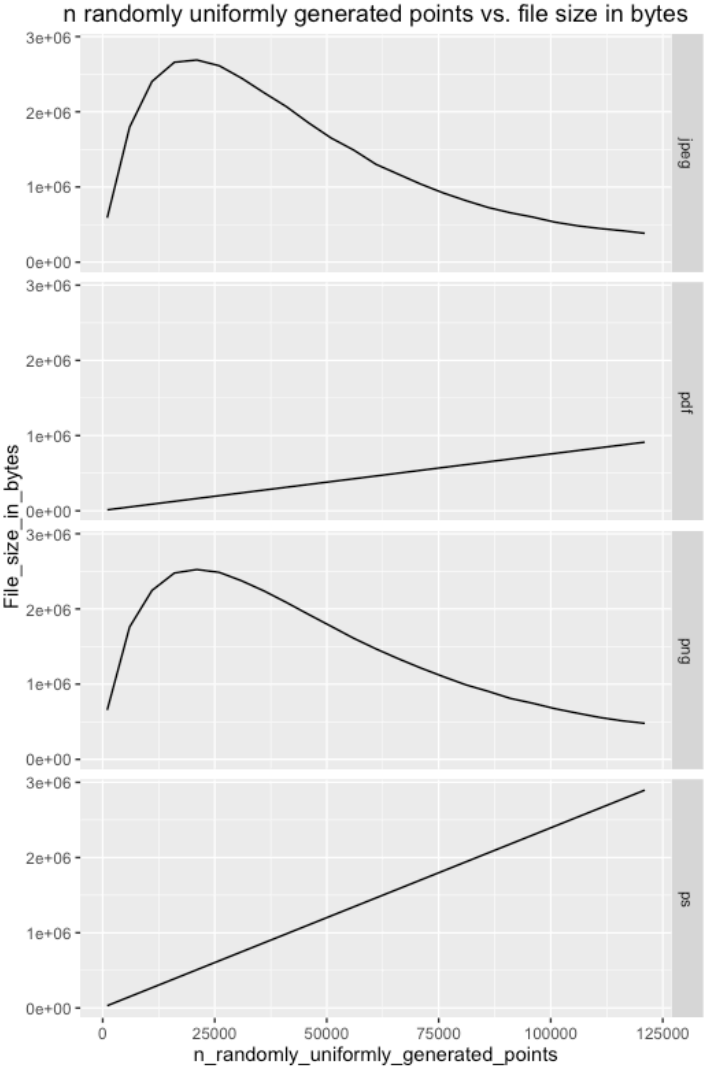
Pros of Histograms:
Histograms are useful if you are trying to summarize numeric data in that they show the rough distribution of data. Easier to read the median and the IQR. Histograms enable you to easily compare data and pair good with large value ranges of data. Apply to continuous, discrete and unordered data.

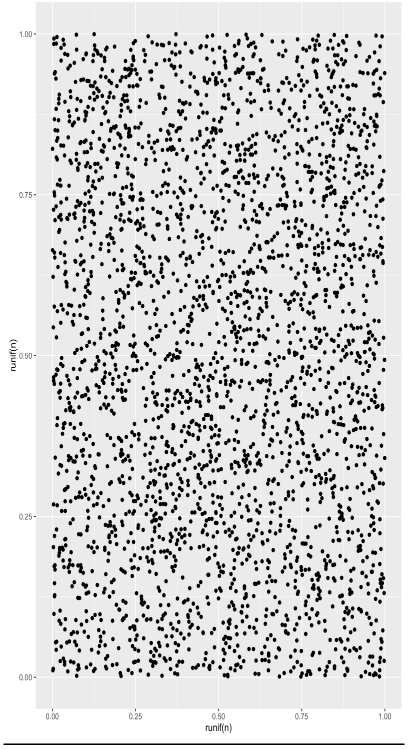
Cons of Histograms:
It is very hard to tell the exact amount of data that is used in it unless it is a frequency histogram. It's hard to display multiple at a time for comparing them against each other. Can set bin width value too big or too small making the graph not very useful.

Sometimes histograms are more useful but in other cases, box plots are more useful, and they reveal information that a histogram does not have. Each are good tools to have in your toolkit you just have to be aware of what situations to use each in.

What kinds of data is most useful to use with:
Box Plot: Large amounts of numerical data.
Histogram: continuous, discrete or unordered data. Data has a large range of values. One dimensional numeric data.
QQ-Plot: continuous data from two distributions.

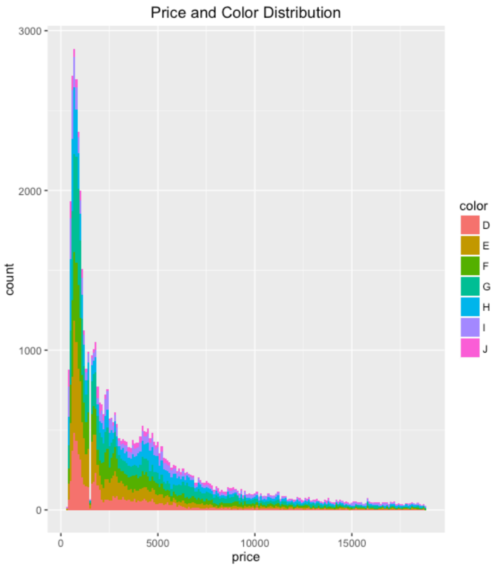
Question #4



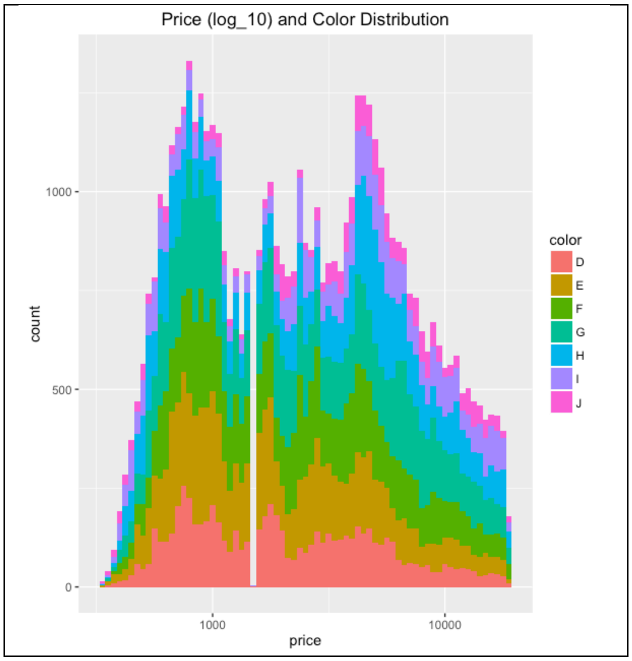


This plot is very interesting, you can see that the file size for PNG and JPEG increase very fast till around 25000 randomly generated points then start decreasing in size and eventually end up leveling off. My best guess as to why this is happening is because of some form of image compression for these two types of images. PDF and PS both increase linearly but PS increases at a higher rate than pdf.

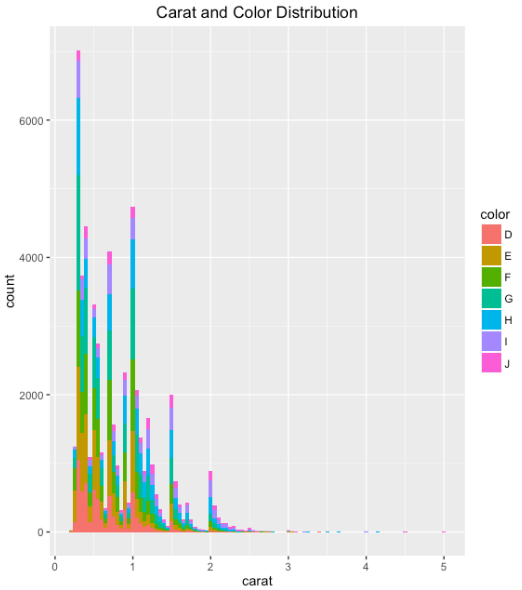
Question # 5

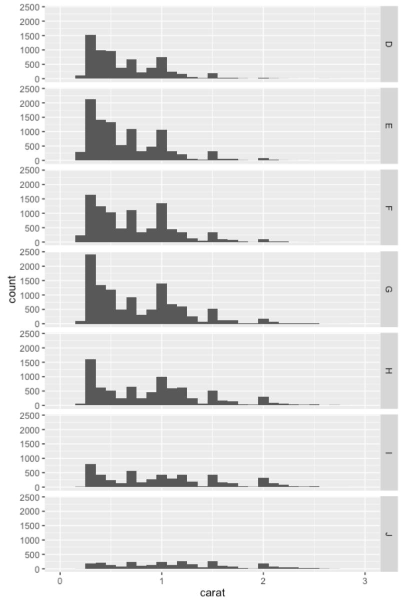


You can see here that the price is skewed to the left. I know that for monetary amounts this is common and that if it is true you should use log base 10. I decided to graph both with and without log base 10 to see what is better.

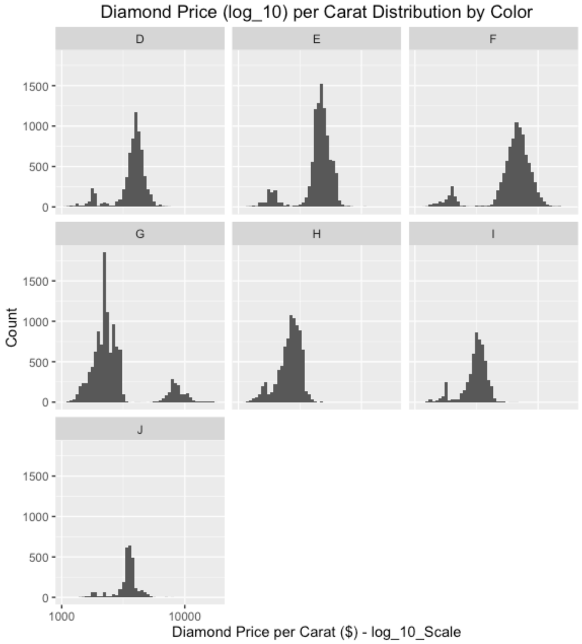


You can see that after I applied the log 10 scale to price that the prices are much closer to a bell curve that you would find in a normal distribution.

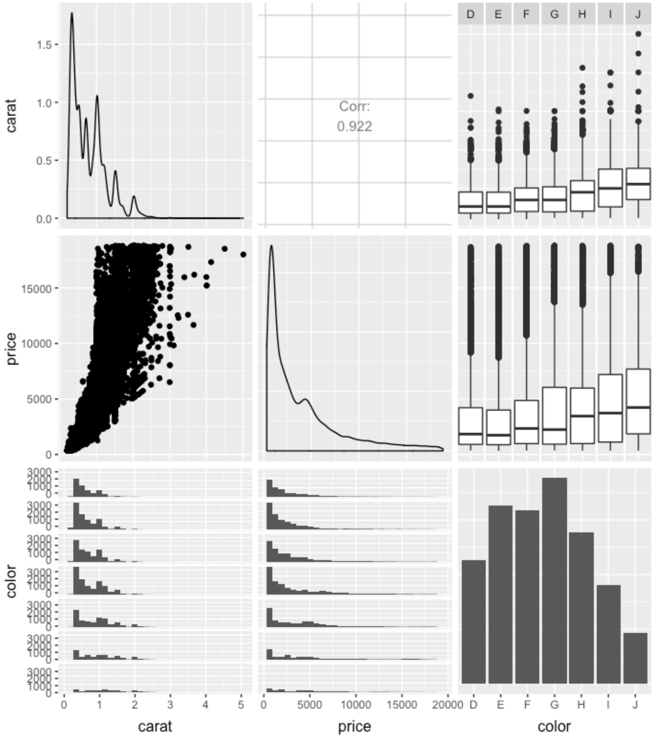


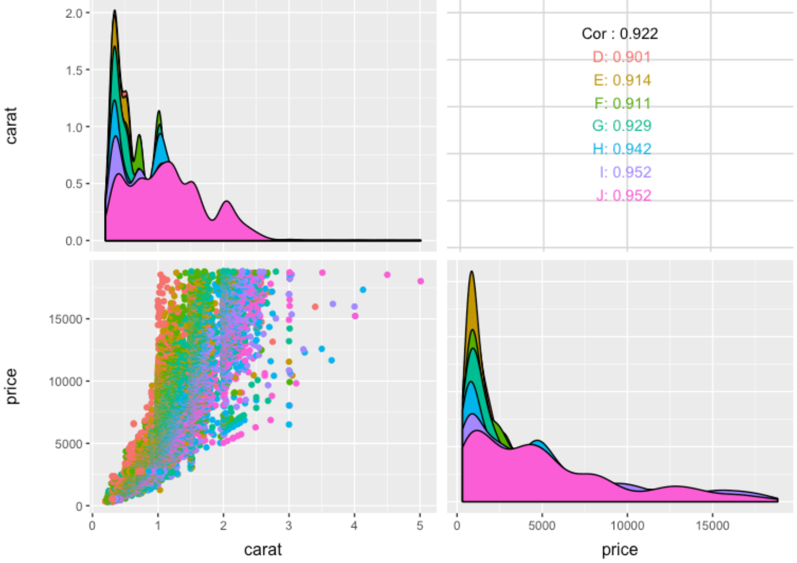


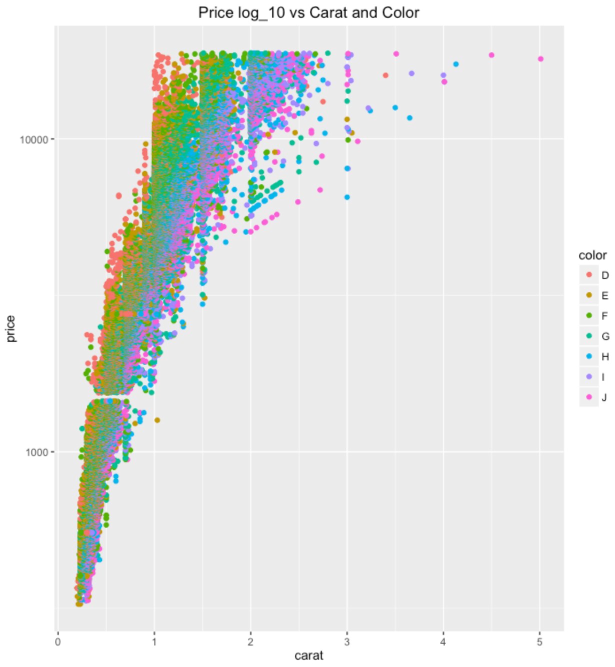
Here you can see the distribution of diamonds by carat is also skewed left, I believe I would have to do more research into this dataset as to tell definitively why this is occurring. You can see that the color D has the least amount of diamonds in the dataset as color J has the most amount of diamonds.



Here using the log 10 scale of diamond price per carat for each color type with a histogram. Here you can see that the distribution of price per carat according to color relationship. You can see that that for each color there seems to be a certain price per carat that each is centered around that is unique to each. Color F's centered price per carat is the highest, the second is located in E, third highest is D, fourth highest is I, fifth highest is J, sixth highest is H and seventh highest is G (the lowest) for color G it's price per carat is the lowest but the count is the highest.







You can see the main attribute that is driving price is the size of the carat. The price between price and carat size is non-linear but seems almost exponential. You can also see that the variance increases dramatically in terms of price when the diamonds are larger than 2 carats. Additionally, you can see by this pairwise plot that the difference in diamond color does explain some of the variance in the price.