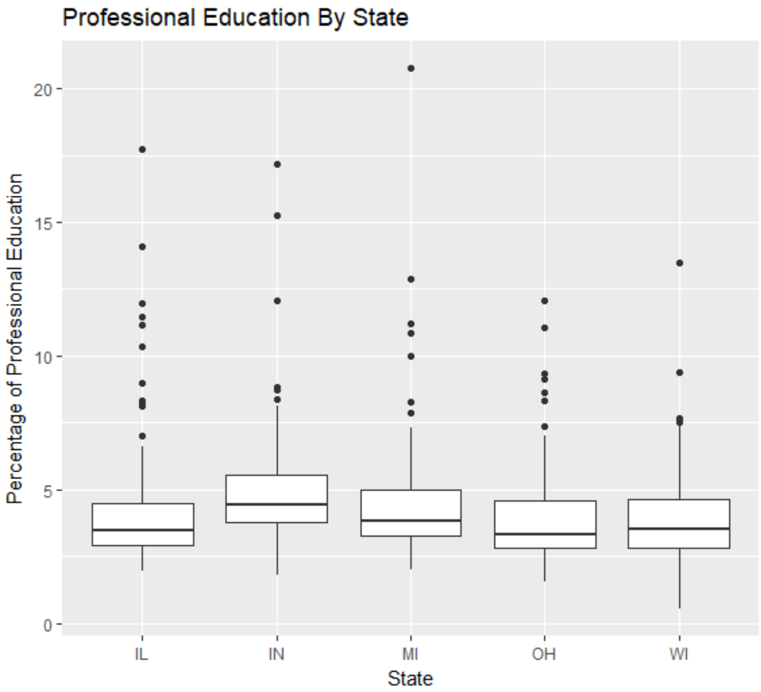
GT Username: sanne31@gatech.edu

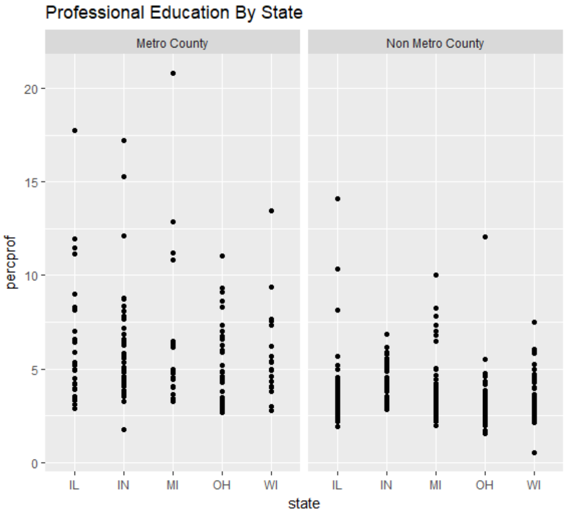
Question 1: Professional Education by State

- Used box plot in ggplot2() to visualize the percentage of professional education (*percprof* is the continuous variable) grouped by state (*state* as the discrete variable)
  - Supporting box plot is shown below



- From the above plot, we can observe the following,
  - The maximum percentage of professional education in any given state seems to be closely around 7%
    - 7 % for IL
    - 8 % for IN
    - 7.2 % for MI
    - 7% for OH
    - 7.5 % for WI
  - However, the median for all given states lies between 3% to 4 %
  - Indiana seems to have a higher spread as compared to the other states
  - There are many outliers noticed in all the states in the higher percentages of professional education as discerned from the box plot

- Another division by county is given below using a faceted point plot



  - The above splits the plot into 2 based on the metro or non-metro county

- Finding the highest and lowest percentage of population with a professional education, there is no direct way of observation. I divided the dataframe by each state first. After that, I calculated the actual population that had a professional education and summed up all the values for each county. Upon dividing this by the total population of the state and multiplying by 100, I obtained the percentage of professional education based on the population by each state. I stored all that data into a data frame.

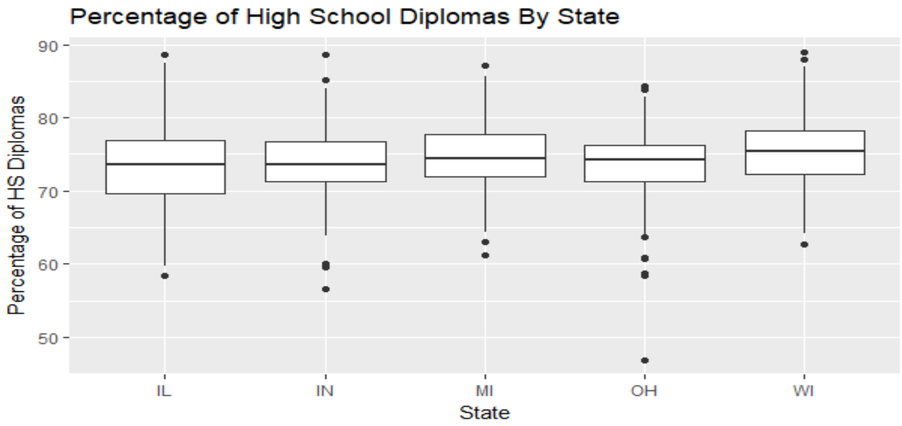
```
 state  edurate
1    IL 7.527751
2    IN 6.427714
3    MI 6.472557
4    OH 5.897535
5    WI 5.636301
```

- Upon further inspection, we arrive at the conclusion that IL has the highest percentage (7.53%) of population that have a professional education whereas WI has the lowest(5.64%)

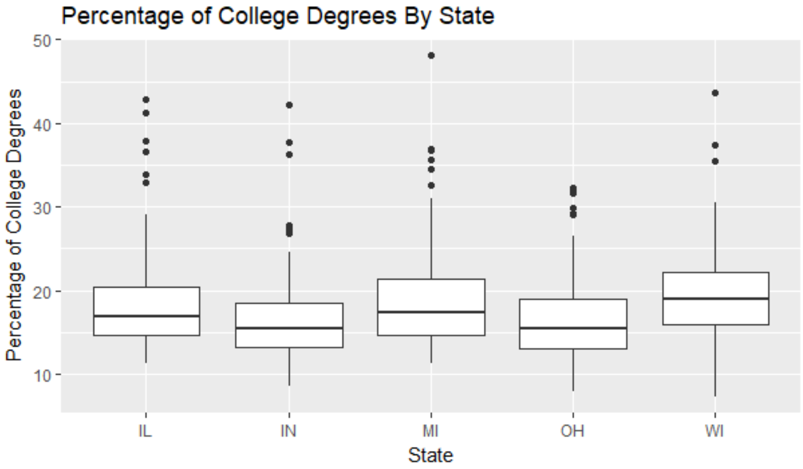
Question 2: School and College Education by State

- Percentage of High School Diplomas vs State
  - Median percentage of population having high school diplomas is observed to be approximately 75% among all states

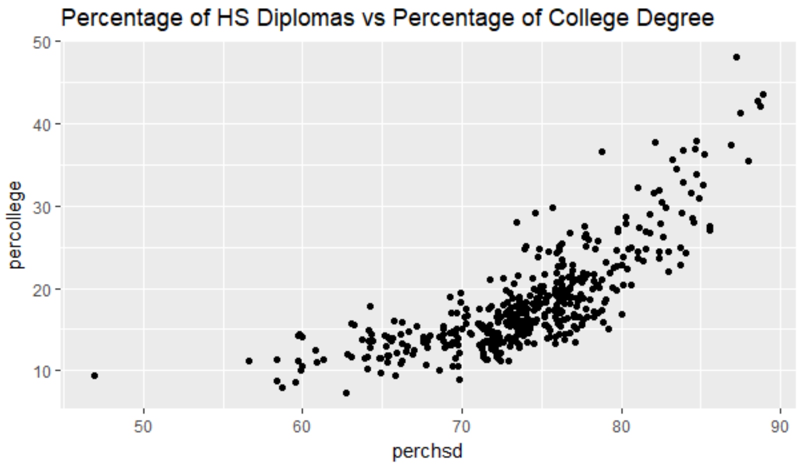
  - However, the state of Illinois seems to have a higher spread leading to the fact that a higher percentage of people have actually completed their high school diplomas in IL, as compared to other states
  - Another interesting observation is that the IQR (inter quartile range) for all states seems to be between 70% - 78% which might also suggest that many people of these observed states have completed their high school diplomas.



- Percentage of College Degrees vs State
  - The median for this plot of percentage of college degree holders for each state seems to be around 15% - 20% among all given states which is significantly lower than the plot observed above
  - One of the main conflicting characteristics observed is that the IQR for the given boxplots for the given states is less than 22% which means that not many people in the observed states have moved towards a college degree
  - These observations, coupled with those above seem to indicate that not many of the population in any of the given states having obtained a high school diploma, tend to attain a college degree. This is further explored in the next plot between high school diplomas and college degrees.

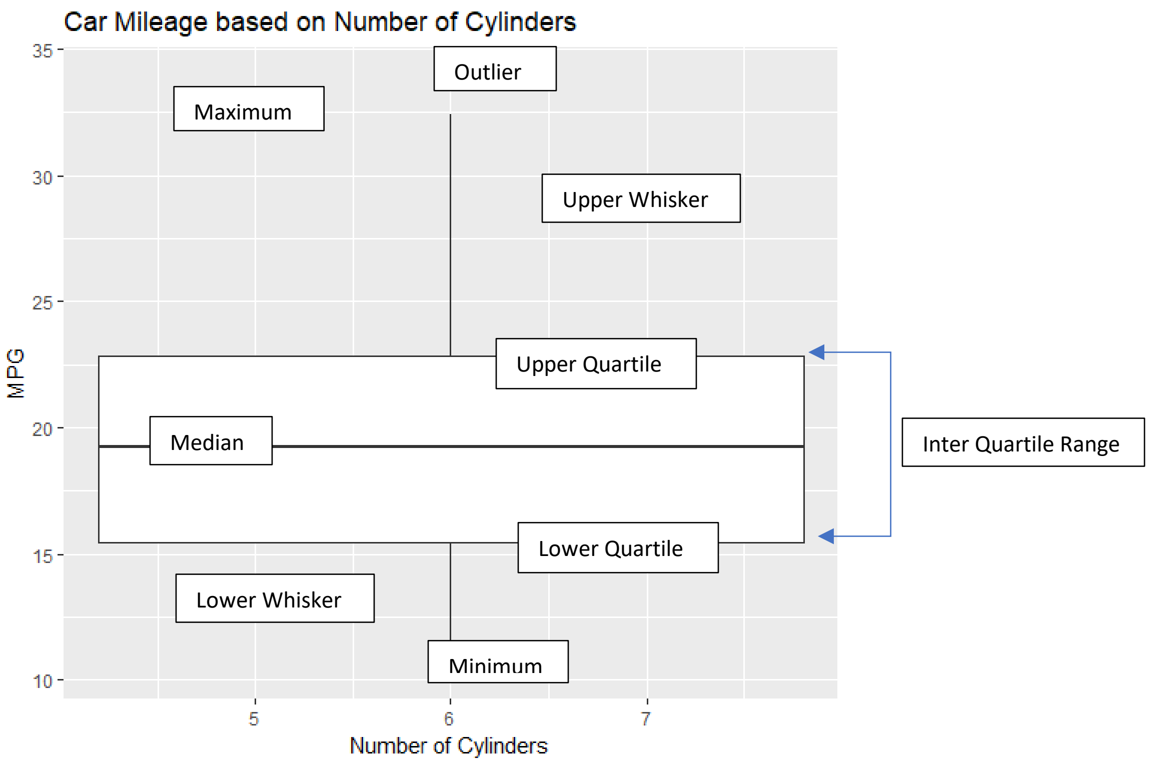


- Percentage of High School Diplomas vs Percentage of College Degrees
    - The scatter plot shows a strong positive correlation between the two quantities
    - We can immediately observe from the graph that not many students who attained a high school diploma went for a college degree.
    - Another interesting fact is that after the 70% mark, the percentage of college degrees starts increasing significantly leading to believe that a high school diploma is strengthening the attainment of a college degree however limited it might be.



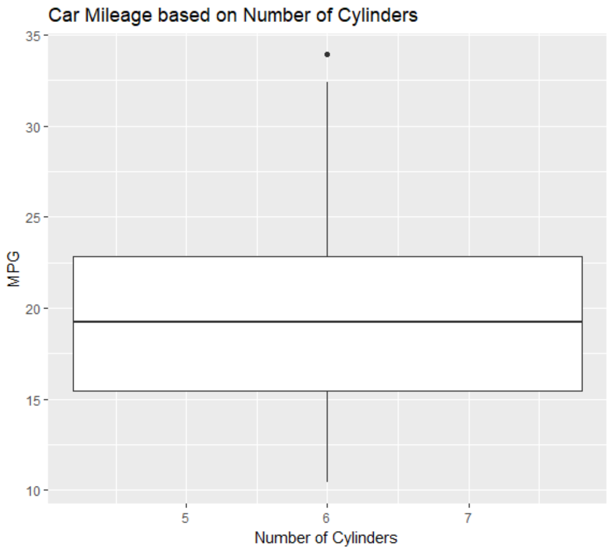
Question 3: Comparison of Visualization Techniques

- **Box Plots**
    - Main aspect of box plots is that they enable us to the study the distributional characteristics of a group as well as the level
    - the plot below shows the relationship between the mpg and the number of cylinders in each of the cars in the mtcars dataset

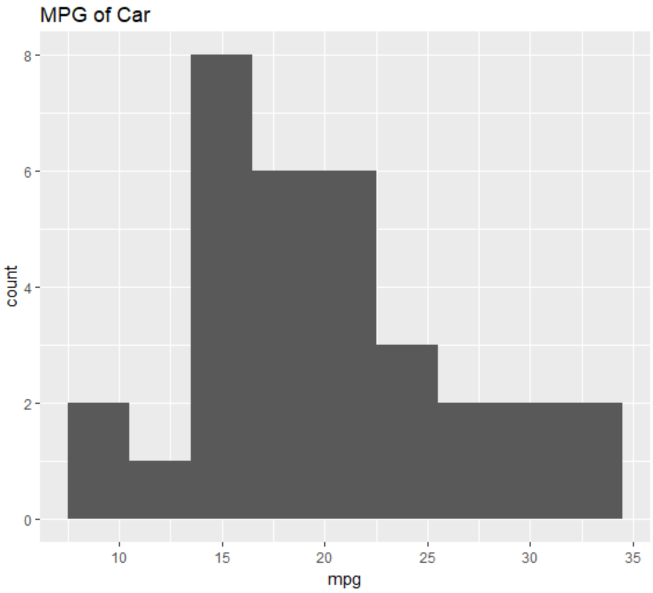


- Median marks the mid-point of the data and is shown above by the line that divides the box plot into 2 parts. Half of MPG values are greater than this value and the remaining half are less.
- Inter quartile range is the middle box which represents the middle 50% of mpg values for the group of cars taken into observation. This also is represented by the range of values between the upper quartile and the lower quartile.
- Upper Quartile – 75% of the values fall below this value
- Lower Quartile – 25% of the values fall below this value
- Whiskers – the upper and lower whiskers represent the mpg values which don't lie in the middle 50% of values and are limited by the maximum and the minimum value

- **Reasons for various plots**
  - Box plots generally don't need the data to be normally distributed. It is generally viewed as a more lossy type given the fact that it has so many outliers. It is useful for being able to identify the differences between the spread of the data and the center when comparing with multiple datasets.
  - A good example for box plots is shown below. (same plot between mpg and number of cylinders in the mtcars dataset.

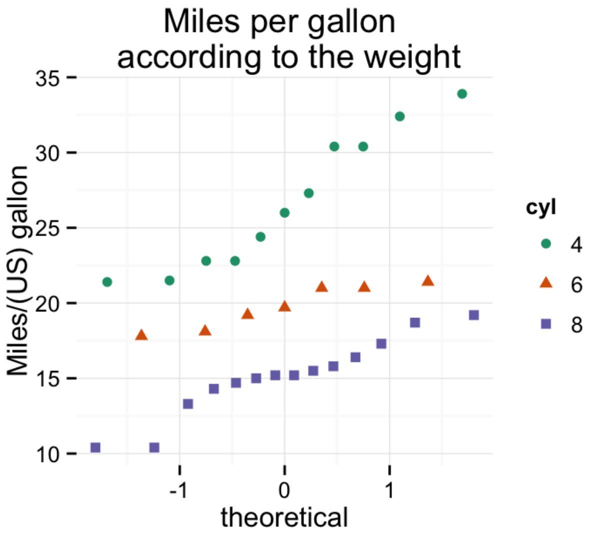


- Histograms can be used to identify the data distributions clearly which can be used to identify the independent and the dependent variables in a dataset.
- Histogram of MPG values is shown below with the frequency and bin width of 3
- We can clearly observe the normal distribution centered around the mpg value of 20



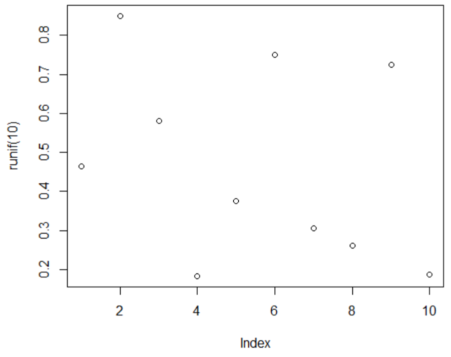
- QQ Plots can be used to identify whether the given dataset follows normal distribution or not. They plot the values in the dataset against the values from a standard normal distribution.
- One thing to remember is that if data is normal, slope of the imaginary line of best fit will be constant

- Example of qq plot between MPG and weight of the car is shown below

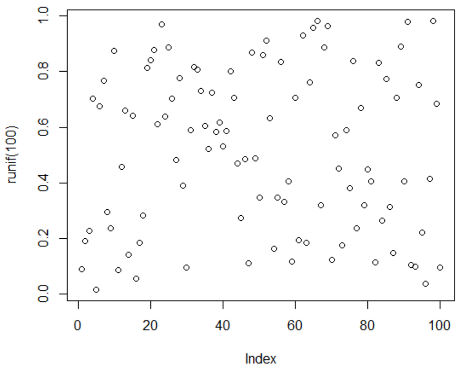


Question 4: Random Scatterplots

Scatter plot using runif(), N = 10

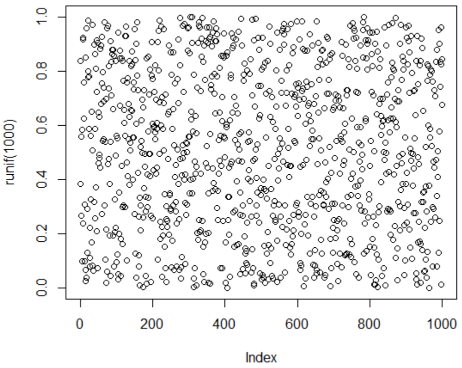


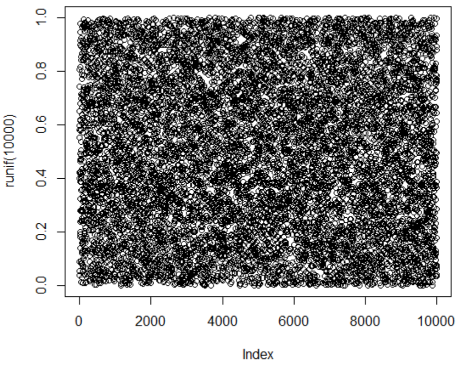
Scatter plot using runif(), N = 100



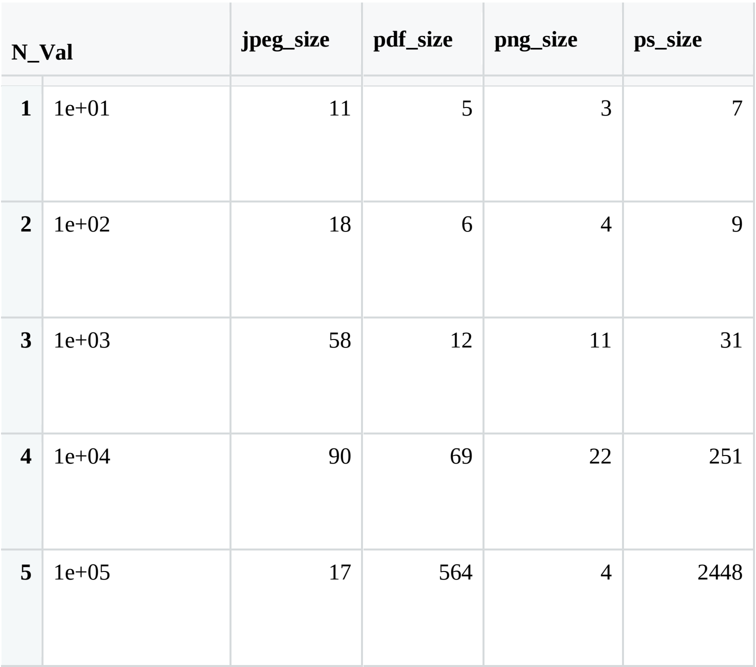
Scatter plot using runif(), N = 1000

Scatter plot using runif(), N = 10000



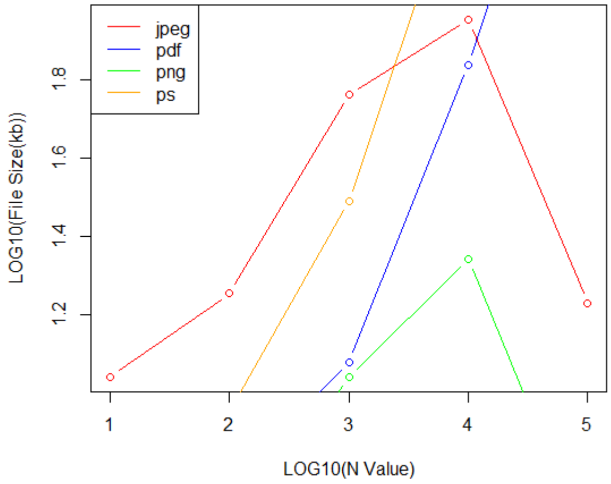


- I saved a plot with N value ranging from 10 to 100000 for each file type, namely: *.jpeg, .pdf, .png, .ps*
- Observed file sizes for each type and for each value and created a dataframe for all the values



	N_Val	jpeg_size	pdf_size	png_size	ps_size
1	1e+01	11	5	3	7
2	1e+02	18	6	4	9
3	1e+03	58	12	11	31
4	1e+04	90	69	22	251
5	1e+05	17	564	4	2448

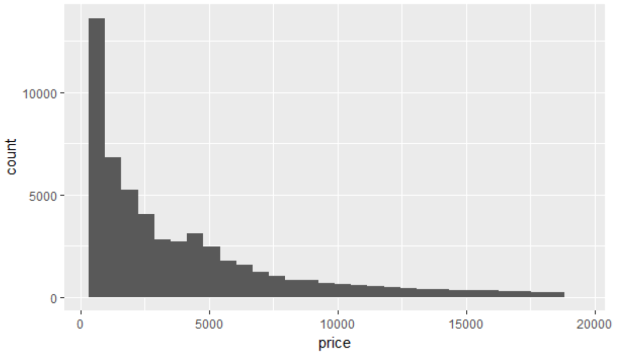
- Sized the dataframe to enable simpler plots by using log base 10 for each field in the data frame
- Plotted the file sizes for variable sizes of N and the plot is shown below



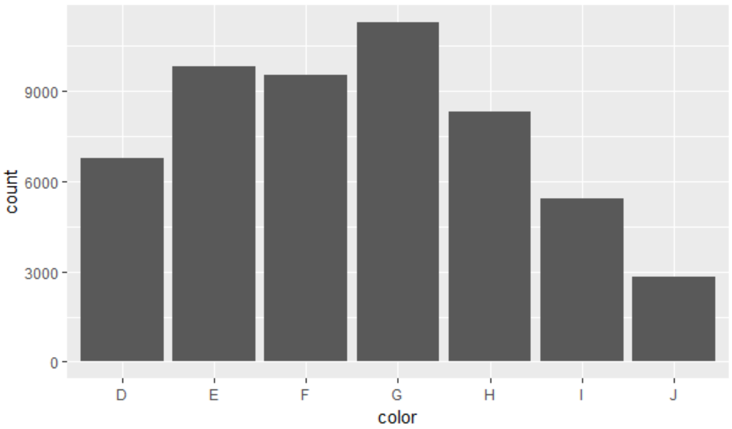
- The *pdf* and *ps* file types show exponential size increase
- However, both the *jpeg* and the *png* show a different case where they increase until N value of 10000 and then they decrease.
  - The reason for this is because as the N value increases, the plot becomes almost completely dark, filled with dots/points. This dense collection of points causes the image to appear solid. However, the same is not reflected in both *pdf* and *ps* types because all the dots/points are written to both files instead of appearing as a single block

Question 5: Diamonds

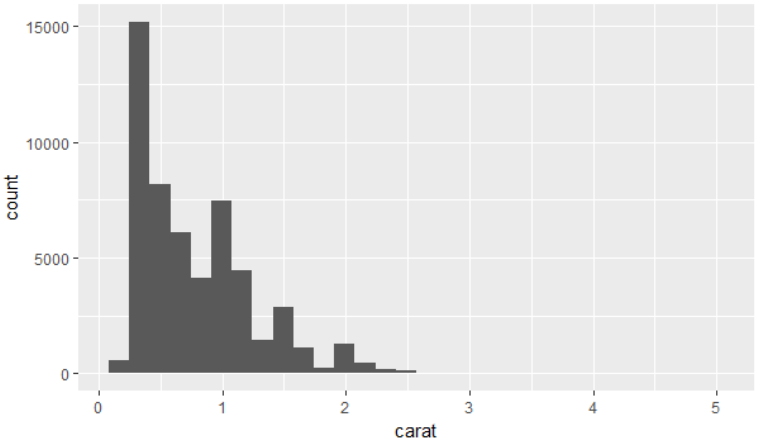
- **Price**
  - By observing the graph below, it is noticeable that higher priced diamonds are rare and fewer in quantity.
  - Decreases exponentially



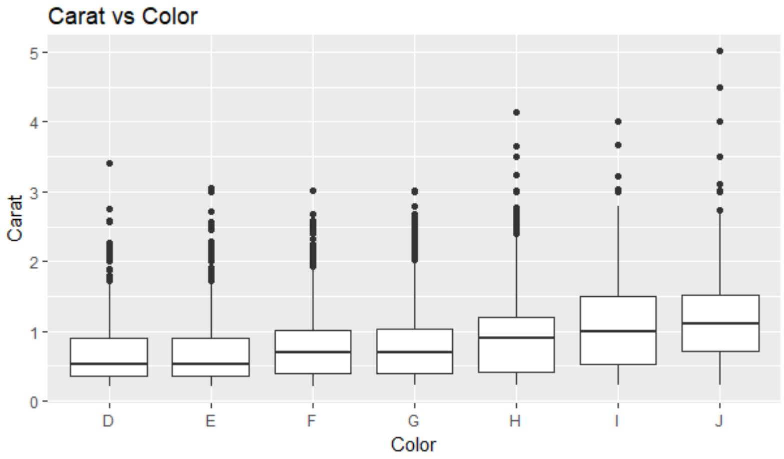
- **Color**
    - Upon direct inspection, there is not much that can be observed about the distribution.
    - It appears that J color diamond is the rarest and G color diamond is the most common one among them all.



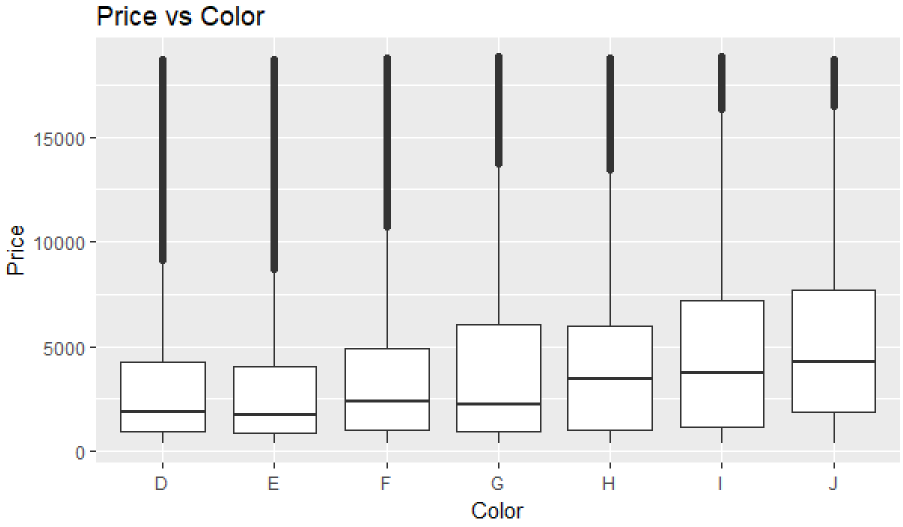
- **Carat**
    - The entire distribution is skewed to one side (the lesser carat value)
    - Centered around 0.5 carats.



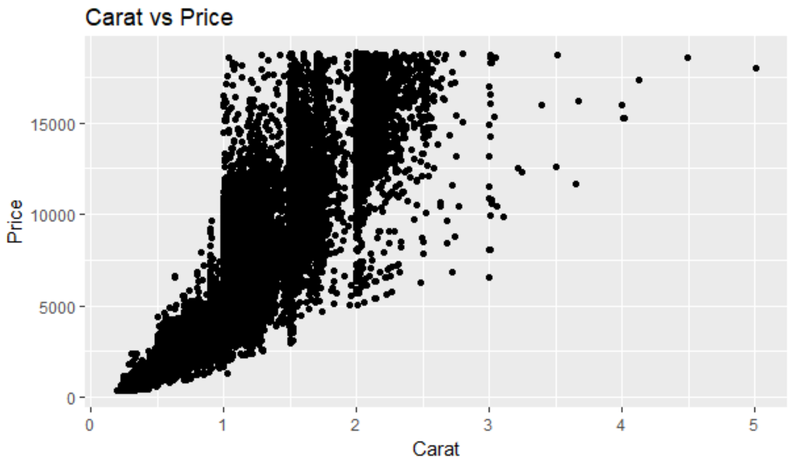
- **Relationships between carat, color and price**
  - Carat vs Color
    - It can be observed that (as seen in the color distribution above) J, being the rarest color, seems to have a higher median carat value.
    - The IQR for colors D,E,F, and G, seems to lie between 0.4 to 1 carat – the lesser pure diamonds whereas the other colors seem to have an upper quartile range up to 1.5 carats.



  - Price vs Color
    - The box plots for price vs color and carat vs color seem to be very similar.
    - Implies that colored diamonds with higher median carat value will have a higher median price and vice versa also holds true. This can be further noticed in the next plot between carat and price.



- Carat vs Price
    - As discerned in the above observations, the higher the carat value, the higher the price of the diamond
    - Price of the diamond increases exponentially with the carat value.



References

- https://stackoverflow.com/questions/47204046/ggplot-show-bin-length-on-bar-stat-bin-must-not-be-used-with-a-y-aesthetic?rq=1
- https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
- http://www.science.smith.edu/~amcnamara/NICAR2016.html
- https://www.stat.berkeley.edu/classes/s133/saving.html
- http://www.sthda.com/english/wiki/ggplot2-qq-plot-quantile-quantile-graph-quick-start-guide-r-software-and-data-visualization
- http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf