CSE6242 Spring 2017 - OMS
HW1: Data Visualization
GT account name: xtao41

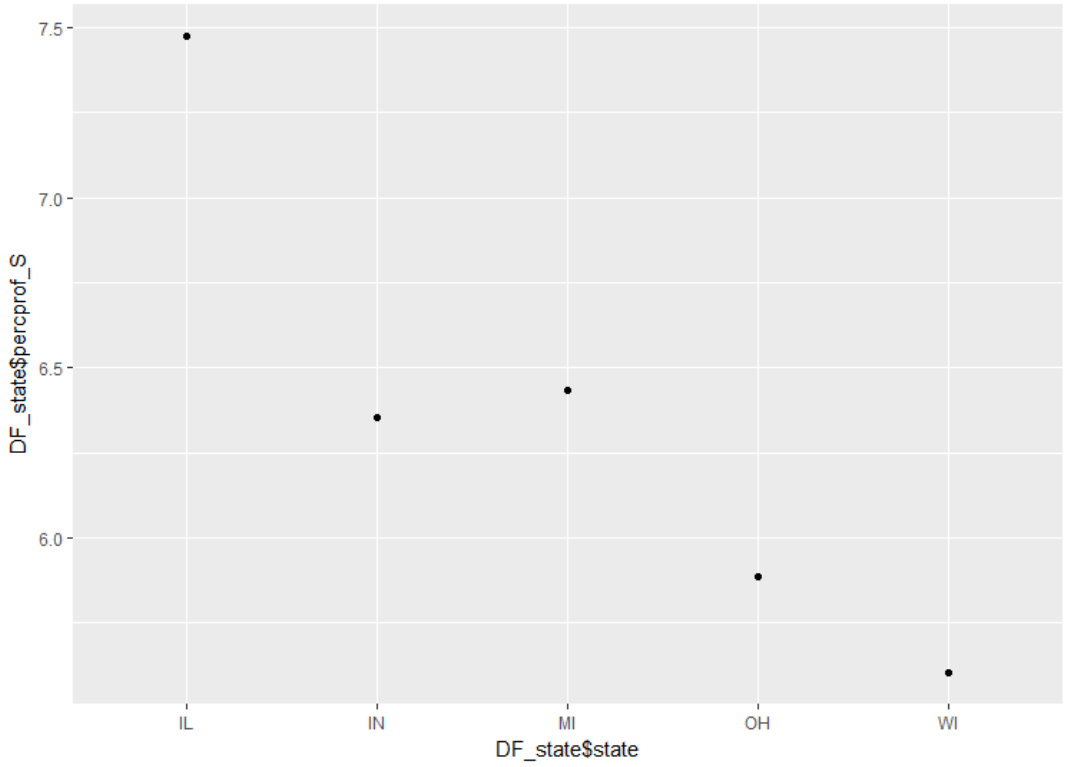
1. Professional Education by State

Calduation based on Aggregation (Interpretation A).

First, I aggregate procprof for each state by combing the vlaues from each county in that state and generate an overall "percprof by state" (percprof_S) value as shown in the table below, using the fomula:

$$\forall s \ \ percprof_s = \frac{\sum_{c \in s} percprof_c \times popadults_c}{\sum_{c \in s} popadults_c}$$

	state	N_prof_S	N_hsd_S	N_college_S	popadults_S	percprof_S	perchsd_S	percollege_S
1	IL	545188	5558141	1956244	7293930	7.474544	76.20228	26.82016
2	IN	221663	2639456	727658	3489470	6.352340	75.64060	20.85297
3	MI	375780	4485883	1406916	5842642	6.431679	76.77833	24.08013
4	OH	407491	5239876	1544480	6924764	5.884547	75.66866	22.30372
5	WI	173367	2432154	769147	3094226	5.602920	78.60299	24.85749

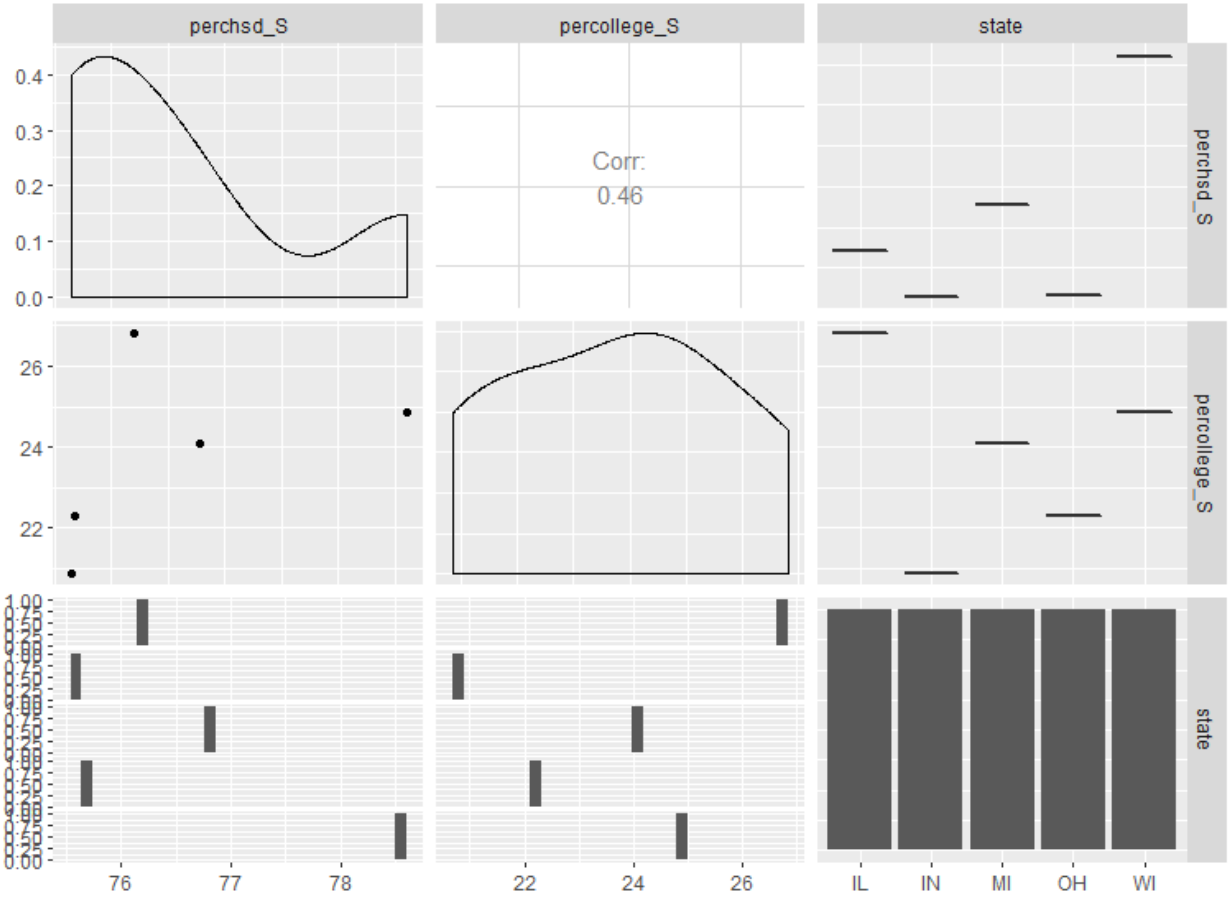


The relationship between the states and percprof by state is illustrated above. Clearly, IL has the highest percprof distribution, and WI has the lowest percprof distribution.

2. School and Colledge Education by State

Calduation based on Aggregation (Interpretation A).

In this method, perchsd and percolledge are aggregated in the similar way as in problem 1. The calculated (aggregated) values are in the table in problem 1, again based on popadults.



As shown in the above all-pair relation plot, we can draw the following conclusions:

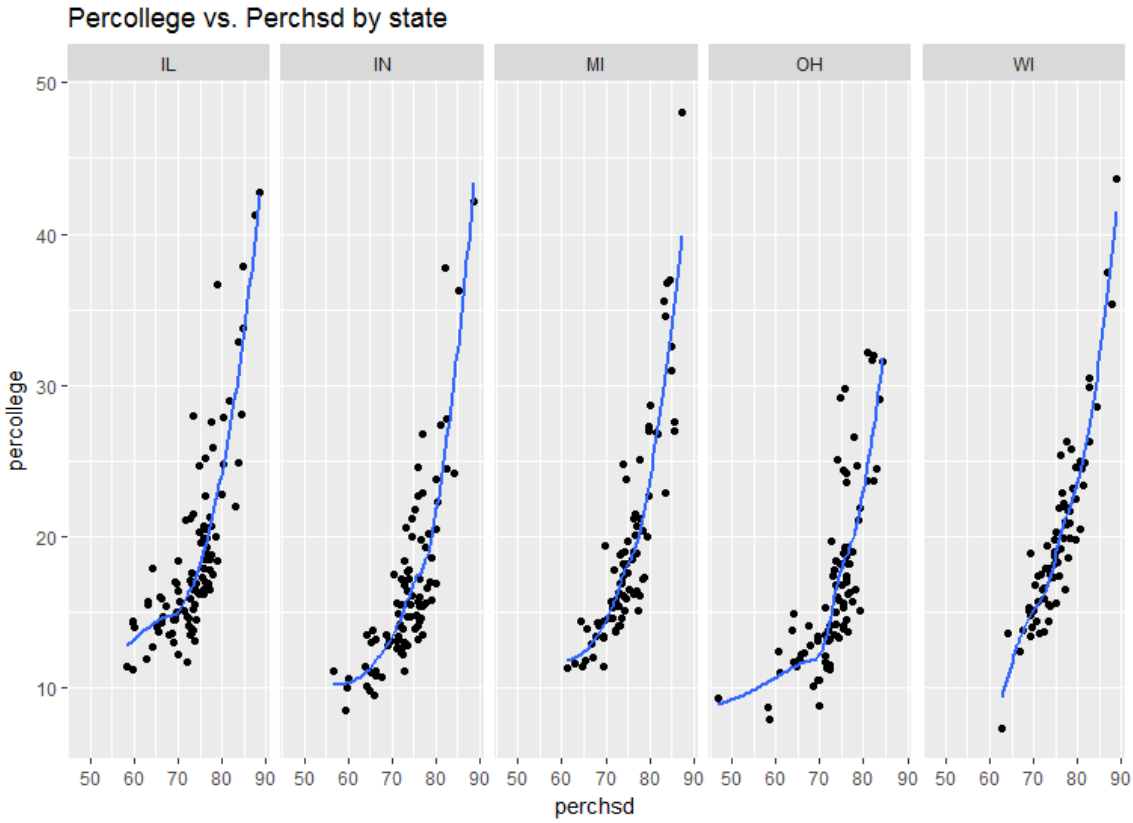
a) Perchsd vs. state: as illustrate in the upper-right figure, WI has the highest percentage of adult population with a high school diploma, and IN has the lowest (refer to the table as well).

b) Percollege vs. state: IL has the highest percentage of college educated adult population, and IN has the lowest. (right-middle plot)

c) Perchsd vs. precollege from aggregated data: I'm trying to draw conclusions based on the 5 pairs from the all-pair relation plot above (middle-left plot). It seems that as perchsd_S (perchsd by state) increase, precollege_S (percollege by state) increases. However, there is only 5 points in the graph, so trending is not that conclusive. Therefore, I pursue a different method, as in d) in the following.

d) **Perchsd vs. percollege within each state**: To illustrate this, I use the scatter plot of **raw data of perchsd vs. percollege** grouped by state (see below). We can see that the overall trending for each state is that as perchsd increases, percollge increases. Furthermore, from perchsd ~70% above, the relationship is almost linear, whereas for perchsd<70%, there are tails that indicating

a slower increase of precollege responding to perchsd. In this regime (perchsd<70%), OH has the slowest response of precollege (the flattest tail) and WI has the fasted response (the sharpest tail). While as perchsd >70%, it looks like the increasing rates of precollege as response to perchsd are not differed much (similar slope).



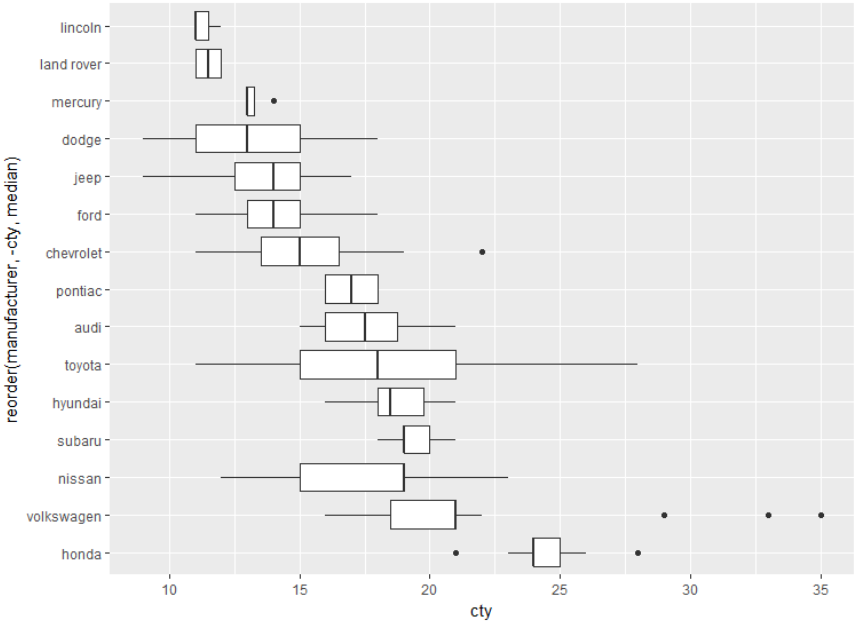
3. Comparison of Visualization Techniques
   1) Define the different elements of a Box Plot (see the plots below for reference).
      A box plot is composed of
      a) a box: denotes the IQR of the dataset
      b) an inner line bisecting the box: denotes the median of the dataset
      c) whisker(s): extends to the most extreme point no further than 1.5 times the length of the IQR away from the edge of the box.
      d) Outliers: graphed as separate points, which are data points outside the box and whiskers' range
   2) How the elements depend on the statistical properties of a sample of numbers.
      a) In general, the position of the inner line (median) is greatly affected by the distribution of the data. If the dataset is symmetric, the median may be in the geometric center of the box. On the other hand, if the data is skewed, the inner line may not be centered in the box.
      b) The width of the box is equal to the IQR value, which is the difference between 25% and 75% percentile. It depends on how spread out the 50% central data are. If the data are

more clustered to the median, then we have a narrower box; if the data is more dispersed, then we have a wider box.

c) The length of whiskers is dependent on the length of box and the edge of the boxes. The edges of the box depend on the first (Q1) the third percentile (Q3) of the distribution. The smaller Q1 or larger Q3, the longer the whisker; the larger the IQR, the longer the whisker.

d) The number of outliers depend on how many points are "too far" from the median in the dataset and are outside the overall distribution. These points are outside the whiskers, and tells they may not be reasonable data.

e) Here I illustrate how these elements depend on the sample by comparing different box plots from "mpg" dataset and plotting "cty" against "manufacturer" as shown below. For example "lincoln" has data cty={11, 11, 12 } with median=11 and is left skewed, thus the inner line lies at the left of the box and the whisker extends to the right. Box width= IQR=11.5-11=0.5 and whisker length=1.5*0.5=0.75. "12" is outside IQR range (>11.5) and therefore there is a whisker to the right. No data outside 11-0.75 or 11.5+0.75 and so no outliers. In contrast, "land rover" has data cty={11,11,12,12} with median=11.5 and is symmetric. Thus we see the inner line is right in the middle of the box, and box length=1. Every point is inside the box and therefore there is no whisker nor outliers. Another example is "Volkswagen", the distribution of the date is skewed to the right. The median is around 21, with the central 50% of the data falling within box less than the median and several points above 26. This results in the median line is on the right edge of the box and several high outliers.



3) Pros and cons of using a box plot or a histogram

a) Box plot:

- Pros: i) very useful for quickly summarizing the statistics (minimum, maximum, (spread), median, first and third quartiles, outliers, etc) of a distribution and quickly indicating whether the distribution is skewed and whether there are any outliers; ii) It is more convenient than histogram to compare data corresponding to different

values of a factor variable (See the side-by-side box plots of city mpg for different make of vehicles above); iii) Box plots handles extremely large datasets easily; iv) by separating outliers from the majority of data (box and whiskers), helps us better understand the sample because it eliminates a few extreme non representative values, making the dataset cleaner to interpret.

- Cons: i) box plot cannot tell as much details of distribution as histogram (not that visually appealing, and thus cannot depict the multimodal nature of the distribution that the histogram shows (see the histogram of eruption example below); ii) the exact values are not retained because they are hidden in the box/whiskers; iii) Not able to identify the mean and mode; iv) can be used only with numerical data

b) Histogram:

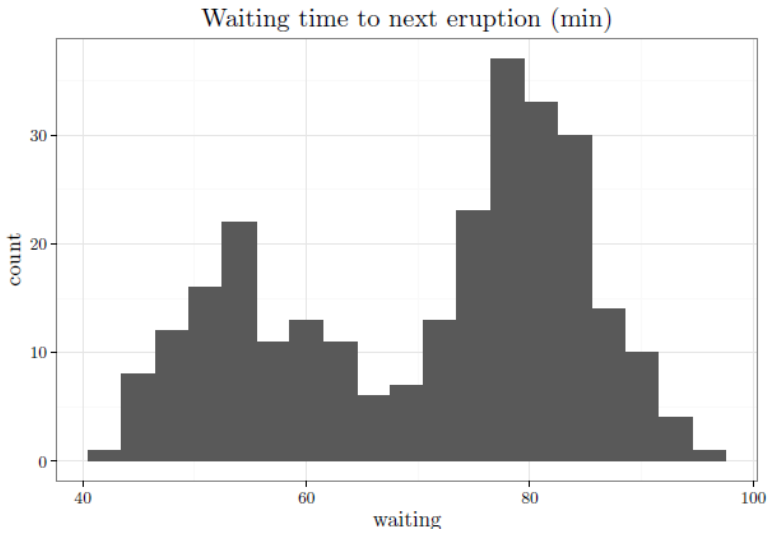
- Pros: i) very good at showing the rough distribution of values- visually strong and easy to determine the shape/mode of data; ii) it is useful and easy, apply to continuous, discrete and even unordered data; iii) It allows easy data transfer from frequency table to histograms. iv) especially useful when dealing with large values of ranges, whereas box plot may generate too many outliers.
- Cons: i) it is extremely difficult or impossible to extract the exact amount of data in the histogram because data is grouped into bins, unless it is a frequency histogram; ii) it is often inconvenient to compare multiple categories of data using histograms, because even you can compare several histograms side by side, it is hard to compare the trends between different categories and segments; iii) the level of details is greatly influenced by the width of bins. If a good bin width is not chosen, the representation might be hard to draw conclusion from; iv) histogram discard the ordering of the data points, and treat samples in the same bin as identical. v) use of lot ink and space to display very little information

4) When to use a Box Plot, a Histogram, or a QQPlot to graphically summarize a sample of numbers.

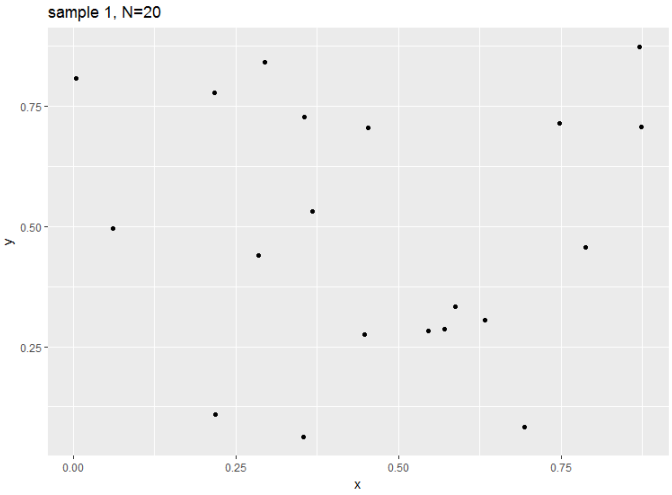
a) Histogram: Most useful with large ranges of inputs or when wide variances exist among the observed frequencies for a particular dataset. For example, if the data is multi-modal, histogram can show the different modes well, whereas box plot only looks roughly normal. Another instance when a histogram is superior over a box plot is when there is little variance among the frequencies observed. In this case, histogram can tell that there is little variance across the groups of data; however, box plot only shows the distribution looks roughly normal. Histogram is also useful with data from frequency tables.

b) Box Plot: More useful when we have multiple data groups from independent sources that are related to each other in some way and we want to compare them side by side. Another scenario that box plot is preferable is when there is moderate variation among the dataset. Plotting by histograms may cause it look ragged and non-symmetrical due to how data is grouped. However, box plot can show it correctly that the data is perfectly normal (non-skewed).

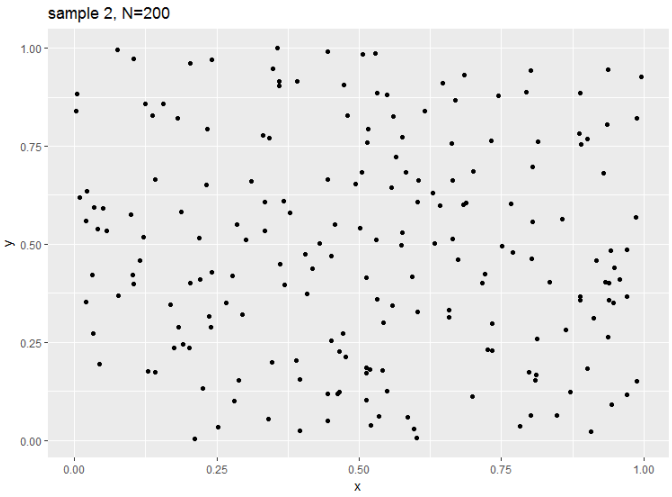
c) QQplots: Extremely useful for comparing two datasets, one of which may be sampled from a certain/theoretical distribution. It is most useful for diagnosing normality because it plots the empirical quantiles against theoretical quintiles.



4. Random Scatterplots
   (1) Sample scatterplots using randomly generated values,
      a) N=20, Save as PNG, File size= 39K.



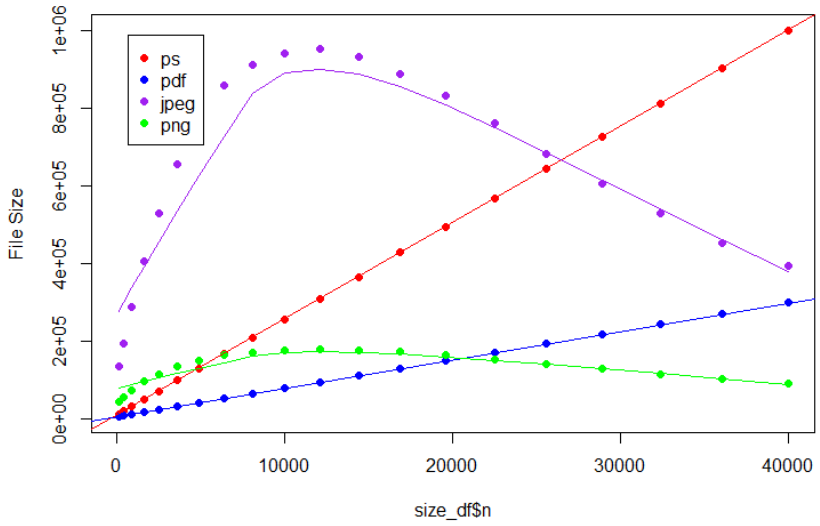
      b) N=200, Save as PNG, File size= 51K.



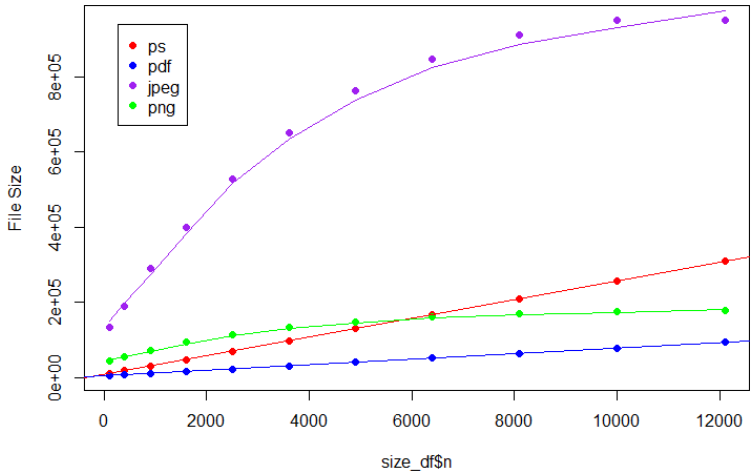
(2) Relationship between file size and N for each of the given file formats.
The data table and plots below (file size Unit: KB) show how file sizes changes as N increases for each format (ps, pdf, jpeg, png). The sampling of N is x^2

n	ps	pdf	jpeg	png
100	11575	5426	135457	43828
400	19012	7832	193069	56089
900	31378	11632	286406	72759
1600	48730	16850	405044	95413
2500	71042	23402	529958	115045
3600	98248	31585	654136	133777
4900	130487	41092	766404	150699
6400	167674	52128	857750	162992
8100	209720	64406	910629	170739
10000	256801	78392	940451	175974
12100	308777	93757	951616	178971
14400	365694	110678	931129	176462
16900	427691	128910	887000	172031
19600	494582	148781	832591	164004
22500	566547	169937	760178	153867
25600	643207	192548	682735	141101
28900	724970	216807	606280	128197
32400	811648	242304	528810	115049
36100	903490	269444	453824	101353
40000	1000026	298105	394491	89941

(n from 100 to 40000)



(n from 100 to 12000)



The conclusions are as follows:

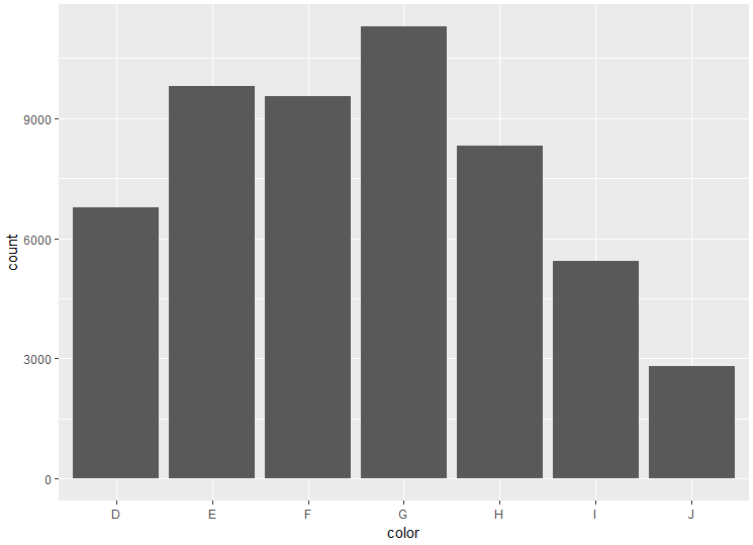
a) For the format of *ps* and *pdf*, the file size is proportional (linearly increase) to the number of values N. The *ps* format has a slope of ~24.8, whereas *pdf* format has a slope of ~7.3. Thus *ps* format has a much larger scaling rate of file size as N increases than *pdf* format. Overall, *pdf*'s file size is smaller than *ps*.

b) For the format of *jpeg* and *png*, they both increases non-linearly. When n<=~12100 (110^2), both format increases as n increases. However, when n>~12100, the file sizes of both format decrease as n increase. Moreover, *jpeg* format increases and decreases much faster than *png* format. Generally speaking, *jpeg* has a larger file size than *png*

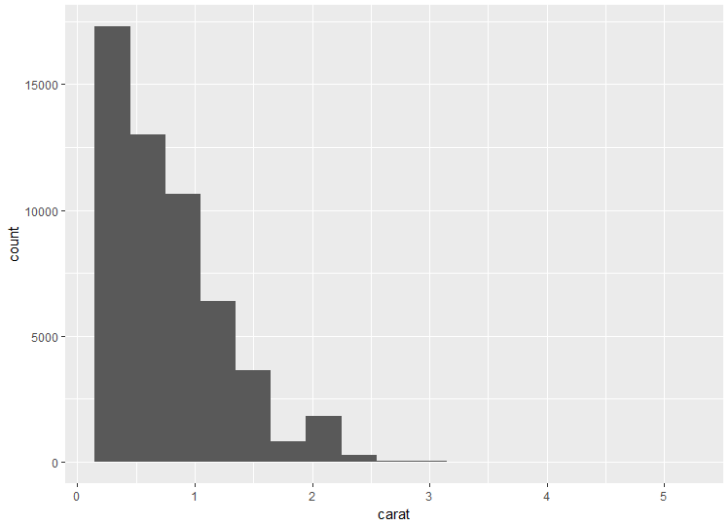
format. The polynomial fittings imply that *jpeg* has higher power increase/decrease than *png* when fitting by *y=a*(x^b)+…* (i.e. larger b)

c) My interpretation of the decreasing file size as N increases for *png* and *jpeg* is that they are compressed picture files, rather than documents. As the number of the random black points increases, the compression of image switched from white background to black background. Thus we could see a peak at somewhere when black and white spaces are nearly identical. There might be other reasons such as how the images compressed or stored, which are different from documents (*pdf* and *ps*).

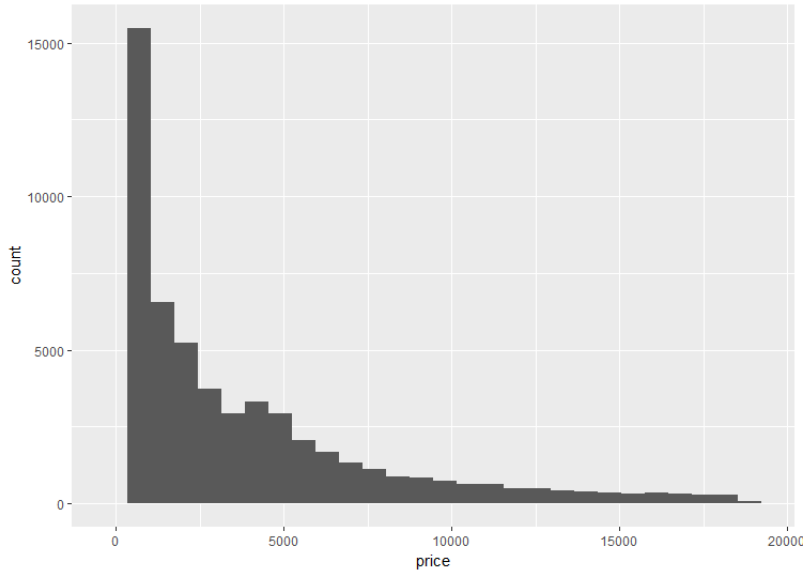
5. Diamonds
   1) Histograms for color, carat and price
      a) Color: from the bar chart below, we can see color G has the most counts and J has the smallest counts. It's hard to tell the shape of the distribution because color is an ordinal type. But it looks like a skewed normal distribution (bell-shaped).



      b) Carat: after trying different bin size, I think binwidth=0.3 best illustrate the distribution. The histogram shows a skewed distribution to the right (positively skewed). A much larger number of diamonds have small carat values than those having high carat values.

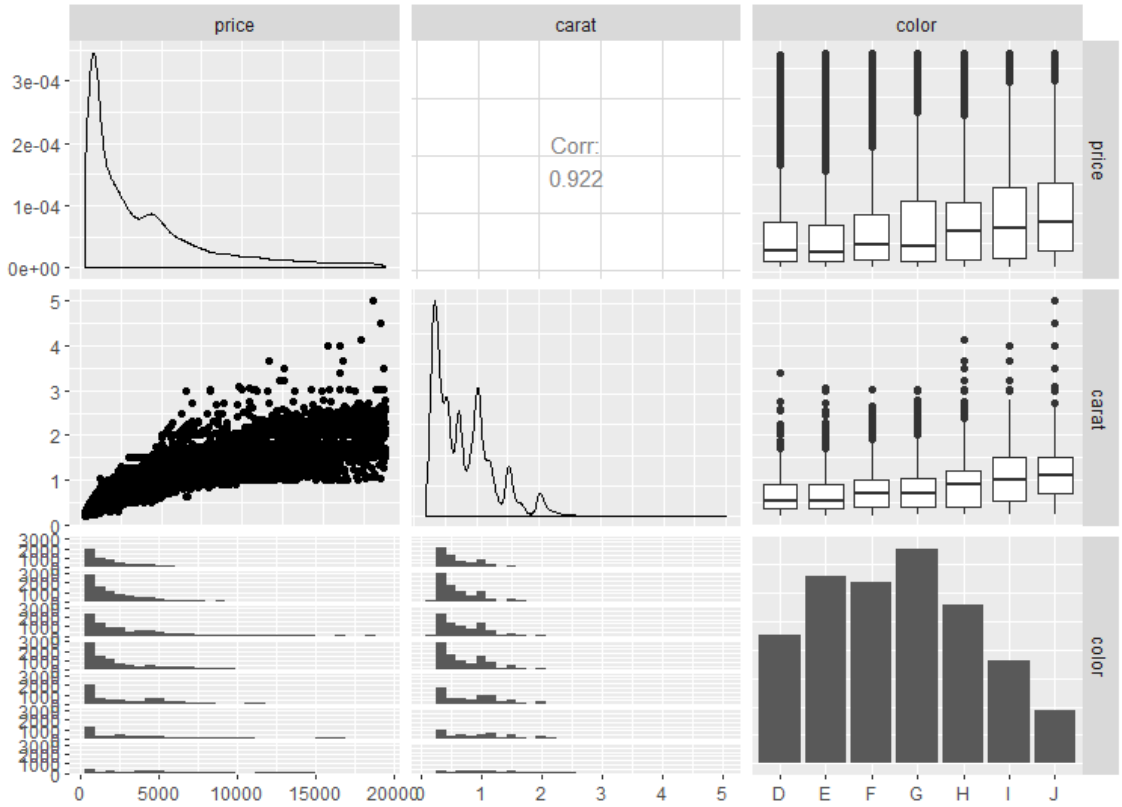


c) Price: here bin width=700 is used since it shows the distribution clearer. The histogram shows a positively skewed distribution. The majority of the price of diamonds are in the lower value cells (left side). The decay of the distribution from the left seems to be exponential.



2) Three-way relationship between price, carat and color.
The three-way relation plot implies that color J has the highest median of carat values and median of price values. Color J also has the largest spread of carat values. Color D and E has the lowest median price and median carat values. The spreads of prices are similar across different colors. The IQRs of price and carat from color I are both the largest.

The relationship between price and carat is close to linear, indicated by their correlation coefficient =0.922. That is, as the carat of the diamond increases, the price increases.



To better illustrate the relation between price and carat, I present a separate scatter plot between them as shown below. The most majority of data falls in 0-2.5 carat region, and in this region the correlation between carat and price is close to linear. From 2.5-5 carat, the sample volume is limited, the correlation is weak. It seems it this range, the increase of price as response to carat is much slower.