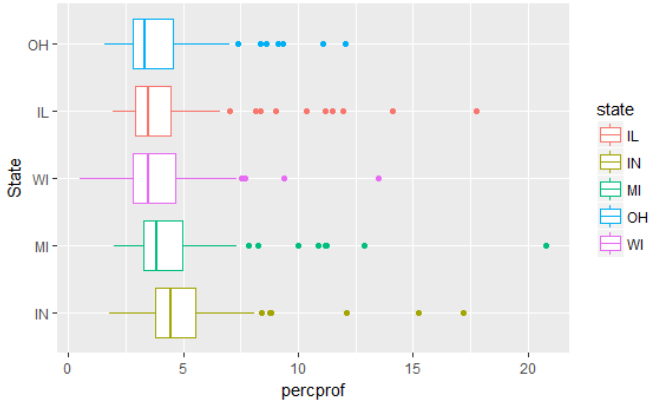
Title: hw2_report

Date: September 24, 2017

1. Professional Education by State [20 points]

Report:



	state <fctr>	median <dbl>	combined mean <dbl>
IL	IL	3.455354	7.474544
IN	IN	4.440127	6.352340
MI	MI	3.827592	6.431679
OH	OH	3.328012	5.884547
WI	WI	3.495100	5.602920

I used boxplot in ggplot2 to visualize the percentage of people that have a professional education for each county, grouped by state. The boxplot shows the median, IQR, spread and outlier of percentage of people having professional education in each state. The middle line in box of this boxplot shows the median value for each state. As shown from the figure above, IN has the highest percentage of population with a professional education, and OH, IL and WI have the lowest median value of percentage of population with a professional education. OH, IL and WI have very similar median value, and OH has a slightly lower median value compared with the other two states, indicating that OH has the lowest percentage of population with a professional education. OH and WI have similar median value and IQR. WI has longer spread than that of OH. OH have more outlier points than that of WI.

The spread of MI is higher than the spread in all other categories. IL has 10 outlying counties, OH has 7 outlying counties, WI has 4 outlying counties, MI have 7 outlying counties, and IN have 5 outlying counties.

I also wrote a small piece of code to calculate the combine mean, median of percentage of people having professional education in different state. The combined mean is calculated by the total number of adults that have a professional education in each state/ the total number of adults in each state.

As shown in the table above, IL has highest percentage of population with a professional education in combined mean (combined mean = 7.474544 for IL) and WI has the lowest percentage of population with a professional education in combined mean (combined mean = 5.602920 for WI). IN has the highest percentage of population with a professional education in median value (median = 4.440127 for IN). OH has lowest percentage of population with a professional education in median value (median = 3.328012 for OH).

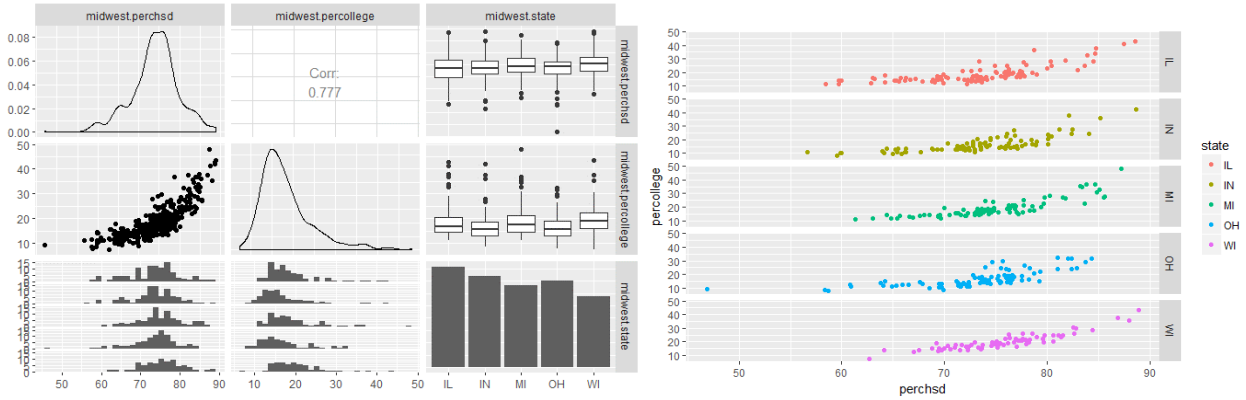
Code:

```
library(ggplot2) # load library ggplot2
data(midwest)
```

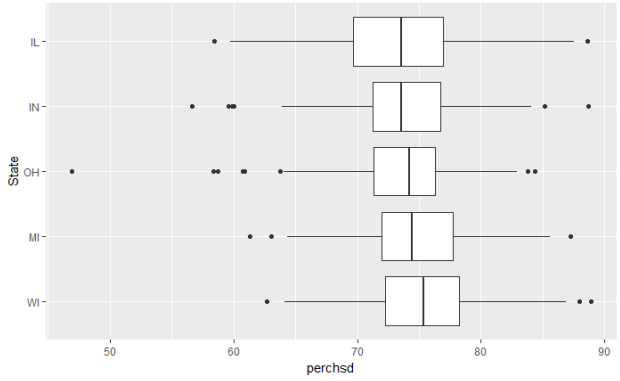
```
# use box plot to visilize the "percprof" of each tountry grouped by state
ggplot(midwest, aes(reorder(state, -percprof, median), percprof,color=state)) +
geom_boxplot() +
coord_flip() +
scale_x_discrete("State")

# calculate combined mean of "percprof" each state
# find all the states in column midwest$state
state_unique=unique(midwest$state)
Comined_mean=c(1:length(state_unique))
Median_Q1=c(1:length(state_unique))
names(Comined_mean)=state_unique
names(Median_Q1)=state_unique
# generate a new column named "num_cprof_county" give initial value =0
midwest$num_cprof_county=0
# calculate the number of prople have professional education for each county
for (i in c(1:length(midwest$percprof))){
  midwest$num_cprof_county[i]=midwest$percprof[i]*midwest$popadults[i]
}
# calculate median and conbined mean for each state
j=1
for (i in state_unique){
  Median_Q1 [j]= median(midwest$percprof[midwest$state==i])
  Comined_mean [j] = sum (midwest$num_cprof_county[midwest$state==i]) /
sum(midwest$popadults[midwest$state==i])
  j=j+1
}
stat_Q1=data.frame(state_unique,Median_Q1,Comined_mean)
names(stat_Q1)=c("state","median","combined mean")
stat_Q1
```

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

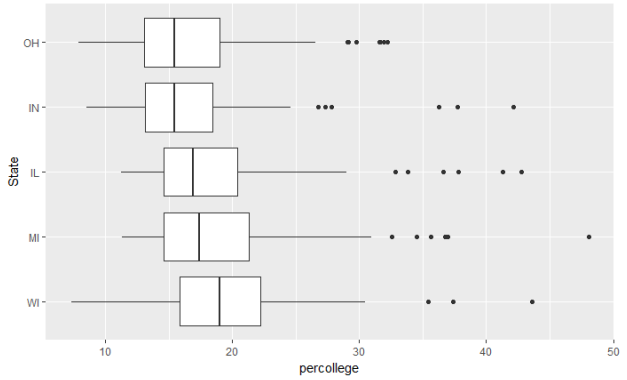
Report:



I explored the three-way relationship between "perchsd", "percollege" and "state" by combined pair-wise plot using ggpairs. Also, I made separate plots of perchsd vs. precollege grouped by each state. We can conclude from above plots that there is a positive nonlinear relationship between "perchsd" and "percollege". "perchsd" and "percollege" increase together nonlinearly (corr = 0.777, calculated by ggpairs).



The boxplot on column 3 row 1 in ggpairs and a separate box plot (perchsd vs. state) show the distribution of percentage of people with a high school diploma in each county ("perchsd") grouped by each state, and WI have the highest median value of "perchsd" among five states. IL and IN have the lowest median value of "perchsd". IL has a larger IQR than that of IN, and IN has more outliners than that of IL.



The boxplot on column 3 row 2 in ggpairs and a separate box plot (percollege vs. stat) show the percentage of college educated population in each county ("percollege") grouped by each state. The median values of "percollege" WI, MI and IL are larger than that of IN and OH. WI has the highest median value of "percollege"

among five states. OH and IN have the lowest median value of "percollege". The median value for OH and IN are similar, IN has larger spread than that of OH. OH has larger IQR than that of IN.

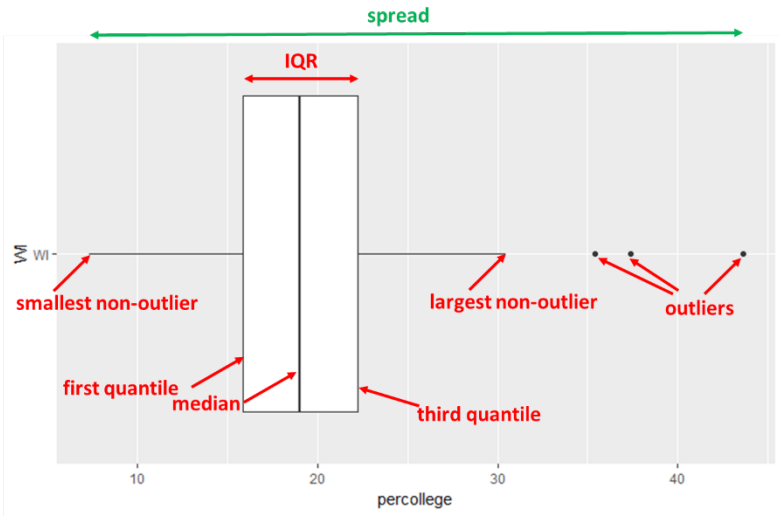
Code:

```
data(midwest)
# a combined pair-wise plot of perchsd,percollege,state in midwest dataframe
school_colledge=data.frame(midwest$perchsd,midwest$percollege,midwest$state)
library(GGally)
ggpairs(school_colledge)

# plot perchsd vs. state, percollege vs. state, perchsd vs. percollege using three different plots
# a boxplot of perchsd vs. state
ggplot(midwest, aes(reorder(state, -perchsd, median), perchsd)) +
geom_boxplot() +
coord_flip() +
scale_x_discrete("State")

# a boxplot of percollege vs. state
ggplot(midwest, aes(reorder(state, -percollege, median), percollege)) +
geom_boxplot() +
coord_flip() +
scale_x_discrete("State")

#plot perchsd vs. percollege
qplot(x=perchsd,y=percollege,data=midwest,xlab="perchsd", ylab="percollege",col=state, facets =state~.)
```

3. Comparison of Visualization Techniques [20 points]

Report:

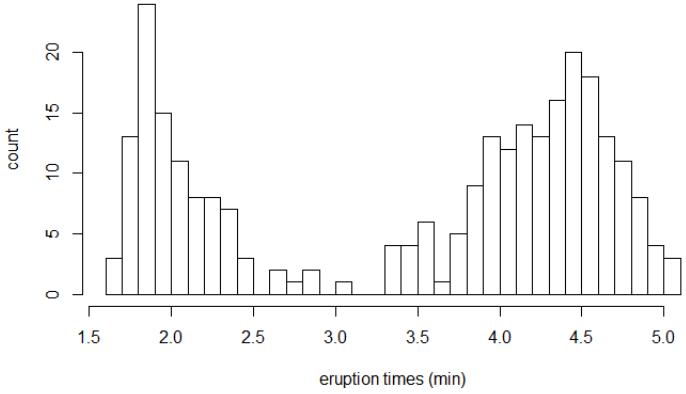
a). discribe different element of boxplot and how they illustrate different statistical properties of a sample.



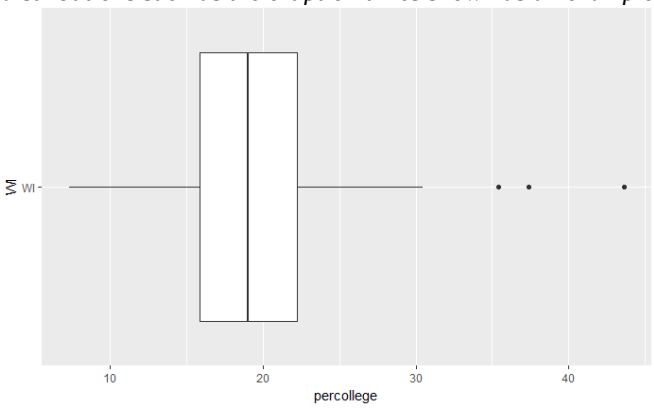
Boxplot shows several distribution features on the numbers in a dataset. As shown in the figure above, median stands for the value below which 50% of the data lie. The first and third quantile shows the value below which 25% and 75% of the data lie. The box between first and third quantile is called inner quantile range (IQR), which stand for the distribution of middle 50% data. The end of whiskers indicate the smallest and largest non-outlier. Additional dots in box plot stand for the outlier in this dataset. Also, the position median in IQR reflect whether the distribution of the data are symmetric or not. Spread stands for the range from the smallest number to the largest number, which is also shown in the boxplot.

b). When would you use a Box Plot, a Histogram, vs. a QQPlot

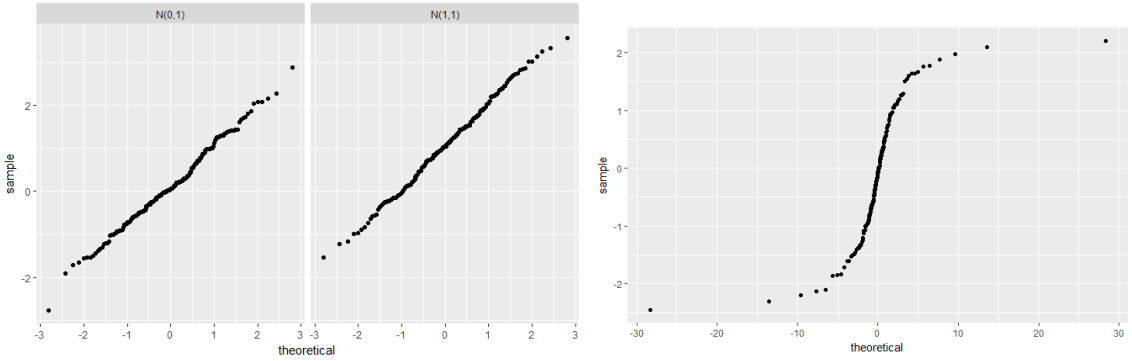
Histogram is an easy and quick way to observe the distribution of a set of numbers. The bins (breaks) in histogram decide how smooth the image looks like. Using large number of bins can see the details of distribution, but sometimes it is very hard to draw conclusion from the histogram with too many bins. On the other hand, using small number of bins can make image smoother. However, using small numbers of bins could cause the loss of important distribution information. Here is an example of histogram plotted from faithful dataset. This histogram shows that the "eruption times (min)" are distributed in two large populations (around 1.8 min and around 4.4 min).



Comparing with histogram, boxplot gives better statistical distribution details of a set of numbers. Boxplot shows the median value, first and third quartiles, IQR, spread and outlier of the data. It is more quantitative to compare the distribution difference of several sets of data by boxplot. Boxplot can see the "extreme" non-presentative values (outlier). Here is an example of boxplot of the "precollege" for each county in state WI from midwest dataset. The boxplot demonstrates that the first quantile, median, and third quantile value are around 16, 18, and 22, respectively. The distribution of inner 50% of the data is symmetric, and this dataset have three outliers. The major limitation of box plot comparing histogram is that box plot cannot deal with data set with multiple clusters of distributions such as the eruption times shown as an example for histogram.



qqplot is good at comparing the distribution of two datasets. One of the dataset can be a sample dataset and the other dataset reflecting theoretical distribution of certain statistical model. The slope and shape of the line in qqplot is related with the distribution and quantiles of a dataset compared with theoretical distribution.
For example, figure below on the left shows a straight line with slope =1 and this line passes the origin indicating that this sample have the same distribution as theoretical distribution. Figure below in the middle shows that the slope =1 but this line does not pass the origin, indicating that the sample dataset has very similar distribution shape and spread compared with theoretical distribution, but the values are shifted from theoretical distribution.
Also, S-shaped curve (figure below on the right) indicates that the distribution of dataset corresponding to the x-axis (theoretical dataset) has heavier tails than that of the sample dataset corresponding to the y-axis.



4. Random Scatterplots [20 points]

Report:

I have generated two sets of N random values using runif(N, 0, 100). These two sets of values were saved in random_1 and random_2. Then, I displayed the scatterplot of these two sets of values using random_1 as x-values and random_1 as y-values.

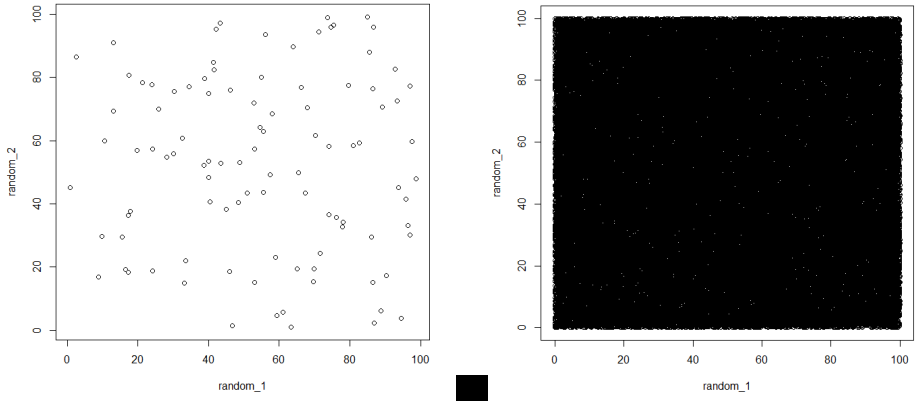
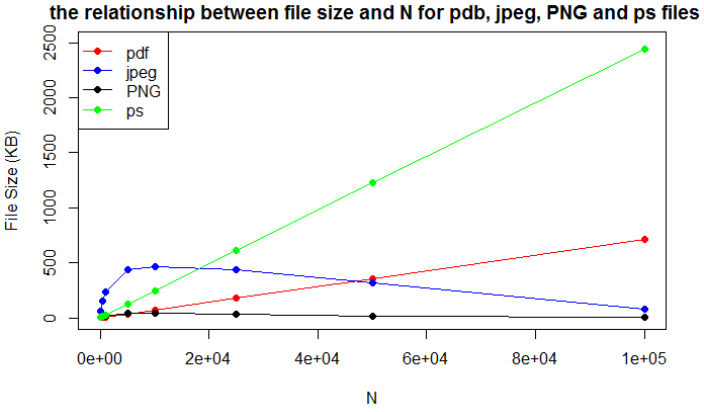


Figure on the left was generated using N = 100 and figure on the right was generated using N= 100000.

I generated 8 sets of pictures (N=c(100,500,1000,5000,10000,25000,50000,100000)) and saved in four different formats (pdf, jpeg, PNG and ps). The relationship between and N and image size for each format were plotted below. The size of pdf file (red) linear increase together with N. The size of ps file (green) also linear increase together with N, and the file size for ps is always larger than that of pdf. The size of jpeg file (blue) quickly increase when N increased from 100 to 5000, then it has a very small increase from 442 KB to 468 KB when N increased from 5000 to 10000. The size of jpeg file decrease from 468 KB to 78 KB when N increased from 10000 to 100000. Compared with other types of files, jpeg file has largest size when N = 100, 500, 1000, 5000. When N = 100000, the size of jpeg file is smaller than that of ps file and pdf file. PNG files always have small size no matter N is small or large. The size of PNG file increased from 6 KB to 44 KB when N increased from 100 to 10000.The size of PNG file decreased from 44 KB to 7 KB when N decreased from 10000 to 100000.



Code:

```
# N stands for number of observation
N = 100
random_1 = runif(N,0,100)
random_2 = runif(N,0,100)
plot(random_1,random_2)
# Type dev.new() in the console to open a new graphics window.
```

```
# Then run plot(random_1,random_2) in the console, save images in different formats
# To generate images with different N, give different N value, plot(random_1,random_2), and save images in different formats.

#generate data frame contain file size and N
N=c(100,500,1000,5000,10000,25000,50000,100000) # different N used for runif
pdf=c(3,6,10,38,74,180,358,715) # size of pdf image uder different N
jpeg=c(60,155,235,442,468,440,315,78) # size of jpeg image under different N
PNG=c(6,12,18,39,44,37,18,7) # size of PNG image under different N
ps=c(7,17,29,126,248,615,1225,2446) # size of ps image under different N
DF=data.frame(N,pdf,jpeg,PNG,ps) # combine all the information in a data frame, named DF

#plot the graph showing the relationship between file size and N
plot(DF$N,DF$pdf,type="o",pch=16,col="red",xlim=c(0,100000),ylim=c(0,2500),xlab="N", ylab="File Size (KB)", main=" the relationship between file size and N for pdb, jpeg, PNG and ps files")
points(DF$N,DF$jpeg,type="o",pch=16,col="blue")
points(DF$N,DF$PNG,type="o",pch=16,col="black")
points(DF$N,DF$ps,type="o",pch=16,col="green")
legend("topleft", c("pdf","jpeg","PNG","ps"),lty=1,col=c("red","blue","black","green"),pch = 16)
```

5. Diamonds [20 points]

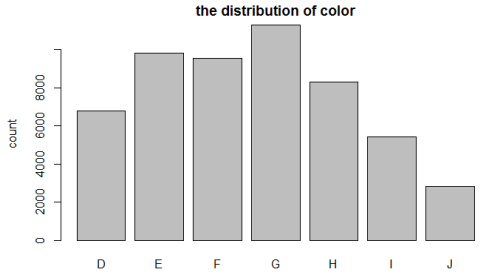
Report:

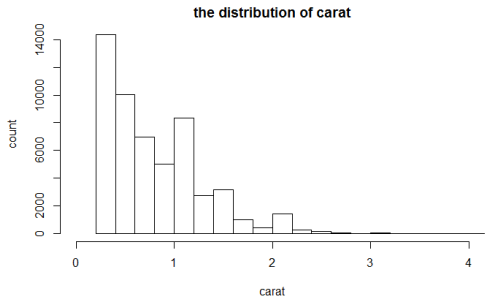
a) Separate distribution of color, carat and price:

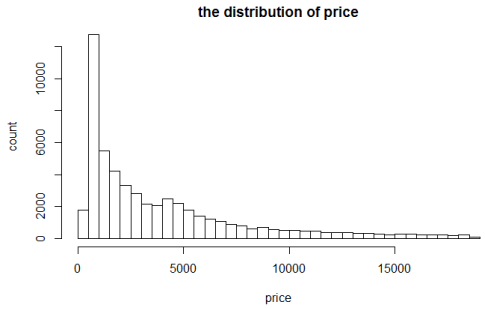
I plotted the bar plot for color and histogram for carat and price. Bar plot of color shows that there are 7 different type of color in this column. Among them, color "G" have the highest number of observation (larger than 10000). Color "J" have the lowest number of observation (less than 4000).

The histogram of carat shows the distribution of diamonds with different carat. More than 50% of diamonds are less than 1 carat. Using 20 breaks in this histogram, we can see that the highest "break" include diamond with carat between 0.2 and 0.4. The number of observation deceases as carat increase except bin for 1-1.2 carat. The overall distribution of carat is skewed to the right. The number of diamonds with carat larger than 1.6 is very low.

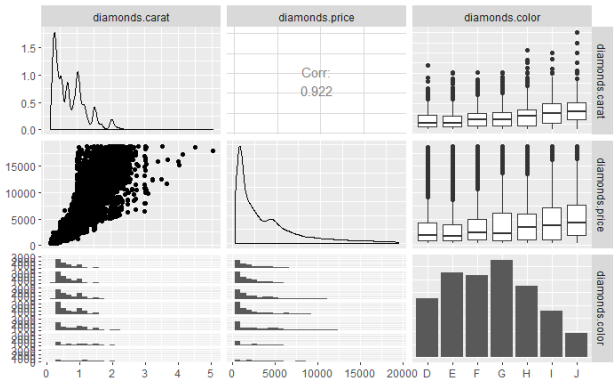
The histogram of price shows that the largest population of diamonds have price between 500 and 1000 (the second "break"). The number of observation deceases as price increase. The overall distribution of price is skewed to the right. The decrease is very slow when the price is higher than 3000.

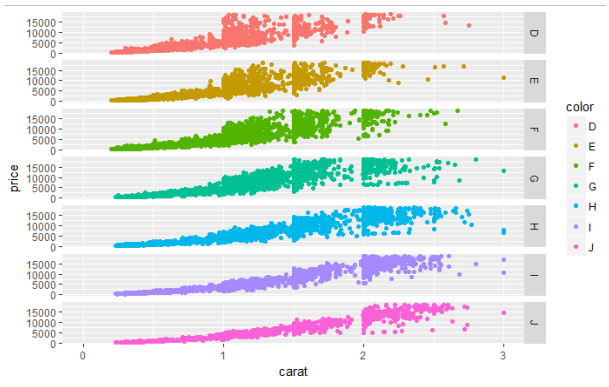






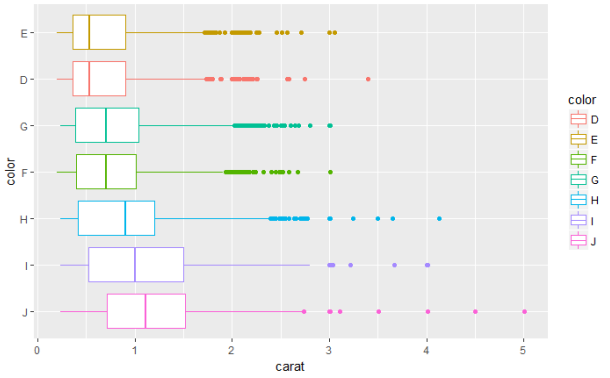
b). Three-way relationship between color, carat and price:

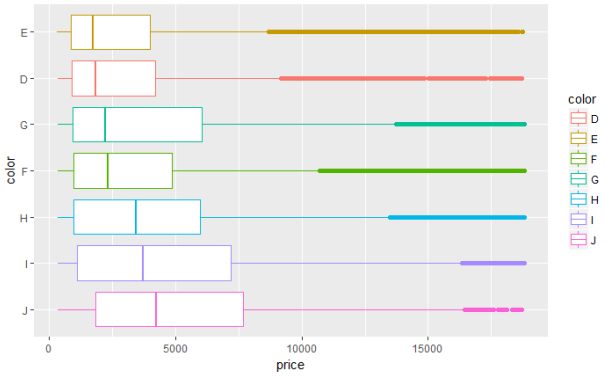




I used gg-pairs to study the three way relationship between color, carat and price. Three is a positive non-linear relationship between color and carat with the corr = 0.922 (figure above left column 1 row 2 and column 2 row 1).

I also plot the carat vs price grouped by each color (Figure above right). All the figures in different color shows the positive relationship between carat and price. The price of diamond increases along with the increase of carat. The color D group has the fastest price increase along with carat increase. The color H, I, J groups have the slowest price increase along with carat increase.
The relationship between carat and color is shown in figure above left column 3 row 1 and figure below on the left. Color J have the highest median value of carat, and color E have the lowest median value of carat. Color J have the largest spread in carat.
The relationship between price and color is shown in figure above left column 3 row 2 and figure below on the right. All of the boxplots grouped by different color have too many outliers. Color J have the highest median value of price, and color E have the lowest median value of price.





Codes:

```
data(diamonds)
head(diamonds)
#plot the individual distritbuion of color, price and carat
plot(diamonds$color, main= "the distribution of color",ylab="count")
hist(diamonds$carat,breaks=20, main="the distribution of carat", xlab="carat",xlim=c(0,4),ylab="count")
hist(diamonds$price,breaks=50, main="the distribution of price", xlab="price",ylab="count")
# three way relationship between price, color and carat using ggpairs
DF=data.frame(diamonds$carat, diamonds$price,diamonds$color)
ggpairs(DF)
# price vs carat grouped by color
qplot(x=carat, y=price,data= diamonds, facets = color~., col=color,xlim=c(0,3))
# carat vs color
ggplot(diamonds, aes(reorder(color, -carat, median), carat,color=color)) +
geom_boxplot() +
```

```
coord_flip() +
scale_x_discrete("color")
# price vs color
ggplot(diamonds, aes(reorder(color, -price, median), price,color=color)) +
geom_boxplot() +
coord_flip() +
scale_x_discrete("color")
```
