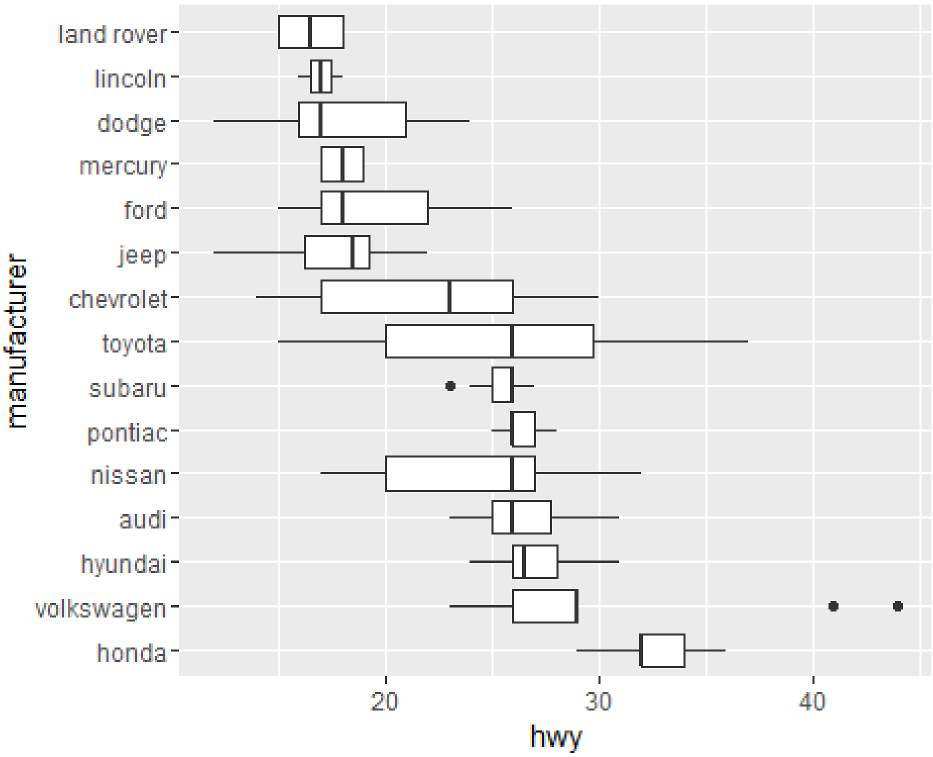
Johnpaul Ogah

August 30, 2016

Question 1

An ideal way to describe the relationship between highway mpg and car manufacturer, is to graph a box plot corresponding to changing mpg values with respect to the car manufacturer as shown in the figure below:

```
library(ggplot2)
ggplot(mpg ,aes(reorder(manufacturer, -hwy, median), hwy)) +
  geom_boxplot() + coord_flip() + scale_x_discrete("manufacturer")
```

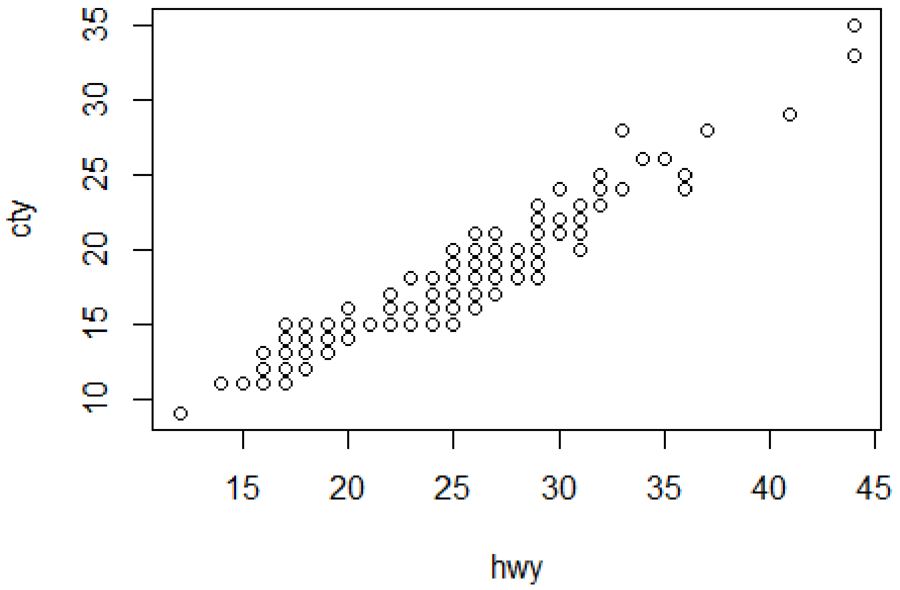


From the graph above, we deduce the following fuel efficiency order among car manufactures: 1)honda , 2)volkswagen with two outliers 3) hyundai 4) both audi, nissan,pontaic and toyota are tied.Toyota have a much larger spread from this group followed by nissan. subaru and pontaic have the shortest spread in this group with subaru having one outlier.5)Chevrolet 6)jeep 7)ford and mecury are tied though ford has the larger spread. 8) dodge and lincoln are tied but lincoln has a much more larger spead. 9) lad rover. Land rover is the least fuel efficient among all car manufacturer, honda is the most fuel effiecient and toyota have the largest spread.

Question 2

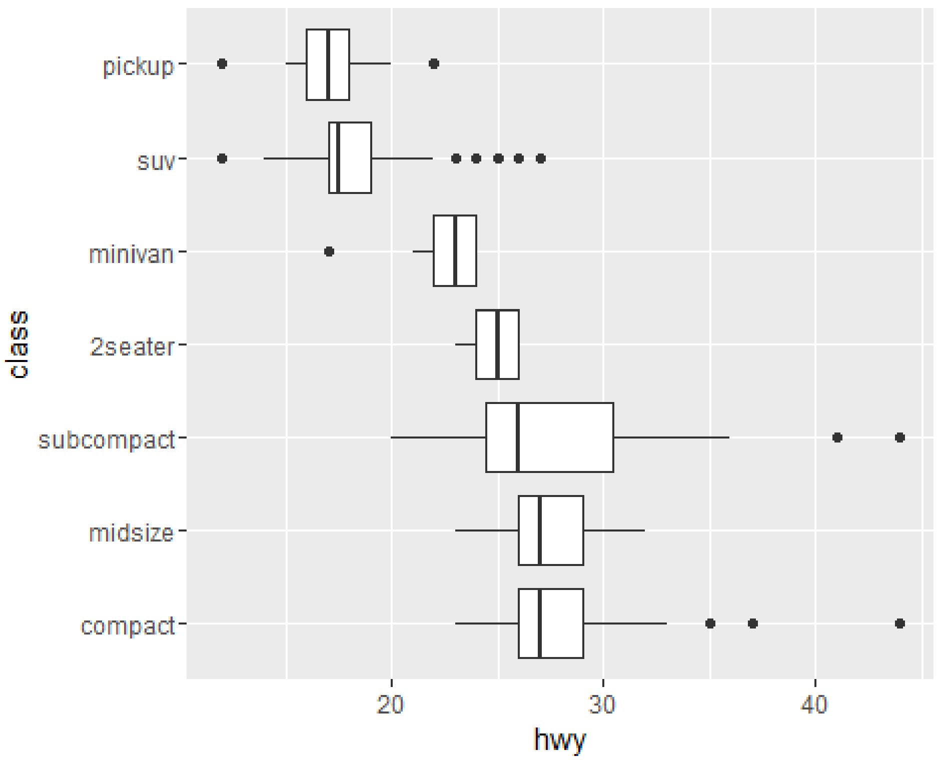
To describe the relationship between highway mpg, city mpg and model class can be done with the aid of a box plot.In this case we have two numeric variable and a factor variable.Lets use a scatter plot to see the relationship between the two numeric variable, as shown below:

```
plot(mpg$hwy,mpg$cty, xlab = "hwy", ylab="cty")
```



As seen in the graph above , highway mpg are generally higher than city mpg and the relationship between the two is linear.Now lets see the relationship between highway mpg and model class as whown in the plot below:

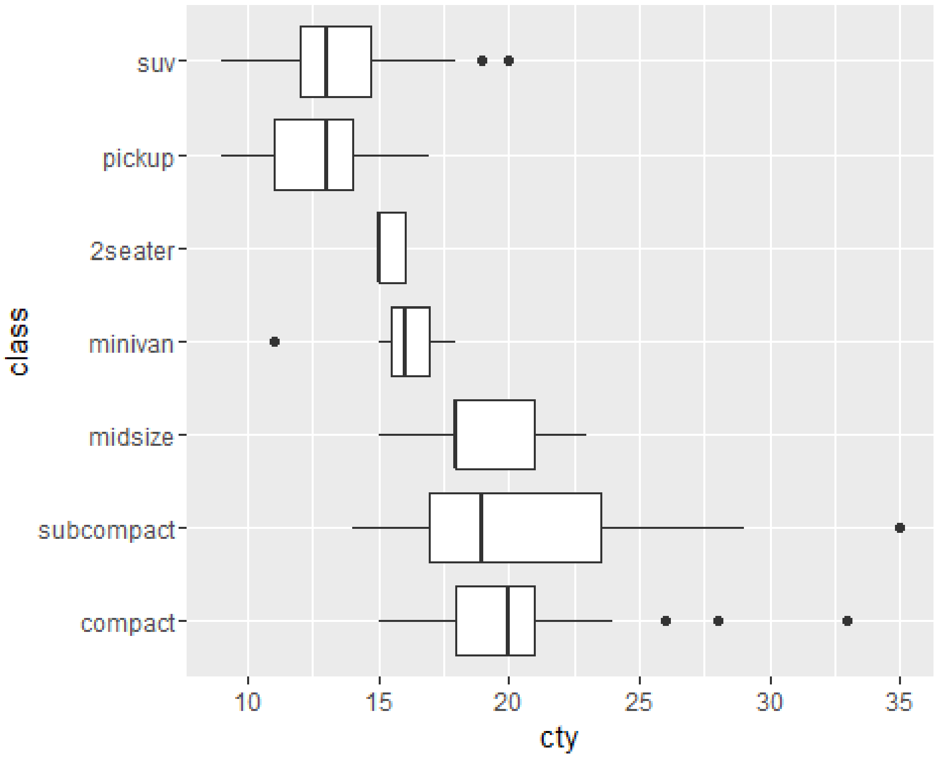
```
ggplot(mpg ,aes(reorder(class, -hwy, median), hwy)) +
  geom_boxplot() + coord_flip() + scale_x_discrete("class")
```



As seen from the plot above the order of fuel efficiency in highway with respect to model class are compact, midsize , subcompact, 2seater, minivan , suv and pickup .Subcompact has a much wider spread with two outliers. The box for both compact and midsize are identical with compact having three outliers.

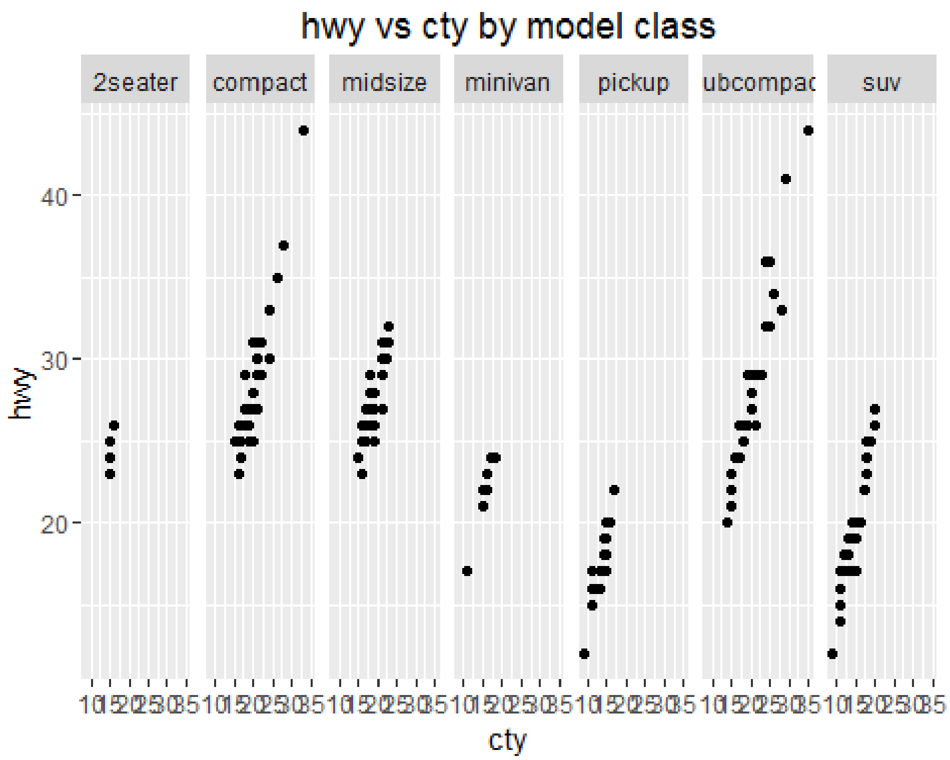
Now let’s look at the relationship between city mpg and model class as shown in the plot below:

```
ggplot(mpg ,aes(reorder(class, -cty, median), cty)) +
  geom_boxplot() + coord_flip() + scale_x_discrete("class")
```



As seen from the above plot , the fuel effieciency order are : compact , subcompact, midsize , minivan, 2seater, pickup and suv.Now lets observe the city and highway mpg side by side using the model class as the facets. This is shown in the graph below:

```
qplot(x=cty,y=hwy,facets = .~class,data=mpg, main= "hwy vs cty by model clas
s")
```



As seen from the graph above , highway mpg are generally higher than city mpg for the same model class.

Question 3

Histogram plot:

1. Histogram can be used to graph continous, discrete and unordered data
2. Mean and mode can easily be determine from a histogram plot.
3. Better suited for small data set
4. we cannot display several histogram at the same time for the purpose of comparison

Box plot:

1. Well suited for large data set.
2. Box plots provide some indication of the data's symmetry and skew-ness.
3. We can see outliers from the box plot.
4. We cannot determine mean and mode from box plot.
5. Not suitable for small number of data point

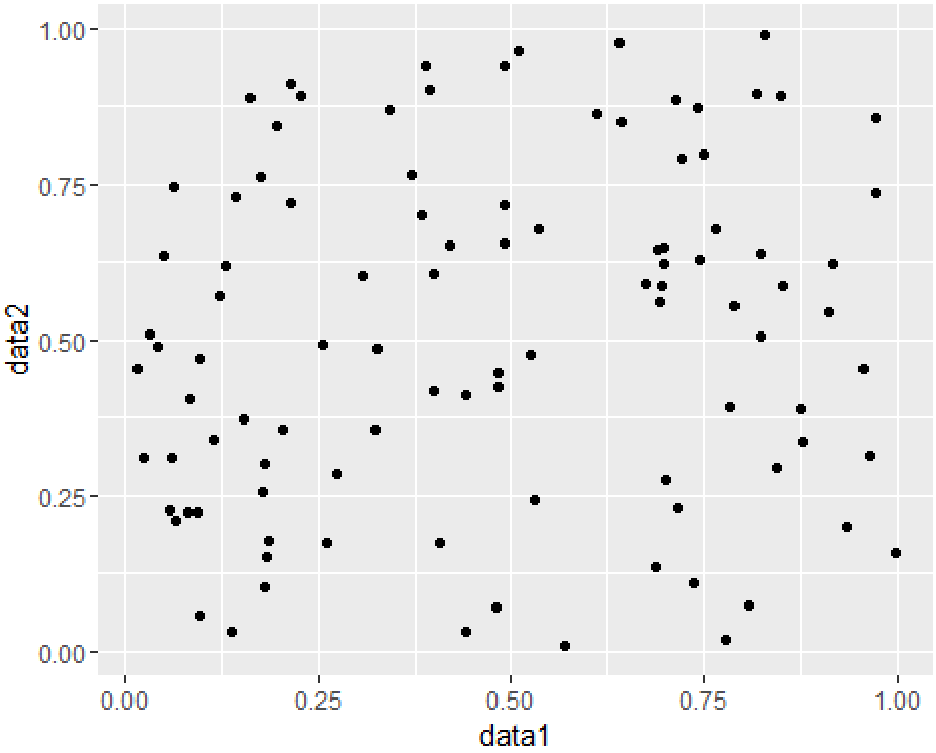
Question 4

```
data1 = runif(100)
data2 = runif(100)
postscript("mydata.ps")
plot(data1,data2)
dev.off()

jpeg("mydata.jpg")
plot(data1,data2)
dev.off()

png("mydata.png")
plot(data1,data2)
dev.off()

qplot(x=data1,y=data2)
```



Saving the plot in the file yield a size of 9179bytes, 17812bytes, 3610bytes and 6043bytes for ps, jpeg , png and pdf respectively. Now let's plot the size of the file for the various format in increasing values of N.

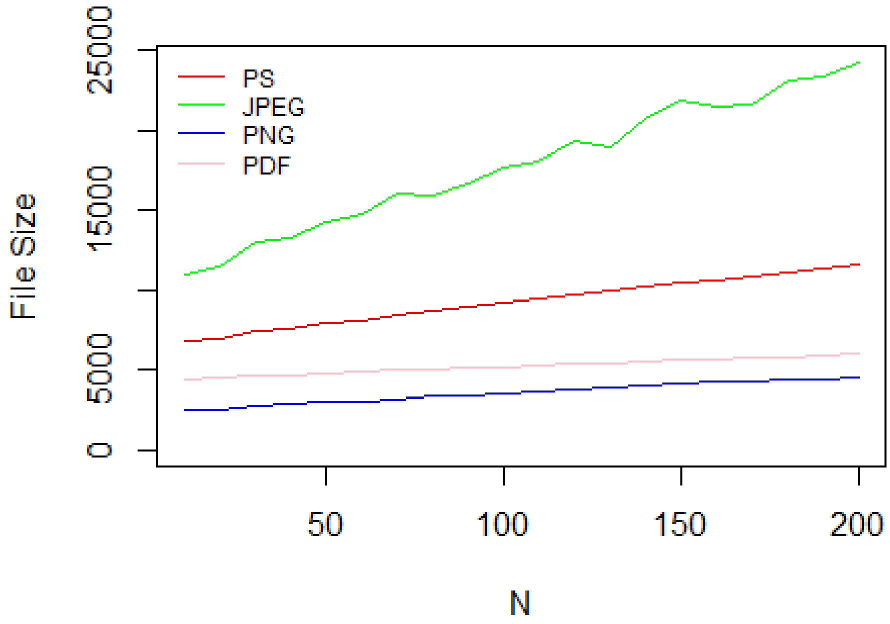
```r
list_ps = vector(mode="numeric", length = 20);
list_jpg= vector(mode="numeric", length = 20);
list_pdf = vector(mode="numeric", length = 20);
list_png = vector(mode="numeric", length = 20);
for (num in seq(10, 200 , by=10)){
  index = num/10;
  data1 = runif(num);
  data2 = runif(num);

#plot(data1,data2, xlab="random data 1" , ylab = "random data 2")
postscript("mydata.ps")
plot(data1,data2)
dev.off()
jpeg("mydata.jpg")
plot(data1,data2)
dev.off()
png("mydata.png")
plot(data1,data2)
dev.off()

pdf("mydata.pdf")
plot(data1,data2)
dev.off()
list_ps[index] = file.info("mydata.ps")$size
list_jpg[index] = file.info("mydata.jpg")$size
list_png[index] = file.info("mydata.png")$size
list_pdf[index] = file.info("mydata.pdf")$size
}
x = seq(10,200 , by = 10)

g_range =range(0, list_ps,list_jpg,list_png,list_pdf)
plot(x,list_ps,type="l",col="red",ylim=g_range,xlab = "N",ylab = "File Size")
lines(x,list_jpg,col="green")
lines(x,list_png,col="blue")
lines(x,list_pdf,col="pink")

legend('topleft', legend = c("PS","JPEG","PNG","PDF") ,
       lty=1, col=c('red', 'green', 'blue','pink'), bty='n', cex=.75)
```

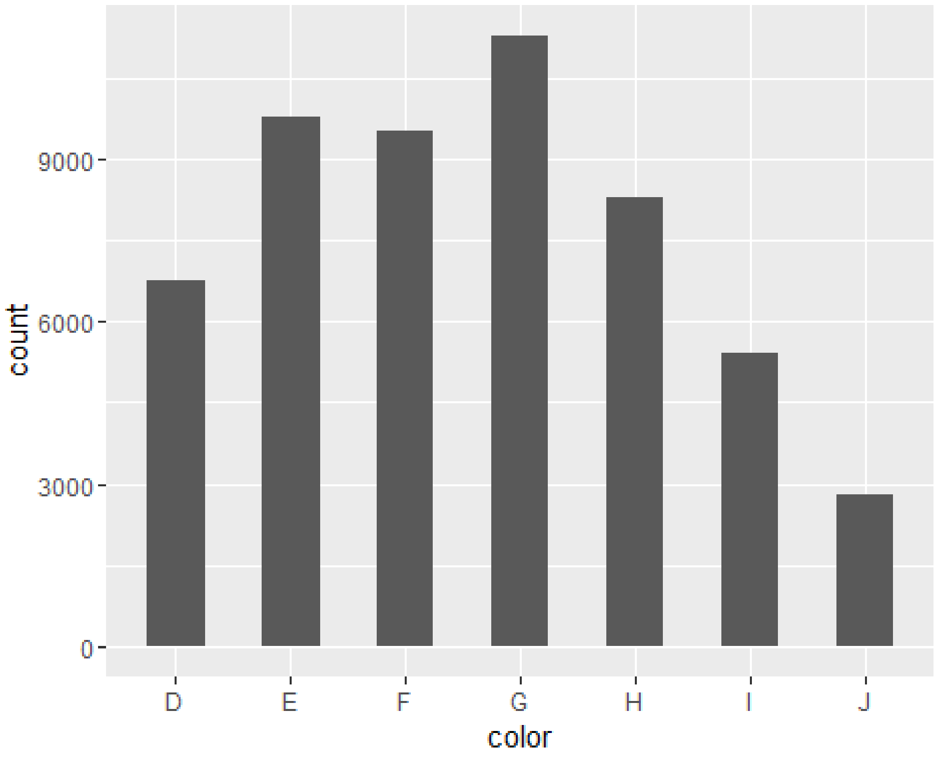


As seen from the graph , png and pdf scaled better with increasing values of N.JPEG scaled worst with increasing N increasing its file size at a much higher rate that the other three.

Question 5

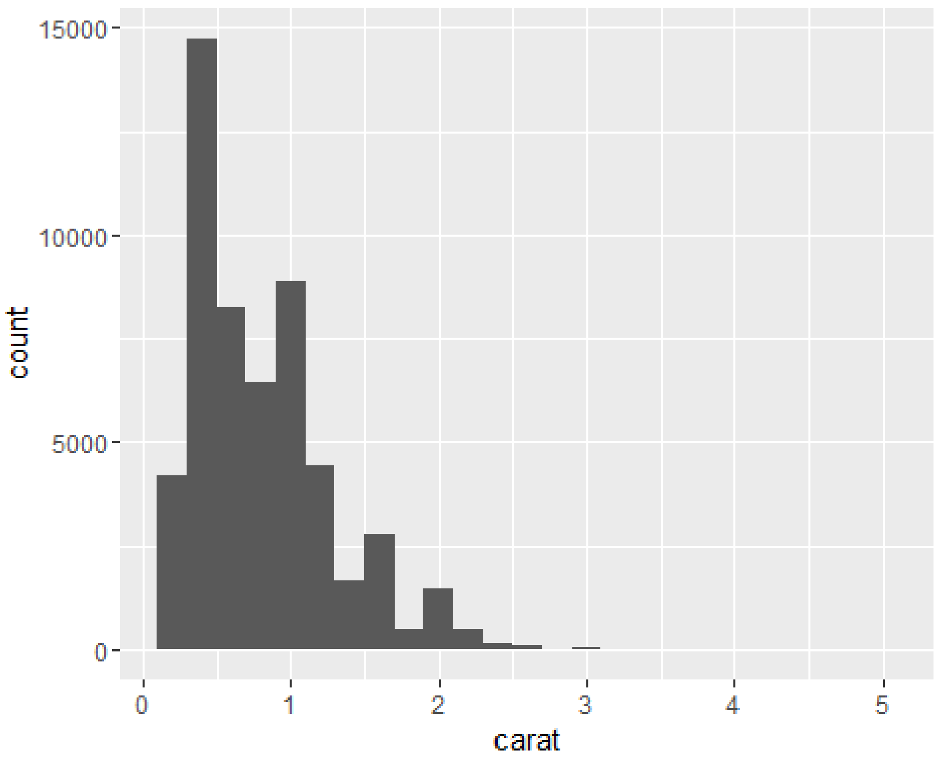
Histogram graph plot for color is shown below:

```
ggplot(diamonds , aes(x=color)) + stat_count(width = 0.5)
```



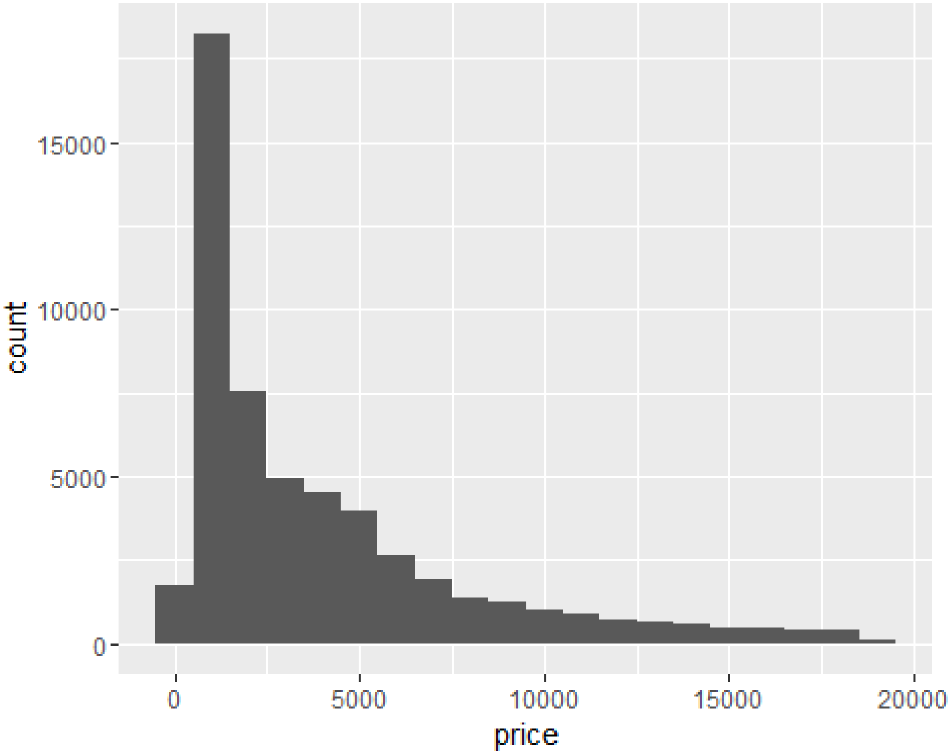
As seen from the graph above lease occuring color of diamond in the data set is color J and the most occuring color in the data set is color G.Now lets look at the histogram plot for carat.

```
ggplot(diamonds , aes(x=carat)) + geom_histogram(binwidth = 0.2)
```



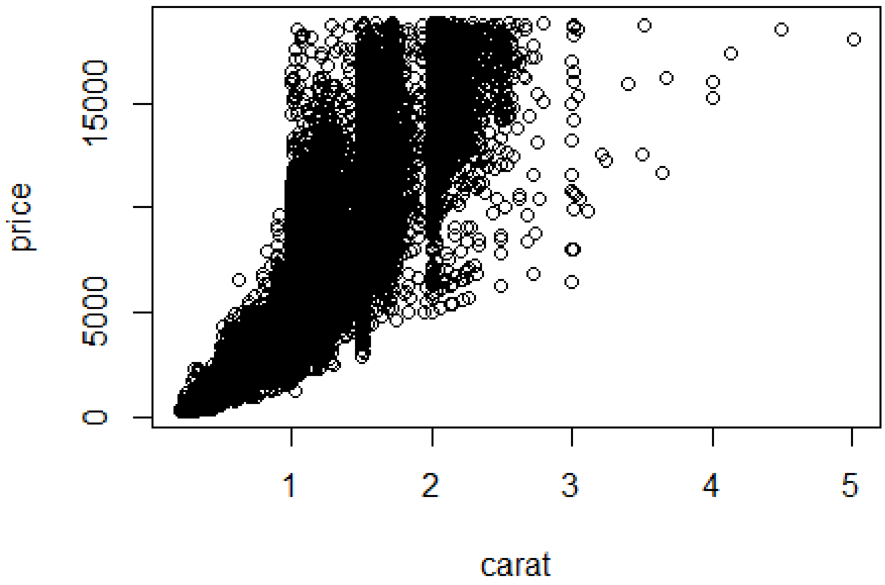
As seen from the graph most diamond in the data set has weight between 0.25 to 5.The histogram plot for various prices are shown below:

```
ggplot(diamonds , aes(x=price)) + geom_histogram(binwidth = 1000)
```



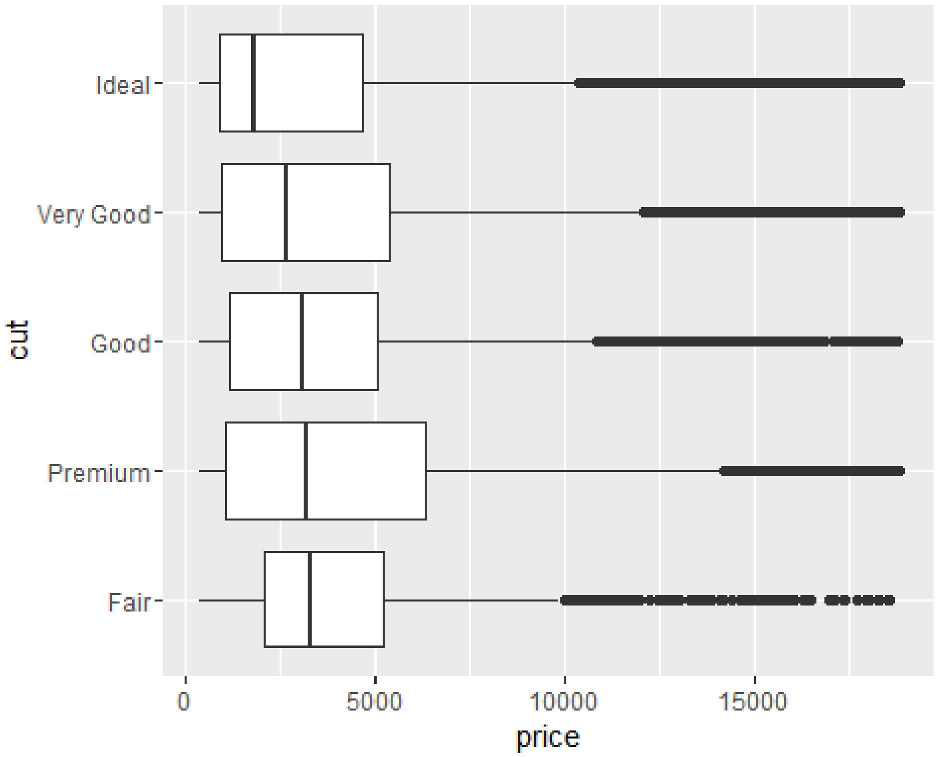
As seen from the graph above, most diamond in the data set cost between $750 - $1750.Let see the relationship between price and carat with the aid of a scatter plot.

```
plot(diamonds$carat,diamonds$price, xlab = "carat", ylab="price")
```



As seen from the scatter plot above there is a linear relationship between the weight of diamond (carat) and the price.Also looking at the plot carefully we can see that for the same weight there are different prices.Some diamond cost significantly higher than others even though their weight are the same, Therefore we can infer from this graph that there are more variables affecting the price .Now lets use a box plot to look at the relationship between cut and prize. This is shown in the figure below:

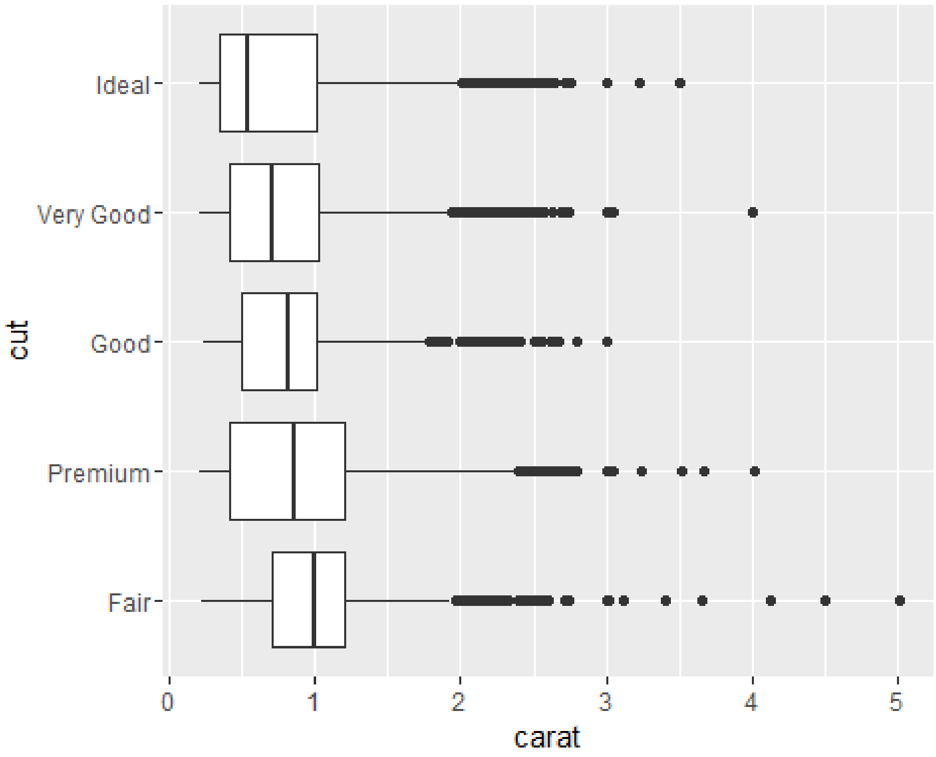
```
ggplot(diamonds ,aes(reorder(cut, -price, median), price)) +
  geom_boxplot() + coord_flip() + scale_x_discrete("cut")
```



from the graph above we see that price increase based on the cut int the following order : Fair , Premium , Good , Very Good and Ideal.Fair diamond has the highest price and Ideal diamond has the lowest prize on the average.

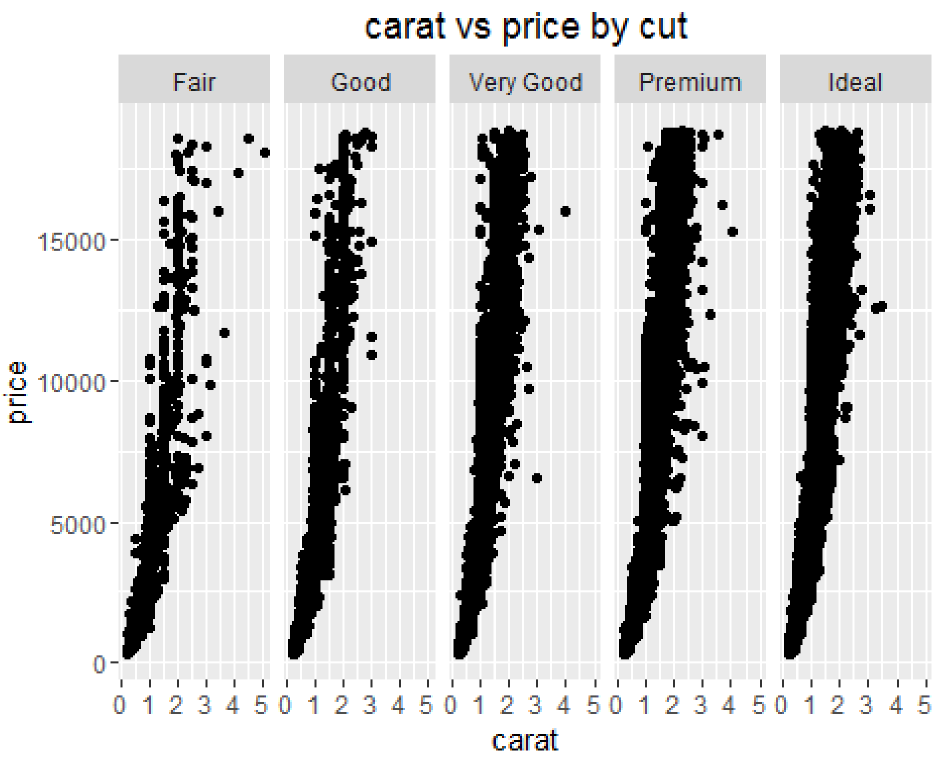
Now let's look at the relationship between cut and weight (carat) as shown in the box plot below:

```
ggplot(diamonds ,aes(reorder(cut, -carat, median), carat)) +
  geom_boxplot() + coord_flip() + scale_x_discrete("cut")
```



This is interesting, now we see that increase in weight follows the same order.Fair diamond which has the greatest prize according to the previous plot also has the highest weight on the average, similarly Ideal diamond also has the lowest average weight. Let's make the cut the facet and plot the carat and price side by side as shown below:

```
 qplot(x=carat,y=price,facets = .~cut,data=diamonds, main= "carat vs price by
cut")
```



As seen the graph above support the earliar conclusion that the cut and the weight together determines the price of a diamond.The higher the weight the higher the price and for the same weight Fair diamond have the highest price on the average and Ideal diamond has the lowest price on the average.