IS457_HW5_127

Class ID: 127

The grading is based on properties of good graph construction:
1.Data stand out
2.Facilitate comparison
3.Information rich
4.Vocabulary (in titles, axes labels, legend names etc)
The grading will be strict, since there are many elements in each plot. The total points for each questions is 15, 10 pts for plotting and 5 for explanation. But there are two bonus questions in the end.

Note: for interpretation questions, you won't get any points only describing the plots.
Use relevant technical terms (from lectures/slides) to EXPLAIN your findings/insights. e.g., for normal distribution, think about mean (center), sd(spread), skewness, outliers etc.
Unless we mentioned using external packages, stick with base R commands.

Part 1. Basic plots

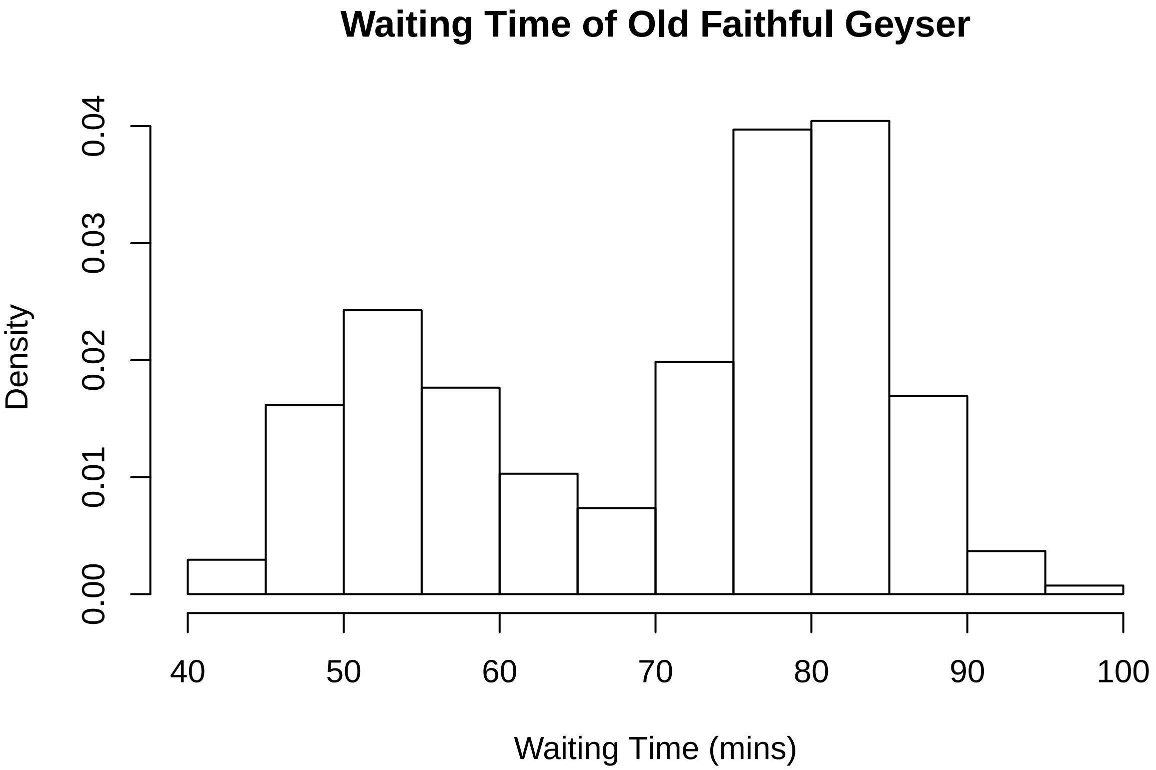
Q1. Show the shape of a distribution.

load the data set "faithful" (we've shown many times before how to load data in base R)

```
data(faithful)
```

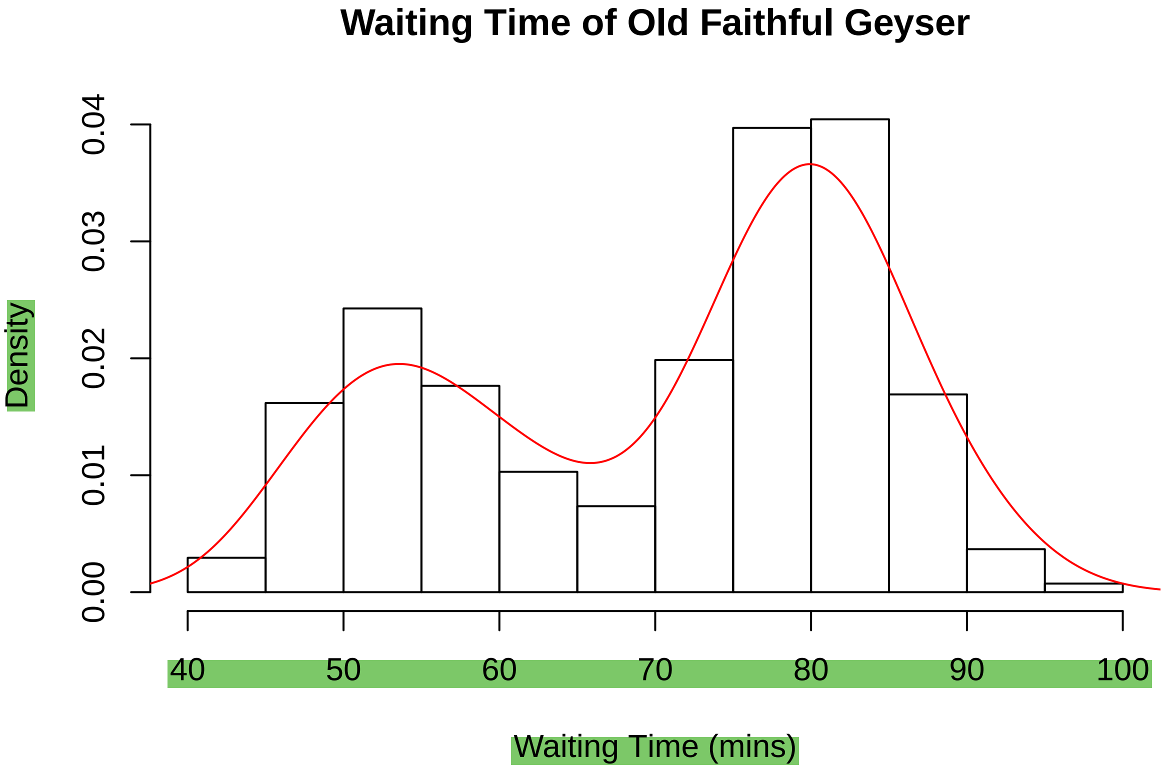
1), make a histogram that shows the distribution of variable "waiting".

```
hist(faithful$waiting, prob = T,main = "Waiting Time of Old Faithful Geyser",
     xlab = "Waiting Time (mins)")
```



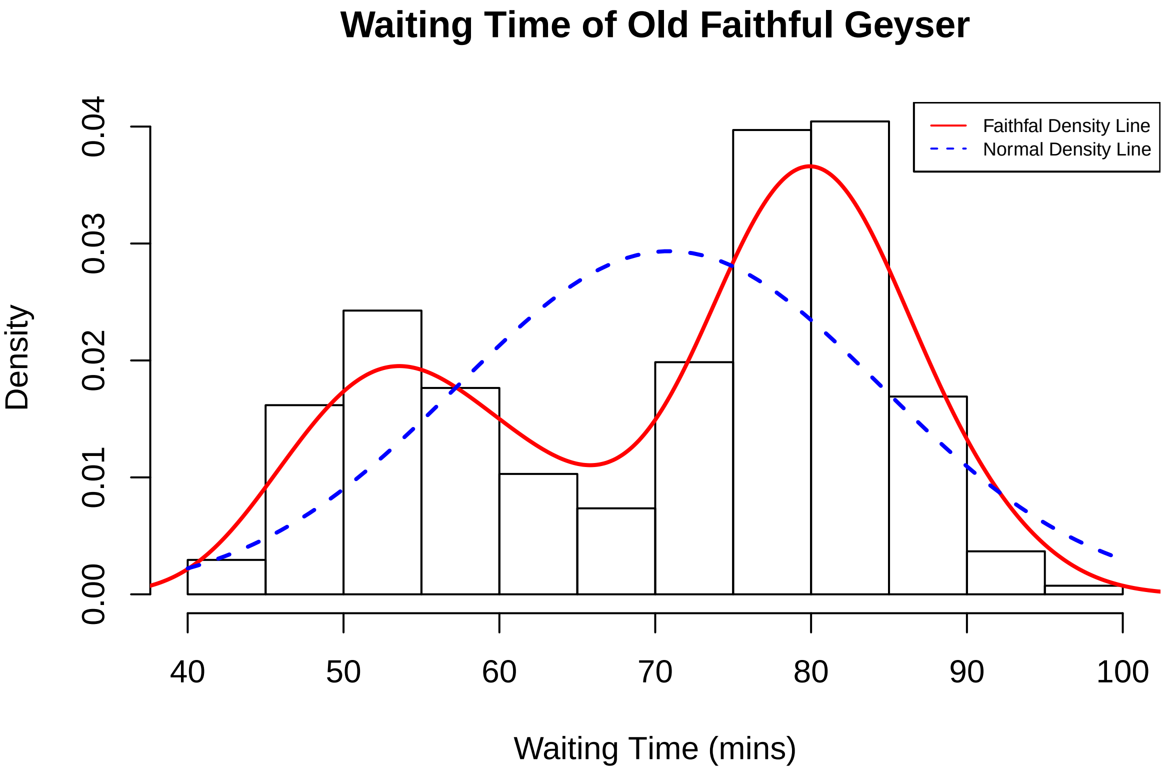
2), add the density curve of waiting. Hint: Adjust arguments of line() to make the line stand out.

```
hist(faithful$waiting, prob = T,main = "Waiting Time of Old Faithful Geyser",
     xlab = "Waiting Time (mins)")
lines(density(faithful$waiting),col= "red")
```



3), add a normal distribution curve on the plot with mean and standard deviation of waiting. Hint: curve() may help, also make the newly added line stand out.

```r
hist(faithful$waiting, prob = T,main = "Waiting Time of Old Faithful Geyser",
     xlab = "Waiting Time (mins)")
lines(density(faithful$waiting),col= "red",lwd = 2)
curve(dnorm(x, mean=mean(faithful$waiting), sd=sd(faithful$waiting)),
      add=TRUE,col="blue",lty=2,lwd = 2)
legend("topright", legend=c("Faithfal Density Line", "Normal Density Line"),
       col=c("red", "blue"), lty=1:2, cex=0.6)
```



What do you see from the histogram? what about after adding the density curve? and after imposing the normal curve?

bimodel distribution

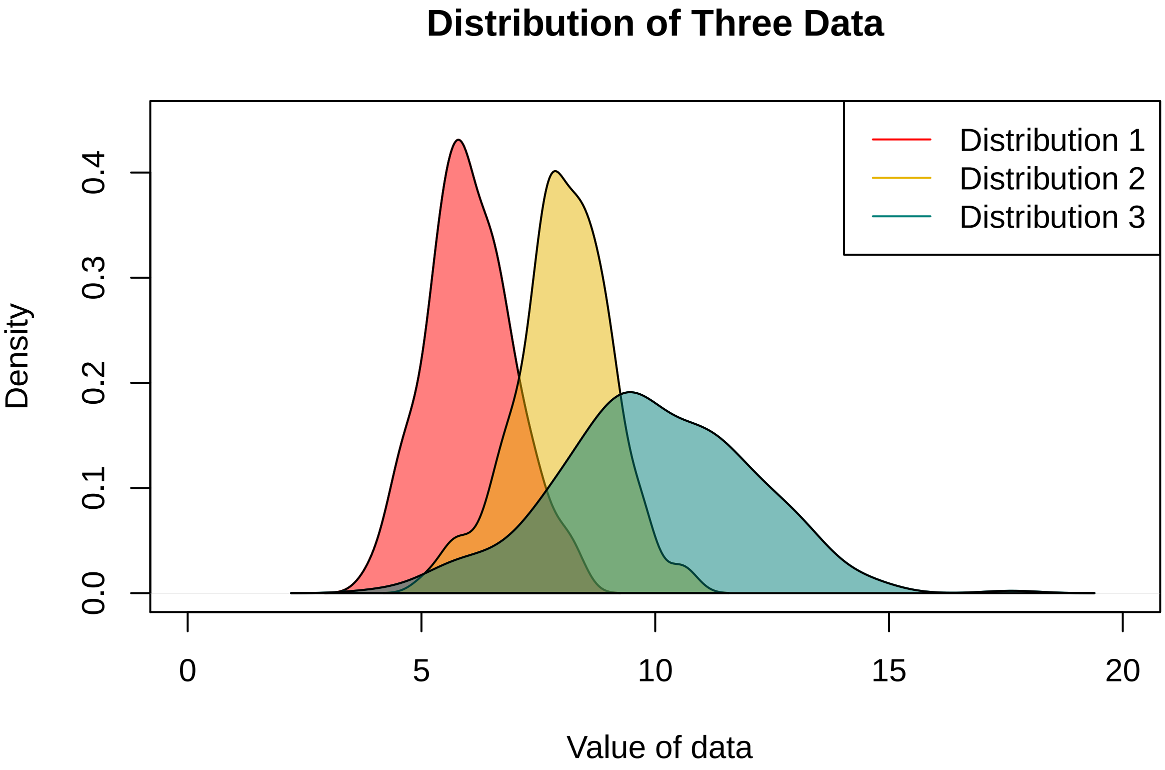
According to the histogram, there are two mode. The minor mode centers around 55, the major mode centers around 80. Thus, it seems the waiting time follow bimodal distribution.
The density curve shows the waiting time follow the bimodal distribution more clearly.
After imposing the normal curve, we can see that the waiting doesn't follow the normal distribution at all.

Q2. Comparing distributions.

generate 3 distributiuons with (sample size, mean, sd) = (200,6,1), (100, 8,1) and (300,10,2). plot them on the same graph, one color each distribution, with rainbow colors. Hint: rgb() function. If your choice of color scheme is correct, overlapping areas should have different/darker colors.

```r
set.seed(1)#set.seed in order to make the shape the same
dist_1<-rnorm(200, mean=6,sd = 1)
dist_2<-rnorm(100, mean=8,sd = 1)
dist_3<-rnorm(300, mean=10,sd = 2)
library(RColorBrewer)
```

```
den1<-density(dist_1)
den2<-density(dist_2)
den3<-density(dist_3)
                                       半透明
plot(den1,col=rgb(1, 0, 0),ylim = c(0, 0.45),xlim = c(0,20),
     main = "Distribution of Three Data",xlab = " Value of data")
polygon(den1$x,den1$y,col=rgb(1, 0, 0,0.5),fillOddEven =T)
lines(den2,col= rgb(230/255, 180/255, 0))
polygon(den2$x,den2$y,col=rgb(230/255, 180/255, 0,0.5))
lines(den3<-density(dist_3),col= rgb(0, 127/255, 120/255))
polygon(den3$x,den3$y,col= rgb(0, 127/255, 120/255,0.5))
legend("topright", legend=c("Distribution 1", "Distribution 2","Distribution 3"),
       col=c(rgb(1, 0, 0), rgb(230/255, 180/255, 0),rgb(0, 127/255, 120/255)), lty=1)
```



#####Comment on the shape of each distribution (effect of sample size, sd);

As we can see from the graph, first, the area under the distribution density line is area(distribution 3) > area(distribution 1)> area(distribution 2), this is because, as for the sample size, distribution 3 >distribution 1>distribution 2; Second, the shape of the distribution 3 is more flat (data points are spread out over a wider range of values) than distribution 1 and distribution 2. This is because the standard deviation of both distribution 1 and distribution 2 is 1, which is smaller than that of distribution 3.

what does the final plot look like, and explain why.

All of three distribution seem symmetric and bell-shaped, since they are generated by normal distribution function; Second, data point of distribution 3 tends to larger than data point of distribution 2, and the data point of distribution 1 is the smallest, this is because we generate these data follow mean(distribution 3) > mean(distribution 2)> mean(distribution 1); Third, three distributions all have their unique color and the area under the density curve also filled with the same color. Since I set the color in the area to be transparent by `rgb(..,...,..,0.5)`, their overlapping areas are shaded.

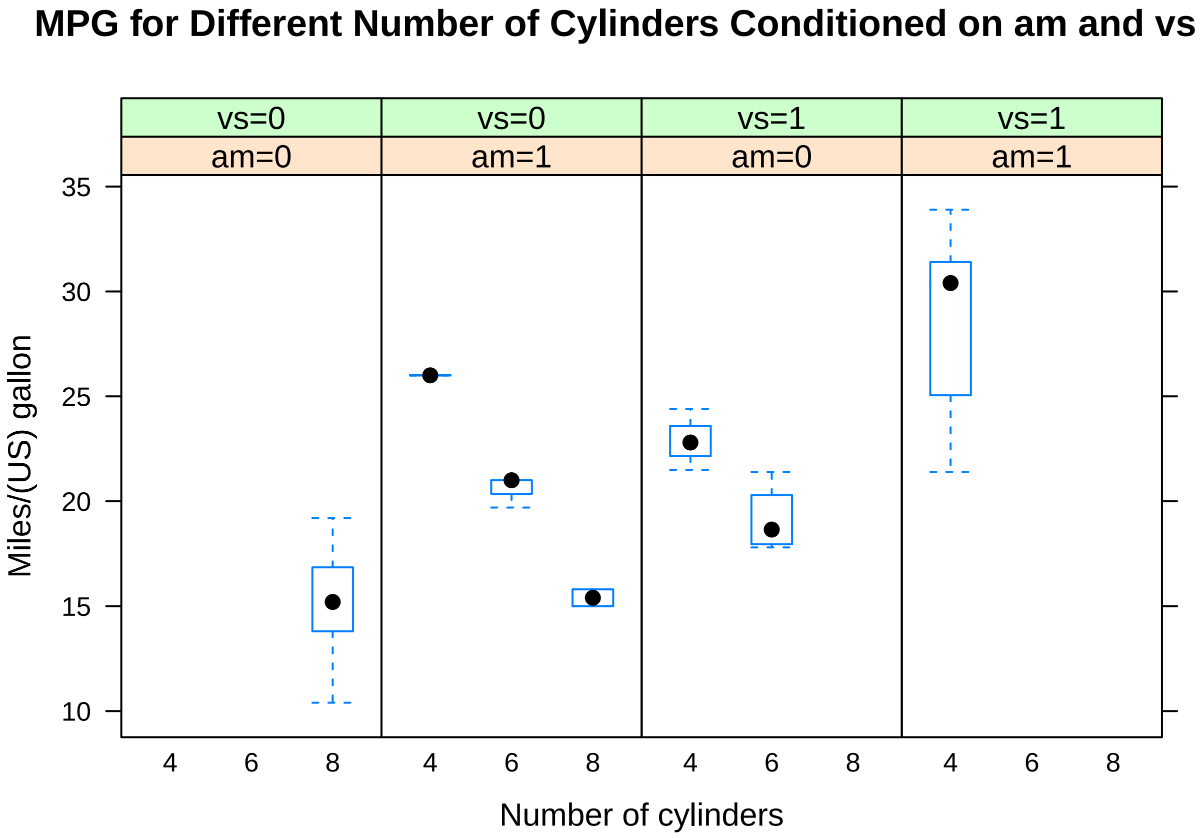
why did the distritbution overlap? is there area overlapped by all three distritbution? if yes, why?

This is because there are data points have same value in the both distribution. Yes, there is a area overlapped by all three distritbutions, this is because all of three distributions have these value.

Q3. Boxplots to display multivariate relationships

We will use the mtcars data set. we've shown you how to use boxplot with one variable with multiple levels in base R command, now let's try with multiple variables using a function from package lattice, look up the manual. make a boxplot to display the variable, mpg, for different values of cylinders, conditioned on am and vs. hint: make sure you read the function documentation of what "condition on" means, your plot should consist of (num. of levels of am X num. of levels of vs) subplots.

```r
library(lattice)
data("mtcars")
mtcars$vs<-factor(mtcars$vs,levels = c(0,1),labels=c("vs=0","vs=1"))
mtcars$am<-factor(mtcars$am,levels = c(0,1),labels=c("am=0","am=1"))
mtcars$cyl <- factor(mtcars$cyl, levels=c(4, 6, 8))
bwplot(mpg~mtcars$cyl|mtcars$am * mtcars$vs, data = mtcars,
       xlab="Number of cylinders",ylab = "Miles/(US) gallon",layout = c(4,1),
       main = "MPG for Different Number of Cylinders Conditioned on am and vs")
```



what information do you get from this plot? anything stand out? explain how/why this kind of plot can be useful.
First, as the number of cylinders increases, generally the `mpg`(Miles/(US) gallon) decreases; Second, for cars with 4 cylinders, `vs=1,am=1` tends to have lager `mpg` values than `vs=0,am=1` and `vs=1,am=0`; Third, for `vs=1` cars with 4 cylinders, `am=1` (manual) cars tend to have larger value of `mpg` than `am=0` (automatic) cars; Forth,

overal the `vs=1,am=1` have largest value of `mpg`.
This kind of plot is useful, since we can compare the data of more explanatory variables clearly at once.

Q4. Stack bar plots with gradient colors

we will use the diamonds data set from ggplot2: first load the package, then load the data set as before. Using two categorical variables, cut and clarity to create a stacked bar chart. Your y axis should be frequency. use the same color with darker shade indicating BETTER cut quality. hint: you can create a contingency table to help you plot. explain what you see from plot in the context of the data set.

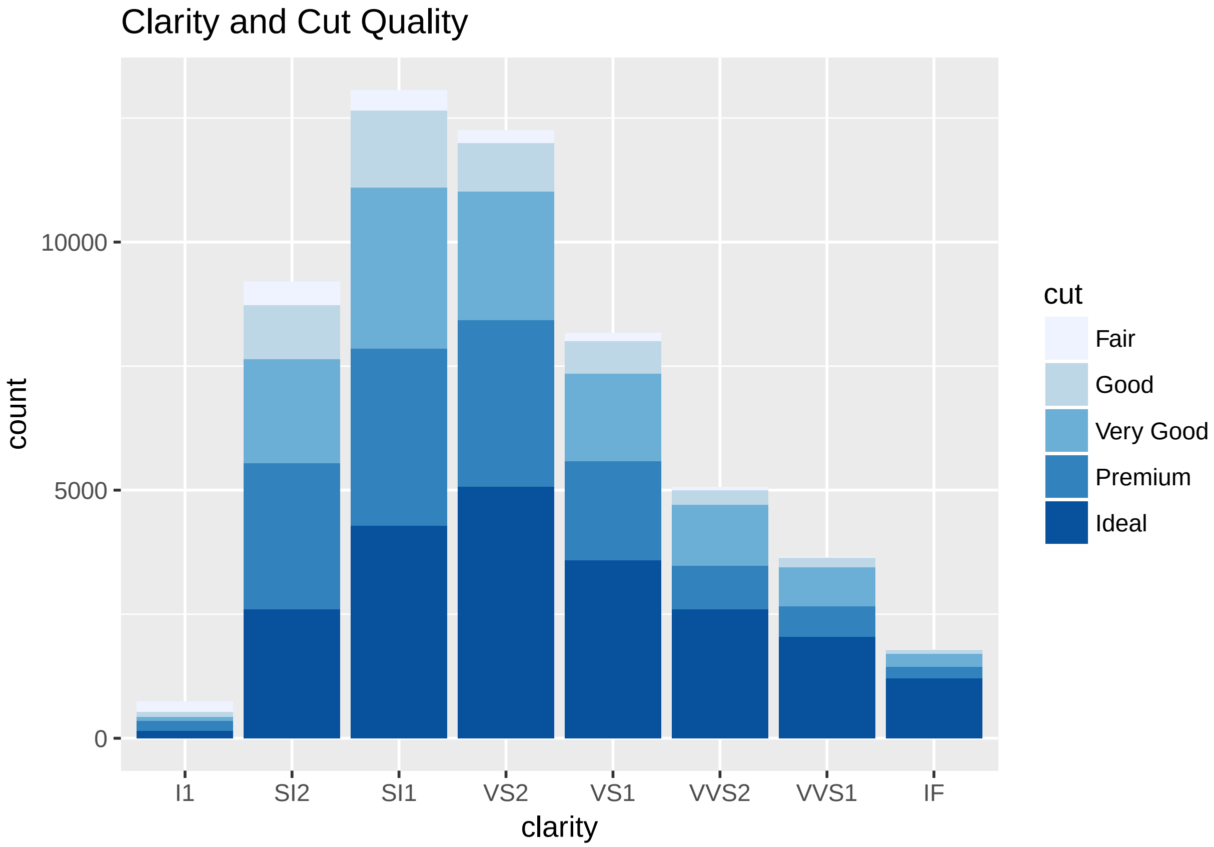
```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
data("diamonds")
table(diamonds$cut,diamonds$clarity)
```

```
##
##               I1  SI2  SI1  VS2  VS1 VVS2 VVS1   IF
##   Fair       210  466  408  261  170   69   17    9
##   Good        96 1081 1560  978  648  286  186   71
##   Very Good   84 2100 3240 2591 1775 1235  789  268
##   Premium    205 2949 3575 3357 1989  870  616  230
##   Ideal      146 2598 4282 5071 3589 2606 2047 1212
```

```r
g <- ggplot(diamonds,aes(clarity))
g+geom_bar(aes(fill=cut))+scale_fill_brewer()+ggtitle("Clarity and Cut Quality")
```



As we can see from above, as the clarity increases (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best)), the propotion of the `Ideal cut` inceases and the propotion of `Fair cut` decreases.

Part 2. Fancy plots with ggplot2 and ggmosaic.

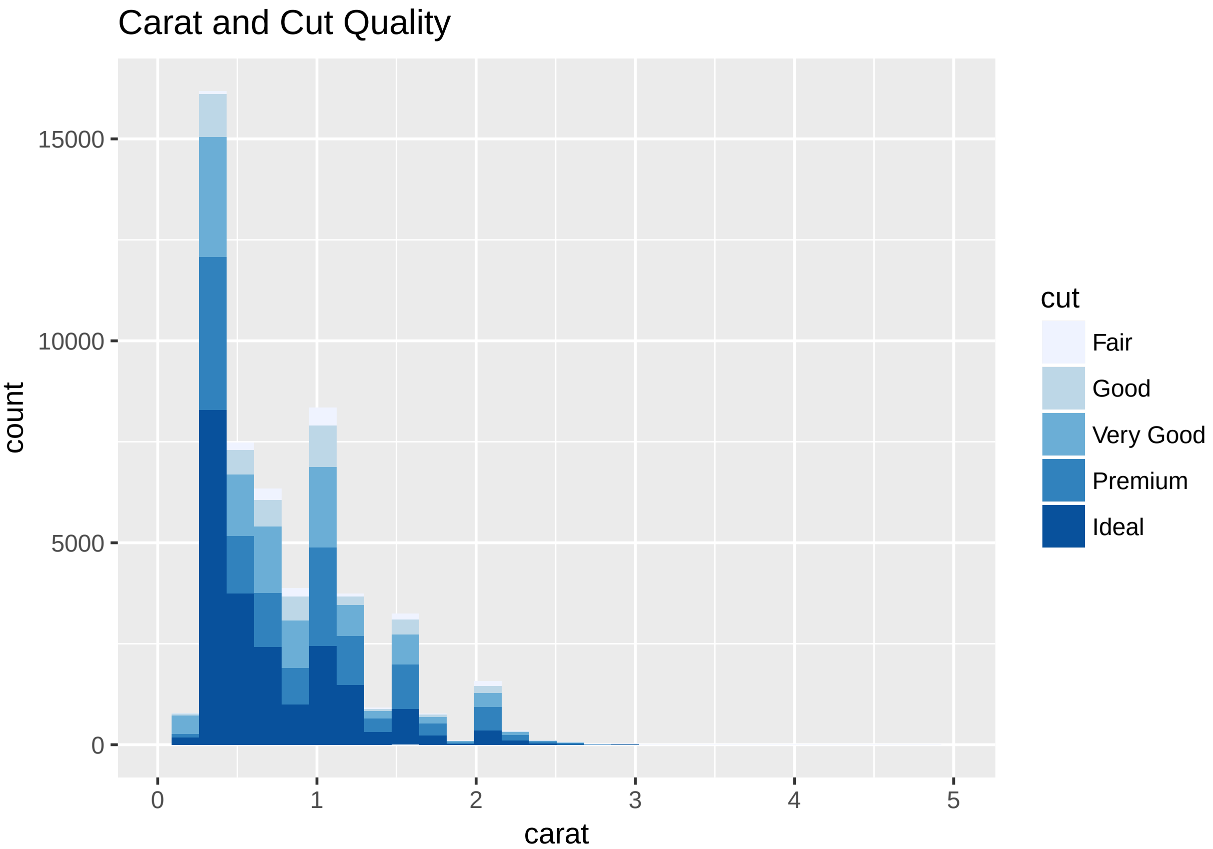
also using diamonds data set for both questions.

Q1. Use ggplot to make a histogram for the carat variable and color it by (levels of) the cut Variable.

explain what you see from the plot in the context of the data set.

Your code here

```
ggplot(diamonds,aes(carat,fill=cut))+
  geom_histogram(bins = 30)+
  ggtitle("Carat and Cut Quality")+xlim(0,5.01)+scale_fill_brewer()
```



As we can see from the plot above, as carat increases the number of diamonds(count) decreases. And in general as the carat increases, the propotion of `ideal cut` decreases and the propotion of `Fair cut` increases.

Q2. Make a mosaic plot by cut and clarity variables.

To create a mosaic plot with ggplot, you will need the ggmosaic package. explain how to interpret the plot, and what you see in the context of the data set.

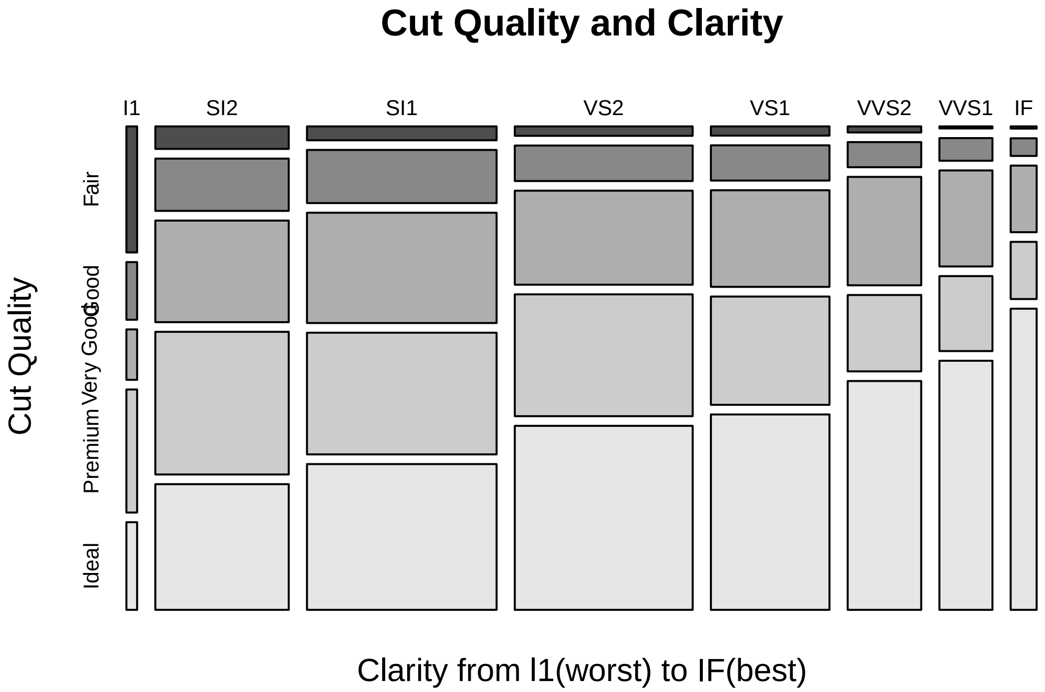
```
require(ggmosaic)
```

```
## Loading required package: ggmosaic

## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'ggmosaic'
```

```
diamonds$clarity<-factor(diamonds$clarity)

mosaicplot( table(diamonds$clarity,diamonds$cut), main="Cut Quality and Clarity",color = TRUE,
            xlab = "Clarity from l1(worst) to IF(best)",
            ylab = "Cut Quality")
```



Accordind to the graph above, we can see that, as the clarity increases (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best)), the propotion of the `Ideal cut` inceases and the propotion of `Fair cut`,`Good` and `Premium` decreases more clearly. Besides, we can also see that the number of diamonds' cut quality are within `S12`, `S11`, `VS2` and `VS1` take the majority of the total diamonds. There are small porpotion of the diamonds have `l1` or `IF` cut quality.

Bonus question:

Include a URL to a "tale" you created that carries out the code you created for this homework in RStudio implemented on the WholeTale platform at wholetale.org . A "tale" is the output of a some code and it includes the code as well. You'll need to log on to Wholetal.org using your UIUC ID. Wholetale is an ongoing research project at UIUC so it would also be useful to hear about any problems you ran into using Wholetale to implement. your homework code (extra bonus there :)) See https://wholetale.readthedocs.io/users_guide/index.html

Please see: `https://tmp-9ovf9hvl0zvm.prod.wholetale.org`

I put .r file and R markdown file which I used to generate report in the directory : `Home/work/home/IS457`.