Quiz 2 Answer Key

Question 1

```
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.1.3
```

```
data("PimaIndiansDiabetes")
```

```
head(PimaIndiansDiabetes) # show first six observations
```

```
##   pregnant glucose pressure triceps insulin mass pedigree age diabetes
## 1        6     148       72      35       0 33.6    0.627  50      pos
## 2        1      85       66      29       0 26.6    0.351  31      neg
## 3        8     183       64       0       0 23.3    0.672  32      pos
## 4        1      89       66      23      94 28.1    0.167  21      neg
## 5        0     137       40      35     168 43.1    2.288  33      pos
## 6        5     116       74       0       0 25.6    0.201  30      neg
```

a) What is the average value of 2-Hour serum insulin?

```
mean(PimaIndiansDiabetes$insulin)
```

```
## [1] 79.79948
```

b) What is the average value of 2-Hour serum insulin in terms of level of diabetes ?

```
aggregate(x = PimaIndiansDiabetes$insulin, by =
list(PimaIndiansDiabetes$diabetes), FUN = "mean")
```

```
##   Group.1        x
## 1     neg  68.7920
## 2     pos 100.3358
```

c) Which level in diabeets variable has more variability around it's average value for 2-Hour serum insulin?

```
aggregate(x = PimaIndiansDiabetes$insulin, by =
list(PimaIndiansDiabetes$diabetes), FUN = "sd")
```

```
##   Group.1         x
## 1     neg  98.86529
## 2     pos 138.68912
```

d) Measure the correlation coefficient between age and glucose of participants.

```
cor(PimaIndiansDiabetes$age,PimaIndiansDiabetes$glucose)
```

```
## [1] 0.2635143
```

e) Create a table with summary command for all variables.

```
summary(PimaIndiansDiabetes)
```

```
##     pregnant         glucose         pressure         triceps
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     insulin           mass          pedigree           age        diabetes
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00   neg:500
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   pos:268
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

Question 2

Read the dataset.

```
ship<-read.table("ship.txt",header=T)
head(ship)
```

```
##          name     line age tonnage passengers length cabin
passenger_density
## 1     Journey  Azamara   6  30.277       6.94   5.94  3.55
42.64
## 2       Quest  Azamara   6  30.277       6.94   5.94  3.55
42.64
## 3 Celebration Carnival  26  47.262      14.86   7.22  7.43
31.80
## 4    Conquest Carnival  11 110.000      29.74   9.53 14.88
36.99
## 5     Destiny Carnival  17 101.353      26.42   8.92 13.21
38.36
## 6     Ecstasy Carnival  22  70.367      20.52   8.55 10.20
34.29
##    crew
## 1  3.55
## 2  3.55
## 3  6.70
## 4 19.10
## 5 10.00
## 6  9.20
```

```
dim(ship)
```

```
## [1] 158   9
```

It is seen that the ship dataset have 158 rows and 9 columns.

a) Please find the name of the ships that has maximum age, maximum tonnage, maximum passenger,maximum length, maximum cabin and maximum crew separetely.

```
Questions=c("Max.Age","Max.Ton","Max.Pass","Max.Length","Max.Cabin","Max.Crew
")

Answers=c(as.character(ship$name[which.max(ship$age)][1]),as.character(ship$n
ame[which.max(ship$tonnage)][1]),as.character(ship$name[which.max(ship$passen
gers)][1]),as.character(ship$name[which.max(ship$length)][1]),as.character(sh
ip$name[which.max(ship$cabin)][1]),as.character(ship$name[which.max(ship$crew
)][1]))

data=data.frame(Questions,Answers)
data

##    Questions    Answers
## 1    Max.Age Marco_Polo
## 2    Max.Ton      Oasis
## 3   Max.Pass      Oasis
## 4 Max.Length      Oasis
## 5  Max.Cabin      Oasis
## 6   Max.Crew      Oasis
```

b) Obtain the summary of the dataset

```
summary(ship)

##      name               line                age           tonnage
##  Length:158         Length:158         Min.   : 4.00   Min.   :  2.329
##  Class :character   Class :character   1st Qu.:10.00   1st Qu.: 46.013
##  Mode  :character   Mode  :character   Median :14.00   Median : 71.899
##                                        Mean   :15.69   Mean   : 71.285
##                                        3rd Qu.:20.00   3rd Qu.: 90.772
##                                        Max.   :48.00   Max.   :220.000
##    passengers        length           cabin        passenger_density
##  Min.   : 0.66   Min.   : 2.790   Min.   : 0.330   Min.   :17.70
##  1st Qu.:12.54   1st Qu.: 7.100   1st Qu.: 6.133   1st Qu.:34.57
##  Median :19.50   Median : 8.555   Median : 9.570   Median :39.09
##  Mean   :18.46   Mean   : 8.131   Mean   : 8.830   Mean   :39.90
##  3rd Qu.:24.84   3rd Qu.: 9.510   3rd Qu.:10.885   3rd Qu.:44.19
##  Max.   :54.00   Max.   :11.820   Max.   :27.000   Max.   :71.43
##       crew
##  Min.   : 0.590
##  1st Qu.: 5.480
##  Median : 8.150
##  Mean   : 7.794
##  3rd Qu.: 9.990
##  Max.   :21.000
```

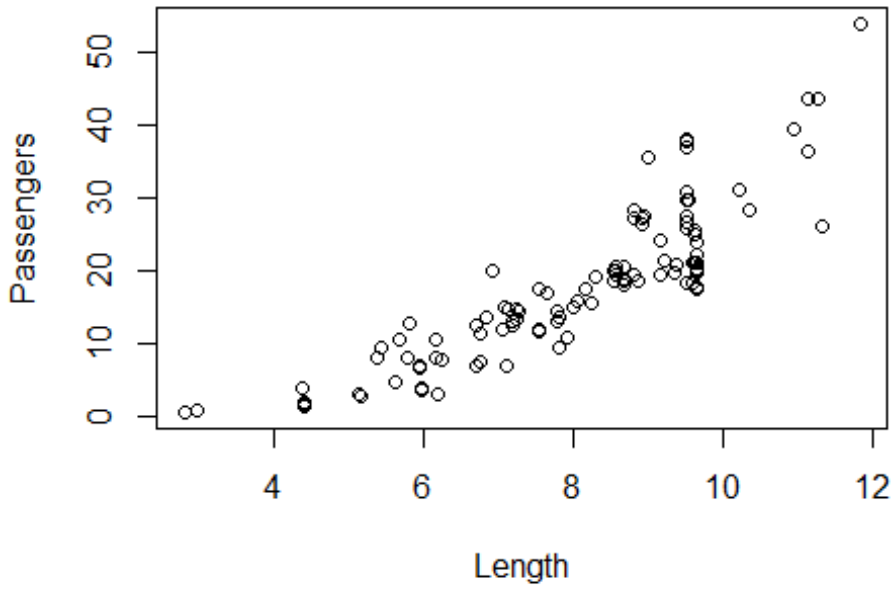
Example Interpretation:

It is seen that the average length of ships is 8.131, while it's median is 8.555. Since median is greater than mean, we can say that length has perhaps left skewed distributon. Also, we can say that the minimum and maximum values of the length variables are 2.79 and 11.82 respectvely. Also, the 75% of the observations are below 9.51 and 75% of the observations are above 7.1. On the other hand, we can say that spirit has the highest frequency in name variable.

c) What is the association between length of the ship and number of passengers?

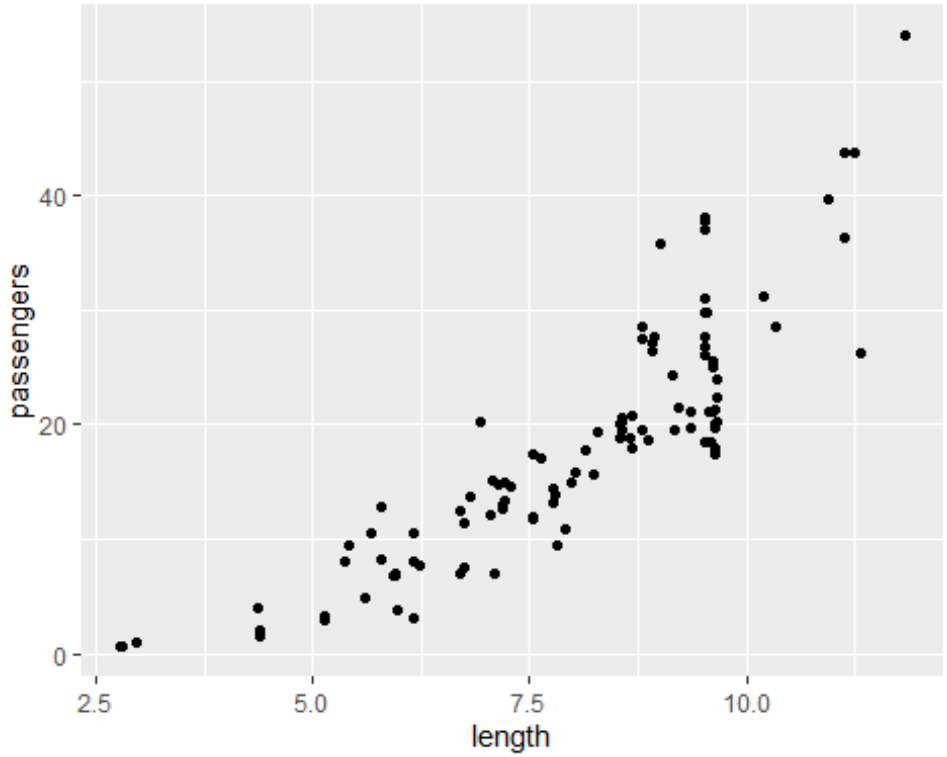
Usual Plot

```
plot(ship$length,ship$passengers,xlab = "Length",ylab="Passengers")
```



ggplot

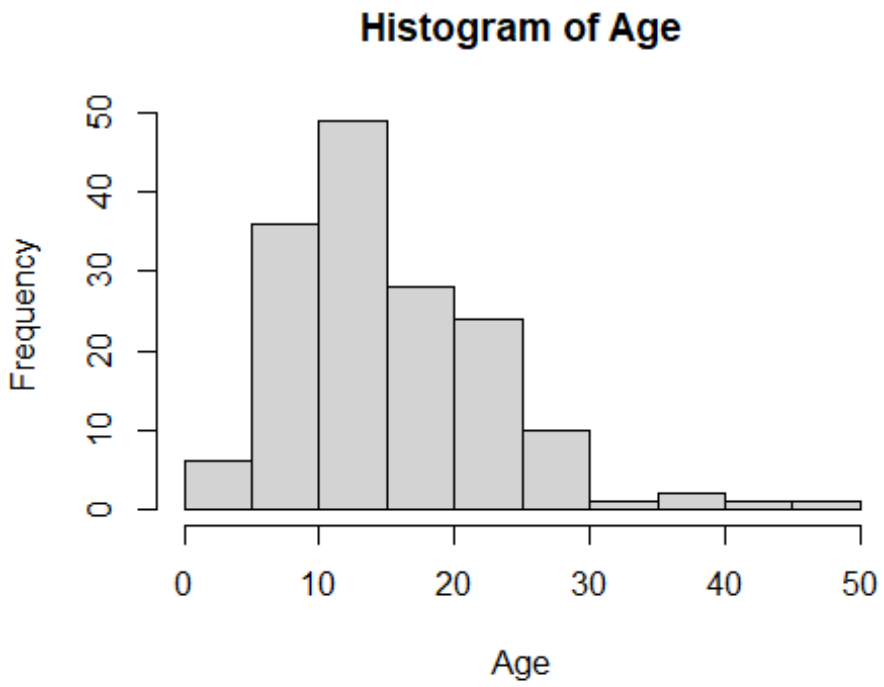
```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.3

ggplot(ship,aes(x=length,y=passengers))+geom_point()
```



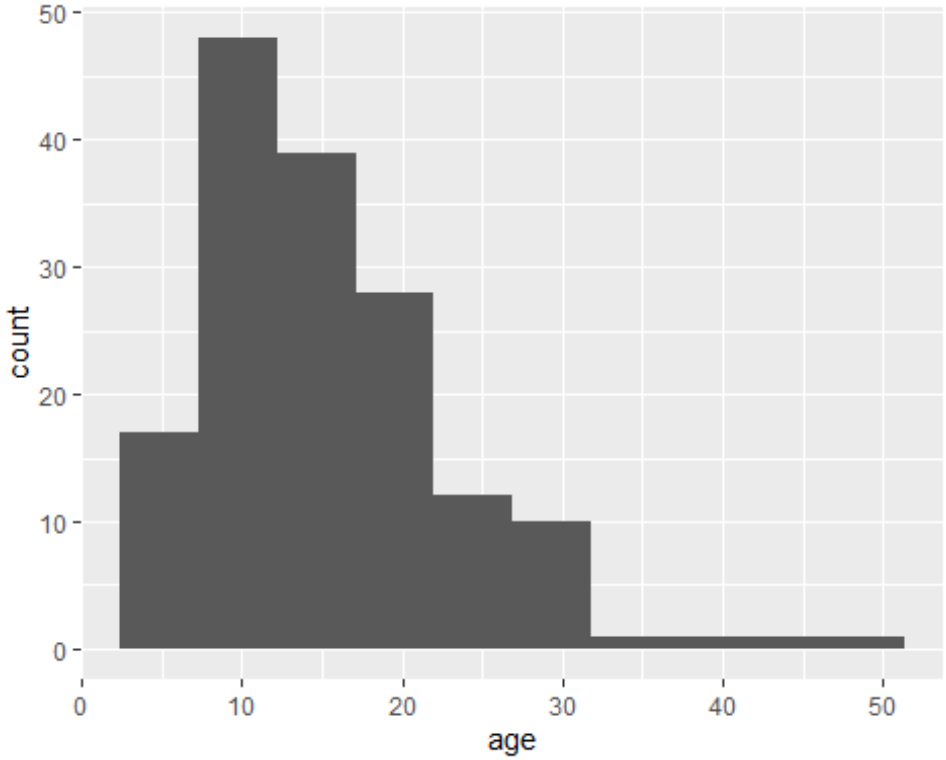
It is seen that there is a positive linear relationship between variables.

d) Draw the histogram of age

```
hist(ship$age,xlab="Age",main="Histogram of Age")
```



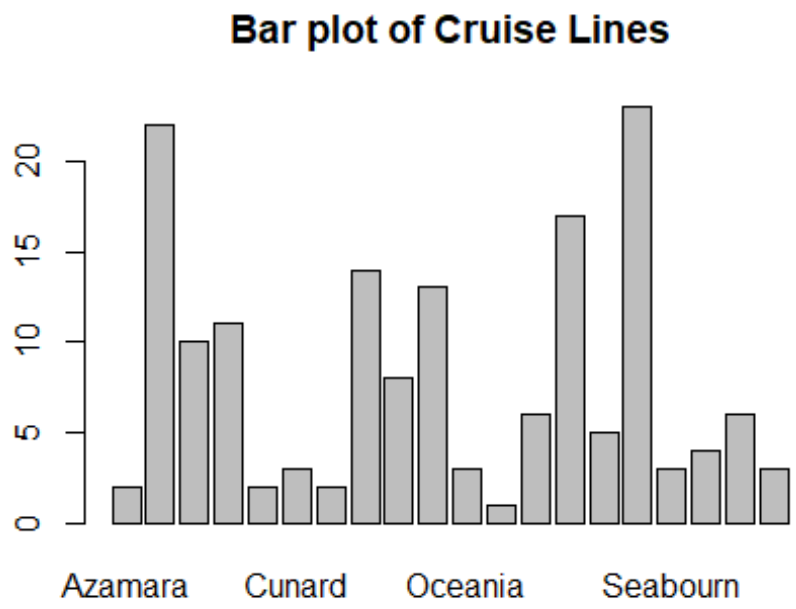

```
library(ggplot2)
ggplot(ship,aes(x=age))+geom_histogram(bins=10)
```



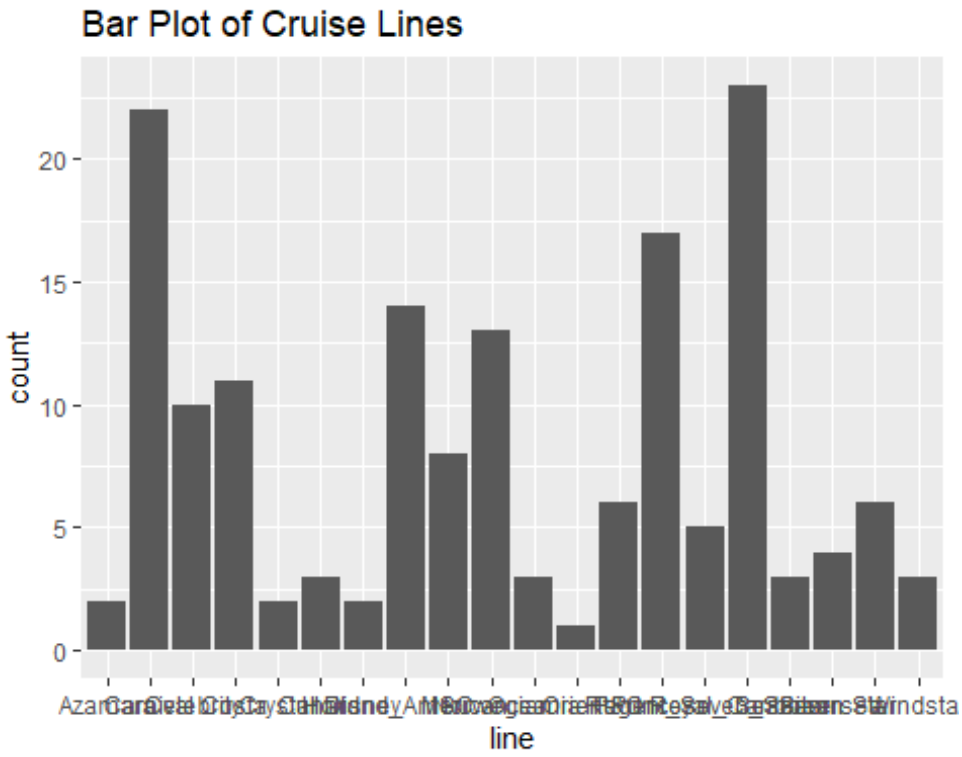
By drawing histogram of the variable of interest, we can say that age variable has right skewed distribution.

e) Draw a bar plot of cruise lines. Then, write the name of most frequent three lines.

```
t<-table(ship$line)

barplot(t,main="Bar plot of Cruise Lines")
```




```
library(ggplot2)
ggplot(ship,aes(x=line))+geom_bar()+labs(title="Bar Plot of Cruise Lines")
```



Royal Caribbean, Carnival and Princess are the most frequent three cruise lines.

f)Consider the three cruise lines that you found in part d. Then, subset the dataset that contains these three cruise lines and corresponding number of passengers for these lines. Having a data, draw a box plot of passengers to compare these three lines.

Take the subset at first.

```
rc<-ship$passengers[ship$line=="Royal_Caribbean"]
ca<-ship$passengers[ship$line=="Carnival"]
p<-ship$passengers[ship$line=="Princess"]
```

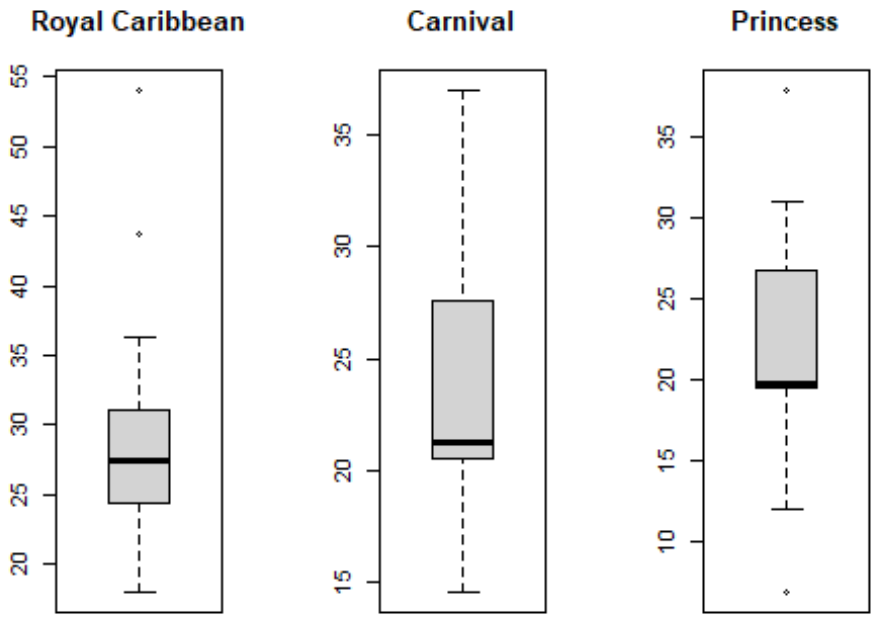
Create dataset. Since the length of subsets are not equal, I add NA terms to have an equal length.

```
data<-data.frame(rc,ca=c(ca,NA),p=c(p,rep(NA,6)))
head(data)
```

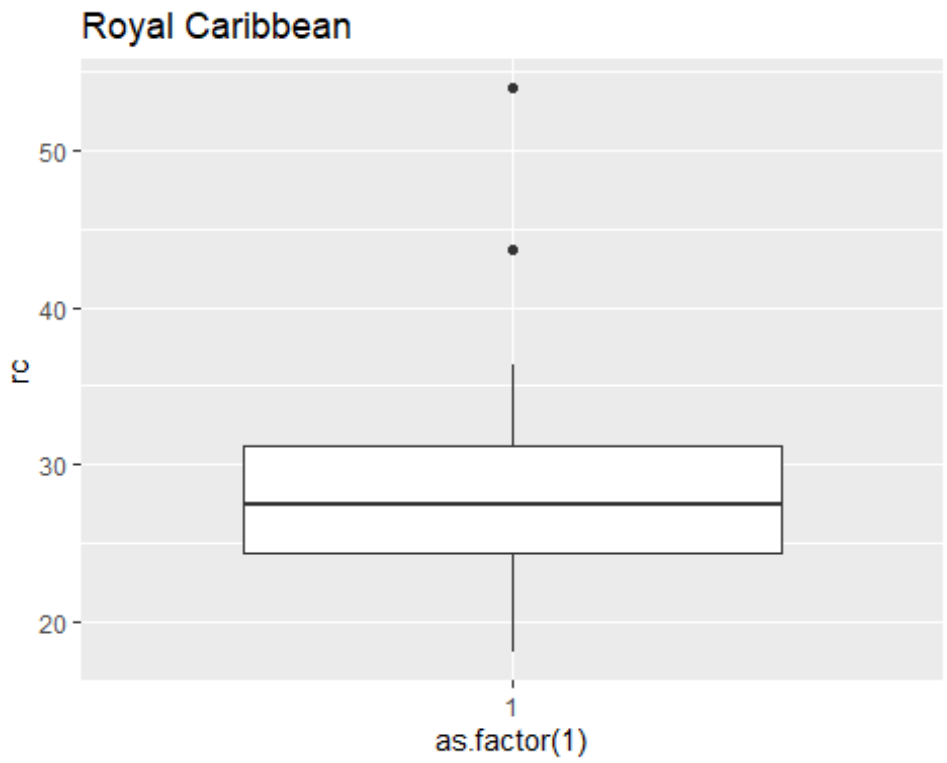
```
##      rc    ca     p
## 1 31.14 14.86 26.00
## 2 25.01 29.74 19.74
## 3 20.20 26.42 31.00
## 4 19.50 20.52 19.50
## 5 31.14 20.52 26.74
## 6 43.70 20.56 37.82
```

```
par(mfrow=c(1,3))
boxplot(rc,main="Royal Caribbean")
```

```
boxplot(ca,main="Carnival")
boxplot(p,main="Princess")
```

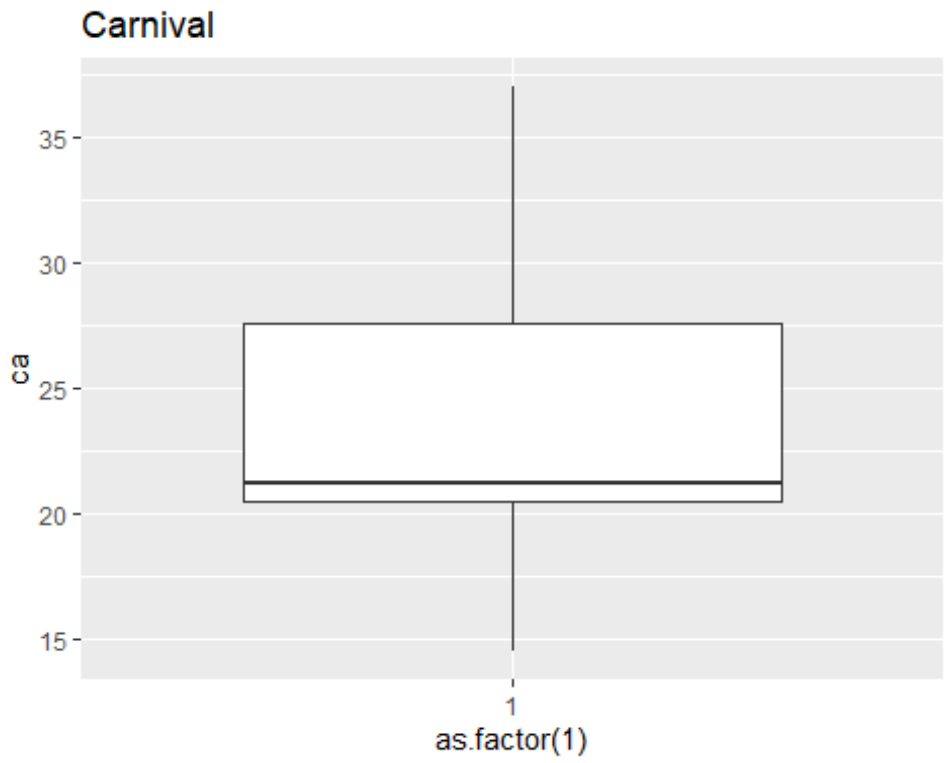



```
library(ggplot2)
ggplot(data,aes(x=as.factor(1),y=rc))+geom_boxplot()+labs(title="Royal
Caribbean")
```



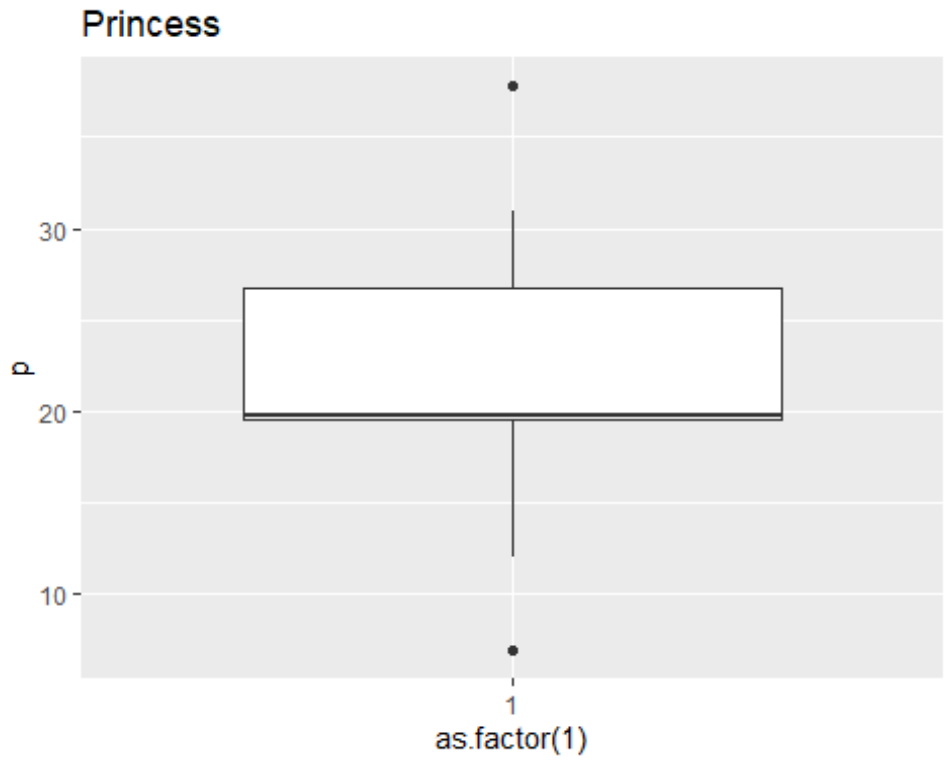

```
ggplot(data,aes(x=as.factor(1),y=ca))+geom_boxplot()+labs(title="Carnival")
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
```



```
ggplot(data,aes(x=as.factor(1),y=p))+geom_boxplot()+labs(title="Princess")
```

```
## Warning: Removed 6 rows containing non-finite values (`stat_boxplot()`).
```



It is seen that royal caribbean has the highest median value, while princess has the lowest value. Also, royal caribbean has almost symetric, carnival has right skewed and princess has bimodel distrbution. Lastly, royal caribbean and carnival have outliers.

g) Create a new variable and call it class of the passenger by using logical operator and for loop with regard to the following conditions. If number of passenger is less than 19, then class them as 0, it it is between 19 and 24, then class them as 1, and if it is greater than 23, then class them as 2. After that, draw a bar plot then write the class that has the highest frequency.

Here, I want to do some data manupulation.

Let's consider,

```
rc<-ship$passengers[ship$line=="Royal_Caribbean"]
ca<-ship$passengers[ship$line=="Carnival"]
p<-ship$passengers[ship$line=="Princess"]
```

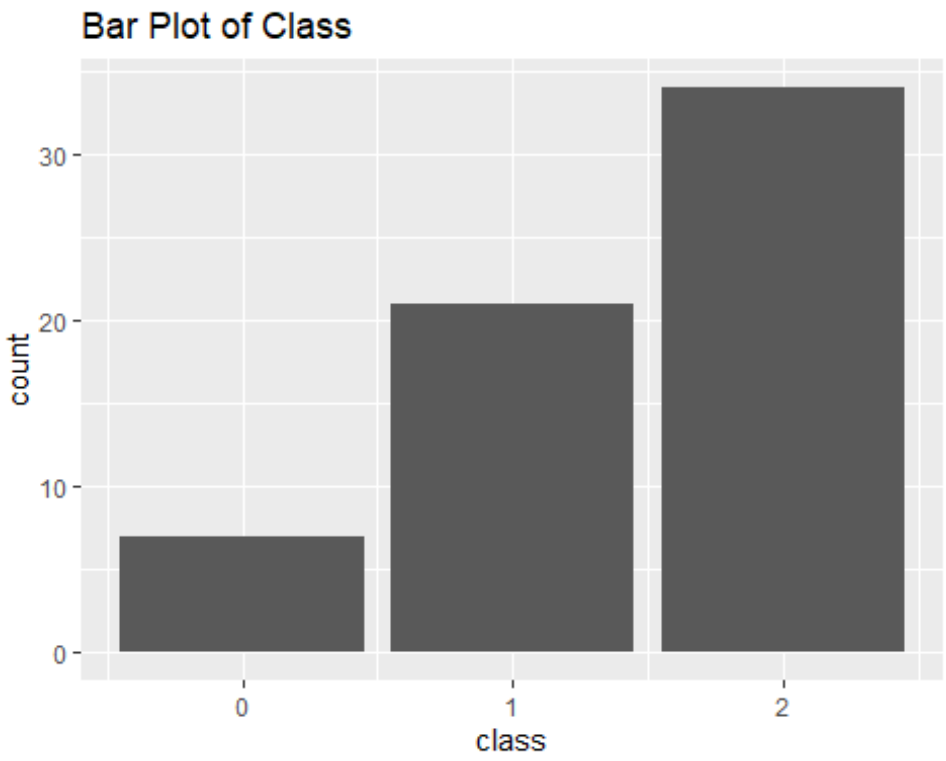
Then, create a new dataset.

```
name=c(rep("rc",length(rc)),rep("ca",length(ca)),rep("p",length(p)))
values=c(rc,ca,p)
newdata<-data.frame(name,values)
head(newdata)
```

```
##   name values
## 1   rc  31.14
## 2   rc  25.01
## 3   rc  20.20
## 4   rc  19.50
## 5   rc  31.14
## 6   rc  43.70
```

```
class=c()
for (i in 1:length(newdata$values)){
  if (newdata$values[i]<19){class[i]=0}
  else if (newdata$values[i]>=19&newdata$values[i]<24){class[i]=1}
    else{class[i]=2}
  }

newdata<-data.frame(newdata,class)

barplot(table(newdata$class),main="Bar Plot of Class")
```


```
library(ggplot2)
ggplot(newdata,aes(x=class))+geom_bar()+labs(title="Bar Plot of Class")
```



It is seen that class 2 has the highest frequency and class 0 has the lowest frequency.

Question 3

a) For observations 10000 to 11000, get the mean of columns 8, 9, 10.

```
library(ggplot2)
apply(diamonds[10000:11000, 8:10], 2, mean)

##        x        y        z
## 6.237852 6.233506 3.851049
```

b) Same as 'a' but round the results to one digit.

```
round(apply(diamonds[10000:11000, 8:10], 2, mean),1)

##   x   y   z
## 6.2 6.2 3.9
```

c) Sort the rounded results in ascending order.

```
sort(round(apply(diamonds[10000:11000, 8:10], 2, mean),1))

##   z   x   y
## 3.9 6.2 6.2
```

d) Calculate the median of table by the cut.

```
tapply(diamonds$table,diamonds$cut,median)
```

```
##      Fair      Good Very Good   Premium     Ideal 
##        58        58        58        59        56
```

e) Use 'apply' to perform a modulo division by 2 on each value in the x,y and z columns of the matrix.

```
subset<-diamonds[,8:10]
head(apply(subset,1:2,function(x) x%%2)) #write head to see first 6 outputs
```

```
##         x    y    z
## [1,] 1.95 1.98 0.43
## [2,] 1.89 1.84 0.31
## [3,] 0.05 0.07 0.31
## [4,] 0.20 0.23 0.63
## [5,] 0.34 0.35 0.75
## [6,] 1.94 1.96 0.48
```