SESSION 8 – ASSIGNMENT 8.1

Date: 29th Jan 2019

Use the package -RcmdrPlugin.IPSUR. data(RcmdrTestDrive) and perform the below operations:

```
install.packages("RcmdrPlugin.IPSUR")
library(RcmdrPlugin.IPSUR)
head(RcmdrTestDrive)
```

```
> head(RcmdrTestDrive)
  order    smoking gender        race before after salary reduction parking
1     1 Nonsmoker Female   Caucasian   72.6  75.2 618.65         9       2
2     2 Nonsmoker   Male AfricanAmer   75.3  73.2 544.56        62       1
3     3 Nonsmoker Female   Caucasian   75.5  74.5 550.24        19       4
4     4 Nonsmoker Female   Caucasian   71.3  74.6 616.16        30       1
5     5 Nonsmoker Female    Hispanic   74.3  73.8 543.39       105       1
6     6 Nonsmoker   Male   Caucasian   73.0  73.6 692.09        43       1
```

```
mean(RcmdrTestDrive$salary)
```

```
> mean(RcmdrTestDrive$salary)
[1] 724.5164
```

```
library(plyr)
library(reshape2)
library(plyr)
library(ggplot2)
```

a) **Calculate the average salary by gender and smoking status.**

```
#of salary
tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, mean)
```

```
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, mean)
  Female     Male
698.0911 743.3915
```

#of smoking status
tapply(RcmdrTestDrive$salary, RcmdrTestDrive$smoking, mean)

```
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$smoking, mean)
Nonsmoker    Smoker
 719.3792  746.3494
```

b) Which gender has the highest mean salary?

genders mean salary respectively
#Female Male
#698.0911 743.3915
#so its the gender male which is highest

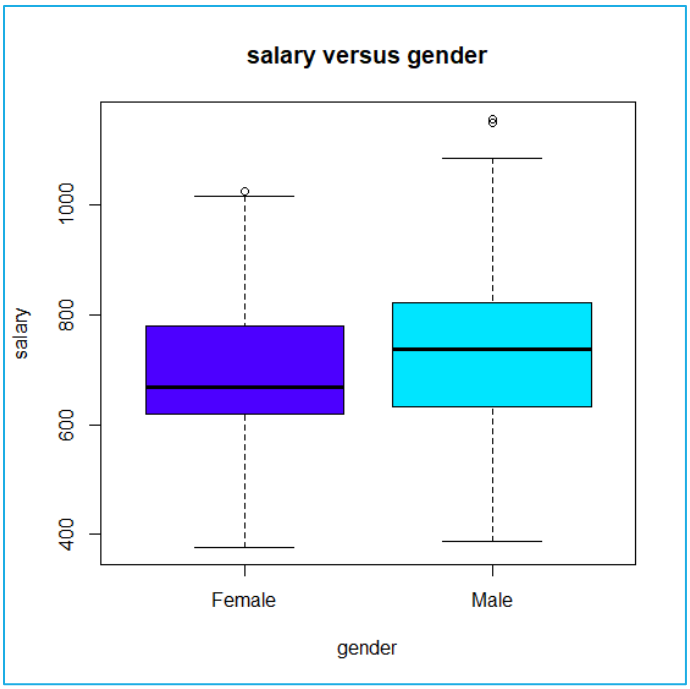
c) Report the highest mean salary.

if we are considering the mean of salary then
mean(RcmdrTestDrive$salary)
#724.5164 #its the mean of salary
if we talk about which has the highest salary of all then
which.max(RcmdrTestDrive$salary)
#152
so at 152 its the highest salary present which is 1156.16

d) Compare the spreads for the genders by calculating the standard deviation of salary by gender.

```
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, sd)
  Female     Male
130.7053 158.5423
```

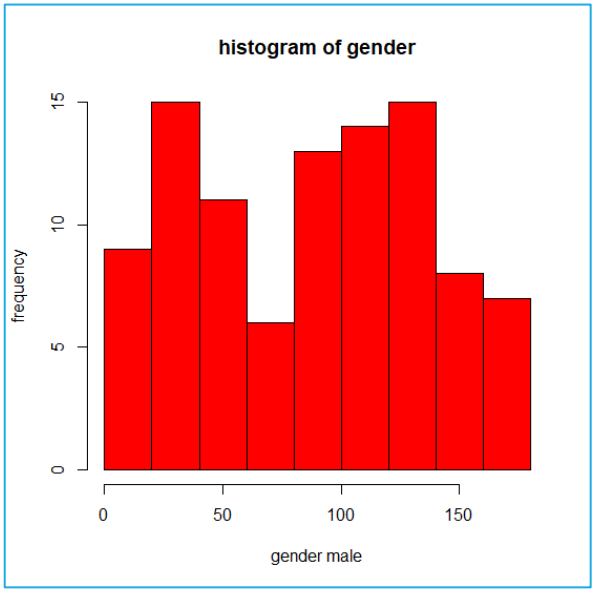
boxplot(salary~gender,data= RcmdrTestDrive,main="salary versus gender",xlab="gender",ylab="salary",col=topo.colors(2))



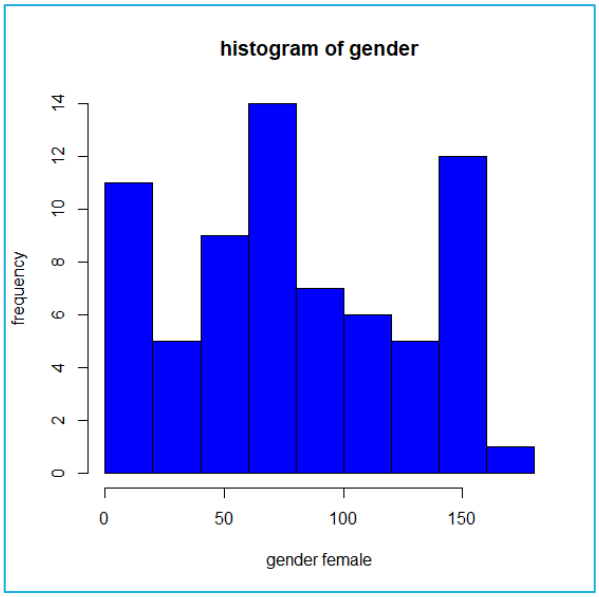
#see mean too
tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, mean)

```
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, mean)
  Female     Male
698.0911 743.3915
>
```

#we can plot histogram by genders to compare spreadness
hist(which(RcmdrTestDrive$gender == "Male") ,xlab = "gender male", ylab = "frequency",main="histogram of gender",col="red")



hist(which(RcmdrTestDrive$gender == "Female") ,xlab = "gender female", ylab = "frequency",main="histogram of gender",col="blue")



#so higher the sd higher the members of a group differ from the mean value for the group
#that the data spreadness in gender male is more comparatively to gender female