SESSION 8: Exploratory Data Analytics

Assignment 1

1. Use the package -RcmdrPlugin.IPSUR.

data(RcmdrTestDrive)

and perform the below operations:

```
install.packages("RcmdrPlugin.IPSUR")
install.packages("rlang")
install.packages("car")
library(rlang)
library(Rcmdr)
library(RcmdrMisc)
library(RcmdrPlugin.IPSUR)

library(sandwich)

library(effects)

library(car)

data("RcmdrTestDrive")

data(BloodPressure)

View(RcmdrTestDrive)

View(BloodPressure)
```

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

```
> # Avg Salary by Gender :
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, mean)

  Female     Male 
698.0911 743.3915 

> # Avg Salary by Smoking Status
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$smoking, mean)

Nonsmoker    Smoker 
 719.3792  746.3494 
```

b. Which gender has the highest mean salary?

Ans : Gender Male has highest mean salary

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

  Female     Male 
698.0911 743.3915 
```

c. Report the highest mean salary.

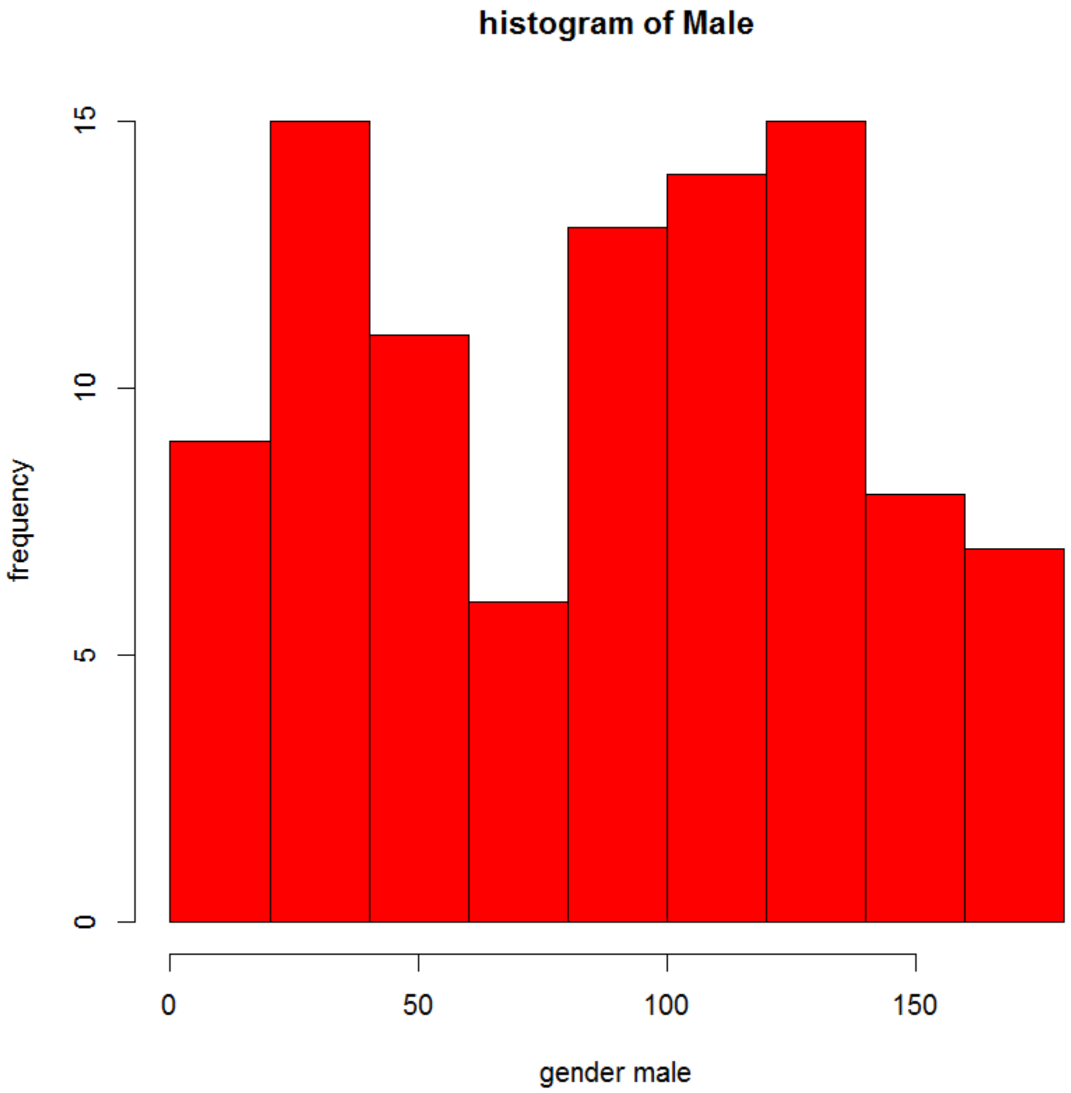
```
> mean(RcmdrTestDrive$salary)
[1] 724.5164
```

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

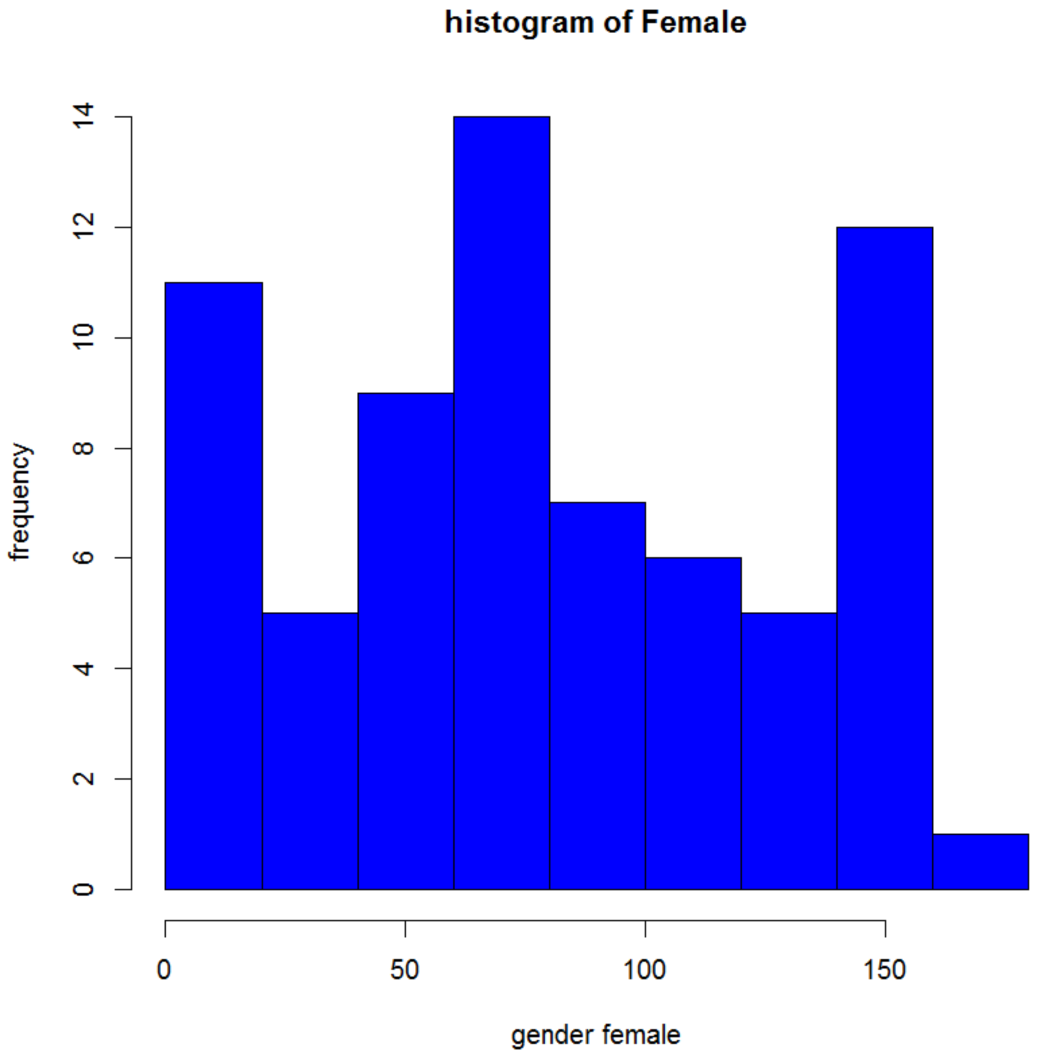
```
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, sd)
  Female     Male 
```

```
130.7053 158.5423
> 
> #for answering the compareness of spreads of genders lets plot boxplot
> boxplot(salary~gender,data= RcmdrTestDrive,main="salary versus gender",x
lab="gender",ylab="salary",col=topo.colors(2))
> 
> #see mean too
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, mean)
  Female     Male 
698.0911 743.3915 
> #as from mean only there is sd deviate takes place
> 
```

```
> #we can aslo 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", y
lab = "frequency",main="histogram of gender",col="blue")
>
```



```
> #as we know standard deviation is a measure that is used to quantify the
amount of variation or dispersion of a set of data values.
> #so higher the sd higher the members of a group differ from the mean val
ue for the group
> #by this we means
> #that the data spreadness in gender male is more comparatively to gender
female
```