Assignment - 1

Session 8 – Exploratory

Data Analytics

1. Use the package RcmdrPlugin.IPSUR.

data(RcmdrTestDrive)

and perform the below operations:

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

Ans:

> #of salary

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

Female Male

698.0911 743.3915

> #of smoking status

> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$smoking, mean)

Nonsmoker Smoker

719.3792 746.3494

b. Which gender has the highest mean salary?

Ans:

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

Female Male

698.0911 743.3915

#so its the gender male which is highest

c. Report the highest mean salary.

Ans:

#if we talk about the mean of salary then here it is

> mean(RcmdrTestDrive$salary)

[1] 724.5164

> #however if we talk about which has the highest salary of all then it is like this

> which.max(RcmdrTestDrive$salary)

[1] 152

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

Ans:

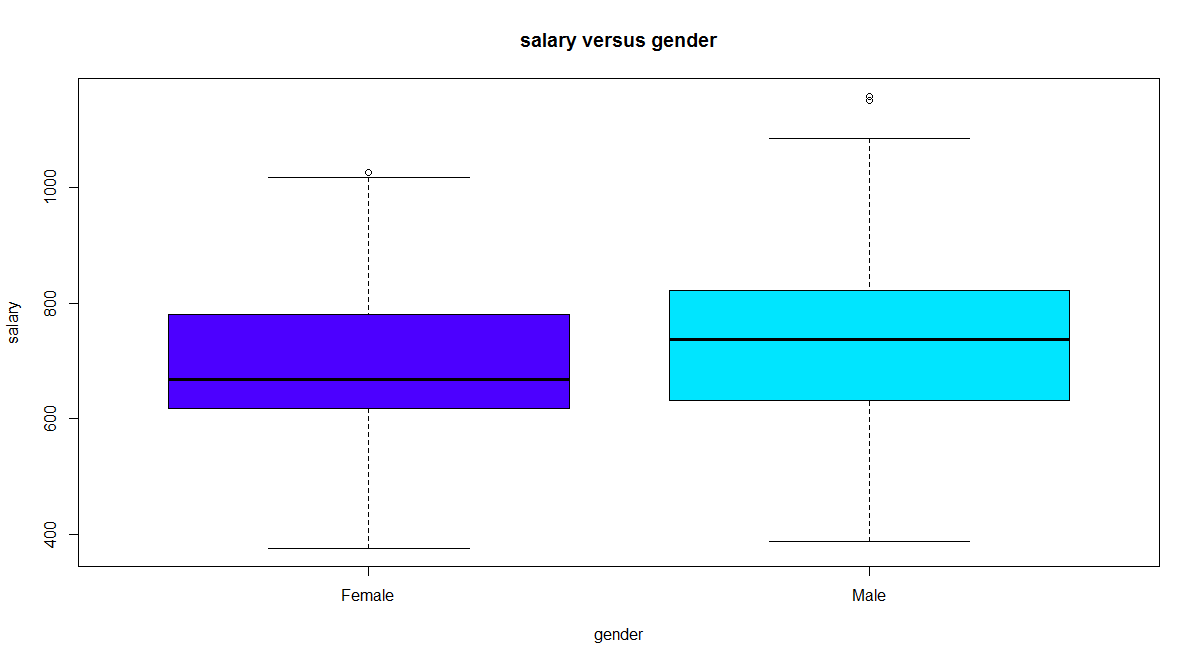
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, sd)

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",xlab="gender",ylab="salary",col=topo.colors(2))



#see mean too

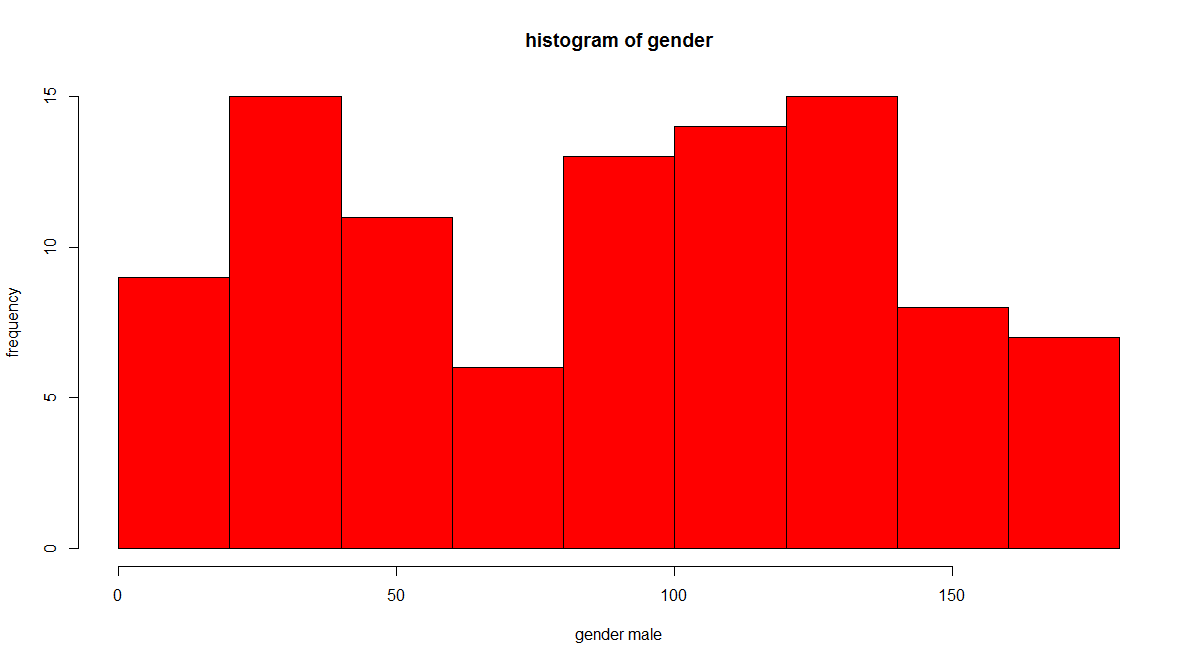
> tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, mean)

Female Male

698.0911 743.3915

> #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", ylab="frequency",main="histogram of gender",col="blue")

