# Problem Statement

1. Use the package RcmdrPlugin.IPSUR. data(RcmdrTestDrive)

and perform the below operations:

library(RcmdrPlugin.IPSUR)

View(RcmdrTestDrive)

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

|  |
| --- |
| > #average salary  > tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, mean)  Female Male  698.0911 743.3915  > #average smoking status  > tapply(RcmdrTestDrive$salary, RcmdrTestDrive$smoking, mean)  Nonsmoker Smoker  719.3792 746.3494 |
|  |
| |  | | --- | | > | |

1. Which gender has the highest mean salary?

# genders mean salary respectively

#Female Male

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

#698.0911 743.3915

#so its the gender male which is highest

1. Report the highest mean salary.

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

mean(RcmdrTestDrive$salary)

#724.5164

#its the mean of salary

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

which.max(RcmdrTestDrive$salary)

#152

RcmdrTestDrive[152,]

#so at 152 its the highest salary present which is 1156.16

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

#salary by gender.

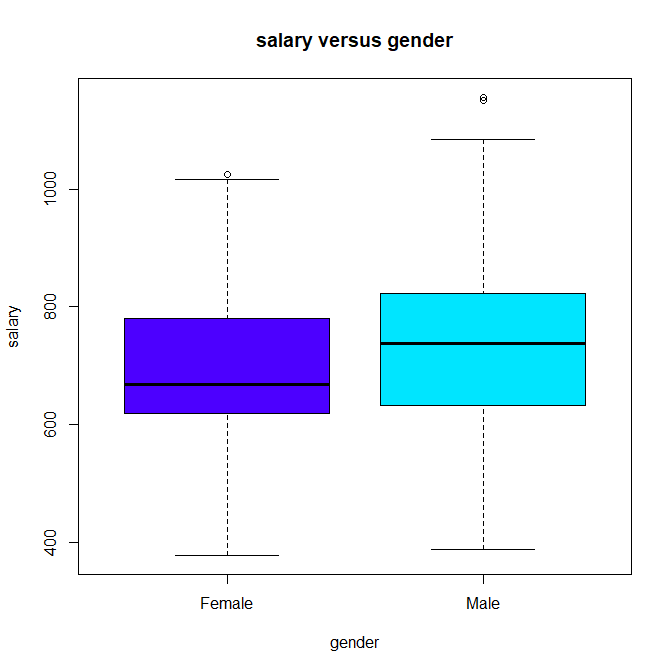
tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, sd)

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



#see mean too

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

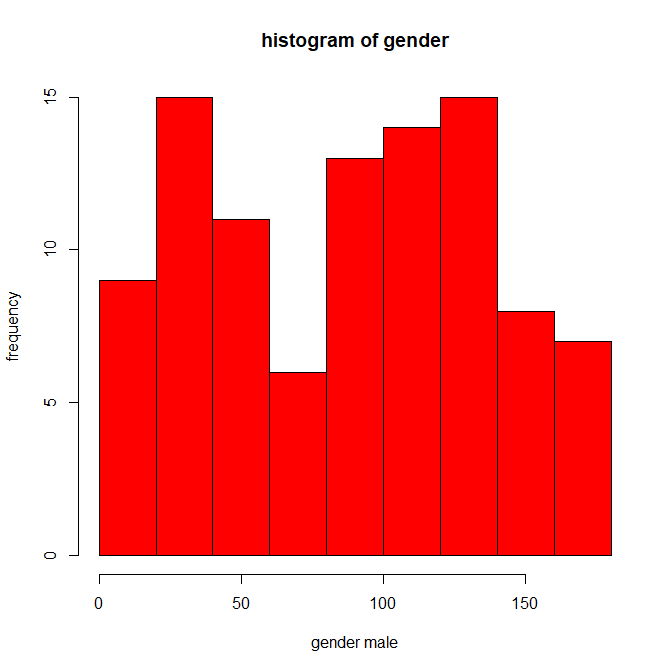
#Female Male

#698.0911 743.3915

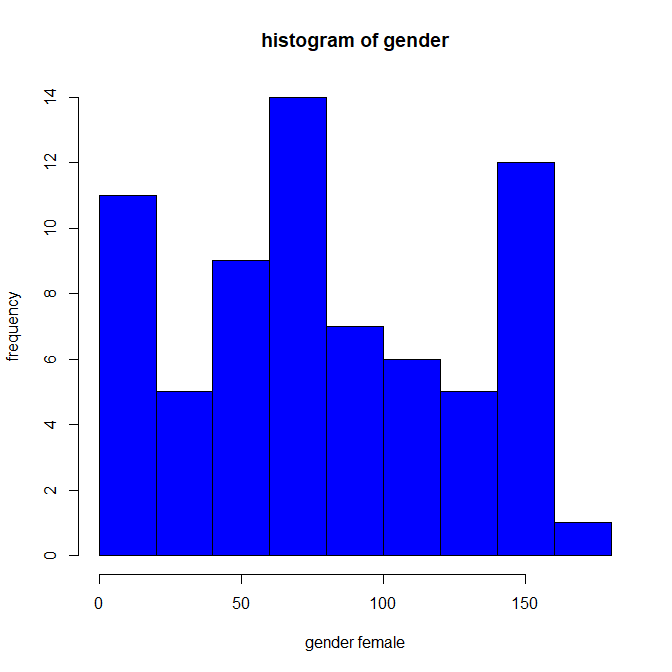
#as from mean only there is sd deviate takes place

#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", ylab = "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 value for the group

#by this we means

#that the data spreadness in gender male is more comparatively to gender female