Session 8 – Exploratory

Data Analytics

Assignment – 1

1. Use the package RcmdrPlugin.IPSUR.

data(RcmdrTestDrive)

and perform the below operations:

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

library(Rcmdr)

library(RcmdrPlugin.IPSUR)

data(RcmdrTestDrive)

RcmdrTestDrive

library(dplyr)

str(RcmdrTestDrive)

Data <- RcmdrTestDrive

AvgSalary <- Data%>%group\_by(gender, smoking)%>%

select(smoking, gender, salary)%>%summarise(mean(salary))

AvgSalary <- as.data.frame(AvgSalary)

AvgSalary$meansalary <- AvgSalary$`mean(salary)`

AvgSalary # Data Frame

stripchart(meansalary ~ gender, vertical=TRUE, method="jitter",

ylab="meansalary", data=AvgSalary)

1. Which gender has the highest mean salary?

with(RcmdrTestDrive, plotMeans(salary, gender, error.bars="se")) # Male has highest meansalary

meansalary <- as.data.frame(Data%>%group\_by(gender)%>%

select(gender,salary)%>%summarise(mean(salary)))

meansalary$meansalary <- meansalary$`mean(salary)`

meansalary

meansalary[which.max(meansalary$meansalary),]

bp <- barplot(meansalary$meansalary, xlab = names(meansalary),

ylab = "Mean Salary",

main = "Mean Salary by Gender",

col = c("Red", "Blue"),

legend = meansalary$gender)

text(bp, 0, meansalary$meansalary, cex = 1, pos = 3)

# Male has the highest mean salary

1. Report the highest mean salary.

meansalary$meansalary <- meansalary$`mean(salary)`

meansalary

meansalary[which.max(meansalary$meansalary),] # gives the maximum mean salary row i.e. Male

# Highest Mean Salary = 743.391

bp <- barplot(meansalary$meansalary, xlab = names(meansalary),

ylab = "Mean Salary",

main = "Mean Salary by Gender",

col = c("Red", "Blue"),

legend = meansalary$gender)

text(bp, 0, meansalary$meansalary, cex = 1, pos = 3)

# Highest Mean Salary = 743.391

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

str(Data)

MaleSalary <- Data%>%select(gender, salary)%>%filter(gender == "Male")

FemaleSalary <- Data%>%select(gender, salary)%>%filter(gender == "Female")

par(mfrow = c(1,2))

M <- density(MaleSalary$salary)

plot(M, type="l", main="Male Salary Distribution", col = "Red")

N <- density(FemaleSalary$salary)

plot(N, type = "l", main = "Female Salary Distribution", col = "Blue")