# Assignment 1 : To calculate mean,trimmed mean,variance, standard deviation

# and draw a Frequency Distribution,Histogram ,Frequency Ploygon,Box plot and Ogive. Data set contains the starting salary for MBA Graduates

# including packages

install.packages('UsingR')

install.packages("ggplot2")

#using the libraries

library(UsingR)

library(ggplot2)

# Importing the .csv data into a dataframe

df = read.csv("C:/Users/pc/Desktop/Data Science/My Workspace/MBA Starting Salaries Data.csv")

age = df[,1] # Extracting the data for age

# Calculating mean,trimmed mean,variance, standard deviation for age

mean\_age = mean(age)

tmean\_age = mean(age,trim=1/10)

variance\_age = var(age)

stddev\_age = sqrt(variance\_age)

#Printing the values

print(paste0("Mean of Age = ", mean\_age))

print(paste0("Trimmed Mean (trimmed by 10%) of Age = ", tmean\_age))

print(paste0("Variance of Age = ", variance\_age))

print(paste0("Standard Deviation of Age = ", stddev\_age))

print("Summary of Age : ")

summary(age)

# Frequency Distribution for age

# calaculate the range

range\_age = range (age)

breaks = seq(20,50,by=5)

age\_cut = cut(age, breaks, right=FALSE)

age\_freq = table(age\_cut) # Frequency Distribution Table

age\_cumfreq = cumsum(age\_freq) # Cumulative Frequency

print("Frequency Distribution for Age : ")

age\_freq

print("Cumulative Frequency Distribution for Age : ")

age\_cumfreq

# Plotting with histogram

hist\_age=hist(age,main="Histogram for Age",xlab="Age (in Years)",freq=TRUE,xlim = c(20, max(age) + 2), col = "Steelblue3", right = F)

#Now we will create our x,y coordinates from the counts and mids variables

mp = c(min(hist\_age$mids) - (hist\_age$mids[2] - hist\_age$mids[1]), hist\_age$mids, max(hist\_age$mids) + (hist\_age$mids[2] - hist\_age$mids[1]))

freq = c(0, hist\_age$counts, 0)

# Plotting Frequency Polygon for age

hist\_age=hist(age,main="Histogram for Age",xlab="Age (in Years)",freq=TRUE,xlim = c(20, max(age) + 2), col = "Steelblue3", right = F)

lines(mp, freq,type = "b", pch = 20, col = "black", lwd = 3)

#Plotting BoxPlot

boxplot(age,main="Box Plot for Age", ylab="Age",ylim=c(20,50))

# Creating ogive

ucl = seq(from = min(hist\_age$breaks), to = max(hist\_age$breaks), by = hist\_age$breaks[2] - hist\_age$breaks[1])

ucl = c(20, ucl[-1])

cf = c(0, cumsum(hist\_age$counts))

plot(ucl, cf, main="Ogive for Age", type = "b", col = "violet", pch = 20) #Plotting Ogive

gmat\_tot = df[,3] # Extracting the data for gmat\_total

# Calculating mean,trimmed mean,variance, standard deviation for gmat\_tot

mean\_gmat = mean(gmat\_tot)

tmean\_gmat = mean(gmat\_tot,trim=1/10)

variance\_gmat = var(gmat\_tot)

stddev\_gmat = sqrt(variance\_gmat)

#Printing the values

print(paste0("Mean of GMAT Total = ", mean\_gmat))

print(paste0("Trimmed Mean (trimmed by 10%) of GMAT Total = ", tmean\_gmat))

print(paste0("Variance of GMAT Total = ", variance\_gmat))

print(paste0("Standard Deviation of GMAT Total = ", stddev\_gmat))

print("Summary of GMAT Total : ")

summary(gmat\_tot)

# Frequency Distribution for gmat\_tot

# calaculate the range , breaks, cut

range\_gmat = range (gmat\_tot)

breaks = seq(400,820,by=70)

gmat\_cut = cut(gmat\_tot, breaks, right=FALSE)

gmat\_freq = table(gmat\_cut) # Frequency Distribution Table

gmat\_cumfreq = cumsum(gmat\_freq) # Cumulative Frequency

print("Frequency Distribution for GMAT Total : ")

gmat\_freq

print("Cumulative Frequency Distribution for GMAT Total : ")

gmat\_cumfreq

# Plotting with histogram

hist\_gmat=hist(gmat\_tot,main="GMAT Score",xlab="GMAT Total",freq=TRUE,xlim = c(400, max(gmat\_tot) + 10), col = "Steelblue3", right = F)

#Now we will create our x,y coordinates from the counts and mids variables

mp = c(min(hist\_gmat$mids) - (hist\_gmat$mids[2] - hist\_gmat$mids[1]), hist\_gmat$mids, max(hist\_gmat$mids) + (hist\_gmat$mids[2] - hist\_gmat$mids[1]))

freq = c(0, hist\_gmat$counts, 0)

# Plotting Frequency Polygon for gmat\_total

hist\_gmat=hist(gmat\_tot,main="GMAT Score",xlab="GMAT Total",freq=TRUE,xlim = c(400, max(gmat\_tot) + 10), col = "Steelblue3", right = F)

lines(mp, freq, type = "b", pch = 20, col = "black", lwd = 3)

#Plotting BoxPlot

boxplot(gmat\_tot,main="Box Plot", ylab="GMAT Score",ylim=c(400,800))

# Creating ogive

ucl = seq(from = min(hist\_gmat$breaks), to = max(hist\_gmat$breaks), by = hist\_gmat$breaks[2] - hist\_gmat$breaks[1])

ucl = c(440, ucl[-1])

cf = c(0, cumsum(hist\_gmat$counts))

plot(ucl, cf, main="Ogive for GMAT Total", type = "b", col = "violet", pch = 20) #Plotting Ogive

work\_yrs = df[,10] # Extracting the data for work\_yrs

# Calculating mean,trimmed mean,variance, standard deviation for work\_years

mean\_workyears = mean(work\_yrs)

tmean\_workyears = mean(work\_yrs,trim=1/10)

variance\_workyears = var(work\_yrs)

stddev\_workyears = sqrt(variance\_workyears)

#Printing the values

print(paste0("Mean of Work Years = ", mean\_workyears))

print(paste0("Trimmed Mean (trimmed by 10%) of Work Years = ", tmean\_workyears))

print(paste0("Variance of Work Years = ", variance\_workyears))

print(paste0("Standard Deviation of Work Years = ", stddev\_workyears))

print("Summary of Work Years : ")

summary(work\_yrs)

# Frequency Distribution for work\_yrs

# calaculate the range , breaks, cut

range\_workyears = range (work\_yrs)

breaks = seq(0,25,by=5)

workyears\_cut = cut(work\_yrs, breaks, right=FALSE)

workyears\_freq = table(workyears\_cut) # Frequency Distribution Table

workyear\_cumfreq = cumsum(workyears\_freq) # Cumulative Frequency

print("Frequency Distribution for Work Years : ")

workyears\_freq

print("Cumulative Frequency Distribution for Work Years : ")

workyear\_cumfreq

# Plotting with histogram

hist\_workyears=hist(work\_yrs,main="Work Years",xlab="Experience (in Years)",freq=TRUE,xlim = c(0, max(work\_yrs) + 3), col = "Steelblue3", right = F)

#Now we will create our x,y coordinates from the counts and mids variables

mp = c(min(hist\_workyears$mids) - (hist\_workyears$mids[2] - hist\_workyears$mids[1]), hist\_workyears$mids, max(hist\_workyears$mids) + (hist\_workyears$mids[2] - hist\_workyears$mids[1]))

freq = c(0, hist\_workyears$counts, 0)

# Plotting Frequency Polygon for age

hist\_workyears=hist(work\_yrs,main="Work Years",xlab="Experience (in Years)",freq=TRUE,xlim = c(0, max(work\_yrs) + 3), col = "Steelblue3", right = F)

lines(mp, freq, type = "b", pch = 20, col = "black", lwd = 3)

#Plotting BoxPlot

boxplot(work\_yrs,main="Box Plot for Experience", ylab="Work Years",ylim=c(0,25))

# Creating ogive

ucl = seq(from = min(hist\_workyears$breaks), to = max(hist\_workyears$breaks), by = hist\_workyears$breaks[2] - hist\_workyears$breaks[1])

ucl = c(0, ucl[-1])

cf = c(0, cumsum(hist\_workyears$counts))

plot(ucl, cf, main="Ogive for Experience", type = "b", col = "violet", pch = 20) #Plotting Ogive

salary = df[,12] # Extracting the data for salary

remove<-c(998,999)

salary = salary[ - which(salary %in% remove)] # removing values 999 and 998

# Calculating mean,trimmed mean,variance, standard deviation for salary

mean\_salary = mean(salary)

tmean\_salary = mean(salary,trim=1/10)

variance\_salary = var(salary)

stddev\_salary = sqrt(variance\_salary)

#Printing the values

print(paste0("Mean of Salary = ", mean\_salary))

print(paste0("Trimmed Mean (trimmed by 10%) of Salary = ", tmean\_salary))

print(paste0("Variance of Salary = ", variance\_salary))

print(paste0("Standard Deviation of Salary = ", stddev\_salary))

print("Summary of Salary : ")

summary(salary)

# Frequency Distribution for Salary

# calaculate the range , breaks, cut

range\_salary = range (salary)

breaks = seq(0,240000,by=40000)

salary\_cut = cut(salary, breaks, right=FALSE)

salary\_freq = table(salary\_cut) # Frequency Distribution Table

salary\_cumfreq = cumsum(salary\_freq) # Cumulative Frequency

print("Frequency Distribution for Salary : ")

salary\_freq

print("Cumulative Frequency Distribution for Salary : ")

salary\_cumfreq

# Plotting with histogram

hist\_salary=hist(salary,main="Salary",xlab="Salary (in Rupees)",freq=TRUE,breaks =7 ,ylim = c(0,100),las=1, col = "Steelblue3", right = T)

#Now we will create our x,y coordinates from the counts and mids variables

mp = c(min(hist\_salary$mids) - (hist\_salary$mids[2] - hist\_salary$mids[1]), hist\_salary$mids, max(hist\_salary$mids) + (hist\_salary$mids[2] - hist\_salary$mids[1]))

freq = c(0, hist\_salary$counts, 0)

# Plotting Frequency Polygon for age

hist\_salary=hist(salary,main="Salary",xlab="Salary (in Rupees)",freq=TRUE,breaks =7 ,ylim = c(0,100),las=1, col = "Steelblue3", right = T)

lines(mp, freq, type = "b", pch = 20, col = "black", lwd = 3)

#Plotting BoxPlot

boxplot(salary,main="Box Plot for Salary", ylab="Salary",ylim=c(0,250000))

# Creating ogive

ucl = seq(from = min(hist\_salary$breaks), to = max(hist\_salary$breaks), by = hist\_salary$breaks[2] - hist\_salary$breaks[1])

ucl = c(0, ucl[-1])

cf = c(0, cumsum(hist\_salary$counts))

plot(ucl, cf, main="Ogive for Salary",type = "b", col = "violet", pch = 20) #Plotting Ogive

## EOP

Insights:

1. Majority of the MBA grads are in the age group of 24 to 28 where some of them are also 45+.
2. Around 80% of students have their GMAT score between 550 and 680.
3. Almost 3/4th of the students have an experience upto 5 years but few have an experience of 20+Years
4. 94% of the students are either not getting any job or getting a starting salary of less than 1.20 LPA.