STAT253 - Homework#1

Ahmet Elburuz Gürbüz - 150116024

Monday, March 23, 2020

1-)Each value measured in this assignment are quantitative. Age is a discrete value. Systolic blood pressure and diastolic blood pressure are continious valueThe measured are discrete variable. Each value measured in this assignment are multivariate.

2-)There are relative frequency histogram, stem-leaf and dotplot graphics. With these graphs, we can detect frequencies, repeating data, outlier data. Determine the shape of the distribution ,unusual features fdifferences in the distribution for males and the distributions for females etc.

3-)

Hw1\_Dataset = read.csv("dataset.csv")  
manSYSBP = Hw1\_Dataset[Hw1\_Dataset$GENDER == 0, "SYSBP"]  
womenSYSBP = Hw1\_Dataset[Hw1\_Dataset$GENDER == 1, "SYSBP"]  
manDIASBP = Hw1\_Dataset[Hw1\_Dataset$GENDER == 0, "DIASBP"]  
womenDIASBP = Hw1\_Dataset[Hw1\_Dataset$GENDER == 1, "DIASBP"]  
#a-----------------------------------------------  
manMean1 = mean(manSYSBP)  
manMean1

## [1] 105.97

womenMean1 = mean(womenSYSBP)  
womenMean1

## [1] 109.89

manMean2 = mean(manDIASBP)  
manMean2

## [1] 72.29

womenMean2 = mean(womenDIASBP)  
womenMean2

## [1] 70.98

#b-----------------------------------------------  
manVar1 = var(manSYSBP)  
manVar1

## [1] 12647.73

manVar2 = var(manDIASBP)  
manVar2

## [1] 113.5615

womenVar1 = var(womenSYSBP)  
womenVar1

## [1] 139.1292

womenVar2 = var(womenDIASBP)  
womenVar2

## [1] 69.91879

#c-----------------------------------------------  
mansd1 = sd(manSYSBP)  
mansd1

## [1] 112.4621

mansd2 = sd(manDIASBP)  
mansd2

## [1] 10.65652

womensd1 = sd(womenSYSBP)  
womensd1

## [1] 11.7953

womensd2 = sd(womenDIASBP)  
womensd2

## [1] 8.361746

#d-----------------------------------------------  
manQuantile1=quantile(manSYSBP,0.25)  
manQuantile1

## 25%   
## 108

manqQuantile2=quantile(manSYSBP,0.75)  
manqQuantile2

## 75%   
## 122.5

manQuantile3=quantile(manDIASBP,0.25)  
manQuantile3

## 25%   
## 66

manQuantile4=quantile(manDIASBP,0.75)  
manQuantile4

## 75%   
## 80

womenQuantile1=quantile(womenSYSBP,0.25)  
womenQuantile1

## 25%   
## 100

womenQuantile2=quantile(womenSYSBP,0.75)  
womenQuantile2

## 75%   
## 118

womenQuantile3=quantile(womenDIASBP,0.25)  
womenQuantile3

## 25%   
## 67.5

womenQuantile4=quantile(womenDIASBP,0.75)  
womenQuantile4

## 75%   
## 76.5

#e-----------------------------------------------  
menmin1 = min(womenSYSBP)  
menmin1

## [1] 86

menmax1 = max(womenSYSBP)  
menmax1

## [1] 150

menmin2 = min(womenDIASBP)  
menmin2

## [1] 40

menmax2 = max(womenDIASBP)  
menmax2

## [1] 94

womenmin1 = min(womenSYSBP)  
womenmin1

## [1] 86

womenmax1 = max(womenSYSBP)  
womenmax1

## [1] 150

womenmin2 = min(womenDIASBP)  
womenmin2

## [1] 40

womenmax2 = max(womenDIASBP)  
womenmax2

## [1] 94

#f-----------------------------------------------  
menRange1 = range(manSYSBP)  
menRange1

## [1] -999 170

menRange2 = range(manDIASBP)  
menRange2

## [1] 50 110

womenRange1 = range(womenSYSBP)  
womenRange1

## [1] 86 150

womenRange2 = range(womenDIASBP)  
womenRange2

## [1] 40 94

#g-----------------------------------------------

For Systolic Blood Pressure,dividing the range by the sample standard deviation.

For Diastolic Blood Pressure,dividing the range by the sample standard deviation.

menRate1 = menRange1/mansd1   
menRate1

## [1] -8.882992 1.511620

menRate2 = menRange2/mansd2   
menRate2

## [1] 4.691961 10.322315

womenRate1 = womenRange1/womensd1  
womenRate1

## [1] 7.291037 12.716926

womenRate2 = womenRange2/womensd2  
womenRate2

## [1] 4.78369 11.24167

#h-----------------------------------------------  
menMedian1 = median(manSYSBP)  
menMedian1

## [1] 116.5

menMedian2 = median(manDIASBP)  
menMedian2

## [1] 70

womenMedian1 = median(womenSYSBP)  
womenMedian1

## [1] 110

womenMedian2 = median(womenDIASBP)  
womenMedian2

## [1] 70

#i-----------------------------------------------  
menIQR1 = IQR(manSYSBP)  
menIQR1

## [1] 14.5

menIQR2 = IQR(manDIASBP)  
menIQR2

## [1] 14

womenIQR1 = IQR(womenSYSBP)  
womenIQR1

## [1] 18

womenIQR2 = IQR(womenDIASBP)  
womenIQR2

## [1] 9

#j-----------------------------------------------  
menSummary1 = fivenum(manSYSBP)  
menSummary1

## [1] -999.0 108.0 116.5 123.0 170.0

menSummary2 = fivenum(manDIASBP)  
menSummary2

## [1] 50 66 70 80 110

womenSummary1 = fivenum(womenSYSBP)  
womenSummary1

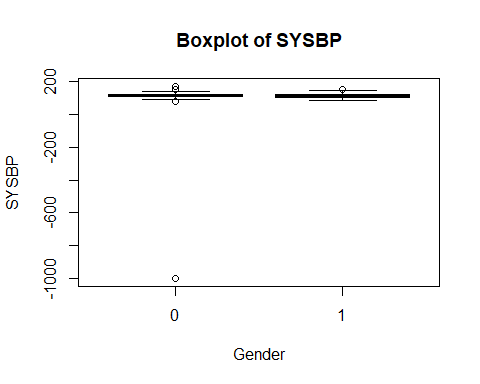
## [1] 86 100 110 118 150

womenSummary2 = fivenum(womenDIASBP)  
womenSummary2

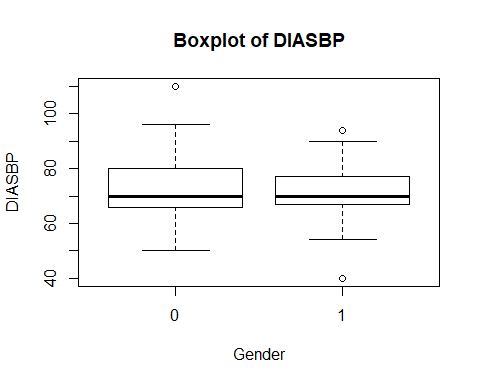
## [1] 40 67 70 77 94

#k-----------------------------------------------  
boxplot(Hw1\_Dataset$SYSBP ~ Hw1\_Dataset$GENDER,   
main = "Boxplot of SYSBP",  
xlab="Gender",   
ylab="SYSBP")

As it can be seen from the ‘Boxplot of SYSBP’ graph has 0 outliers and graphs is skewed to left.And on the other hand ‘Boxplot of DIASBP’ graph has 3 outliers and it is skewed to left.



boxplot(Hw1\_Dataset$DIASBP ~ Hw1\_Dataset$GENDER,  
main = "Boxplot of DIASBP",  
xlab="Gender",  
ylab="DIASBP")



#l-----------------------------------------------  
stem(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==0])

##   
## The decimal point is 2 digit(s) to the right of the |  
##   
## -10 | 0  
## -8 |   
## -6 |   
## -4 |   
## -2 |   
## -0 |   
## 0 | 89999000000000011111111111111111111111122222222222222222222222222222+19

stem(Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==0])

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 5 | 002668  
## 6 | 0000000000224444456666666888888  
## 7 | 00000000000000000022222245666666888  
## 8 | 00000000000000568  
## 9 | 0000000026  
## 10 |   
## 11 | 0

stem(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==1])

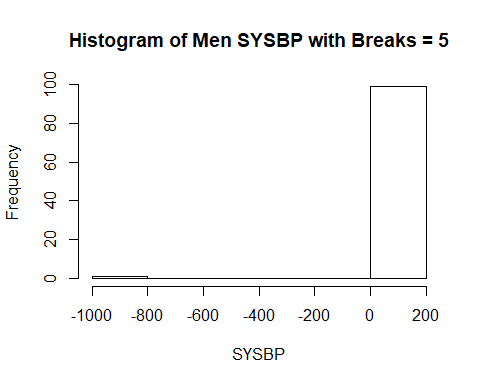
##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 8 | 688  
## 9 | 0000688888  
## 10 | 00000000000000222244456688888888  
## 11 | 00000000000000000000224444568888  
## 12 | 00000000000244568  
## 13 | 004  
## 14 | 04  
## 15 | 0

stem(Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==1])

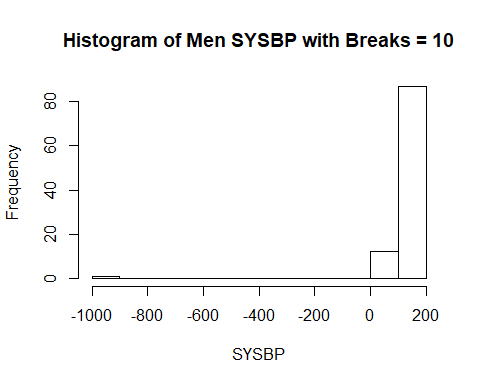
##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 4 | 0  
## 5 | 458  
## 6 | 0000000000000024446668888  
## 7 | 00000000000000000000000000000022444445666666668888  
## 8 | 0000000000000022244  
## 9 | 04

#m-----------------------------------------------  
hist(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==0],  
main = "Histogram of Men SYSBP with Breaks = 5",  
xlab= "SYSBP",  
breaks = 5)

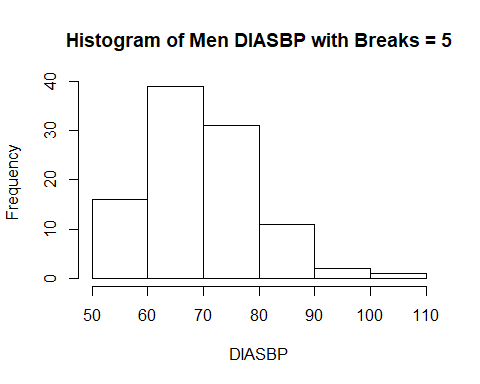
If we compare two graphs for both cases,it is obvious that we can see the distribution better by using more sub-intervals.



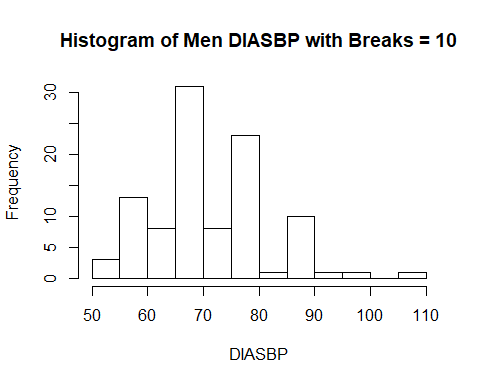
hist(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==0],  
main = "Histogram of Men SYSBP with Breaks = 10",  
xlab= "SYSBP",  
breaks = 10)



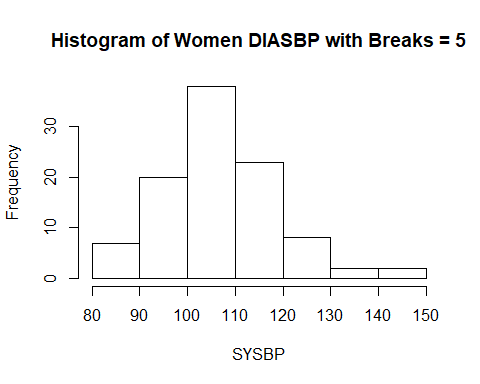
hist(Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==0],  
main = "Histogram of Men DIASBP with Breaks = 5",  
xlab= "DIASBP",  
breaks = 5)



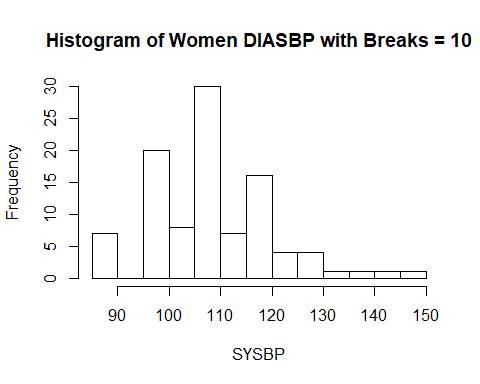
hist(Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==0],  
main = "Histogram of Men DIASBP with Breaks = 10",  
xlab= "DIASBP",  
breaks = 10)



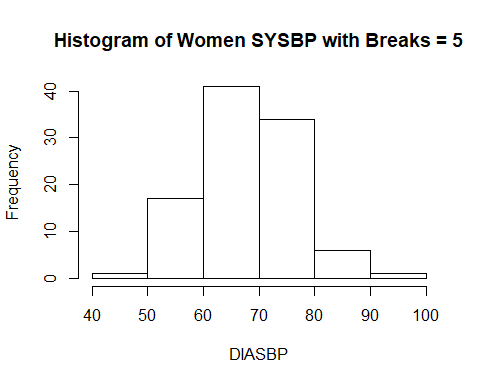
hist(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==1],  
main = "Histogram of Women DIASBP with Breaks = 5",  
xlab= "SYSBP",  
breaks = 5)



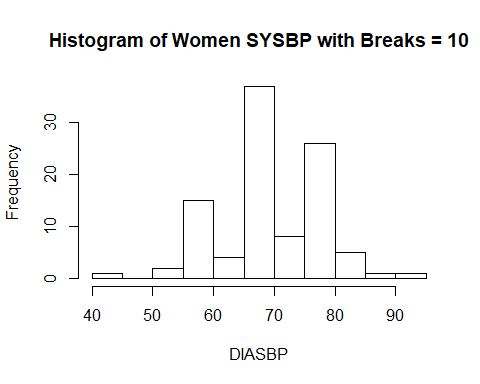
hist(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==1],  
main = "Histogram of Women DIASBP with Breaks = 10",  
xlab= "SYSBP",  
breaks = 10)



hist(Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==1],  
main = "Histogram of Women SYSBP with Breaks = 5",  
xlab= "DIASBP",  
breaks = 5)

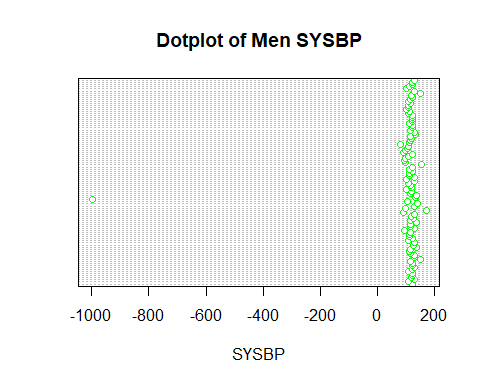


hist(Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==1],  
main = "Histogram of Women SYSBP with Breaks = 10",  
xlab= "DIASBP",  
breaks = 10)

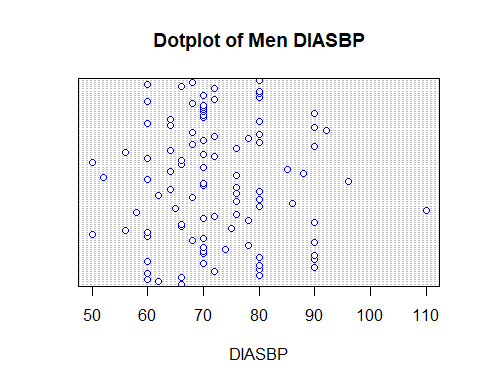


#n-----------------------------------------------  
dotchart(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==0],  
color = "green",  
main = "Dotplot of Men SYSBP",  
xlab= "SYSBP")

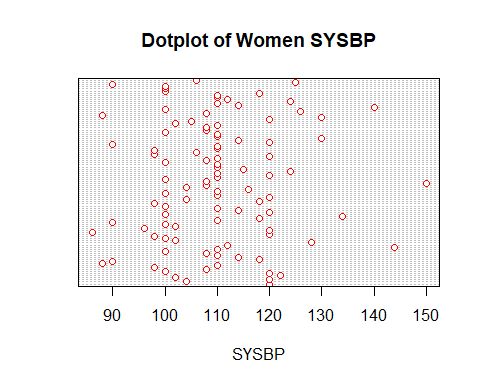
As it can be seen from dot plot graphs,both of them has some outliers and shapes of the graphs are not mound shaped.



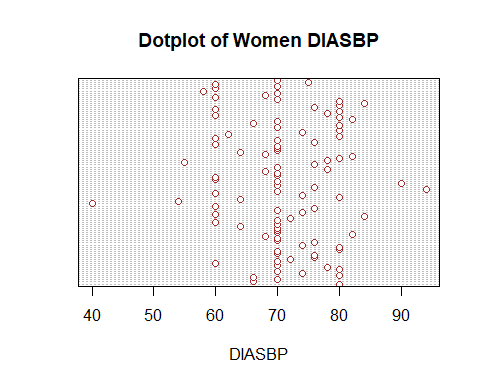
dotchart(Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==0],  
color = "blue",  
main = "Dotplot of Men DIASBP",  
xlab= "DIASBP")



dotchart(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==1],  
color = "red",  
main = "Dotplot of Women SYSBP",  
xlab= "SYSBP")



dotchart(Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==1],  
color = "brown",  
main = "Dotplot of Women DIASBP",  
xlab= "DIASBP")



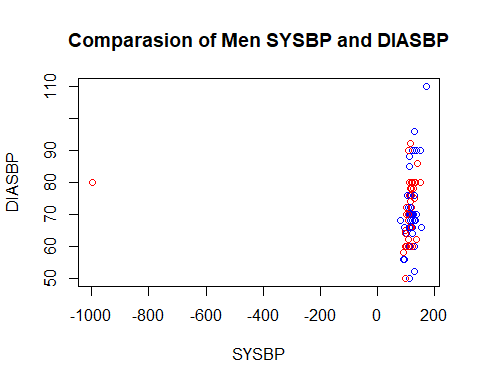
## o) Tchebysheff’s Theorem

Tchebysheff’s Theorem can be applied to any set of measurements(large or small,mound-shaped or skewed) and it is always satisfied.So we can use Tchebysheff’s Theorem for both cases.

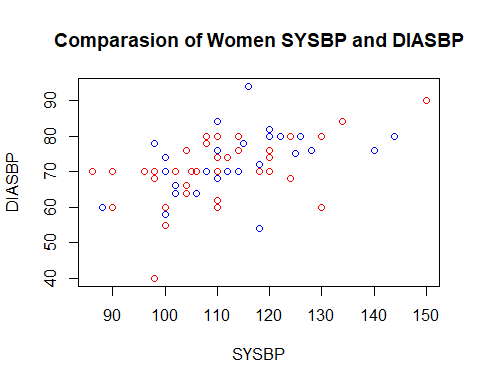
## p) The Empirical Rule

The Empirical Rule can be applied only when the data tend to be roughly mound-shaped.Since the data is not mound-shaped(bell-shaped,symmetric) and we have outliers in both cases we can not use Empirical Rule.

#q-----------------------------------------------  
plot(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==0],Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==0],  
col=c("blue","red"),  
main = "Comparasion of Men SYSBP and DIASBP",  
xlab= "SYSBP",  
ylab = "DIASBP")



plot(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==1],Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==1],  
col=c("blue","red"),  
main = "Comparasion of Women SYSBP and DIASBP",  
xlab= "SYSBP",  
ylab = "DIASBP")

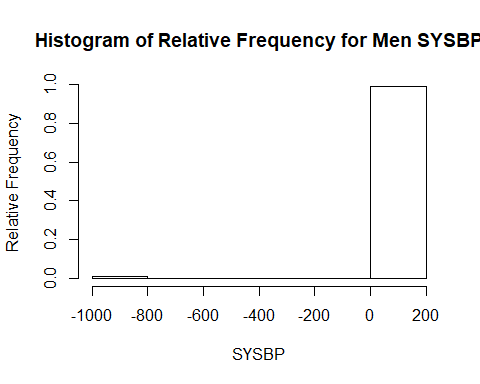


#r-----------------------------------------------  
library(HistogramTools)

## Warning: package 'HistogramTools' was built under R version 3.6.3

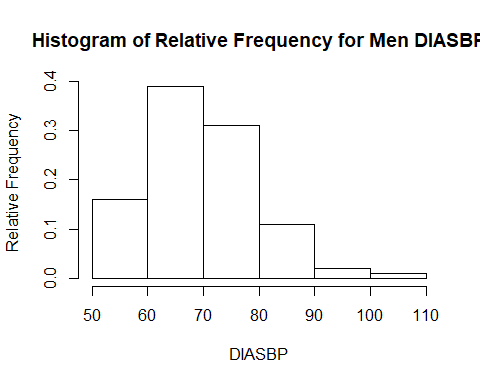
PlotRelativeFrequency(hist(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==0],plot = F),  
main = "Histogram of Relative Frequency for Men SYSBP",  
xlab= "SYSBP")

Skewed Left



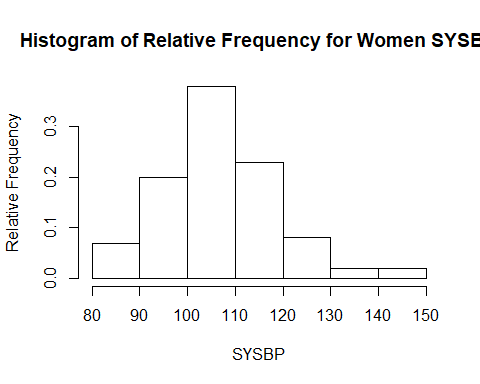
PlotRelativeFrequency(hist(Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==0],plot = F),  
main = "Histogram of Relative Frequency for Men DIASBP",  
xlab= "DIASBP")

Skewed Right



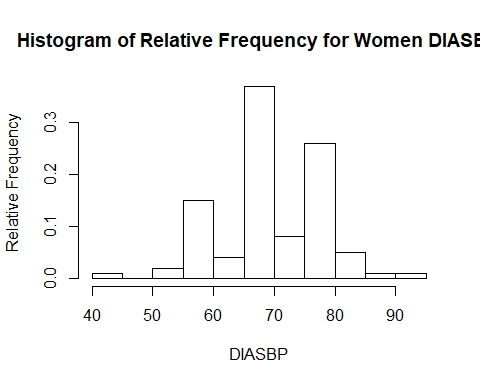
PlotRelativeFrequency(hist(Hw1\_Dataset$SYSBP[Hw1\_Dataset$GENDER==1],plot = F),  
main = "Histogram of Relative Frequency for Women SYSBP",  
xlab= "SYSBP")

Skewed Right



PlotRelativeFrequency(hist(Hw1\_Dataset$DIASBP[Hw1\_Dataset$GENDER==1],plot = F),  
main = "Histogram of Relative Frequency for Women DIASBP",  
xlab= "DIASBP")

Skewed Left



#s-----------------------------------------------  
zScoreMenLargestSYSBP = (menmax1-manMean1)/mansd1  
zScoreMenLargestSYSBP

## [1] 0.3915096

zScoreMenLargestDIASBP = (menmax2-manMean2)/mansd2  
zScoreMenLargestDIASBP

## [1] 2.03725

zScoreMenSmallestSYSBP = (menmin1-manMean1)/mansd1  
zScoreMenSmallestSYSBP

## [1] -0.1775709

zScoreMenSmallestDIASBP = (menmin2-manMean2)/mansd2  
zScoreMenSmallestDIASBP

## [1] -3.030069

zScoreWomenLargestSYSBP = (womenmax1-womenMean1)/womensd1  
zScoreWomenLargestSYSBP

## [1] 3.400506

zScoreWomenLargestDIASBP = (womenmax2-womenMean2)/womensd2  
zScoreWomenLargestDIASBP

## [1] 2.753014

zScoreWomenSmallestSYSBP = (womenmin1-womenMean1)/womensd1  
zScoreWomenSmallestSYSBP

## [1] -2.025382

zScorWomenSmallestDIASBP = (womenmin2-womenMean2)/womensd2  
zScorWomenSmallestDIASBP

## [1] -3.704968

## Systolic Blood Pressure

## z-scores between -2 and 2 are not unusual.z-scores shouldn’t be more than 3 in absolute value.z-scores larger than 3 in absolute value would indicate a possible outlier.So in that case we can indicate smallest observation(-2.025382) as a somewhat unusual and largest observation(3.400506) as a outlier.

## Diastolic Blood Pressure

As i mentioned about the reason above we can indicate smallest observation(-3.704968) as a somewhat unusual and largest observation(2.753014) as a outlier.

# R code