class practice

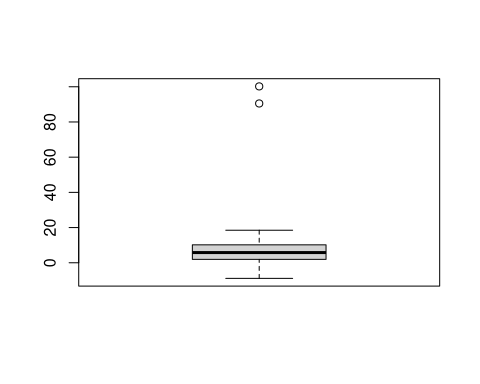
Justin Markey

9/13/2021

#import data

data <- read.csv("class practice sep-13.csv")

with (data,  
 boxplot(X1))



#define range

bottomq<-with(data,quantile(X1,0.01,na.rm=TRUE))  
topq<-with(data,quantile(X1,0.99,na.rm=TRUE))

#identify outliers

data$outlier.label<-with(data,  
 ifelse(X1>topq,"Outlier",  
 ifelse(X1<bottomq,"Outlier","Normal Data")))  
subset(data,outlier.label=="Outlier")

## X1 X2 outlier.label  
## 7 -8.88 B Outlier  
## 17 100.20 A Outlier

#winsorize

data$X1.Winsorized<-with(data,  
 ifelse(X1>topq,topq,  
 ifelse(X1<bottomq,bottomq,X1)))

#imputation

unconditional.mean<-with(data,mean(data$X1.Winsorized,na.rm=TRUE))

#replace

data$X1.Winsorized.Imputed<-with(data,  
 ifelse(is.na(X1.Winsorized),  
 unconditional.mean,  
 X1.Winsorized))

#check cleaned data

summary(data$X1.Winsorized.Imputed)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -8.356 2.030 6.155 12.280 11.500 98.066

#conditional mean

mean.A<-with(subset(data,X2=="A"),  
 mean(X1.Winsorized,na.rm=TRUE))  
mean.B<-with(subset(data,X2=="B"),  
 mean(X1.Winsorized,na.rm=TRUE))

data$X1.Winsorized.Imputed.V2<-with(data,  
ifelse(  
 is.na(X1.Winsorized)&X2=="A",  
 mean.A,  
 ifelse(  
 is.na(X1.Winsorized)&X2=="B",  
 mean.B,  
 X1.Winsorized  
 )  
)  
)

summary(data)

## X1 X2 outlier.label X1.Winsorized   
## Min. : -8.88 Length:24 Length:24 Min. :-8.356   
## 1st Qu.: 1.92 Class :character Class :character 1st Qu.: 1.920   
## Median : 5.78 Mode :character Mode :character Median : 5.780   
## Mean : 12.35 Mean :12.280   
## 3rd Qu.: 10.20 3rd Qu.:10.200   
## Max. :100.20 Max. :98.066   
## NA's :1 NA's :1   
## X1.Winsorized.Imputed X1.Winsorized.Imputed.V2  
## Min. :-8.356 Min. :-8.356   
## 1st Qu.: 2.030 1st Qu.: 2.030   
## Median : 6.155 Median : 6.155   
## Mean :12.280 Mean :12.273   
## 3rd Qu.:11.500 3rd Qu.:11.465   
## Max. :98.066 Max. :98.066   
##