HW2

Justin Markey

9/15/2021

# Import the Data

setwd("C:/Projects/R/bizanalytics/data")  
dirtydata <- read.csv("HW2 Data.csv", stringsAsFactors = TRUE)

# Question 1

## define range using normal boxplot ranges

q1 = with(dirtydata,quantile(X,0.25,na.rm=TRUE))  
q3 = with(dirtydata,quantile(X,0.75,na.rm=TRUE))  
  
lowerrange<-q1-1.5\*(q3-q1)  
upperrange<-q3+1.5\*(q3-q1)

upperrange = 5.1 lowerrange = -0.46 q1 = 1.63 q3 = 3.02

## treat outliers using winsor method

dirtydata$X1 <- with(dirtydata,  
 ifelse(X > upperrange, upperrange,  
 ifelse(X < lowerrange, lowerrange, X)))  
summary(dirtydata)

## X Z X1   
## Min. :-1.036 A:348 Min. :-0.4596   
## 1st Qu.: 1.626 B:152 1st Qu.: 1.6264   
## Median : 2.320 Median : 2.3204   
## Mean : 2.316 Mean : 2.3135   
## 3rd Qu.: 3.017 3rd Qu.: 3.0171   
## Max. : 5.901 Max. : 5.1030   
## NA's :23 NA's :23

### Conclusion

The upperrange and lowerrange values are equal to the X1(winsorized data) Max and Min. This means that the outliers have been treated successfully.

# Question 2

## define range of upper 99% and lower 1%

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

topq = 5.14 bottomq = -0.33

## treat outliers using winsor method

dirtydata$X2 <- with(dirtydata,  
 ifelse(X > topq, topq,  
 ifelse(X < bottomq, bottomq, X)))  
summary(dirtydata)

## X Z X1 X2   
## Min. :-1.036 A:348 Min. :-0.4596 Min. :-0.3295   
## 1st Qu.: 1.626 B:152 1st Qu.: 1.6264 1st Qu.: 1.6264   
## Median : 2.320 Median : 2.3204 Median : 2.3204   
## Mean : 2.316 Mean : 2.3135 Mean : 2.3150   
## 3rd Qu.: 3.017 3rd Qu.: 3.0171 3rd Qu.: 3.0171   
## Max. : 5.901 Max. : 5.1030 Max. : 5.1356   
## NA's :23 NA's :23 NA's :23

### Conclusion

Again the top 99% value is equal to X2 Max and lower 1% value is equal to X2 Min. This proves every value outside the range has been winsorized.

# Question 3

## Calculate unconditional mean and replace na values with that value

X1mean.unconditional<-with(dirtydata, mean(X1, na.rm=TRUE))  
  
dirtydata$X3 <- with(dirtydata,  
 ifelse(is.na(X1),  
 X1mean.unconditional,  
 X1))  
  
summary(dirtydata)

## X Z X1 X2 X3   
## Min. :-1.036 A:348 Min. :-0.4596 Min. :-0.3295 Min. :-0.4596   
## 1st Qu.: 1.626 B:152 1st Qu.: 1.6264 1st Qu.: 1.6264 1st Qu.: 1.6714   
## Median : 2.320 Median : 2.3204 Median : 2.3204 Median : 2.3135   
## Mean : 2.316 Mean : 2.3135 Mean : 2.3150 Mean : 2.3135   
## 3rd Qu.: 3.017 3rd Qu.: 3.0171 3rd Qu.: 3.0171 3rd Qu.: 2.9795   
## Max. : 5.901 Max. : 5.1030 Max. : 5.1356 Max. : 5.1030   
## NA's :23 NA's :23 NA's :23

Unconditional mean = 2.313479331

### Conclusion

In the X3 column we can see that there is no NA values because they’ve been successfully replaced with the unconditional mean.

# Question 4

## Calculate conditional mean for each factor

mean.A<-with(subset(dirtydata, Z == "A"),  
 mean(X1, na.rm=TRUE))  
mean.B<-with(subset(dirtydata, Z == "B"),  
 mean(X1, na.rm=TRUE))

Mean of A = 2.022240891 Mean of B = 1.980314933

## Replace na values with the mean for each categorical variable

dirtydata$X4 <- with(dirtydata,  
 ifelse(  
 is.na(X1) & Z =="A",  
 mean.A,  
 ifelse(  
 is.na(X1) & Z == "B",  
 mean.B,  
 X1)))  
  
summary(dirtydata)

## X Z X1 X2 X3   
## Min. :-1.036 A:348 Min. :-0.4596 Min. :-0.3295 Min. :-0.4596   
## 1st Qu.: 1.626 B:152 1st Qu.: 1.6264 1st Qu.: 1.6264 1st Qu.: 1.6714   
## Median : 2.320 Median : 2.3204 Median : 2.3204 Median : 2.3135   
## Mean : 2.316 Mean : 2.3135 Mean : 2.3150 Mean : 2.3135   
## 3rd Qu.: 3.017 3rd Qu.: 3.0171 3rd Qu.: 3.0171 3rd Qu.: 2.9795   
## Max. : 5.901 Max. : 5.1030 Max. : 5.1356 Max. : 5.1030   
## NA's :23 NA's :23 NA's :23   
## X4   
## Min. :-0.4596   
## 1st Qu.: 1.6714   
## Median : 2.2835   
## Mean : 2.3135   
## 3rd Qu.: 2.9828   
## Max. : 5.1030   
##

### Conclusion

Good news is that the imputation was successful because their is no NA values, however this imputation is slightly different than the unconditional imputation. The values remained the same except for the Median and 3rd Quantile. The dataset has little change when comparing the unconditional mean imputation and the conditional mean imputation.