Assignment4

Aishat Olatunji

2022-09-19

#Number one  
#file.choose()  
income\_datapre<-read.csv( "/Users/aishatolatunji/Downloads/Income Dirty Data.csv")  
head(income\_datapre)

## ID gender age income tax..15..  
## 1 1 Women 21 147168 22075.20  
## 2 2 Female 29 119595 17939.25  
## 3 3 Female 56 87770 13165.50  
## 4 4 <NA> 21 54259 8138.85  
## 5 5 Male 28 NA 160230.00  
## 6 6 Woman 0 128326 19248.90

summary(income\_datapre)

## ID gender age income   
## Min. : 1.0 Length:1000 Min. :-1.00 Min. : -1170   
## 1st Qu.: 250.8 Class :character 1st Qu.:24.00 1st Qu.: 66567   
## Median : 500.5 Mode :character Median :37.00 Median : 93078   
## Mean : 500.5 Mean :35.09 Mean : 90436   
## 3rd Qu.: 750.2 3rd Qu.:49.00 3rd Qu.:122344   
## Max. :1000.0 Max. :60.00 Max. :149991   
## NA's :109   
## tax..15..   
## Min. : -7119   
## 1st Qu.: 11128   
## Median : 15694   
## Mean : 55098   
## 3rd Qu.: 19720   
## Max. :1362450   
## NA's :93

prod(dim(income\_datapre))# The total numbers of completedata set before imputations

## [1] 5000

attach(income\_datapre)  
# This helps in understanding the numbers of observations and numbers of attributes in the dataset  
dim(income\_datapre)

## [1] 1000 5

colnames(income\_datapre)

## [1] "ID" "gender" "age" "income" "tax..15.."

# The above results helps to know the names of each attributes.  
  
sum(is.na(income\_datapre))#The total number of missing values in the data set is 290 variables

## [1] 290

# The total numbers of NA's availaible in each attributes.  
sum(is.na(gender))

## [1] 88

sum(is.na(age))

## [1] 0

sum(is.na(income))

## [1] 109

sum(is.na(tax..15..))

## [1] 93

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

### Number 1b. .....  
No\_na\_income=na.omit(income\_datapre)  
summary(No\_na\_income)

## ID gender age income   
## Min. : 1.0 Length:733 Min. :-1.00 Min. : -1170   
## 1st Qu.: 243.0 Class :character 1st Qu.:25.00 1st Qu.: 65468   
## Median : 523.0 Mode :character Median :38.00 Median : 90281   
## Mean : 509.1 Mean :35.89 Mean : 88757   
## 3rd Qu.: 764.0 3rd Qu.:49.00 3rd Qu.:118489   
## Max. :1000.0 Max. :60.00 Max. :149991   
## tax..15..   
## Min. : -7119   
## 1st Qu.: 10846   
## Median : 15264   
## Mean : 57939   
## 3rd Qu.: 19672   
## Max. :1362450

dim(No\_na\_income)

## [1] 733 5

prod(dim(income\_datapre)) # the total numbers of observations present in the data set.

## [1] 5000

#The sum of the total numbers of complete data based on the dimensions and the percentage of the total numbers are given below;  
prod(dim(income\_datapre)) - sum(is.na(income\_datapre))

## [1] 4710

# the output above shows that the total numbers of complete data is 4710.  
# and the percentage is  
(prod(dim(income\_datapre))-sum(is.na(income\_datapre)))/prod(dim(income\_datapre))\*100

## [1] 94.2

#The result below shows that the percentage of complete data set is 94.2%

#number2  
library(deducorrect)

## Loading required package: editrules

## Loading required package: igraph

##   
## Attaching package: 'igraph'

## The following objects are masked from 'package:dplyr':  
##   
## as\_data\_frame, groups, union

## The following objects are masked from 'package:stats':  
##   
## decompose, spectrum

## The following object is masked from 'package:base':  
##   
## union

##   
## Attaching package: 'editrules'

## The following objects are masked from 'package:igraph':  
##   
## blocks, normalize

## The following object is masked from 'package:dplyr':  
##   
## contains

cr<- correctionRules(expression(if(!is.na(age) & age <18) age <- NA,  
if(!is.na(income)& income <1) income <-NA,   
if(!is.na(tax..15..)& tax..15..<1) tax..15.. <-NA  
))  
cr

## Object of class 'correctionRules'  
## ## 1-------  
## if (!is.na(age) & age < 18) age <- NA   
## ## 2-------  
## if (!is.na(income) & income < 1) income <- NA   
## ## 3-------  
## if (!is.na(tax..15..) & tax..15.. < 1) tax..15.. <- NA

corule<- correctWithRules(cr,income\_datapre)  
data2<-corule$corrected  
dim(data2)

## [1] 1000 5

sum(is.na(data2))

## [1] 442

#The numbers of missing value in data2 is 442  
#NROW(na.omit(data2))  
prod(dim(data2))

## [1] 5000

# The result below shows the total numbers of observation without the missing values which is 4558.  
(prod(dim(data2))-sum(is.na(data2)))

## [1] 4558

# the percentage of numbers without errors is 91.16 based on the output below.  
(prod(dim(data2))-sum(is.na(data2)))/prod(dim(data2))\*100

## [1] 91.16

#number 3  
cr2<- correctionRules (expression(if(!is.na(age) & age <10)age <- NA,  
if(!is.na(gender) & gender =="Man")gender <- "Male",  
if(!is.na(gender) & gender =="Men")gender <- "Male",  
if(!is.na(gender) & gender =="Woman")gender <- "Female",  
if(!is.na(gender) & gender =="Women")gender <- "Female",  
if(!is.na(income) & income<1) income <-NA,   
if(!is.na(tax..15..) & tax..15..<1) tax..15.. <-NA,   
if(!is.na(tax..15..) & is.na(income)) income <- (tax..15../0.15),  
if(!is.na(income) & is.na(tax..15..)) tax..15.. <- (income\*0.15)  
))  
  
corules<- correctWithRules(cr2, income\_datapre)  
data3 <-corules$corrected  
sum(is.na(data3))

## [1] 216

(prod(dim(data3))-sum(is.na(data3)))

## [1] 4784

#The total number of completed data here is 4784  
  
#we still have 216 empty values, after we fixed/derived some tax and income values  
# what i observed based on the result above is that, we did some data cleansing to make our data consistence and change all our iregular gender patterns to the right format.  
#Based on the answer yes i believe we can obtain from the income column to reduce the the numbers of missing values contained in the data set.

#number 4  
summary(data3)

## ID gender age income   
## Min. : 1.0 Length:1000 Min. :18.00 Min. : 70   
## 1st Qu.: 250.8 Class :character 1st Qu.:28.00 1st Qu.: 70634   
## Median : 500.5 Mode :character Median :40.00 Median : 97838   
## Mean : 500.5 Mean :38.96 Mean : 111995   
## 3rd Qu.: 750.2 3rd Qu.:50.00 3rd Qu.: 124674   
## Max. :1000.0 Max. :60.00 Max. :5200320   
## NA's :98 NA's :15   
## tax..15..   
## Min. : 10.5   
## 1st Qu.: 11173.2   
## Median : 15849.6   
## Mean : 52111.7   
## 3rd Qu.: 19722.8   
## Max. :1362450.0   
## NA's :15

dim(data3)

## [1] 1000 5

# the summary before the imputations showws that we have a total number of 5000 observations  
  
library(VIM)

## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##   
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':  
##   
## sleep

data\_imputation <- kNN(data3)  
sum(is.na(data\_imputation))

## [1] 0

summary(data\_imputation)

## ID gender age income   
## Min. : 1.0 Length:1000 Min. :18.00 Min. : 70   
## 1st Qu.: 250.8 Class :character 1st Qu.:28.00 1st Qu.: 70872   
## Median : 500.5 Mode :character Median :40.00 Median : 98365   
## Mean : 500.5 Mean :38.93 Mean : 111911   
## 3rd Qu.: 750.2 3rd Qu.:49.00 3rd Qu.: 124320   
## Max. :1000.0 Max. :60.00 Max. :5200320   
## tax..15.. ID\_imp gender\_imp age\_imp   
## Min. : 10.5 Mode :logical Mode :logical Mode :logical   
## 1st Qu.: 11226.0 FALSE:1000 FALSE:912 FALSE:902   
## Median : 15838.6 TRUE :88 TRUE :98   
## Mean : 51559.6   
## 3rd Qu.: 19706.7   
## Max. :1362450.0   
## income\_imp tax..15..\_imp   
## Mode :logical Mode :logical   
## FALSE:985 FALSE:985   
## TRUE :15 TRUE :15   
##   
##   
##

#After imputations we have no missing values as all our missing values has been replaced by infered metthod.