R Notebook

Code

3/4/2020

Hide

```
# PROBLEM 1 :

custdatarev <- read.table("custdata.txt",header = TRUE, sep = '\t')

custdatarev
```

custid <int>	sex <fctr>	is.employed <lgl>	inco… <int>	marital.stat <fctr>	health.ins <lgl>	housing.type <fctr>
2068	F	*NA*	11300	Married	TRUE	Homeowner free and clear
2073	F	*NA*	0	Married	TRUE	Rented
2848	M	TRUE	4500	Never Married	FALSE	Rented
5641	M	TRUE	20000	Never Married	FALSE	Occupied with no rent
6369	F	TRUE	12000	Never Married	TRUE	Rented
8322	F	TRUE	180000	Never Married	TRUE	Homeowner with mortgage/loan
8521	M	TRUE	120000	Never Married	TRUE	Homeowner free and clear
12195	M	TRUE	40000	Married	TRUE	Rented
14989	M	*NA*	9400	Married	TRUE	Rented
15917	F	TRUE	24000	Divorced/Separated	TRUE	Homeowner free and clear

1-10 of 1,000 rows | 1-7 of 11 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
str(custdatarev$num.vehicles) #an int type
```

```
 int [1:1000] 2 3 3 0 1 1 1 3 2 1 ...
```

Hide

```
custdataNoNA <- subset(custdatarev, custdatarev$num.vehicle >= 0) #get rid of NA values
custdataNoNA
```

custid <int>	sex <fctr>	is.employed <lgl>	inco… <int>	marital.stat <fctr>	health.ins <lgl>	housing.type <fctr>

	custid <int>	sex <fctr>	is.employed <lgl>	inco… <int>	marital.stat <fctr>	health.ins <lgl>	housing.type <fctr>
1	2068	F	*NA*	11300	Married	TRUE	Homeowner free and clear
2	2073	F	*NA*	0	Married	TRUE	Rented
3	2848	M	TRUE	4500	Never Married	FALSE	Rented
4	5641	M	TRUE	20000	Never Married	FALSE	Occupied with no rent
5	6369	F	TRUE	12000	Never Married	TRUE	Rented
6	8322	F	TRUE	180000	Never Married	TRUE	Homeowner with mortgage/l
7	8521	M	TRUE	120000	Never Married	TRUE	Homeowner free and clear
8	12195	M	TRUE	40000	Married	TRUE	Rented
9	14989	M	*NA*	9400	Married	TRUE	Rented
10	15917	F	TRUE	24000	Divorced/Separated	TRUE	Homeowner free and clear

1-10 of 944 rows | 1-8 of 11 columns Previous **1** 2 3 4 5 6 … 95 Next

Hide

```
mean(custdataNoNA[,9]) #the mean of the 9th column (num.vehicle column)
```

```
[1] 1.916314
```

Hide

```
for(i in 1:length(custdatarev[,9])) {

    if(is.na(custdatarev[i,9])) {

       custdatarev[i,9] <- mean(custdataNoNA[,9]) # the mean of the data w/o NA values
    }
}

custdatarev
```

custid <int>	sex <fctr>	is.employed <lgl>	inco… <int>	marital.stat <fctr>	health.ins <lgl>	housing.type <fctr>
2068	F	*NA*	11300	Married	TRUE	Homeowner free and clear
2073	F	*NA*	0	Married	TRUE	Rented
2848	M	TRUE	4500	Never Married	FALSE	Rented
5641	M	TRUE	20000	Never Married	FALSE	Occupied with no rent

custid <int>	sex <fctr>	is.employed <lgl>	inco… <int>	marital.stat <fctr>	health.ins <lgl>	housing.type <fctr>
6369	F	TRUE	12000	Never Married	TRUE	Rented
8322	F	TRUE	180000	Never Married	TRUE	Homeowner with mortgage/loan
8521	M	TRUE	120000	Never Married	TRUE	Homeowner free and clear
12195	M	TRUE	40000	Married	TRUE	Rented
14989	M	*NA*	9400	Married	TRUE	Rented
15917	F	TRUE	24000	Divorced/Separated	TRUE	Homeowner free and clear

1-10 of 1,000 rows | 1-7 of 11 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
NA
```

Hide

```
#Problem 2:
str(custdatarev$housing.type) #a factor type
```

```
 Factor w/ 4 levels "Homeowner free and clear",..: 1 4 4 3 4 2 1 4 4 1 ...
```

Hide

```
#converting column to string data type so I can assign row values to "unknown":
custdatarev$housing.type <- as.character(custdatarev$housing.type)

for(i in 1:length(custdatarev[,7])) {

    if(is.na(custdatarev[i,7])) {

       custdatarev[i,7] <- "unknown"
    }
}

#converting column back to factor data type:
custdatarev$housing.type <- as.factor(custdatarev$housing.type)

custdatarev
```

custid <int>	sex <fctr>	is.employed <lgl>	inco… <int>	marital.stat <fctr>	health.ins <lgl>	housing.type <fctr>
2068	F	*NA*	11300	Married	TRUE	Homeowner free and clear
2073	F	*NA*	0	Married	TRUE	Rented

custid <int>	sex <fctr>	is.employed <lgl>	inco… <int>	marital.stat <fctr>	health.ins <lgl>	housing.type <fctr>
2848	M	TRUE	4500	Never Married	FALSE	Rented
5641	M	TRUE	20000	Never Married	FALSE	Occupied with no rent
6369	F	TRUE	12000	Never Married	TRUE	Rented
8322	F	TRUE	180000	Never Married	TRUE	Homeowner with mortgage/loan
8521	M	TRUE	120000	Never Married	TRUE	Homeowner free and clear
12195	M	TRUE	40000	Married	TRUE	Rented
14989	M	*NA*	9400	Married	TRUE	Rented
15917	F	TRUE	24000	Divorced/Separated	TRUE	Homeowner free and clear

1-10 of 1,000 rows | 1-7 of 11 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
NA
```

Hide

```
#Problem 3:

for(i in 1:length(custdatarev[,8])) {

    custdatarev$recent.move[which(custdatarev$recent.move == TRUE)] <- "recent"
    custdatarev$recent.move[which(custdatarev$recent.move == FALSE)] <- "not recent"

    if(is.na(custdatarev[i,8])) {

       custdatarev[i,8] <- "unknown"
    }
}

#converting column to factor data type:
custdatarev$recent.move <- as.factor(custdatarev$recent.move)

custdatarev
```

custid <int>	sex <fctr>	is.employed <lgl>	inco… <int>	marital.stat <fctr>	health.ins <lgl>	housing.type <fctr>
2068	F	*NA*	11300	Married	TRUE	Homeowner free and clear
2073	F	*NA*	0	Married	TRUE	Rented
2848	M	TRUE	4500	Never Married	FALSE	Rented
5641	M	TRUE	20000	Never Married	FALSE	Occupied with no rent

custid <int>	sex <fctr>	is.employed <lgl>	inco… <int>	marital.stat <fctr>	health.ins <lgl>	housing.type <fctr>
6369	F	TRUE	12000	Never Married	TRUE	Rented
8322	F	TRUE	180000	Never Married	TRUE	Homeowner with mortgage/loan
8521	M	TRUE	120000	Never Married	TRUE	Homeowner free and clear
12195	M	TRUE	40000	Married	TRUE	Rented
14989	M	*NA*	9400	Married	TRUE	Rented
15917	F	TRUE	24000	Divorced/Separated	TRUE	Homeowner free and clear

1-10 of 1,000 rows | 1-7 of 11 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
NA
```

Hide

```
#Problem 4:

for(i in 1:length(custdatarev[,3])) {

    custdatarev$is.employed[which(custdatarev$is.employed == TRUE)] <- "employed"
    custdatarev$is.employed[which(custdatarev$is.employed == FALSE)] <- "unemployed and seeking"

    if(is.na(custdatarev[i,3])) {

       custdatarev[i,3] <- "unemployed and not seeking"
    }
}

#converting column to factor data type:
custdatarev$is.employed <- as.factor(custdatarev$is.employed)

custdatarev
```

custid <int>	sex <fctr>	is.employed <fctr>	inco… <int>	marital.stat <fctr>	health.ins <lgl>
2068	F	unemployed and not seeking	11300	Married	TRUE
2073	F	unemployed and not seeking	0	Married	TRUE
2848	M	employed	4500	Never Married	FALSE
5641	M	employed	20000	Never Married	FALSE
6369	F	employed	12000	Never Married	TRUE
8322	F	employed	180000	Never Married	TRUE

custid <int>	sex <fctr>	is.employed <fctr>	inco… <int>	marital.stat <fctr>	health.ins <lgl>
8521	M	employed	120000	Never Married	TRUE
12195	M	employed	40000	Married	TRUE
14989	M	unemployed and not seeking	9400	Married	TRUE
15917	F	employed	24000	Divorced/Separated	TRUE

1-10 of 1,000 rows | 1-6 of 11 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
NA
NA
NA
```

Hide

```
#Problem 5:

income.class <- c(custdatarev$income) #populating a new column with data from custdatarev datafr
ame.


#making new dataframe with same name...adding new income.class column
custdatarev <- data.frame(custdatarev$custid,custdatarev$sex, custdatarev$is.employed, custdatar
ev$income, custdatarev$marital.stat, custdatarev$health.ins, custdatarev$housing.type, custdatar
ev$recent.move, custdatarev$num.vehicles, custdatarev$age, custdatarev$state.of.res, income.clas
s)

#renaming the columns:
colnames(custdatarev) <- c("custid","sex","is.employed","income","marital.stat","health.ins","ho
using.type","recent.move","num.vehicles", "age", "state.of.res", "income.class")



custdatarev
```

custid <int>	sex <fctr>	is.employed <fctr>	inco… <int>	marital.stat <fctr>	health.ins <lgl>
2068	F	unemployed and not seeking	11300	Married	TRUE
2073	F	unemployed and not seeking	0	Married	TRUE
2848	M	employed	4500	Never Married	FALSE
5641	M	employed	20000	Never Married	FALSE
6369	F	employed	12000	Never Married	TRUE
8322	F	employed	180000	Never Married	TRUE

custid <int>	sex <fctr>	is.employed <fctr>	inco… <int>	marital.stat <fctr>	health.ins <lgl>
8521	M	employed	120000	Never Married	TRUE
12195	M	employed	40000	Married	TRUE
14989	M	unemployed and not seeking	9400	Married	TRUE
15917	F	employed	24000	Divorced/Separated	TRUE

1-10 of 1,000 rows | 1-6 of 12 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
custdatarev$income.class <- as.character(custdatarev$income.class)


for(i in 1:length(custdatarev[,12])) {

    if(custdatarev$income[i] < 30000) {

        custdatarev$income.class[i] <- "Poor Class"
    }

    else if(custdatarev$income[i] >= 30000 & custdatarev$income[i] < 45000) {

            custdatarev$income.class[i] <- "Lower Middle Class"
    }

    else if(custdatarev$income[i] >= 45000 & custdatarev$income[i] < 140000) {

                custdatarev$income.class[i] <- "Middle Class"
    }

    else if(custdatarev$income[i] >= 140000 & custdatarev$income[i] < 200000) {

                custdatarev$income.class[i] <- "Upper Class"
    }

    else if(custdatarev$income[i] >= 200000){

                custdatarev$income.class[i] <- "Rich Class"
    }
}


#converting column to factor data type:
custdatarev$income.class <- as.factor(custdatarev$income.class)

custdatarev
```

custid <int>	sex <fctr>	is.employed <fctr>	inco… <int>	marital.stat <fctr>	health.ins <lgl>
2068	F	unemployed and not seeking	11300	Married	TRUE
2073	F	unemployed and not seeking	0	Married	TRUE
2848	M	employed	4500	Never Married	FALSE
5641	M	employed	20000	Never Married	FALSE
6369	F	employed	12000	Never Married	TRUE
8322	F	employed	180000	Never Married	TRUE
8521	M	employed	120000	Never Married	TRUE
12195	M	employed	40000	Married	TRUE
14989	M	unemployed and not seeking	9400	Married	TRUE
15917	F	employed	24000	Divorced/Separated	TRUE

1-10 of 1,000 rows | 1-6 of 12 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
NA
NA
NA
NA
```

Hide

```
#Problem 6:

norm.age <- c(custdatarev$age)

custdatarev <- data.frame(custdatarev$custid,custdatarev$sex, custdatarev$is.employed, custdatarev$income, custdatarev$marital.stat, custdatarev$health.ins, custdatarev$housing.type, custdatarev$recent.move, custdatarev$num.vehicles, custdatarev$age, custdatarev$state.of.res, custdatarev$income.class, norm.age)


#renaming the columns:
colnames(custdatarev) <- c("custid","sex","is.employed","income","marital.stat","health.ins","housing.type","recent.move","num.vehicles", "age", "state.of.res", "income.class","norm.age")

custdatarev
```

custid <int>	sex <fctr>	is.employed <fctr>	inco… <int>	marital.stat <fctr>	health.ins <lgl>
2068	F	unemployed and not seeking	11300	Married	TRUE
2073	F	unemployed and not seeking	0	Married	TRUE

custid <int>	sex <fctr>	is.employed <fctr>	inco… <int>	marital.stat <fctr>	health.ins <lgl>
2848	M	employed	4500	Never Married	FALSE
5641	M	employed	20000	Never Married	FALSE
6369	F	employed	12000	Never Married	TRUE
8322	F	employed	180000	Never Married	TRUE
8521	M	employed	120000	Never Married	TRUE
12195	M	employed	40000	Married	TRUE
14989	M	unemployed and not seeking	9400	Married	TRUE
15917	F	employed	24000	Divorced/Separated	TRUE

1-10 of 1,000 rows | 1-6 of 13 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
mn <- mean(custdatarev[,10])

sd <- sd(custdatarev[,10])

for(i in 1:length(custdatarev$norm.age)) {

  n <- ((custdatarev[i,13] - mn) / sd)
  custdatarev$norm.age[i] <- n
}

custdatarev
```

custid <int>	sex <fctr>	is.employed <fctr>	inco… <int>	marital.stat <fctr>	health.ins <lgl>
2068	F	unemployed and not seeking	11300	Married	TRUE
2073	F	unemployed and not seeking	0	Married	TRUE
2848	M	employed	4500	Never Married	FALSE
5641	M	employed	20000	Never Married	FALSE
6369	F	employed	12000	Never Married	TRUE
8322	F	employed	180000	Never Married	TRUE
8521	M	employed	120000	Never Married	TRUE
12195	M	employed	40000	Married	TRUE

custid <int>	sex <fctr>	is.employed <fctr>	inco… <int>	marital.stat <fctr>	health.ins <lgl>
14989	M	unemployed and not seeking	9400	Married	TRUE
15917	F	employed	24000	Divorced/Separated	TRUE

1-10 of 1,000 rows | 1-6 of 13 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
#Problem 7:

summary(custdatarev)
```

```
    custid        sex                         is.employed      income      
 Min.   :   2068   F:440   employed                  :599   Min.   : -8700  
 1st Qu.: 345667   M:560   unemployed and not seeking:328   1st Qu.: 14600  
 Median : 693403           unemployed and seeking    : 73   Median : 35000  
 Mean   : 698500                                            Mean   : 53505  
 3rd Qu.:1044606                                            3rd Qu.: 67000  
 Max.   :1414286                                            Max.   :615000  

            marital.stat health.ins                            housing.type
 Divorced/Separated:155   Mode :logical   Homeowner free and clear    :157  
 Married           :516   FALSE:159       Homeowner with mortgage/loan:412  
 Never Married     :233   TRUE :841       Occupied with no rent       : 11  
 Widowed           : 96                   Rented                      :364  
                                          unknown                     : 56  


    recent.move   num.vehicles        age              state.of.res
 not recent:820   Min.   :0.000   Min.   :  0.0   California  :100  
 recent    :124   1st Qu.:1.000   1st Qu.: 38.0   New York    : 71  
 unknown   : 56   Median :2.000   Median : 50.0   Pennsylvania: 70  
                  Mean   :1.916   Mean   : 51.7   Texas       : 56  
                  3rd Qu.:2.000   3rd Qu.: 64.0   Michigan    : 52  
                  Max.   :6.000   Max.   :146.7   Ohio        : 51  
                                                  (Other)     :600  
            income.class    norm.age       
 Lower Middle Class:161   Min.   :-2.74074  
 Middle Class      :325   1st Qu.:-0.72626  
 Poor Class        :438   Median :-0.09011  
 Rich Class        : 37   Mean   : 0.00000  
 Upper Class       : 39   3rd Qu.: 0.65207  
                          Max.   : 5.03516  
```

Hide

```
write.csv(custdatarev,"custdatarev.txt")
```