Cluster Testing

Jerrie Kumalah

February 8, 2018

library(klaR)

The *klaR* package has the kmodes() fundtion which allows us to cluster categorical data. Before using the function it is important to remove the ID column within the dataset

all\_tgp\_test <- all\_tgp\_data[, -1]

## Addressing NAs within the data set

Having NAs will throw erros in our analysis so we need to address the NAs through imputations.Since I am not yet familiar witht he process I am going to experiement with a number of packages and processes

#inspect the data  
summary(all\_tgp\_test)

We can see that the *Hispanic, Homelessness, and Benefits* variables all have missing data. These are also the data points that are TRUE/FALSE. We can either code the information as FALSE or we can use packages that will help us fill it out. THe MICE package fills out NAs by looking at the probability of observed data points.

### Identifying the Patterns of Missing Data

Install MICE package and then load

library(mice)

Important things to think about when looking at missing data. Upon inspection does it look like data is missing at random or is data missing not at random and might be due to poor question design or something like that.

Check for columns and rows that have more that 5% of data missing

pMiss <- function(x){sum(is.na(x))/length(x)\*100}  
  
apply(all\_tgp\_test, 2, pMiss) #Columns data

## Fname Lname DOB Race multipleR   
## 0.00000 0.00000 0.00000 0.00000 0.00000   
## Health\_Issues multipleH Hispanic Homelessness M\_Status   
## 0.00000 0.00000 10.60105 12.29333 0.00000   
## Household Benefits MaxIncome Employment Visits   
## 0.00000 11.51527 0.00000 0.00000 0.00000   
## Age   
## 0.00000

The code shows that 11% of Hispanic variable data missing, 12% of Homelessness variable data missing and 12% of Benefits variable data missing.

To see what the pattern is we will use a function from the mice package.

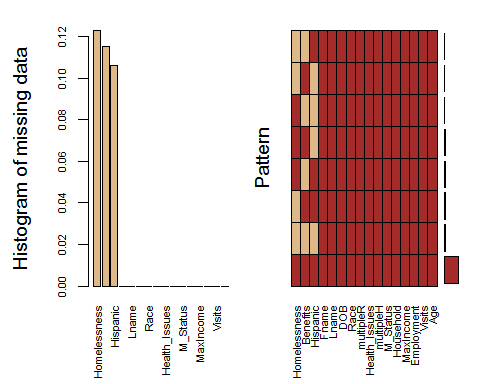
md.pattern(all\_tgp\_test)

## multipleR multipleH MaxIncome Visits Age Hispanic Benefits  
## 4101 1 1 1 1 1 1 1  
## 153 1 1 1 1 1 0 1  
## 250 1 1 1 1 1 1 1  
## 207 1 1 1 1 1 1 0  
## 45 1 1 1 1 1 0 1  
## 48 1 1 1 1 1 0 0  
## 38 1 1 1 1 1 1 0  
## 299 1 1 1 1 1 0 0  
## 0 0 0 0 0 545 592  
## Homelessness Fname Lname DOB Race Health\_Issues M\_Status Household  
## 4101 1 0 0 0 0 0 0 0  
## 153 1 0 0 0 0 0 0 0  
## 250 0 0 0 0 0 0 0 0  
## 207 1 0 0 0 0 0 0 0  
## 45 0 0 0 0 0 0 0 0  
## 48 1 0 0 0 0 0 0 0  
## 38 0 0 0 0 0 0 0 0  
## 299 0 0 0 0 0 0 0 0  
## 632 5141 5141 5141 5141 5141 5141 5141  
## Employment   
## 4101 0 8  
## 153 0 9  
## 250 0 9  
## 207 0 9  
## 45 0 10  
## 48 0 10  
## 38 0 10  
## 299 0 11  
## 5141 42897

This shows you how complete the data is. Wherever you see a 0 tells you that the data is missing for that.

To visualize the pattern we can use the *VIM* package

library(VIM)  
aggr\_plot <- aggr(all\_tgp\_test, col=c('brown','burlywood'),   
 numbers=TRUE,   
 sortVars=TRUE,   
 labels=names(all\_tgp\_test),  
 cex.axis=.7, gap=3,   
 ylab=c("Histogram of missing data","Pattern"))



##   
## Variables sorted by number of missings:   
## Variable Count  
## Homelessness 0.1229333  
## Benefits 0.1151527  
## Hispanic 0.1060105  
## Fname 0.0000000  
## Lname 0.0000000  
## DOB 0.0000000  
## Race 0.0000000  
## multipleR 0.0000000  
## Health\_Issues 0.0000000  
## multipleH 0.0000000  
## M\_Status 0.0000000  
## Household 0.0000000  
## MaxIncome 0.0000000  
## Employment 0.0000000  
## Visits 0.0000000  
## Age 0.0000000

The plot shows us that approximately 80% of our data is not missing

### Imputing the Missing Data

We are going to use the mice() to complete our imputations

?mice #to get information about the function  
tempData <- mice(all\_tgp\_test, m = 5, method = 'pmm')  
  
summary(tempData) #to get an overview of the data   
tempData$imp$Hispanic # Example of how to inspect imputation

Completing your data

compl\_tgp\_data <- complete(tempData, 1) # complete data without any NAs