CHARITY PROJECT ASSIGNMENT - PART 1

# 1. Read data from csv file

1. Here I use the na.strings="NA" argument to encode the missing values.
2. I set stringsAsFactors=TRUE
3. Finally, I use the colClasses argument to update DONR, HOME, and HINC from integers to factors.

setwd('/Users/stevenfutter/Dropbox/NU/MACHINE\_LEARNING/charity\_project')  
  
# a. Use na.strings='NA' to encode the missing values  
# b. stringsAsFactors=TRUE is set  
# c. colClasses argument used to set other integer values to factors for DONR,  
# HOME, and HINC.  
charity = read.csv('projectDataPart1.csv', header=T, na.strings="NA")  
  
# ,   
# stringsAsFactors = TRUE,  
# colClasses=c("DONR"="factor", "HOME"="factor","HINC"="factor"))  
  
# check that column types are set up correctly  
str(charity)

## 'data.frame': 3684 obs. of 21 variables:  
## $ ID : int 39 315 333 377 461 557 711 776 792 893 ...  
## $ DONR : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ DAMT : num 20 5 10 5 10 25 15 30 10 36 ...  
## $ AGE : int 43 85 24 76 83 43 87 77 40 76 ...  
## $ HOME : int 1 0 0 1 1 0 1 1 1 1 ...  
## $ HINC : int 6 1 6 7 1 NA 5 4 4 5 ...  
## $ GENDER : Factor w/ 6 levels " ","A","F","J",..: 3 3 5 3 3 3 3 5 3 3 ...  
## $ MEDAGE : int 46 42 38 42 40 29 42 42 50 48 ...  
## $ MEDPPH : int 228 151 223 292 189 125 174 140 172 193 ...  
## $ MEDHVAL : int 3909 689 731 3359 1207 684 2907 1139 504 1178 ...  
## $ MEDINC : int 870 235 350 878 284 272 420 303 238 482 ...  
## $ MEDEDUC : int 153 120 120 160 128 140 140 141 120 145 ...  
## $ NUMPROM : int 14 70 13 24 22 20 23 27 50 45 ...  
## $ NUMPRM12: int 12 10 8 15 12 10 12 10 13 14 ...  
## $ RAMNTALL: num 20 102 15 29 29 27 27 35 102 109 ...  
## $ NGIFTALL: int 1 21 1 5 4 2 3 2 12 7 ...  
## $ MAXRAMNT: num 20 8 15 10 11 20 15 20 11 23 ...  
## $ LASTGIFT: num 20 5 15 6 11 20 9 20 11 23 ...  
## $ TDON : num 16 18 20 16 18 ...  
## $ RFA\_97 : Factor w/ 14 levels "L1E","L1F","L1G",..: 2 7 2 12 8 2 5 2 4 9 ...  
## $ RFA\_96 : Factor w/ 58 levels " ","A1D","A1E",..: 22 49 22 42 12 4 9 4 46 13 ...

# confirm size of data table  
dim(charity) # 3684 21

## [1] 3684 21

# take a peek at the data  
head(charity)

## ID DONR DAMT AGE HOME HINC GENDER MEDAGE MEDPPH MEDHVAL MEDINC MEDEDUC  
## 1 39 1 20 43 1 6 F 46 228 3909 870 153  
## 2 315 1 5 85 0 1 F 42 151 689 235 120  
## 3 333 1 10 24 0 6 M 38 223 731 350 120  
## 4 377 1 5 76 1 7 F 42 292 3359 878 160  
## 5 461 1 10 83 1 1 F 40 189 1207 284 128  
## 6 557 1 25 43 0 NA F 29 125 684 272 140  
## NUMPROM NUMPRM12 RAMNTALL NGIFTALL MAXRAMNT LASTGIFT TDON RFA\_97 RFA\_96  
## 1 14 12 20 1 20 20 16 L1F F1F  
## 2 70 10 102 21 8 5 18 L3D S3D  
## 3 13 8 15 1 15 15 20 L1F F1F  
## 4 24 15 29 5 10 6 16 L4E N4E  
## 5 22 12 29 4 11 11 18 L3E A3E  
## 6 20 10 27 2 20 20 18 L1F A1F

# 2. Data Quality Check