Shapeslosion data cleaning

Yuan Wang

5/11/2017

## Introduction

#### Goal

* We want to clean the raw Shapessplosion data. The goal is to filter out all rows that contain gender information, and then add a gender column with 0 indicating female and 1 be male.

#### Method

* We first filter out a gender data using v1label1.
* Then, we mutate a column to the gender data where if the user was male we denote as 1, and if female we denote as 0.
* We use a for loop to select all groupID's whose sample size is less then 50 and greater than 5. Also we use only groups that have more than 1 male and 1 female observations. (Otherwise t.test would fail).
* We save each group to an individual .csv file.
* We then bind all .csv file together and obtain a clean gender data.

## Import and trim data

We use only untimed data with non-zero time used recording.

Original <- read.csv("original\_raw\_data.csv")  
shape <- Original  
shape\_untime <- shape[shape$requestedTime==0,]  
shape\_untimed <- shape\_untime[shape\_untime$timeUsed!=0,]

We changed some numerical variables to factors since they have only a fixed number of levels. We also changed the unit of the time data from milliseconds to seconds.

shape\_untimed$numShapes <- as.factor(shape\_untimed$numShapes)  
shape\_untimed$matchingScheme <- as.factor(shape\_untimed$matchingScheme)  
shape\_untimed$requestedTime <- as.factor(shape\_untimed$requestedTime)  
shape\_untimed$timeUsed <- as.numeric(shape\_untimed$timeUsed)  
shape\_untimed$timerDisplay <- as.factor(shape\_untimed$timerDisplay)  
shape\_untimed$numErrors <- as.numeric(shape\_untimed$numErrors)  
  
shape\_untimed <- mutate(shape\_untimed, TimeUsedSec = shape\_untimed$timeUsed/1000)

We further trimed the data to include a varible column named Gender.

gender <- filter(shape\_untimed, tolower(strtrim(shape\_untimed$v1label,3))=="gen" |  
 tolower(strtrim(shape\_untimed$v1label,3))=="sex" |  
 tolower(strtrim(shape\_untimed$v1label,4))=="male" |  
 tolower(strtrim(shape\_untimed$v1label,1))=="f")  
  
case <- (tolower(strtrim(gender$v1value,1)) == "m") & (tolower(strtrim(gender$v2value,1)) =="f")  
gender1 <- gender[(strtrim(gender$v1value,1) != "1") &  
 (gender$v1value != "2") &  
 (gender$v1label != "Gender\rOrder") &  
 (gender$v1label != "female\rorder") &  
 (gender$v1value != "0") &  
 (tolower(gender$v1value) != "morf") &  
 (gender$studentID != "mb") &  
 (gender$studentID != "31207") &  
 (gender$v1value != "cat") &  
 (gender$studentID != "3659") &  
 (tolower(gender$v1value) != "attempt") &  
 (!case),]  
  
gender1 <- mutate(gender1,   
 gender = as.factor(ifelse(  
 pmax((tolower(strtrim(gender1$v1label,1)) == "m"),  
 (tolower(strtrim(gender1$v1value,1)) == "m")),   
 1,   
 ifelse(pmax(tolower(strtrim(gender1$v1label,1)) == "f",  
 (tolower(strtrim(gender1$v1value,1)) =="f")),   
 0,   
 NA))))  
  
gender2 <- na.omit(gender1)

## Extract GroupID

Then, within the gender data, we identified the group names with sample size between 5 and 50. We make a vector of groupID of our desired groups.

tb <- as.data.frame(table(gender2$groupID))  
  
##Select groups that are under size 50  
tb2 <- tb[tb$Freq >= 5 & tb$Freq <= 50,]  
  
##Create a vector of groupID's whose size is between 25 and 50  
selected\_groupID <- as.character(tb2$Var1)

## Create .csv files

Using a for loop, we traverse through the vector of groupID and, for each group with the corresponding groupID, we write out the data to a .csv file.

par(mar=c(1,1,1,1))  
par(mfrow = c(5,5))  
groupName <- c()  
pvalues <- c()  
for (i in 1:length(selected\_groupID)) {  
 female <- gender2[gender2$groupID == selected\_groupID[i] & gender2$gender == 0,]$TimeUsedSec  
 male <- gender2[gender2$groupID == selected\_groupID[i] & gender2$gender == 1,]$TimeUsedSec  
 if (length(female) > 1 & length(male) > 1) {  
 write.csv(gender2[gender2$groupID == selected\_groupID[i],], file =paste(selected\_groupID[i], "group.csv"))  
 }  
}

## Create clean data

We then read each .csv file back into data files, and then use rbind to bind them together to produce one clean data set.

data <- read.csv("hj190f14 group.csv")  
data <- rbind(data, read.csv("MA376C1 group.csv"))  
data <- rbind(data, read.csv("msp2013 group.csv"))  
data <- rbind(data, read.csv("MA376I3 group.csv"))  
data <- rbind(data, read.csv("MAT336S15 group.csv"))   
data <- rbind(data, read.csv("MA376B1 group.csv"))  
data <- rbind(data, read.csv("MATH22015 group.csv"))  
data <- rbind(data, read.csv("MATH22018 group.csv"))  
data <- rbind(data, read.csv("mth22602 group.csv"))  
data <- rbind(data, read.csv("stats2 group.csv"))  
data <- rbind(data, read.csv("mth32601 group.csv"))  
data <- rbind(data, read.csv("mth32602 group.csv"))  
data <- rbind(data, read.csv("MAED550 group.csv"))  
data <- rbind(data, read.csv("336S14 group.csv"))  
data <- rbind(data, read.csv("MAT336S14 group.csv"))  
data <- rbind(data, read.csv("mth22601 group.csv"))  
data <- rbind(data, read.csv("psych3700 group.csv"))  
data <- rbind(data, read.csv("USCOTS15 group.csv"))  
  
write.csv(data, file = "cleaned\_gender.csv")