R - data preparation manual

# How to prepare datasets for using them in tableau

This is a brief manual on how we prepared the data for our prototype at the example of one dataset. The process can be transferred to all other datasets from DHS. The DHS data was made available in the SPSS file format “.sav” but did not come with meaningful header names and included many variables that are not relevant for the topic of mobility. Moreover, it included some variables that are not consistently filled. To get insight jnto which variables are included and how the raw data is structured, please refer to: [Link](https://dhsprogram.com/pubs/pdf/DHSG4/Recode7_DHS_10Sep2018_DHSG4.pdf)

# Install and load required packages

#install.packages("dyplr")  
#install.packages("data.table")  
#install.packages("foreign")  
#install.packages("tidyr")  
#install.packages("readxl")  
#install.packages('writexl')  
  
library("dplyr")  
library("data.table")  
library("foreign")  
library("tidyr")  
library("readxl")  
library("writexl")

## Load Data into R (Example: Household data 2016)

setwd("C:\\Users\\janho\\OneDrive\\TechChallenge\_Connection People\\Prototype\\Tableau Map")  
data\_sav <- read.spss("ETHR71FL.SAV")  
  
HR\_dt\_1 <- as.data.table(data\_sav)

# Rename the variables

## 1. First load the excel file that contains the mapping from variable code to variable name

You can find the file “Mobility Indicators DHS\_universal.xlsx” in the file repository

indicators\_dt <-read\_excel("C:\\Users\\janho\\OneDrive\\TechChallenge\_Connection People\\Prototype\\Tableau Map\\Mobility Indicators DHS\_universal.xlsx", sheet = "Mobility")  
setDT(indicators\_dt)  
class(indicators\_dt)  
View(indicators\_dt)

## 2. Extract the rows that contain the relevant identifier. In this case: “Household\_hh” from “indicators\_dt”.

household\_dt <- indicators\_dt[Dataset == "Household\_hh"]  
Individual\_fem\_dt <- indicators\_dt[Dataset == "Individual\_fem"]  
Men\_m\_dt <- indicators\_dt[Dataset == "Men\_m"]   
  
View(household\_dt)

## 3. Extract the variable codes and short names from “household\_dt” and remove NAs

codes\_names\_dt\_NA <- household\_dt[, c("Variable code", "Short Name")]  
View(codes\_names\_dt\_NA)  
codes\_names\_dt <- codes\_names\_dt\_NA[!is.na(codes\_names\_dt\_NA$"Variable code"), ]  
View(codes\_names\_dt)

## 4. Get vectors of variable code and short name

codes <- codes\_names\_dt$`Variable code`  
names <- codes\_names\_dt$`Short Name`

## 5. Change column names

renamed\_HR\_dt <- setnames(HR\_dt\_1, codes, names, skip\_absent=TRUE)  
#manual change:  
setnames(renamed\_HR\_dt, c("HV001", "HV002"), c("DHSCLUST", "Household Number"))

## 6. transform into data table

renamed\_dt2 <- as.data.table(renamed\_HR\_dt)

# Cleaning the data: get rid of variables unrelated to mobility (as specified in the mapping file)

HR\_end\_dt <- renamed\_HR\_dt %>% select(-contains("HV"))  
HR\_end\_dt\_nonum <- HR\_end\_dt %>% select (-ends\_with("1"),   
 -ends\_with("2"),  
 -ends\_with("3"),  
 -ends\_with("4"),  
 -ends\_with("5"),  
 -ends\_with("6"),  
 -ends\_with("7"),  
 -ends\_with("8"),  
 -ends\_with("9"),  
 -ends\_with("0"))   
clean\_HR\_2016 <-HR\_end\_dt\_nonum %>% select(-starts\_with("SH"), -starts\_with("IDX"))  
View(clean\_HR\_2016)  
  
# Check for NAs and if necessary, delete the variables with many NAs in the previous step  
na\_count <-sapply(clean\_HR\_2016, function(y) sum(length(which(is.na(y)))))  
na\_count <- data.frame(na\_count)  
na\_count

# Export the tansformed datafile into .csv and .xlsx

write.csv(clean\_HR\_2016,"C:\\Users\\janho\\OneDrive\\TechChallenge\_Connection People\\Prototype\\Tableau Map\\clean datasets\\clean\_HR\_2016.csv")  
  
write\_xlsx(clean\_HR\_2016, "C:\\Users\\janho\\OneDrive\\TechChallenge\_Connection People\\Prototype\\Tableau Map\\clean datasets\\clean\_HR\_2016.xlsx")