Preparación y Manipulación de Datos

## 1. Objetivo:

Limpieza y preparación del dataset para un posterios analisis.

El objetivo es eliminar toda la información que no sea util.

## 2. Lectura del dataset en R.

Creo un directorio y descargo en el los datos que voy a utilizar

#limpio el espacio de trabajo  
rm(list=ls())  
ls()

## character(0)

#Instalo y cargo las librerias que voy a utilizar  
libs <- c("gsheet","knitr","dplyr","zoo","stringr")  
   
for (i in libs){  
 print(i)  
 if(!require(i, character.only = TRUE)) { install.packages(i, character.only = TRUE); }  
 library(i, character.only = TRUE)  
}

## [1] "gsheet"  
## [1] "knitr"  
## [1] "dplyr"  
## [1] "zoo"  
## [1] "stringr"

#Cambio el espacio de trabajo si es necesario  
#path\_source <- "./RStudio/preparacion"  
#setwd(path\_source)  
  
getwd()

## [1] "C:/Users/alvaro/Documents/RStudio/preparacion"

currentDir <- getwd()  
  
#Descargo el fichero que voy a limpiar uasando gsheet2tbl  
url <- 'https://docs.google.com/spreadsheets/d/1CDWBeqpUTBd1TkmDz\_M6UGRWdHgU7LOcoiGRTvIttKA/edit#gid=0'  
dataset <- gsheet2tbl(url)  
  
#Dimensiones del datset  
dim(dataset)

## [1] 5279 5

names(dataset)

## [1] "Year"   
## [2] "Area..use.fill.down."   
## [3] "Street..use.find.and.replace.to.replace.the.odd.character.with.a.space."  
## [4] "Street.2..in.Refine.use.titlecase.and.cluster.and.edit."   
## [5] "Strange.HTML..use.unescape.HTML."

#Sustituyo los nombres de las variables por unos que sean mas faciles de manejar  
names(dataset) = c("year", "area", "street", "street2", "strange\_html")  
  
#Elimino la ultima variable ya que no veo que tenga ningun valor para el dataset  
dataset <- dataset %>% select(year, area, street, street2)  
  
#Paso todo el dataset a minusculas  
dataset <- mutate\_each(dataset, funs(tolower))  
  
#Sustituyo el caracter estraño de la variable street por un espacio  
dataset$street <- gsub("å", " ", dataset$street)  
  
#A simple vista las variable street y la variable street2 parecen iguales, voy a comprobar si es asi y puedo eliminar una de las 2  
#Creo una combinacion de los nombres de todas las variables de dos en dos  
combs <- combn(names(dataset),2)  
combs

## [,1] [,2] [,3] [,4] [,5] [,6]   
## [1,] "year" "year" "year" "area" "area" "street"   
## [2,] "area" "street" "street2" "street" "street2" "street2"

#Compruebo cual de estas combinaciones de columnas son iguales  
dups <- mapply(identical,  
 dataset[combs[1,]],  
 dataset[combs[2,]])  
dups

## year year year area area street   
## FALSE FALSE FALSE FALSE FALSE TRUE

#El resultado me dice que street y street2 son iguales, asique selecciono una de las dos para eliminarla  
drop <- combs[2,][dups]  
drop

## [1] "street2"

#Me quedo solo con las variables que no estan repetidas  
dataset <- dataset[ !names(dataset) %in% drop ]  
str(dataset)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 5279 obs. of 3 variables:  
## $ year : chr "2011" "2011" "2011" "2011" ...  
## $ area : chr "birmingham" "" "" "" ...  
## $ street: chr "acocks green" "adderly street" " alcester raod, mosley" "alcester road" ...

#En la variable area convierto los valores vacios en NA para poder rellenarlos con la funcion na.locf que sustituye el NA de una variable por el valor de la mas reciente anterior con valor  
is.na(dataset$area) <- dataset$area == ""  
  
dataset$area <- na.locf(dataset$area)  
  
#Voy a eliminar en la variable street caracteres tales como espacios innecesarios, signos de puntuación ...  
  
#Elimino signos de puntuación  
dataset$street <- gsub('[[:punct:]]','',dataset$street)  
  
#Espacios en blanco al principio y al final.  
dataset$street <- str\_trim(dataset$street)  
  
#Mas de un espacio seguido  
dataset$street <- gsub('[[:blank:]]{2,}','',dataset$street)  
  
#Sustituyo las abreviaturas de Road (RD) y Street (ST)  
dataset$street <- gsub('\\<rd\\>','road',dataset$street)  
dataset$street <- gsub('\\<st\\>','street',dataset$street)  
  
#elimino filas repetidas  
clean.dataset <- unique(dataset)  
  
#Dimensiones del datset resultado  
dim(clean.dataset)

## [1] 3324 3

## 4. Exportar desde R a un fichero local, el dataset limpio resultante.

outputDir <- "./output"  
if (!file.exists(outputDir)) {  
 dir.create(outputDir)  
 }  
write.table(clean.dataset,file=paste(outputDir,"clean.dataset.csv", sep="/"), sep=";",row.names=FALSE)

## 5. Conclusiones.

Partiamos de un dataset con 5 variables y 5279 lineas, he conseguido reducirlo a un dataset de 3 variables y 3324 lineas. Para ello he eliminado una variable que no tenia informacion relacionada con las otras, una variable que estaba duplicada y despues de limpiar la variable street de espacios, signos de puntuación y abreviaturas he podido eliminar muchas filas que aparecian repetidas.