practica_preparacion_datos

Jesus Manuel Vicente Garcia

7 de diciembre de 2016

Leemos el dataset. El archivo se ha descargado como CSV

currentDir <- getwd() path <- currentDir dir <- "datos" file <- "messyData.csv" fileURL <- "https://docs.google.com/spreadsheets/d/1CDWBeqpUTBd1TkmDz_M6UGRWdHgU7LOcoiGRTvIttKA/edit#gid=0" if (!file.exists(paste0(path,"/", dir, "/", "pp"))) { library(downloader) download(fileURL, paste0(path,"/", dir, "/", file), mode= "wb") }

dataToClean <- read.csv2("./input/messyData.csv", row.names=NULL, sep=",", header=TRUE)

Cambiamos el nombre de las columnas que lo requieren

names(dataToClean) <- c("Year","Area","Street","Street2","StrangeHTML")

Comprobamos el la clase de las columnas y modificamos las que lo requieran

lapply(dataToClean, class)

as.character(dataToClean$Area)

as.character(dataToClean$Street)

as.character(dataToClean$Street2)

as.character(dataToClean$StrangeHTML)

Comprobamos y eliminamos los elementos NA en caso de que los haya

rowNadata <- dataToClean[rowSums(is.na(dataToClean))>0,] dim(rowNadata)

colNadata <- dataToClean[colSums(is.na(dataToClean))>0,] dim(colNadata)

Convertimos a minúscula los nombres de las columnas

names(dataToClean) <- tolower(names(dataToClean))

Las columnas 'street' y 'strangehtml' no aportan información relevante, ya que la información necesaria para nuestro análisis estará en las columnas 'year' 'area' y 'street2'

library(data.table) streetLower <- tolower(dataToClean$street2)$tidyData < -data.table(year = dataToCleanyear,area=dataToClean$area,street=streetLower)

Podemos ver en la columna 'street' que algunos valores incluyen además el barrio al que pertenecen. Crearemos una nueva columna para el barrio, separando a partir de la ','

library(tidyr) tidyData2 <- data.table(separate(tidyData, street, c("street","neighbourhood"),sep=','))

Al haber creado la columna para el barrio, vemos que la columna 'area' en realidad contiene valores de ciudades. Con lo cual, cambiaremos el nombre de esa columna a 'city'

setnames(tidyData2, "area", "city")

Eliminamos los valores NA de la columna 'neighbourhood'

nomissingTidyData <- complete.cases(tidyData2) tidyDataNoNA <- tidyData2[nomissingTidyData,]

Vemos que la columna 'city' tiene la mayoría de las celdas vacías, y no es posible deducir como rellenarlas a partir de los datos que tenemos, por lo que prescindiremos de esa columna.

tidyDataNoNANoCity <- data.table(year=tidyDataNoNA$year, street = tidyDataNoNA$street,neighbourhood=tidyDataNoNA

Ahora eliminamos las filas duplicadas

tidyDataNoNANoCityNoDup <- data.table(year=tidyDataNoNANoCity$year, street = unique(tidyDataNoNANoCity$street),

Quitamos los '.' y caracteres especiales en los valores de la columna 'neighbourhood'

noPointNeighb <- gsub('\.',",tidyDataNoNANoCityNoDup$neighbourhood)tidyDataNoNANoCityNoDup <
$-data.table(year = tidyDataNoNANoCityNoDup$year,street=tidyDataNoNANoCityNoDup$street, neighbourhood =$
$noPointNeighb)noCarEspNeighb < -gsub("[[: punct :]]",'', tidyDataNoNANoCityNoDup$neighbourhood)$
tidyDataNoNANoCityNoDup <- data.table(year=tidyDataNoNANoCityNoDup$year, street = tidyDataNoNANoCityNoDup

Una vez eliminamos los valores repetidos y los caracteres especiales, ya tenemos un dataset limpio para poder realizar análisis. Los datos de este dataset no nos permiten realizar análisis muy exhaustivos ni obtener información de especial relevancia, ya que unicamente contamos con años calles y barrios. Nos permitiría realizar análisis sencillos, como obtener los barrios que cuentan con mayor número de calles.

Exportamos nuestro tidy dataset resultante a un archivo CSV

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