Tarea2

Daniel Camarena

24/8/2017

Tarea Esta tarea consta de dos partes, la primera parte se concentra en la limpieza de datos y se entregará el lunes 28.

PRIMERA PARTE En la carpeta de la tarea encontrarás un archivo de excel (m_013.xls), este archivo contiene información de causas de mortalidad en México entre 2000 y 2008. Contesta las siguientes preguntas: 1. ¿Cuáles son las variables en esta base de datos? 2. ¿La base de datos cumple con los principios de datos limpios? ¿Qué problemas presenta? 3. La información del archivo de excel se ha guardado también en archivos de texto (csv) 2001-2008, limpia los datos para que cumplan los principios de datos limpios. Recuerda que las modificaciones deben de ser reproducibles, para esto guarda tu trabajo en un script.

Observaciones: * Puedes filtrar/eliminar los valores a Total si crees que es más claro. * Intenta usar las funciones que estudiamos en la clase (gather, separate, select, filter). * Si aún no te sientes cómodx con las funciones de clase (y lo intentaste varias veces) puedes hacer las manipulaciones usando otra herramienta (incluso Excel, una combinación de Excel y R o cualquier software que conozcas); sin embargo, debes documentar tus pasos claramente, con la intención de mantener métodos reproducibles.

Respuestas

1. ¿Cuáles son las variables en esta base de datos?

Las variables son:

- Entidad Federativa
- Año
- Sexo
- Tipo de Enfermedad
- Tasa de Mortalidad

2. ¿La base de datos cumple con los principios de datos limpios? ¿Qué problemas presenta?

- La base no esta en un formato *tidy*, para que sea *tidy* debe cumplir con lo siguiente:
- Cada variable es una columna
- Cada observacion es una fila
- Cada unidad observacional es una tabla

3. Limpieza de datos

Leamos los datos

```
library(tidyverse)
paths <- dir( pattern = "\\.csv$", full.names = TRUE)
paths <- set_names(paths, basename(paths))

mortalidad <- map_df(paths, ~read_csv(., col_types = "ddddddddddd"), .id = "filename")

mortalidad <- mutate(mortalidad, year=parse_number(filename)) %>% select(-filename)
```

Quitemos los totales y limpiemos los datos i.e. dejemoslos en un formato *tidy*. Dejemos todo en un CSV

```
mortalidad_stg <- select(mortalidad,  -contains("_Total"))
mortalidad_hombres <- select(mortalidad_stg,edo,year,contains("_Hombres"))
```

```
mortalidad_mujeres <- select(mortalidad_stg,edo,year,contains("_Mujeres"))
mortalidad_hombres <- mutate(mortalidad_hombres, sexo = "M")
mortalidad_mujeres <- mutate(mortalidad_mujeres, sexo = "F")
mortalidad_hombres<-set_names(mortalidad_hombres,c("edo","year","trans","noTrans","lesiones","sexo"))
mortalidad_mujeres <- set_names(mortalidad_mujeres,c("edo","year","trans","noTrans","lesiones","sexo"))
mortalidad_clean <- union_all(mortalidad_hombres,mortalidad_mujeres)
mortalidad_clean <- gather(mortalidad_clean,"tipo_enfermedad","tasa",-edo,-year,-sexo)
head(mortalidad_clean)
```

```
## # A tibble: 6 × 5
##                   edo  year  sexo tipo_enfermedad  tasa
##                 <chr> <dbl> <chr>           <chr> <dbl>
## 1      Aguascalientes  2000     M           trans  72.5
## 2     Baja California  2000     M           trans 115.1
## 3 Baja California Sur  2000     M           trans  91.3
## 4            Campeche  2000     M           trans  64.1
## 5            Coahuila  2000     M           trans  69.3
## 6              Colima  2000     M           trans  72.1
```

```
write.csv(mortalidad_clean, file = "mortalidad.csv", append=FALSE)
```