Data tables handling for NeuroColombia research project

Data analysis notebook

Daniel Manrique-Castano
Digital Research Alliance of Canada

Tuesday, March 11, 2025

Table of contents

Install and load packages

We install the required packages to open and handle the data tables.

```
if (!requireNamespace("dplyr", quietly = TRUE)) install.packages("dplyr")
if (!requireNamespace("readxl", quietly = TRUE)) install.packages("readxl")
if (!requireNamespace("ggplot2", quietly = TRUE)) install.packages("ggplot2")
if (!requireNamespace("here", quietly = TRUE)) install.packages("here")
if (!requireNamespace("tidyr", quietly = TRUE)) install.packages("tidyr")
if (!requireNamespace("tibble", quietly = TRUE)) install.packages("tibble")

library(dplyr)
library(readxl)
library(ggplot2)
library(here)
library(tidyr)
library(tibble)
```

Load data tables

We load three data tables associated with neuropsichological diagnostics (`Ministery_DiagnosticData.xlsx`) and population data for each city in Colombia (`DANE_PopulationData_2005-2019.xlsx` and `DANE_PopulationData_2005-2019.xlsx`).

```
Ministery_Data <- read_excel(here("Data_Processed/
Ministery_DiagnosticData.xlsx"))
Dane_Data_2019 <- read_excel(here("Data_Raw/
DANE_PopulationData_2005-2019.xlsx"))
```

```
Dane_Data_2035 <- read_excel(here("Data_Raw/
DANE_PopulationData_2020-2035.xlsx"))
```

Data processing

Subseting of data tables

Next, we subset the `Ministery_Data` to obtain the age (Edad) of interest for the current analysis.

```
# We select the rows containing ages from 0 to 11 years old
Ministery_Data <- Ministery_Data %>%
  filter(Edad >= 0, Edad <= 11)
```

We inspect the results

```
head(Ministery_Data)
```

```
# A tibble: 6 × 8
    Año Año_Cod Departamento Municipio Diagnostico  Edad Sexo    Cantidad
  <dbl>   <dbl> <chr>        <chr>     <chr>       <dbl> <chr>      <dbl>
1  2016       0 Antioquia    Medellín  F700            2 Hombres        1
2  2016       0 Antioquia    Medellín  F700            4 Mujeres        2
3  2016       0 Antioquia    Medellín  F700            4 Hombres        1
4  2016       0 Antioquia    Medellín  F700            6 Mujeres        1
5  2016       0 Antioquia    Medellín  F700            6 Hombres        1
6  2016       0 Antioquia    Medellín  F700            7 Mujeres        3
```

Now, we merge the Dane_Data_2019 and Dane_Data_2035 in a single dataset (`Dane_Data_Total`) and then subset the rows of interest, which includes the counts of men (Hombres) and women (Mujeres) from 0 to 11 years old..

```
# We merge the datasets
Dane_Data_Total <- rbind(Dane_Data_2019, Dane_Data_2035)

# We select the rows containing total counts for ÁREA GEOGRÁFICA (Including urban
and rural areas)
Dane_Data_Total <- subset(Dane_Data_Total, `ÁREA GEOGRÁFICA` == "Total")

# We subset ID variables and columns containing counts for population between 0
and 11 years old
Dane_Data_Total <- Dane_Data_Total %>%
   select(DP, DPNOM, DPMP, MPIO, AÑO, `ÁREA GEOGRÁFICA`, matches("^(Hombres|
Mujeres)_([0-9]|1[0-1])$"))
```

Here, we inspect the resulting data for `Dane_Data_Total`.

```
head(Dane_Data_Total)
```

```
# A tibble: 6 × 30
  DP    DPNOM  DPMP  MPIO    AÑO `ÁREA GEOGRÁFICA` Hombres_0 Hombres_1 Hombres_2
  <chr> <chr>  <chr> <chr> <dbl> <chr>                 <dbl>     <dbl>     <dbl>
1 05    Antio… Mede… 05001  2005 Total                 14301     14726     15179
2 05    Antio… Mede… 05001  2006 Total                 14149     14476     14889
3 05    Antio… Mede… 05001  2007 Total                 13926     14328     14657
4 05    Antio… Mede… 05001  2008 Total                 13750     14123     14510
5 05    Antio… Mede… 05001  2009 Total                 13633     13971     14333
6 05    Antio… Mede… 05001  2010 Total                 13608     13911     14231
# ℹ 21 more variables: Hombres_3 <dbl>, Hombres_4 <dbl>, Hombres_5 <dbl>,
#   Hombres_6 <dbl>, Hombres_7 <dbl>, Hombres_8 <dbl>, Hombres_9 <dbl>,
#   Hombres_10 <dbl>, Hombres_11 <dbl>, Mujeres_0 <dbl>, Mujeres_1 <dbl>,
#   Mujeres_2 <dbl>, Mujeres_3 <dbl>, Mujeres_4 <dbl>, Mujeres_5 <dbl>,
#   Mujeres_6 <dbl>, Mujeres_7 <dbl>, Mujeres_8 <dbl>, Mujeres_9 <dbl>,
#   Mujeres_10 <dbl>, Mujeres_11 <dbl>
```

Relating information between data tables

In this step, we transform the `Dane_Data_Total` from wide to long format and setup the variable `Edad` as numeric.

```
# Reshape Dane_Data_Total to long format:
Dane_Data_Total_long <- Dane_Data_Total %>%
  pivot_longer(
    cols = c(starts_with("Hombres_"), starts_with("Mujeres_")),
    names_to = c("Sexo", "Edad"),
    names_sep = "_",
    values_to = "Poblacion"
  )

Dane_Data_Total_long$Edad <- as.numeric(Dane_Data_Total_long$Edad)
```

We create a new column called “Poblacion”, which includes the reference population for the number of diagnostics in the `Ministery_Data` dataset.

```
head(Dane_Data_Total_long)
```

```
# A tibble: 6 × 9
  DP    DPNOM     DPMP     MPIO    AÑO `ÁREA GEOGRÁFICA` Sexo     Edad Poblacion
  <chr> <chr>     <chr>    <chr> <dbl> <chr>             <chr>   <dbl>     <dbl>
1 05    Antioquia Medellín 05001  2005 Total             Hombres     0     14301
2 05    Antioquia Medellín 05001  2005 Total             Hombres     1     14726
3 05    Antioquia Medellín 05001  2005 Total             Hombres     2     15179
```

```
4 05    Antioquia Medellín 05001  2005 Total            Hombres     3     15651
5 05    Antioquia Medellín 05001  2005 Total            Hombres     4     16109
6 05    Antioquia Medellín 05001  2005 Total            Hombres     5     16545
```

Then, we pair the counts per age from `Dane_Data_Total_long` to the `Ministery_Data` dataset, considering city (`Municipio`), year (`Año`), gender (`Sexo`) and age (`Edad`)

```
# Merge the reshaped reference population with Ministery_Data:
Ministery_Data_with_pop <- Ministery_Data %>%
  left_join(Dane_Data_Total_long,
            by = c("Municipio" = "DPMP",    # City match: Municipio vs. DPMP
                   "Año" = "AÑO",           # Year match: Año vs. AÑO
                   "Sexo",                  # Gender match
                   "Edad"                   # Age match
            ))
```

Next, we generate a column to indicate if the city (`Municipio`) is a Departemental capital or not.

```
# We create a lookup table for department capitals
capitals <- tibble(
  Departamento = c("Amazonas", "Antioquia", "Arauca", "Atlántico", "Bolívar",
                  "Boyacá", "Caldas", "Caquetá", "Casanare", "Cauca", "Cesar",
                   "Chocó", "Córdoba", "Cundinamarca", "Guainía", "Guaviare",
                   "Huila", "La Guajira", "Magdalena", "Meta", "Nariño",
                   "Norte de Santander", "Putumayo", "Quindío", "Risaralda",
                   "San Andrés y Providencia", "Santander", "Sucre", "Tolima",
                   "Valle del Cauca", "Vaupés", "Vichada", "Bogotá, D.C."),
    CapitalMunicipio  =  c("Leticia",  "Medellín",  "Arauca",  "Barranquilla",
"Cartagena",
                        "Tunja", "Manizales", "Florencia", "Yopal", "Popayán",
"Valledupar",
                       "Quibdó", "Montería", "Bogotá", "Inírida", "San José del
Guaviare",
                          "Neiva", "Riohacha", "Santa Marta", "Villavicencio",
"Pasto",
                       "Cúcuta", "Mocoa", "Armenia", "Pereira", "San Andrés",
                         "Bucaramanga", "Sincelejo", "Ibagué", "Cali", "Mitú",
"Puerto Carreño", "Bogotá, D.C.")
)

# We perform a left join with the capitals lookup and then create the new
"Capital" column.
Ministery_Data_with_capital <- Ministery_Data_with_pop %>%
  left_join(capitals, by = "Departamento") %>%
  mutate(Capital = if_else(Municipio == CapitalMunicipio, "Si", "No"))
```

Finally, we clean the data subseting the columns of interest and elimintating NAS

```
Ministery_Data_Compiled <- Ministery_Data_with_capital %>%
  select(Año, Año_Cod, Departamento, Municipio, Capital, Diagnostico, Sexo, Edad,
Cantidad, Poblacion)

Ministery_Data_Compiled <- na.omit(Ministery_Data_Compiled)
```

We inspect the resulting dataset for analysis

```
head(Ministery_Data_Compiled)
```

```
# A tibble: 6 × 10
    Año Año_Cod Departamento Municipio Capital Diagnostico Sexo    Edad Cantidad
  <dbl>   <dbl> <chr>        <chr>     <chr>   <chr>       <chr>  <dbl>    <dbl>
1  2016       0 Antioquia    Medellín  Si      F700        Hombr…     2        1
2  2016       0 Antioquia    Medellín  Si      F700        Mujer…     4        2
3  2016       0 Antioquia    Medellín  Si      F700        Hombr…     4        1
4  2016       0 Antioquia    Medellín  Si      F700        Mujer…     6        1
5  2016       0 Antioquia    Medellín  Si      F700        Hombr…     6        1
6  2016       0 Antioquia    Medellín  Si      F700        Mujer…     7        3
# ℹ 1 more variable: Poblacion <dbl>
```

And write a .cvs file in the “Data_Processed” folder.

```
write.csv(Ministery_Data_Compiled,                                "Data_Processed/
Ministery_DiagnosticData_Compiled.csv")
```