Homicides in Colombia (2003-2023)

Reynell Badillo Sarmiento

07 diciembre 2024

Contents

Daily Homicides in Colombia (2003-2023)

For this project, the first big task will be to compile all events of homicides of Colombia between 2003-2023 in a single dataset. The Colombian National Police publishes yearly an Excel spreadsheet with all the homicides that have occurred in the country. However, these datasets are not necessarily consistent among different years, as they sometimes include new variables or write with different configurations (all caps, accents, etc...). My first task, then, will be to analyze these datasets and merge all of them in a single big dataset that can be used to compare homicides trends in Colombia. After that, I will show the usefulness of the new dataset in two ways: first, showing some descriptive trends of homicides in the country, and, second, in showing how we can use the dataset to answer some substantive questions about criminal violence and homicidal violence. I must say, I am not answering in particular any question with certainty. Rather, my purpose is to create this nice dataset that may be actually quite useful for my future research. In the third section, I will briefly try to explore some questions.

Cleaning the data

Cleaning the data (2003-2010)

First, let´s try to clean the data, as it is not ready for analysis. First, let´s get the data and have a peak of what we have. I will start with the 2003-2010 dataset because it is all included in a single one. Then I will try to identify patterns that I can use in a single function. The first thing is that the first 10 rows are just unimportant stuff (as the poilice logo, the person who created the dataset, etc...), so I will skip then. Then, I realize that, in the end, I have some irrelevant columns with additional information. I identify them as "NAs" in one of the variables, as the dataset does not contain any NA. I will also delete them.

```
homicides_2003_2010 <- read_excel("./raw_data/homicidios_20032010.xlsx", skip = 10) |>
  filter(!is.na(MUNICIPIO))
```

Now, this dataset has a somewhat different structure than other datasets, as it includes different variables and exclude others. I will first rename some of the variables and drop those that are irrelevant:

```
#Renaming variables
homicides_2003_2010 <- homicides_2003_2010 |>
  rename(
    departamento = "DEPARTAMENTO",
    municipio = "MUNICIPIO",
    date = "FECHA DE HECHO",
    quantity = "CANTIDAD"
  )

#Dropping irrelevant variables
homicides_2003_2010 <- homicides_2003_2010 |>
  select(-"DIA SEMANA", -"AÑO")
```

Now, I want to expand the "quantity" column, as I would want to get a single observation per homicide. Also, I want that instead of a single "date" column, I have "day", "week" and "year" for better analysis and grouping. I created a function that can help me with this, as this is something that I will need to do in every dataset:

```
expand_data <- function(data, date_col, quantity_col, sep = "-") {
  # Expands the data frame by duplicating rows based on a quantity column and separates a date column i

  # Args:
  #   data (tibble): input data
  #   date_col (character): name of the column with the date to be separated
  #   quantity_col (character): name of the column with quantities used for duplicating rows
  #   sep (character): separator used in the date column (default is "/")

  # Returns:
  #   tibble: a data frame with rows duplicated based on `quantity_col` and date column split into "day

  expanded_data <- data |>
    uncount(!!sym(quantity_col)) |>
    separate(!!sym(date_col), into = c("year", "month", "day"), sep = sep)

  return(expanded_data)
}
```

Let´s try!

```
head(expand_data(homicides_2003_2010, "date", "quantity"))
```

```
## # A tibble: 6 x 5
##   departamento municipio    year  month day
##   <chr>        <chr>        <chr> <chr> <chr>
## 1 AMAZONAS     Leticia (CT) 2003  01    01
## 2 AMAZONAS     Leticia (CT) 2003  03    04
## 3 AMAZONAS     Leticia (CT) 2003  03    22
## 4 AMAZONAS     Leticia (CT) 2003  03    22
## 5 AMAZONAS     Leticia (CT) 2003  05    17
## 6 AMAZONAS     Leticia (CT) 2003  05    19
```

It worked! So let´s save the new dataset

```
homicides_2003_2010 <- expand_data(homicides_2003_2010, "date", "quantity")
```

Cleaning the data (2011-2019)

Unfortunately, the dataset from 2003-2010 is a little bit different from the rest of the datasets, but still we can do little tweaks on the function to get similar results. Let´s try now with 2011-2019 (as the structure changes after 2020). We will do some renaming and we can repeat this process from 2011 to 2019 as they are the exact same. So I will include in my function that renaming.

```
homicides_2011 <- read_excel("./raw_data/homicidios_2011.xlsx", skip = 10) |>
  filter(!is.na(MUNICIPIO))
```

```
#Renaming variables
homicides_2011 <- homicides_2011 |>
  rename(
    departamento = "DEPARTAMENTO",
    municipio = "MUNICIPIO",
    municipio_code = "CODIGO DANE",
    weapon = "ARMAS MEDIOS",
    date = "FECHA HECHO",
    gender = "GENERO",
    age_group = "AGRUPA EDAD PERSONA",
    quantity = "CANTIDAD"
  )
```

Let´s create a function that renames and expand date and quantity

```
expand_data <- function(data, date_col, quantity_col, sep = "/") {
  # Expands the data frame by duplicating rows based on a quantity column and separates a date column i
  # Args:
  #   data (tibble): input data
  #   date_col (character): name of the column with the date to be separated
  #   quantity_col (character): name of the column with quantities used for uncounting/duplicating rows
  #   sep (character): separator used in the date column (default is "/")
  # Returns:
```

```
  #   tibble: a data frame with rows duplicated based on `quantity_col` and date column split into "day

  renamed_data <- data |>
    rename(
      departamento = "DEPARTAMENTO",
      municipio = "MUNICIPIO",
      municipio_code = "CODIGO DANE",
      weapon = "ARMAS MEDIOS",
      date = "FECHA HECHO",
      gender = "GENERO",
      age_group = "AGRUPA EDAD PERSONA",
      quantity = "CANTIDAD"
    )

  expanded_data <- renamed_data |>
    uncount(!!sym(quantity_col)) |>
    separate(!!sym(date_col), into = c("year", "month", "day"), sep = "-")

  return(expanded_data)
}
```

```
head(expand_data(read_excel("./raw_data/homicidios_2011.xlsx", skip = 10) |>
                     filter(!is.na(MUNICIPIO)),
                 "date",
                 "quantity"))
```

```
## # A tibble: 6 x 9
##   departamento municipio         municipio_code weapon  year  month day   gender
##   <chr>        <chr>                      <dbl> <chr>   <chr> <chr> <chr> <chr>
## 1 ANTIOQUIA    AMAGÁ                    5030000 ARMA B~ 2011  01    01    MASCU~
## 2 ANTIOQUIA    MEDELLÍN (CT)            5001000 ARMA B~ 2011  01    01    MASCU~
## 3 ANTIOQUIA    MEDELLÍN (CT)            5001000 ARMA B~ 2011  01    01    MASCU~
## 4 ANTIOQUIA    TURBO                    5837000 ARMA B~ 2011  01    01    MASCU~
## 5 ATLÁNTICO    BARRANQUILLA (CT)        8001000 ARMA B~ 2011  01    01    MASCU~
## 6 ATLÁNTICO    BARRANQUILLA (CT)        8001000 ARMA B~ 2011  01    01    MASCU~
## # i 1 more variable: age_group <chr>
```

Now, from analyzing the Excel documents, it can be observed that, at least until 2019, this structure is useful. So let´s try to create a function for that. First, I realized that while until 2019 the extension is "xlsx", from 2020 it is "xls", so I will take advantage of that:

```
paths_list_2011_2019 <- as.list(dir(path = "./raw_data/",
                          pattern = "xlsx$",
                          full.names = T))
#As I know that the 2003-2010 dataset is different, I will exclude it
paths_list_2011_2019 <- paths_list_2011_2019[-1]
```

Now let´s create a function to import the dataset skipping problametic columns and rename all the variables:

```
import_homicides_2011_2019 <- function(path) {
  # Import a dataset, eliminate irrelevant columns and rows, and rename variables
  # Arg:
```

```
  #   path: Relative path to the dataset
  # Returns:
  #   The data without irrelevant columns and rows
  data_cleaned <- read_excel(path, skip = 10) |> #read the data without first columns
   filter(!is.na(MUNICIPIO)) |> #select all columns except the one called "...67"
    rename( #rename the variables
      departamento = "DEPARTAMENTO",
      municipio = "MUNICIPIO",
      municipio_code = "CODIGO DANE",
      weapon = "ARMAS MEDIOS",
      date = "FECHA HECHO",
      gender = "GENERO",
      age_group = "AGRUPA EDAD PERSONA",
      quantity = "CANTIDAD"
    )
  return(data_cleaned)
}
```

And the function to expand the quantity and date

```
expand_data_2011_2019 <- function(data, date_col, quantity_col, sep = "-") {
  # Expands the data frame by duplicating rows based on a quantity column and separates a date column i
  # Args:
  #   data (tibble): input data
  #   date_col (character): name of the column with the date to be separated
  #   quantity_col (character): name of the column with quantities used for duplicating rows
  #   sep (character): separator used in the date column (default is "-")
  # Returns:
  #   tibble: a data frame with rows duplicated based on `quantity_col` and date column split into "yea

  expanded_data <- data |>
    uncount(!!sym(quantity_col)) |>
    separate(!!sym(date_col), into = c("year", "month", "day"), sep = "-")
  return(expanded_data)
}
```

Trying:

```
# Process for the year 2013
head(expand_data_2011_2019(import_homicides_2011_2019("./raw_data/homicidios_2013.xlsx"),
                      "date",
                      "quantity",
                      "-"))
```

```
## # A tibble: 6 x 9
##   departamento municipio      municipio_code weapon     year  month day   gender
##   <chr>        <chr>                  <dbl> <chr>      <chr> <chr> <chr> <chr>
## 1 ANTIOQUIA    BETULIA              5093000 ARMA BLAN~ 2013  01    01    MASCU~
## 2 ANTIOQUIA    BURITICÁ             5113000 ARMA BLAN~ 2013  01    01    MASCU~
## 3 ANTIOQUIA    MEDELLÍN (CT)        5001000 ARMA BLAN~ 2013  01    01    MASCU~
## 4 ATLÁNTICO    SOLEDAD              8758000 ARMA BLAN~ 2013  01    01    MASCU~
## 5 BOLÍVAR      CARTAGENA (CT)      13001000 ARMA BLAN~ 2013  01    01    MASCU~
## 6 BOLÍVAR      CARTAGENA (CT)      13001000 ARMA BLAN~ 2013  01    01    MASCU~
## # i 1 more variable: age_group <chr>
```

It is working! Now, let´s do a for loop to get all the datasets in our path

```
homicides_2011_2019 <- data.frame() #I created a data.frame to save all my datasets

# For loop to process and combine all datasets
for (i in seq_along(paths_list_2011_2019)) {
  # Import and clean the dataset
  imported_data <- import_homicides_2011_2019(paths_list_2011_2019[[i]])

  # Expand the cleaned dataset
  expanded_data <- expand_data_2011_2019(imported_data, "date", "quantity", "-")

  # Bind the datasets into the combined dataset
  homicides_2011_2019 <- rbind(homicides_2011_2019, expanded_data)
}

# Print the combined dataset to verify
head(homicides_2011_2019)
```

```
## # A tibble: 6 x 9
##   departamento municipio         municipio_code weapon  year  month day   gender
##   <chr>        <chr>             <chr>          <chr>   <chr> <chr> <chr> <chr>
## 1 ANTIOQUIA    AMAGÁ             5030000        ARMA B~ 2011  01    01    MASCU~
## 2 ANTIOQUIA    MEDELLÍN (CT)     5001000        ARMA B~ 2011  01    01    MASCU~
## 3 ANTIOQUIA    MEDELLÍN (CT)     5001000        ARMA B~ 2011  01    01    MASCU~
## 4 ANTIOQUIA    TURBO             5837000        ARMA B~ 2011  01    01    MASCU~
## 5 ATLÁNTICO    BARRANQUILLA (CT) 8001000        ARMA B~ 2011  01    01    MASCU~
## 6 ATLÁNTICO    BARRANQUILLA (CT) 8001000        ARMA B~ 2011  01    01    MASCU~
## # i 1 more variable: age_group <chr>
```

Cleaning the data (2020-2023)

Perfect! Now we "only" :) have to discover the structure from 2020 to 2023 that, hopefully, will be similar. It seems like the only change is that we only need to skip 9 columns instead of 10:

```
homicidios_2020 <- read_excel("./raw_data/homicidios_2020.xls", skip = 9) |>
  filter(!is.na(MUNICIPIO))

head(homicidios_2020)
```

```
## # A tibble: 6 x 8
##   DEPARTAMENTO MUNICIPIO `CODIGO DANE` `ARMAS MEDIOS` `FECHA HECHO`       GENERO
##   <chr>        <chr>     <chr>         <chr>          <dttm>              <chr>
## 1 AMAZONAS     LETICIA ~ 91001000      ARMA BLANCA /~ 2020-03-14 00:00:00 MASCU~
## 2 AMAZONAS     LETICIA ~ 91001000      ARMA BLANCA /~ 2020-10-20 00:00:00 MASCU~
## 3 AMAZONAS     LETICIA ~ 91001000      ARMA BLANCA /~ 2020-12-31 00:00:00 MASCU~
## 4 ANTIOQUIA    ABEJORRAL 05002000      ARMA BLANCA /~ 2020-07-05 00:00:00 MASCU~
## 5 ANTIOQUIA    ABEJORRAL 05002000      ARMA BLANCA /~ 2020-08-02 00:00:00 MASCU~
## 6 ANTIOQUIA    AMAGÁ     05030000      ARMA BLANCA /~ 2020-02-09 00:00:00 MASCU~
## # i 2 more variables: `AGRUPA EDAD PERSONA` <chr>, CANTIDAD <dbl>
```

As that is the case, then let´s directly slightly modify our original function and apply it here:

```r
import_homicides_2020 <- function(path) {
  # Import a dataset, eliminate irrelevant columns and rows, and rename variables
  # Arg:
  #   path: Relative path to the dataset
  # Returns:
  #   The data without irrelevant columns and rows
  data_cleaned <- read_excel(path, skip = 9) |> #read the data without first columns
   filter(!is.na(MUNICIPIO)) |> #select all columns except the one called "...67"
    rename( #rename the variables
      departamento = "DEPARTAMENTO",
      municipio = "MUNICIPIO",
      municipio_code = "CODIGO DANE",
      weapon = "ARMAS MEDIOS",
      date = "FECHA HECHO",
      gender = "GENERO",
      age_group = "AGRUPA EDAD PERSONA",
      quantity = "CANTIDAD"
    )
  return(data_cleaned)
}
```

And we can still use our function "expand_data_2011_2019", as nothing changes there

```r
expand_data_2011_2019(import_homicides_2020("./raw_data/homicidios_2020.xls"),
                      "date",
                      "quantity",
                      "-")
```

```
## # A tibble: 12,127 x 9
##    departamento municipio    municipio_code weapon      year  month day   gender
##    <chr>        <chr>        <chr>          <chr>       <chr> <chr> <chr> <chr>
##  1 AMAZONAS     LETICIA (CT) 91001000       ARMA BLANC~ 2020  03    14    MASCU~
##  2 AMAZONAS     LETICIA (CT) 91001000       ARMA BLANC~ 2020  10    20    MASCU~
##  3 AMAZONAS     LETICIA (CT) 91001000       ARMA BLANC~ 2020  12    31    MASCU~
##  4 ANTIOQUIA    ABEJORRAL    05002000       ARMA BLANC~ 2020  07    05    MASCU~
##  5 ANTIOQUIA    ABEJORRAL    05002000       ARMA BLANC~ 2020  08    02    MASCU~
##  6 ANTIOQUIA    AMAGÁ        05030000       ARMA BLANC~ 2020  02    09    MASCU~
##  7 ANTIOQUIA    AMAGÁ        05030000       ARMA BLANC~ 2020  05    14    MASCU~
##  8 ANTIOQUIA    AMAGÁ        05030000       ARMA BLANC~ 2020  11    22    MASCU~
##  9 ANTIOQUIA    ANDES        05034000       ARMA BLANC~ 2020  01    17    MASCU~
## 10 ANTIOQUIA    ANDES        05034000       ARMA BLANC~ 2020  03    12    MASCU~
## # i 12,117 more rows
## # i 1 more variable: age_group <chr>
```

Thus, let´s get the paths and apply the function into a for loop!

```r
paths_list_2020_2023 <- as.list(dir(path = "./raw_data/",
                          pattern = "xls$",
                          full.names = T))
```

```r
homicides_2020_2023 <- data.frame() #I created a data.frame to save all my datasets
```

```r
# For loop to process and combine all datasets
for (i in seq_along(paths_list_2020_2023)) {
  # Import and clean the dataset
  imported_data <- import_homicides_2020(paths_list_2020_2023[[i]])

  # Expand the cleaned dataset
  expanded_data <- expand_data_2011_2019(imported_data, "date", "quantity", "-")

  # Bind the datasets into the combined dataset
  homicides_2020_2023 <- rbind(homicides_2020_2023, expanded_data)
}

# Print the combined dataset to verify
head(homicides_2020_2023)
```

```
## # A tibble: 6 x 9
##   departamento municipio    municipio_code weapon       year  month day   gender
##   <chr>        <chr>        <chr>          <chr>        <chr> <chr> <chr> <chr>
## 1 AMAZONAS     LETICIA (CT) 91001000       ARMA BLANCA~ 2020  03    14    MASCU~
## 2 AMAZONAS     LETICIA (CT) 91001000       ARMA BLANCA~ 2020  10    20    MASCU~
## 3 AMAZONAS     LETICIA (CT) 91001000       ARMA BLANCA~ 2020  12    31    MASCU~
## 4 ANTIOQUIA    ABEJORRAL    05002000       ARMA BLANCA~ 2020  07    05    MASCU~
## 5 ANTIOQUIA    ABEJORRAL    05002000       ARMA BLANCA~ 2020  08    02    MASCU~
## 6 ANTIOQUIA    AMAGÁ        05030000       ARMA BLANCA~ 2020  02    09    MASCU~
## # i 1 more variable: age_group <chr>
```

Combining our three datasets

Before continuing, it may be useful fill in the "municipio_code" variable from 2003-2010, as we have that information actually from the other datasets (and in the future it will also be super handy to combine with other datasets). First of all, there are two errors in the police code that we need to solve in the two datasets:

There is problem because the homicide datasets have three extra zeros in the code for no reason :) So I will try to remove them. Also, the Police omitted a "0" before municipalities whose code only has 4 digits, so I will add it:

```r
homicides_2020_2023 <- homicides_2020_2023 |>
  mutate(
    # Remove extra zeros
    municipio_code = str_remove(municipio_code, "000$"),
    # Add a zero for codes with exactly 4 digits
    municipio_code = ifelse(nchar(municipio_code) == 4,
                            paste0("0", municipio_code),  # Add leading zero
                            municipio_code)
  )

homicides_2011_2019 <- homicides_2011_2019 |>
  mutate(
    # Remove extra zeros
    municipio_code = str_remove(municipio_code, "000$"),
    # Add a zero for codes with exactly 4 digits
    municipio_code = ifelse(nchar(municipio_code) == 4,
                            paste0("0", municipio_code),  # Add leading zero
```

```
                            municipio_code)
  )

#Some municipios have 3 random extra-digits for no reason, so I will remove them
homicides_2020_2023 <- homicides_2020_2023 |>
  mutate(municipio_code = substr(as.character(municipio_code), 1, 5))

homicides_2011_2019 <- homicides_2011_2019 |>
  mutate(municipio_code = substr(as.character(municipio_code), 1, 5))
```

Now we can add those municipio_codes to our dataset from 2003-2010

```
homicides_2011_2019_codes <- homicides_2011_2019 |>
  mutate(
    departamento = toupper(departamento), # all caps so they are consistent
    municipio = toupper(municipio)
  ) |>
  select(departamento, municipio, municipio_code) |> #The three as there are municipios with the same n
  distinct() |>
  group_by(departamento, municipio) |>
  summarise(municipio_code = first(municipio_code), .groups = 'drop')


#Left join so I get the municipio_code for 2003-2010
homicides_2003_2010 <- homicides_2003_2010 |>
 mutate(
    departamento = toupper(departamento), # all caps so they are consistent
    municipio = toupper(municipio)
  ) |>
  left_join(homicides_2011_2019_codes, by = c("departamento", "municipio"))
```

Now, there are still 23 municipalities without code due to errors writing the name of the municipalities or the fact that there were no homicides in the other dataset. I will add those manually looking at the webpage of the DANE in Colombia.

```
homicides_2003_2010 |>
  filter(is.na(municipio_code)) |>
  distinct() |>
   group_by(municipio) |>
  summarise(municipio_code = first(municipio_code),
            departamento = first(departamento)) # Ensure only unique municipio
```

```
## # A tibble: 27 x 3
##    municipio  municipio_code departamento
##    <chr>      <chr>          <chr>
##  1 AGUADA     <NA>           SANTANDER
##  2 ARATOCA    <NA>           SANTANDER
##  3 BERBEO     <NA>           BOYACÁ
##  4 BRICEÑO    <NA>           BOYACÁ
##  5 CALIFORNIA <NA>           SANTANDER
##  6 CEPITÁ     <NA>           SANTANDER
##  7 CHIVOLO    <NA>           MAGDALENA
```

```
##  8 CHIVOR     <NA>           BOYACÁ
##  9 CIÉNEGA    <NA>           BOYACÁ
## 10 COLÓN      <NA>           PUTUMAYO
## # i 17 more rows
```

```
homicides_2003_2010 <- homicides_2003_2010  |>
  mutate(municipio = case_when(
    municipio == "CIÉNEGA" ~ "CIÉNAGA",
    municipio == "FÚNEQUE" ~ "FÚQUENE",
    TRUE ~ municipio
  ))
```

```
homicides_2003_2010 <- homicides_2003_2010 |>
  mutate(municipio_code = case_when(
    municipio == "SOPLAVIENTO" & departamento == "BOLÍVAR" ~ "13760",
    municipio == "AGUADA" & departamento == "SANTANDER" ~ "68013",
    municipio == "BERBEO" & departamento == "BOYACÁ" ~ "15090",
    municipio == "BRICEÑO" & departamento == "BOYACÁ" ~ "05107",
    municipio == "CALIFORNIA" & departamento == "SANTANDER" ~ "68132",
    municipio == "CEPITÁ" & departamento == "SANTANDER" ~ "68160",
    municipio == "CHIVOLO" & departamento == "MAGDALENA" ~ "47170",
    municipio == "CHIVOR" & departamento == "BOYACÁ" ~ "15236",
    municipio == "CIÉNEGA" & departamento == "BOYACÁ" ~ "15189",
    municipio == "CIÉNAGA" & departamento == "BOYACÁ" ~ "15189",
    municipio == "ARATOCA" & departamento == "SANTANDER" ~ "68051",
    municipio == "CORRALES" & departamento == "BOYACÁ" ~ "15215",
    municipio == "EL CALVARIO" & departamento == "META" ~ "50245",
    municipio == "EL ESPINO" & departamento == "BOYACÁ" ~ "15248",
    municipio == "EL GUACAMAYO" & departamento == "SANTANDER" ~ "68245",
    municipio == "FÚQUENE" & departamento == "CUNDINAMARCA" ~ "25288",
    municipio == "GAMA" & departamento == "CUNDINAMARCA" ~ "25299",
    municipio == "LA CAPILLA" & departamento == "BOYACÁ" ~ "15380",
    municipio == "MACARAVITA" & departamento == "SANTANDER" ~ "68425",
    municipio == "PALMAR" & departamento == "SANTANDER" ~ "68522",
    municipio == "RONDÓN" & departamento == "BOYACÁ" ~ "15621",
    municipio == "SAN JUANITO" & departamento == "META" ~ "50686",
    municipio == "SATIVANORTE" & departamento == "BOYACÁ" ~ "15720",
    municipio == "TENZA" & departamento == "BOYACÁ" ~ "15798",
    municipio == "UBAQUE" & departamento == "CUNDINAMARCA" ~ "25841",
    municipio == "COLÓN" & departamento == "PUTUMAYO" ~ "86219",
    municipio == "CONCEPCIÓN" & departamento == "SANTANDER" ~ "68207",
    municipio == "SAN MIGUEL" & departamento == "SANTANDER" ~ "68686",
    TRUE ~ as.character(municipio_code) # Retain existing values as character for other rows
  ))

homicides_2003_2010 |>
  filter(is.na(municipio_code)) |>
  distinct() |>
   group_by(municipio) |>
  summarise(municipio_code = first(municipio_code),
            departamento = first(departamento)) # Ensure only unique municipio
```

```
## # A tibble: 0 x 3
## # i 3 variables: municipio <chr>, municipio_code <chr>, departamento <chr>
```

Perfect! Now we are (finally) ready to bind the datasets. As the 2003-2010 dataset does not have data for some variables, I will use "bind_rows", that merges them and creates NA values for all missing columns.

```
homicides_2003_2023 <- bind_rows(homicides_2003_2010, homicides_2011_2019, homicides_2020_2023)
```

It worked! So now we have an entire single dataset with all homicides between 2003 and 2023 in Colombia. We can have very interesting descriptive data here. But before, given that some words are written in all caps and others not, I will just use a code to change everything to caps to make it consistent:

```
homicides_2003_2023 <- homicides_2003_2023 |>
  mutate(
    departamento = toupper(departamento),
    municipio = toupper(municipio),
    weapon = toupper(weapon),
    gender = toupper(gender),
    age_group = toupper(age_group)
  )
```

And, just because it may be useful, I created a new variable for every region of the country:

```
#Creating a new variable per region
homicides_2003_2023 <- homicides_2003_2023 |>
    mutate(region = case_when(
    departamento %in% c("CUNDINAMARCA", "ANTIOQUIA", "BOYACÁ", "CALDAS",
                        "HUILA", "NORTE DE SANTANDER", "QUINDÍO", "RISARALDA",
                        "SANTANDER", "TOLIMA") ~ "Andina",
    departamento %in% c("AMAZONAS", "CAQUETÁ", "GUAINÍA", "GUAVIARE",
                        "PUTUMAYO", "VAUPÉS") ~ "Amazonia",
    departamento %in% c("VALLE DEL CAUCA", "VALLE", "CHOCÓ", "CAUCA",
                        "NARIÑO") ~ "Pacifico",
    departamento %in% c("ATLÁNTICO", "BOLÍVAR", "CESAR", "CÓRDOBA",
                        "LA GUAJIRA", "GUAJIRA", "MAGDALENA", "SUCRE", "SAN ANDRÉS") ~ "Caribe",
    departamento %in% c("ARAUCA", "CASANARE", "META", "VICHADA") ~ "Orinoquia",
    TRUE ~ NA_character_
  ))
```

As the dataset is ready, let´s save it as a csv:

```
write_csv(homicides_2003_2023, "homicides_2003_2023.csv")
```

Showing the usefulness of the dataset for research

Example 1: Descriptive data about homicides in Colombia

Trends using the total number of homicides

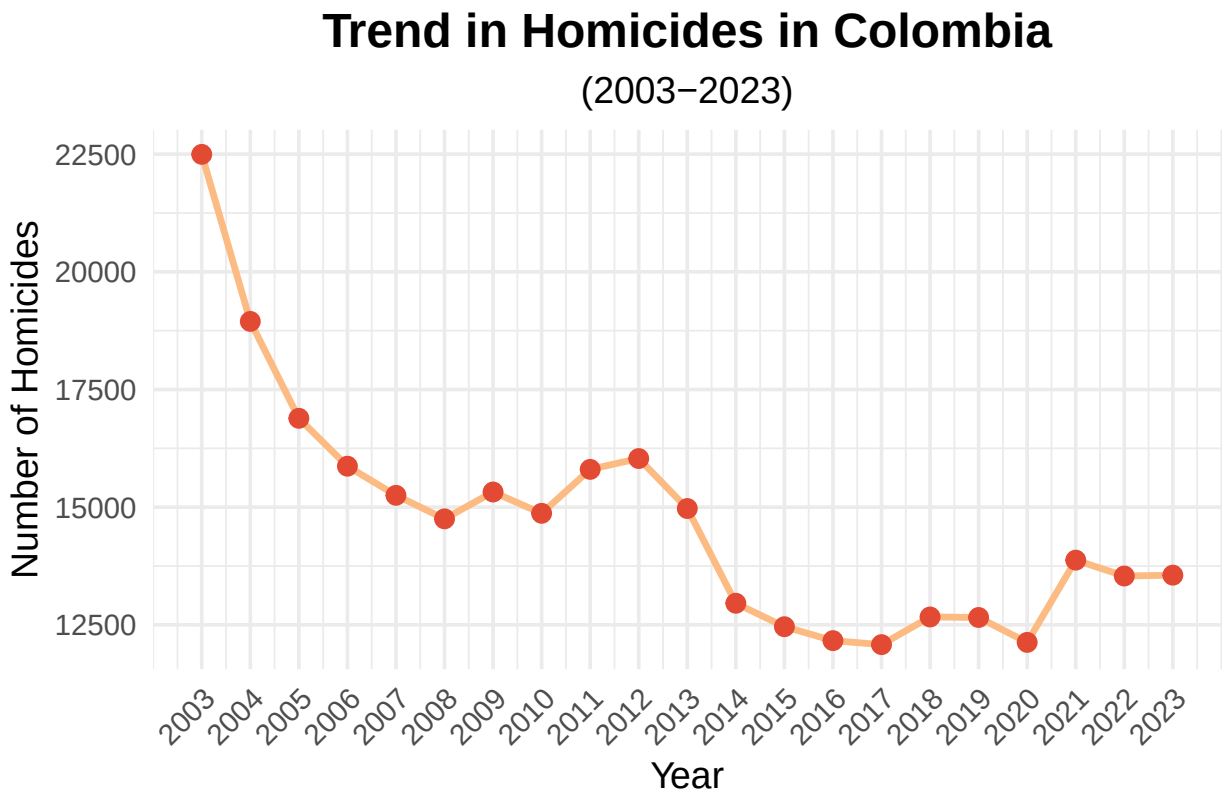
Before showing how useful the dataset can be for research purposes, let´s show first that, for people interested in mapping homicide trends, the dataset works very smoothly.

First, let´s show a simple trend line graph about homicides in the country during all the years of the dataset:

```r
# Summarize data to get the count of homicides per year
graph1_homicides_trends <-
  homicides_2003_2023 |>
  mutate(year = as.numeric(year)) |>
  group_by(year) |>
  summarize(count = n()) |>
# Now I will plot homicides by year
ggplot(aes(x = year, y = count, group = 1)) +
  geom_line(color = "#fdbb84", size = 1.2) +
  geom_point(color = "#e34a33", size = 3) +
  labs(
    title = "Trend in Homicides in Colombia",
    subtitle = "(2003-2023)",
    caption = "Data Source: Policía Nacional de Colombia",
    x = "Year",
    y = "Number of Homicides"
  ) +
  theme_minimal(base_size = 14) +  # Adjust base font size
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),  # Center title and increase size
    plot.subtitle = element_text(hjust = 0.5, size = 14),  # Center subtitle
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels for better readability a
    plot.caption = element_text(hjust = 1, size = 10)   # Right align caption
  ) +
  scale_x_continuous(breaks = seq(min(homicides_2003_2023$year), max(homicides_2003_2023$year), by = 1)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

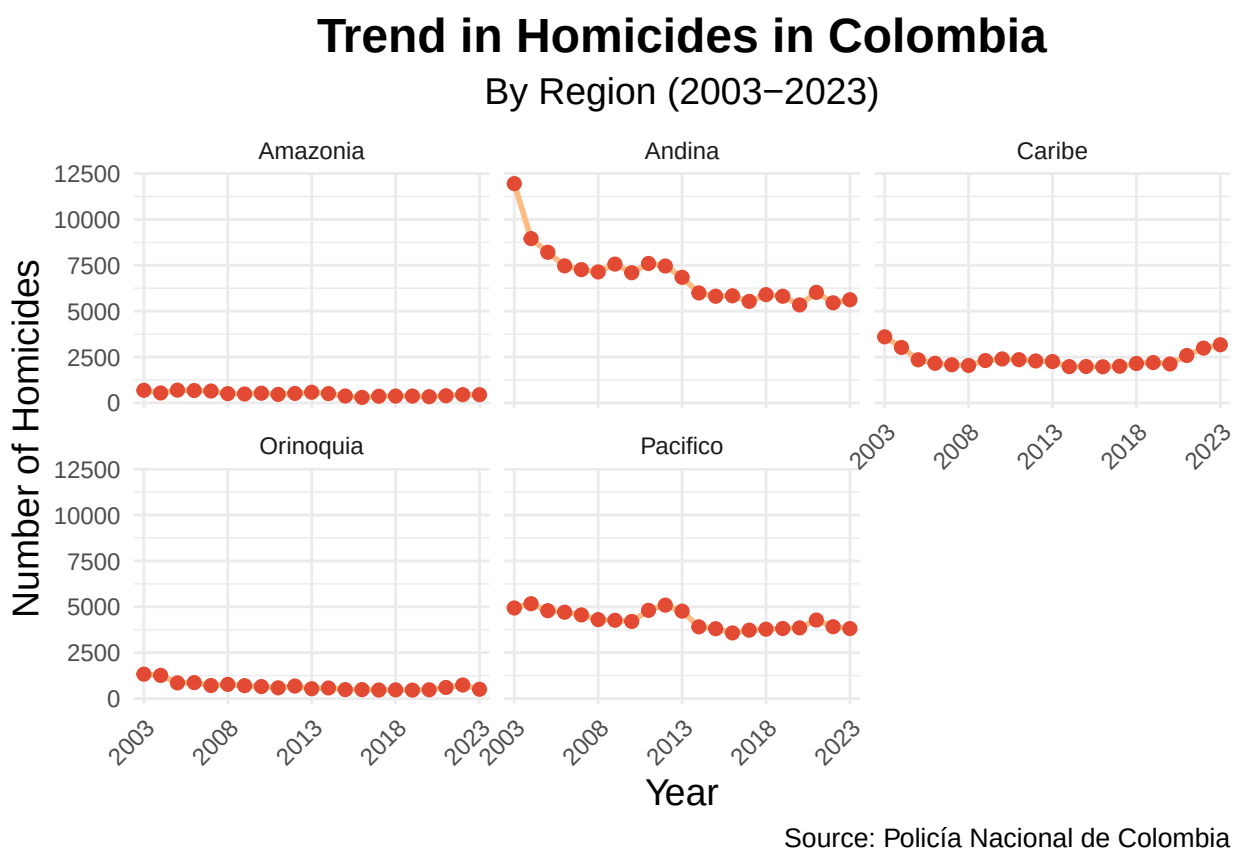
```r
print(graph1_homicides_trends)
```



As observed, there is a very interesting trend of notable decrease in homicides since 2003. Between 2003-2006, a paramilitary group, the AUC, got demobilized, which can help to understand the magnitude of the decrease. Homicides remained relatively stable, with some increases between 2008 and 2012, and then started decreasing until 2020, when we see a new peak. Now, let´s analyze homicide numbers by regions (I will use similar structures as before, in case I do not explain the code in detail):

```
graph2_homicides_regions <-
  homicides_2003_2023 |>
  group_by(year, region) |>
  summarize(count = n(), .groups = 'drop') |>
ggplot(aes(x = factor(year), y = count)) +
  geom_line(color = "#fdbb84", size = 1, group = 1) +
  geom_point(color = "#e34a33", size = 2, group = 1) +
  facet_wrap(~ region) +
  labs(
    title = "Trend in Homicides in Colombia",
    subtitle = "By Region (2003-2023)",
    x = "Year",
    y = "Number of Homicides",
    caption = "Source: Policía Nacional de Colombia"
  ) +
  scale_x_discrete(breaks = seq(2003, 2023, by = 5)) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14),
    axis.title.x = element_text(size = 14),
```

```
    axis.title.y = element_text(size = 14),
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.caption = element_text(hjust = 1, size = 10)
  )

print(graph2_homicides_regions)
```



Trends using the homicide rates per 100,000 inhabitants

Getting municipality-year data Now, it must be considered that just taking into account the number of homicides can be misleading, so I will also merge this dataset with the population dataset (obtained from the webpage of the DANE, the Statistical department of Colombia) by year and get a homicide rate:

```
#Importing data
population_2003_2017 <- read_excel("./raw_data/population/poblacion_2003_2017.xls")
population_2018_2023 <- read_excel("./raw_data/population/poblacion_2018_2023.xlsx")


#Let´s make this tidy!
population_2003_2017 <- population_2003_2017 |>
  pivot_longer(
    cols = `2003`:`2017`,     # range of years I will use
    names_to = "year",         # new column
    values_to = "poblacion"   # Population counts
```

```
  )

#Making year character just to make it consistent with my previous dataset
population_2003_2017$year <- as.character(population_2003_2017$year)
population_2018_2023$year <- as.character(population_2018_2023$year)

#And make a single dataset with all years:
population_total <- bind_rows(population_2003_2017, population_2018_2023)

#selecting just the relevant variables
population_total <- population_total  |>
  select(municipio_code, year, poblacion)
```

Now, I will first create a new dataset in which I group homicides by municipality-year

```
homicides_yearly <- homicides_2003_2023 |>
  group_by(departamento, municipio, year)  |>
  summarize(
    total_homicides = n(),                  # Count homicides per municipality-year
    municipio_code = first(municipio_code), # Just a single code per group
    region = first(region)                  # Just a single region per group
  ) |>
  ungroup() # Remove grouping after summarizing

head(homicides_yearly)
```

```
## # A tibble: 6 x 6
##   departamento municipio    year  total_homicides municipio_code region
##   <chr>        <chr>        <chr>           <int> <chr>          <chr>
## 1 AMAZONAS     LETICIA (CT) 2003               10 91001          Amazonia
## 2 AMAZONAS     LETICIA (CT) 2004                5 91001          Amazonia
## 3 AMAZONAS     LETICIA (CT) 2005               11 91001          Amazonia
## 4 AMAZONAS     LETICIA (CT) 2006               10 91001          Amazonia
## 5 AMAZONAS     LETICIA (CT) 2007                8 91001          Amazonia
## 6 AMAZONAS     LETICIA (CT) 2008                9 91001          Amazonia
```

Now let´s add the population!

```
homicides_yearly <- homicides_yearly |>
  left_join(population_total, by = c("municipio_code", "year")) |>
  rename(
    population = poblacion
  )
```

```
municipality_year <- homicides_yearly |>
  mutate(homic_rate = (total_homicides / population) * 100000)  |>
  mutate(homic_rate = replace(homic_rate, is.infinite(homic_rate), NA))

head(municipality_year)
```

```
## # A tibble: 6 x 8
```

```
##   departamento municipio  year  total_homicides municipio_code region population
##   <chr>        <chr>      <chr>           <int> <chr>          <chr>       <dbl>
## 1 AMAZONAS     LETICIA (~ 2003               10 91001          Amazo~      37049
## 2 AMAZONAS     LETICIA (~ 2004                5 91001          Amazo~      37459
## 3 AMAZONAS     LETICIA (~ 2005               11 91001          Amazo~      37832
## 4 AMAZONAS     LETICIA (~ 2006               10 91001          Amazo~      38234
## 5 AMAZONAS     LETICIA (~ 2007                8 91001          Amazo~      38609
## 6 AMAZONAS     LETICIA (~ 2008                9 91001          Amazo~      38957
## # i 1 more variable: homic_rate <dbl>
```

Let´s save this dataset as I think it may be useful

```
write_csv(municipality_year, "municipality_year.csv")
```

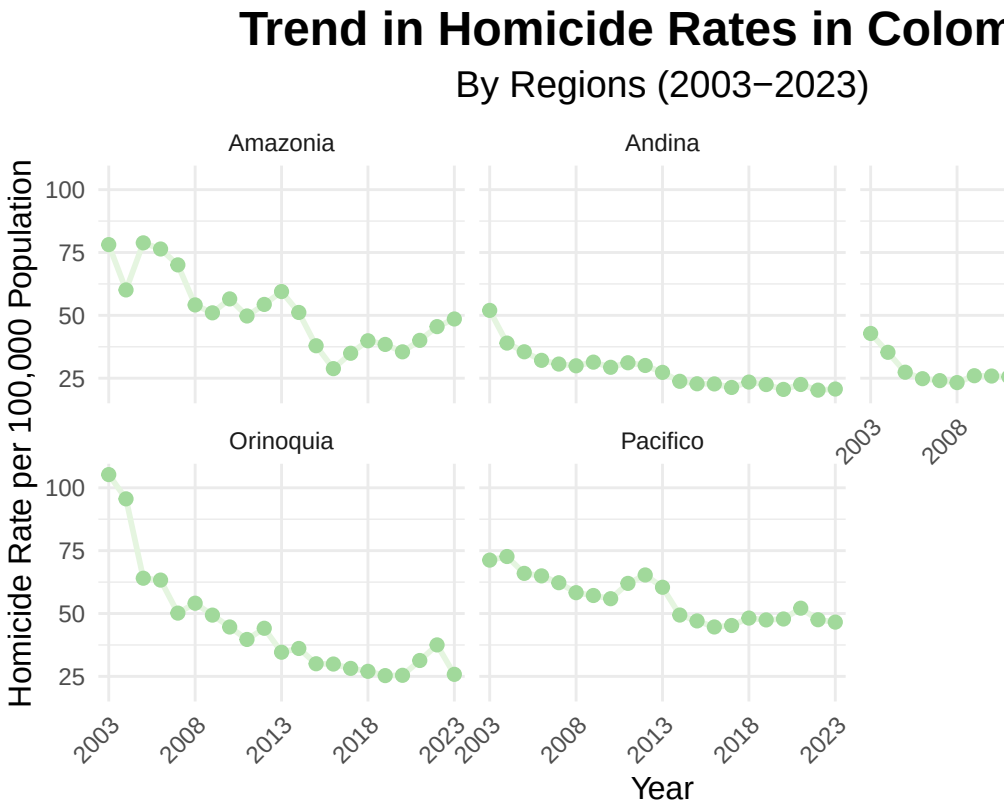
Getting region-year data I will do a similar procedure, but just for region-year data:

```
region_year <- municipality_year |>
  group_by(region, year) |>
  summarize(
    total_homicides = sum(total_homicides, na.rm = TRUE),
    total_population = sum(population, na.rm = TRUE)
  ) |>
  ungroup() |>
  mutate(homicide_rate = (total_homicides / total_population) * 100000)  # Per 100,000 population

head(region_year)
```

```
## # A tibble: 6 x 5
##   region   year  total_homicides total_population homicide_rate
##   <chr>    <chr>           <int>            <dbl>         <dbl>
## 1 Amazonia 2003              685           876720          78.1
## 2 Amazonia 2004              541           900021          60.1
## 3 Amazonia 2005              693           879070          78.8
## 4 Amazonia 2006              673           881048          76.4
## 5 Amazonia 2007              645           920469          70.1
## 6 Amazonia 2008              504           930794          54.1
```

```
graph3_homicides_region <-
  region_year |>
  ggplot(aes(x = factor(year), y = homicide_rate)) +
  geom_line(color = "#e5f5e0", size = 1, group = 1) +
  geom_point(color = "#a1d99b", size = 2, group = 1) +
  facet_wrap(~ region) +
  labs(
    title = "Trend in Homicide Rates in Colombia",
    subtitle = "By Regions (2003-2023)",
    x = "Year",
    y = "Homicide Rate per 100,000 Population",
    caption = "Source: Policía Nacional de Colombia"
```

```
  ) +
  scale_x_discrete(breaks = seq(2003, 2023, by = 5)) +
  theme_minimal() +
    theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.caption = element_text(hjust = 1, size = 10)
  )

print(graph3_homicides_region)
```

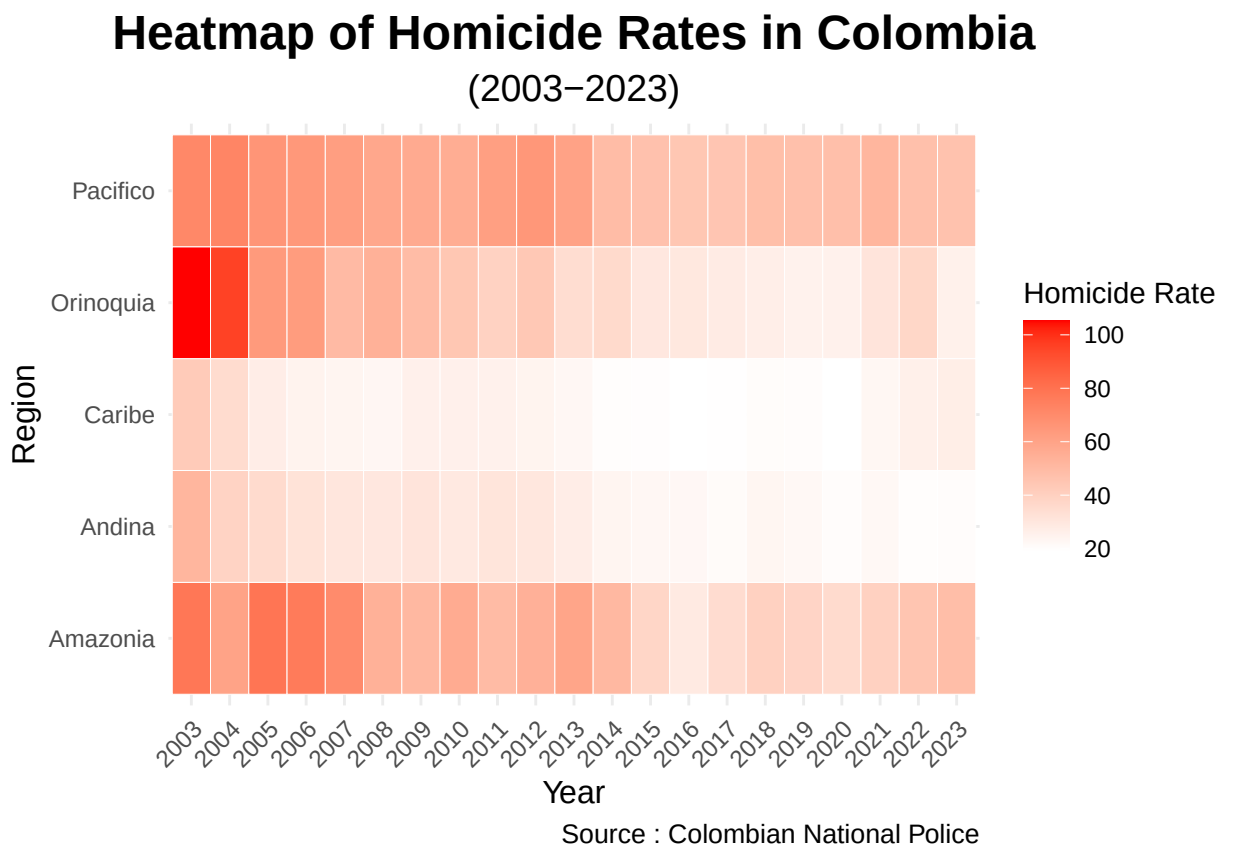


Source: Policía Nac

Descriptive graphs region level

```
graph4_heatmap_homicides_regions <-
  region_year |>
  ggplot(aes(x = factor(year), y = region)) +
  geom_tile(aes(fill = homicide_rate), color = "white") + #division color in every year
  scale_fill_gradient(low = "white", high = "red") + #Gradient of homicide rates
  labs(
    title = "Heatmap of Homicide Rates in Colombia",
    subtitle = "(2003-2023)",
    x = "Year",
    y = "Region",
    fill = "Homicide Rate",
```

```
    caption = "Source : Colombian National Police"
  ) +
  theme_minimal() +
     theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.caption = element_text(hjust = 1, size = 10)
  )

print(graph4_heatmap_homicides_regions)
```

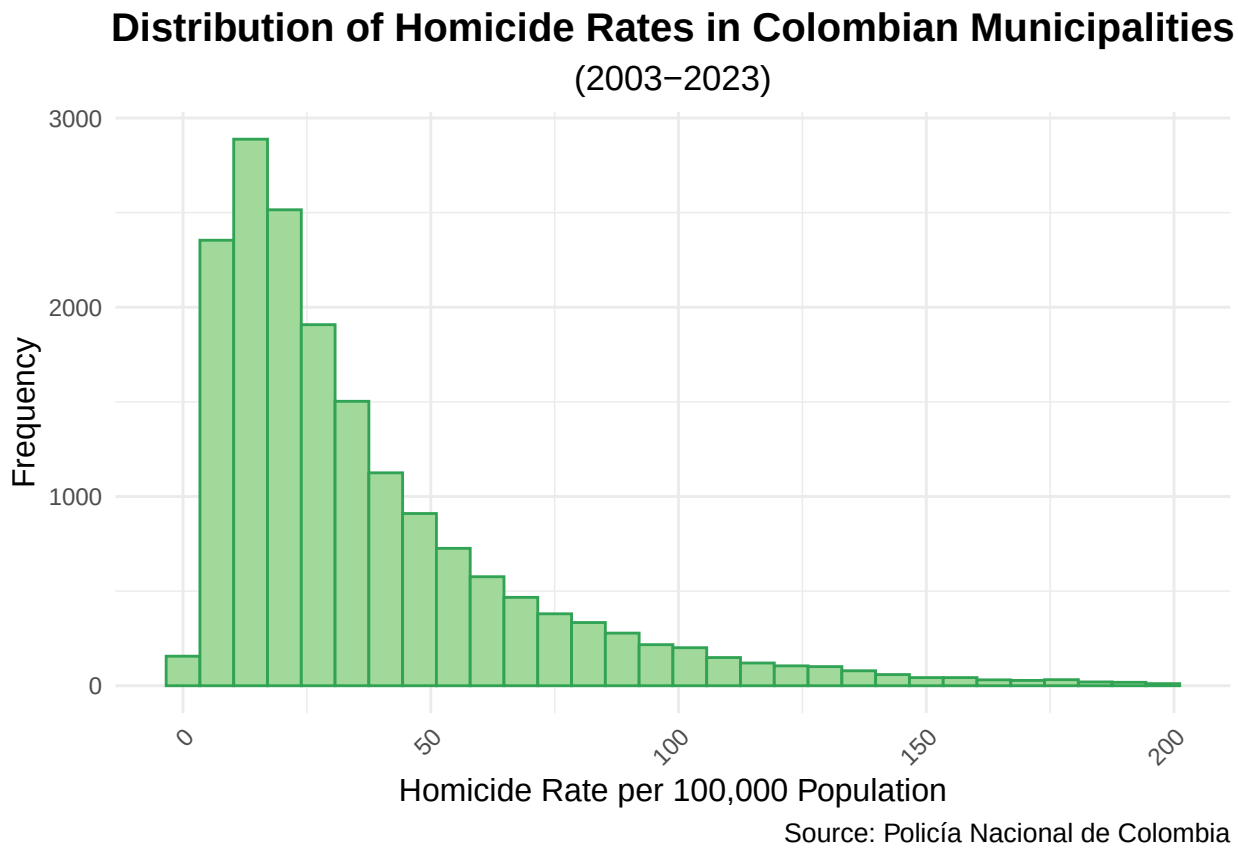


The heatmap using homicide rates is very interesting because then it shows that, even though homicides are higher in the Caribe and Andina region, when controlling for population the effect is not that significant. Amazonia and Orinoquia, although with less total homicides, have a considerably high homicide rate in comparison with other regions. It must also be acknowledge that, in general, homicide rates are particularly high in the country: the variation from 20 to 100 is actually notably high if compared to other countries of the region or the world.

Descriptive graphs municipality level We can also plot the distribution of homicide rates in Colombia across different municipalities during the last 20 years:

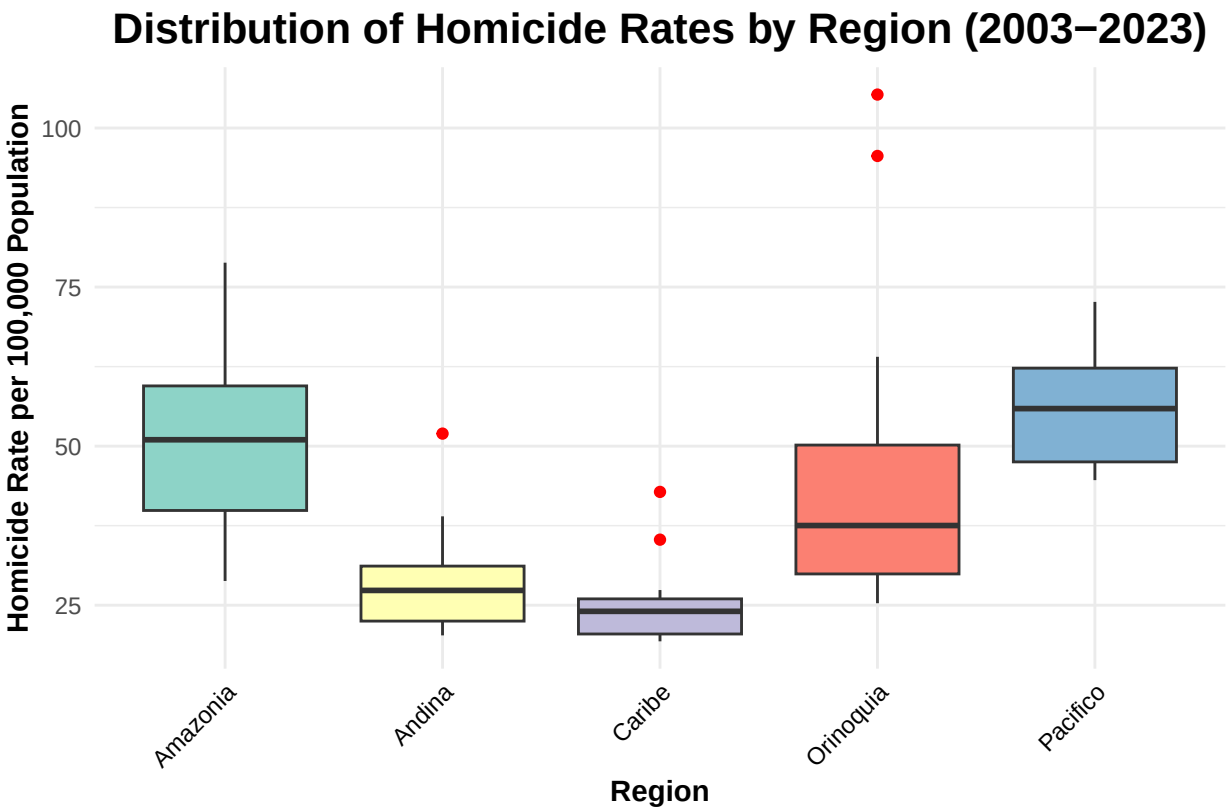
```r
graph5_distribution_homicide_rates <-
  municipality_year |>
  filter(homic_rate < 200) |>
  ggplot(aes(x = homic_rate)) +
  geom_histogram(fill = "#a1d99b", color = "#31a354") +
  labs(
    title = "Distribution of Homicide Rates in Colombian Municipalities",
    subtitle = "(2003-2023)",
    x = "Homicide Rate per 100,000 Population",
    y = "Frequency",
    caption = "Source: Policía Nacional de Colombia"
  ) +
  theme_minimal() +
      theme(
    plot.title = element_text(hjust = 0.5, size = 15, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 13),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.caption = element_text(hjust = 1, size = 10)
  )

print(graph5_distribution_homicide_rates)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As observed (and expected), most of the homicide rates in Colombia during the entire panel data have been between 0 and 35 homicides per 100,000 people. Given that countries homicide rates have been around 20-40 during the last 20 years, it makes total sense to observe this distribution.

```
graph6_boxplot_regions <- ggplot(region_year,
                                 aes(x = region,
                                     y = homicide_rate,
                                     fill = region)) +
  geom_boxplot(outlier.color = "red") +
  scale_fill_brewer(palette = "Set3", name = "Region") +
  labs(
    title = "Distribution of Homicide Rates by Region (2003-2023)",
    x = "Region",
    y = "Homicide Rate per 100,000 Population",
    caption = "Source: Policía Nacional de Colombia"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, color = "black"),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    plot.caption = element_text(size = 10, face = "italic"),
    legend.position = "none"  # Hide legend if not needed
  )

print(graph6_boxplot_regions)
```



Source: Policía Nacional de Colombia

Through the boxplots, we can also see some interesting variations. The Caribe region and Andina region are the ones with the smallest averages and also with the smallest variation. The Pacifico region is the one with the highest average, very close to the Amazonia region. There are some outliers, but they do not seem to be driving the variation, as they are not a lot.

Descriptive graphs department level (2023) Now, clearly mixing homicides during 20 years implies a lot of variation that is really hard to explain. So why don´t we just take a snapshot of homicides during last year in the country by department:

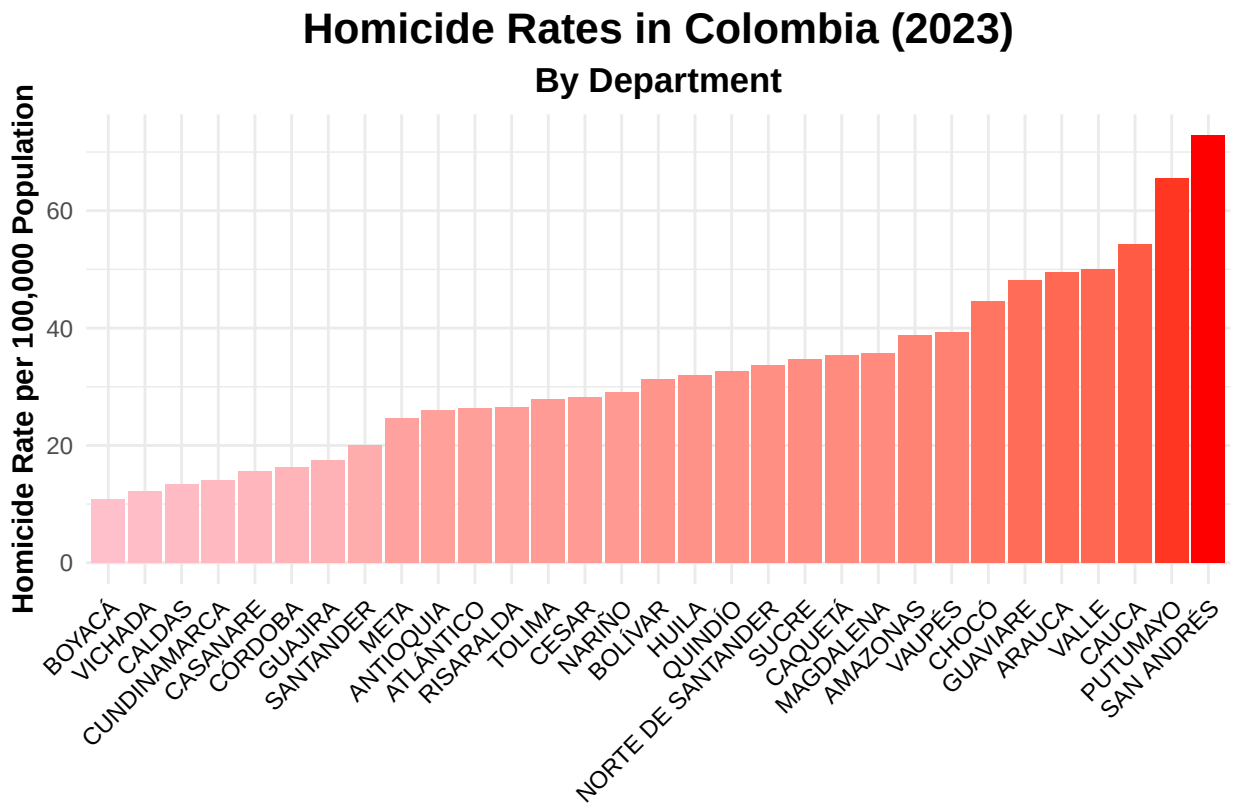
```
department_year <- municipality_year |>
  group_by(departamento, year) |>
  summarize(
    total_homicides = sum(total_homicides, na.rm = TRUE),
    total_population = sum(population, na.rm = TRUE)
  ) |>
  ungroup() |>
  mutate(homicide_rate = (total_homicides / total_population) * 100000)  # Per 100,000 population

head(department_year)
```

```
## # A tibble: 6 x 5
##   departamento year  total_homicides total_population homicide_rate
##   <chr>        <chr>           <int>            <dbl>         <dbl>
## 1 AMAZONAS     2003               13            43689          29.8
## 2 AMAZONAS     2004                5            37459          13.3
## 3 AMAZONAS     2005               11            37832          29.1
```

```
## 4 AMAZONAS     2006               10            38234          26.2
## 5 AMAZONAS     2007                8            38609          20.7
## 6 AMAZONAS     2008                9            38957          23.1
```

```r
#getting the csv
write_csv(department_year, "department_year.csv")
```

```r
graph7_2023_department <- department_year |>
    filter(year == 2023) |>
    arrange(homicide_rate)  |>
ggplot(aes(x = reorder(departamento, homicide_rate), y = homicide_rate, fill = homicide_rate)) +
  geom_bar(stat = "identity") +
  scale_fill_gradient(low = "pink", high = "red") +
  labs(
    title = "Homicide Rates in Colombia (2023)",
    subtitle = "By Department",
    x = NULL,
    y = "Homicide Rate per 100,000 Population",
    fill = "Homicide Rate",
    caption = "Source: Policía Nacional de Colombia"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, color = "black"),
    axis.title.y = element_text(face = "bold"),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    plot.subtitle = element_text(hjust = 0.5, face = "bold", size = 13),
    plot.caption = element_text(size = 10, face = "italic"),
    legend.position = "none"  # Hide legend if not needed
  )

print(graph7_2023_department)
```



Source: Policía Nacional de Colombia

This very same graph can be produced for every year, month, week or day, which may be interesting for people trying to understand more localized variation.

Example 2: Explaining competition between armed groups

This dataset, more than just helping out with some descriptive statistics about the distribution of homicides in Colombia or the geographical patterns, can also be useful to answer interesting questions about criminal violence. I will make this case using an example from my own research: how criminal groups use violence (or not) to compete. I study three cities of Colombia: Barranquilla, Santa Marta and Cartagena. Interestingly, I got data from official entities in which they coded years in which there was competition between criminal groups vs. years in which in these cities armed groups were competing (competition), they were the single armed group (monopoly) or they were sharing the illicit markets with other groups without competing (duopoly). In the literature, it is usually mentioned that when criminal groups compete, homicidal violence increases. On the contrary, when criminal groups get the monopoly of violence or establish a duopoly of violence, criminal violence usually decreases.

Let´s try to observe this behavior in these three cities. One of the hypotheses that I am trying to defend in another work is that this behavior of homicidal violence during competition depends on the strategy of states´ crackdowns. In other words: if the state makes a "conditional crackdown", in which it shows criminal groups that if they increase homicides the state will crackdown on them, they may actually compete, but using less violence. In those scenarios, after conditional crackdowns, even though we may observe competition, we will not necessarily observe sharp increases in homicides. Barranquilla, Santa Marta and Cartagena are really nice cases to show that because the state actually made a conditional crackdown in 2010 in the first two, but not in the third one. They are really similar cities, so we can assume these cases to be a very small-n comparison in which, if my idea is true, homicidal violence during competition will behave different than in our established conceptions of violence. Let´s see that in the data!

First, let´s get a municipality-year dataset only with our three cities. I will select the years 2007-2023 because I am not interested in homicides committed by paramilitary groups until 2006, and also because I do not have data of competition before 2006.

```
BCS_homicide_rates <- municipality_year |>
  filter(year > 2007 & (municipio == "BARRANQUILLA (CT)" |
                          municipio == "CARTAGENA (CT)" |
                          municipio == "SANTA MARTA (CT)"))
```

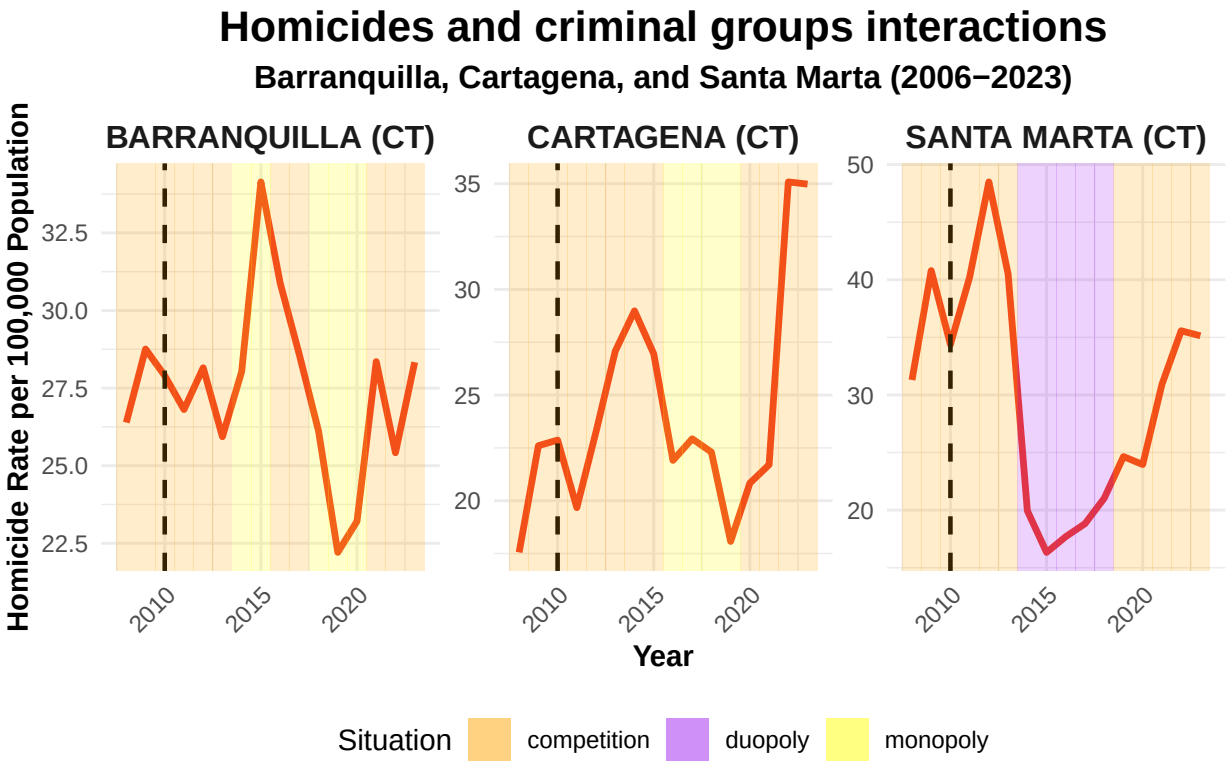
Now I will include the data on competition doing a left join

```
#Getting the dataset and only using relevant variables
competition <- read_excel("./raw_data/competition/competition_BSC.xlsx") |>
  select(municipio, year, situation, ocgs, variation, result)


#making year a character so it is congruent with our dataset
competition$year <- as.character(competition$year)

#merging the datasets
BCS_homicide_rates <- BCS_homicide_rates |>
  left_join(competition, by = c("municipio", "year"))
```

Now let´s produce a graph that can summarize a complex argument or, at least, provide some interesting intuition:

```
#Making year integer because then it is easier to manipulate
BCS_homicide_rates$year <- as.integer(BCS_homicide_rates$year)

# Faceted plot for every municipality
graph8_homicides_competition <-
  BCS_homicide_rates |>
  ggplot(aes(x = year, y = homic_rate, group = municipio)) +
  geom_line(size = 1.2, color = "#f03b20") +
  facet_wrap(~ municipio, scales = "free_y") +  # Separate plots for each municipality

# Adding vertical line for conditional crackdown in 2010
  geom_vline(xintercept = 2010, linetype = "dashed", color = "black", size = 0.8) +

# Coloring background for 'situation' periods
  geom_rect(data = subset(BCS_homicide_rates, situation == 'Competition'),
            aes(xmin = year - 0.5, xmax = year + 0.5, ymin = -Inf, ymax = Inf, fill = 'competition'),
            alpha = 0.2, inherit.aes = FALSE) +
  geom_rect(data = subset(BCS_homicide_rates, situation == 'Monopoly'),
            aes(xmin = year - 0.5, xmax = year + 0.5, ymin = -Inf, ymax = Inf, fill = 'monopoly'),
            alpha = 0.2, inherit.aes = FALSE) +
  geom_rect(data = subset(BCS_homicide_rates, situation == 'Duopoly'),
            aes(xmin = year - 0.5, xmax = year + 0.5, ymin = -Inf, ymax = Inf, fill = 'duopoly'),
            alpha = 0.2, inherit.aes = FALSE) +

# Setting colors for the background
  scale_fill_manual(values = c("competition" = "orange",
                               "monopoly" = "yellow",
```

```
                                "duopoly" = "purple"), name = "Situation") +

# Labels
   labs(
    title = "Homicides and criminal groups interactions",
    subtitle = "Barranquilla, Cartagena, and Santa Marta (2006-2023)",
    x = "Year",
    y = "Homicide Rate per 100,000 Population",
 caption = "Source: Policía Nacional de Colombia\nVertical dashed line represents conditional crackdown
 ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5, size = 16),
    plot.subtitle = element_text(face = "bold", hjust = 0.5, size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    plot.caption = element_text(size = 10, face = "italic"),
    strip.text = element_text(face = "bold", size = 12),
    legend.position = "bottom"
  )

print(graph8_homicides_competition)
```



Source: Policía Nacional de Colombia
Vertical dashed line represents conditional crackdown in 2010

As observed, the trend is not as expected: after the conditional crackdown, homicides in Barranquilla continued decreasing, but armed groups were still competing. In Santa Marta, even though there is a period

of increase, the homicides start decreasing radically in 2012. In Cartagena, on the contrary, the crackdown does not seem to produce any visible decrease in homicides, as they radically increased. Even though we cannot make any inference on this, at least with the dataset it was posible to evaluate our priors about how criminal groups interact during times of competition or monopoly.

Example 3: Homicides and Two Covariates

Homicides and Poverty

Now, a second example of the usefulness will be to answer what is the association between poverty and homicididal violence. For that purpose, and only as a toy example, I will use a dataset with an indicator that the Colombian government uses to measure poverty (unsatisfied basic needs, in Spanish: necesidades básicas insatisfechas). I have information of the value of the indicator for every municipality during 2021. So, let´s try to show some examples in graphs.

First, I will get the dataset. Now, this "municipio_code" has the same issue as the original Police code (the 4 digit codes miss the initial zero), so I will do that using the same code:

```
nbi_2021 <- read_excel("./raw_data/nbi/NBI_2021.xlsx") |>
  mutate(municipio_code = ifelse(nchar(municipio_code) == 4,
                                 paste0("0", municipio_code),  # Add leading zero
                                 municipio_code)) |>
  select(municipio_code, NBI)
```

Now, let´s filter our "municipality_year" data to get just 2021, and I will also don´t consider outliers (more than 100 homicide rate):

```
municipality_year_nbi <- municipality_year |>
  filter(year == 2021 & homic_rate < 100)
```

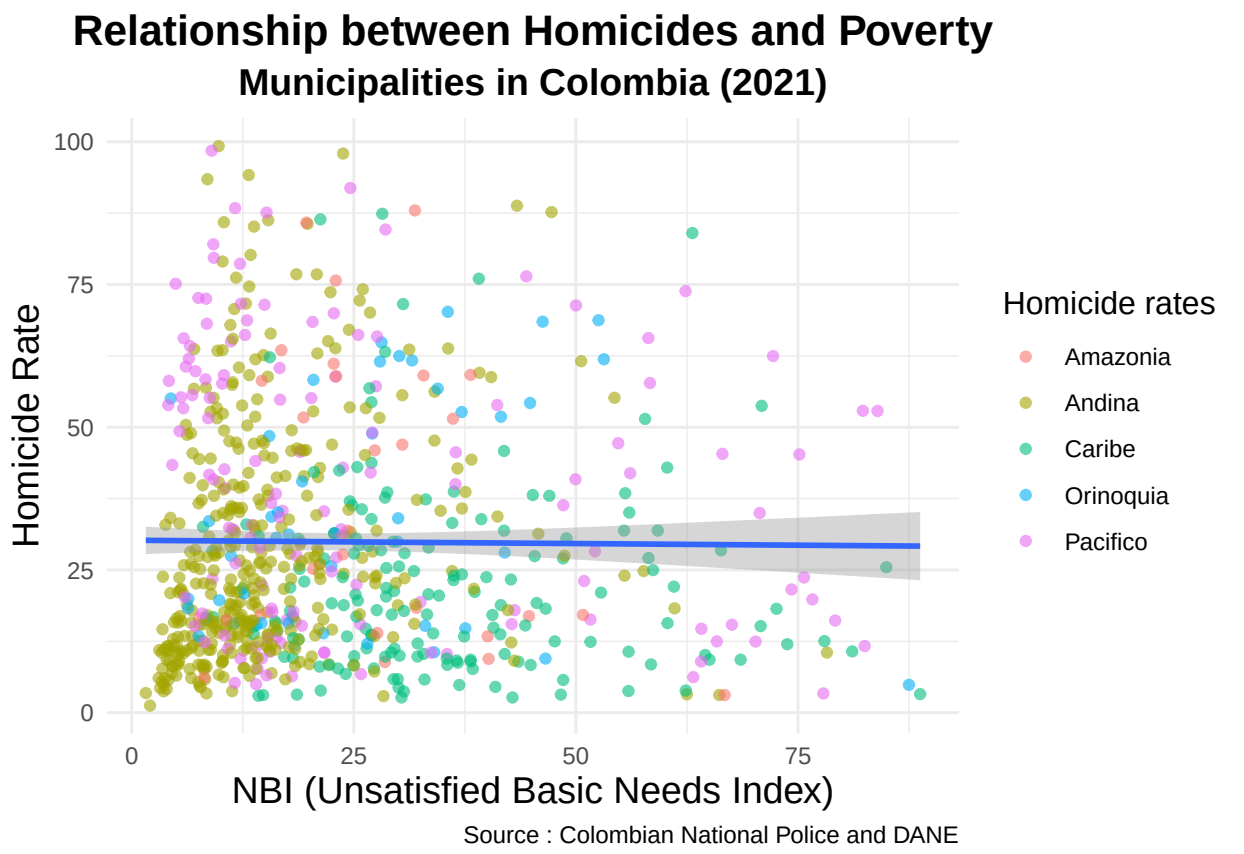
Now we can do a left_join with our municipality_year data

```
municipality_year_nbi <- municipality_year_nbi |>
  left_join(nbi_2021, by = "municipio_code")
```

Then, let´s show a scatter plot to see the association!

```
graph9_homicides_nbi <-
municipality_year_nbi |>
  ggplot(aes(x = NBI, y = homic_rate)) +
  geom_point(aes(color = region), alpha = 0.6) +   # Scatter points with color based on homic_rate
  geom_smooth(method = "lm") +
  labs(title = "Relationship between Homicides and Poverty",
       subtitle = "Municipalities in Colombia (2021)",
       x = "NBI (Unsatisfied Basic Needs Index)",
       y = "Homicide Rate",
       color = "Homicide rates",
       caption = "Source : Colombian National Police and DANE") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14, face = "bold"),
```

```
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    legend.title = element_text(size = 12)
  )
print(graph9_homicides_nbi)
```



Well, the results do not show any clear pattern between poverty and homicide rates. Of course, we would need to consider some covariates to say something meaningful. That exceeds my capacity right now, but still shows how useful it is to have homicide rates in a single dataset. In the third example, I will show another covariate (state capacity). However, none of these should be assume as a causal association.

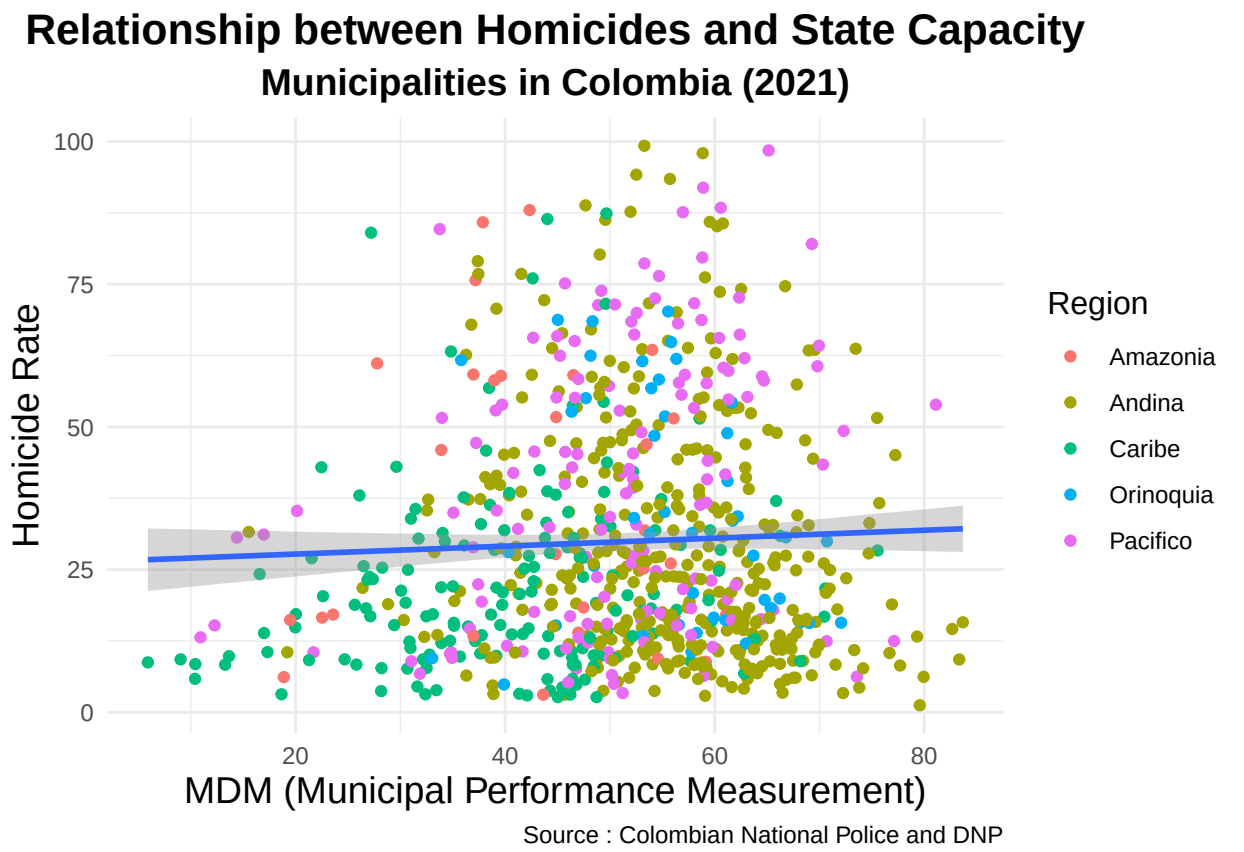
Homicides and State Capacity

Finally, a similar excercise we can do is to consider a different covariate: state capacity. The Colombian government has a year dataset called "Medición de Desempeño Municipal" (MDM); in English: "Municipal Performance Measurement". Another story that can be said is that when the state is weak, homicides tend to be higher as it cannot control them. Let´s evaluate that hypothesis with the new dataset for 2021 and to run a regression including also our NBI data.

```
#Importing the data
mdm_2021 <- read_excel("./raw_data/nbi/MDM_2021.xlsx") |>
  select(municipio_code, MDM)

# Merging datasets
```

```r
municipality_year_nbi <- municipality_year_nbi |>
  left_join(mdm_2021, by = "municipio_code")
```

```r
graph10_homicides_stcap <-
  municipality_year_nbi |>
  ggplot(aes(x = MDM, y = homic_rate)) +
  geom_point(aes(color = region)) +   # Scatter points with color based on homic_rate
  geom_smooth(method = "lm") +
  labs(title = "Relationship between Homicides and State Capacity",
       subtitle = "Municipalities in Colombia (2021)",
       x = "MDM (Municipal Performance Measurement)",
       y = "Homicide Rate",
       caption = "Source : Colombian National Police and DNP",
       color = "Region") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    legend.title = element_text(size = 12)
  )

print(graph10_homicides_stcap)
```



```r
regression <- lm(homic_rate ~ NBI + MDM, municipality_year_nbi)

regression_summary <- summary(regression)

print(regression_summary)
```

```
##
## Call:
## lm(formula = homic_rate ~ NBI + MDM, data = municipality_year_nbi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.555 -16.682  -6.125  12.525  69.357
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.67474    3.98236   6.447 2.02e-10 ***
## NBI          0.01293    0.04884   0.265    0.791
## MDM          0.07658    0.06475   1.183    0.237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.33 on 767 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.001892,   Adjusted R-squared:  -0.0007103
## F-statistic: 0.7271 on 2 and 767 DF,  p-value: 0.4836
```

Well, the relationship, if anything, does not seem to be particularly strong. There is a significant variation in the dependent variable (homicide rates) as well as in the independent variable (MDM). It is really hard to say something meaningful just with two variables, as naturally homicidal violence is one of those phenomena that are likely explained by a huge range of other covariates. However, even when trying to find some association through a linear regression, the results do not show any significant effect of any of the two variables.

However, this small exercise is useful to show how the new big dataset can provide with interesting opportunities for researchers to explain violence in Colombia. Some very interesting opportunities are related to areas with civil war/non-civil war, criminal groups/non-criminal groups, the size and kind of armed group, and a lot of other possibilities.