Analysis of the telephone line for gender-based violence cases in Argentina 2020-2022.

RNCP

Angela SALGADO BELTRÁN

DATA ANALYTICS BOOTCAMP

December 2022

Table of Contents

INTRODUCTION 2

1. DATA AND DATA SOURCES 3

2. EXPLORATORY DATA ANALYSIS 3

3. EXPLORATORY DATA ANALYSIS 5

4. DATABASE 10

5. CONCLUSION 17

6. BIBLIOGRAPHY 18

INTRODUCTION

1. *Context*. In the last 10 years, different governments in Latin America have created telephone lines to attend to gender-based violence[1] and violence against women cases. For example, Mexico City, Bogota, and Montevideo now have lines where people who suffer these type violence can be assisted and directed to other state institutions, such as the police or the ministries of social protection. In the case of Argentina, since 2013 a national telephone line (Line 144) aims to "provide care, support and advice in situations of gender-based violence".[2]

   These telephone lines became more important during the COVID-19 pandemic. Most of them, reported significant increases in the number of calls, especially in 2020 and 2021, and have become a valuable tool to face the negative consequences of measures such as lockdowns, where some victims must live together with their aggressors all the time.

2. *Business Use Case Objective*. This final project will take as a case study Argentina's telephone line (Line 144), focusing on public data for the period January 2020 - July 2022. The objective is to build an unsupervised machine learning model, that will allow to understand their behavior to generate an optimal classification of persons accessing the helpline services.

3. *Plan*. This document is divided into 5 parts. The first part will explain the collection and sources of the data. Then we will cover the cleaning process that was performed before data manipulation could take place. In the third part, we will present the results of the data exploration.

   In order to develop this project, it was necessary to create a database; the explanation of this process as well as the reasons for choosing a SQL database will be the subject of a fourth part of this document. We will then display the entity-relationship diagram of our database, as well as the procedures for its creation and feeding. Finally, we will conclude.

[1] Gender-based violence have a more broad definition than violence against women. "Gender-based violence refers to any type of harm that is perpetrated against a person or group of people because of their factual or perceived sex, gender, sexual orientation and/or gender identity." COUNCIL OF EUROPE. *What is gender-based violence?* En ligne : https://www.coe.int/en/web/gender-matters/what-is-gender-based-violence [consulté le].

[2] GOBIERNO DE ARGENTINA. *Línea 144*. 2022.

1. DATA AND DATA SOURCES

4. *Data sources.* Emergency telephone calls are a matter of great care in terms of protecting the identity of the victim who reports violence. This is why many of the data is not public. This is the case in Uruguay, Ecuador and Chile for example. However, I found three databases from the websites of national statistics institutions and/or ministries of equality or women's rights: the emergency lines in Mexico City, Bogotá (*Purple line*) and the national line in Argentina (*Line 144*).

5. *Data collection.* For the data collection, I downloaded the public files of these three telephone lines. These three datasets were very dissimilar from each other. Unfortunately it was impossible to unify them because the variables available were not the same in each of them. The public data for the Argentina line had variables that were more attractive to analyse. Thus, this dataset covers six types of violence as well as six types of exercise modalities of this violence. For these reasons, I chose the public data for the Argentina line, which is available at the following link, from January 2022 to the first semester of 2022: https://www.argentina.gob.ar/generos/linea-144/informacion-estadistica

2. EXPLORATORY DATA ANALYSIS

6. *Data cleaning tools.* Data cleaning was performed in Python, using the Pandas, Plistib, Numpy, and Math libraries.

7. *Initial data conditions.* The starting point was three csv files, one for 2020 (29.707 rows), one for 2021 (25.302 rows), and one for the first half of 2022 (12.399 rows). All three contained the same variables, so their unification into a single file did not cause significant problems. Each line of the tables represents a call made to Line 144.

   Columns of the datasets :
   - date. (date on which the call was made)
   - province of residence of the person in a situation of violence. (place of residence of the person reporting violence)
   - gender of the person in situation of violence. (gender with which the caller self-identifies)

- country of birth person in violent situation.
- age of the person in a situation of violence.
- link with the aggressor
- gender of the aggressor. (gender with which the caller identifies the aggressor)

The violent type columns contain only "yes" values for the presence or "no" values for the absence of the type of violence

- type of physical violence.
- type of psychological violence
- type of sexual violence
- type of economic and patrimonial violence.
- type of symbolic violence.
- type of domestic violence

The modalities columns contain only "yes" values for the presence or "no" values for the absence of the type of violence. One modality refers to the methods or places where this violence is exercised. For example, the type of sexual violence can be perpetrated by an institutional authority, when the aggressor is a public servant and the violence is committed within the framework of the exercise of his or her functions.

- modality of institutional violence.
- modality of violence at work
- modality of violence against reproductive freedom
- modality of obstetric violence
- modality of media violence
- other modalities of violence.

8. *Unification of datasets*. The first step was to import the three CSV files using Pandas in Python, and concatenate them into a single dataframe that we called gvb_calla_data. The data cleaning was performed on this dataframe. The concatenated dataset had **67405 observations (calls)** and 19 columns in total.

9. *Missing values*. The next step was to check the missing values in each column. I found missing values in the following columns: province of residence of the person in a situation of violence 974 rows; gender of the person in situation of violence 1274 rows; country of birth person in violent situation **22772 rows;** age of the person in a situation of violence **12991 rows ;** link with the aggressor **2345 rows; and** gender of the aggressor **6630 rows.**

10. *Drop missing values*. The criteria for removing some elements from the dataset was the percentage that these values represented in the total variable, and also whether or not they could be replaced. The 'country of birth of the person in situation of violence' column was deleted because it had more than 30% missing values and could not be replaced. The rows with missing values were deleted from the following columns: 'gender of the aggressor', 'link with the aggressor', and 'gender of the person in situation of violence'.

11. *Replace missing values*. To avoid losing more data, it was decided that for the missing values in the victim's age column, the average age according to the person's region would be inserted. For example, for each missing age value for a person living in Buenos Aires, the average age value for Buenos Aires was assigned. Thus, for example, for each missing age value for a person living in Buenos Aires, the average age value for the entire province of Buenos Aires was assigned.

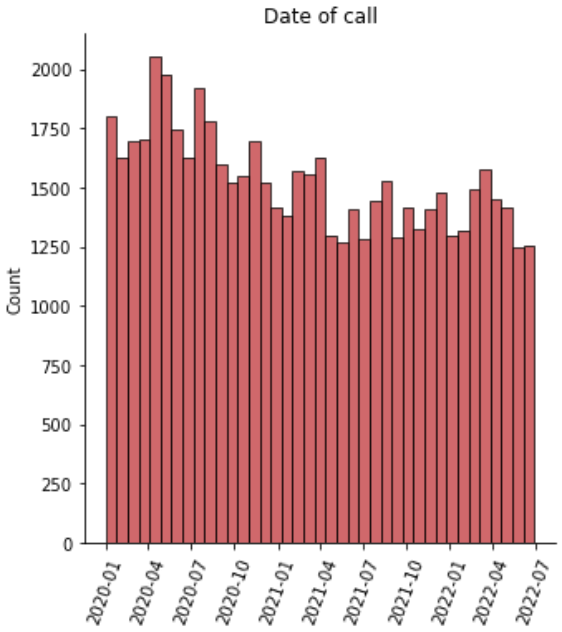
12. *Inconsistency*. Since 18 of the 19 columns contain string data, all columns were checked for inconsistencies in the typing. Indeed, it was necessary to replace several texts in order to have unified data. There were duplicate province names due to differences in capitalisation and syntax. The same problems were found in the columns of types of violence and modalities.

13. *Results*. The final result was a dataframe with 59,360 observations and 18 columns.

3. EXPLORATORY DATA ANALYSIS

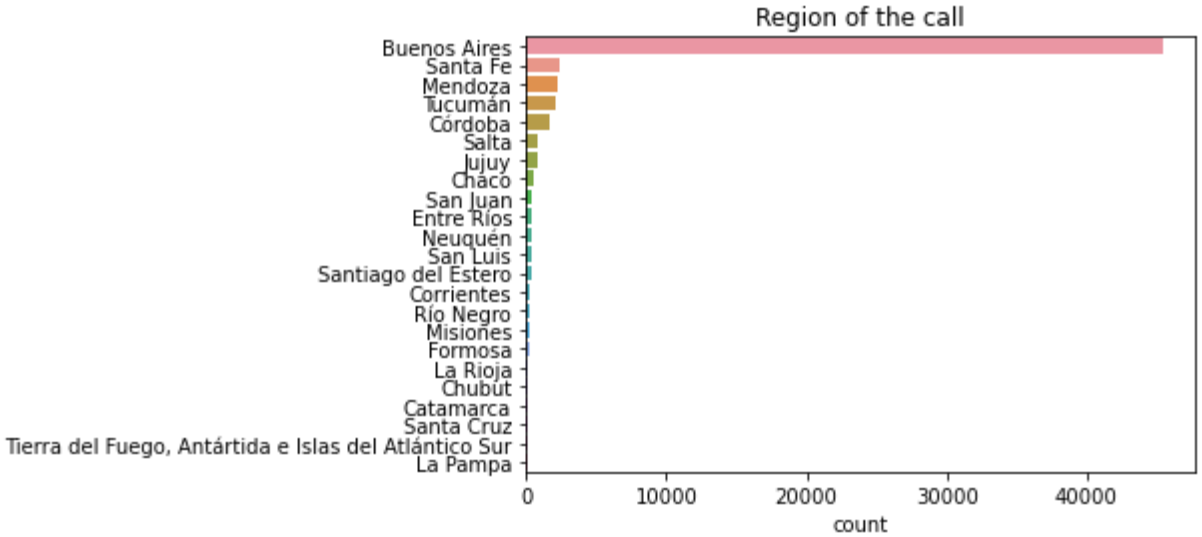
14. *Exploratory Data Analysis*. After data cleaning, the exploratory analysis of the data was carried out. Since there was only one numerical variable (age) the analysis tried to focus on counting the values of different variables.

15. *Date of calls*. Graph N1 shows a general downward trend in calls, especially between May 2020 and July 2022. The peak in the number of calls is located in March and April 2020, which coincides with the start of the Covid 19 pandemic (start of national containment on



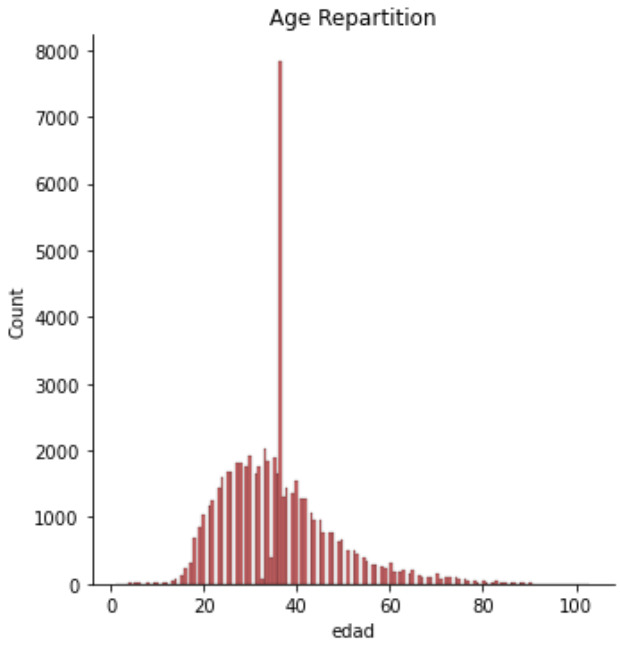
N1

16. *Geographical distribution*. Calls are concentrated in the province of Buenos Aires, where the city of Buenos Aires is located. This province is home to 35% of the country's population. This may also be linked to issues of access to information and telecommunications, and to the visibility campaigns of the 144 line.



N2

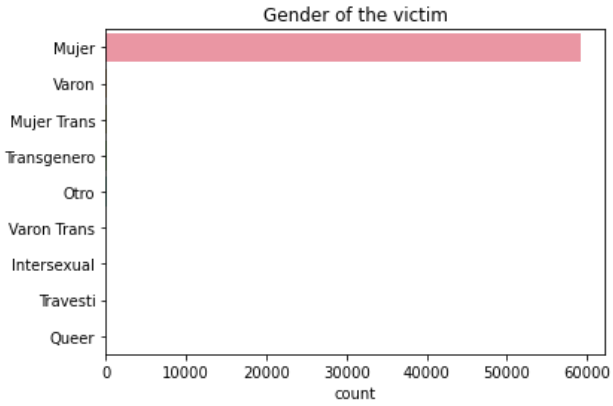
17. *Age of the victim*. The average age of the persons who called to report an act of violence is 35.85 years old. The age of the victims has a positive skew, with few records for persons under 15, and over 60 years of age. This distribution can be seen in graph N3.



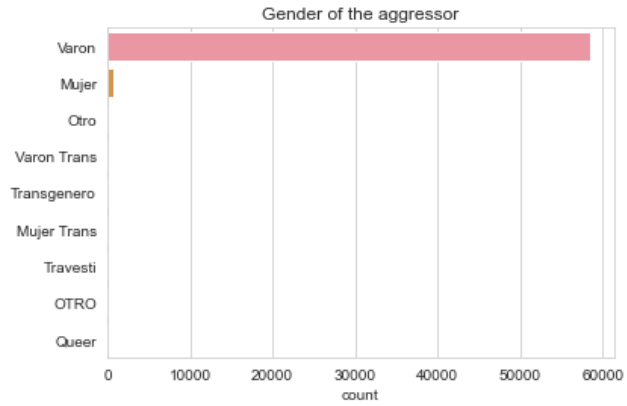
N3

As a consequence of the decision to replace the missing values in the age column by the mean of the province of each call, and the concentration of calls in the Buenos Aires region, we have a large number of observations whose age is close to the mean of Buenos Aires, which is 36 years.

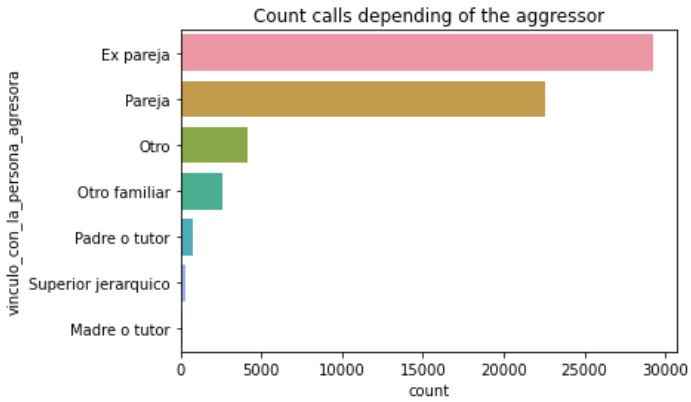
18. *Gender*. What we can see in graphs N4 and N5 is the opposite distribution in terms of the gender of the person reporting violence and the perpetrator. Despite the fact that the hotline is designed to receive reports of gender-based violence, which is broader than violence against women, the vast majority of reports correspond to women who have been violent by men. To this we can add graph N6 which shows the number of cases in relation to the relationship of the victim with the aggressor. Thus, in 52.8% of the calls, the aggressor is the victim's ex-partner, and in 40.46% it is the victim's current partner.



N4

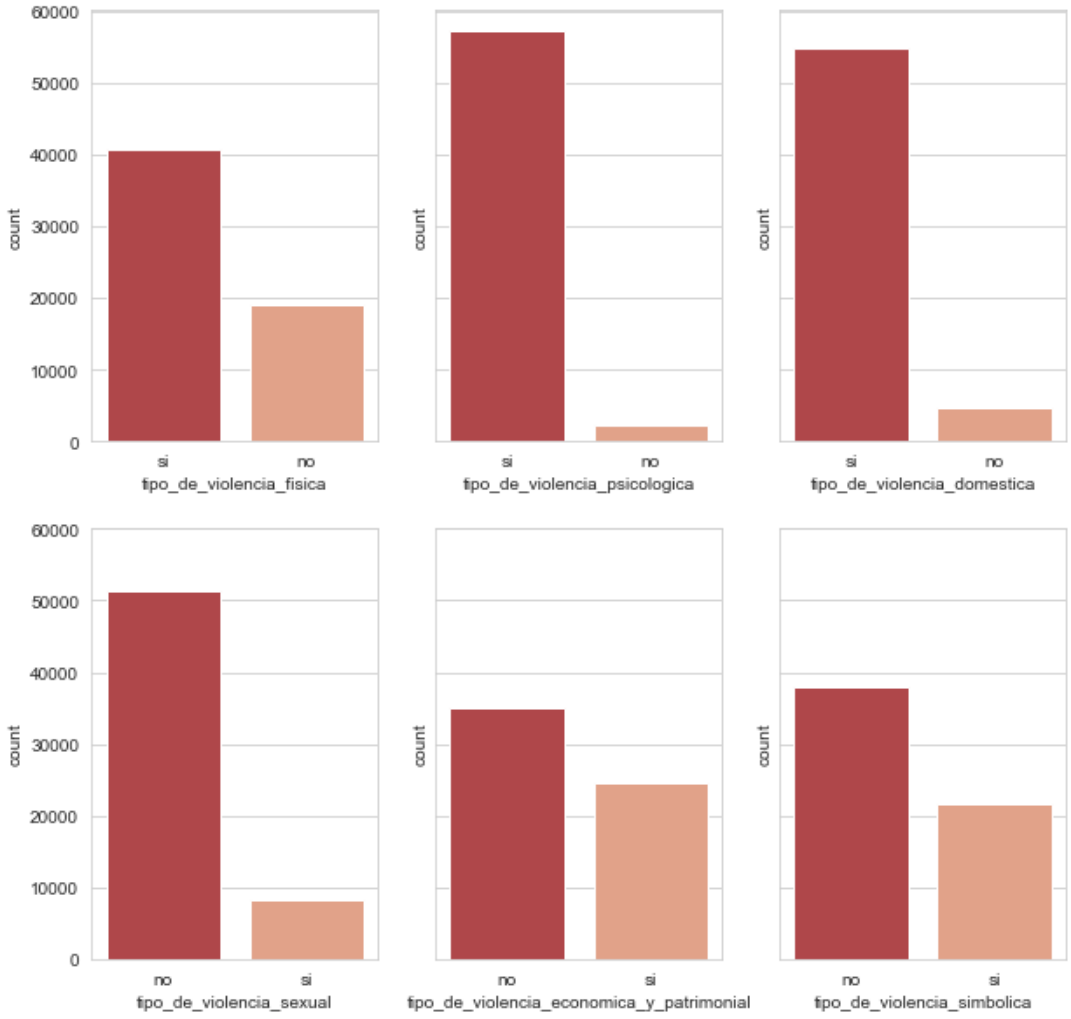


N5

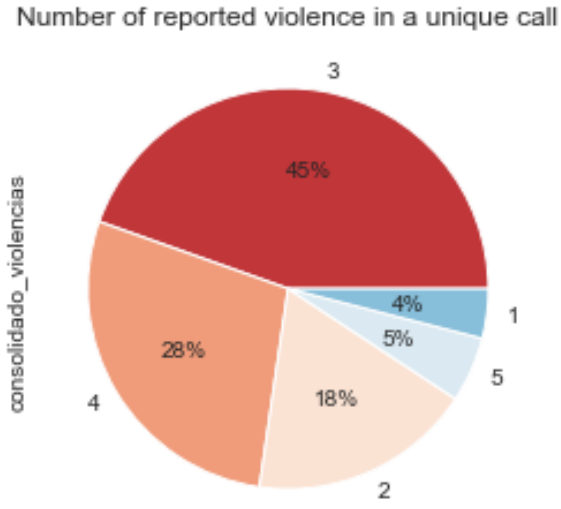


N5

19. *Occurrence of violence.* The 6 columns corresponding to the type of violence reported by the caller are binary variables: "yes" or "no". As mentioned above, a caller can call to report several types of violence. Two graphs are presented below, the first one N6 showing the proportion of yes and no for each violence, and the second one N7 prints the percentage of calls according to the number of accumulated violence.


60000
50000
40000
30000
20000
10000
0
count
si
no
tipo_de_violencia_fisica
tipo_de_violencia_psicologica
tipo_de_violencia_domestica
tipo_de_violencia_sexual
tipo_de_violencia_economica_y_patrimonial
tipo_de_violencia_simbolica

N6


Number of reported violence in a unique call
consolidado_violencias
3
45%
1
4%
5
5%
2
18%
4
28%

N7

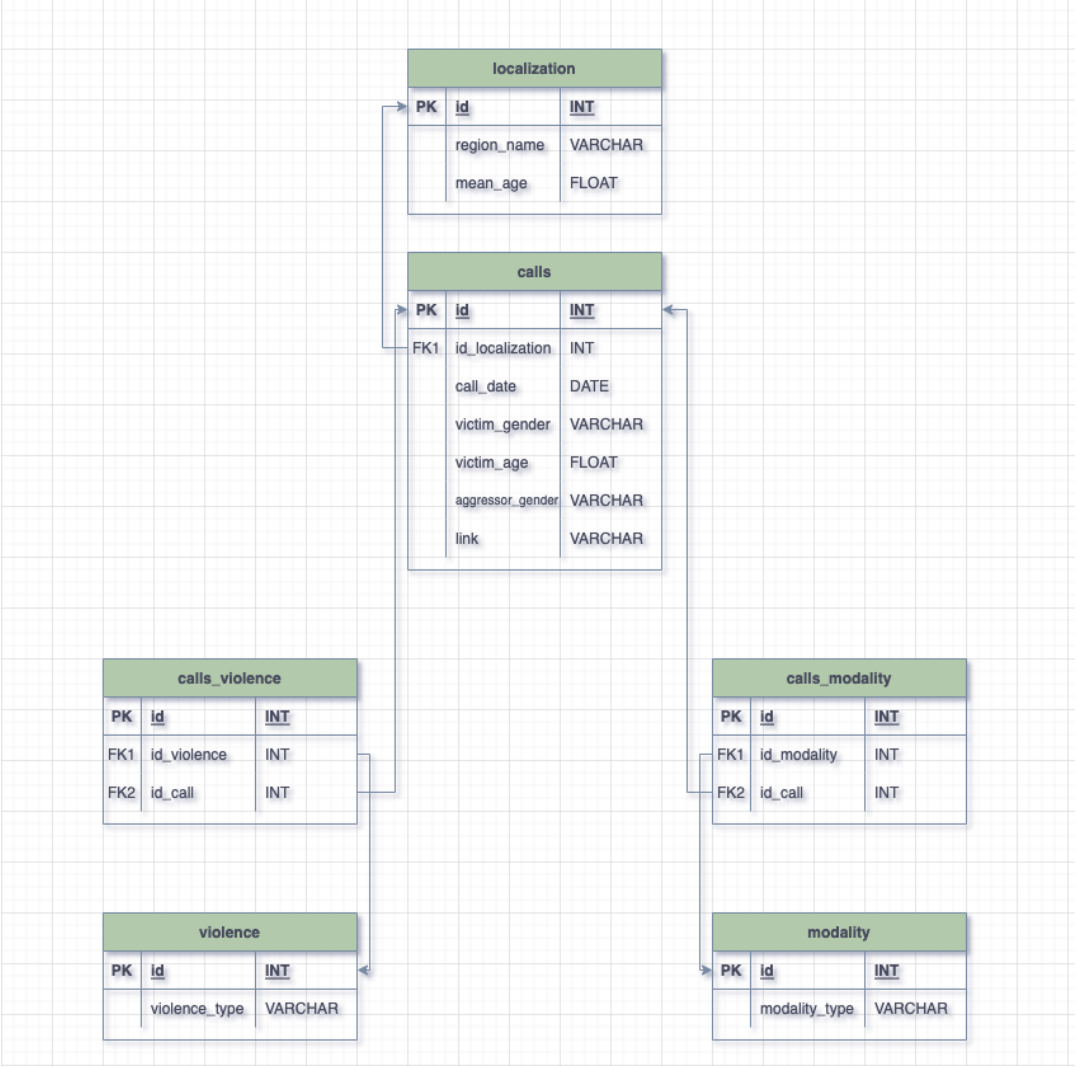
The most reported violence is psychological violence, followed by domestic and physical violence. Sexual violence is the least reported violence to this help line. Also, in 45% of the calls, 3 types of violence occurred in the same events, and in 33% of the cases up to 5 types of violence were reported.

4. DATABASE

20. *Choice of database type.* The choice of the type of database to be used took into account the fact that we initially had tabular data. Following the definition of an SQL database as "a collection of highly structured tables, wherein each row reflects a data entity, and every column defines a specific information field. Relational databases are built using the structured query language (SQL) to create, store, update, and retrieve data,"[3] it was decided to use this type of database because it was the most convenient. MySql was used as the database management system for its construction and implementation.

21. *Entity-relationship diagram.* For the structuring of the database it was taken into account that, after analysing our table, we have two types of relationships, one from "one to many", and others from "many to many". For one call there can be several types of violence occurring at the same time as well as different types of modalities, and on the other hand, one modality or one type of violence can have several related calls. These relationships in our table led to the following database structure:

[3] SOLARWINDS. *What is a SQL Database?* En ligne : https://www.solarwinds.com/resources/it-glossary/sql-database [consulté le].



22. *Database creation.* The creation of the database was done from MySql script. My database is called "gender_base_violence". The first step was to load a table we call "staging" which is the table resulting from the data cleaning. For the creation of the database the primary keys as well as the foreign keys were established. The database is interlinked from the id of the "staging" table, which was kept throughout the data import, in order to be able to separate the variables without losing the unit of analysis, i.e. each call.

```
CREATE DATABASE IF NOT EXISTS gender_base_violence;

USE gender_base_violence;

CREATE TABLE
```

```
  localization(
    id INT AUTO_INCREMENT,
    region_name VARCHAR(250) UNIQUE,
    mean_age FLOAT,
    PRIMARY KEY (id)
  );

CREATE TABLE
  IF NOT EXISTS violence(
    id INT AUTO_INCREMENT,
    violence_type VARCHAR(250) UNIQUE,
    PRIMARY KEY (id)
  );

CREATE TABLE
  IF NOT EXISTS modality(
    id INT AUTO_INCREMENT,
    modality_type VARCHAR(250) UNIQUE,
    PRIMARY KEY (id)
  );

CREATE TABLE
  IF NOT EXISTS calls(
    id INT AUTO_INCREMENT,
    call_date DATE NULL,
    id_localization INT null,
    victim_gender VARCHAR(250) null,
    victim_age FLOAT null,
    aggressor_gender VARCHAR(250) null,
    link VARCHAR(250)  NULL,
    PRIMARY KEY (id),
    FOREIGN KEY(id_localization) REFERENCES localization(id)
  );

CREATE TABLE IF NOT EXISTS calls_violence (
 id INT AUTO_INCREMENT,
 call_id INT,
```

```
  violence_id INT,
  PRIMARY KEY (id),
  FOREIGN KEY (call_id) REFERENCES calls(id),
  FOREIGN KEY (violence_id) REFERENCES violence(id)
);

CREATE TABLE IF NOT EXISTS calls_modality (
  id INT AUTO_INCREMENT,
  call_id INT,
  modality_id INT,
  PRIMARY KEY (id),
  FOREIGN KEY (call_id) REFERENCES calls(id),
  FOREIGN KEY (modality_id) REFERENCES modality(id));
```

23. *Database feed*. The violence and modalities tables were fed manually due to the small number of elements in each table. Then, from the "staging" table, the "calls" table was fed, storing the ID of each call. Finally, the "calls-violence" and "call-modality" tables were fed.

```
-- VIOLENCE TYPE
INSERT
    IGNORE INTO violence(violence_type)
VALUES ('physical'), ('psychological'), ('sexual'), ('economical'), ('symbolical'), ('domestical');

-- VIOLENCE MODALITY
INSERT
    IGNORE INTO modality(modality_type)
VALUES ('institutional'), ('labour'), ('media'), (
        'against reproductive freedom'
    ), ('obstetric'), ('others');

-- CALLS
INSERT INTO
    calls (
        id,
        call_date,
        id_localization,
        victim_gender,
```

```
    victim_age,
    aggressor_gender,
    link
  ) (
    SELECT
      staging.id,
      STR_TO_DATE(fecha, "%Y-%m-%d"),
      localization.id,
      genero_persONa_en_situaciON_de_violencia,
      edad,
      genero_de_la_persONa_agresora,
      vinculo_cON_la_persONa_agresora
    FROM staging
      JOIN localization ON localization.region_name = staging.prov_residencia_persONa_en_situaciON_violencia
    WHERE NOT EXISTS (
        SELECT c.id
        FROM
          calls c,
          staging stg
        WHERE
          c.id = stg.id
        LIMIT 1
      )
  );

-- CALLS_VIOLENCE
INSERT INTO
  calls_violence(call_id, violence_id) (
    SELECT
      staging.id AS call_id,
      violence.id AS violence_id
    FROM
      staging,
      violence
    WHERE ( (
          staging.tipo_de_violencia_fisica = 'si'
          AND violence.violence_type = 'physical'
```

```
            )
            OR (
                staging.tipo_de_violencia_psicologica = 'si'
                AND violence.violence_type = 'psychological'
            )
            OR (
                staging.tipo_de_violencia_sexual = 'si'
                AND violence.violence_type = 'sexual'
            )
            OR (
                staging.tipo_de_violencia_ecONomica_y_patrimONial = 'si'
                AND violence.violence_type = 'ecONomical'
            )
            OR (
                staging.tipo_de_violencia_simbolica = 'si'
                AND violence.violence_type = 'symbolical'
            )
            OR (
                staging.tipo_de_violencia_domestica = 'si'
                AND violence.violence_type = 'domestical'
            )
        )
        AND NOT EXISTS (
            SELECT *
            FROM
                calls_violence cv2,
                staging stg2,
                violence v2
            WHERE
                cv2.call_id = stg2.id
                AND cv2.violence_id = v2.id
            LIMIT 1
        )
    );

-- CALLS_MODALITY
INSERT INTO
```

```
calls_modality(call_id, modality_id) (
    SELECT
        staging.id AS call_id,
        modality.id AS modality_id
    FROM
        staging,
        modality
    WHERE ( (
                staging.modalidad_de_violencia_instituciONal = 'si'
                AND modality.modality_type = 'institutiONal'
            )
            OR (
                staging.modalidad_de_violencia_laboral = 'si'
                AND modality.modality_type = 'labour'
            )
            OR (
                staging.modalidad_violencia_cONtra_libertad_reproductiva = 'si'
                AND modality.modality_type = 'against reproductive freedom'
            )
            OR (
                staging.modalidad_de_violencia_obstetrica = 'si'
                AND modality.modality_type = 'obstetric'
            )
            OR (
                staging.modalidad_de_violencia_mediatica = 'si'
                AND modality.modality_type = 'media'
            )
            OR (
                staging.modalidad_de_violencia_otrAS = 'si'
                AND modality.modality_type = 'others'
            )
        )
        AND NOT EXISTS (
            SELECT *
            FROM
                calls_modality cm,
                staging stg,
```

```
        modality m
    WHERE
        cm.call_id = stg.id
        AND cm.modality_id = m.id
    LIMIT 1
));
```

5. CONCLUSION

Despite the difficulties that exist in accessing data from telephone lines that deal with cases of gender-based violence due to their private nature, finding and collecting the data for the case of the telephone line in Argentina was not a particularly complicated exercise.

The data has some missing values, probably generated as a result of human error of not digitising them at the time the call is received. Nevertheless, the database has consistent information for most of the variables. Consulting the reports that the Argentinean government generates, it is also evident that there is a lot of relevant data that is not shared publicly, so the analysis may be somewhat restricted, especially in terms of providing possible explanations for the occurrence of certain types of violence.

The exploration of the data showed that the calls are concentrated in Buenos Aires, that most of the calls are made by women between 30 and 40 years of age, and that the aggressors are mostly men with whom they had or maintain a relationship as partners. We also concluded that in more than 70% of the cases, three or more types of violence are reported at the same time, which shows the interrelationship between them. An important point to highlight is that the most reported violence is not necessarily physical violence as one might think at first. This is interesting from the point of view of the prevention of other types of violence, where physical integrity may be compromised, such as feminicide. Early reporting of psychological violence can enable early prevention policies to be put in place.

In technical terms, My Sql is really an ideal tool for working with tabular data. The decomposition of the database made it possible to preserve all the relationships of the observations while at the same time organising the different variables linked to the telephone calls in a much clearer way for any user.

6. BIBLIOGRAPHY

- COUNCIL OF EUROPE. *What is gender-based violence?* En ligne : https://www.coe.int/en/web/gender-matters/what-is-gender-based-violence [consulté le].

- GOBIERNO DE ARGENTINA. *Línea 144*. 2022.

- SOLARWINDS. *What is a SQL Database?* En ligne : https://www.solarwinds.com/resources/it-glossary/sql-database [consulté le].