LUCIA CATALAN GRIS

LUCAS GARCÍA DE VIEDMA

Data Integration, Bias and Fairness

How was the evolution of the COVID-19 virus in “Comunidad de Castilla y León” in relation to the vaccination of the population?

INDEX

[1 Specific scenario 1](#_Toc101710958)

[2 Datasets selected 1](#_Toc101710959)

[2.1 Data source 1 1](#_Toc101710960)

[2.2 Data source 2 1](#_Toc101710961)

[3 Conflicts identified 1](#_Toc101710962)

[3.1 Data source 1 conflicts 1](#_Toc101710963)

[3.1.1 Conflict A 2](#_Toc101710964)

[3.1.1 Conflict B 2](#_Toc101710965)

[3.2 Data source 2 conflicts 3](#_Toc101710966)

[3.2.1 Conflict A 3](#_Toc101710967)

[3.2.2 Conflict B 4](#_Toc101710968)

[3.3 Merging conflicts 4](#_Toc101710969)

[3.3.1 Conflict 1 4](#_Toc101710970)

[3.3.2 Conflict 2 5](#_Toc101710971)

[4 Reports from tools for bias and fairness 5](#_Toc101710972)

# Specific scenario

Scenario: How was the evolution of the COVID-19 virus in “Comunidad de Castilla y León” in relation to the vaccination of the population?

# Datasets selected

We selected two datasets to answer the query proposed, one about the epidemiological situation of the covid virus in Castilla y León and the other about the number of vaccinated people by date.

## Data source 1

The first dataset selected was “situacion-epidemiologica-coronavirus-en-castilla-y-leon.xls” from [datos.gob.es](https://datos.gob.es/es/catalogo/a07002862-situacion-epidemiologica-coronavirus-covid-19-en-castilla-y-leon-por-provincias1). It contains data related to confirmed cases, discharges, and deceases in hospitals, due to COVID-19, in the A.C. of Castilla y León. The data regarding the number of positive people is provided by the “*Red de Vigilancia Epidemiológica”* at 20:00 hours of the day before. The data is in Spanish.

## Data source 2

The second dataset was “personas-vacunadas-covid.xls” from [datos.gob.es](https://datos.gob.es/es/catalogo/a07002862-personas-vacunadas-covid-19). It contains information about the vaccine doses that have been administered in the A.C. of Castilla y León.

# Conflicts identified

While the obtention of an integrated dataset is not the goal, we will show the different steps we have followed to obtain it to help understanding the difficulties that we encountered in the process.

## Data source 1 conflicts



### 3.1.1 Conflict A

Issue

The poorly assigned names to the different variables, together with no further descriptions of them in the source, lead to confusion (Do the variables refer to daily data? Or is it rather accumulated data up until that day?). Spanish names also make it harder for foreigners to work with the data.

Solution

Analyzing the data, we can determine the true meaning of each variable and assign a precise name, in English, that represents it properly.

1. Fecha -> Date: date of the record.
2. Provincial -> Province: province in Castilla y León.
3. Casos\_confirmados -> Total\_positives: accumulated number of confirmed cases of covid.
4. Nuevos\_positivos -> New\_positives: number of new positive cases.
5. Altas -> Total\_discharges: accumulated number of discharges.
6. Fallecimientos -> Total\_deceases: accumulated number of deceases due to covid.
7. Codigo\_ine -> Ine\_code: ine code of the municipality.
8. Posicion -> Position: position.

The negative values in the column “New\_positives” correspond to patients whose result was initially positive but was discarded in the confirmation test.

### Conflict B

Issue

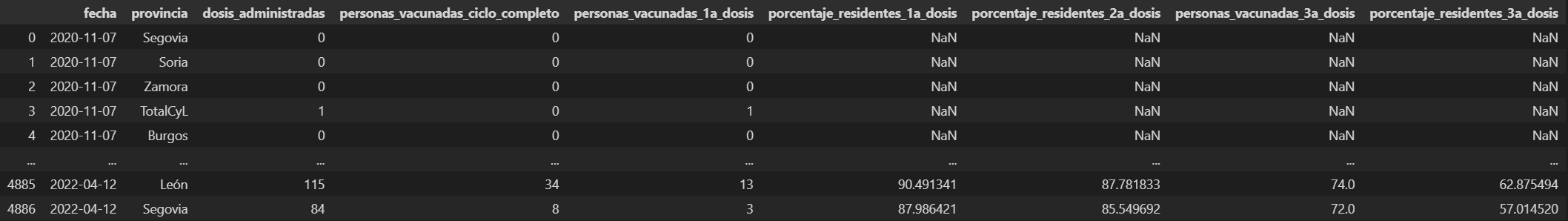
Only for a single variable the daily new values have been registered, while the accumulated values have been saved for all of them. For our scenario we would need both the daily and accumulated registers.

Solution

With the data from the accumulated values, we obtain the daily new data for the remaining variables. We obtain two new columns:

1. New\_deceases: number of new deceases due to covid.
2. New\_discharges: number of new discharges.

## Data source 2 conflicts



### Conflict A

Issue

Some variables have been registered since the creation of the dataset and others have been included later, deriving in confusing names and numerous missing values. Spanish names also make it harder for foreigners to work with the data.

Solution

Analyzing the data, we can determine the true meaning of each variable and assign a precise name, in English, that represents it properly. Additionally, we can determine the cause of the missing values and determine how to deal with them.

1. Fecha -> Date: date of the record.
2. Provincial -> Province: province in Castilla y León.
3. Dosis\_administradas-> new\_doses\_administered: new doses administered
4. Personas\_vacunadas\_ciclo\_completo-> New\_persons\_two\_doses: number of new persons with at least two doses or the equivalent protection (single dose of Jansen or single dose with previous covid infection)
5. Personas\_vacunadas\_1a\_dosis-> New\_persons\_one\_dose: number of persons with at least one dose
6. Porcentaje\_residentes\_1a\_dosis-> Percentage\_persons\_one\_dose: percentage of the population with at least one dose.
7. Porcentaje\_residentes\_2a\_dosis-> Percentage\_persons\_two\_doses: percentage of the population with at least two doses or the equivalent protection (single dose of Jansen or single dose with previous covid infection) .
8. Personas\_vacunadas\_3a\_dosis-> New\_persons\_three\_dose: number of persons with three doses or the equivalent protection (one dose plus the previous cases stated in the two doses situations).
9. Porcentaje\_residentes\_3a\_dosis-> Percentage\_persons\_three\_doses: percentage of persons with three doses or the equivalent protection (one dose plus the previous cases stated in the two doses situations).

The missings on the ‘percentage’ columns' were caused because these values were not measured since the beginning. Nonetheless, their value can be obtained from the original variables. However, we can not do it yet because we will need to obtain the population of each province and the accumulated values of each day for the columns ‘New\_persons\_one\_dose’, ‘New\_persons\_two\_doses’ and ‘New\_persons\_three\_doses’.

On the other hand, for the missings in ‘New\_persons\_three\_doses’ we can just input zeros, since third doses were not given until the summer of 2021.

### Conflict B

Issue

Similar to the previous dataset, we need both the daily and accumulated values for each variable. Thus, in this case we will need to obtain the accumulated number of doses of each type from the data we already have.

Solution

As previously said, we obtain the accumulated values of the ‘New\_persons\_one\_dose’, ‘New\_persons\_two\_doses’ and ‘New\_persons\_three\_doses’ for each day.

Additionally, we solve conflict A using the new accumulated values to determine the percentages that were missing in the ‘percentage’ columns.

## Merging conflicts

When trying to merge the two datasets we encounter two new conflicts.

### Conflict 1

Issue

The vaccination dataset includes a value (Total\_CyL) in the ‘Provinces’ column that the epidemiology dataset doesn’t have. This generates a conflict when trying to merge the data by date and provinces.

Solution

Generate the Total\_CyL (total values from Castilla y Leon) values for the epidemiology dataset, by adding the values from all provinces for every date of the dataset.

### Conflict 2

Issue

Secondly, the temporal span of both datasets differs, as the epidemiology dataset has data from the 13th of March 2020 to the 19th of May 2022, while the vaccination dataset has data from the 7th of November 2020 to the 12th of May 2022.

Solution

We want to observe the impact the vaccination had on the covid pandemic in Castilla y Leon, which means that we have to consider all the timeline from the epidemiology dataset. Thus, once we merge the two datasets, we will need to input zeros in all the missings from the variables of the vaccination dataset that go from the 13th of March 2020, when the epidemiological register began, to the 7th of November 2020, when the first vaccine dose was administrated (because there is no data about the vaccination before).

However, the missing values from the 12th of May 2022 to the 19th of May 2022 we will force us to delete all the data regarding that interval, as we don’t have the necessary information to obtain metrics about that period of time.

# Reports from tools for bias and fairness