Abstract

In this research an exhaustive study is conducted over the Covid-19 database, taking into consideration different aspects of our society in which the pandemics had a significant impact. We may present some introductory studies, with a brief interpretation of the data with some visualizations to illustrate our explanation. Following up we may introduce the groupings we have gathered. To end up with this introductory presentation we may include our napkin design.

1 Exploratory Data Analysis

1.1 The data

Let us start by exploring and describing the data, trying to find relationships and interesting attributes. At first glance, we notice that our data consists of temporal series classified by different geographical regions, from country level to city. The main part of the dataset is the COVID-19 epidemiology data, i.e. new cases, deaths, hospitalizations, vaccination, etc. We can also find other categories, such as economics, demographics, health, sex, gender, weather and so on. Those will allow us to interpret and relate the data and extract conclusions. Foremost, we are going to study the main data, that is, new cases, vaccines and hospitalizations.

1.1.1 New cases

Let us explore the data corresponding to new cases.

- We have the real new cases and new deaths of COVID-19 per day, shown for all time, last week, last 14 days or last month. We can also visualize the data depending on the geographic location of the cases, grouped by countries and regions. In addition, we have the punctual new cases per day, and the seventh day average data, as we can see on Figure 1.
  Therefore, we have numerical data depending on time, in other words we have qualitative observations of real new cases of COVID-19, which could be treated and understood as time series. It does have a clearly temporal and geographic nature.

- The range of values is given by natural numbers from 0 to even 150k, depending clearly on the filtered data we are looking at (cases or deaths per days, months, for whole countries or just small regions). The units used are given by number of cases or deaths. The precision required is one unit of cases or death, i.e. one person.

- The data life span is one day (for the general data), i.e. it should be updated daily, since it gives the cases per day.

- If we study the distribution of the data for the last 30 days, for most of the geographic regions we can visualize a stationary tendency, i.e. it has a periodicity, as shown on Figure 1. In fact, we could see a weekly periodicity, characterized by a decrease of cases during the weekends (or it could be a decrease of documentation of the cases) and a remarkable increase on Mondays.

- Linking this aspect with the previous one, we can see that on Mondays we have values greater than the weekly mean. Similarly, during the weekends, we have smaller values than the weekly mean. Therefore, depending on the quantile used to define the outliers condition, we could have them as outliers when looking at the monthly data. Looking at the whole dataset's timeline, see Figure 2, we see an outstanding interval with remarkable outliers during the last month of January. We also have high values outstanding the tendency at that time during the even previous January.

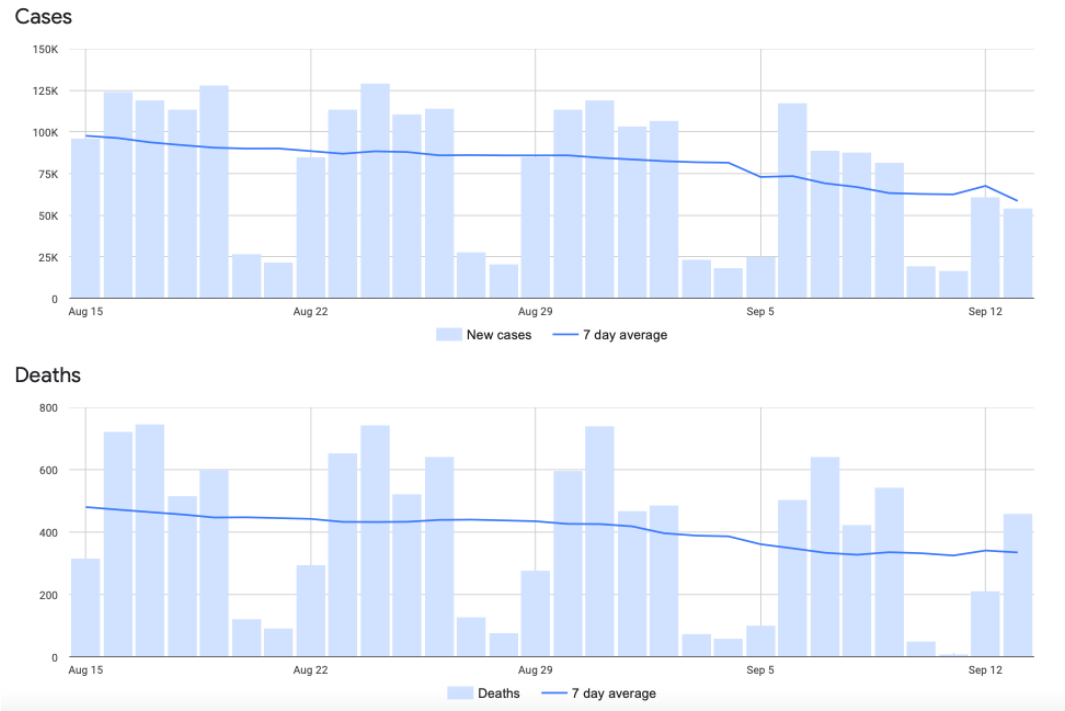


Figure 1: General visualization of new cases and deaths for the past 30 days.

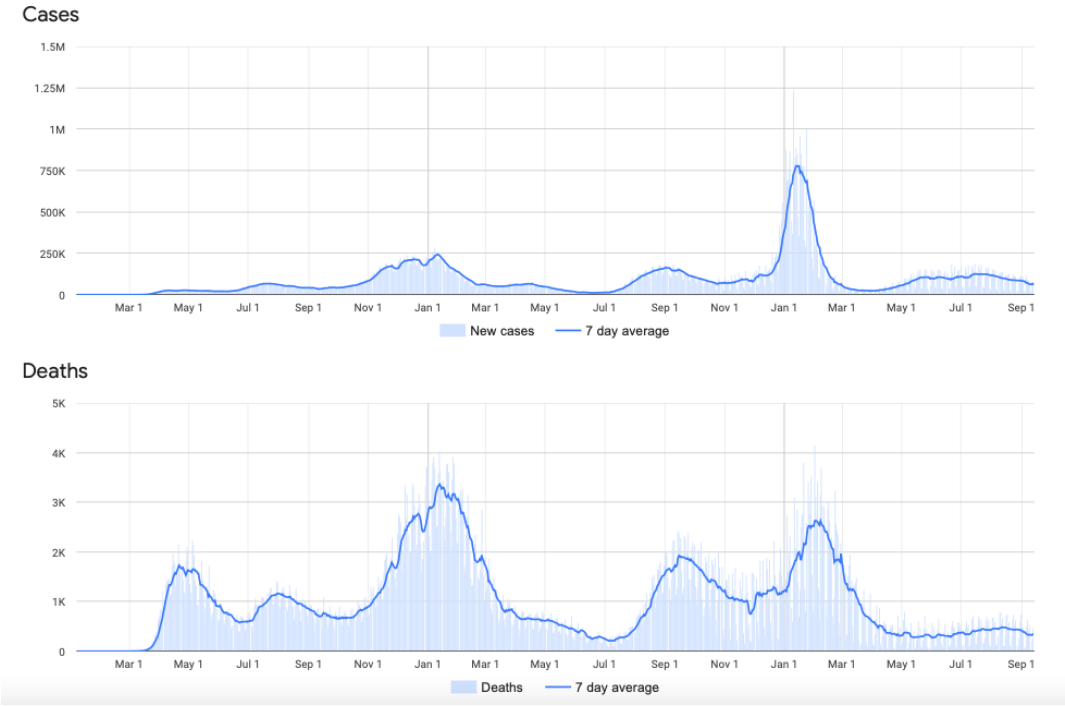


Figure 2: General visualization of new cases and deaths for the whole data set.

1.1.2 Vaccines

Let us explore the data corresponding to vaccines.

- In the same way that before, we have the received vaccines per day, shown for all time, last week, last 14 days or last month. Moreover, the received vaccines are divided in two groups, the ones who have at least one doses and the ones who have the complete vaccination schedule. Furthermore, if we deep into the data, we can see this information for every vaccine, Pfizer, Janssen, Moderna, etc. In addition, we have the received vaccines per day and the seventh day average data. Since the vaccines did not commercialize until December/January of 2020/2021 we have 0 value until that time. We also observe the geographical nature of our data since we can visualize it depending on the country, region and in some cases, municipalities. We give an example of the accumulated number of received vaccines among the time in US.

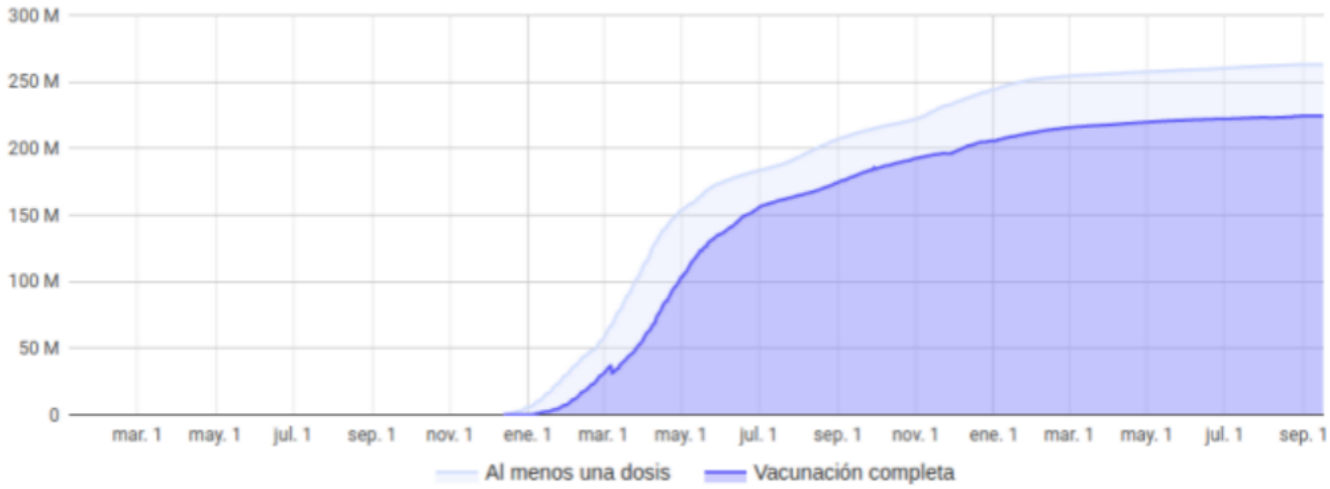


Figure 3: Accumulated number of received vaccines in US from 2020 to 2022.

- The range of values is given by natural numbers from 0 to 5M, depending clearly on the filtered data we are looking at (vaccines per days, months, for whole countries or just small regions). We can also look for the accumulated values among the time, then the numbers will come from 0 to the population of the country. Observe that if we look at these values filtered into this last month (or for a shorter period of time), we will notice that the values are nearly the same among the time, since nowadays most of the population is vaccinated (see Figure 4). For that reason, if we look for the values among all time, we will notice, that in most of the countries, it appears a curve which has the value 0 until the vaccine appears, then increases in a logarithmic scale, and then becomes constant (see Figure 3).The units used in the statistics are given by number of doses received or by the number of complete vaccination schedule.

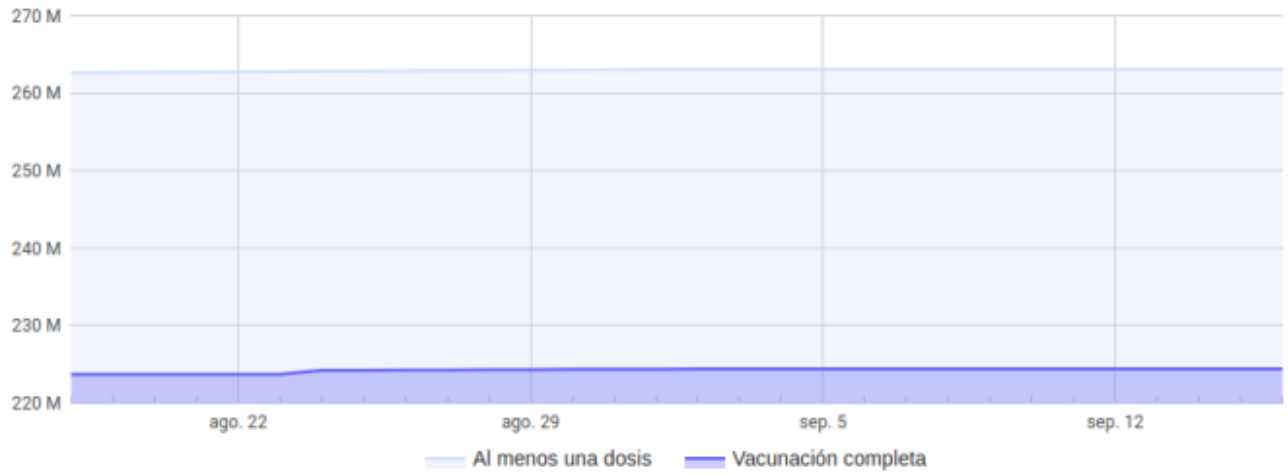


Figure 4: Accumulated number of received vaccines in US of this last month.

- As in the previous section, the data life span is one day (for the general data), i.e. it should be updated daily, since it gives the number of vaccines or the number of complete vaccination schedule per day.
- As we mentioned before, we can create two distinct curves, corresponding to the number of accumulated doses received per day among the time and the other corresponding to the accumulated number of complete vaccination schedule received per day (see Figure 3 to see this dichotomy). Both curves have similar behavior. They are 0 until the vaccine appears, that is December/January of 2020/2021, then, they start increasing in a logarithmic scale (in most of the cases), until most of the population of the region are vaccinated, in which they become constant. Of course, there are countries that the curve increases following another behavior (see Figure 5) but most of the big countries has a logarithmic scale. At the time this project is done, there are still people without the vaccine, so the curve is non constant at all.

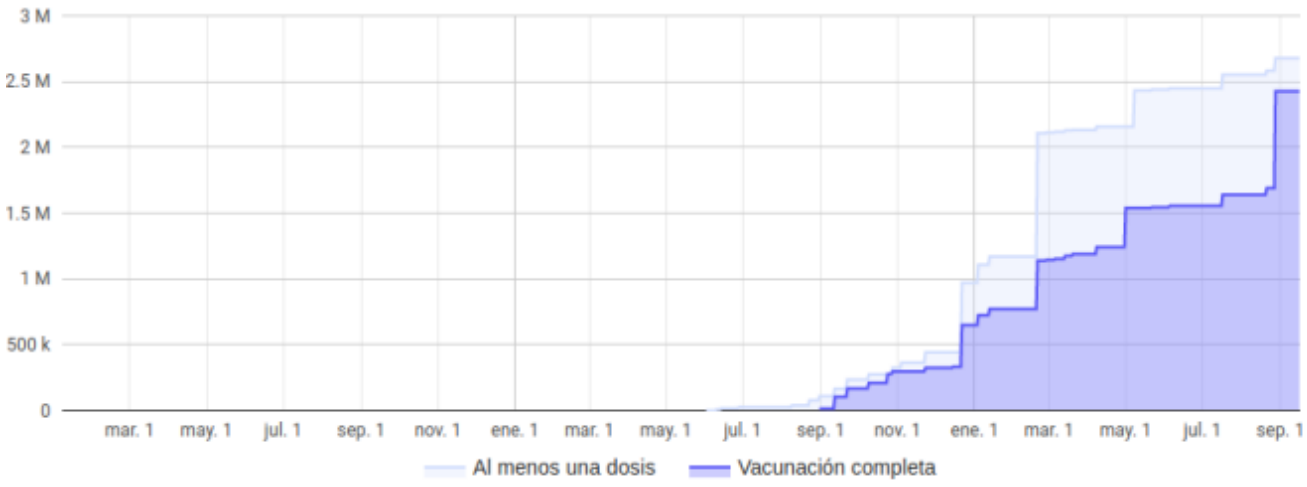


Figure 5: Visualization of the number of received vaccines in Burkina Faso from 2020 to 2022.

- In the case of the vaccines, there does not seem to be any outliers, since the curves present a regular

behavior along the timeline.

1.1.3 Hospitalizations

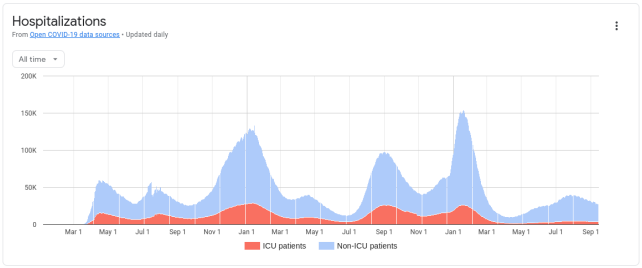
Finally, let's explore the data corresponding to hospitalizations.

- In this part of the dataset we are given the amount of hospitalizations per day, and there's a build in data visualizer that allow us to see the time series weekly, every two weeks, monthly, or all time, as before. The data is also classified per country, region and even municipalities, but for many places we have missing values. We are given the different types of hospitalizations, which are the **hospitalized** patients, the ones that received **intensive care**, and those who needed a **ventilator**. For each of these categories, we are given three different time series, the **new**, the **cumulative**, and the **current** hospitalizations.

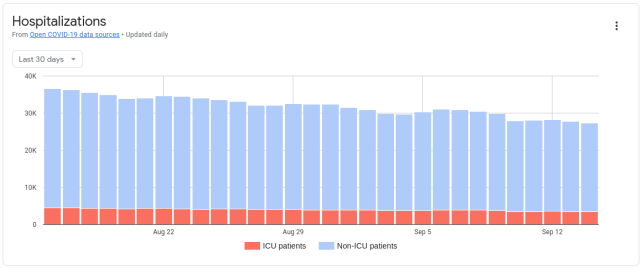
- As values, we receive a range of integers that go from 0 to even more than 150k hospitalizations and fluctuate over time. We can see a clear correlation between the new cases and vaccines graphics, since hospitalizations increase almost at the same time as new cases but at a slower pace when vaccines got commercialized.

- The data life span in this section is very varied, depending on the region we are given daily data, weekly, or even just some days of the week (we assume that these are the working days). Ideally, we should be given daily data to have a proper visualization of how hospitalizations evolve.

- When looking at a specific region for the last thirty days, we observe a monotonically decreasing function, see Figure 6b. But at a larger scale we can observe periodicity over time and some months like January or February have a larger amount of hospitalizations, see Figure 6a. Also, we can spot correlations with the charts of the former topics, that will be further analyzed.



(a) All-time domain.



(b) 30-day time domain.

Figure 6: USA hospitalizations.

- Regarding the outliers, we do not seem to have any of them but, as said before, there's a lot of missing data, as shown on Figure 7. That may be because of privacy policies of different countries.

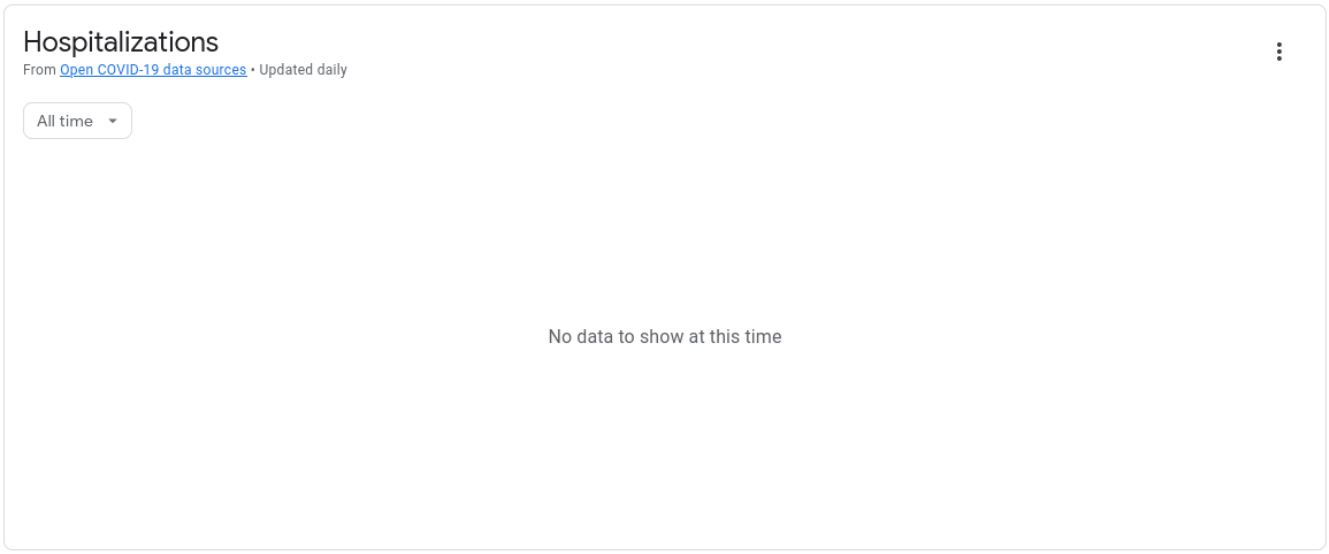
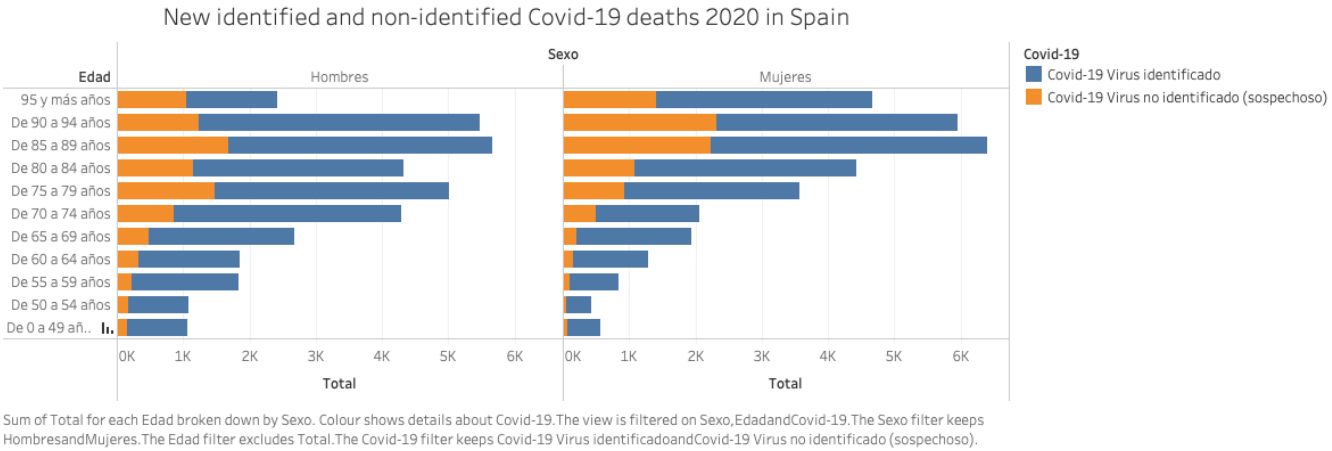


Figure 7: All time Italy, missing values.

1.2 Identify relations and groupings

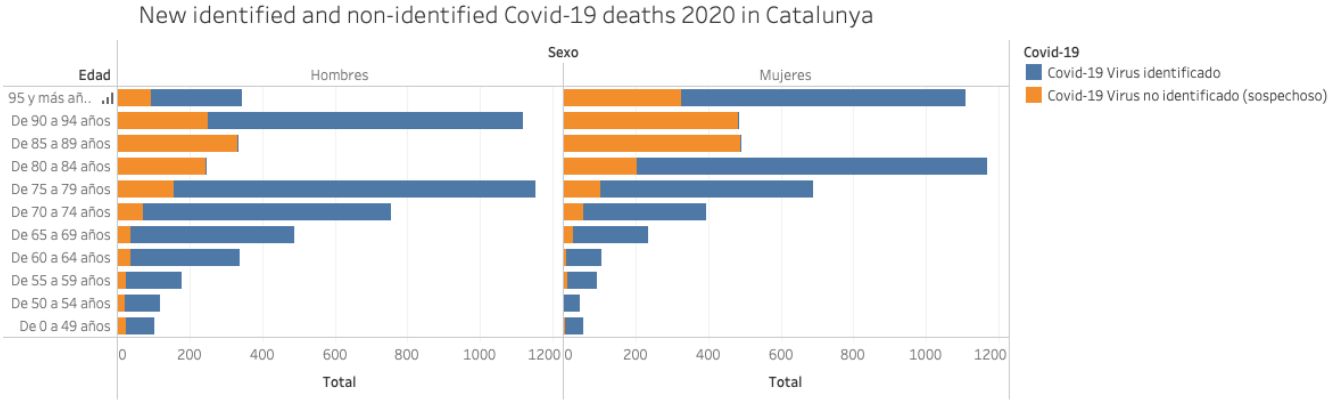
Since there are many flaws and missing values in the former categories, we will focus on analyzing how COVID-19 impacted mainly in Spain. Our goal is to show the impact of COVID-19 in Spain by autonomous communities (CCAA hereinafter). In order to do that, we will visualize the data. Foremost, we will do a deep study on COVID-19 by CCAA relating it with gender and age. With that, we will be able to see if COVID-19 affects young people or older people more, as well as, in gender. Following this epidemics part, we will also try to relate the number of vaccines received, with the new cases and with the hospitalizations by CCAA. This will allow us to conclude if vaccines have been a real solution for the reduction of new cases and hospitalizations among the time. Finally, we are going to study the economy of every CCAA to see how COVID-19 affects to this topic and, in the same way, see how the topic of every CCAA affects COVID-19. To see this kind of relationships, we will study Gross domestic product (GDP since now) and rate of unemployment.

Now, we are going to show some charts about what we were talking about. In this first part of the project, we will only show the data of Catalunya, Comunitat Valenciana and of the whole country. Nevertheless, the idea is to do a final dashboard studying all this topics for every CCAA. Let us start with the epidemics part.



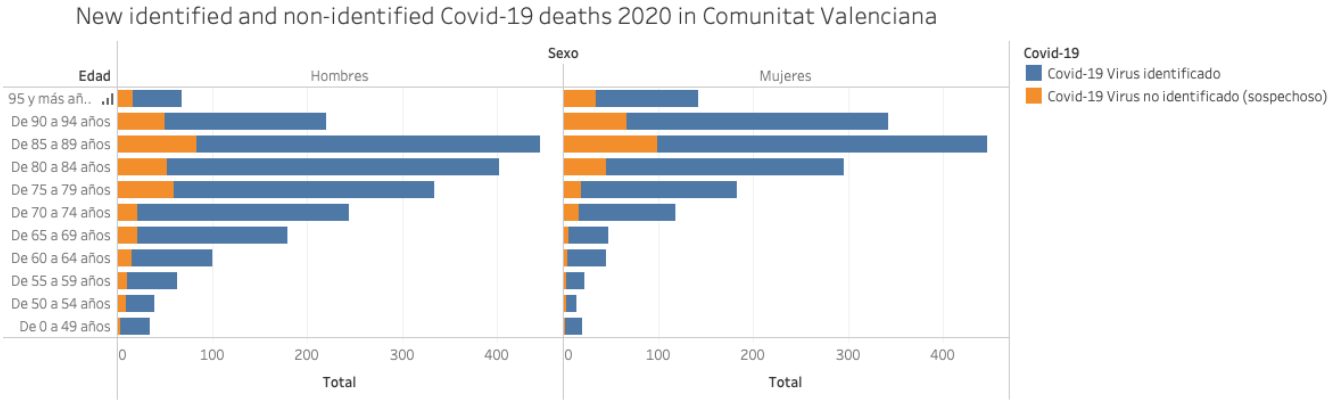
Sum of Total for each Edad broken down by Sexo. Colour shows details about Covid-19.The view is filtered on Sexo,EdadandCovid-19.The Sexo filter keeps HombresandMujeres.The Edad filter excludes Total.The Covid-19 filter keeps Covid-19 Virus identificadoandCovid-19 Virus no identificado (sospechoso).

Figure 8: New identified and non-identified COVID-19 deaths 2020 in Spain.



Sum of Total for each Edad broken down by Sexo. Colour shows details about Covid-19.The data is filtered on Comunidades y Ciudades Autónomas, which keeps Cataluña.The view is filtered on Sexo,Covid-19andEdad.The Sexo filter keeps HombresandMujeres.The Covid-19 filter keeps Covid-19 Virus identificadoandCovid-19 Virus no identificado (sospechoso).The Edad filter excludes Total.

Figure 9: New identified and non-identified COVID-19 deaths 2020 in Catalunya.



Sum of Total for each Edad broken down by Sexo. Colour shows details about Covid-19.The data is filtered on Comunidades y Ciudades Autónomas, which keeps Comunitat Valenciana.The view is filtered on Sexo,Covid-19andEdad.The Sexo filter keeps HombresandMujeres.The Covid-19 filter keeps Covid-19 Virus identificadoandCovid-19 Virus no identificado (sospechoso).The Edad filter excludes Total.

Figure 10: New identified and non-identified COVID-19 deaths 2020 in Comunitat Valenciana.

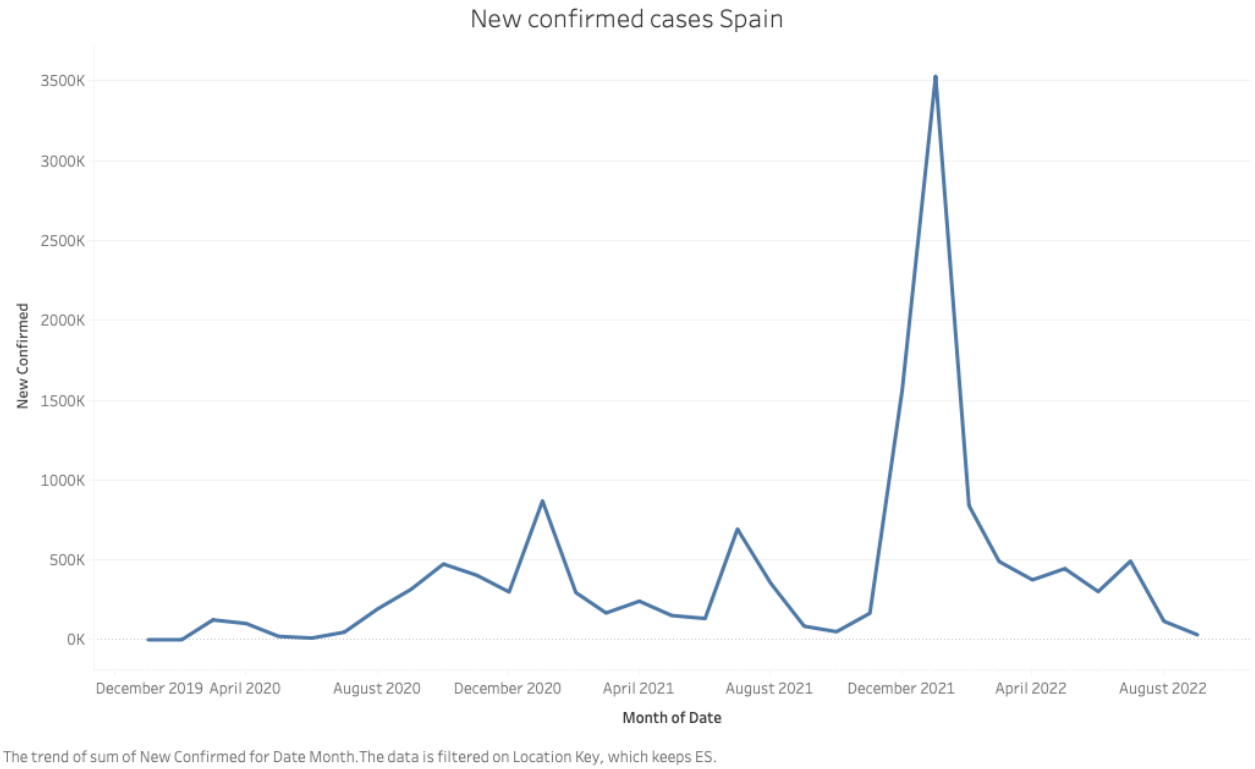


Figure 11: New confirmed cases Spain

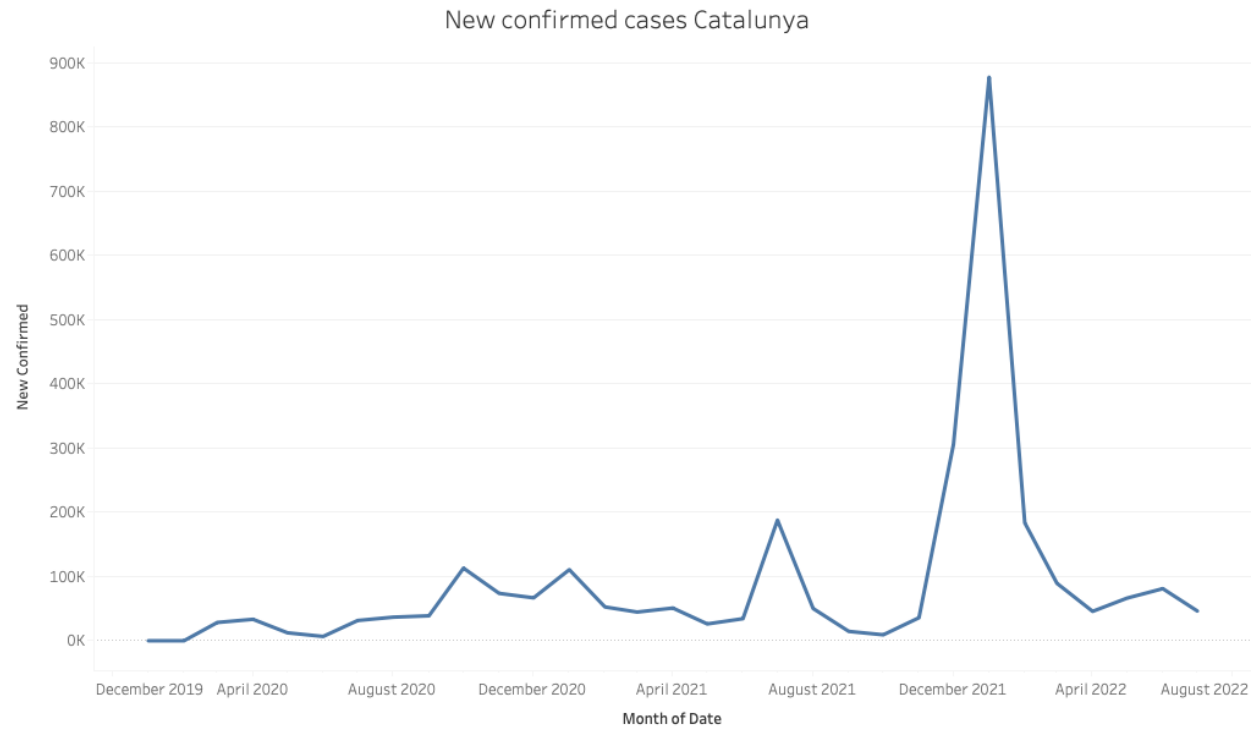


Figure 12: New confirmed cases Catalunya

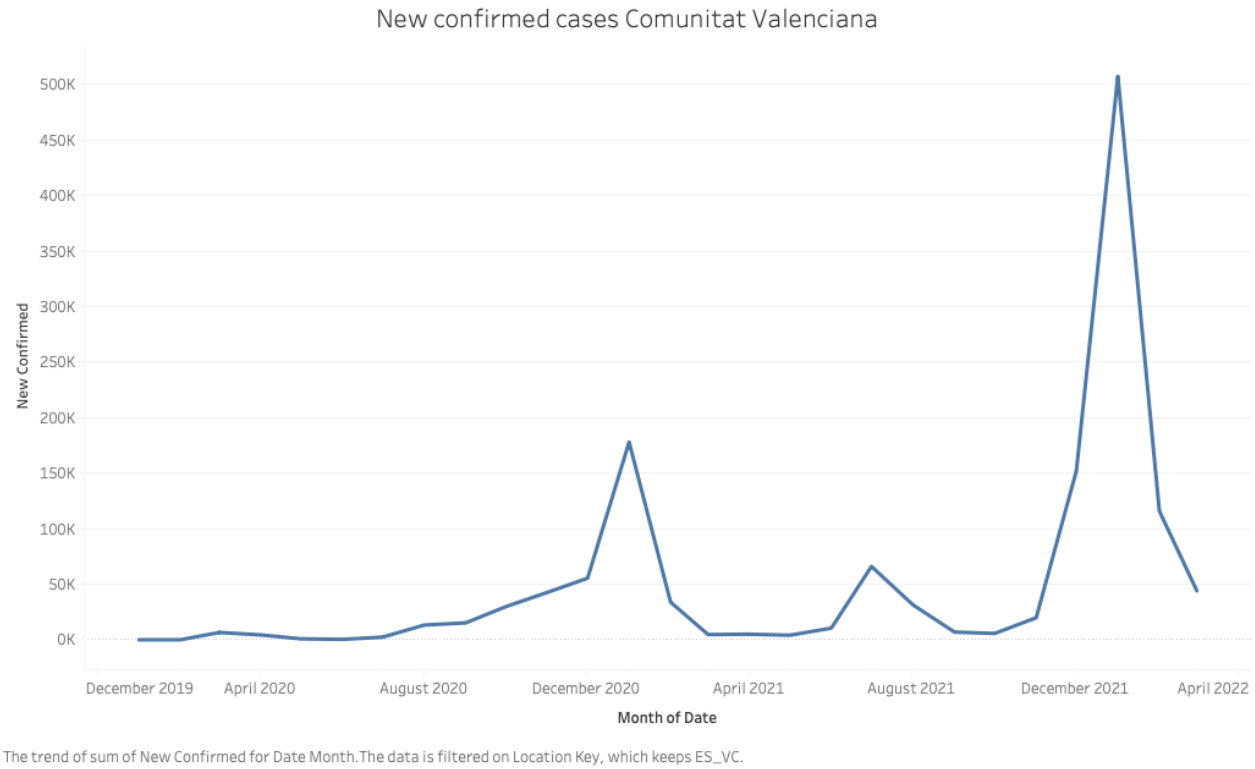


Figure 13: New confirmed cases Comunitat Valenciana

The first three charts, see Figures 8, 9, 10, are the new identified and non-identified (suspicious) deaths of COVID-19. Visualizing the data, we can see that the population group that dies the most in Spain is the 85-89 years old group, both men and women. This behavior continues in the case of Comunitat Valenciana, but in Catalunya, it seems that population group that dies the most is the 75-79 years old group for men and 80-84 years old group for women (we could have missing values for this CCAA and get wrong conclusions, but since deal with missing values is not the goal of this project, we will believe in the giving data). The following three charts, see Figures 11, 12, 13, represent the new confirmed cases of COVID-19. As we can see, there is big uptick in cases in Spain, on December/January of 2021, that is for the arrival of the Omicron variant, which, as we can see, it was very contagious. Let us continue with charts of vaccination and hospitalizations.

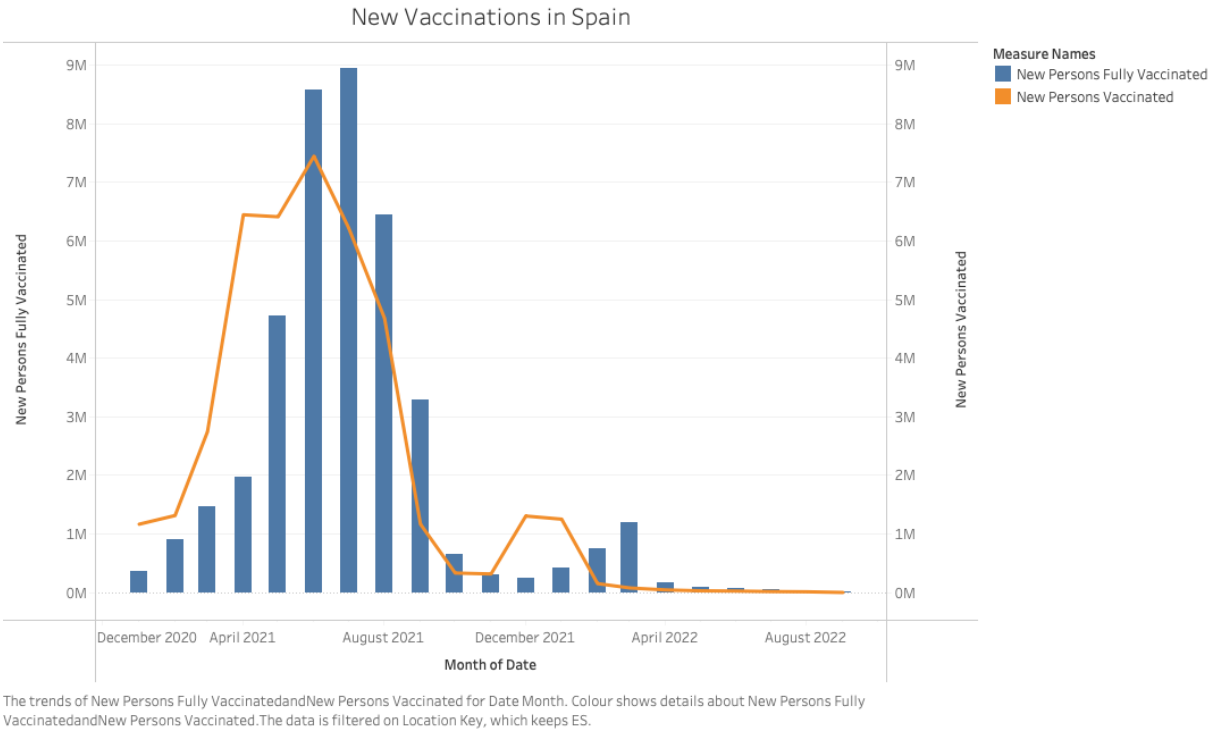


Figure 14: New fully persons vaccinated and new persons vaccinated in Spain

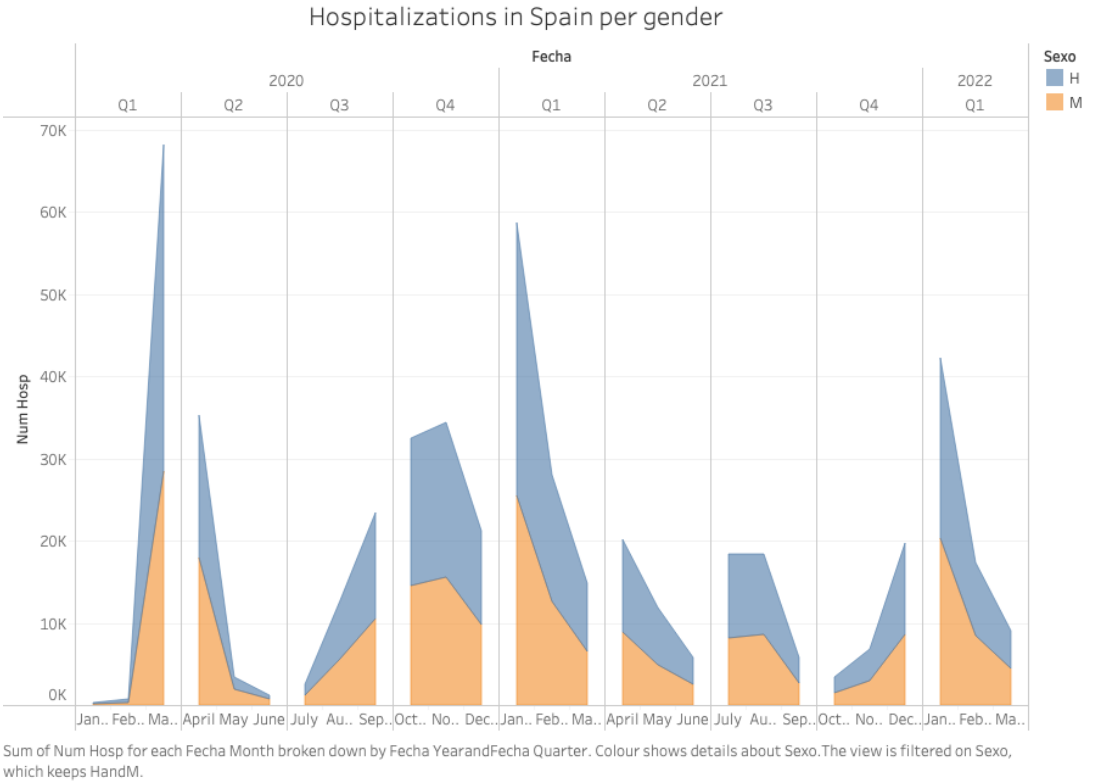


Figure 15: Hospitalizations in Spain per gender

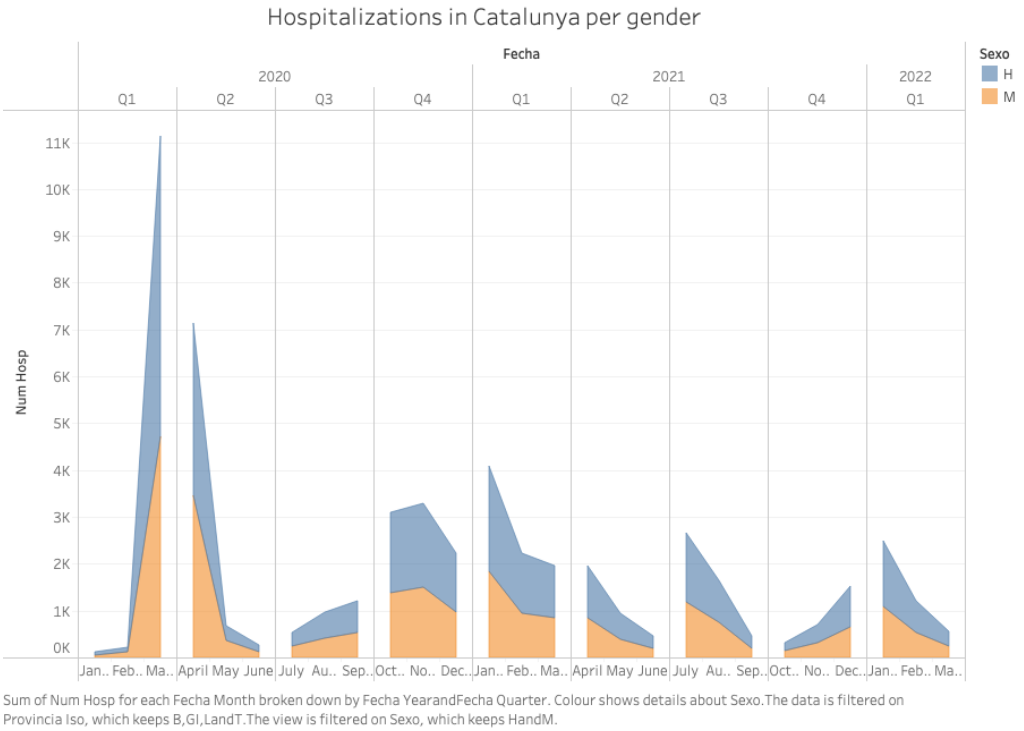


Figure 16: Hospitalizations in Catalunya per gender

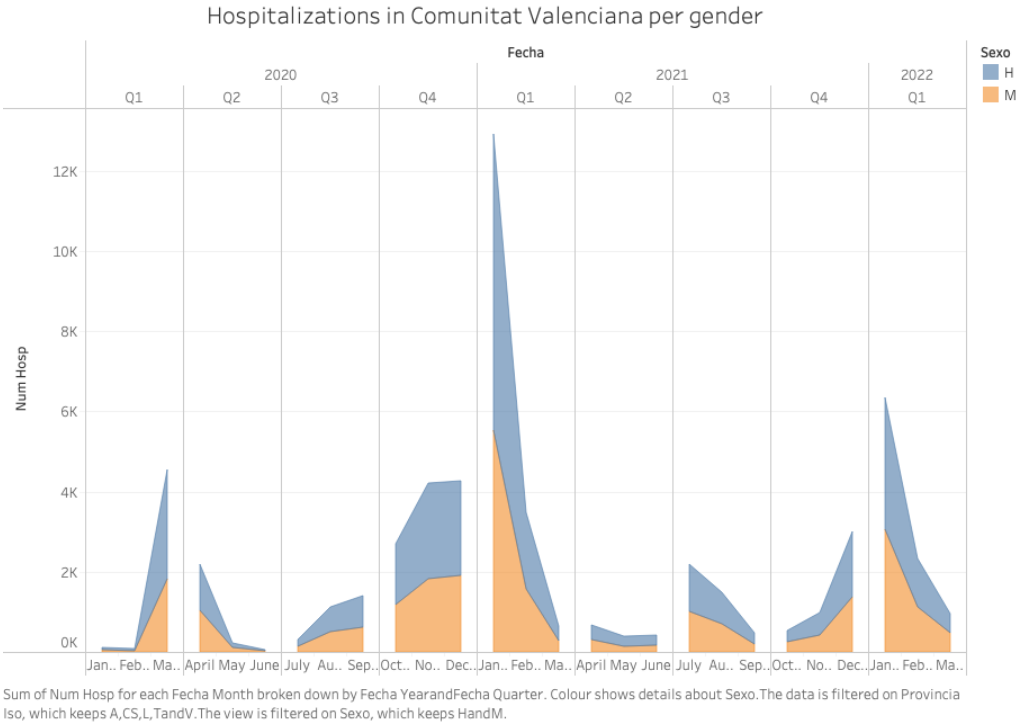


Figure 17: Hospitalizations in Comunitat Valenciana per gender

The first chart, Figure 14, is the new fully vaccinated and new persons vaccinated in Spain. Notice

that when the vaccine appears, the new people vaccinated are obviously more than the people with fully vaccination scheme, but at some point of the summer of 2021, there were more people fully vaccinated than only vaccinated. This is reasonable, since people who received the first doses of the vaccine have to wait about six months to receive the second one. The following three charts, see Figures 15, 16, 17 represent hospitalizations per gender. As we can see, in Spain and Catalunya there is a big peak of hospitalizations in the beginnings of the pandemic, with more effect on men than women. On the other hand, in Comunitat Valenciana, this big beak is presented not in the beginnings of the pandemic, but in the beginnings of 2021, with the new Omicron variant. We can see with this exploration that the Omicron variant, in general cases (Spain), although it was very contagious, it was not so dangerous in terms of hospitalizations comparing with the beginning of the pandemic, except for Comunitat Valenciana. The reason for this may come out for the fact that at the time that Omicron variant appears, most of the population was vaccinated, at least with one dose. To finish this data exploration, we show charts regarding economics.

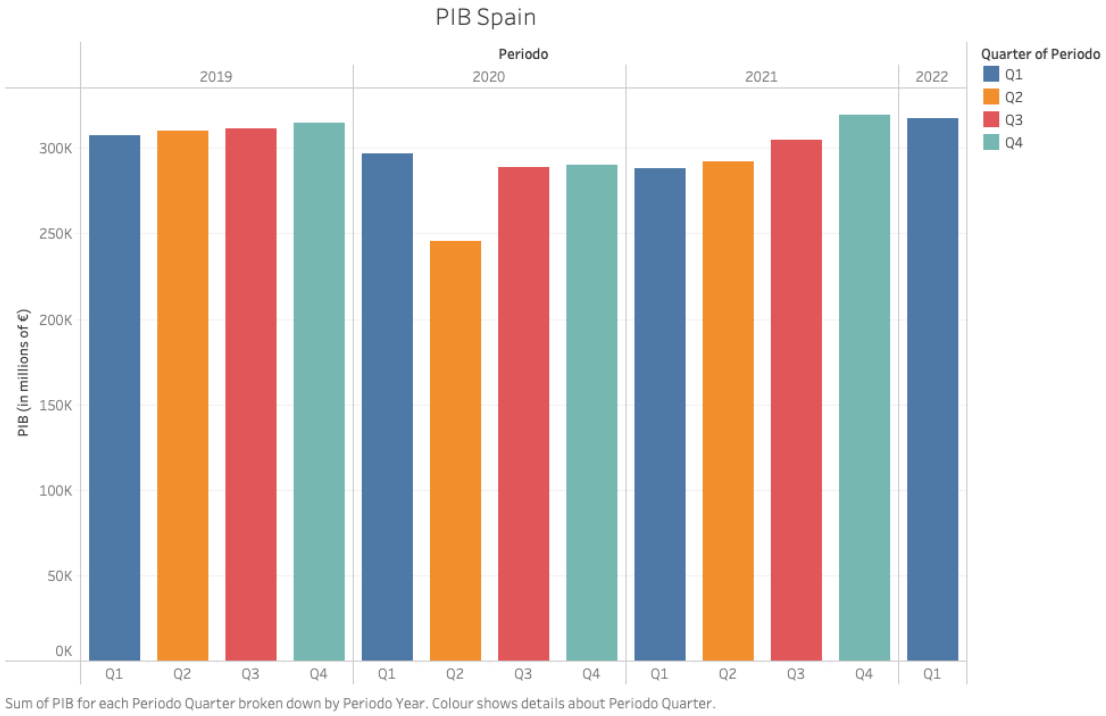


Figure 18: GDP in Spain per trimesters

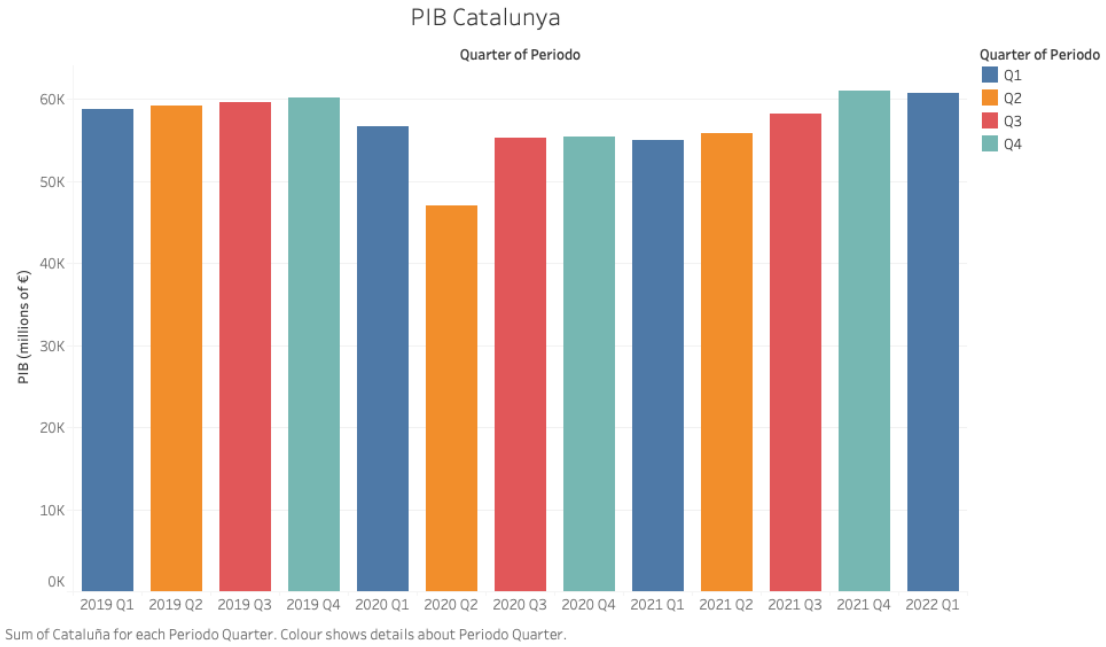


Figure 19: GDP in Catalunya per trimseters

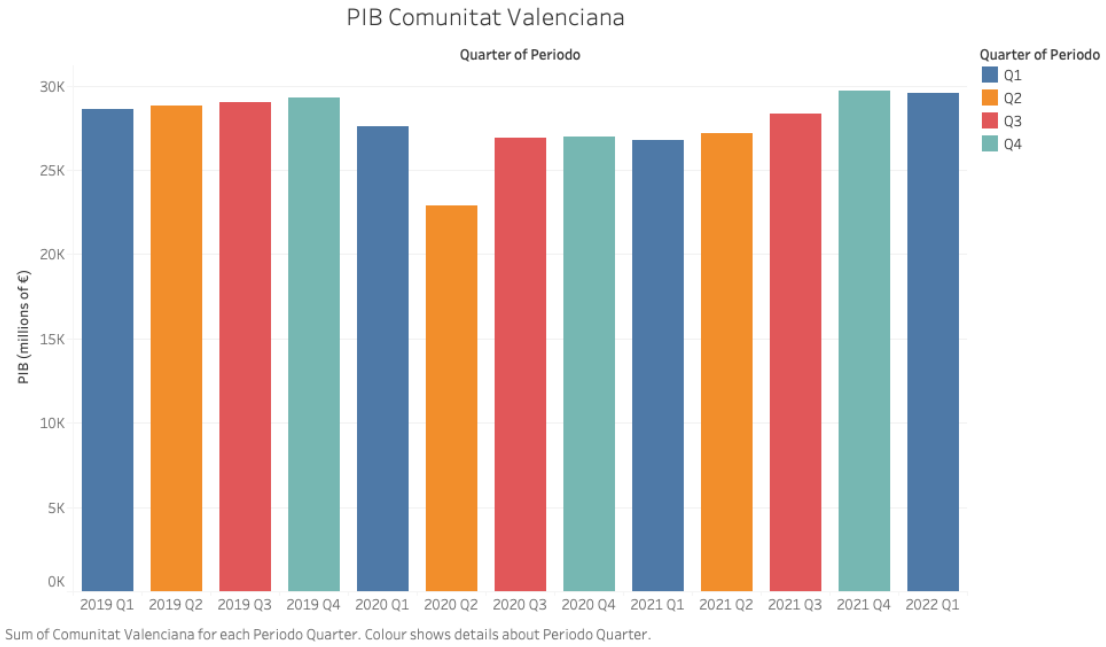


Figure 20: GDP in Comunitat Valenciana per trimseters

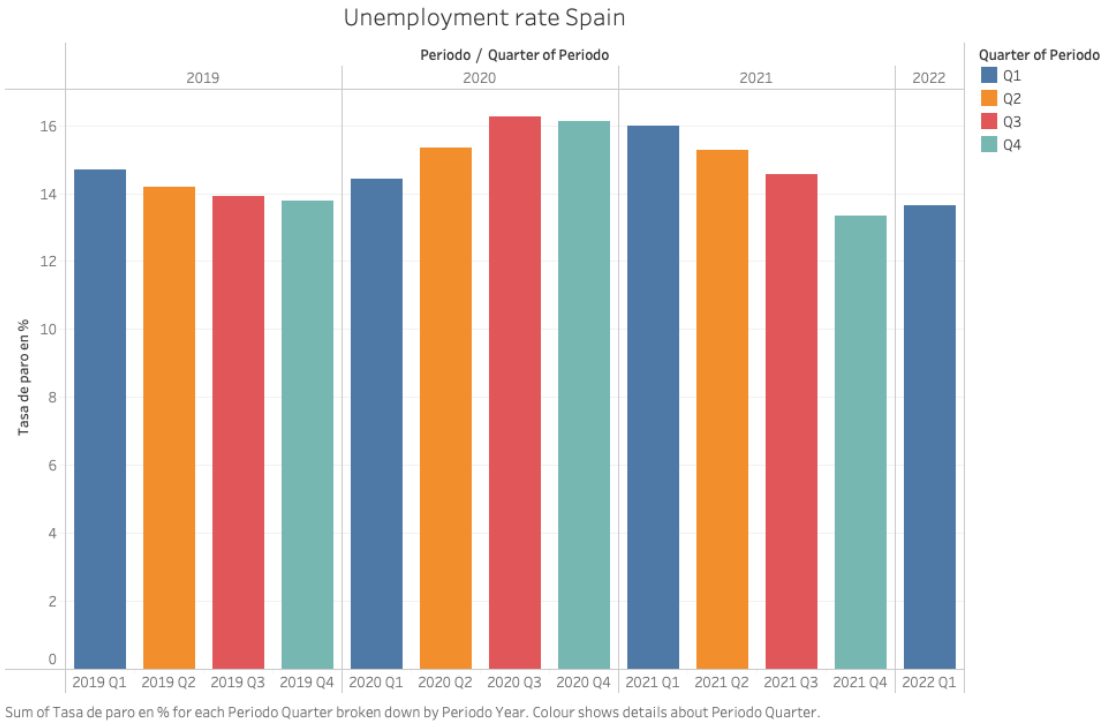


Figure 21: Unemployment rate in Spain per trimesters

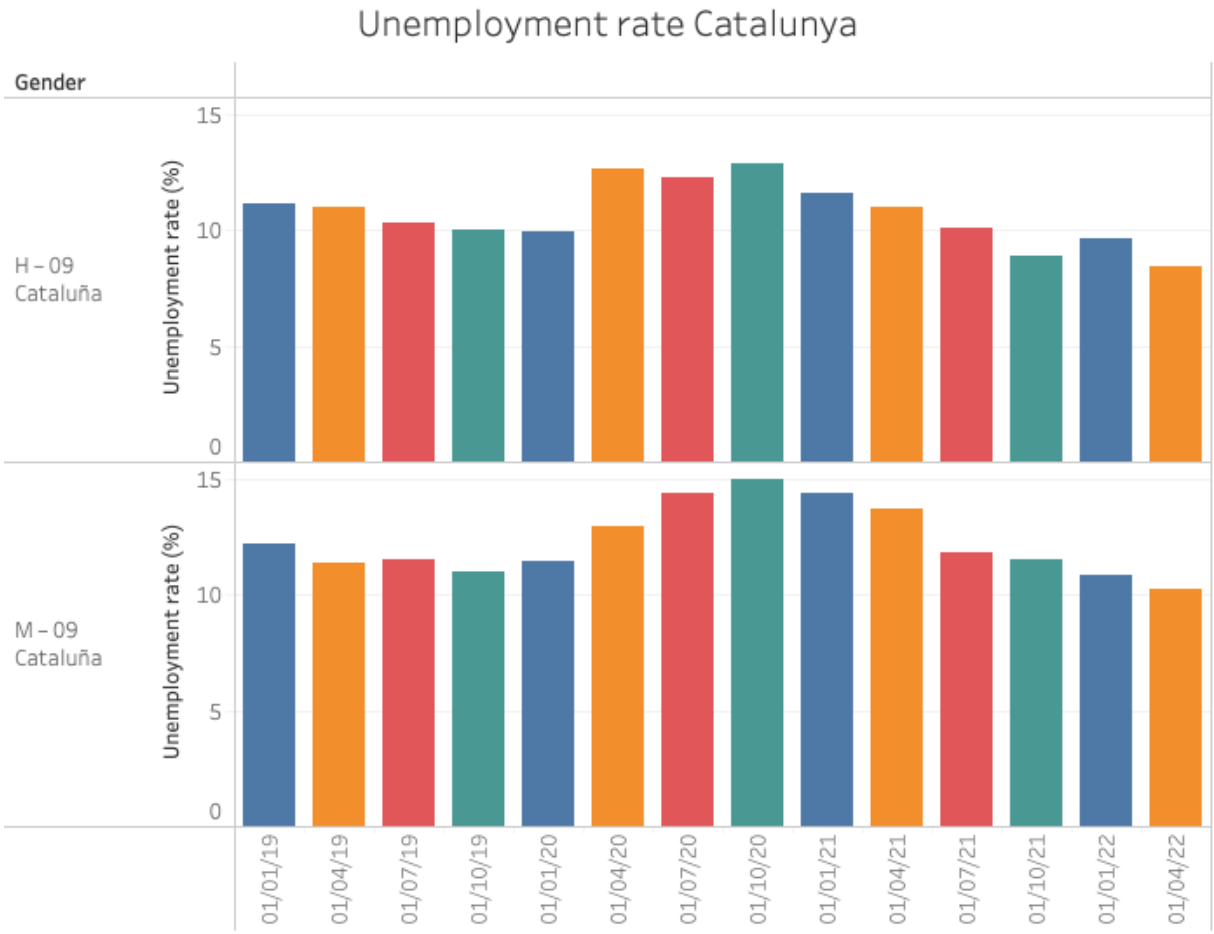


Figure 22: Unemployment rate in Catalunya per trimesters

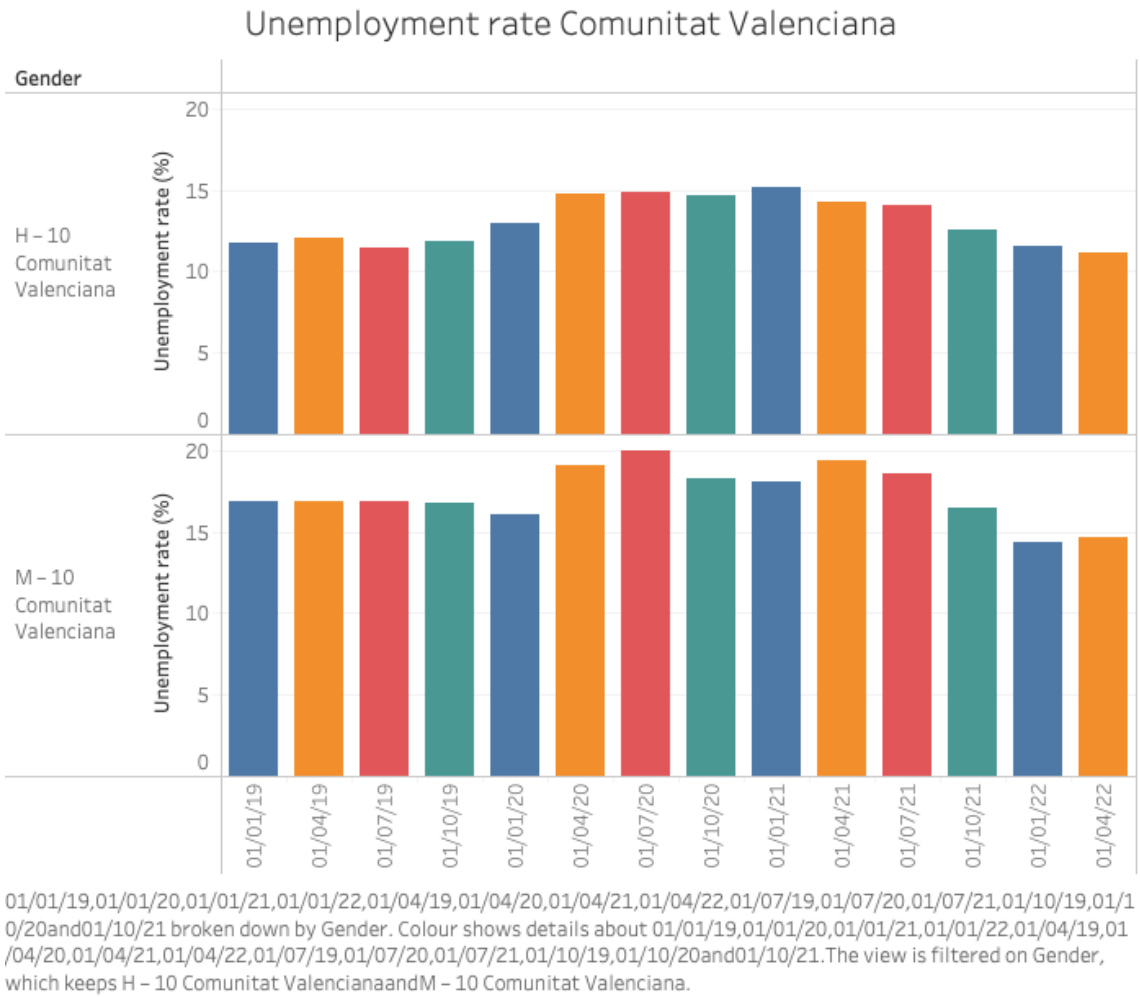


Figure 23: Unemployment rate in Comunitat Valenciana per trimesters

In generals terms, we can observe a big descent of the GDP in the second trimester of 2020, when COVID-19 starts. This is reasonable, since the pandemic stopped the world for a couple of months, being locked up at home. Moreover, we can notice a rise in the unemployment rate in this period of time, due also to the pandemic. As we can see in the charts, this behavior is similar in Catalunya and Comunitat Valenciana, the CCAA we are studying for this first part.

2 Who is your audience

For the past two years, there has been a radical change in the life of most of us. We were forced to live isolated, sent back to our homes, some of us by ourselves, and even went through an unknown illness in the middle of a worldwide chaos. Lockdowns, mandatory masks, mobility restrictions and uncertainty have had a major influence in many levels of society, such as Economics and Demographics. Some of us may have experienced changes in family structures, in habits of consumption or even in mobility.

During most of that period of time, the epidemic was almost everything we could hear of, in the news, at home. However, throughout this two years, we have gathered a lot of information about this previously unknown virus, its effects, its mortality, the way it spreads and how to prevent massive contagions.

Thanks to all this study, we have overcome the brutality of this epidemic, reducing its effects in our society. Nonetheless, we have faced a good deal of fake news or believed what our uncle said about the virus. Therefore, for a common citizen not familiarized with the real data, there are still a lot of flaws and questions to be answered.

Our study is directed to people like us that need validated reasons to comprehend how COVID-19 really affected the population. Our goal is to answer questions such as, 'How the illness affected the health of the population by sex and age?, 'At which level the region or the density of population affected the spread of the epidemic?', 'How it influenced the economy by unemployment rates?', 'Where the governmental policies effective? At which cost?'. And not only answer those, but present the information in a clear and visual way so that everything is fitted in a three chart dashboard.

The idea is to build a principal chart with the epidemics' evolution per region, sex and gender and then two charts that correlate the former with Economics and with Vaccines and Hospitalizations. Since we will use data mainly from the pasts years 2020 and 2021 to build an informative dashboard, clients do not need to visit the document that often, since once build won't be updated.

3 Napkin Design

To finish this first part of the project, we have created a napkin design of our dashboard. The dashboard has four main parts, as observed in Figure 24. In the upper left side, we will find a map of Spain divided by CCAA in which we can see the number of new cases per CCAA depending on the saturation of the color, i.e., the more intensity, the more cases. Moreover, we can see the evolution of this from 2020 to 2022, for that, we will find a slider to go over the time. On the upper right side, we will find a chart that relates number of new cases (in proportion on the total of the population), gender and age. The same slider will work for this chart too. Notice that the time slider is only for the upper part, that is why we clearly distinct both parts, with different colors and with the black mark. In the bottom left side, we will find a chart that relates economy categories as GDP and rate of unemployment with new cases of COVID-19. Finally, at the bottom right side, we will find a chart that relate the number of new cases, vaccines and hospitalizations. The main idea is that once you open the Dashboard, you have all the data described in a general way, i.e., data from the whole country and then, if you click on one CCAA, all the data described will be the data of the clicked CCAA.

This, will allow us to extract some conclusions with only this data visualization. We will see if vaccines were the real solution for the reductions of new cases and hospitalization and how economy affects to COVID-19 and vice versa.
The napkin design is the following.

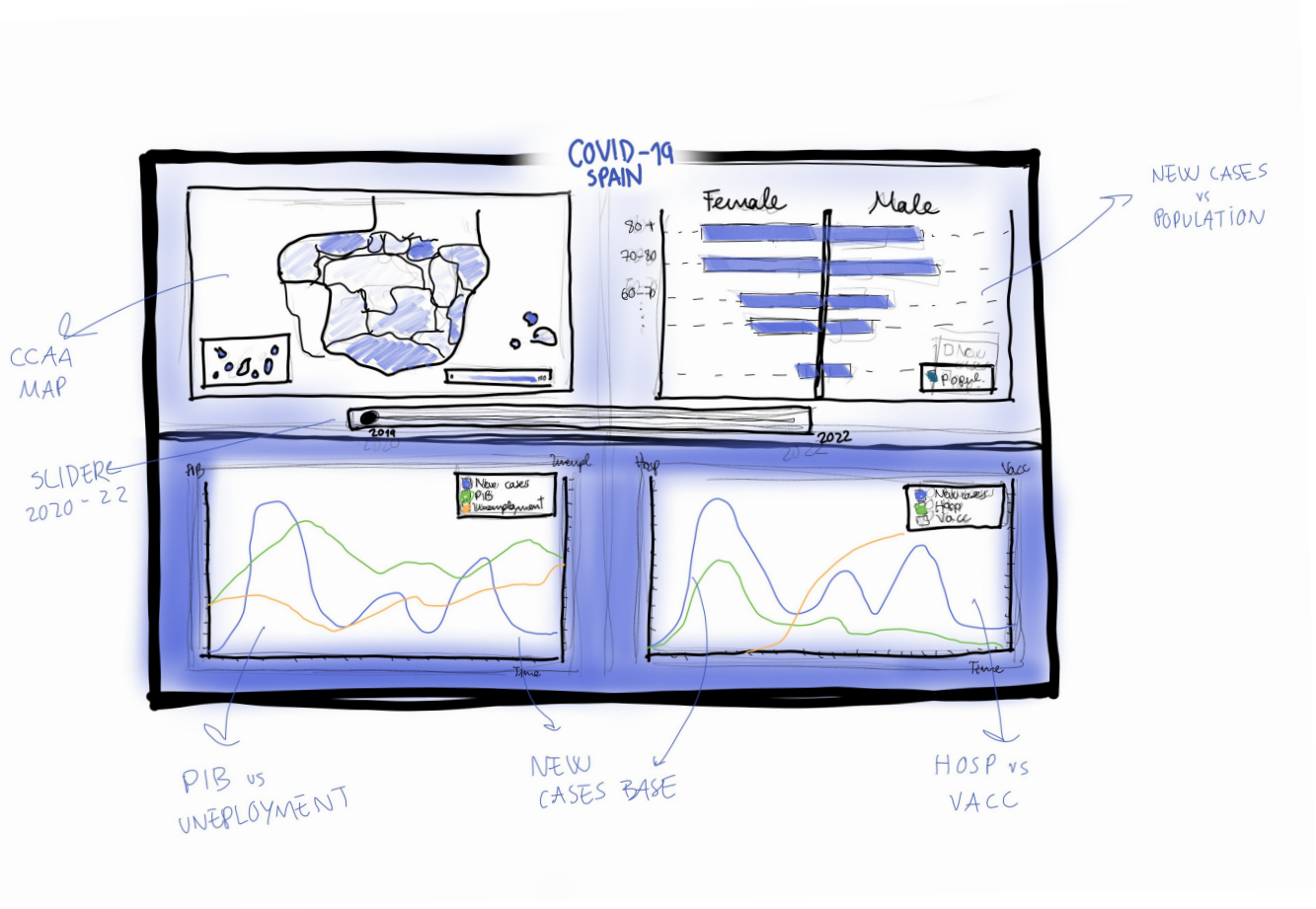


Figure 24: Napkin design.