EU CoVis-19: visual analysis of Covid-19 effects in Europe

Valerio Coretti
Engineering in Computer Science
Sapienza University of Rome
Rome, Italy
coretti.1635747@studenti.uniroma1.it

Fabio Caputo
Engineering in Computer Science
Sapienza University of Rome
Rome, Italy
caputo.1695402@studenti.uniroma1.it

Weihao Peng
Engineering in Computer Science
Sapienza University of Rome
Rome, Italy
peng.1713518@studenti.uniroma1.it

Abstract*—We live in the era of big data. For each topic, we have a huge amount of data that we can analyze. Powerful tools have been created over the years to manage big data. In this document we will try to use these tools to make an in-depth analysis of one of the largest pandemics the world has ever suffered. We are talking about Covid-19. In the past two years of the pandemic, a vast amount of epidemiological data has been collected. We have created a platform for visualizing this data, using the latest available Visual Analytics techniques. We have come up with a solution that can help users better understand information about COVID-19 deaths, cases and vaccines with a focus for the European countries. The repo containing all the material is accessible at the following link: *https://github.com/EU-CoVis-19

***Index Terms*—Visual Analytics, Covid-19, Vaccine**

I. Introduction

II. Related Work

III. Dataset

Before we started implementing our system, we needed a lot of information about COVID-19 and therefore we took the *Our World in Data* [1]. The Dataset is very huge (AS index greather than 6 milion), it contains the collected data for all the world. This dataset was built by collecting data from different sources:

1) COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)
2) European Centre for Disease Prevention and Control
3) Government sources

Dataset is composed by a total o 65 columns, which are the features, and more than 125.000 rows.

A. Preprocessing

The size is very large and for this reason we have decided to select only the European countries. Furthermore analysing the data with some python scripts we see that there was European countries with very few data and for this reason we have also decided to discard the following nations: *['Guernsey', 'Jersey', 'Vatican', 'Andorra', 'Faeroe Islands', 'Gibraltar', 'Isle of Man', 'Kosovo', 'Liechtenstein', 'Monaco', 'San Marino', 'North Macedonia']*. So the number of rows now is about 40 thousand.

Finally we select only a part of the features, the ones related to vaccines, death and cases:

- *name:* Country name
- *continent:* Continent of the geographical location
- *date:* Date of observation
- *population:* Population (latest available values).
- *population_density:* Number of people divided by land area, measured in square kilometers, most recent year available
- *median_age:* Median age of the population, UN projection for 2020
- *gdp_per_capita:* Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
- *cardiovasc_death_rate:* Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)
- *diabetes_prevalence:* Diabetes prevalence (% of population aged 20 to 79) in 2017
- *female_smokers:* Share of women who smoke, most recent year available
- *male_smokers:* Share of men who smoke, most recent year available
- *life_expectancy:* Life expectancy at birth in 2019
- *human_development_index:* A composite index measuring average achievement in three basic dimensions of human development a long and healthy life, knowledge and a decent standard of living.
- *stringency_index*: Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)
- *total_cases*: Total confirmed cases of COVID-19
- *new_cases*: New confirmed cases of COVID-19
- *new_cases_smoothed*: New confirmed cases of COVID-19 (7-day smoothed)
- *total_deaths:* Total deaths attributed to COVID-19
- *new_deaths:* New deaths attributed to COVID-19
- *new_deaths_smoothed:* New deaths attributed to COVID-19 (7-day smoothed)

- *people_vaccinated:* Total number of people who received at least one vaccine dose
- *people_fully_vaccinated:* Total number of people who received all doses prescribed by the vaccination protocol
- *new_vaccinations:* New COVID-19 vaccination doses administered (only calculated for consecutive days)
- *new_vaccinations_smoothed:* New COVID-19 vaccination doses administered (7-day smoothed). For countries that don't report vaccination data on a daily basis, we assume that vaccination changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window
- *total_boosters:* Total number of COVID-19 vaccination booster doses administered (doses administered beyond the number prescribed by the vaccination protocol)

B. Data management

Due to the fact that we have to manage a very huge amount of data, we have chosen to store them inside a non relational DataBase, making the accessibility easier.

C. Principal component analyses (PCA)

Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension [2].

Specifically, for this task, we decided to use PCA (Principal component analysis), a linear technique for dimensionality reduction that performs a linear mapping of the data to a lower-dimensional space so that the variance of the data in the low-dimensional representation is maximized. Therefore, this method allows us to plot each multidimensional tuple on a bidimensional space, still maintaining all the underlying properties. The algorithm is applied on all the attributes and the results are showed inside a scatterplot.

IV. Technologies

Covid19 Visualizer is a platform and it is built as a proper web application with the following technologies:

A. NodeJS and MongoDB

We used NodeJS [4] to built our Back-end, where we do all the computation and where we retrieve the data from the DB. To store the data we have chosen the widely used MongoDB [5], a non relational database that is very easy to use with Node.

B. D3.js

The D3.js framework [6] have been used for the development of the visualizations that compose the service.

C. React

V. Visualizations and Interactions

EU CoVis-19 is composed by a set of different visualizations. In this chapter we want to analyze each component individually to understand what it is showing and how it interacts with the others. As mentioned before, the strength of our platform compared to the existing ones is the interaction between the various charts. In fact, each of them is connected to the others in order to make the user experience simpler and more direct. There are three main views that we are using:

- The first view shows data about the deaths (Fig. 1);
- The second view shows data about the cases (Fig. 2);
- The third view shows data about the vaccinations (Fig. 3);

A. Navbar selection

Opening the platform the first data that are showed are the one related to the Europe. They are an aggregation of the data for all the european countries. Then we hav a navbar where the user can select the views, the countries and also the time span. Furthermore, clicking the circle with the europe flag we can reset the selection.

B. Choropleth Map

C. Line chart

A line chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments [11].

D. Bar chart

A bar chart is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. Therefore, it is used to show comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value [8].

In our platform we used a *stacked bar chart* with the aims to compare different values and different countries at the same time. It shows the following data:

- *Cases view*: It shows the percentage of positives over population in comparison with the percentage of deaths over population and with the stringency index. The purpose of this plot is to have a clear idea of how the stringency index of a nations influences the cases and the deaths.
- *Deaths view*: It shows the percentage of positives over population in comparison with the percentage of deaths over population and with the percentage of deaths over positives. The purpose of this plot is to have a focus on the deaths with respect the cases.
- *Vaccinations view*: It shows the percentages of vaccinated with one, two and three doses of vaccine.

Note that the data showed are related to the date selected as the end of the interval of time, so they are data of one day. Obviously this date can be changed.

Interactions:

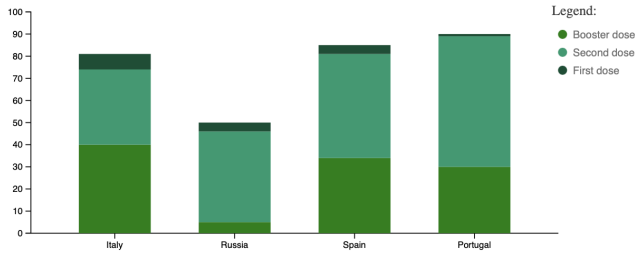

Fig. 1. Bar Chart

- *On mouse over*: It shows a tooltip with the data showed.
- *On Click*: It focus the countries selected. This event has effect also in all the other visualizations that will focus the selected country.

E. PCA chart

A scatter plot is a type of plot or mathematical using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded (color/shape/size), one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis [9].

In EU CoVis-19 PCA chart shows the two principal components (2D visualization) of the data for each selected country. This chart is very useful because showing all the data for a country it makes possible to do comparison between countries and understand well that many nations have provided very few data, and as we have known for some time now, this is a factor that must be taken into account when comparing two or more countries about the covid pandemic.

PCA is also integrated as analytics, indeed it changes based on the countries selected.

Based on the views, the data taken in considerations is different:

- *Cases and Deaths view*: [”new cases”, ”new cases smoothed”, ”total deaths”, ”new deaths”, ”new deaths smoothed”, ”stringency index”]
- *Vaccinations view*: [”new vaccinations smoothed”, ”people fully vaccinated”, ”people vaccinated”, ”total boosters”]

F. Parallel Coordinates chart

Parallel plot or parallel coordinates plot allows to compare the feature of several individual observations (series) on a set of numeric variables. Each vertical bar represents a variable and often has its own scale. (The units can even be different). Values are then plotted as series of lines connected across each axis. [10].

In our platform we used a *parallel coordinates chart* with the aims to compare different values of different countries at the same time. It shows the following data:

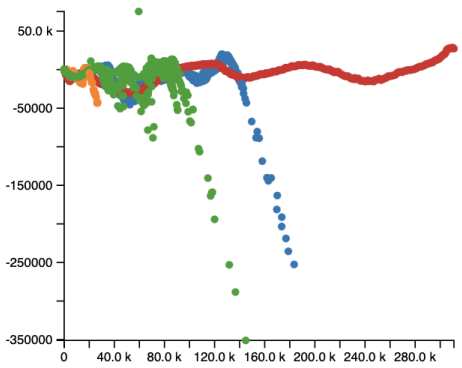

Fig. 2. PCA Chart

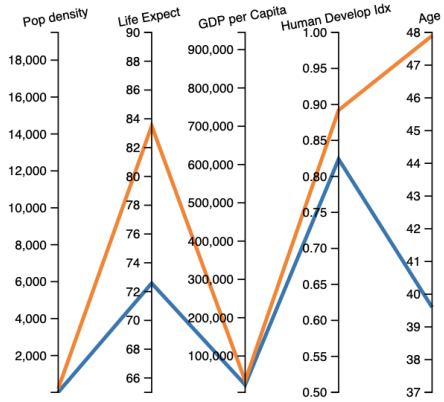

Fig. 3. Parallel Coordinates Chart

- *Cases view*: [’Country’, ’Population’, ’Population density’, ’Smokers’, ’Cardiovasc death rate’, ’Diabetes prevalence’, ’Median Age’].
- *Deaths and Vaccines views*: [’Country’, ’Population’, ’Population density’, ’Life Expectancy’, ’GDP pro capita’, ’Human Development Index’, ’Median Age’].

In all three views we show the basic values that describe a nation, and which can be useful to better understand the pandemic numbers: Population, population density and median age. In addition, in the Cases views we wanted to give greater focus to factors (Smokers, Cardiovasc death rate and Diabetes prevalence) that could affect the respiratory tract and therefore cause a possible infection.

In Vaccines and Deaths views, on the other hand, we wanted to give greater relevance to three factors (Life Expectancy, Human Development Index and GDP per capita) which could be more easily attributable to causes of death or could highlight problems such as not being able to afford suitable treatments.

These factors also allow us to make a comparison based on the wealth of a particular nation.

G. Table chart

We have decided to use the table chart as a big legend for the current view. Indeed in the table are showed the names of

the selected countries each of which with a color, that is the same color the nations have in all the other visualizations. With this choices we want simplify the views to the user and to make uniform the visualizations. The table shows also the data presented in the parallel coordinates chart. The table is also interactive, so it has the ability to change the other visualizations. Interactions:

- *On Click*: It focus the countries selected. This event has effect also in all the other visualizations that will focus the selected country.

Country	Population	Pop. density	Life Expect	GDP	Median age	HDI
Italy	60.4 M	205.859	83.51	35220.1	47.9	0.892
Portugal	10.2 M	112.371	82.05	27936.9	46.2	0.864
Spain	46.7 M	93.105	83.56	34272.4	45.5	0.904
Russia	145.9 M	8.823	72.58	24766	39.6	0.824

VI. Analytics

In our platform we will develop the following analytics:

A. Simple: Comparisons, Aggregations and Operations

EU CoVis-19 has been designed and built to make the comparison between data from different European countries as simple as possible. All the data that are shown when one or more countries is selected from the map (or from the navbar) are the result of a computation that aggregates all the data of the selection and shows them in the various views. So the aggregations are based on countries selected and interval of time (Date) and they are computed at the selection (Not precomputed). This means also that there are a very high number of possible combination of selections that a user could compute.

About *Parallel Coordinates* and *TableChart* they simple aggregation, while in the *LineChart* we have an aggregation plus a summation of the data because to show the trend during the interval of time select we have to sum the data of every day. Furthermore, all the percentages shown in the *BarChart* are the result of a computation, and they are computed at the selection and not predefined.

Europe data, shown at the beginning or when the Europe button is clicked, are computed at the selection, they are the sum of all other countries present in the dataset.

B. Complex: PCA

In this project we have decided not to use Principal Component analysis only as a dimensionality reduction of the dataset, and to show all the data, but also in this case we have tried to make sense of the platform by making the data shown by PCA useful for making comparisons. As already mentioned above, PCA is applied to different data based on the views chosen. This choice helps us to understand how much data we have of each country, because looking at the visualization we realize that some countries provide much more data than others (At least from the sources from which the dataset collects data). This can be a useful yardstick for not reaching hasty conclusions on the data of some countries. PCA is an analytics because the computation is done at each selection and is not precomputed. It is an expensive computation made up of several steps, which is why for us it is a complex analytics.

VII. Case of study

VIII. Conclusion and Future Work

Over the past year, there has been work to visualize and analyze different aspects of big data related to COVID-19. However, most existing viewers focus on visualizing temporal and / or spatial trends, case numbers and mortality, but none take into account all the other possible factors that can affect a nation's pandemic picture. Plus it is very difficult for viewers to make comparisons across multiple views since almost all of them are multi-page and not one-page structured. In this project, we focus on the already existing factors of the various countries and make the comparison of the data as efficient as possible. Our key contributions include the design and development of a big data visualization and visual analysis tool for epidemiological data of COVID-19. By incorporating PCAs as analytics we also provide a tool that helps to ideally understand how different country data is and how much data is being provided. We give users the flexibility to make as many comparisons as they want. The results of the visualizations can help researchers, politicians, but also ordinary people to gain a better understanding of COVID-19 and thus enable them to fight the disease.

Our tool can be applicable to other real-world applications, by changing the data in the dataset, having such a broad view of an issue can help combat it. As a work in progress and in the future, we explore the possibility of incorporating other tools into the project, with particular reference to machine learning techniques that can make predictions on trends or help to better understand the data in possession.

References

[1] https://github.com/owid/covid-19-data
[2] https://en.wikipedia.org/wiki/Dimensionality_reduction
[3] https://www.javascript.com/
[4] https://nodejs.org/it/
[5] https://www.mongodb.com
[6] https://d3js.org/
[7] https://en.wikipedia.org/wiki/Choropleth_map
[8] https://en.wikipedia.org/wiki/Bar_chart
[9] https://en.wikipedia.org/wiki/Scatter_plot
[10] https://www.data-to-viz.com/graph/parallel.html
[11] https://en.wikipedia.org/wiki/Line_chart

Fig. 4. Cases view

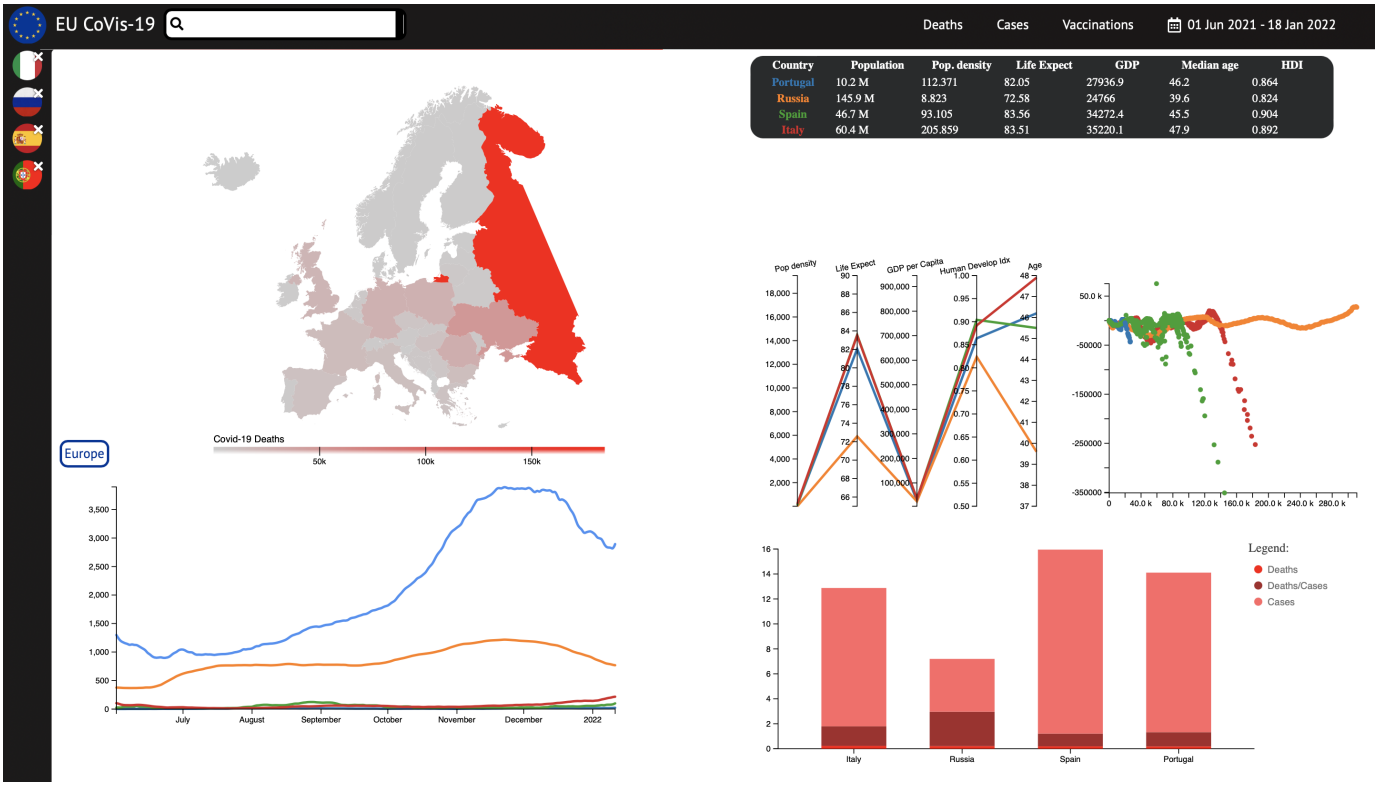

Fig. 5. Deaths view

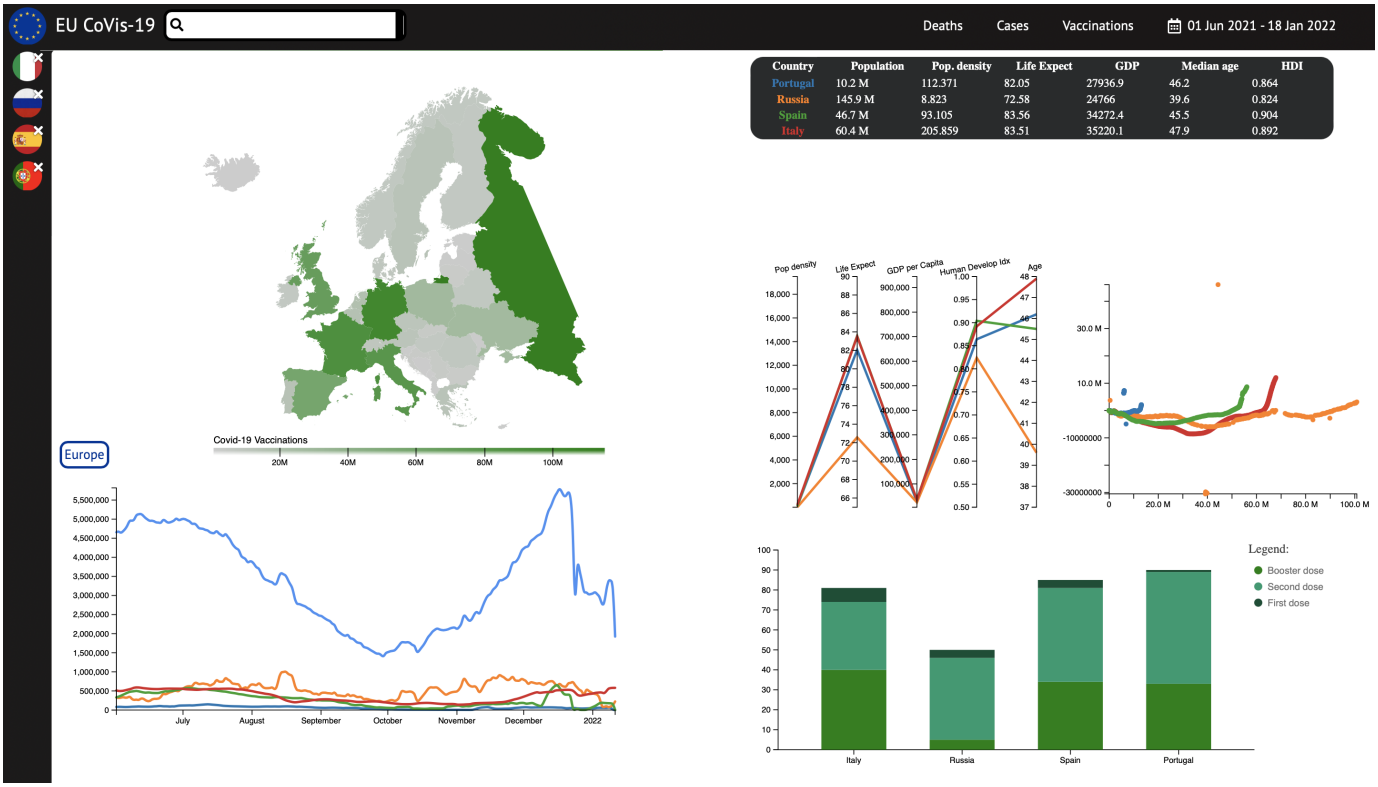

Fig. 6. Vaccinations view