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 on 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

In the last two years, we have been dealing with one of the biggest pandemics in history: The Coronavirus or Covid-19. Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus (SARS-CoV-2). The first known case was identified in Wuhan, China, in December 2019. The disease has since spread worldwide, leading to an ongoing pandemic [23].

As we are living in an era of big data, we have a huge amount of data that we can analyze. Embedded in these big data are useful information and valuable knowledge. So we can use the healthcare and epidemiological data such as data related to the Covid-19 pandemic. Knowledge discovered from these data helps researchers, epidemiologists, and policymakers to get a better understanding of the disease, which may inspire them to come up with ways to detect, prevent and control the disease.

Unfortunately, due to the huge amount of data, we cannot analyze the pandemic for all the countries in the world. So we restrict our analysis only to Europe. For the Covid-19 pandemic, many researchers have focused on different aspects of the Covid-19 disease. A majority of the existing visualizers focused on showing the number of confirmed cases and mortality. However, in addition to these two aspects, other important knowledge can be discovered from the epidemiological data. We will concentrate our analysis on three main aspects: cases, deaths, and vaccinations. Combined with other data(population density, median age, etc.) we will have a better understanding of the Covid-19 situation in these countries.

We have created EU CoVis-19 with two objectives, the first is to make the comparison between different European countries as easy and effective as possible by analyzing different factors, the second is to make the visualizations interactive to make the analysis clearer. In the following sections, we will present our work and at the end, we will also demonstrate a practical case of use, but first, it is necessary to analyze the already existing literature on the subject.

II. Related Works

A. COVID-19 Dashboards

Due to the COVID-19 pandemic, many viewers and dashboards have been developed in the past year. Some of them [13]–[15] viewed literature related to COVID-19 research and others [16] viewed the economic impact of COVID-19. However, most of them [17] focused on actual COVID-19 cases. World-class viewers include a dashboard from the World Health Organization (WHO) Coronavirus Disease 2019 (COVID-19) [18], COVID-19 dashboard from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) [19], and COVID-19 dashboard of the European Center for Disease Prevention and Control (ECDC) [20].

What these dashboards have in common is that they all show the data of cases and deaths related to the covid, but most of them are only concerned with showing the data for what it is, in other words, they perform only information visualization. The total number of confirmed cases and deaths in different countries (or regions/sovereignties) are usually represented by a *bubble map*. In a bubble map, the total number of confirmed cases for each country is indicated by the radius of the bubble that represents the country. While the severity of COVID-19 in many countries can be accounted for by the size of the radius of the bubbles these countries represent, many bubbles overlap. The overlapping and containment of the bubbles makes it difficult to visualize the severity of the disease in countries in dense regions. So a better choice, which is the one made in our project, is to represent these data through a *colorpleth map*, which colors the map with lighter or darker shades based on the severity of the situation under analysis.

Platform	*Comparisons*	*Interactions*	*Analytics*	*PCA*	*Structure*
JHU	❌	❌	❌	❌	multi-page
ECDC	❌	❌	❌	❌	multi-page
WHO	❌	poor	❌	❌	multi-page
CSSE	only textual	❌	❌	❌	multi-page
OWID	✅	tooltip	only aggregations	❌	multi-page
EU CoVis-19	✅	✅	✅	✅	single-page

Fig. 1. Comparison table

Another very useful platform to find out about world covid data is the one from which we downloaded the dataset that is used in EU CoVis-19: Our World in Data [21]. This platform has resulted in a more successful attempt of visual analytics than the other platforms. In fact, on this site, we can start some computations by making comparisons, or using barcharts with time-history which are very useful when it comes to temporal data such as covid data. It also provides some interactions, such as tooltips on mouseover.

In any case, the main problem that all the platforms mentioned have in common is that they all have a multi-page structure, so you can only interact with one view at a time, and therefore it is more difficult to make comparisons on different aspects. Furthermore, this structure involves the non-interactivity of the visualizations which is instead a very important component to facilitate the understanding of the data and to have better comparisons between the various countries.

B. EU CoVis-19

Once these platforms have been analyzed, what we have tried to do in our work is to start from these problems we have just encountered and try to solve them, also adding views that could give a clear idea of the data being analyzed (PCA). EU CoVis-19 is a platform designed to make comparisons between nations clearer not only by showing data relating to cases and deaths but also to other factors such as median age, population, GDP pro capita, etc., which could influence assessments in these cases. In addition, the tool is *single page*, which makes viewing easier and greatly increases the ability to make graphs interactive. *Comparisons* and *interactions* are the strong point of the project.

What have we done more? We have added a *PCA chart* which is fundamental since it gives a clear image of the data available for each nation, which as we know is a factor to be taken into account when doing this type of analysis, as we will show in the case of study there may be countries that provide little data and therefore may be less reliable.

In the table in Fig1, we resume what we have described before to have a clear picture of the position of our work with respect to the already existent.

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 greater than 6 million), 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 of 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 analyzing the data with some python scripts we see that there were European countries with very little 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 daily, we assume that vaccination changed equally daily 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

Because 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 analysis (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 to all the attributes and the results are shown 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

NodeJS is an asynchronous event-driven JavaScript runtime, designed to build scalable network applications. We used NodeJS [4] to build 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 [6], a non-relational database that is very easy to use with Node.

B. D3.js

The D3.js framework [7] has been used for the development of the visualizations that compose the service.

C. ReactJS

ReactJS is an open-source JavaScript library that allows developers to build user interfaces through a component-based approach; it uses Virtual DOM with JSX to create efficient web applications that update components only when necessary [5].

The idea of using reusable, independent, and integrable UI plus compatibility with d3.js was fundamental in our choice; furthermore, ReactJS's hooks suites perfectly with the concept of an interactive and dynamic ecosystem, allowing the developer to manage logic states between components smoothly.

V. Visualizations and Interactions

EU CoVis-19 is composed of 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. Each of them is connected to the others 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. 8);
- The second view shows data about the cases (Fig. 9);
- The third view shows data about the vaccinations (Fig. 10);

A. Navbar selection

Opening the platform the first data that are shown are the ones related to Europe. They are an aggregation of the data for all the European countries. On the top-right side of the navbar, the user can select among vaccination, cases and deaths views and also the period time for the analysis. On the top-left side instead, we find a country selector by name and a button containing the European flag, this can be used to reset the selected countries showing only Europe data and eventually restore the decision. Selected countries will be displayed on the left panel, where the user can easily deselect by clicking on the related flag button.

B. Choropleth Map

A choropleth map is a type of thematic map in which a set of pre-defined areas, the countries in our case of study, is colored or patterned in proportion to a statistical variable that represents an aggregate summary of a geographic characteristic within each area [22]. It is a simple view that allows us to visualize how the pandemic goes in Europe through the time.

In our visualization we use three parameters: cases (blue), deaths (red) and vaccinations (green). Instead of using the usual Covid-19 data about the total number of cases/deaths/vaccinations like many other visualizers, our color of countries is based on the percentage of these parameters compared to the population of each country. We

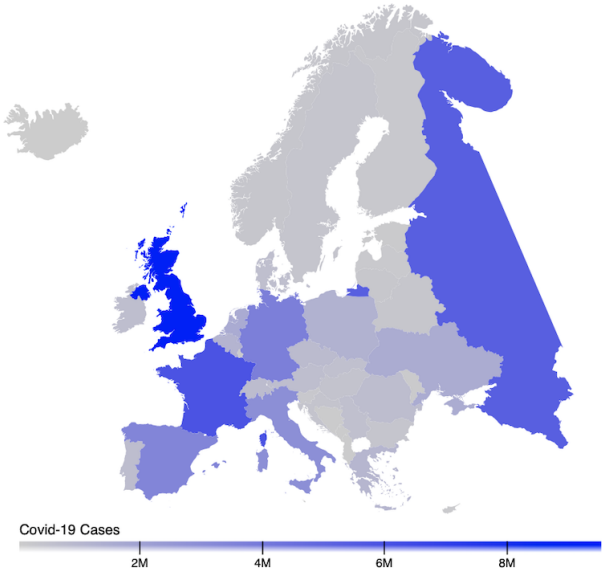

Fig. 2. Chropleth Map

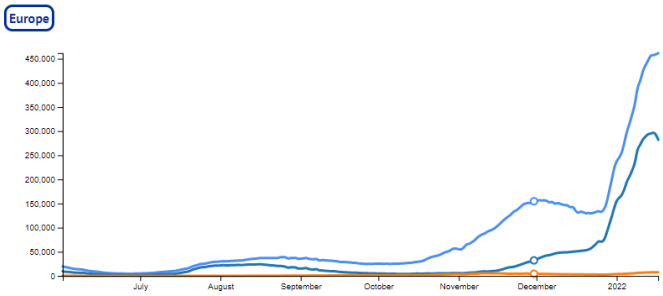

Fig. 3. Line Chart

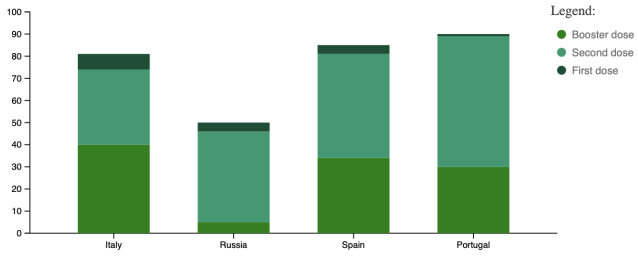

Fig. 4. Bar Chart

use differences in shading to represent it. The darker is the shading, the percentage is high. There is also a legend below the map with the color scale where we can easily check the number associated with the color.

Interactions:

- *On Click*: It will select the country so the other charts will show the data related to the selected country. The selected countries will be displayed on the left panel with their country flag, clicking on the country again will remove it from the selected countries.
- *Zoom*: when a country is selected in the table or bar chart, a zoom / projection on the selected country takes place.

C. Line chart

A line chart is a type of chart that displays information as a series of data points called 'markers' connected by straight line segments [12]. This kind of chart it's very useful to represent temporal information about the displayed data, in our case vaccinations, deaths, and cases among selected countries. Given that the information regarding Europe's data overperforms every single country, we have placed a specific button to hide those data. The intrinsic power of this visualization it's that we can easily check which countries affect more on the overall data for every interval of time. To achieve a smoother visualization of the data, each marker refers to the median value of the last seven days.

Interactions:

- *On mouse over*: Shows a tooltip containing the visualized data for each country containing also Europe and also a circle in the selected point.

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 [9].

In our platform, we used a *stacked bar chart* 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 country 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 to the cases.
- *Vaccinations view*: It shows the percentages of vaccinated with one, two, and three doses of vaccine.

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

Interactions:

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

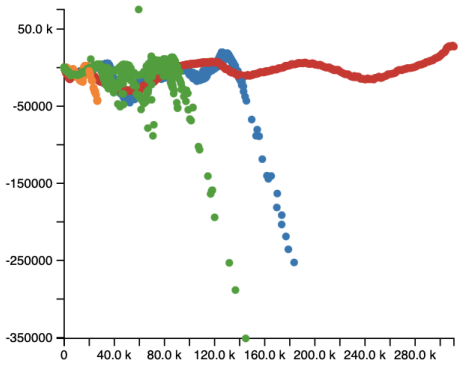

Fig. 5. PCA Chart

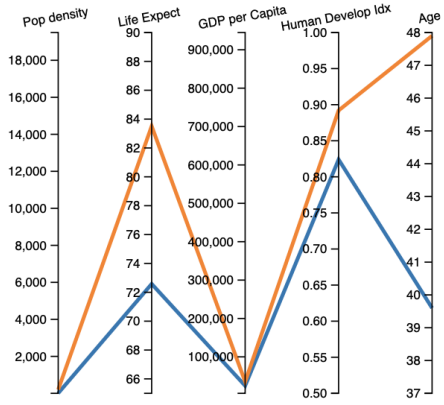

Fig. 6. Parallel Coordinates Chart

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 [10].

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 a comparison between countries and understand well that many nations have provided very little 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

A 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 scale. (The units can even be different). Values are then plotted as a series of lines connected across each axis. [11].

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

- *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 visualizations, 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 shown 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 these choices, we want to 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 can change the other visualizations. Interactions:

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

VI. Analytics

Our platform provides the following analytics:

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

Fig. 7. Table Chart

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 combinations 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 worked to visualize and analyze different aspects of big data related to COVID-19. However, most existing viewers focus on visualizing temporal 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 big data visualization and visual analysis tools 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, and ordinary people to gain a better understanding of COVID-19 and thus enable them to fight the disease.

Our tool can be applied 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, 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://reactnative.dev/
[6] https://www.mongodb.com
[7] https://d3js.org/
[8] https://en.wikipedia.org/wiki/Choropleth_map
[9] https://en.wikipedia.org/wiki/Bar_chart
[10] https://en.wikipedia.org/wiki/Scatter_plot
[11] https://www.data-to-viz.com/graph/parallel.html
[12] https://en.wikipedia.org/wiki/Line_chart
[13] P. Le Bras, et al., Visualising COVID-19 research 2020, CoRR abs/2005.06380
[14] J. Tu, M. Verhagen, B. Cochran, J. Pustejovsky, "Exploration and discovery of the COVID-19 literature through semantic visualization," 2020, CoRR abs/2007.01800
[15] F. Wolinski, "Visualization of diseases at risk in the COVID-19 literature," 2020, CoRR abs/2005.00848
[16] F. Zuo, et al., "An interactive data visualization and analytics tool to evaluate mobility and sociability trends during COVID-19," in ACM KDD Workship on UrbComp 2020, pp. 5:1-5:5
[17] S. Zhang, Y. Cai, J. Li, "Visualization of COVID-19 spread based on spread and extinction indexes," Sci. China Inf. Sci. 63(6), 2020, pp. 164102:1-164102:3
[18] WorldHealthOrganization(WHO),WHOcoronavirusdisease(COVID- 19) dashboard. https://covid19.who.int/
[19] (CSSE) at Johns Hopkins University (JHU) https://coronavirus.jhu.edu/map.html
[20] European Center for Disease Prevention and Control https://www.ecdc.europa.eu/en/covid-19
[21] Our World in Data, Coronavirus Pandemic (COVID-19): https://ourworldindata.org/coronavirus
[22] https://en.wikipedia.org/wiki/Choroplethmap
[23] https://en.wikipedia.org/wiki/COVID-19

Fig. 8. Cases view

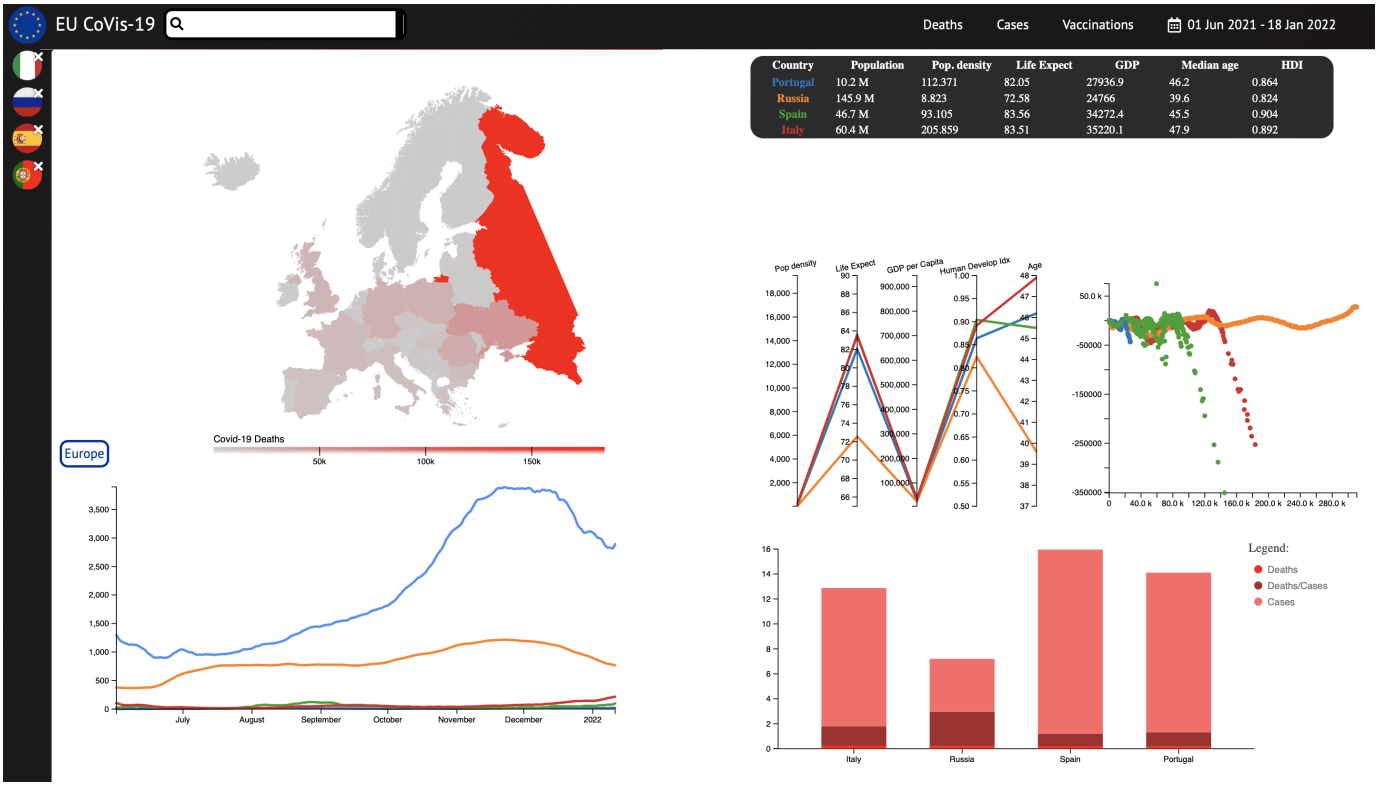

Fig. 9. Deaths view

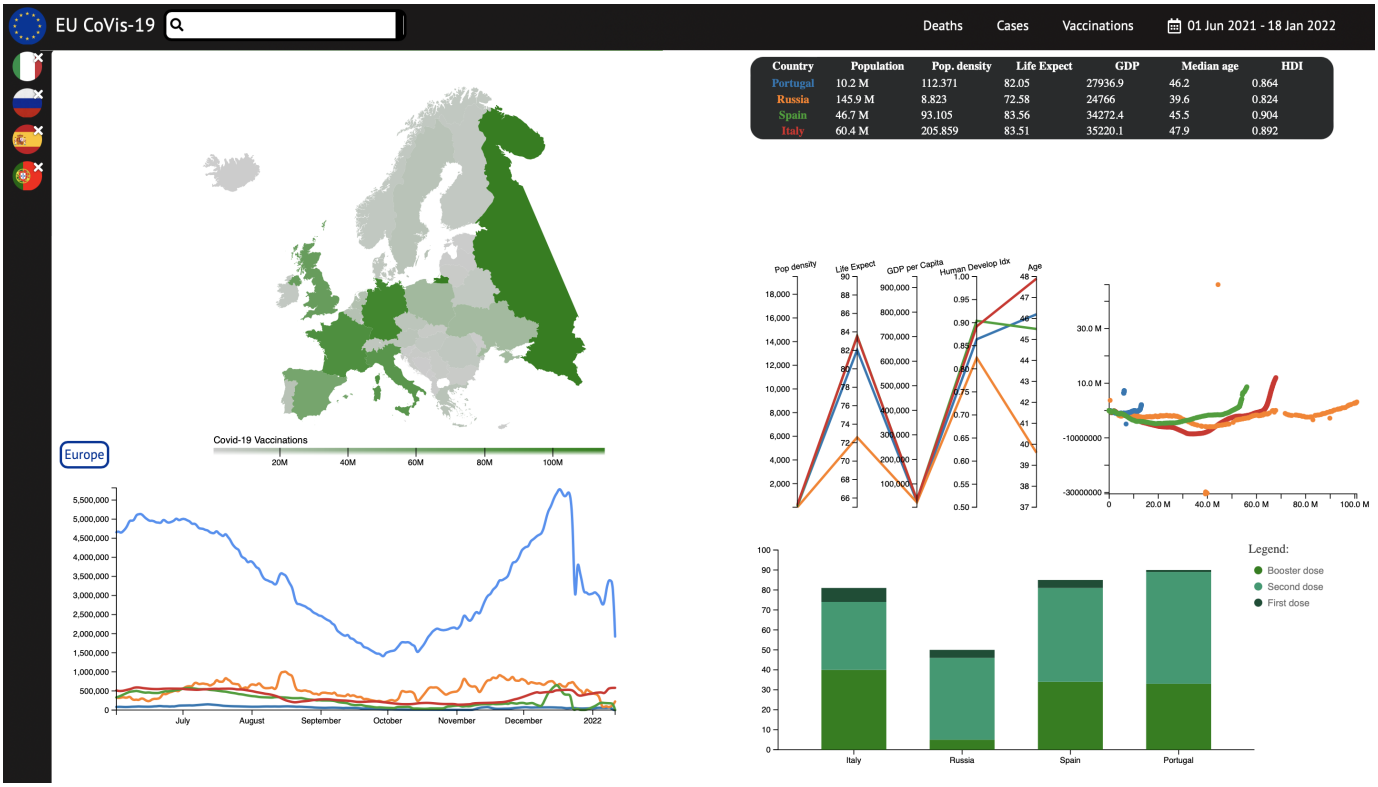

Fig. 10. Vaccinations view