COVID-19 in Data

Data Visualization Final Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Maria Arbelaez, 20221381

Luca Frias Loureiro, 20221750

Lukas Sebastian Gross, 20221363

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1. INTRODUCTION

The COVID-19 pandemic has affected millions of people around the world, with widespread social, economic, and health implications. As the pandemic continues to evolve, it is increasingly important to understand and communicate the complex data associated with COVID-19, from infection rates to vaccination efforts. Our COVID-19 data visualization project aims to communicate the global evolution and impact of the pandemic through interactive visual aids. We were inspired by a personal connection to the issue and recognized the power of data visualization in communicating complex information. Our project includes various visualizations, such as bar charts, scatter plots, and maps, which can be customized and manipulated by users. Additionally, interactive features, such as filters, zooms, and tooltips, allow users to explore the data in the most relevant ways. The goal is to make the data accessible and understandable to a wide range of users of any background.

2. DATASET AND METHODS

2.1. Data description

The dataset we are using for our COVID-19 data visualization project was taken from Our World in Data and it includes a wide range of variables related to the pandemic, as well as other demographic and economic factors that may impact its spread and its global and local respective significance.

For our interactive visualization we selected only some of the metrics contained in the dataset such as: information on the total confirmed cases and deaths attributed to COVID-19, as well as daily new cases and deaths for each country over time; the Government Response Stringency Index, which is a composite measure of a country's response to the pandemic based on nine indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100; data on the number of people fully vaccinated or with at least one shot, and the number of boosters administered per one hundred people in the total population; and finally we use the GDP per capita to let the user explore if there is any correlation between this indicator and the COVID-19 metrics.

All variables in the dataset are broken down by country and date, allowing for detailed analysis of how different factors may impact the spread and impact of COVID-19 across countries and over time. The time frame of the data is three years from January 2020 to December 2022.

2.2. Technical aspects

We utilized the Python programming language and its associated library Plotly, to generate different types of charts and plots. We experimented with various visualizations, including line charts, bar charts, choropleth maps and scatterplots, to find the most effective way to display the data.

Once we had created our visualizations, we used the Dash library, also built in Python, to create a web-based dashboard in pyCharm. The dashboard allows users to interact with the visualizations and customize the data displayed. We incorporated different types of user inputs, such as drop-down menus and radio buttons to make the dashboard more interactive and user-friendly. After that we implemented a Dash based Web Application with Heroku Platform to make our visualization project accessible to a wider audience.

2.3. Visualization and interaction choices

Since our dataset type was a table with items and quantitative attributes, the idioms chosen to visualize it were Bar Charts when comparing one qualitative value attribute with one quantitative value attribute and Scatterplots when compering two quantitative value attributes to explore possible existing correlations. Our data set contained spatial data as well, since it contained quantitative attributes per country, therefore we made use of the Choropleth Map idiom. When creating our visualization, we followed the principles of expressiveness and effectiveness to make sure the visualizations share the right information and communicate the information accurately and efficiently. Our visualization project can be considered of the Interactive Slideshow type as it has a predefined ordering and somewhat strong messaging, but it encourages audience participation to keep them engaged and interested in the story.

3. RESULTS AND DISCUSSION: Reading the visualization using a DV perspective.

3.1. Data encoding and Filtering

3.1.1. Bar Plot: The Evolution of Covid-19

This bar plot lets the user visualize the evolution of daily new Covid-19 cases, new Covid-19 related deaths and death rate (Daily New Deaths / Daily New Cases) worldwide and per country. These bar charts encode two attributes using a line mark with the vertical spatial position channel for the quantitative attribute, in our case the new Covid-19 cases, new Covid-19 related deaths or the death rate, and the horizontal spatial position channel for the date attribute. A drop-down menu lets the users choose whether they want to visualize the data in a worldwide level or for a specific country, and a set of options buttons let them filter the quantitative attribute to visualize. Each bar in the chart includes a tooltip that gives further information on the specific date and quantitative attribute values for that specific bar. The color channel was used to identify each bar plot, using dark blue for cases and dark red for deaths and death rate, as these last two attributes represent more criticality.

3.1.2. Choropleth Map: Covid-19 in a Map

These choropleth maps let the user visualize total Covid-19 cases per million, total Covid-19 related deaths per million, and the Stringency Index per country. For the first two attributes it is important to note that we chose to use the rate per million people in the population instead of the absolute values, so it is easier to compare between countries, otherwise, the most populated countries would have naturally shown more cases. The Stringency Index, being rescaled to a value from 0 to 100, already allows this comparability. A set of option buttons let the user filter the quantitative attribute they want to visualize.

Each choropleth map encodes a quantitative attribute using color saturation variations per region, with darker colors representing higher values and lighter colors representing lower values, and where each region represents a categorical attribute: a country. The visual channel of color is used to represent the quantitative attribute, and the spatial position channel is used to represent the categorical attribute. We used the *color_continuous_scale* parameter to divide the data into classes and provided a legend to help interpret the quantitative values, as well as tooltips for each country where the user can access information regarding the country name and value for the respective quantitative attribute. The resulting maps show the distribution of the quantitative attributes across different countries. Each map used a different color channel: blues for cases, reds for deaths and yellows for stringency index. Note that the first two attributes have the same color hue as in the last visualization to provide consistency over the dashboard. Finally, the user also has the possibility to zoom in and zoom out over the maps, guaranteeing access to visualize even the smaller countries.

3.1.3. Scatter Plot: Comparing Stringency Index and GDP with Covid-19 Severity

This visualization lets the user explore possible relationships between GDPs per capita and Stringency Index with Covid-19 severity per country, measured in total cases and total Covid-19 related deaths. The questions we were hoping to be answered were: were countries with higher stringency measures less affected by Covid-19? Is there a relationship between the economic strength of a country and the severity of the pandemic? We provide two dropdown menus where the user can select which two attributes they want to compare, with four possible combinations resulting in four different plots.

Each scatter plot encodes two quantitative attributes using point marks for each country and both vertical and horizontal spatial positions for each quantitative attribute selected. The horizontal axis represents total cases per million or total deaths per million and the vertical axis represents the stringency index or the GDP per capita. Size and color channels where additionally used to improve the understanding of the chart: the size channel was applied to the y axis with bigger sizes representing countries with larger GDPs or stringency indexes, the color saturation channel was applied to the x axis with darker colors representing countries with more cases or deaths, and the color hue channel was applied to the x axis for the user to differentiate whether they are visualizing cases (blue) or deaths (red). Interactive tooltips allow the user to access information regarding country names and quantitative attribute values for every point mark.

3.1.4. Bar Plot: Countries leading the Vaccination Race

These final bar charts let the users visualize the top 20 countries by people vaccinated per hundred, people fully vaccinated per hundred, and total boosters per hundred. A dropdown menu gives the user the choice to filter which metric they want to visualize.

Each chart encodes two attributes using a line mark with the horizontal spatial position channel for the selected quantitative attribute, and the horizontal spatial position channel for the qualitative attribute or country name. Each bar in the chart includes a tooltip that gives further information on the specific country name and quantitative attribute values for that specific bar. The color channel was used to identify all the metrics represented here as part of the group of measures taken by the countries to combat Covid-19, they all use a yellow color just as the stringency index does in the choropleth map.

4. CONCLUSION

The main objective of our data visualization project was to use visualization techniques so that the user can easily and interactively understand the information contained in the dataset, while always considering the fundamental principles and concepts that underlie data visualizations. We hope that our visual representations will help answering questions for the user such as: how has the pandemic evolved? Is Covid-19 still as contagious and deadly? Which countries have been more affected, and which have taken better measures to combat the pandemic? Among others. We expect that the user can extract this information from our dashboard effortlessly and to confirm this, further validation will be required. The next step of our project would be to design an experiment where our target users can interact with the visualizations, perform a set of tasks, and answer a questionnaire regarding their feelings and judgments about the dashboard. With the findings and insights of this experiment, our visualization project could be further improved.

5. REFERENCES

- Munzner, T. (2014). Visualization Analysis and Design. CRC Press.
- Plotly Technologies Inc. (2023). Plotly Python Graphing Library. Retrieved April 9, 2023, from https://plotly.com/python/