**Ruben López, rjlopez2@gmail.com**

**Andrea Gómez,** [**andrea.gomez.chamorro@gmail.com**](mailto:andrea.gomez.chamorro@gmail.com)

**Data Science Project Module 1**

**Government strategies in containing the pandemic outbreak**

**16th of October 2020**

# Abstract

Many countries are having trouble to deal with the current pandemic situation due to the COVID-19. Coronavirus cases are expected to be associated with the measures and response of the countries to the outbreak. Different measures have been described in the literature ranging from strict lockdowns to optional social distancing. Assessing the overall effect of the measures on the coronavirus cases and whether they were successful is challenging. As part of a collective effort, the measures have been summarized as a stringency index. Moreover, the measures and response of the countries are related to the socioeconomic status of the countries. Therefore, we expected countries with larger GDP and income levels to be able to cope more efficiently with coronavirus cases. At the same time, this will mean that the stringency measurement taken by countries is inversely linked to the number of coronavirus cases. In this project, we imported and performed some data management on the official data sources of these variables. The data showed that the most developed countries are not particularly dealing more efficiently with the virus. On the contrary, the top 10 most affected countries (by number of cases) belong to high income countries, which could be potentially explained by their testing capacity. Following this hypothesis, the least affected countries (by number of cases) belong in general to low income countries, which have a lower testing capacity.

# Table of Contents

[Abstract 0](#_Toc53694709)

[Table of Contents 1](#_Toc53694710)

[Background 2](#_Toc53694711)

[Project Objectives 2](#_Toc53694712)

[Methods 3](#_Toc53694713)

[Data 4](#_Toc53694714)

[Metadata 5](#_Toc53694715)

[Data Quality 6](#_Toc53694716)

[Data Flow 7](#_Toc53694717)

[Data Model 8](#_Toc53694718)

[Risks 10](#_Toc53694719)

[Preliminary Studies 10](#_Toc53694720)

[Conclusions 12](#_Toc53694721)

[References and Bibliography 12](#_Toc53694722)

[Suuplementary information 12](#_Toc53694723)

# 

# 

# 

# 

# 

# Background

Many countries are having trouble to deal with the current pandemic situation due to The COVID-19. Governments have been using a wide range of measures in response, however, is not totally clear what measurement or set of measurement are the more successful in combating the spread of the disease.

The government of Wuarilandia is currently having an uncontrolled outbreak of the disease, with an outstanding number of tolls plus catastrophic consequences in the economy. Like many other countries, they are desperate and lack the experience in dealing with a pandemic. The president of Warilandia, Nikivan Maduque has recently hired the most talented data scientist to address the situation and find out potential solutions. They wanted to check if developed countries are dealing properly to contain the disease and if that is the case, what kind of measures are they taking to success. They were required to address the following questions:

1. Are the more developed countries doing well in containing the disease? 2. If that is the case, which strategy to restrain the outbreak is working the best?

# Project Objectives

The general aim of this project was to evaluate the effect of the measures made by governments on the coronavirus situation. Specifically, we aim to understand whether the socioeconomic status of the countries had an association to a best strategy to contain the pandemic. For this, we have extracted, cleaned and organized the most up-to-date and trustworthy data on the COVID pandemic worldwide, socioeconomic indicators and stringency indices to contain the pandemic outbreak. We have used visualization and descriptive statistics to understand the data and

We hypothesize that the larger stringency measurement taken by countries the smaller number of cases of Covid will occur. We also hypothesize that countries with larger GDP will perform better coping with covid cases. In order to address our working hypothesis, we will perform the following tasks:

* Extract, clean and combine valuable information of time series on up-to-date Covid cases reports, together with government stringency indicators and socioeconomic indicators from most countries around the world.
* Visualize the current situation related to the pandemic dynamics worldwide.
* Find correlation in the stringency indicators, socioeconomic indicator and the pandemic dynamics.
* Identify patterns in countries with the best response against the pandemic outbreak.

# Methods

We collect the raw data from thrustful and recognized online sources that include two universities and one international organization. The three data sources offer the advantages of being publicly available and are maintained and updated on a regular basis for public service. The three datasets can be divided according to the piece of information they provide as follow:

1. Covid epidemiology time series: this data was collected directed from the git repository of COVID-19 Dashboard application by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (1).
2. Covid government response indicator: this dataset was collected from the git repository of the Oxford COVID-19 Government Response Tracker, Blavatnik School of Government (2).
3. Socioeconomic indicator: this data was collected via API protocol communication with the World Bank Open database (3).

We use the program python under the Jupyter notebook environment to write a program that imports the above mentioned raw data from the respective sources. The program collects, organize, arrange, clean and merge the obtained data and finally perform exploratory statistics analysis together with visualizations.

The Covid epidemiology, stringency index and countries’ socioeconomic datasets were imported from the original sources and stored in our program repository as backup where they can be run in case the original data sources web link are down or changed.

We clean, format and organize each dataset to be able to create a final working relational dataset that contains all relevant information on the variables of our interest. Departing from the final working dataframe we conduct exploratory analysis and visualization of Covid time course evolution. We did visualizations on the number of confirmed coronavirus cases and deaths (normalized by 100.000 people) given the income level or stringency index of the countries (details on project module 2).

To perform most of the tasks in our script we use the well-known libraries dedicated for data analysis and scientific computing such as pandas, numpy and matplotlib.pyplot. We also use other libraries for easy working with directories, loading internet content and communicating with APIs such as os, wget and requests.

# Data

The coronavirus data was obtained from the coronavirus resource center ([Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering](https://coronavirus.jhu.edu/map.html)) based at the John Hopkins University. The data is collected from the official and publicly available sources per country. This time series database has been the baseline for COVID infections worldwide (1). We used this time series dataset to analyse the infection rate, death cases and recovery by country (Table 1). We expected the coronavirus cases to be associated with the measures and response of the countries to the outbreak. The measures and response of the countries could be related to their socioeconomic status. Therefore, we used socioeconomic indicators such as the GDP and the income level stratification from the last year. These socioeconomic indicators can be obtained from the World Bank Open Data (2; Table 2). There has been a wide range of measures made by the governments to the outbreak. The coronavirus government response tracker (OxCGRT) based at University of Oxford created the stringency and policy indices to record the number of government policies (3). The data is collected from publicly available sources using 17 indicators (the details of the indicators are properly explained on the website of the coronavirus government response tracker). The indicators are aggregated and reported as a number between 1 and 100 reflecting the government action. We used this time series dataset to analyse the stringency index per country (Table 3).

Table 1. Coronavirus data selected from the coronavirus resource center at the John Hopkins University.

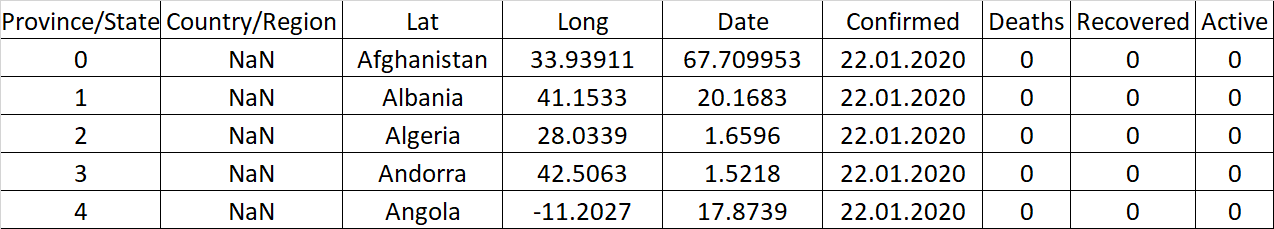
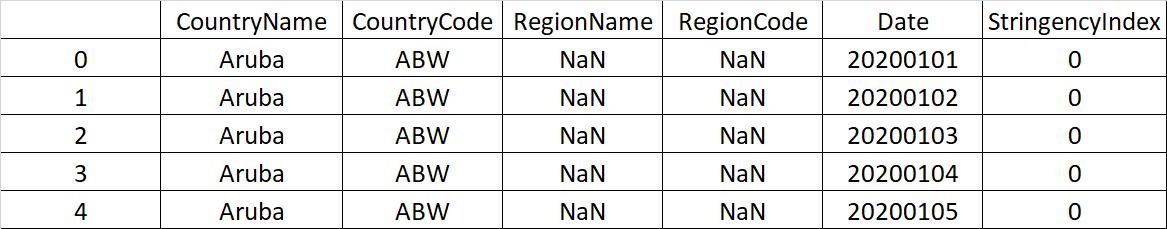


Table 2. Income Group indicator selected from the World Bank Open Data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Country Code | Region | IncomeGroup | TableName |
| 0 | ABW | Latin America & Caribbean | High income | Aruba |
| 1 | AFG | South Asia | Low income | Afghanistan |
| 2 | AGO | Sub-Saharan Africa | Lower middle income | Angola |
| 3 | ALB | Europe & Central Asia | Upper middle income | Albania |
| 4 | AND | Europe & Central Asia | High income | Andorra |

Table 3. Stringency Index selected from the coronavirus government response tracker at University of Oxford



The three datasets were carefully organized, cleaned and merged. New variables were also created departing from the merged dataset such as active cases and new daily cases of infection, deaths and recovery. Furthermore, in order to make reasonable comparisons among country’s performance, we normalize all our infections’ related variables by 100000 people based on each country's population. So the final working dataframe consists of a time series spanning from 22 January 2020 until the present day with normalized COVID cumulative cases as well as new daily reported cases. Additionally, the final dataset contains information on the severity of the government response defined as the stringency index. Finally, it also contains socioeconomic indicators of all the countries considered in our study.

# Metadata

The metadata of this project can be accessed in the supplementary folder of the git repository dedicated for this project in the following link: <https://github.com/rjlopez2/ADS_CAS_Bern_2020/tree/main/Projects/M1%20and%20M2/Suppl_info>. This folder contains information on detailed metadata of the three used datatest as well as the metadata generated in our script and final working dataframe. Metadata was directly taken from the respective repositories or websites of each dataset source. The metadata can be summarized as described in table 4.

Table 4. Summarized metadata of the three datasets used in this project.

|  |  |
| --- | --- |
| Dataset | Variables |
| COVID infections | Date, country names, country codes, confirmed, deaths, recovery. |
| Socioeconomic indicators | Countries, country codes, PIB, income level. |
| Stringency index | Date, countries, country codes, stringency index. |

Detailed information on the complete variable names and description from our final working dataset can be found in the excel file located at the supplementary information. This directory also contains metadata related to the script in an additional excel file.

# Data Quality

The data quality of this project relies on the data quality of its repositories sources. Here, we used three trustful data repositories. However, we encounter with a number of limitations from using online sources that we can list as follow:

1. Multiple data sources do not always use the same name or format for similar variables, for example the country name or the date. To overcome this issue we used a similar date format and standardized the country names of each dataset by using ISO country codes and make in this way multiple datasets compatible based on one (or more) key variable.
2. In the case of covid epidemiology and stringency index datasets the cases are reported for some countries at national and regional level. We aggregate the regional data and focus only on national level reports.
3. In the case of the of covid epidemiology dataset, some extra information not relevant for our study was filtered out. For example we found record cases for three ship cruises (Grand Princess, Diamond Princess and MS Zaandam) that were removed because of incompatibility for the rest of the dataset. Data from Canada became in a different format that was not fulfilling the data structure of the rest of the dataset, so we filtered information from this country out to facilitate the tidying process.
4. As expected, not all information on infection reports was complete for all countries in all days. In those cases we replace with ceros the instances where data was not found or missed (NaN).
5. In the case of the stringency index and socioeconomic datasets, we consider only those countries that were present in the Covid dataset. However not all countries from the stringency index or socioeconomic dataset have missed information for some countries.

In summary, we obtain an enriched dataset that consists of 180 countries with covid cases reported since January the 22nd 2020 until today. From those 180 countries 141 have no information yet on government response or stringency index, 24 countries have no information on the GDP economic indicator, and 3 countries have no information on Income level.

# Data Flow

As we use three different datasets that we will later combine, each one has to be collected from the respective source:

1. Data related to the Covid infections. Once a new case of infection, death or recovery is confirmed from each national medical center, the data is reported to each governmental health institution. The Johns Hopkins Coronavirus Research Center aggregates from many international institutions the national data all around the world and uploads this information in a git repository. This data is publicly available and is updated on a daily basis. For a detailed list of source institutions where Johns Hopkins University obtained the data from, please look at the readme file in their repository (1). The dataset can be then obtained to any local system using our python program.
2. Data related to the Covid government policy indicator. Data is collected from publicly available sources such as news, articles and government press releases and briefings. These are identified via internet searches by a team from the Oxford University. This data after being standardized and organized by a systematic method that they defined (2), is uploaded to a git t repository and updated on a daily basis. The dataset can then be obtained to any local system using our python program.
3. Data related to the socioeconomic indicators. Data is collected from each nation by the World Bank and the Organisation for Economic Co-operation and Development (OECD) (see metadata files in the supplementary information for details). The data is publicly available in the World Bank data catalog on development indicators. This data is then accessed programmatically via direct communication with the “Indicators” API and stored locally.

The figure 1 shows a diagram indicating the dataflow followed by our three datasets from their source to the final working station.

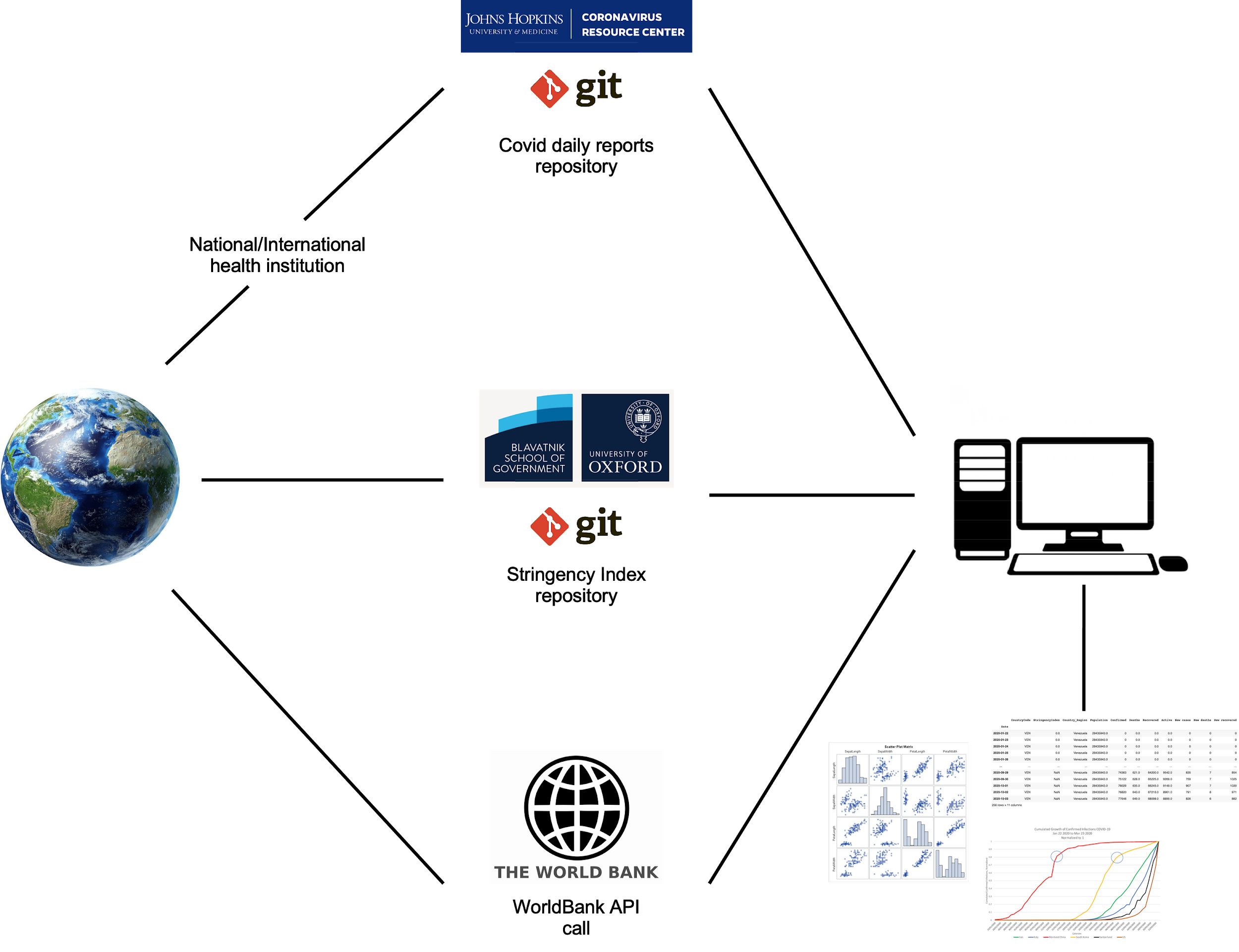


Figure 1. Diagram of data flow from original sources to the final working station. From left to right: data flows from original primary sources and then is made publicly available by being stored in git repositories as is the case for Covid infectious reports and Covid government policy response. The socioeconomic indicators are obtained via API communication request. Thus, the three datasets are stored from any local system for producing the final analysis and visualization.

# Data Model

The data model at the conceptual, logical and physical level in our work flow is specified as described below.

**Conceptual**: We create a new relational database by combining the databases from COVID infections, socioeconomic indicators and stringency index. The new database is an enriched dataframe with relevant information that facilitates our further analysis. The figure 2 shows a diagram of the logical data model.

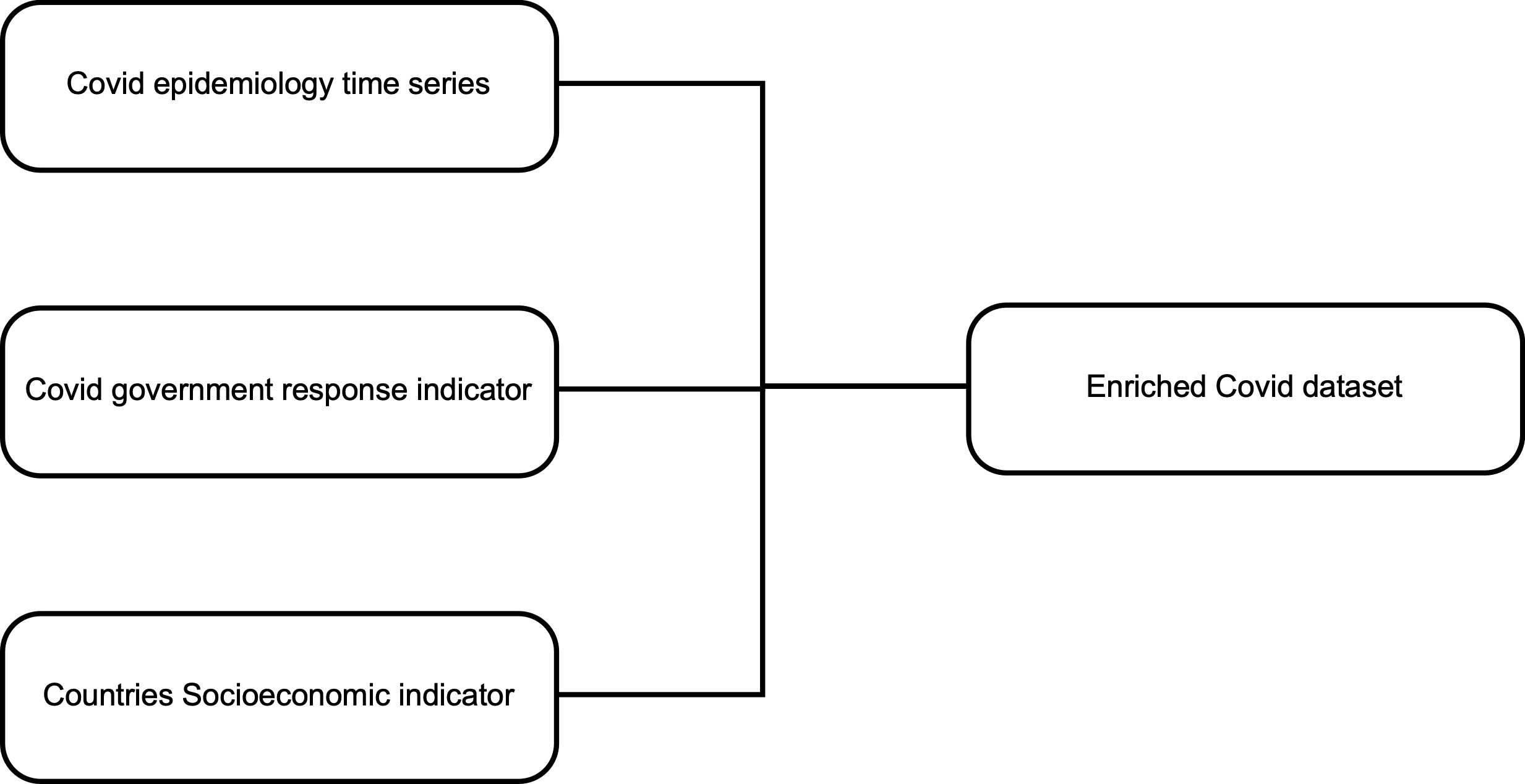


Figure 2. Conceptual data model for our final dataset. We construct our final dataset (right) by combination of three different datasets (left).

**Logical**: We use the ISO country code (“CountryCode”) as our standard variable to identify specifically and unequivocally each country, since country names alone given in each dataset were not equally standard. This key variable was found in the three datasets and used to relate each other and create a new relational database. The “Date” was also an important key variable to join epidemiology covid data with government response data and make a single merged time series. The figure 3 shows a diagram of our logical data model highlighting in bold the key variables used for each dataset to create our new relational dataset.

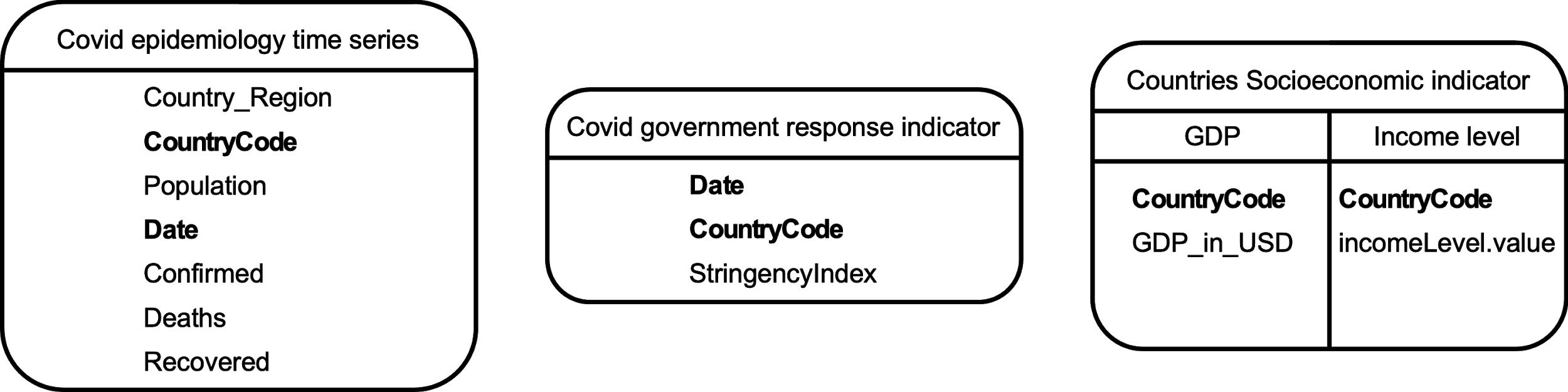


Figure 3. Logical data model for our final dataset. The three boxes represent the primary datasets in the form of dataframes, containing many variables of interest as well as the key variables (bold) which will be used for building our final relational dataset.

**Physical**: Primary data sources live in their respectives repositories or in the case for the World Bank in their own servers. Therefore, we rely on their physical infrastructure (servers, connectivity, availability) to store and get access, at least, to the primary datasets. Thereafter, a computer able to have access to the internet and running the program is the next level required to import the primary data sources, clean and tidy the data, perform the final analysis and make the final visualization. Most modern computers have the minimum capacity in terms of RAM memory, CPU power and disk storage space to perform smoothly all operations required to run the current program. From the repository where our program lives, the size of the script is less than 1 MB and the final dataframe is currently (to the actual date 11.09.2020) 9.9 MB growing at a rate of 44 KB per date as every case is reported. So approximately 10 MB of script + data, when projected to one year of tracking analysis this will require a minimum space of 26 MB of minimus storage in any local system, a negligible amount of storage/memory space from modern computer systems.

# Risks

Physical Level: our program relies heavily on third party data sources. Therefore, the success of running the script and provide the final desired output depends on the availability of these three mentioned data sources. Nevertheless, in the repository where our script is stored, a copy of the raw data sources is saved from the last time the script has been running. This backup of the raw data can be used to recreate the analysis if desired. Since our program is stored in a Git repository, the public access to the program will also depend on the Git website availability. If the link for this repository is down or unavailable for any reason the interested people would have to require a copy of the script via email to the authors.

The script was written in python program language using specific versions and libraries. To avoid problems with running the script in different platforms we also provide a link to run the script via colab, a service from google that allows us to run our Jupyter notebook scripts platform independently. Just a connection to the internet is required.

# Preliminary Studies

The data showed that the most developed countries are not particularly dealing more efficiently with the virus. On the contrary, the top 10 most affected countries (by number of cases) belong to high income countries, which could be potentially explained by their testing capacity (Figure X). Following this hypothesis, the least affected countries (by number of cases) belong in general to low income countries, which have a lower testing capacity.

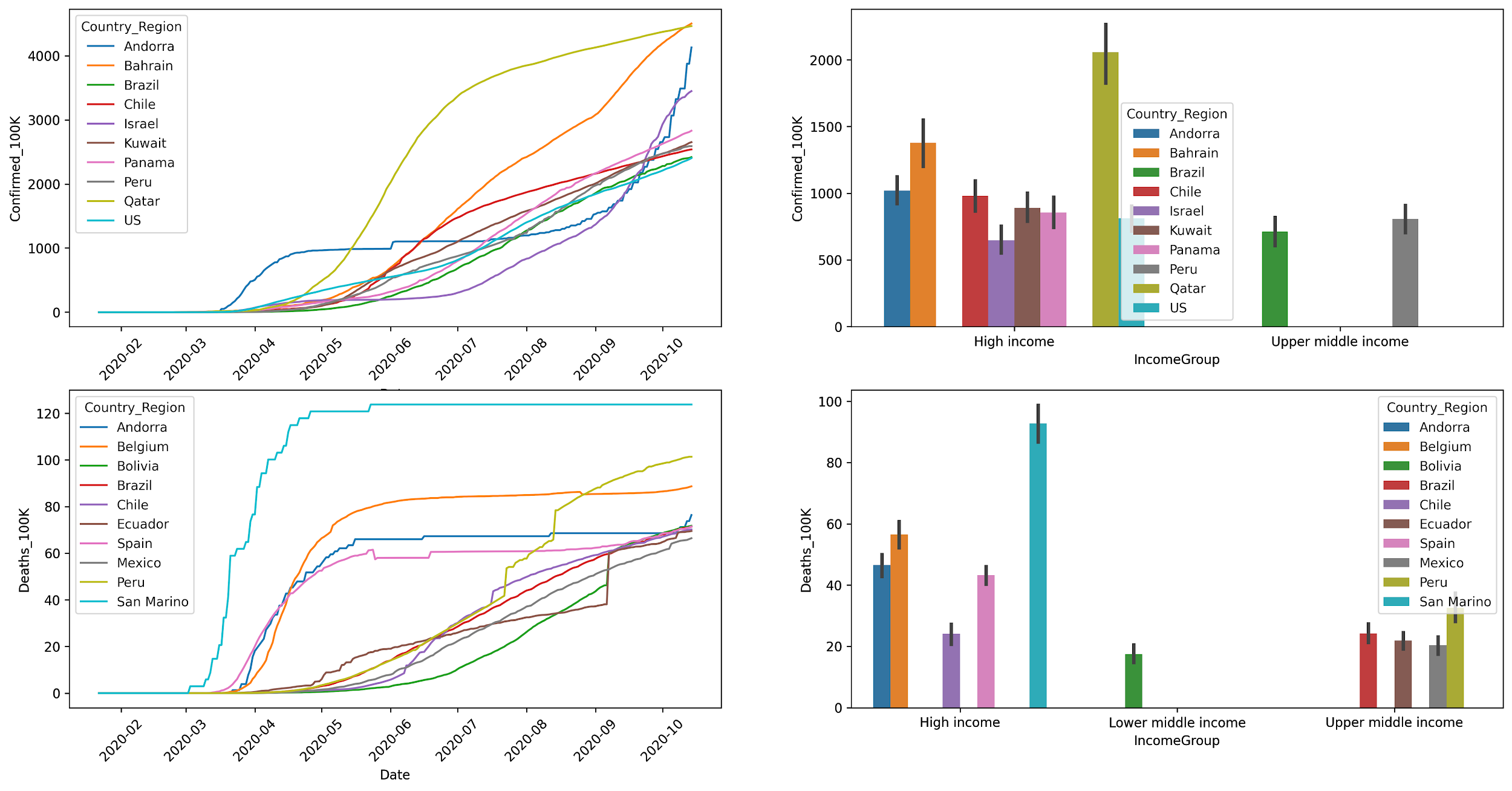


Figure 1. On the right, confirmed cases per 100K and deaths cases per 100K over time for the 10 most affected countries. On the left, income level related to the selected countries.

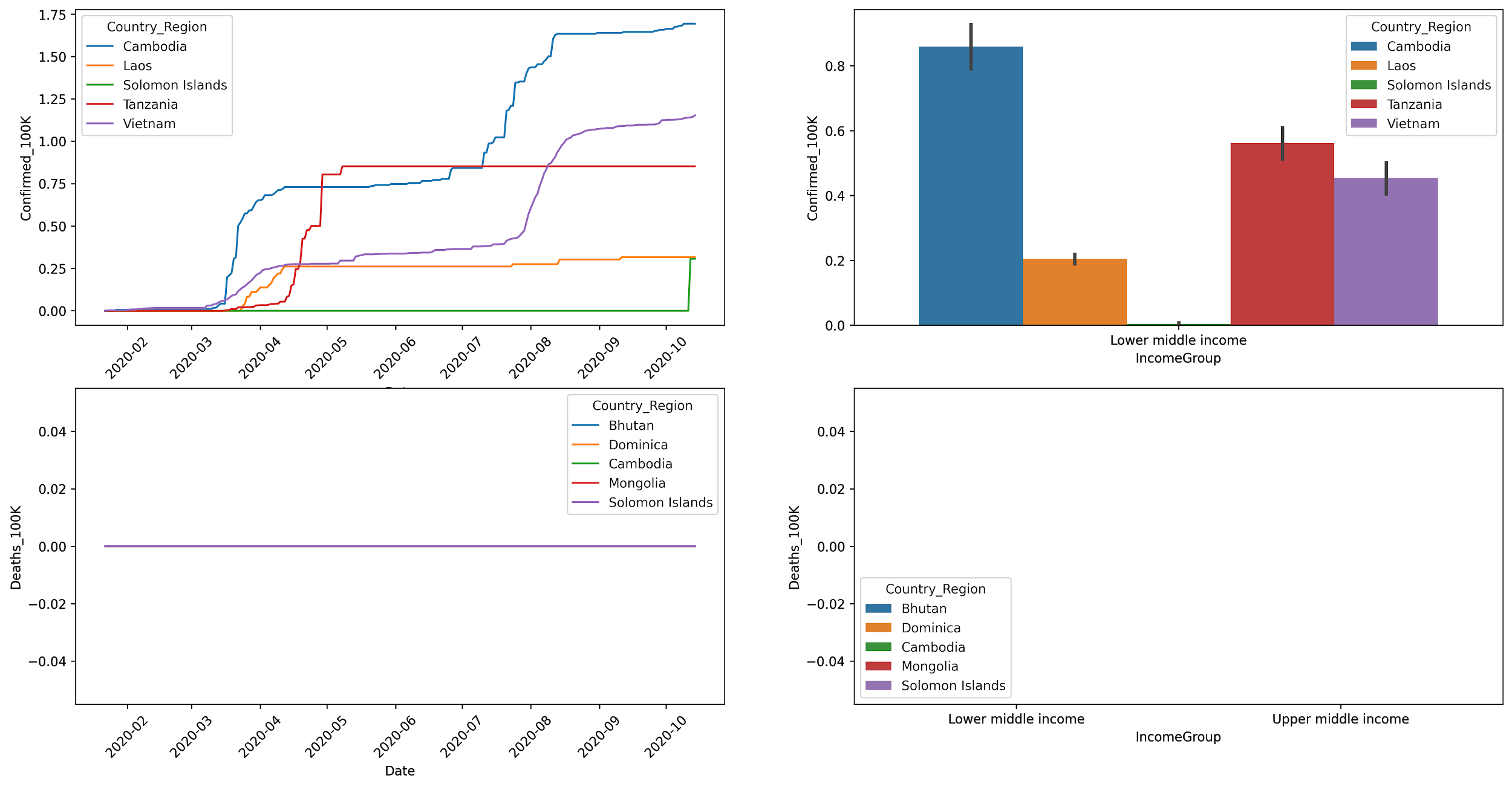


Figure 2. On the right, confirmed cases per 100K and deaths cases per 100K over time for the 10 least affected countries. On the left, income level related to the selected countries.

# Conclusions

In this project, we used recognized data sources for COVID-19 and government strategies. However, these datasets rely on the official sources for each country. This brings challenges for the analysis and the interpretation of the results that should be taken cautiously. There are a large number of factors that could drive the actual crisis. The data showed that the most developed countries are not particularly dealing more efficiently with the virus. On the contrary, the top 10 most affected countries (by number of cases) belong to high income countries, which could be explained by their testing capacity. This hypothesis seems to stand as the least affected countries (by number of cases) belong in general to low income countries. The high income countries that have dealt most properly the crisis vary in a large number of factors, but in general their measurements against the virus were early enough to avoid transmission.

# References and Bibliography

[1] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1

[2] Thomas Hale, Tilbe Atav, Laura Hallas, Beatriz Kira, Toby Phillips, Anna Petherick, Annalena Pott. Variation in US states’ responses to COVID-19. Blavatnik School of Government.

[3] World Bank data on world development indicators. 2020. <https://data.worldbank.org/>

# Suuplementary information

The git repository for this study including scripts and backups datasets, figures generated, metadata and other can be accessed in the following url: <https://github.com/rjlopez2/ADS_CAS_Bern_2020>