Smart Analysis in Big Data Systems

COVID-19's Influence on Global Socio-Economic Landscape

MEI - 2023/2024

Francisco Claudino
PG50380

Afonso Bessa
PG53597

University of Minho

Contents

1 Introduction **7**

2 Definition **8**

2.1 Gross Domestic Product 8

2.2 Inflation 8

2.3 Migration 9

2.4 Population 9

2.5 Taxes 9

2.6 Unemployment 9

2.7 G8 Countries 10

2.8 G20 Countries 10

3 Data Breakdown **10**

3.1 Search Criteria 11

3.2 Selected Datasets 11

4 Methods **12**

4.1 Architecture 12

4.2 Pre-Processing 13

4.2.1 COVID-19 13

4.2.2 Gross Domestic Product 15

4.2.3 Inflation 16

4.2.4 Migration 17

4.2.5 Population 18

4.2.6 Tax 19

4.2.7 Unemployment . . . 21
4.2.8 Merge Process . . . 22
4.3 Pandas Vs Spark . . . 23
4.4 MongoDB . . . 24
4.5 PowerBI . . . 25

5 Results **26**

5.1 Variable . . . 26
5.1.1 Countries . . . 27
5.1.2 Covid Cases . . . 27
5.1.3 Gross Domestic Product . . . 28
5.1.4 Inflation . . . 29
5.1.5 Migration . . . 30
5.1.6 Population . . . 31
5.1.7 Taxes . . . 32
5.1.8 Unemployment . . . 33
5.2 G8 . . . 35
5.2.1 G8 vs CovidCases . . . 36
5.2.2 G8 vs Gross Domestic Product . . . 37
5.2.3 G8 vs Inflation . . . 38
5.2.4 G8 vs Migration . . . 39
5.2.5 G8 vs Population . . . 40
5.2.6 G8 vs Taxes . . . 41
5.2.7 G8 vs Unemployment . . . 43
5.3 Continents . . . 44
5.3.1 Continents vs CovidCases . . . 45

5.3.2 Continents vs Gross Domestic Product 46
5.3.3 Continents vs Inflation 47
5.3.4 Continents vs Migration 48
5.3.5 Continents vs Population 49
5.3.6 Continents vs Taxes 49
5.3.7 Continents vs Unemployment 50
5.4 Queries 52

6 Discussion **53**
6.1 Variable Results 53
6.2 G8 Results 54
6.3 G20 Results 56
6.4 Continents Results 56

7 Conclusion and Critial Analysis **58**

List of Figures

1 Dataset Files 11

2 Architecture 13

3 Dashboard Categorical Variables 28

4 Dashboard CovidCases Variable 29

5 Dashboard GDP Variables 30

6 Dashboard Inflation Variable 31

7 Dashboard Migration Variable 32

8 Dashboard Population Variable 33

9 Dashboard Taxes Variable 34

10 Dashboard Unemployment Variable 35

11 Dashboard G8 Covid Variable 37

12 Dashboard G8 GDP Variable 38

13 Dashboard G8 Inflation Variable 39

14 Dashboard G8 Migration Variable 41

15 Dashboard G8 Population Variable 42

16 Dashboard G8 Taxes Variable 43

17 Dashboard G8 Unemployment Variable 44

18 Dashboard Continent Numerical Variables 45

19 Dashboard Continent CovidCases Variables 46

20 Dashboard Continent GDP Variables 47

21 Dashboard Continent Inflation Variables 48

22 Dashboard Continent Migration Variables 49

23 Dashboard Continent Population Variables 50

24 Dashboard Continent Taxes Variables 51

25 Dashboard Continent Unemployment Variables 51
26 Dashboard Globally Variable Comparision 54
27 Dashboard G8 Variable Comparision 55
28 Dashboard G20 Variable Comparision 56
29 Dashboard Continents Variable Comparision 58

1 Introduction

COVID-19, also known as the coronavirus disease, is a highly contagious respiratory illness caused by the severe acute respiratory syndrome coronavirus 2 (*SARS-CoV-2*). It was first identified in December 2019 in Wuhan, China, and rapidly spread worldwide, leading to a global pandemic declared by the World Health Organization (*WHO*) in March 2020.

The disease is characterized by a wide range of symptoms, which can vary from mild to severe and include fever, cough, shortness of breath, fatigue, muscle aches, headache, loss of taste and smell, and in more severe cases, pneumonia, acute respiratory distress syndrome, organ failure, and death.

COVID-19 is primarily transmitted through respiratory droplets when an infected person coughs, sneezes, or talks, and it can also be spread by touching contaminated surfaces and then touching the face. The spread of the virus has been exacerbated by globalization and high global connectivity, posing a significant challenge to healthcare systems and public health authorities worldwide.

Since the onset of the pandemic, various measures have been implemented to control the virus's spread, including social distancing, wearing face masks, frequent handwashing, and mass vaccination. However, the emergence of virus variants and challenges in vaccine distribution and acceptance continue to influence the trajectory of the pandemic. The *SARS-CoV-2* pandemic has led to significant challenges across various aspects of society, including public health, economic instability, disruptions in education, limitations on travel, and profound impacts on people's lifestyles worldwide. These challenges have been compounded by varying responses to the pandemic, influenced by differing policies and capacities among countries, leading to a multitude of adverse outcomes and experiences.

The pandemic's influence on economic performance is profound, affecting global supply chains, consumer demand, and investment flows. Inflation rates have fluctuated due to disrupted production and changes in consumer behavior. Migration patterns have also been altered as travel restrictions and economic uncertainty impact population movements. Population growth rates have been influenced by changes in mortality and birth rates. Tax revenues have been affected by economic slowdowns, impacting government budgets and public services. Unemployment rates have risen in many regions due to lockdowns and reduced economic activities, exacerbating social inequalities and poverty levels.

The main objective of this report is to analyze the impacts of the pandemic on various socio-economic factors, including *Gross Domestic Product* (**GDP**), *Inflation*, *Migration*, *Population*, *Taxes*, and *Unemployment*, on a global scale. This study aims to provide a comprehensive overview of how different countries have been affected by the pandemic, highlighting the variations in socio-economic impacts across regions and economies. Understanding these impacts is crucial for policymakers, economists, and

researchers to develop strategies that can mitigate the adverse effects and promote recovery.

By examining these socio-economic factors, the report seeks to draw meaningful comparisons and insights into the varying effects of the pandemic. The study's findings will contribute to a deeper understanding of the pandemic's long-term consequences and inform future research and policy development aimed at fostering economic resilience and social well-being in the face of global crises.

2 Definition

In this section, we define and discuss the key indicators relevant to our study: **Gross Domestic Product**, **Inflation**, **Migration**, **Population**, **Taxes**, and **Unemployment**. These measures are crucial for understanding the comprehensive impact of the COVID-19 pandemic across different regions, providing a foundation for our analysis of their variations and the ensuing socio-economic consequences.

2.1 Gross Domestic Product

Gross Domestic Product (**GDP**) is a measure of the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period [1]. **GDP** serves as a broad indicator of a country's overall economic health. It encompasses all private and public consumption, government outlays, investments, and the foreign balance of trade, which significantly impacts GDP by increasing it during trade surpluses and decreasing it during trade deficits. Real **GDP**, adjusted for inflation, offers a more accurate reflection of a country's economic performance over time [2].

2.2 Inflation

Inflation refers to the rate at which the general level of prices for goods and services rises, eroding the purchasing power of money. It is expressed as a percentage and signifies that a unit of currency buys fewer goods and services than before [3]. Inflation impacts the cost of living and can decelerate economic growth if it outpaces productivity. Central banks often manage inflation through monetary policies to maintain economic stability. Understanding inflation is crucial as it affects consumer behavior, savings, and investments. [4]

2.3 Migration

Migration is the movement of people across international borders or within a country for periods exceeding one year. It includes both voluntary and involuntary migration, influenced by various factors such as economic opportunities, conflicts, and environmental changes. Migration significantly affects socio-economic measures, impacting labor markets, demographic dynamics, **GDP**, tax revenues, and unemployment rates. Analyzing migration patterns helps understand shifts in population and workforce distribution. [5]

2.4 Population

Population represents the total number of people living in a specific area, including all nationals and permanently settled non-nationals. Population data is crucial for understanding demographic trends, planning public services, and assessing economic development. Governments typically use censuses to measure population size, which informs policy decisions and resource allocation. Population growth rates, resulting from births, deaths, and net migration, provide insights into demographic changes over time. [6]

2.5 Taxes

Taxes are mandatory financial charges imposed by governments on individuals and corporations to fund public expenditures. They are essential for financing public services such as healthcare, education, infrastructure, and social security [7].Taxes influence economic activities by determining the financial burden on businesses and consumers. Understanding tax structures and revenues helps analyze government fiscal policies and their impact on economic stability and growth. [8]

2.6 Unemployment

Unemployment refers to individuals who are capable of working, actively seeking work, but are unable to find employment. The unemployment rate, calculated as the percentage of the labor force that is unemployed, is a critical indicator of labor market health. High unemployment rates can indicate economic distress, whereas low rates suggest a robust economy. Unemployment impacts economic productivity, consumer spending, and overall economic health. [9]

These features are crucial to our study, providing the framework to analyze the socio-economic impacts of the **COVID-19** pandemic. Understanding these concepts allows us to examine the pandemic's effects on economic performance, demographics,

labor markets, and fiscal policies across regions, enabling meaningful comparisons and global insights.

2.7 G8 Countries

The **Group of Eight (G8)** is an intergovernmental political forum comprising eight of the world's major advanced economies: the United States, the United Kingdom, France, Germany, Italy, Canada, Japan, and Russia. Established to foster dialogue and coordinate policies on economic and political issues of mutual concern, the *G8* plays a significant role in global governance. The G8's primary activities revolve around annual summits where leaders gather to discuss and strategize on global economic policies, trade, climate change, security, and other pressing international issues. These summits provide a platform for the world's leading economies to address common challenges, seek consensus, and promote cooperative solutions.

2.8 G20 Countries

The **Group of Twenty (*G20*)**, established in 1999, is an international forum of 19 countries and the European Union, representing around 85% of global **GDP**, over 75% of international trade, and about two-thirds of the world's population. Initially focused on economic issues, the G20 has expanded to address climate change, global health, and sustainable development. Leaders meet annually at summits, supplemented by finance ministers, central bank governors, and working group meetings, to foster international cooperation and address global economic stability and growth. The G20's comprehensive approach ensures that a wide range of global issues are addressed, making it a critical player in international economic governance.

3 Data Breakdown

In this section, we provide a comprehensive overview of the datasets essential for analyzing socio-economic dynamics, particularly during the period from 2010 to 2023, which includes the COVID-19 pandemic period. The datasets encompass a range of socio-economic indicators critical for understanding the pandemic's impact.

As depicted in Figure 1, a selection of seven datasets was made, comprising six in *CSV* format and one in *XLSX* format. These datasets were chosen based on specific criteria to ensure their relevance, comprehensiveness, and data quality.

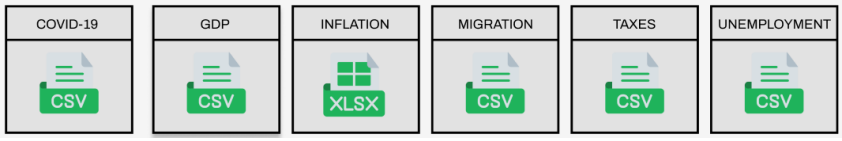

Figure 1 Dataset Files

3.1 Search Criteria

The selection of datasets for this study was guided by several key criteria to ensure comprehensiveness, accuracy, and relevance:

1. **Relevance to the Problem Domain:** The datasets directly address the research questions, offering insights into the socio-economic impacts of the COVID-19 pandemic.
2. **Size and Variety:** The datasets are extensive, qualifying as *Big Data*, and provide diverse insights at scale.
3. **Data Quality:** High consistency and low levels of missing data were prioritized to ensure reliable and accurate analysis.

3.2 Selected Datasets

Below is a detailed description of the selected datasets, emphasizing their relevance and the type of information they contain:

COVID-19: This dataset offers a comprehensive overview of the COVID-19 pandemic, including critical metrics such as confirmed cases, deaths, recoveries, and other pertinent data across the globe. It is sourced from Our World in Data, providing reliable and up-to-date information essential for analyzing the pandemic's trajectory and impact. [10]

Gross Domestic Product: This dataset includes GDP data for various countries from 1980 to 2028, with future projections. It allows for the analysis of global economic trends, cross-country comparisons, and understanding economic growth patterns. Sourced from the International Monetary Fund (IMF), it ensures the reliability and comprehensiveness needed for economic research. [11]

Inflation: This dataset provides the annual inflation rate based on the average consumer price index for various countries from 1980 to 2028. It includes data on the percentage change in inflation, which is crucial for understanding economic stability and purchasing power over time. [12]

Migration: This dataset contains net migration data for various countries and regions from 2000 to 2025. Net migration is calculated as the number of immigrants minus the number of emigrants, providing insights into population changes and demographic shifts. [13]

Tax: This dataset includes tax revenue as a percentage of GDP for various countries from 1960 to 2022. It provides information on government revenue generation and fiscal policies, which are essential for understanding economic policies and their impacts. [14]

Population: This dataset offers total population figures for a range of countries and regions from 2000 to 2024. It is crucial for demographic analysis and understanding population growth trends. [15]

Unemployment: This dataset provides unemployment rates as a percentage of the total labor force for various countries and regions from 1960 to 2022. It includes data on the total number of unemployed individuals, which is vital for analyzing labor market conditions and economic health. [16]

These datasets form the backbone of our study, providing the necessary data to explore the socio-economic impacts of the **COVID-19** pandemic comprehensively. By analyzing these diverse indicators, we aim to understand the pandemic's effects on global economic performance, demographic changes, and fiscal policies across different regions.

4 Methods

This section details the architecture of the implemented system, beginning with the selection of relevant columns from datasets to align with research objectives. The preprocessing phase addresses missing values to ensure data integrity, followed by merging datasets to create a unified analytical framework. A comparative analysis of **Pandas** and **Apache Spark** highlights their respective efficiencies in handling large datasets and complex transformations. The database used for storing the data is **MongoDB**, a scalable and flexible database. Finally, the section explains the data migration process from pre-processed datasets to **MongoDB**, and then to **Microsoft PowerBI** for advanced visualization and analysis, ensuring a robust and efficient data pipeline.

4.1 Architecture

The architecture utilized in this project encompasses all the datasets described in the previous section. Initially, each dataset is pre-processed individually, utilising the Pandas library. This step includes data cleaning, handling missing values, changing the

format of the dataset to better encapsulate the data and normalizing the data formats. Following the individual pre-processing, the datasets are merged into a unified dataset to facilitate comprehensive analysis. Subsequently, the consolidated data is stored in a MongoDB database, chosen for its flexibility and scalability in managing large and diverse datasets. The final step involves exporting the data from **MongoDB** to **Microsoft PowerBI**. This enables the creation of dynamic dashboards and visualizations, which aid in the detailed analysis and presentation of the data insights. Figure 2 provides a visual representation of the described architecture, illustrating the workflow from data extraction and to pre-processing to visualization.

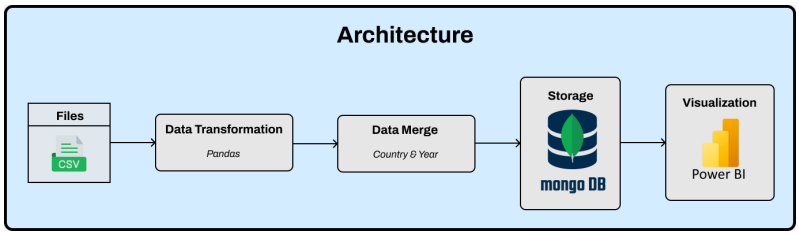

Figure 2 Architecture

4.2 Pre-Processing

Pre-Processing is a crucial step in this project as it ensures data quality and consistency. The data is cleaned, transformed and prepared for analysis. Each variable undergoes specific preprocessing steps, outlined below:

4.2.1 COVID-19

The pre-processing of the **COVID-19** dataset involves several systematic steps to transform the raw data into a structured and clean format suitable for analysis. This dataset includes various metrics related to the **COVID-19** pandemic such as:

1. **date:** The date of the recorded data (in YYYY-MM-DD format), between *2020-01-05* to *2024-02-04*.
2. **country:** The name of the country.
3. **confirmed_cases:** The cumulative number of confirmed COVID-19 cases.
4. **deaths:** The cumulative number of deaths due to COVID-19.
5. **recovered:** The cumulative number of recoveries from COVID-19.
6. **active_cases:** The number of active COVID-19 cases.
7. **new_cases:** The number of new COVID-19 cases reported on that date.

8. **new_deaths:** The number of new deaths due to COVID-19 reported on that date.
9. **new_recoveries:** The number of new recoveries from COVID-19 reported on that date.

The Pre-Processing of the **COVID-19** dataset involves several key steps aimed at cleaning and structuring the data for analysis. Initially, the data is loaded into a DataFrame from a *CSV* file, with a timer set to measure the total preprocessing time.

The first step is the selection of relevant columns from the dataset. Specifically, the columns for `location`, `date` and `total cases` are extracted (which are the columns with the relevant information for the study we are conducting) to form a new DataFrame. This focuses the analysis on essential information, eliminating unnecessary data.

Next, the date column is converted into a datetime format. This conversion facilitates time-based operations and ensures that the data can be grouped and manipulated accurately. Following this, the dataset is grouped by location and year. Within each group, the last record of the year is selected. This step is crucial as it captures the cumulative total cases at the end of each year for each location, providing a clear annual snapshot of the pandemic's progression.

Subsequently, a new column is created to store the year extracted from the date column. The date column is then dropped as it is no longer needed. The columns are renamed to more intuitive names, such as changing `location` to `country` and `total_cases` to `covid_cases`, enhancing the readability of the dataset.

To ensure consistency and accuracy, the country names are standardized. Some country names in the dataset are replaced with their official or more commonly accepted versions, such as *Timor* to *Timor-Leste*. This step addresses any inconsistencies and discrepancies in the naming conventions, which is essential for reliable data analysis.

The Pre-Processing also involves removing aggregated entries and non-country entities from the dataset. These include continents, income groups and the world as a whole. By focusing solely on individual countries, the dataset becomes more relevant for country-specific analysis.

Finally, the cleaned and processed dataset is saved to a new *CSV* file. This file serves as the final output, ready for further analysis or reporting. The total execution time of the preprocessing steps is measured and printed, providing insight into the efficiency of the process.

4.2.2 Gross Domestic Product

The Pre-Processing of the GDP dataset involves several systematic steps to transform the raw data into a structured and clean format, suitable for analysis.

The raw *CSV* dataset contains columns representing the GDP values for various countries from 1980 to 2028. Each row corresponds to a different country and the columns represent annual GDP values for each year.

Below is a full breakdown of the collumns of this dataset:

1. **Country:** The name of the country or region.
2. **1980, ..., 2028:** These columns represent the GDP values for the corresponding years from 1980 to 2028.

The first task is to reshape the DataFrame. This is achieved by converting the wide format data, where each year is a separate column, into a long format with three columns: `Country`, `Year`, and `Value`. The melt function in Pandas is used for this transformation. By applying the *melt* function, the DataFrame is restructured so that each row represents a single observation with the corresponding country, year, and value. This transformation is essential for managing and analyzing the data more effectively. After reshaping, the DataFrame is sorted by country and then by year within each country to organize the data chronologically. The long format resulting from the *melt* function simplifies the structure of the dataset, making it easier to perform time-series analysis, merge with other datasets, and visualize trends over time.

Next, the columns are renamed to more intuitive names. The column previously labeled as `Value` is renamed to `GDP`, `Country` is renamed to `country` and `Year` is renamed to `year`. This renaming enhances the readability and clarity of the dataset.

Standardizing country names is a crucial step to ensure consistency across the dataset. Various country names are replaced with their official or commonly accepted versions. For instance, *China, People's Republic of* is standardized to *China*, *Micronesia, Fed. States of* to *Micronesia* and *Türkiye, Republic of* to *Turkiye*. This standardization addresses any inconsistencies and discrepancies in the naming conventions.

Furthermore, the dataset includes several aggregated regions or non-country entities, such as continents, economic regions and the world as a whole. These entries are removed to focus the dataset solely on individual countries. Examples of removed entities include *Africa (Region)*, *Emerging and Developing Asia*, *European Union* and *World*.

Finally, the cleaned and processed DataFrame is saved to a new *CSV* file. This output file contains the GDP data in a structured and consistent format, suitable for

further analysis. The total execution time for the preprocessing steps is measured and printed, providing insight into the efficiency of the process.

4.2.3 Inflation

Preprocessing the dataset containing inflation rates for average consumer prices involves systematically transforming raw, wide-format data into a structured, long-format dataset suitable for analysis.

This dataset provides annual inflation rate data for various countries from 1980 to 2028. Each row in the raw data corresponds to a specific country and the columns represent the annual inflation rates for each year.

Below is a full breakdown of the collumns of this dataset:

1. **Country:** The name of the country or region.
2. **1980, ..., 2028:** These columns represent the annual inflation rates for the corresponding years from 1980 to 2028.

The processing begins with loading the data from an Excel file into a pandas DataFrame. The first task is to rename the column containing country names to `country` for clarity and consistency.

This is followed by reshaping the DataFrame from a wide format to a long format using the *melt* function, like the previous Pre-Processement. In the original wide format, each year is a separate column and in the long format, there are three columns: `country`, `year` and `inflation`. This transformation facilitates easier data manipulation and analysis.

After reshaping, the DataFrame is sorted by country and then by year within each country to organize the data chronologically. This ensures that the data for each country is in a sequential order by year, making it more intuitive to analyze.

The standardization of country names followed the same format as the previous dataset pre-process methods.

Next, the dataset includes several aggregated regions or non-country entities, such as continents and economic regions, which are removed to focus solely on individual countries. Rows with missing values in the 'country' column are also removed, following the same processes as in the above preprocessing methods to ensure data integrity.

Finally, the cleaned and processed DataFrame is saved to a new *CSV* file, ensuring that all files have the same format prior to the merging process. This output file contains the inflation rate data in a structured and consistent format, ready for further

analysis. The total execution time for the preprocessing steps is measured and printed, providing insight into the efficiency of the process.

4.2.4 Migration

Pre-Processing the dataset containing net migration data aims to convert the raw, unstructured data into a well-organized format suitable for comprehensive analysis.

This dataset offers detailed annual net migration figures for various countries and regions from 2000 to 2025. Each record includes fields like country name, country code, and annual migration numbers, all needing careful transformation to ensure clarity, consistency, and usability.

Below is a full breakdown of the collumns of this dataset:

1. **Series Name:** Indicates the type of data series, which in this case is *Net migration*.
2. **Series Code:** A unique code for the data series.
3. **Country Name:** The name of the country or region.
4. **Country Code:** A unique code for the country or region.
5. **2000 [YR2000], ..., 2025 [YR2025]:** These columns represent the annual net migration numbers for the corresponding years from 2000 to 2025. The values are the net migration numbers for each year.

To begin with, the dataset is loaded from a *CSV* file into a pandas DataFrame. This initial step involves reading the data from its raw form into a structured data frame, setting up the foundation for further manipulations. Several columns that are not essential for the analysis, such as the series name, series code and country code, are removed to streamline the dataset and focus only on the relevant data.

The columns containing year-specific data are then renamed for simplicity. The original format, which included brackets around the year labels, is modified to a straightforward year-only format. This renaming enhances the readability and usability of the data.

Following the renaming, the DataFrame is reshaped in the same way as the previous datasets, following the same initial format and for the exact same purpose as the ones mentioned above.

To ensure chronological organization, the data is sorted by country and then by year within each country. This sorting arranges the records in a logical sequence, facilitating more intuitive analysis and visualization of migration trends over time.

Standardizing country names followed the same process as the ones above, replacing various country names with their official or commonly accepted versions to maintain consistency and accuracy, facilitating easier comparison and analysis.

Following the same process as before, entries for aggregated regions and non-country entities were removed to focus solely on individual countries. This refinement ensures the dataset remains relevant and specific to the analysis of country-level migration data.

Additionally, any rows with missing values in the country column are eliminated. This step is essential to preserve the dataset's integrity and prevent potential issues during analysis.

Finally, the cleaned and processed data is saved to a new *CSV* file. This output file, now structured and consistent, is ready for detailed analysis. The entire preprocessing sequence is timed, providing insights into the efficiency of the process.

4.2.5 Population

Preprocessing the dataset containing total population data is vital for converting raw, unorganized information into a structured format that is ready for comprehensive analysis.

This dataset includes annual total population figures for various countries and regions from 2000 to 2025. Each record initially includes multiple fields such as the country name, country code, series name and series code, along with annual population numbers.

Below is a full breakdown of the collumns of this dataset:

1. **Country Name:** The name of the country or region.
2. **Country Code:** A unique code for the country or region.
3. **Series Name:** Indicates the type of data series, which in this case is 'Population, total'.
4. **Series Code:** A unique code for the data series.
5. **2000 [YR2000], ..., 2025 [YR2025]:** These columns represent the annual total population numbers for the corresponding years from 2000 to 2025. The values are the total population for each year.

To start, the dataset is loaded from a *CSV* file into a pandas DataFrame. This sets up the initial structure for further manipulations. Several columns that are not essential for the analysis, such as the series name, series code and country code, are removed to streamline the dataset and focus only on the relevant data.

The columns containing year-specific data are then renamed for simplicity. The original format, which included brackets around the year labels, is modified to a straightforward year-only format. This renaming enhances the readability and usability of the data.

Following the renaming, the DataFrame is reshaped following the same reshape process as previsouly mentioned. As a result, the dataset now comprises three primary columns: country, year, and population. This long format is more conducive to time-series analysis and easier to manipulate for various analytical tasks.

To ensure chronological organization, the data is sorted by country and then by year within each country. This sorting arranges the records in a logical sequence, facilitating more intuitive analysis and visualization of population trends over time.

The standardization process of country names is taken in a similar way as in the datasets described above. This standardization addresses discrepancies and ensures that the data remains consistent and accurate, making it easier to compare and analyze.

The dataset also contains entries for aggregated regions and non-country entities, which are not necessary for the intended analysis. These entries are removed to focus solely on individual countries. This refinement further streamlines the dataset, ensuring that it remains relevant and specific to the analysis of country-level population data.

Additionally, missing values are eliminated in the same way as the previous mentioned datasets for the same reasons.

Finally, the cleaned and processed data is saved to a new *CSV* file. This output file, now structured and consistent, is ready for detailed analysis. The entire preprocessing sequence is timed, providing insights into the efficiency of the process.

4.2.6 Tax

Transforming the dataset containing tax revenue data is crucial to convert raw, unstructured information into a well-organized format, making it suitable for detailed analysis.

This dataset includes annual tax revenue as a percentage of GDP for various countries and regions from 1960 to 2022. Each record initially includes multiple fields such as the country name, country code, indicator name, and indicator code, along with annual tax revenue percentages

Below is a full breakdown of the collumns of this dataset:

1. **Country Name:** The name of the country or region.
2. **Country Code:** A unique code for the country or region.
3. **Indicator Name:** The type of data series, which in this case is 'Tax revenue (% of GDP)'.
4. **Indicator Code:** A unique code for the data series.
5. **1960, ..., 2022:** These columns represent the annual tax revenue as a percentage of GDP for the corresponding years from 1960 to 2022. The values are the tax revenue percentages for each year.

Initially, the dataset is loaded from a *CSV* file into a pandas DataFrame. This step establishes the foundation for further manipulations. Several columns that are not essential for the analysis, such as the country code, indicator code, and indicator name, are removed to streamline the dataset and focus only on the relevant data.

The column containing country names is renamed to 'country' to enhance clarity and consistency. This renaming makes the dataset more intuitive to work with and aligns with common data conventions.

Following the renaming, the DataFrame is reshaped. This process is similar to the ones described above, as the format of the datasets is identical. This long format is more conducive to time-series analysis and easier to manipulate for various analytical tasks.

To ensure chronological organization, the data is sorted by country and then by year within each country. This sorting arranges the records in a logical sequence, facilitating more intuitive analysis and visualization of tax revenue trends over time.

Rows with missing values in the tax column are removed to ensure the dataset's integrity and prevent potential issues during analysis.

Standardizing country names was done in the same way as in the previous datasets. This standardization addresses discrepancies, ensuring the data remains consistent and accurate, making it easier to compare and analyze.

The dataset also contains entries for aggregated regions and non-country entities, which are not necessary for the intended analysis, thus being removed.

Then, we remove the missing values as done in the previous Pre-Processement methods for the other datasets.

Finally, the cleaned and processed data is saved to a new *CSV* file. This output file, now structured and consistent, is ready for detailed analysis. The entire preprocessing sequence is timed, providing insights into the efficiency of the process.

4.2.7 Unemployment

Preprocessing the dataset containing unemployment data is crucial for converting raw, unstructured information into a well-organized format that is suitable for detailed analysis.

This dataset includes annual unemployment rates as a percentage of the total labor force for various countries and regions from 1960 to 2022. Each record initially includes multiple fields such as the country name, country code, indicator name, and indicator code, along with annual unemployment percentages.

Below is a full breakdown of the collumns of this dataset:

1. **Country Name:** The name of the country or region.
2. **Country Code:** A unique code for the country or region.
3. **Indicator Name:** The type of data series, which in this case is 'Unemployment total (% of total labor force) (modeled ILO estimate)'.
4. **Indicator Code:** A unique code for the data series.
5. **1960, ..., 2022:** These columns represent the annual unemployment rates as a percentage of the total labor force for the corresponding years from 1960 to 2022. The values are the unemployment rates for each year.

Initially, the dataset is loaded from a *CSV* file into a pandas DataFrame, establishing the foundation for further manipulations. Several columns that are not essential for the analysis, such as the country code, indicator code, and indicator name, are removed to streamline the dataset and focus only on the relevant data.

The column containing country names is renamed to 'country' to enhance clarity and consistency. This renaming makes the dataset more intuitive to work with and aligns with common data conventions.

Following the renaming, the DataFrame is reshaped, with th reshaping process applied in the same way as the datasets prior to it, for the same purposes.

To ensure chronological organization, the data is sorted by country and then by year within each country. This sorting arranges the records in a logical sequence, facilitating more intuitive analysis and visualization of unemployment trends over time.

Rows with missing values in the unemployment column are removed to ensure the dataset's integrity and prevent potential issues during analysis.

Standardization methods take place in this dataset in the same way as in the previous ones.

The dataset also contains entries for aggregated regions and non-country entities, which are not necessary for the intended analysis. These entries are removed to focus solely on individual countries. This refinement further streamlines the dataset, ensuring that it remains relevant and specific to the analysis of country-level unemployment data.

Additionally, any rows with missing values in the country column are eliminated.

Finally, the cleaned and processed data is saved to a new *CSV* file. This output file, now structured and consistent, is ready for detailed analysis. The entire preprocessing sequence is timed, providing insights into the efficiency of the process.

4.2.8 Merge Process

After each dataset has been Pre-Processed, as described in the subsections above, the merge process takes place to obtain a single dataset witha ll the information about all the variables described.

The process begins by reading the preprocessed *CSV* files into separate pandas DataFrames. These files include data on inflation, **COVID-19** cases, tax revenue, migration, population, **GDP**, and unemployment. Each dataset is loaded into a DataFrame using the pd.read_csv function, which sets up the initial structure for further manipulation.

Once the data is loaded, the data types of the 'country' and 'year' columns in each DataFrame are explicitly converted to strings and integers, respectively. This conversion ensures consistency across the datasets and prepares them for merging.

The data is then filtered to include only the years between 2010 and 2023. This step narrows down the datasets to a specific timeframe, ensuring that only relevant data is included in the final merged dataset.

Next, the individual DataFrames are merged into a single DataFrame using a series of outer joins on the 'country' and 'year' columns. This merging process combines all the datasets while preserving all data points from each DataFrame. The merged DataFrame is then sorted by country and year to organize the data chronologically within each country.

Certain entries, corresponding to non-country entities or aggregated regions, are removed from the merged dataset. This step ensures that the dataset remains focused on individual countries, enhancing the relevance and accuracy of the analysis. Additionally, any remaining entries with missing values in the 'country' column are dropped to maintain the dataset's integrity.

Some country names are standardized to ensure consistency across the dataset. For example, 'Korea' is replaced with 'North Korea' and 'Pacific island small states'

is replaced with 'Pacific Islands'. This standardization addresses discrepancies and ensures that the data remains uniform and accurate.

Any instances of 'no data' within the dataset are replaced with NaN values to handle missing data appropriately during analysis. This step is crucial for maintaining the dataset's quality and preventing errors in subsequent analyses.

The final merged DataFrame is then sorted again by country and year to ensure proper organization. Finally, the cleaned and merged dataset is saved to a new *CSV* file named 'FinalFilePP.csv'. This output file provides a comprehensive and structured view of the socioeconomic data, ready for detailed analysis.

4.3 Pandas Vs Spark

The Pre-Processing of multiple datasets was approached using two different tools: the **Pandas Library** and the **Apache Spark Framework**. This analysis aimed to evaluate the efficiency and execution times of these tools when consolidating and processing multiple data sources.

Pandas is a powerful data manipulation and analysis library for Python. It offers data structures and functions needed to manipulate structured data seamlessly. Pandas is well-suited for handling small to moderately sized datasets in-memory, providing an intuitive interface and efficient performance for data preprocessing tasks.

Apache Spark is a unified analytics engine for large-scale data processing. It utilizes distributed computing to handle vast datasets across clusters of computers, offering speed and scalability. Spark is designed to perform both batch and streaming data processing efficiently and can handle large-scale data preprocessing tasks.

To compare the execution times between **Pandas** and **Spark**, the pre-processing and merging steps of the datasets were consolidated into a single notebook called *PreProcessingAllInOne*. This approach was chosen to address the inefficiencies observed when initializing the Spark session multiple times in individual notebooks. The repeated initialization was time-consuming, impacting the overall performance.

By merging the preprocessing steps into one file, the session initialization occurs only once, significantly reducing the overhead. Additionally, handling all preprocessing steps in a single file ensures that data is read once at the beginning and written once at the end, maintaining all datasets in memory. This method is more efficient as it minimizes *I/O* operations, allowing for streamlined data manipulation.

After running the consolidated preprocessing notebook using both **Pandas** and **Spark**, the results were as follows:

In this particular case, the **Pandas** library demonstrated significantly better performance compared to the **Apache Spark** framework (1.3 seconds compared to 14.7

Tool	Time (seconds)
Apache Spark	14.703
Pandas	1.314

Table 1 Tool Execution Time Comparison

seconds). The execution time for Pandas was substantially lower, with a significant diference of 13.4 seconds, making it the more efficient option for the data preprocessing needs, and thus the one chosen.

4.4 MongoDB

MongoDB is a NoSQL database known for its flexibility, scalability, and performance. Unlike traditional relational databases, MongoDB stores data in JSON-like documents, which allows for dynamic schemas and easy data integration. This characteristic makes it highly suitable for handling unstructured and semi-structured data, providing an efficient way to store and query large volumes of data.

MongoDB's document-oriented storage, schema flexibility, and support for rich queries make it a robust choice for modern data applications. Its distributed nature ensures high availability and scalability, which are critical for managing extensive datasets like the one used in this project. Furthermore, **MongoDB**'s ability to index data efficiently enables fast query performance, making it ideal for real-time analytics and data retrieval.

MongoDB was chosen for this project due to the reasons mentioned above. Its schema-less nature allows for the easy accommodation of evolving data structures without requiring significant modifications, which aligns perfectly with the diverse and dynamic nature of our dataset. **MongoDB** supports horizontal scaling, distributing data across multiple servers to handle large volumes of data efficiently, ensuring high availability and performance for both read and write operations. Additionally, its powerful querying capabilities enable complex data retrieval and analysis, making it ideal for deriving insights from our data. The JSON-like document structure of **MongoDB** simplifies data integration and manipulation, creating a seamless data pipeline from preprocessing to storage.

When comparing **MongoDB** to **Cassandra**, both are NoSQL databases designed for high availability and scalability. However, **MongoDB**'s flexible document structure is more adaptable to changing data schemas compared to Cassandra's more rigid data model. This flexibility is crucial for our dataset, which requires handling evolving and diverse data types. While Cassandra excels in scenarios requiring high write throughput and can handle large-scale data across multiple data centers, it lacks the same level of querying capabilities that **MongoDB** offers. Thus, for applications needing

dynamic schema adjustments and complex queries, **MongoDB** is the more suitable choice.

In contrast to **Hadoop**, which is a framework for distributed storage and processing of large datasets using the MapReduce programming model, **MongoDB** is optimized for real-time queries and analytics. Hadoop excels at batch processing and handling massive datasets but requires a more complex setup and maintenance. MongoDB provides a more straightforward approach to database management, making it easier to use and integrate into our existing workflow. Its ability to perform high-speed data retrieval and complex queries without the overhead of Hadoop's ecosystem makes **MongoDB** the better option for our data, which demands real-time analytics and flexible data manipulation.

The data extraction script used for the data extraction to **MongoDB** connects to a local **MongoDB** server and specifies the target database and collection. It reads the merged *CSV* file containing the preprocessed data into a pandas DataFrame and converts the DataFrame into a list of dictionaries, aligning with **MongoDB**'s document-oriented structure. Finally, the script inserts these dictionary records into the specified MongoDB collection using the insert_many method. This process ensures an efficient and seamless transfer of data from the preprocessing stage into **MongoDB**, facilitating further analysis and querying.

4.5 PowerBI

PowerBI is a powerful business analytics tool developed by **Microsoft** that enables users to visualize and share insights from their data. It provides interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards. **PowerBI** connects to a wide range of data sources, making it a versatile tool for data analysis and visualization.

The advantages of using **PowerBI** for data visualization are numerous. It offers a comprehensive suite of visualization tools that allow for the creation of detailed and interactive dashboards. These visualizations help in understanding complex datasets by presenting data in a graphical format, making it easier to identify trends, patterns, and outliers. **PowerBI**'s ability to integrate with various data sources, including databases, cloud services, and spreadsheets, ensures that users can consolidate all relevant data into one platform. Additionally, it supports real-time data updates, allowing users to make data-driven decisions promptly. Its user-friendly interface, combined with advanced analytics features, makes **PowerBI** an excellent choice for transforming raw data into actionable insights.

PowerBI was chosen for our project due to these powerful visualization capabilities, combined with its ease of integration with multiple data sources, including **MongoDB**. The ability to create dynamic and interactive dashboards allows us to explore and present our data in a way that is both insightful and easy to understand.

PowerBI's compatibility with MongoDB via Open Database Connectivity (*ODBC*) driver further enhances its utility, as it allows direct connection to the data stored in **MongoDB**, ensuring that the visualizations are based on the most current data available.

To import the data from **MongoDB** into **PowerBI**, the *ODBC* driver from Devart was used. The process begins with setting up the *ODBC* connection to MongoDB. This involves configuring the *ODBC* data source name (*DSN*) with the connection details to the MongoDB database. Once the *ODBC DSN* is set up, PowerBI can connect to MongoDB using this DSN. In PowerBI, the data source is added by selecting the *ODBC* option and choosing the configured *DSN*. The tables and data from **MongoDB** are then accessible within **PowerBI**, where they can be selected and loaded into the PowerBI environment. This seamless integration allows for the data to be visualized and analyzed directly within **PowerBI**, leveraging its powerful analytics and visualization tools to derive meaningful insights from the data.

5 Results

After importing the data into **PowerBI**, we began constructing dashboards to explore and extract information and insights about the data and their relationships.

5.1 Variable

The first step was to perform an individual analysis of each variable mentioned previously. For each one of the variables, except for Countries, a dashboard was created to allow observation of various visualizations presenting the distribution and statistics of the variable globally. The key elements of this dashboard include:

- **Interactive Map**: The map shows the geographical distribution of the variable across countries. Each country has a circle whose size corresponds to the average number of the variable; the larger the circle, the higher the average number of the variable.
- **Filters**: At the top of the dashboard, there are filters for selecting specific years, countries and continents for detailed analysis. This is useful for focusing on particular regions, countries, or time periods of interest.
- **Sum of Variable by Year** - This line chart, in the top left, displays the total number of variable reported each year from 2010 to 2023.
- **Average of Variable by Year** - This line chart, below the sum chart, shows the average number of variable per year.

- **Max and Min Covid Cases by Year**: These two line charts, in the top right, represent the maximum and minimum number of variable reported by any country each year.
- **Data Table** - The table in the center lists the total, average, maximum and minimum variable for each country, providing users with the ability to see the values for each selected country, as well as the total value for each one.
- **Average of Variable by Country and Year**: This bar chart, in the bottom left, displays the average number of variable by country for each year from 2010 to 2023.
- **Count of Variable by Category**: The pie chart, in the bottom right, categorizes countries based on their variable counts. The categories are divided into *Very Low*, *Low*, *Medium* and *High*.

5.1.1 Countries

For the variable **Countries**, a different dashboard, as shown in Figure 3 was built since the label is categorical:

- **Interactive Map**: The map shows the geographical location of countries. Each country is colored differently, allowing for quick identification and distinction between them.
- **Filters**: At the top of the dashboard, there are filters for selecting specific years, countries and continents for detailed analysis. This is useful for focusing on particular regions, countries or time periods of interest.
- **Stacked Bar Chart**: This chart displays the number of countries by continent over the years from 2010 to 2023. Each bar is segmented by colors representing different years, making it easy to compare changes over time.
- **Bar Chart**: Below the stacked bar chart, there is a simple bar chart showing the total number of countries by year from 2010 to 2023. It is important to note that, in addition to the 193 countries recognized by the United Nations, there are also some countries that are not UN members, such as Greenland, Aruba or Kosovo, bringing the total number of recognized countries to over 200.

5.1.2 Covid Cases

The first numeric label analyzed was **Covid Cases**. Next, an explanation of the dashboard in Figure 4 will be provided.

The *Sum of Covid Cases by Year* chart shows significant increases in cases, with the total number rising from approximately 0.08 billion in 2020 to over 0.77 billion

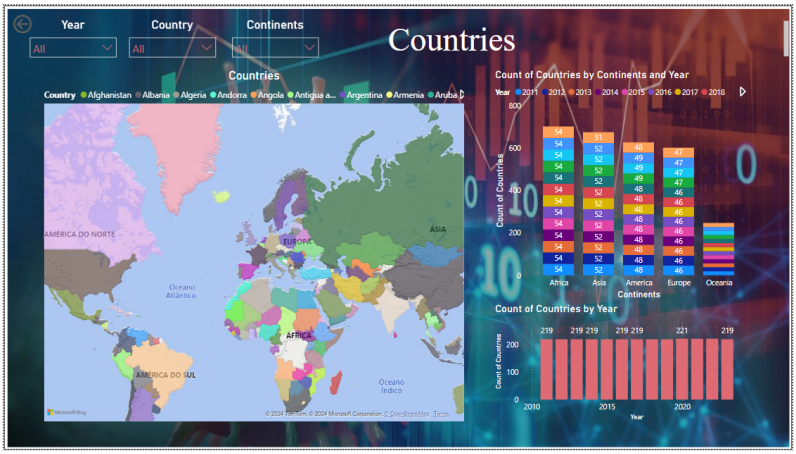

Figure 3 Dashboard Categorical Variables

in 2023. For instance, the year 2022 saw a notable peak, with cases dramatically increasing from 0.28 billion to 0.70 billion.

The *Average of Covid Cases by Year* line chart highlights trends in average cases, increasing from about 0.4 million in 2020 to 3.8 million in 2023, indicating how the virus spread over time and affected more people each subsequent year.

The *Max and Min Covid Cases by Year* charts show the highest and lowest case counts per year, respectively. The maximum chart shows the highest case counts peaking at approximately 103 million cases in 2023. The minimum chart indicates the lowest case counts, starting from zero cases in 2020 and increasing to 2943 cases in 2023, showing the broad range of case counts reported globally.

The *Average of Covid Cases by Country and Year* bar chart shows that the United States, India, and Brazil exhibit the highest averages, providing insights into which regions were most affected. For example, the average cases for the United States were significantly higher compared to other countries.

The *Count of Covid Cases by Category* pie chart illustrates that the majority of the cases (78.76%) fall into the Very Low category, while the High, Medium, and Low categories each account for 7.08% of the total, reflecting the varied impact of Covid-19 across different regions.

5.1.3 Gross Domestic Product

The second numeric label analyzed was **Gross Domestic Product**. Next, an explanation of the dashboard in Figure 5 will be provided.

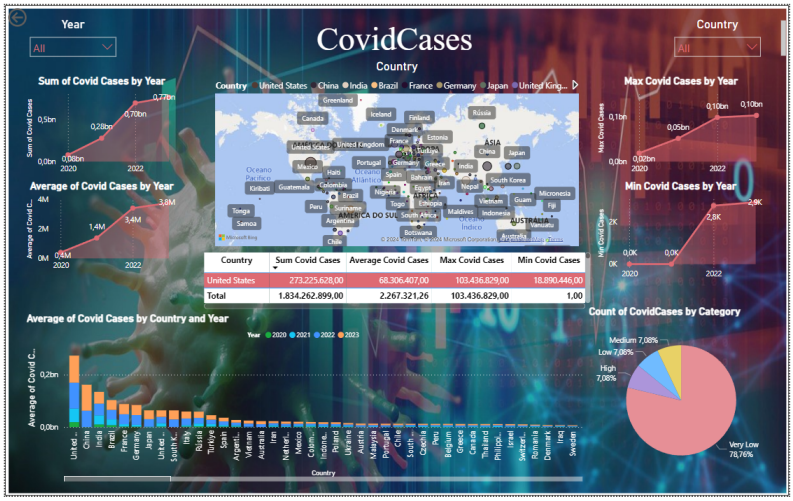

Figure 4 Dashboard CovidCases Variable

The *Sum of GDP by Year* line chart starts at 2.8 million in 2010, showing a significant drop to 2.5 million in 2015 and 2016, and then a rise during the pandemic years, from 2.7 million in 2020 to 3.4 million in 2023.

The *Average of GDP by Year* line chart highlights trends in the average GDP, increasing from about 12.9K in 2015 and 2016 to 17.1K in 2023, indicating changes in economic performance over time. Notably, in the first year of the pandemic, there was a decrease from 14.4K to 13.5K in one year.

The maximum chart shows the highest GDP values peaking at approximately 135K in 2021 and 2023. The minimum chart shows the lowest GDP values, with zero indicating countries for which no values were recorded.

The *Average of GDP by Country and Year* bar chart shows that Luxembourg, Switzerland, and Norway exhibit the highest averages, providing insights into which countries had the highest GDP on average.

The *Count of GDP by Category* pie chart demonstrates that the majority category is Very Low with 32.95%, while the High, Medium, and Low categories each represent 22.36% of the total, reflecting the varied economic performance across different regions.

5.1.4 Inflation

Inflation was the third numeric feature analyzed. Next, an explanation of the dashboard in Figure 6 will be provided.

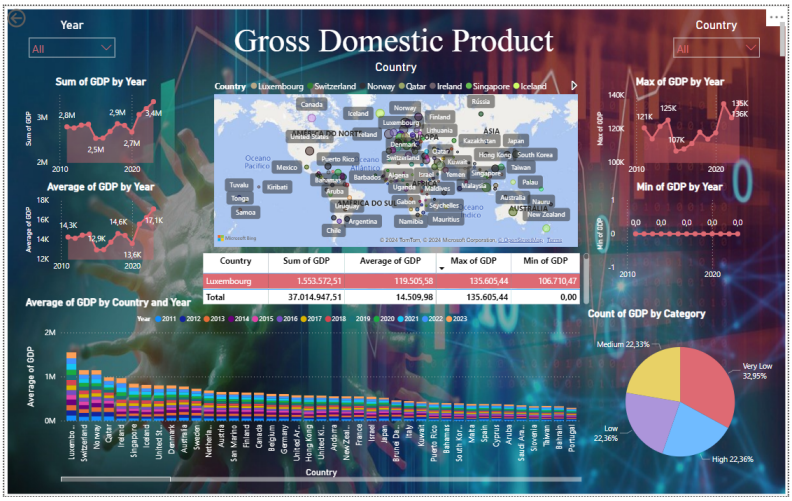

Figure 5 Dashboard GDP Variables

The *Sum of Inflation by Year* line chart begins with 1K in 2010, showing a significant peak at 66K in 2018 and a sharp drop to 3K in 2020.

The *Average of Inflation by Year* line chart highlights trends in average inflation, increasing from about 6 in 2010 to 339 in 2018, indicating fluctuations in inflation rates over time.

The maximum chart shows the highest inflation rates per year, peaking at approximately 65K in 2018. The minimum chart shows the lowest inflation rates, highlighting 2018 with a value of -44.

The *Average of Inflation by Country and Year* bar chart shows that Venezuela exhibits the highest averages, significantly surpassing all other countries.

The *Count of Inflation by Category* pie chart demonstrates that the largest category falls into the Very Low category, while the High, Medium, and Low categories each represent about 22% of the total.

5.1.5 Migration

Migration was the fourth numeric feature analyzed. Next, an explanation of the dashboard in Figure 7 will be provided.

The *Sum of Migration by Year* line chart starts with a positive migration sum in 2010, experiencing fluctuations between negative and positive values in the following years and then a major drop to -0.25 million in 2019.

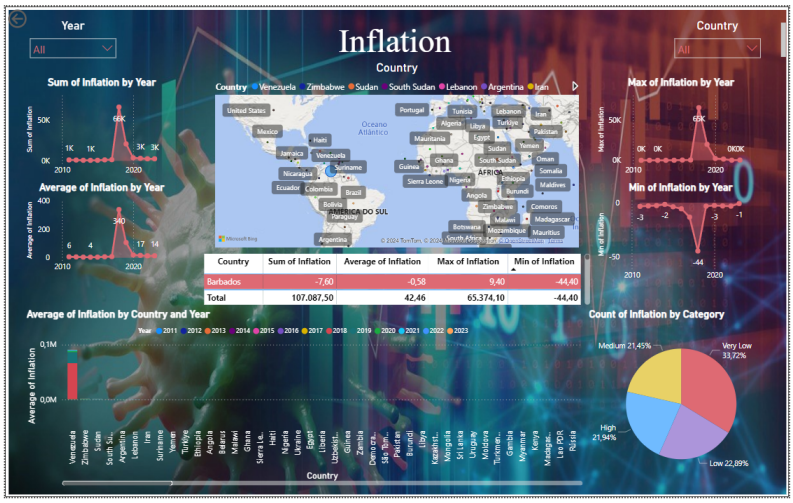

Figure 6 Dashboard Inflation Variable

The *Average of Migration by Year* line chart displays trends in average migration, with values varying from 0.5K in 2016 to -1.2K in 2019, indicating fluctuations in migration rates over time.

The maximum chart shows the highest migration rates per year, peaking at approximately 3.4 million in 2022. The minimum chart shows the lowest migration rates, starting from -0.9 million in 2010, dropping to -6.7 million in 2022, and then reaching -0.5 million in 2023.

The *Average of Migration by Country and Year* bar chart shows that the United States, Russia, and Germany exhibit the highest averages, providing insights into which countries had the highest migration rates on average.

The *Count of Migration by Category* pie chart shows that all the values are well distributed, with each category having values close to 25%.

5.1.6 Population

The fifth numeric feature analyzed was the **Population**. Next, an explanation of the dashboard in Figure 8 will be provided.

The *Sum of Population by Year* line chart shows a steady increase in the global population, starting from 7.1 billion in 2010 and reaching 8.1 billion in 2022, but having a drop to 8.0 billion in 2023.

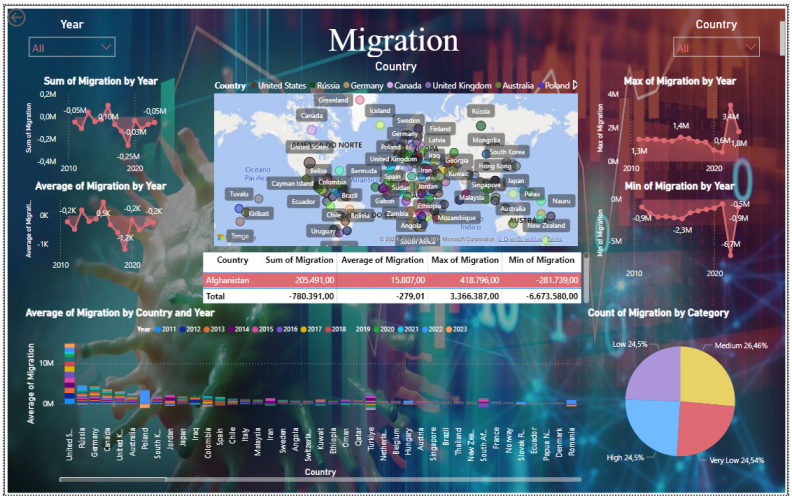

Figure 7 Dashboard Migration Variable

The *Average of Population by Year* line chart shows the average population per year. It highlights trends in average population, increasing from about 33.2 million in 2010 to 37.4 million in 2023, indicating a gradual rise in population over the years.

Similar to the previous charts, the maximum chart a increase by time with the highest value reaching approximately 1.43 billion in 2023. The minimum chart shows the lowest population counts, remaining relatively constant at around 11 thousand throughout the years, indicating the range of population sizes reported globally.

By looking at the *Average of Population by Country and Year* bar chart, it is evident that China and India surpass other countries, being the two most populous by a wide margin.

The pie chart titled *Count of Population by Category* is very similar to the migration data. All the values are well distributed, with each category having values close to 25

5.1.7 Taxes

The sixth numeric feature analyzed was **Taxes**. Next, an explanation of the dashboard in Figure 9 will be provided.

The *Sum of Taxes by Year* line chart illustrates the trend in the total amount of taxes collected globally. Starting from approximately 2.3K in 2010, the sum slightly increased to around 2.4K by 2022, with a notable drop starting in 2020 until 2023, reaching a value of 0.4K.

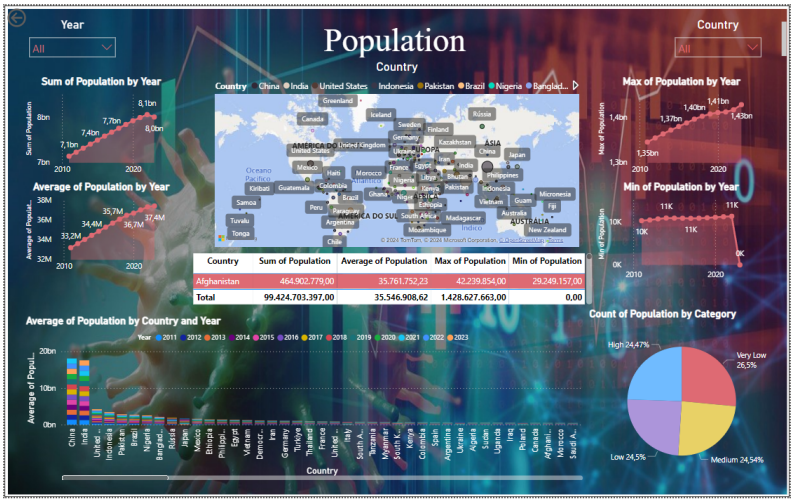

Figure 8 Dashboard Population Variable

The *Average of Taxes by Year* line chart highlights the average taxes collected each year. This metric shows values within the same range from 2010 to 2020, but from 2021 to 2023, the value reached a new high of about 19.1 in 2023, suggesting an increase in average tax collection over the years.

Similar to the previous charts, the *Max of Taxes by Year* line chart indicates an increasing trend over time, reaching the highest value of 148 in 2012 but dropping to 32 in 2023. Conversely, the *Min of Taxes by Year* chart shows minimal values ranging from 0 in 2010 to approximately 12 in 2022, demonstrating the variability in the minimum taxes collected globally.

By examining the *Average of Taxes by Country and Year* bar chart, it becomes evident that certain countries consistently show higher average tax collections, such as Timor-Leste and Denmark. The chart reveals how average taxes vary significantly across different countries and over the years.

The pie chart titled *Count of Taxes by Category* provides a categorical breakdown of the tax data. The Very Low category shows the highest value at 60.6%, while the rest of the categories each represent 13.21%.

5.1.8 Unemployment

The last numeric feature to be analyzed was **Unemployment**. Next, an explanation of the dashboard in Figure 10 will be provided.

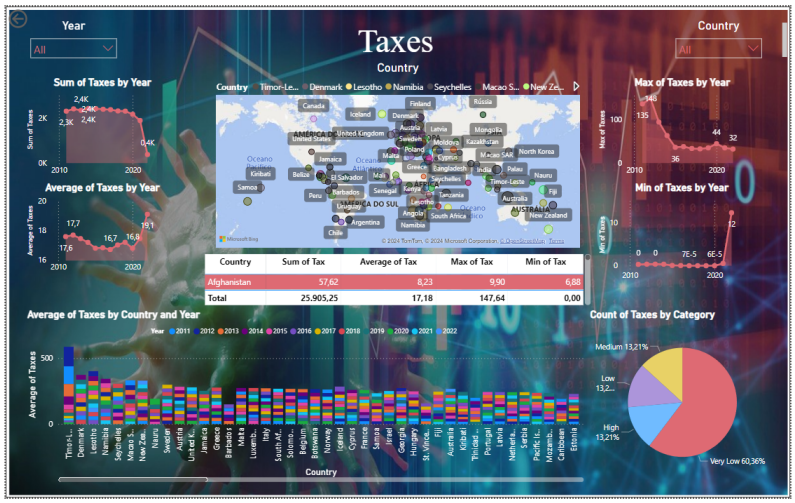

Figure 9 Dashboard Taxes Variable

The *Sum of Unemployment by Year* line chart starts at approximately 1546 in 2010, there is a gradual decline to 1351 in 2019, then a major increase to 1553 with a high decline to the lowest value 1346 in 2023.

The *Average of Unemployment by Year* line chart highlights the average number of unemployed individuals per year. This trend shows a peak of 8.3 in 2010 and 2015, with a subsequent decline to 7.9 in 2022, indicating a reduction in average unemployment rates over the period.

The *Max of Unemployment by Year* line chart indicates the highest recorded unemployment numbers each year, reaching a peak of 31.5 in 2010 and slightly decreasing to 26 in 2016. Conversely, the *Min of Unemployment by Year* chart shows the lowest unemployment numbers, with a peak in 2012 and then remains relatively stable around 0.1 to 0.15 throughout the years, reflecting the variability in minimum unemployment levels globally.

By examining the *Average of Unemployment by Country and Year* bar chart, it becomes evident that certain countries consistently show higher average unemployment values. Countries such as Djibouti, South Africa, and Eswatini, which are considered less developed, exhibit notably high average unemployment rates. This indicates a significant economic challenge in these regions, where unemployment remains a persistent issue over the years.

By examining the *Average of Unemployment by Country and Year* bar chart, it becomes evident that certain countries consistently show higher average unemployment values, such as Djibouti, South Africa and Eswatini.

The pie chart titled *Count of Unemployment by Category* provides a categorical breakdown of unemployment data. The majority being in the Very Low category at 41.08%, followed by other categories such as Low, Medium and High which each account for approximately 19.56% to 19.76% of the total data.

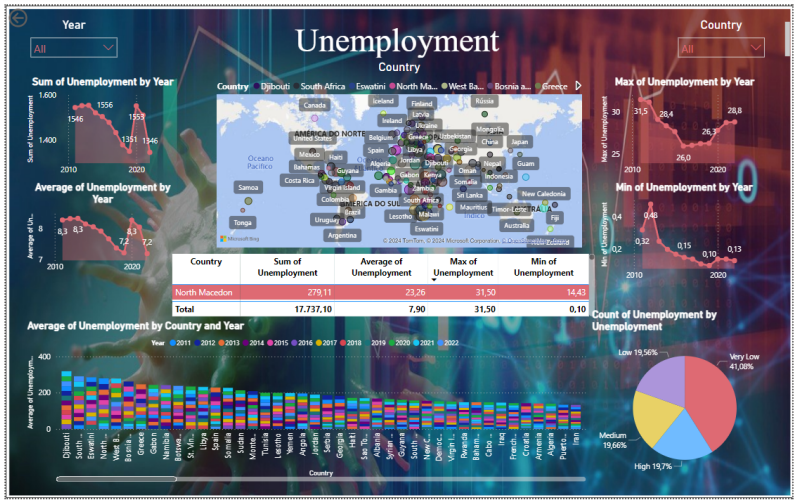

Figure 10 Dashboard Unemployment Variable

5.2 G8

After analyzing each variable globally, we aimed to delve deeper into the impact of these variables on individual countries, specifically comparing the G8 nations with those not in this group. The G8 countries, comprising some of the world's largest and most advanced economies, typically display different economic dynamics compared to less developed nations. This next section will explore these contrasts in detail, providing insights into how all the variables affect the G8 countries relative to non-G8 countries. This analysis will help identify unique patterns, challenges and opportunities within these diverse economic groups.

The key elements of the G8 dashboard, as shown in Figure **??**, include:

- **Filters**: At the top of the dashboard, there are filters for selecting specific years and G8 countries for detailed analysis.
- **Average of Variable in G8** - The pie chart, in the top left, shows the distribution of average variable among the G8 countries. Each segment represents a country's proportion of the total average of the variable, highlighting the relative impact across the G8.

- **Average of Variable by G8 and Year** - This bar chart, below the pie chart, displays the average number of variable in each G8 country for the years 2020 to 2023, providing a clear comparison across the years.
- **Average of Variable by Year and BoolG8**: This clustered column chart, in the top right, shows the average of variable by year, comparing G8 countries (True) to non-G8 countries (False).
- **Average of Variable by Year and G8**: This line chart, in the bottom left, shows the average number of variable Variableby year, differentiating between each G8 countries and non-G8 countries. There is a trendline to represent how the values are changing over the years.
- **Average of Variable by Year and BoolG8**: This line chart, in the bottom right, represents the average number of the variable by year, comparing G8 countries (True) against non-G8 countries (False). There is a trendline, similar to the previous line chart, to illustrate how the values change year by year.

5.2.1 G8 vs CovidCases

Following the same order as the previous global analysis, the first numeric variable to be analyzed was **CovidCases**. Next, an explanation of the dashboard in Figure 11 will be provided.

The *Average of CovidCases in G8* pie chart illustrates that the United States accounts for the majority of the cases, with 38.51%. In contrast, non-G8 countries collectively represent less than 1%. The remaining G8 countries each show values ranging between 8% and 13%.

The clustered column chart, named *Average of CovidCases by G8 and Year*, highlights the significant impact of COVID-19 in the United States, especially in 2022 and 2023. Another notable observation is that Canada, despite being a neighbor of the United States, registered the lowest number of COVID-19 cases among all G8 countries. As mentioned before, non-G8 countries recorded values that are almost negligible compared to the G8 countries.

That major difference is better illustrated by the other clustered column chart labeled *Average of CovidCases by Year and BoolG8*, which clearly evidences this disparity. For example, in 2023, the average number of COVID-19 cases in G8 countries was 37 million, compared to only 2.5 million in non-G8 countries.

The line chart, located in the bottom left, continues to highlight the impact of COVID-19, particularly in the United States. It also demonstrates that from 2021 to 2022, the number of cases increased significantly, presenting a stark contrast and higher values compared to previous years.

Similarly, the line chart in the bottom right shows the difference in the impact of COVID-19 between more developed countries and the rest of the world.

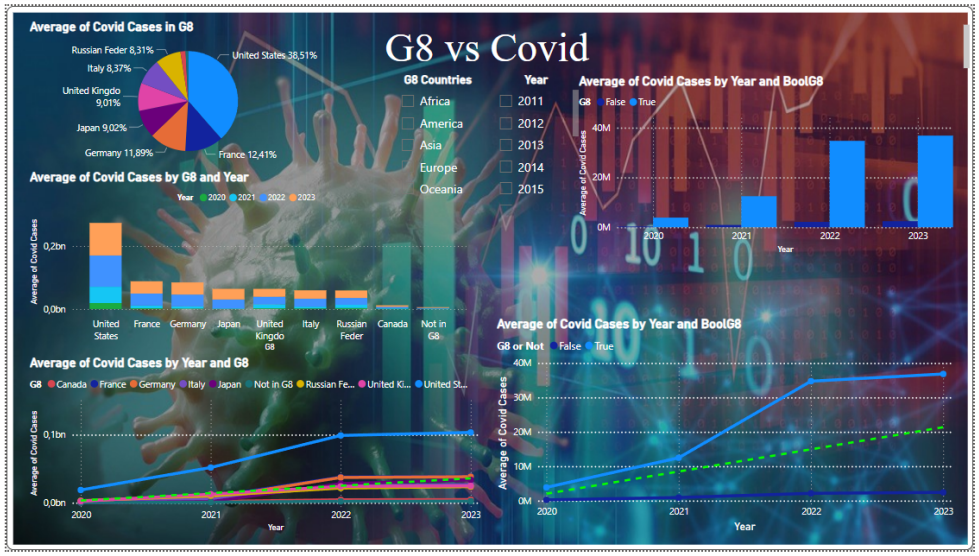

Figure 11 Dashboard G8 Covid Variable

5.2.2 G8 vs Gross Domestic Product

The second numeric variable to be analyzed is **GDP**. Next, an explanation of the dashboard in Figure 12 will be provided.

The *Average of GDP in G8* pie chart illustrates that almost all G8 countries have similar average GDP values, ranging from 11.61 in Japan to 17.96 in the United States. In contrast, non-G8 countries collectively represent less than 4%.

The clustered column chart, named *Average of GDP by G8 and Year*, similarly to the CovidCases, highlights the significant impact of GDP in the United States. Another notable observation is that Russia surprisingly registered the lowest GDP among all G8 countries, performing worse than the non-G8 countries.

As illustrated by the clustered column chart labeled *Average of GDP by Year and BoolG8*, the GDP values for all G8 countries remain stable around 40k, whereas the non-G8 countries have GDP values ranging between 11k and 16k.

The line chart, located in the bottom left, continues to highlight the impact of GDP, particularly in the United States, where in 2023, there was a significant improvement

to 80k. It also demonstrates that from 2010 to 2023, the GDP of Russia and non-G8 countries has remained very close.

Similarly, the line chart in the bottom right shows the difference in the impact of GDP more developed countries and the rest of the world.

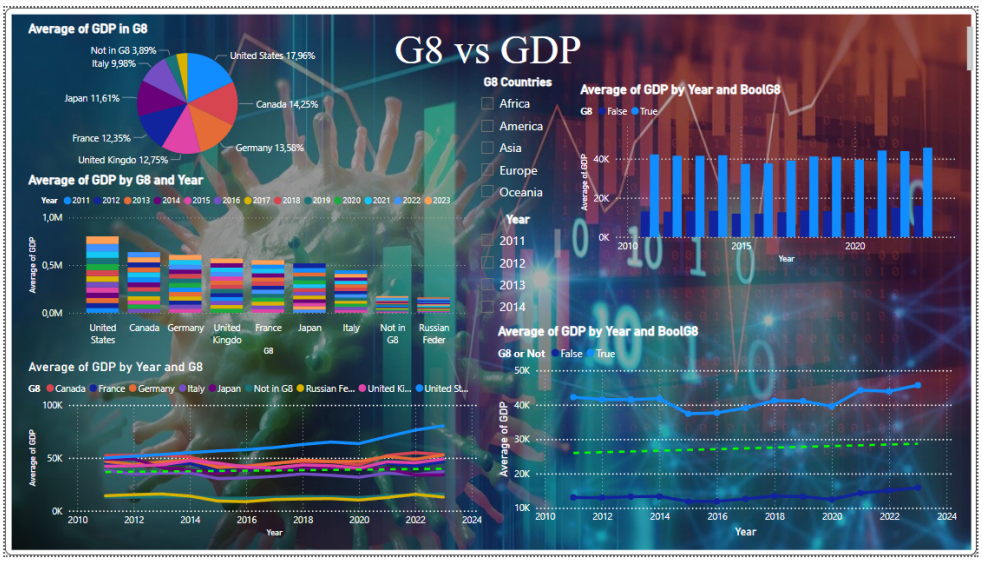

Figure 12 Dashboard G8 GDP Variable

5.2.3 G8 vs Inflation

The third numeric variable to be analyzed is **third**. Next, an explanation of the dashboard in Figure 13 will be provided.

The *Average of Inflation in G8* pie chart illustrates that the majority of G8 countries have relatively similar average inflation values, ranging from 3.15% in Italy to 10.52% in the Russia Federation. Notably, non-G8 countries collectively represent a significant portion, making up 66.47% of the total average inflation values.

The clustered column chart, named *Average of Inflation by G8 and Year*, shows that Russia experienced the highest average inflation rates among the G8 countries, while countries like Italy, Canada, and Japan registered lower average inflation rates. However, none of these values compare to the non-G8 countries, especially in the year 2018. This disparity was highlighted in the previous topic, *Inflation*, where Venezuela exhibited an extremely high inflation rate.

As illustrated by the clustered column chart labeled *Average of Inflation by Year and BoolG8*, the inflation values for G8 countries show a distinct trend when compared to non-G8 countries. The G8 countries' inflation rates are relatively stable, whereas the non-G8 countries exhibit more volatility in their inflation rates over the years, specially in 2018.

The line chart, located in the bottom left, continues to highlight the inflation trends, particularly showing notable spikes and drops. For instance, in 2018, there was a significant spike in inflation rates, which was more pronounced in non-G8 countries.

Similarly, the line chart in the bottom right illustrates the differences in inflation trends between G8 and non-G8 countries. This chart shows that the G8 countries tend to have more controlled inflation rates, while non-G8 countries experience more significant fluctuations, indicating varying economic stability levels.

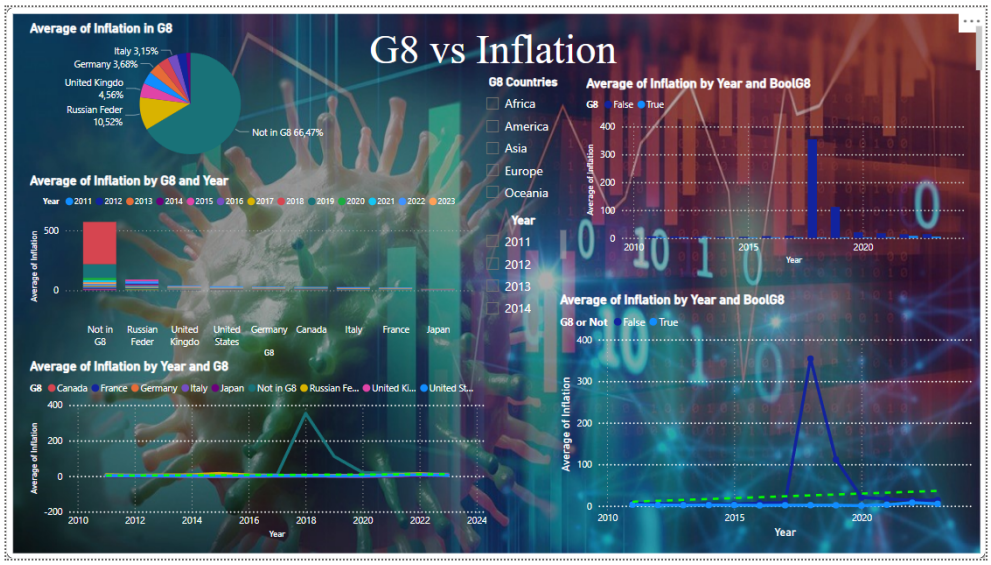

Figure 13 Dashboard G8 Inflation Variable

5.2.4 G8 vs Migration

Migration is the fourth variable to be analyzed. Next, an explanation of the dashboard in Figure 14 will be provided.

The *Average of Migration in G8* pie chart illustrates that the United States has the highest average migration among the G8 countries, accounting for 44.55% of the

total average migration. Other notable contributions come from the Russia (13.19%), Germany (12.87%), and Canada (10.57%).

The clustered column chart, named *Average of Migration by G8 and Year*, highlights the significant impact of migration in the United States. It shows that while the United States consistently has high migration rates, other G8 countries such as Germany, the Russian Federation, and Canada also show notable average migration values.

As illustrated by the clustered column chart labeled *Average of Migration by Year and BoolG8*, the migration values for G8 countries are generally higher and more stable compared to non-G8 countries. The chart shows that G8 countries maintain relatively high migration rates over the years, whereas non-G8 countries experience negative values.

The line chart, located in the bottom left, continues to highlight the impact of migration, particularly in the United States. It demonstrates that from 2011 to 2023, migration rates in the United States remained significantly higher compared to other G8 and non-G8 countries. Russia registered a high peak in 2022, reaching nearly 1 million, coming very close to the United States' figures.

The line chart in the bottom right shows the differences in migration trends between G8 and non-G8 countries. This chart indicates that G8 countries have more consistent and higher migration rates, whereas non-G8 countries exhibit more variability and lower overall migration rates. Another notable observation is the significant decrease in migration rates in G8 countries during 2020 and 2021, dropping from 0.34 million to 0.22 million. However, in 2022, the migration rates in G8 countries rebounded to levels consistent with those seen in the earlier years.

5.2.5 G8 vs Population

Population is the fifth variable to be analyzed. Next, an explanation of the dashboard in Figure 15 will be provided.

The *Average of Population in G8* pie chart illustrates that the United States has the highest average population among the G8 countries, accounting for 34.54% of the total average population. Other notable contributions come from the Russian Federation (15.33%), Japan (13.49%), and Germany (8.76%).

The clustered column chart, named *Average of Population by G8 and Year*, highlights the significant population sizes in the United States. It shows that while the United States consistently has a high population, other G8 countries such as the Russian Federation, Japan, and Germany also show substantial average populations.

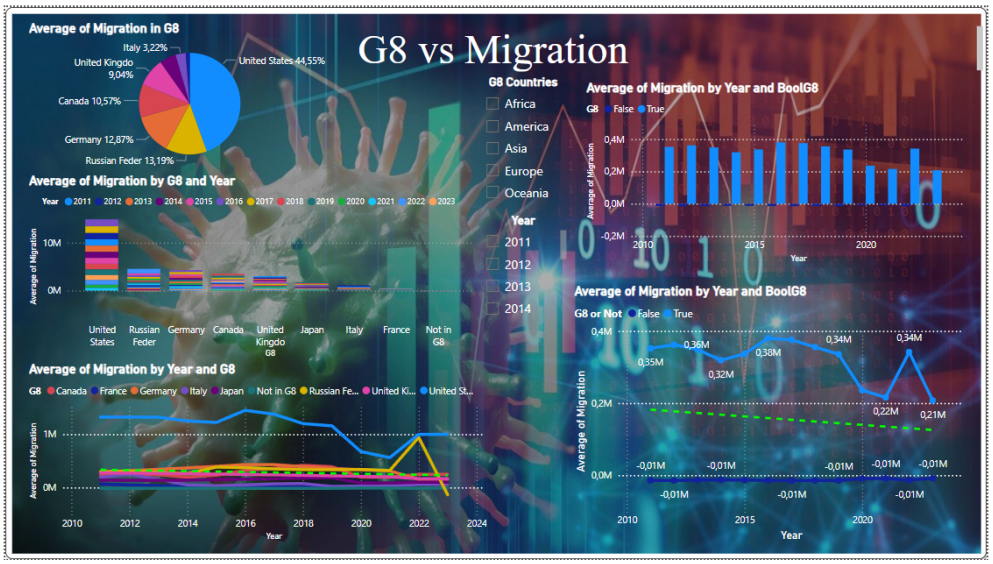

Figure 14 Dashboard G8 Migration Variable

As illustrated by the clustered column chart labeled *Average of Population by Year and BoolG8*, the population values for G8 countries are a lot higher compared to non-G8 countries in each year. The chart shows that G8 countries maintain relatively high population figures, above 100 million, whereas non-G8 countries exhibit more values between 30 and 35 million.

The line chart, located in the bottom left, continues to highlight the population trends, particularly in the United States. It demonstrates that from 2011 to 2023, the population in the United States remained significantly higher compared to other G8 countries and non-G8 countries.

The line chart in the bottom right shows the differences in population trends between G8 and non-G8 countries. This chart indicates that G8 countries have more consistent and higher population figures, whereas non-G8 countries have lower overall population figures, which is expected.

5.2.6 G8 vs Taxes

The penultimate variable to be analyzed, following **Population**, is **Taxes**. Next, an explanation of the dashboard in Figure 9 will be provided.

The *Average of Taxes in G8* pie chart illustrates that the United Kingdom has the highest average tax rate among the G8 countries, accounting for 18.5% of the

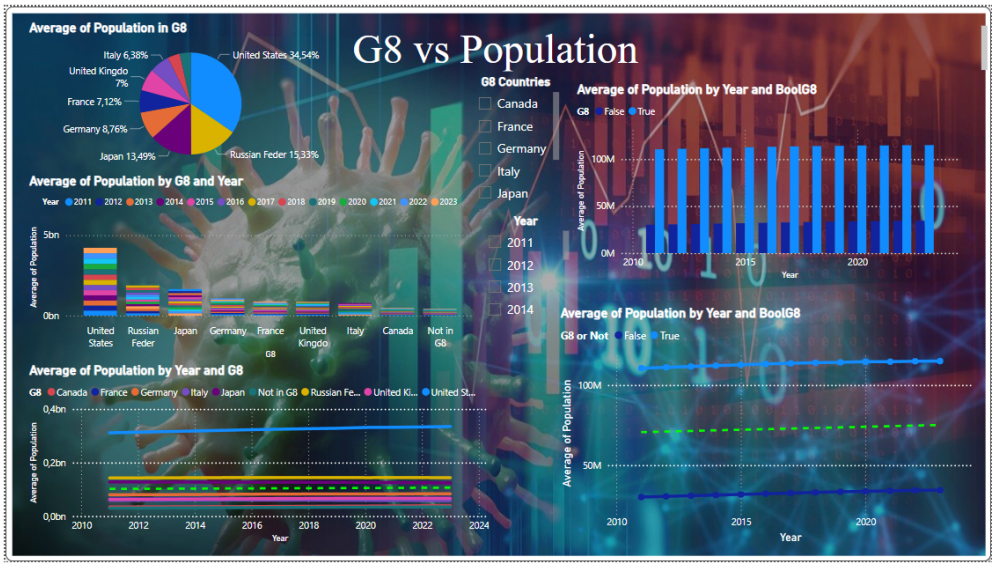

Figure 15 Dashboard G8 Population Variable

total average taxes. Other notable contributions come from Italy (18.03%) and France (17.12%).

The clustered column chart, named *Average of Tax by G8 and Year*, highlights the significant tax rates in the United Kingdom and Italy. It shows that while these countries consistently have high tax rates, France and non-G8 countries are not far behind.

As illustrated by the clustered column chart labeled *Average of Tax by Year and BoolG8*, the tax values for G8 and non-G8 countries are very similar, ranging between 15 and 20. However, in the last year, 2022, the biggest difference occurred, with the tax values for G8 countries dropping to below 15, while non-G8 countries' values almost reached 20.

The line chart, located in the bottom left, continues to highlight the tax trends, particularly in the United Kingdom and Italy. It shows with the trendline in dashed green that the values remain stable below 20 from 2010 to 2022.

Similarly, the line chart in the bottom right shows the differences in tax trends between G8 and non-G8 countries in the last year, as previously mentioned.

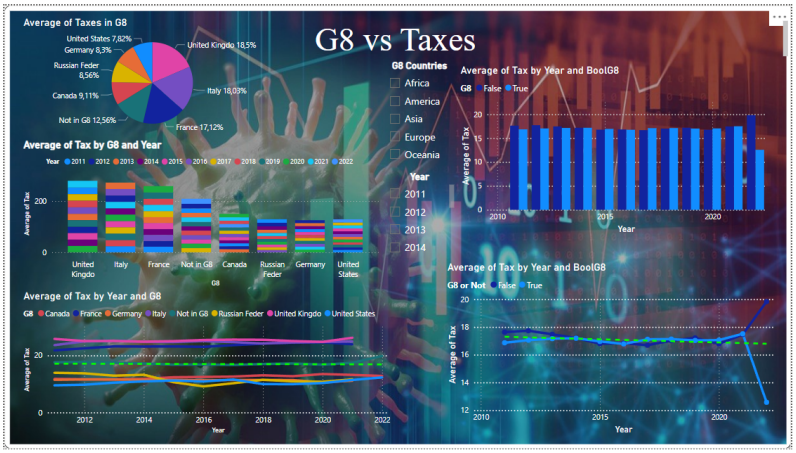

Figure 16 Dashboard G8 Taxes Variable

5.2.7 G8 vs Unemployment

The last numeric feature to be analyzed was **Unemployment**. Next, an explanation of the dashboard in Figure 17 will be provided.

The *Average of Unemployment in G8* pie chart illustrates that Italy has the highest average unemployment rate among the G8 countries, accounting for 17.95% of the total average unemployment. Other notable contributions come from France (15.63%), non-G8 countries (13.63%), Canada (11.91%) and the United States (9.93%).

The clustered column chart, named *Average of Unemployment by G8 and Year*, highlights the significant unemployment rates in Italy and France compared to Germany and Japan, which have the lowest registered values.

As illustrated by the clustered column chart labeled *Average of Unemployment by Year and BoolG8*, the unemployment values for G8 countries are generally higher and more stable compared to non-G8 countries. The chart shows that G8 countries maintain relatively high unemployment figures over the years, with values between 7 and 8. In contrast, non-G8 countries exhibit more fluctuations, with values ranging from 7 in 2012 to 5 in 2019.

The line chart, located in the bottom left, continues to highlight the unemployment trends, particularly in Italy and France. Italy registered a peak in 2014 with almost 13%, while on the other side, Japan recorded its lowest value in 2019 with only 2%.

Similarly, the line chart in the bottom right shows the differences in unemployment trends between G8 and non-G8 countries. This chart indicates that G8 countries generally have higher unemployment rates, whereas non-G8 countries exhibit lower overall

unemployment rates. However, as the dashed trendline shows, both groups experienced a decline in unemployment rates until 2019, followed by an increase, reaching the highest values in 2020.

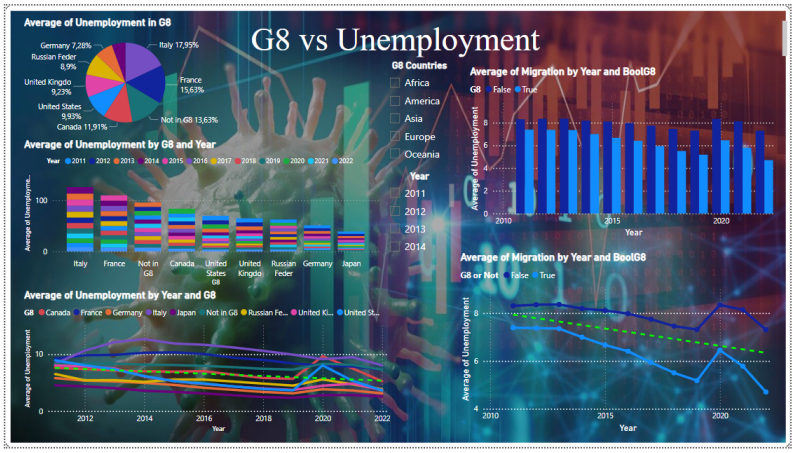

Figure 17 Dashboard G8 Unemployment Variable

5.3 Continents

After analyzing the variables globally, as well as focusing on the world's leading powers, the next essential step was to conduct an analysis by continents. This would allow us to understand the impact of the Covid-19 pandemic across different continents, providing a broader view of the global pandemic's effects. By examining the data continent-wise, we can identify specific trends, challenges and responses unique to each region, thereby gaining a comprehensive understanding of the pandemic's worldwide footprint.

The key elements of the Continent dashboard, as shown in Figure **??**, include:

- **Interactive Map**: The map shows the geographical location of countries. Each continent is colored differently, allowing for quick identification and distinction between them.
- **Filters**: At the top corners of the dashboard, there are filters for selecting specific continents and years for detailed analysis.
- **Sum of Variable by Year** - This line chart, in the top left, displays the total number of variable reported each year from 2010 to 2023 for a selected continent.
- **Average of Variable by Year** - This line chart, below the sum chart, shows the average number of variable per year for a selected continent.

- **Max and Min Covid Cases by Year**: These two line charts, in the top right, represent the maximum and minimum number of variable reported by selected continent each year.
- **Data Table** - The table in the center lists the total, average, maximum and minimum variable for each selected continent.
- **Average of Variable by Continent and Year**: This clustered column chart, in the bottom left, shows the average of variable by year, comparing each continent.
- **Average of Variable in Continent** - The pie chart, in the bottom middle, shows the distribution of average variable among the continents. Each segment represents a continent's proportion of the total average of the variable.
- **Average of Variable per Continent by Year**: This line chart, in the bottom right, shows the average number of variable by year, differentiating between each continent. There is a trendline to represent how the values are changing over the years.

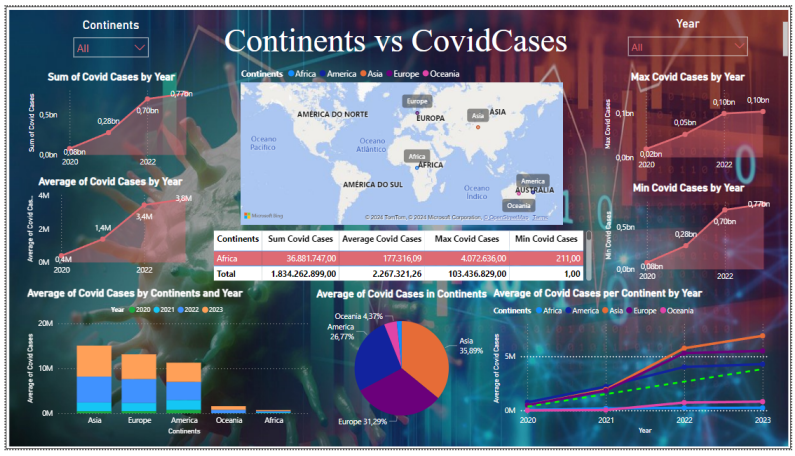

Figure 18 Dashboard Continent Numerical Variables

5.3.1 Continents vs CovidCases

In the Figure 19, the *Average of Covid Cases by Continents and Year* bar chart highlights the average number of Covid cases across different continents for the years 2020 to 2023. Asia, Europe and America exhibit the highest average cases, with Africa and Oceania showing relatively lower figures.

The pie chart titled *Average of Covid Cases in Continents* illustrates the distribution of average Covid cases among the continents. As previously mentioned, Asia accounts for the largest proportion at 35.89%, followed by Europe at 31.29% and

America at 26.77%. Africa and Oceania represent smaller shares at 4.37% and 1.68%, respectively.

The *Average of Covid Cases per Continent by Year* line chart provides a year-by-year comparison of average Covid cases for each continent. In 2021, Asia, Europe, and America registered almost the same values, but in 2022, Asia showed a distinct increase, continuing to rise until 2023. The trendline illustrates the mean progress of this data.

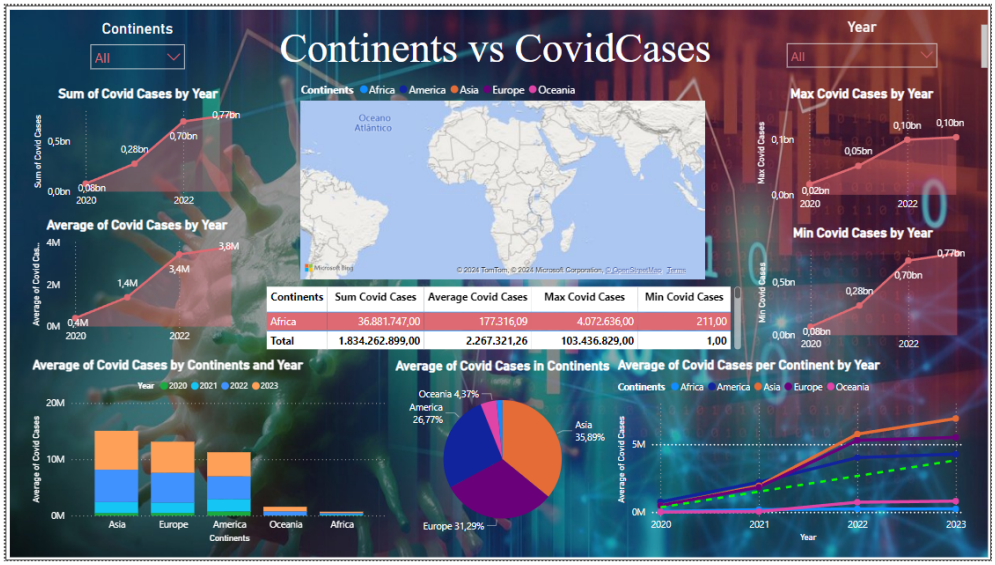

Figure 19 Dashboard Continent CovidCases Variables

5.3.2 Continents vs Gross Domestic Product

In the Figure 20, the bar chart titled *Average of GDP by Continents and Year* showcases the average GDP values across different continents from 2011 to 2020. Europe consistently demonstrates the highest average GDP during this period. Following Europe, America and Asia exhibit similar average GDP figures. Africa and Oceania, on the other hand, present relatively lower average GDP values compared to the other continents.

The *Average of GDP in Continents* pie chart illustrates the distribution of average GDP among the continents. Europe accounts for the largest proportion at 45.6%, followed by Asia (18.54%) and America (17.05%). Africa and Oceania represent smaller shares at 3.77% and 15.04%, respectively.

The *Average of GDP per Continent by Year* line chart offers a year-by-year comparison of average GDP for each continent. The chart highlights a significant disparity over time between Europe and the other continents, with Europe consistently leading in average GDP. In contrast, Africa displays the lowest average GDP, closely approaching the bottom line. These values accurately reflect the differing levels of development between countries in Europe and those in Africa.

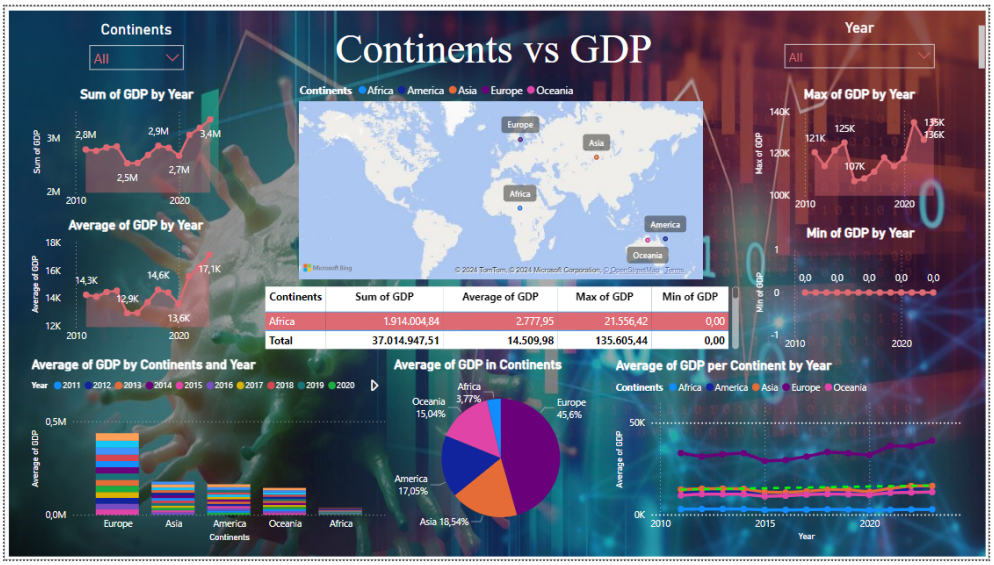

Figure 20 Dashboard Continent GDP Variables

5.3.3 Continents vs Inflation

In the Figure 21, the *Average of Inflation by Continents and Year* bar chart highlights the average inflation rates across different continents for the years 2011 to 2020. America, especially in 2018, exhibits the highest average inflation rate, with a significant margin compared to all the other continents, which show relatively lower figures.

The *Average of Inflation in Continents* pie chart illustrates the distribution of average inflation among the continents. America completely dominates, accounting for the largest proportion at 88.96%, followed by Africa at 5.2%, with smaller shares from other continents.

The *Average of Inflation per Continent by Year* line chart provides a year-by-year comparison of average inflation for each continent. This chart is unusual because the only values significantly above zero are for America, with almost 2k of inflation rate in

2018 and about 500 in 2019. These results align with the analysis made in the previous sections.

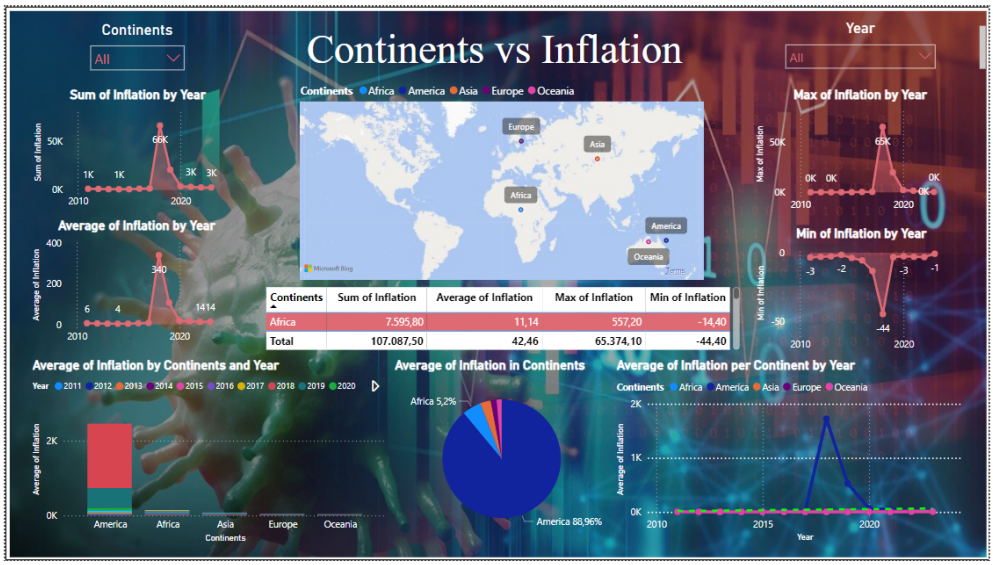

Figure 21 Dashboard Continent Inflation Variables

5.3.4 Continents vs Migration

In the Figure 22, the *Average of Migration by Continents and Year* bar chart highlights the average migration across different continents for the years 2011 to 2020. Europe and America exhibit the highest average migration, followed by Oceania, with Africa and Asia showing negative values indicating net emigration.

The *Average of Migration in Continents* pie chart illustrates the distribution of average migration among the continents. Europe accounts for the largest proportion at 43.3%, followed by America (38.26%) and smaller shares from Oceania (18.4%). Africa and Asia are not included in this analysis because their migration rates are negative, indicating net emigration, which is not represented in the pie chart.

The *Average of Migration per Continent by Year* line chart provides a year-by-year comparison of average migration for each continent. This chart highlights the trends over time, showing that Africa and Asia consistently had values below zero from 2010 to 2023.

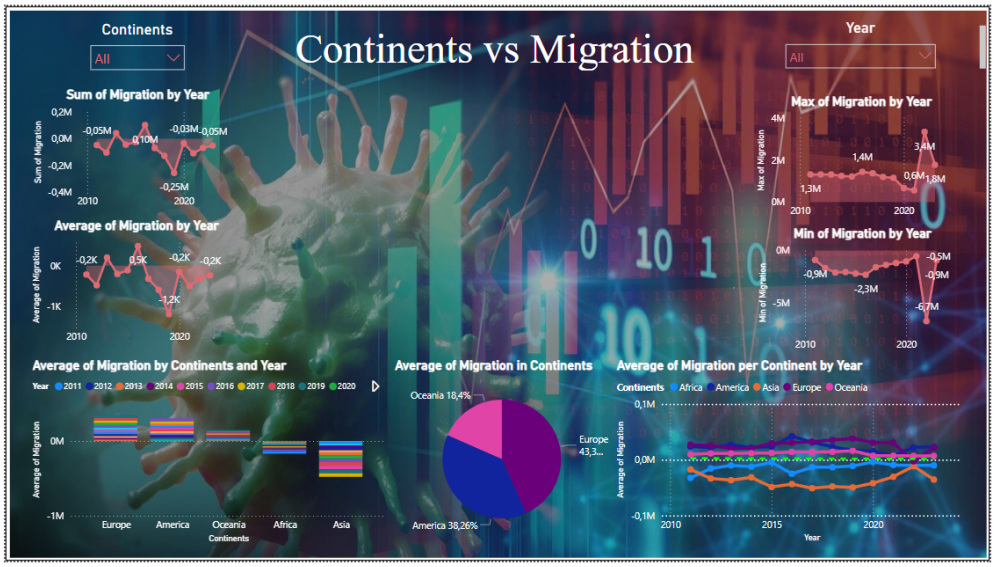

Figure 22 Dashboard Continent Migration Variables

5.3.5 Continents vs Population

In the Figure 23, the *Average of Population per Continent by Year* bar chart highlights the average population across different continents for the years 2010 to 2022. Asia exhibits the highest average population, because of China and India as previous seen, followed by Africa and Europe, with America and Oceania showing relatively lower figures.

The *Average of Population in Continents* pie chart illustrates the distribution of average population among the continents. Asia accounts for the largest proportion at 57.63%, followed by Africa (14.4%), America (16.36%), and Europe (10.71%). Oceania represents a smaller share.

The *Average of Population per Continent by Year* line chart provides a year-by-year comparison of average population for each continent. This chart highlights Asia's significant dominance compared to the other continents.

5.3.6 Continents vs Taxes

In the Figure 24, the *Average of Tax by Continents and Year* bar chart highlights the average tax across different continents for the years 2011 to 2020. For the first time, Oceania exhibits the highest average value in a variable, in this case tax, followed by Europe and Africa, with America and Asia showing relatively lower figures.

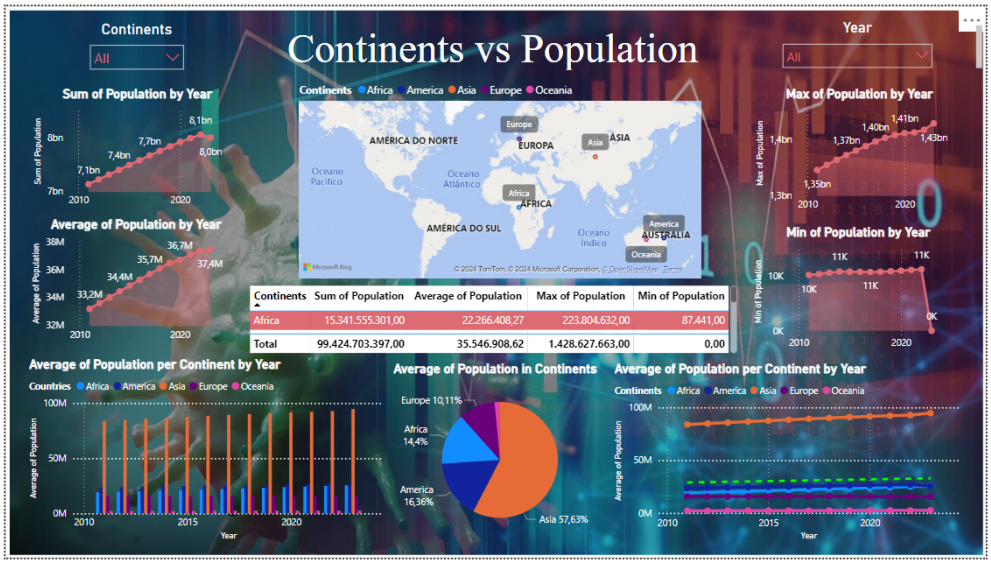

Figure 23 Dashboard Continent Population Variables

The *Average of Taxes in Continents* pie chart illustrates the distribution of average taxes among the continents. Oceania accounts for the largest proportion at 23.98%, followed by Oceania (23.45%) and Asia (17.07%). Africa and America represent smaller shares at 18.22% and 17.28%, respectively.

The Average of Taxes per Continent by Year line chart provides a year-by-year comparison of average taxes for each continent, clearly showing an increase in taxes during the pandemic.

5.3.7 Continents vs Unemployment

In the Figure 25, the *Average of Unemployment by Continents and Year* bar chart highlights the average unemployment across different continents for the years 2011 to 2020. Africa exhibits the highest average unemployment, followed by Europe and America, with Asia and Oceania showing relatively lower figures.

The *Average of Unemployment in Continents* pie chart illustrates the distribution of average unemployment among the continents. Asia accounts for the largest proportion at 57.63%, followed by America (16.36%) and Africa (14.4%). Europe and Oceania represent smaller shares at 10.11% and 1.5%, respectively.

The *Average of Unemployment per Continent by Year* line chart provides a year-by-year comparison of average unemployment for each continent. This chart highlights

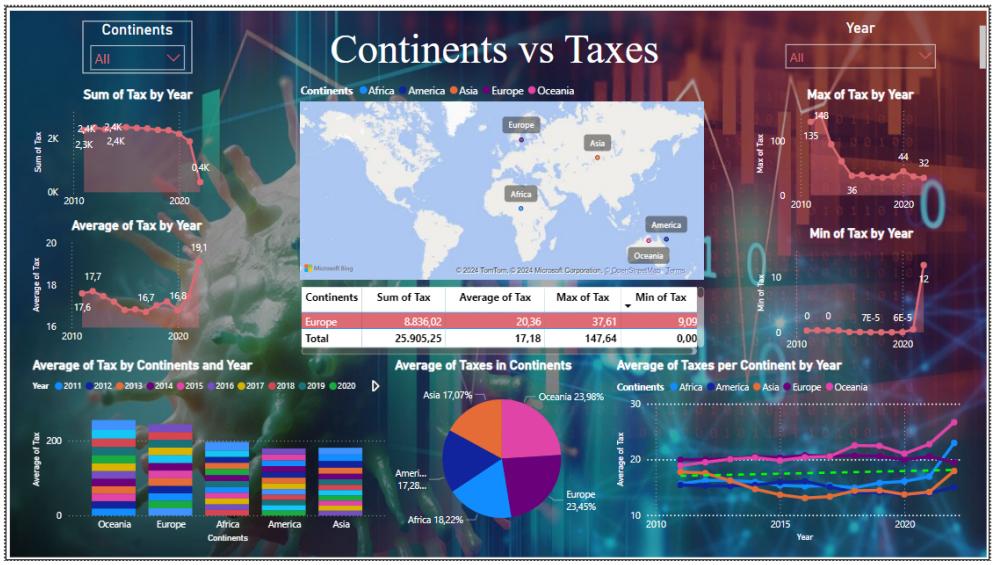

Figure 24 Dashboard Continent Taxes Variables

the trends over time, showing how different continents' unemployment values evolved during the specified period.

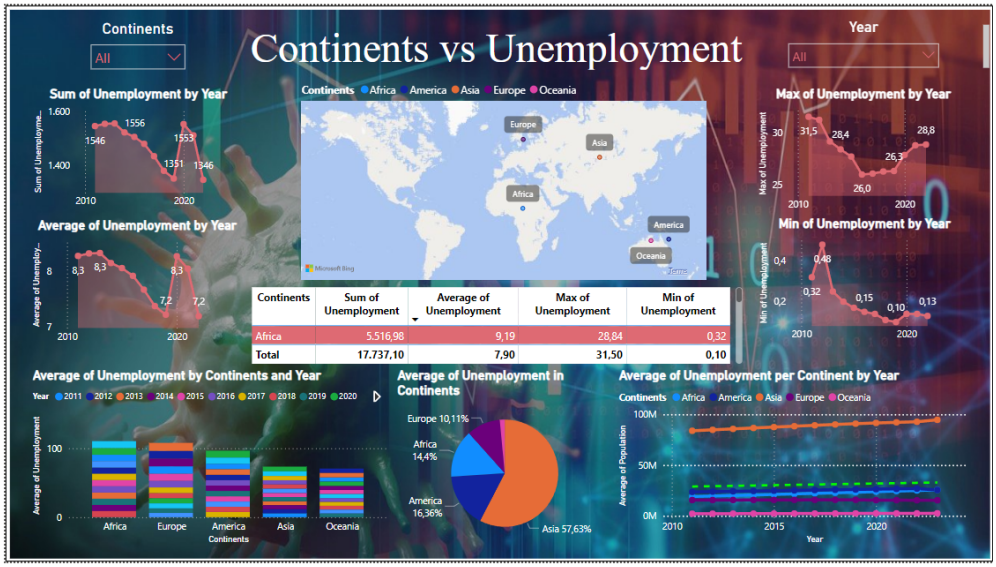

Figure 25 Dashboard Continent Unemployment Variables

5.4 Queries

To generate the required visualizations, it was necessary to develop specific queries to identify countries belonging to groups such as the G8, G20, and Continents. Additionally, boolean flags were created to indicate whether a country belongs to the G8 or G20.

Below is an example of a DAX formula used to label countries as part of the G20:

Listing 1 G20 Formula

```
G20 =
IF(
    'data'[Country] IN {"Argentina", "Australia", "Brazil", "Canada", "
        China", "France", "Germany", "Indonesia", "Italy", "Japan", "
        Mexico", "Russian Federation", "Saudi Arabia", "South Korea", "
        Turkey", "United Kingdom", "United States"},
    'data'[Country],
    "Not in G20"
)
```

This formula checks if a country is part of the G20 and labels it accordingly. If a country is not in the G20, it is labeled as *Not in G20*.

To create a boolean flag indicating G20 membership, the following DAX formula was implemented:

Listing 2 G20 Boolean Formula

```
G20 =
IF(
    'data'[Country] IN {"Argentina", "Australia", "Brazil", "Canada", "
        China", "France", "Germany", "Indonesia", "Italy", "Japan", "
        Mexico", "Russian Federation", "Saudi Arabia", "South Korea", "
        Turkey", "United Kingdom", "United States"},
    'data'[Country],
    "Not in G20"
)
```

This formula returns *True* if the country is part of the G20 and *False* otherwise.

For categorizing a label, as shown in the previous pie charts, the following DAX formula was implemented:

Listing 3 Unemployment Category Formula

```
Unemployment Category =
VAR Q1 = PERCENTILE.INC('data'[Unemployment], 0.25)
```

```
VAR M = PERCENTILE.INC('data'[Unemployment], 0.50)
VAR Q3 = PERCENTILE.INC('data'[Unemployment], 0.75)
RETURN
    SWITCH(
        TRUE(),
        'data'[Unemployment] <= Q1, "Very Low",
        'data'[Unemployment] > Q1 && 'data'[Unemployment] <= M, "Low",
        'data'[Unemployment] > M && 'data'[Unemployment] <= Q3, "Medium
            ",
        'data'[Unemployment] > Q3, "High"
    )
```

This formula categorizes the unemployment rate into four categories: *Very Low*, *Low*, *Medium* and *High* based on the quartiles.

6 Discussion

Based on the results mentioned in the previous section, several conclusions can be drawn about the data.

6.1 Variable Results

Regarding the variables from a global perspective, as shown in Figure 26, the study concluded that the onset of COVID-19 had a significant impact on several economic parameters. For instance, the average GDP of each country increased from around 12k-14k to 17k in 2023. In terms of Inflation, there was a notable rise during the pandemic compared to the values recorded between 2010 and 2017, which ranged from 4 to 8. During the COVID-19 years, Inflation rates surged to between 14 and 19. However, none of these figures compare to the inflation peak in 2018, where the average inflation rate soared to 340. This peak was primarily due to Venezuela, which experienced hyperinflation because of severe economic mismanagement, declining oil prices, and political instability. Venezuela's situation significantly contributed to the global inflation rate increase as it was the most extreme example of inflation during that period.

Similarly, the emergence of COVID-19 led to an increase in the average number of Migrations, rising from about -1200 to nearly neutral (150). However, the following year, the number oscillated, dropping to -500. In the case of Population, the pandemic had little effect, as the population consistently increased from 34 million in 2010 to approximately 38 million in 2023. Taxes, which varied between 17 and 16 before COVID, reached a new high of over 19 in 2023. Finally, Unemployment, which had been decreasing since 2013, from 8.32 to 7.22 (2019), rose again in 2020 to 8.26 due

to the global lockdowns. However, in the subsequent years, it resumed its downward trend, reaching a record low of 7.2 in 2023, the lowest since 2010.

In conclusion, the COVID-19 pandemic had a profound impact on various global economic parameters. It significantly influenced GDP growth, leading to notable increases. Inflation rates surged during the pandemic, affecting economic stability. Migration patterns were disrupted, with noticeable fluctuations in movement trends. Despite the pandemic, population growth remained steady, showing resilience. Tax rates saw an upward trend, reflecting changes in economic policies. Unemployment rates initially spiked due to lockdowns but eventually resumed a downward trend, demonstrating recovery efforts. Overall, the pandemic brought extensive socio-economic changes, affecting multiple aspects of the global economy.

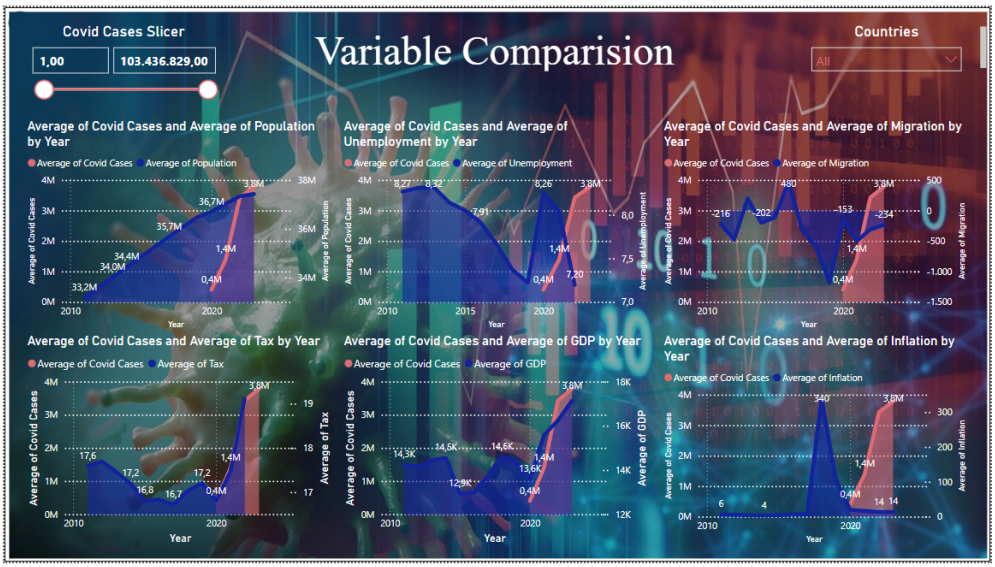

Figure 26 Dashboard Globally Variable Comparision

6.2 G8 Results

Focusing now on the G8 countries, we can observe a similar trend to the global case, as seen in the Figure 27. There was a significant drop in GDP during the first year of the pandemic, with values reaching 39.5k compared to 40k or more in previous years. However, in the following years, there was a substantial increase, culminating in a peak in 2023 at 45k, the highest since 2010. Regarding Inflation, there was a notable difference; between 2010 and 2020, inflation rates ranged from 3.3% to 0.89%. With the onset of COVID-19, this rate surged to unprecedented levels, peaking at 7.94% in 2022. By 2023, it had decreased to 5.2%.

Migration was another factor that saw significant changes. Since 2016, Migration numbers had been declining from 380k to 216k in 2021. However, in 2022, there was a sharp increase back to 342k, followed by a drop to the lowest average since 2010, at just 208k. In terms of Population, there was little change compared to previous years, with an increase of about 10 million from 2010 to 2023.

Regarding Taxes, significant differences were observed. Before the pandemic, tax values ranged from 16% to 18%. However, in 2022, there was an abrupt drop to an unprecedented 12.6%, reflecting the severe impact of the disease in the G8 Countries. Finally, Unemployment followed a similar trend, peaking at 6.5% in 2020, the first year of the pandemic, and then decreasing in the following years to a recent low of 4.7%.

In conclusion, the COVID-19 pandemic had a significant impact on the G8 countries' economies, similar to global trends. GDP experienced a substantial drop in the first year of the pandemic but recovered in subsequent years, peaking in 2023. Inflation surged to unprecedented levels during the pandemic, reaching a peak in 2022 before decreasing in 2023. Migration patterns were disrupted, with a sharp decline in numbers followed by fluctuations. Population growth remained steady, showing resilience over the years. Taxes saw a dramatic drop in 2022, reflecting the severe economic impact of the pandemic. Unemployment spiked in 2020 but gradually decreased in the following years, demonstrating economic recovery efforts. Overall, the pandemic brought extensive socio-economic changes to the G8 countries, affecting GDP, inflation, migration, population growth, tax rates and unemployment.

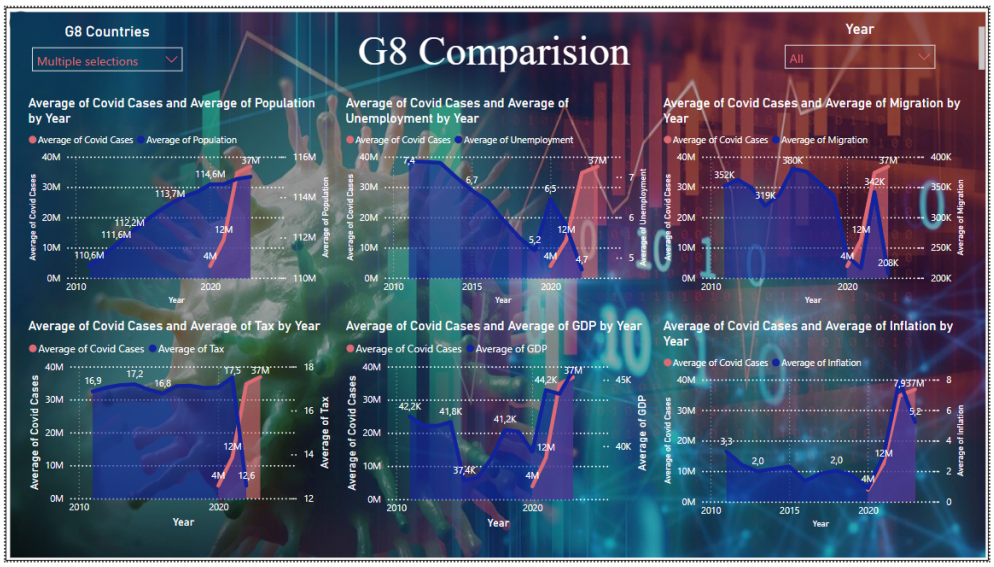

Figure 27 Dashboard G8 Variable Comparision

6.3 G20 Results

After conducting an in-depth study on the G8, we decided to include the G20 as well. As previously mentioned, the G20 consists of the world's 20 largest economies. Our aim was to determine if the impact on economies outside the G8 differed or if they followed the same pattern. As observed in Figure 28, the recorded values for both the G20 and the G8 show little variation. The only notable difference was in the average tax rates. In the case of the G8, this value decreased significantly from 2020 to 2022, while in the G20, there was an increase from 14.58 to an unprecedented 15.93, compared to the pre-pandemic values ranging from 14.24 to 14.81.

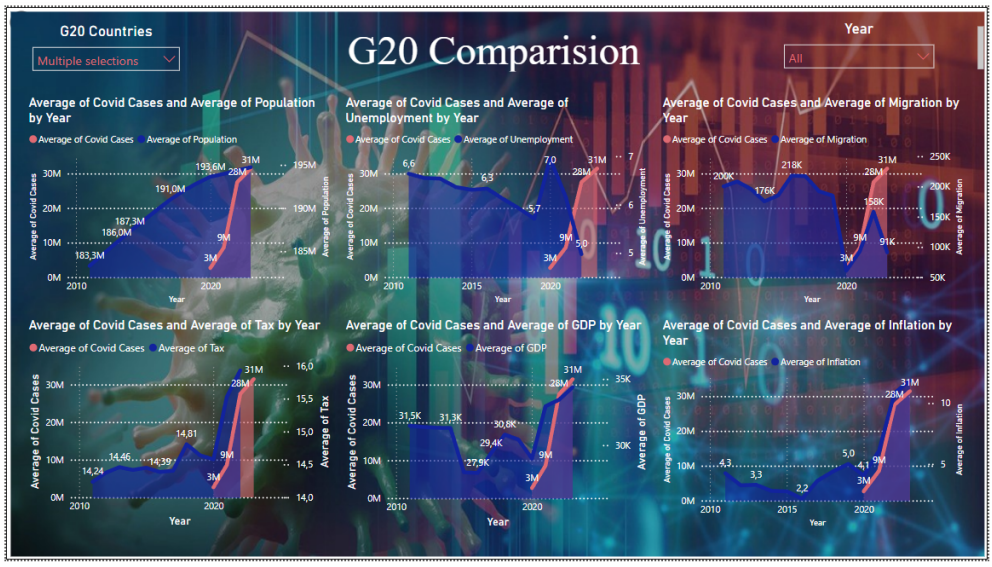

Figure 28 Dashboard G20 Variable Comparision

6.4 Continents Results

Finally, we were able to conclude the impact of the pandemic on each continent and determine which ones were most affected. The dashboard used, was similar to the previous ones, but we used the Filter to select each individual continent (Figure 29).

The GDP values during the pandemic followed a similar trend across all countries. In the pre-pandemic years, especially from 2014 to 2016, the GDP reflected the global economy on a large scale, with one of the main reasons being [17]. For example, Europe's GDP was 33.4k in 2014, but it fell to below 30k in the following two years. In the first year of the COVID-19 pandemic, a drop in the economy occurred worldwide,

with all continents showing a decline from 2019 to 2020. However, in the following years, there was a significant increase in GDP, with most countries reaching record values. For instance, Asia's GDP rose from 12k in 2020 to almost 16k in 2023.

Inflation is a special case. As shown in the previous section, America, specifically Venezuela in 2018, exhibited values completely different from other countries and continents. Consequently, America is the only continent where there is no major difference in inflation during the COVID years compared to the pre-pandemic period. In contrast, the rest of the continents showed a substantial increase in inflation during the COVID years. For example, in Asia, the inflation rate rose from 3.88% in 2019 to 14.11% in just three years.

Migration shows two different kinds of impacts. In countries with values below zero, such as Africa and Asia, the numbers rose closer to zero, for example, from -25k in 2016 in Africa to only -4k. On the other hand, the continents that had positive values experienced a drop. For instance, in Europe, the values went from 38k before the pandemic to 30k in 2021.

The impact of COVID-19 on taxes was global, with every continent experiencing an increase from 2019 to 2023. In Africa, the average values rose from 14.9% in 2018 to 22.96% in 2022. The only significant drop occurred in Europe, where the value fell to a new low of only 19.49% in 2022, compared to over 20% in the pre-pandemic period.

Unemployment around the world suffered due to the pandemic, with each continent showing a similar pattern. In the first year of the pandemic, there was a major increase in the unemployment rate. For example, in America, the rate was 7.7% in 2019, but it rose to 10% in just one year. However, in the following years, 2022 and 2023, the rates improved and dropped. In most countries, the unemployment rate in 2023 reached the lowest value recorded from 2010 to 2023. For instance, in Oceania, the pre-pandemic average was around 6%, but in 2023, it dropped to 5.49%.

In terms of Population, Europe and America were the most affected by the pandemic. In Europe, there had been a progressive increase in average values since 2010, but in 2020, the population dropped from 15.83 million to 15.53 million, continuing to decline until it reached 15.38 million in 2023. In America, despite improvements in the first two years of the pandemic, there was a significant drop in 2023, with the population decreasing from 26.9 million to 25.5 million in just one year.

In conclusion, the COVID-19 pandemic had varied impacts across different continents, highlighting both common trends and unique regional effects. Globally, GDP saw a significant decline in the first year of the pandemic but rebounded in subsequent years, with many regions reaching record levels by 2023. Inflation rates spiked on most continents during the pandemic, except in America, where the pre-existing situation in Venezuela overshadowed pandemic-induced changes. Migration patterns shifted, with continents like Africa and Asia experiencing reduced negative migration, while Europe saw a decrease in positive migration values. Tax rates generally increased worldwide,

except in Europe, which experienced a notable decrease. Unemployment rates initially surged across all continents but improved in the following years, reaching new lows by 2023. Population trends varied, with Europe and America facing declines during the pandemic years, contrasting with other continents' steady or increasing populations. Overall, the pandemic brought significant economic and social changes, affecting GDP, inflation, migration, taxes, unemployment, and population growth in distinct ways across the globe.

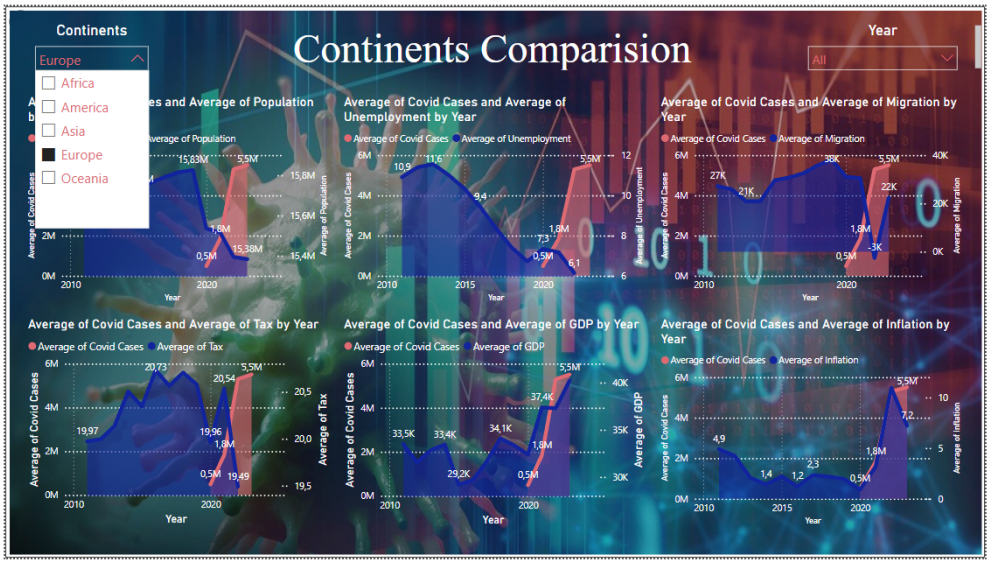

Figure 29 Dashboard Continents Variable Comparision

7 Conclusion and Critial Analysis

In this project, a comprehensive analysis of the impacts of COVID-19 on the global socio-economic landscape was conducted using Big Data systems. Utilizing PowerBI, we developed an extensive and interactive dashboard that allowed for detailed data visualization and analysis.

The comprehensive preprocessing of the various datasets was a critical step in ensuring the accuracy and reliability of our subsequent analysis. Each dataset, encompassing key socio-economic indicators underwent a meticulous transformation process. This involved cleaning the data, addressing missing values, standardizing country names and reshaping the datasets into a consistent long-format structure.

By consolidating these diverse datasets into a unified analytical framework, we facilitated a robust analysis of the socio-economic impacts of the COVID-19 pandemic.

The use of the Pandas library proved to be efficient in handling these tasks, resulting in a well-organized dataset stored in MongoDB, which was then seamlessly integrated into Microsoft PowerBI for advanced visualization. This preprocessing pipeline ensured that the data was not only clean and consistent but also readily accessible for detailed analysis, enabling us to derive meaningful insights and support the study's objectives effectively.

Throughout all the dashboards, various charts were presented, such as maps, pie charts, line charts, histograms, ranking charts and even tables to display data in terms of sum, average, minimum, and maximum. Additionally, users were allowed to choose country, continent, COVID-19 cases interval, year, and much more to make the experience more interactive and personalized.

All of this culminated in an in-depth analysis on each socio-economic variable at various scopes (globally, g8-wise, g20-wise and continent wise). The extracted insights were as follows.

Globally, **GDP** experienced a substantial decline in the first year of the pandemic, reflecting widespread economic disruptions. This trend was evident in the G8 countries, where GDP dropped significantly in 2020 before rebounding in subsequent years, peaking in 2023. Similar patterns were observed in the G20, indicating a broad economic recovery post-pandemic. Continental analysis also confirmed these trends, with all regions showing an initial GDP decline followed by recovery, underscoring the pandemic's extensive economic impact.

Inflation surged to unprecedented levels during the pandemic, driven by supply chain disruptions and increased demand for certain goods. This was particularly notable in the G8 and G20 countries, where inflation rates peaked in 2022 before stabilizing in 2023. Continental analysis revealed that while most regions experienced inflation spikes, unique conditions in America, particularly due to Venezuela's pre-existing economic situation, led to varying inflationary impacts.

Migration patterns were significantly disrupted, with sharp declines followed by fluctuations. In the G8 and G20 countries, migration numbers dropped during the pandemic's peak but showed signs of recovery in the following years. Continental analysis indicated that Africa and Asia experienced reduced negative migration, while Europe saw a decrease in positive migration values, highlighting the pandemic's differential impact on migration flows.

Population growth remained relatively steady despite the pandemic, demonstrating resilience. Both the G8 and G20 countries showed consistent population trends, with continental data further supporting this stability. Europe and America faced slight declines in population growth during the pandemic years, contrasting with steady or increasing populations in other regions.

Taxes revenues experienced significant changes during the pandemic. The G8 countries saw a dramatic drop in taxes rates in 2022, reflecting the severe economic

impact. In contrast, the G20 countries exhibited an increase in taxes rates, suggesting varied fiscal responses. Continental analysis showed that while tax rates generally increased worldwide, Europe experienced a notable decrease, emphasizing regional fiscal policy differences.

Unemployment rates initially spiked across all regions due to lockdowns and economic slowdowns but gradually decreased in the following years as economies recovered. The G8 countries demonstrated this trend clearly, with unemployment peaking in 2020 and then declining. The G20 and continental analyses confirmed these patterns, showing a universal struggle with unemployment during the pandemic, followed by recovery efforts that eventually led to lower unemployment rates by 2023.

In conclusion, this project meticulously developed a robust framework for analyzing the socio-economic impacts of the COVID-19 pandemic, utilizing comprehensive data preprocessing, efficient data storage in MongoDB, and advanced data visualization with PowerBI. The analysis revealed critical insights, such as the substantial GDP decline and subsequent recovery, inflation spikes, disrupted migration patterns, resilient population growth, fluctuating tax revenues, and initial unemployment surges followed by recovery. These findings underscore the extensive economic and social disruptions caused by the pandemic. Strategic recommendations emerging from these insights include the need for robust economic recovery plans, enhanced supply chain resilience, targeted migration policies, adaptive fiscal measures, and comprehensive support for labor markets to mitigate future crises and promote sustained economic stability.

References

[1] (IMF), I.M.F.: Gross Domestic Product (GDP). https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/gross-domestic-product-GDP

[2] Investopedia: Gross Domestic Product (GDP). https://www.investopedia.com/terms/g/gdp.asp

[3] Investopedia: Inflation. https://www.investopedia.com/terms/i/inflation.asp

[4] (IMF), I.M.F.: Inflation. https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/Inflation: :text=Inflation

[5] Commission, E.: Migration. https://home-affairs.ec.europa.eu/networks/european-migration-network-emn/emn-asylum-and-migration-glossary/glossary/migration$_e n$

[6] OECD: Population. https://data.oecd.org/pop/population.htm

[7] Foundation, T.: Tax. https://taxfoundation.org/taxedu/glossary/tax/

[8] Investopedia: Taxes. https://www.investopedia.com/terms/t/taxes.asp

[9] Investopedia: Unemployment. https://www.investopedia.com/terms/u/unemployment.asp

[10] Data, O.W.: COVID-19 Cases. https://ourworldindata.org/covid-cases

[11] Kanchana1990: IMF's GDP Data (1980-2028): Global Trends. https://www.kaggle.com/datasets/kanchana1990/imfs-gdp-data-1980-2028-global-trends

[12] Fund, I.M.: World Economic Outlook: Inflation (Consumer Prices, Annual % Change). https://www.imf.org/external/datamapper/PCPIPCH@WEO/WEOWORLD/VEN

[13] Bank, W.: World Bank Data: Net Migration. https://databank.worldbank.org/reports.aspx?source=2series=SM.POP.NETMcountry=

[14] Bank, W.: World Bank Data: Tax Revenue (% of GDP). https://data.worldbank.org/indicator/GC.TAX.TOTL.GD.ZS

[15] Bank, W.: World Bank Data: Total Population. https://databank.worldbank.org/reports.aspx?source=2series=SP.POP.TOTLcountry=WLD

[16] Bank, W.: World Bank Data: Unemployment, Total (% of Total Labor Force). https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS

[17] Investopedia: European Sovereign Debt Crisis. https://www.investopedia.com/terms/e/european-sovereign-debt-crisis.asp