25933337_GlobalDevelopmentIndicators_report

SI 348 Assignment Report

Project 1: Global Development Indicators

Introduction

The data for project 1 contains a dataset of various countries and their global development indicators. These indicators include values by year for each country, such as GDP, Control of Corruption, Gini index, population, and electric power consumption. While the dataset does contain certain missing values, conducting an exploratory data analysis on the data revealedinteresting patterns and correlations between values such as GDP and year (Percentage change in GDP per year), GDP and Gini index, and GDP and Population (GDP per capita). The following report will examine these patterns and relationships in terms of the following questions:

1. How do factors such as year illustrate the variation in a country's GDP?
2. How does the variation in factors such as population effect a country's GDP?
3. What is the correlation between factors such as Gini index and a country's GDP?

Data Cleaning and Preparation

A variety of steps were taken in the process of cleaning the data and preparing it for the exploratory data analysis process. The steps are outlined below as follows.

Loading the Provided Dataset into R for Project 1

The relevant dataset provided for Project 1 was first loaded into R Studio using a relative path to ensure the code is fully reproducible. The relevant dataset was in a `.csv` format, and was loaded accordingly using the `read_csv()` function.

Performing the Initial Inspection

An initial inspection of the dataset was then performed. This was acheived through the use of functions such as `head()`, `glimpse()`, and `view()`. The dataset could then be inspected in its raw form. This initial inspection made it immediately clear that the data would need to be pivoted longer in order for a successful exploratory data analysis to be conducted.

Table 1: The First Five Rows and Eight Columns of the Data in its Raw Form

Country Name	Country Code	Series Name	Series Code	1983 [YR1983]	1984 [YR1984]	1985 [YR1985]	1986 [YR1986]
Afghanistan	AFG	Gini index	SI.POV.GINI	..	..	..	..
Afghanistan	AFG	Adolescents out of school (% of lower secondary school age)	SE.SEC.UNER.LO.ZS		..	..	..
Afghanistan	AFG	Average precipitation in depth (mm per year)	AG.LND.PRC	327	327	327	327
Afghanistan	AFG	Central government debt, total (% of GDP)	GC.DOD.TOTL.GD.ZS		..	..	..
Afghanistan	AFG	Compulsory education, duration (years)	SE.COM.DUF	..	..	..	..

Inspecting Column Values

Due to the large amount of columns in the dataset, it was deemed essential that the dataset be pivoted before it could be fully inspected. Once the data was pivoted, the values for each column could be better inspected. In the inspection of the column values, it was observed that '..' was a fixed value which denoted a missing value, and once the data was pivoted these values were filtered out.

Table 2: The First Five Rows of the Pivoted Data

Country Name	Country Code	Series Name	Series Code	Year	Total

Afghanistan	AFG	Average precipitation in depth (mm per year)	AG.LND.PRC	1983	327
Afghanistan	AFG	Average precipitation in depth (mm per year)	AG.LND.PRCP.MM	1984	327
Afghanistan	AFG	Average precipitation in depth (mm per year)	AG.LND.PRC	1985	327
Afghanistan	AFG	Average precipitation in depth (mm per year)	AG.LND.PRCP.MM	1986	327
Afghanistan	AFG	Average precipitation in depth (mm per year)	AG.LND.PRC	1987	327

Cleaning the Data

Now that the dataset has been successfully pivoted, it could be cleaned using regular expressions. In order to be fully tidied, the column values needed to be split to only contain one value each. This was achieved through the use of regular expressions, wherein regular expressions were used to detect the Series Name from each row and extract the unit of measurement for the series into another column named Measurement. The original Series Name value was then tidied to remove the unit of measurement.

Table 3: The First Five Rows of the Cleaned Data

Country Name	Country Code	Series Name	Series Code	Year	Total	Measurement
Afghanistan	AFG	Average precipitation in depth	AG.LND.PRC	1983	327	(mm per year)
Afghanistan	AFG	Average precipitation in depth	AG.LND.PRCP.MM	1984	327	(mm per year)
Afghanistan	AFG	Average precipitation in depth	AG.LND.PRC	1985	327	(mm per year)
Afghanistan	AFG	Average precipitation in depth	AG.LND.PRCP.MM	1986	327	(mm per year)

Afghanistan	AFG	Average precipitation in depth	AG.LND.PRC	1987	327	(mm per year)

Filtering the Data

The clean dataset was then filtered and grouped by Country Name.

Exploratory Data Analysis

Summary Statistics

The first step in the exploratory data analysis process involved creating a variety of summary statistics for key variables from the dataset to explore the relationships both withinand between variables. This was an iterative process, as certain summary statistics revealed interesting trends, while some were deemed irrelevant to the research questions. Any notable patterns were then visualized using either scatterplots, heat maps, or bar graphs. The key variables relevant tto the research question were identified as GDP, year, population, and Gini Index.

Summary statistics such as mean were explored for variables with non-linear values, due to the nature of the format of the dataset. These variables included Control of Corruption, Expenditure on Primary, Secondary, and Tertiary Education, Military Expenditure, Gini index, Electric power consumption, Electricity produced from coal sources, International tourism, and Research and development expenditure. The summary statistics can be viewed in the tables below.

Table 4: The First Five Rows of the Control of corruption Summary Statistic

Country Name	mean_corrup_control
Denmark	2.309121
Finland	2.253437
Netherlands	2.003860
Iceland	1.958645
Canada	1.890934

Table 5: The First Five Rows of the Expenditure on primary education Summary Statistic

Country Name	mean_prim_edu
Cambodia	56.88139
Solomon Islands	56.29228
Nepal	55.52886
Afghanistan	54.50428
Kenya	52.88722

Table 6: The First Five Rows of the Expenditure on secondary education Summary Statistic

Country Name	mean_sec_edu
Afghanistan	24.59907
Bangladesh	41.15424
Belarus	50.45858
Belgium	46.72427
Botswana	43.02956

Table 7: The First Five Rows of the Expenditure on tertiary education Summary Statistic

Country Name	mean_ter_edu
Libya	52.72631
Canada	32.29491
Liberia	31.56345
Netherlands	30.51171
Ukraine	30.12521

Table 8: The First Five Rows of the Military expenditure Summary Statistic

Country Name	mean_militaryexp
United Arab Emirates	5.484465
Pakistan	4.673181
United States	4.348598
Libya	3.792103
Ethiopia	3.628593

Table 9: The First Five Rows of the Gini index Summary Statistic

Country Name	mean_gini
South Africa	61.88333
Botswana	58.70000
Brazil	56.20833
Chile	50.54375
Mexico	49.66667

Table 10: The First Five Rows of the Electric power consumption Summary Statistic

Country Name	mean_power_consump
Iceland	27801.92
Canada	16114.89

Finland	14128.02
United States	12494.84
United Arab Emirates	10112.86

Table 11: The First Five Rows of the Electricity production from coal sources Summary Statistic

Country Name	mean_elec_coal
Botswana	94.28963
South Africa	93.98530
Poland	93.64987
Czechia	69.14619
India	67.66250

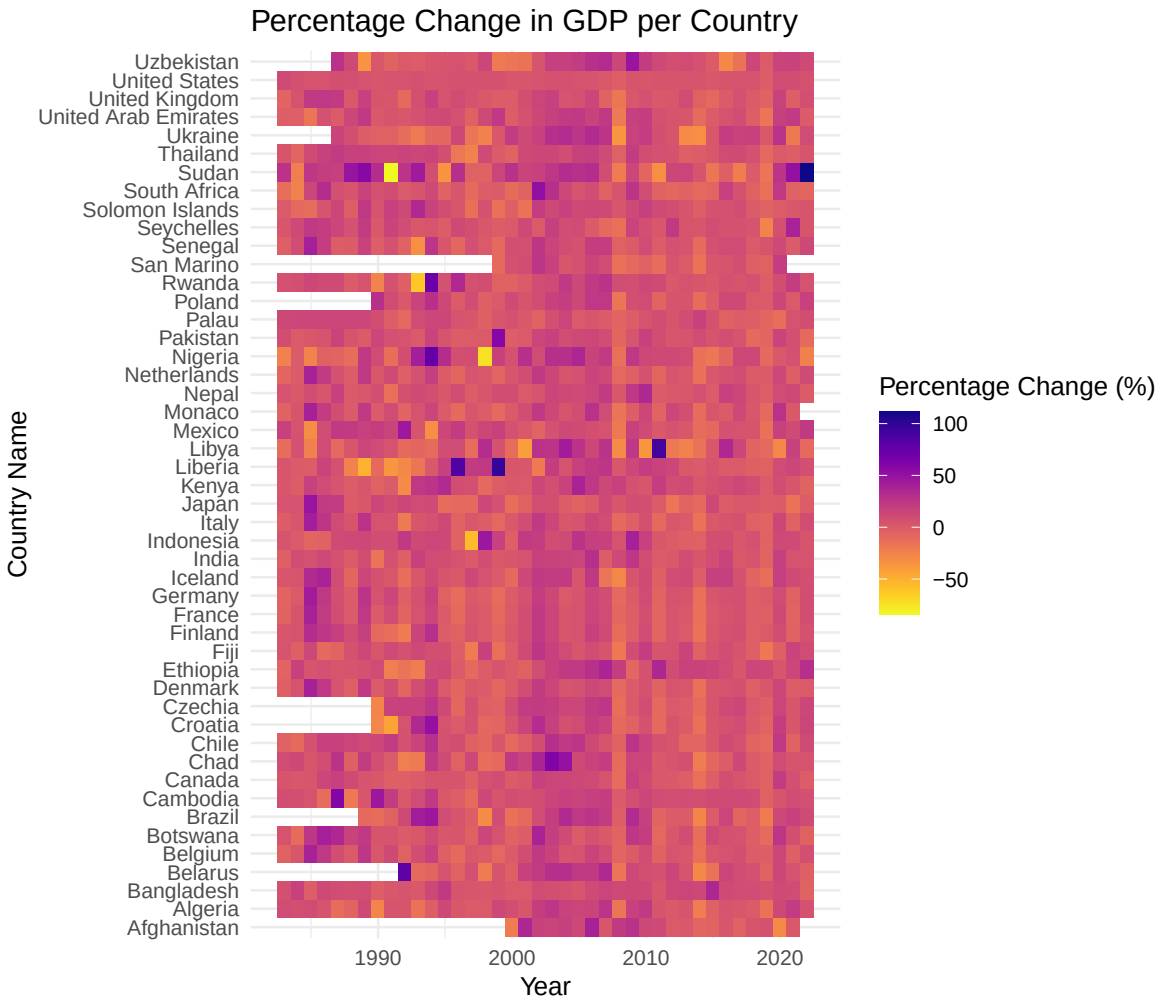
Table 12: The First Five Rows of the International tourism Summary Statistic

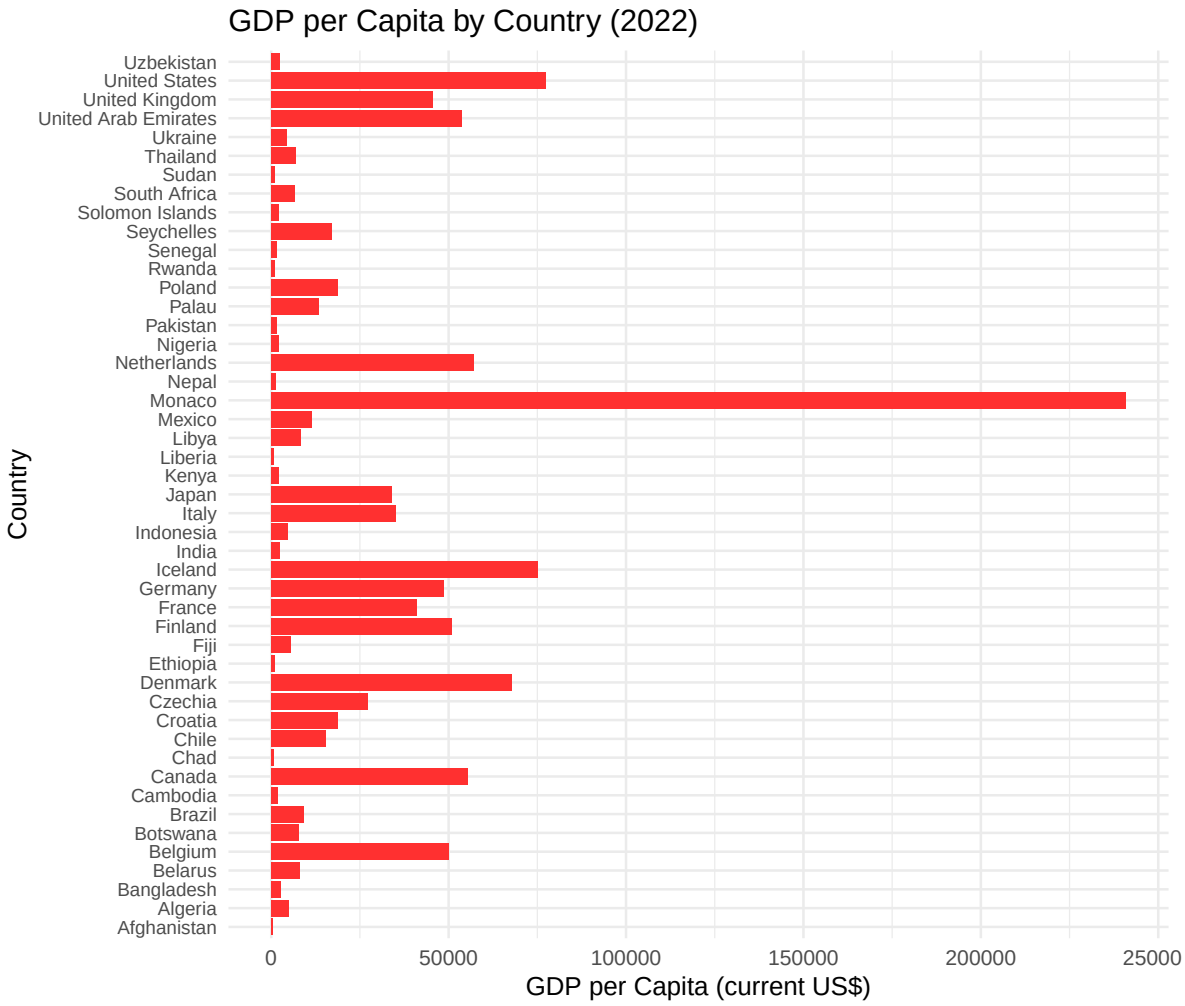
Country Name	mean_tourism
France	190032632
United States	125009356
Mexico	90666192
Poland	72186480
Italy	69282062

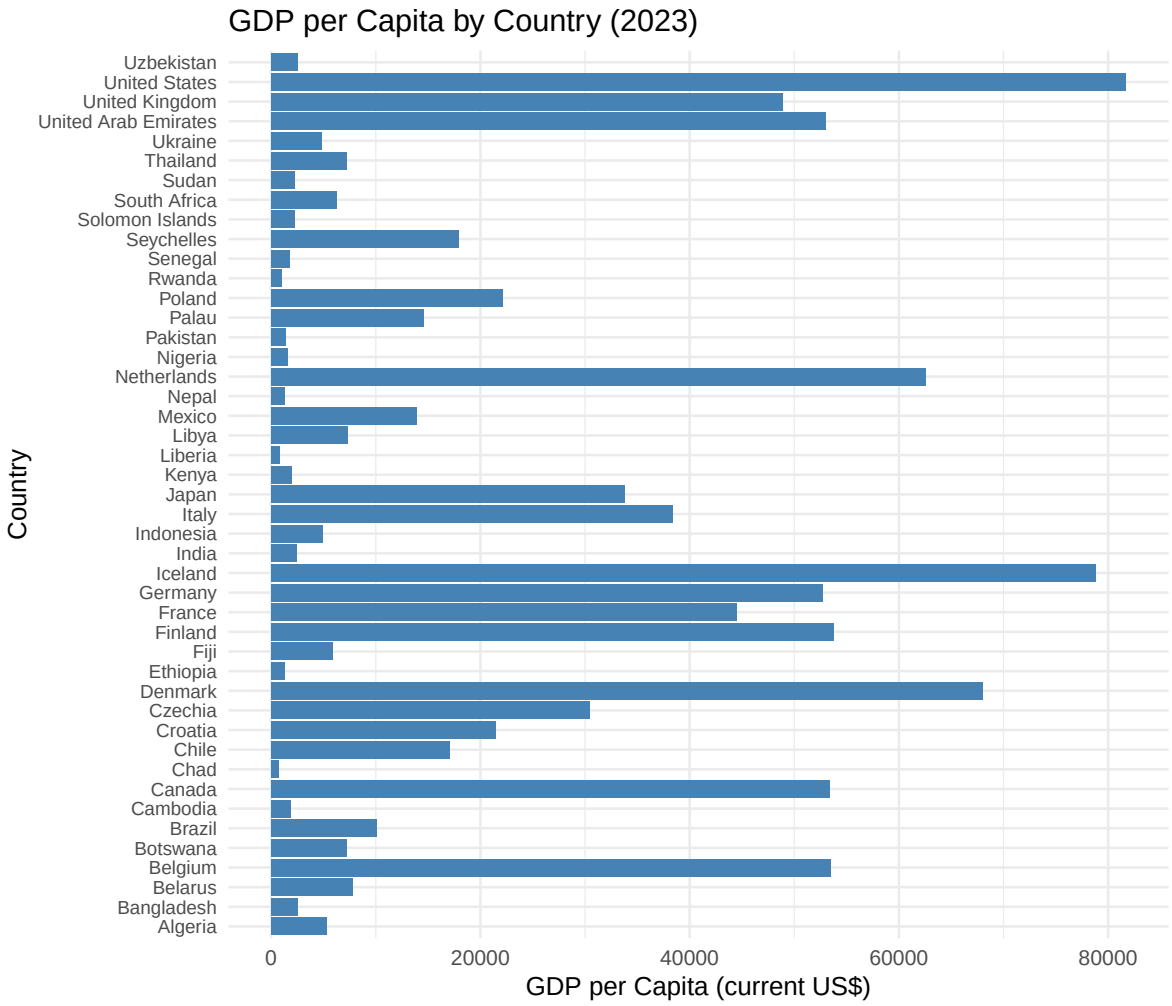
Table 13: The First Five Rows of the Research and development expenditure Summary Statistic

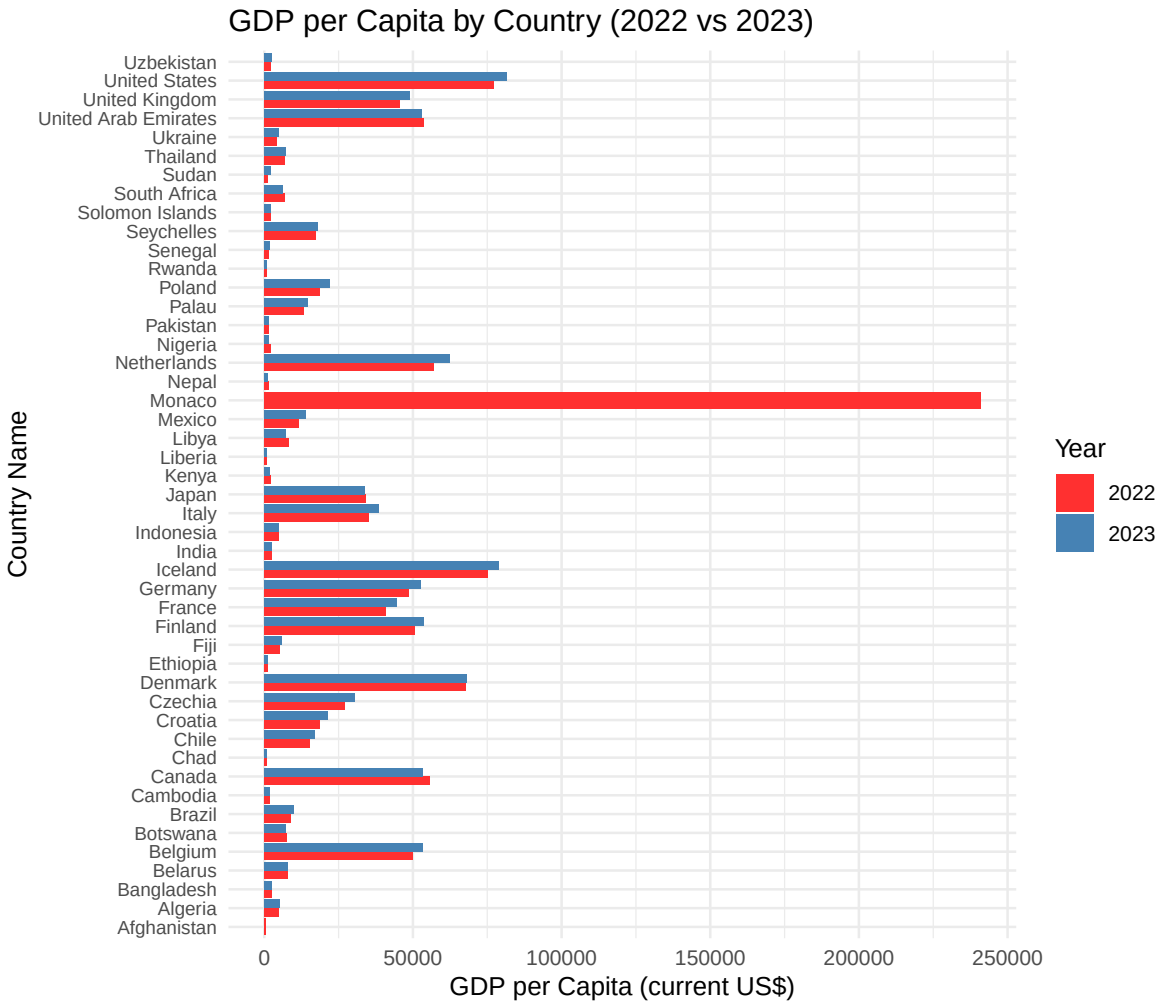
Country Name	mean_research_dev
Finland	3.130685
Japan	3.097945
United States	2.738588
Germany	2.668934
Denmark	2.627044

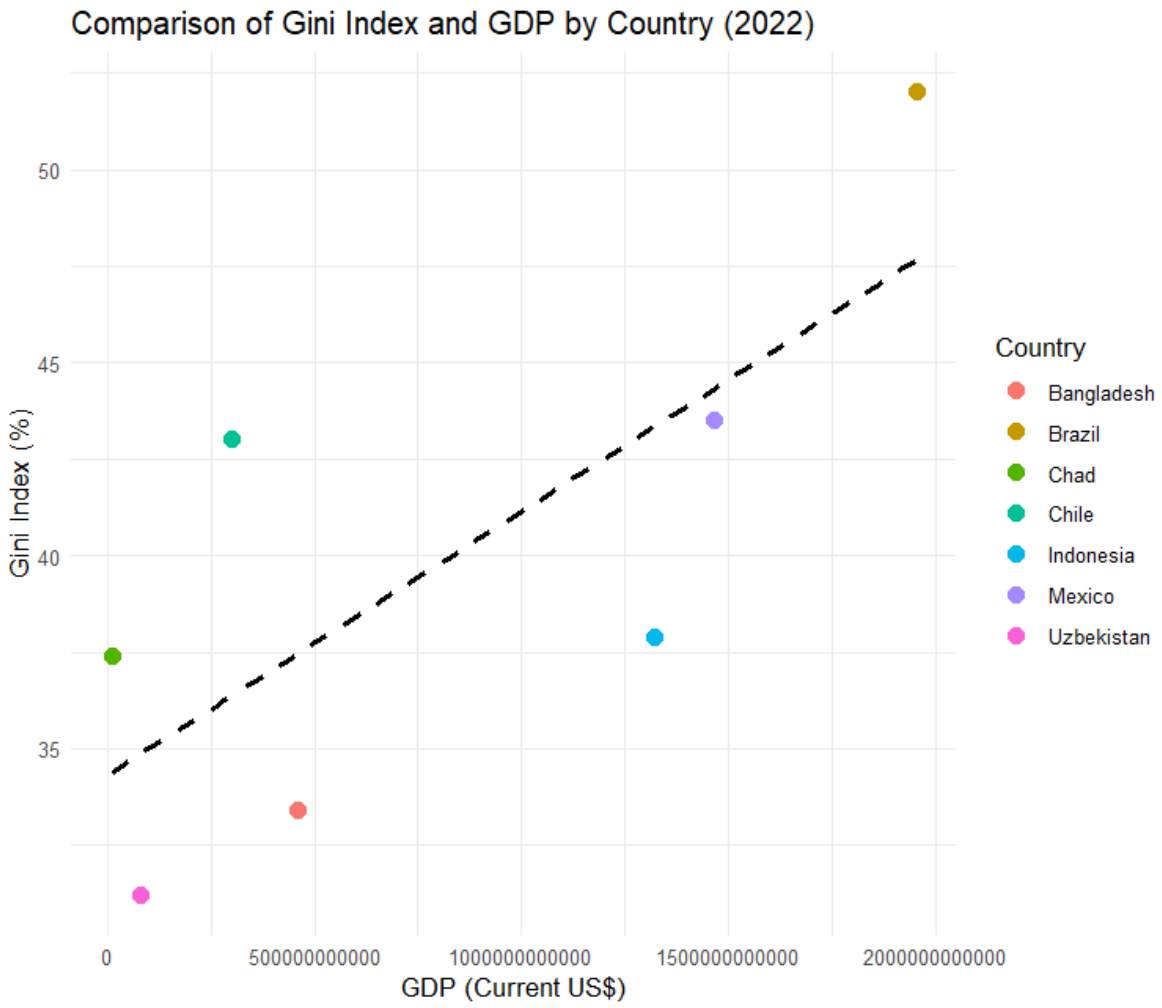
Visualizations


Percentage Change in GDP per Country
Uzbekistan
United States
United Kingdom
United Arab Emirates
Ukraine
Thailand
Sudan
South Africa
Solomon Islands
Seychelles
Senegal
San Marino
Rwanda
Poland
Palau
Pakistan
Nigeria
Netherlands
Nepal
Monaco
Mexico
Libya
Liberia
Kenya
Japan
Italy
Indonesia
India
Iceland
Germany
France
Finland
Fiji
Ethiopia
Denmark
Czechia
Croatia
Chile
Chad
Canada
Cambodia
Brazil
Botswana
Belgium
Belarus
Bangladesh
Algeria
Afghanistan
Country Name
1990
2000
2010
2020
Year
Percentage Change (%)
100
50
0
−50


GDP per Capita by Country (2022)
Country
Uzbekistan
United States
United Kingdom
United Arab Emirates
Ukraine
Thailand
Sudan
South Africa
Solomon Islands
Seychelles
Senegal
Rwanda
Poland
Palau
Pakistan
Nigeria
Netherlands
Nepal
Monaco
Mexico
Libya
Liberia
Kenya
Japan
Italy
Indonesia
India
Iceland
Germany
France
Finland
Fiji
Ethiopia
Denmark
Czechia
Croatia
Chile
Chad
Canada
Cambodia
Brazil
Botswana
Belgium
Belarus
Bangladesh
Algeria
Afghanistan
0
50000
100000
150000
200000
25000
GDP per Capita (current US$)

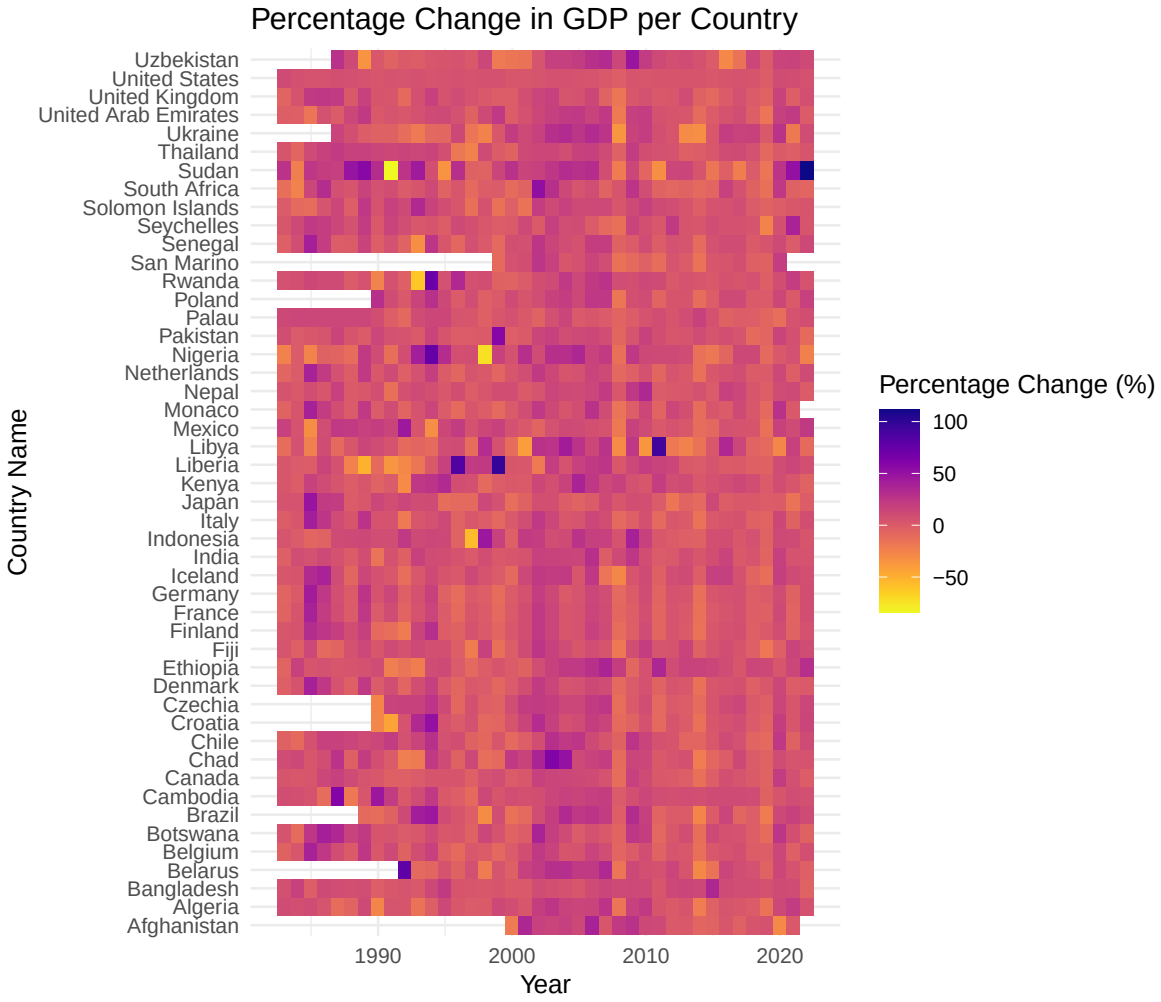

GDP per Capita by Country (2023)
Country
Uzbekistan
United States
United Kingdom
United Arab Emirates
Ukraine
Thailand
Sudan
South Africa
Solomon Islands
Seychelles
Senegal
Rwanda
Poland
Palau
Pakistan
Nigeria
Netherlands
Nepal
Mexico
Libya
Liberia
Kenya
Japan
Italy
Indonesia
India
Iceland
Germany
France
Finland
Fiji
Ethiopia
Denmark
Czechia
Croatia
Chile
Chad
Canada
Cambodia
Brazil
Botswana
Belgium
Belarus
Bangladesh
Algeria
0
20000
40000
60000
80000
GDP per Capita (current US$)


GDP per Capita by Country (2022 vs 2023)
Country Name
Uzbekistan
United States
United Kingdom
United Arab Emirates
Ukraine
Thailand
Sudan
South Africa
Solomon Islands
Seychelles
Senegal
Rwanda
Poland
Palau
Pakistan
Nigeria
Netherlands
Nepal
Monaco
Mexico
Libya
Liberia
Kenya
Japan
Italy
Indonesia
India
Iceland
Germany
France
Finland
Fiji
Ethiopia
Denmark
Czechia
Croatia
Chile
Chad
Canada
Cambodia
Brazil
Botswana
Belgium
Belarus
Bangladesh
Algeria
Afghanistan
0
50000
100000
150000
200000
250000
GDP per Capita (current US$)
Year
2022
2023



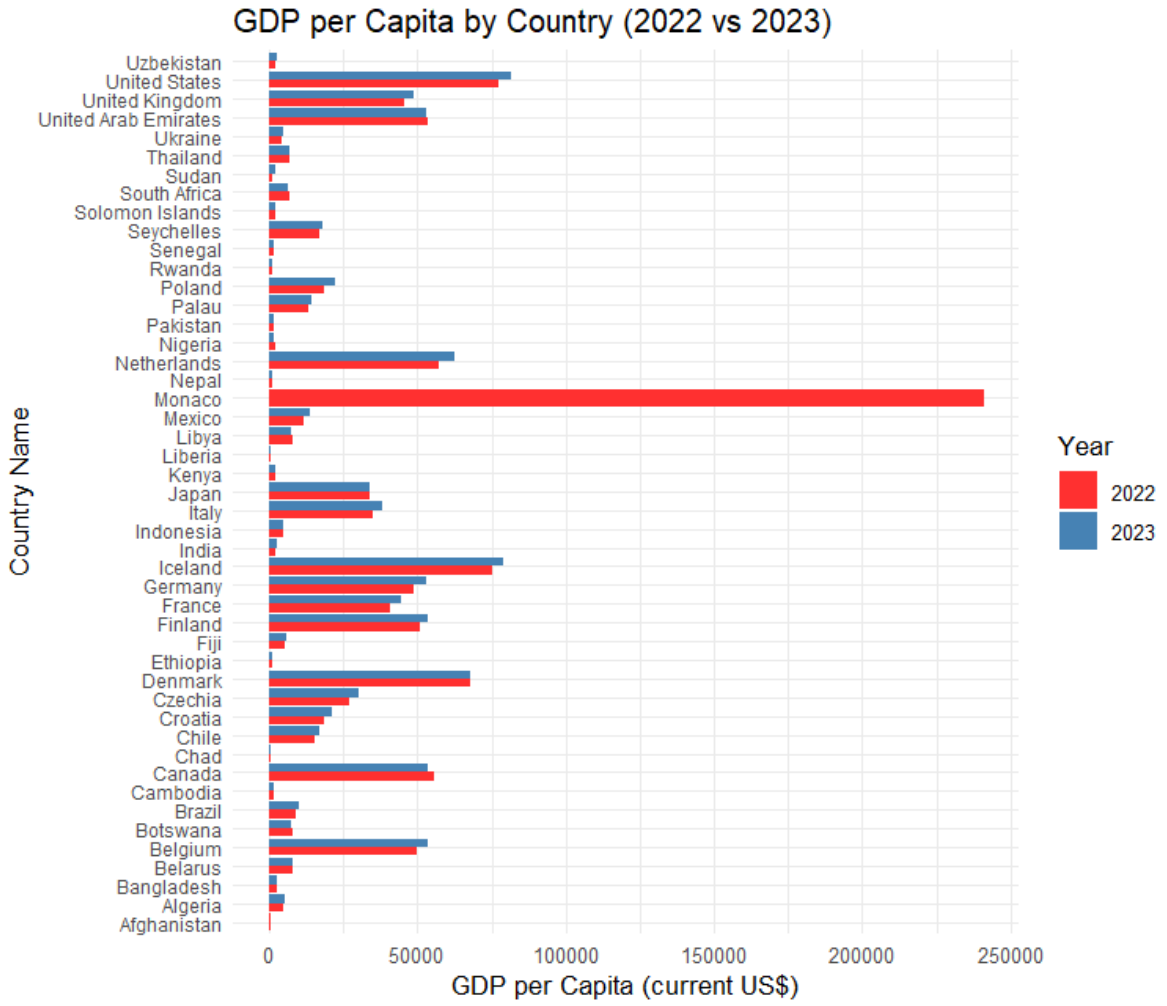
Reporting

How do factors such as year illustrate the variation in a country's GDP?



The graph presented above is a heat map comparing a country's percentage change in GDP per year. Unfortunately, not all countries had GDP values to work with for all years, but it was decided that these countries would not be excluded- as the information that is present in the data is still useful. While the general colour trends in the heat map reveal that the percentage change in GDP generally ranged between -50% to 50%, there are clear outliers. One such example of this that is evident in the graph is Liberia, which has values changes ranging from -50% to 100% in a matter of year. Additionally, the column of light yellow which appears through the graph just before 2010 after a period of more purple values implies the impact of the 2008 Financial Crisis on country's percentage change in GDP. The graph was initialized using percentage change in GDP due to the vastly fluctuating differences both between and within country's GDP values.

How does the variation in factors such as population effect a country's GDP?

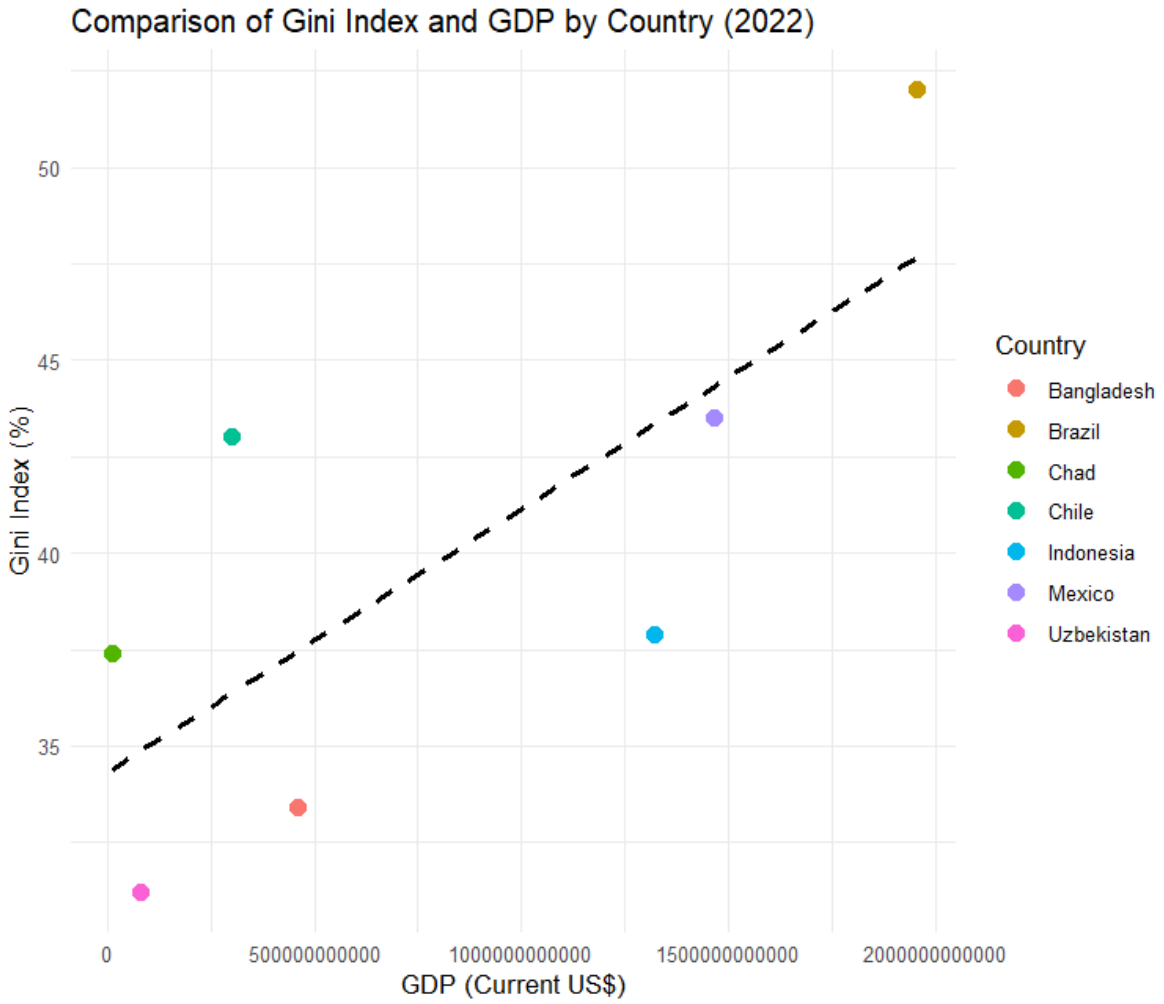


The above stacked bar graph considers the relationship between a country's population and it's GDP per capita coloured by year. The decision to colour the graph based on each country's 2022 and 2023 GDP per capita values was based in the vast difference in GDP per capita in some countries. The most interesting example of this is Monaco. As can be seen from the above graph, Monaco did not have a GDP value present in the dataset for 2023, but had a GDP per capita out ranking the United States in 2022. While the cause of the missing value is unknown, further research reveals that Monaco's GDP per capita was at an all time high in 2022. Although the United States has the highest GDP over 2022 and 2023, it is Monaco's small population and rich economy that makes it an outlier.

Aside from general annual growth in GDP, it is assumed that the trend occurring in some country's GDP per capita wherein their values for 2023 out rank 2022 are still recovering from the effects of COVID-19 on many country's economys. However, this is not the case for every country, as certain countries observed on the above graph have higher values for GDP per

capita in 2022 than 2023, such as Canada. Additionally, it appears from the graph that some country's GDPs per capita remained consistent between 2022 and 2023, such as Denmark and Japan.

What is the correlation between factors such as Gini index and a country's GDP?



The data available for the graph presented was limited to the seven countries which had values present for both Gini index and GDP in 2022. Even less data was available to illustrate the relationship between the two variables for 2023, and so the next best year was selected. It is also important to note that most of the country's listed in the above graph are low income, however, Brazil is considered a developing nation.

The line of best fit in the graph above illustrates the linear relationship between a country's GDP and it's Gini index ranking. The Gini index however, is measured as a percentage in which

the closer the value is to 0% the more equal a country is. This means that it can inferred from the graph above that the higher a country's GDP is, the more unequal their society. While this may be unexpected, one of the factors determining a country's Gini index is wealth distribution, meaning that country's with a low GDP may have a lower average income per person, but that members of the population are equal in their low income. Conversely, in countries with a higher GDP, the wealth disparities are more drastic, as is demonstrated in the graph above. However, it is important to note that the presence of any European countries is lacking from the graph above, as these country's tend to have higher GDPs and Gini index rates.

Conclusion

The report presented above explored and identified how factors such as year, population, and Gini index effect various measures of a country's GDP. This was achieved by conducting an exploratory data analysis on the dataset presented for Project 1. This required the cleaning and preparation of the data before the exploratory data analysis could be conducted, which involved loading the given dataset into R Studio, performing an initial inspection, inspecting column values, cleaning the dataset and then filtering it. The data could then be explored through the use of various summary statistics, such as a country's mean Gini index, and then further investigated using visualisations.

The visualisations presented in the above report explored the most interesting correlations and relationships that were presented in the given data in terms of a country's GDP. These correlations included the ways in which a country's percentage change in GDP varied by year, the variation in a country's GDP per capita by year, and the liner relationship between a country's GDP and Gini index rating.

While the exploratory data analysis process was successful, there were limitations- such as the amount of data missing from the World Bank Developement Indicators Service dataset. This was visible in all the graphs presented above, but particularly in the last graph which explored the relationship between a country's Gini index rating and GDP, in which only seven countries contained data for both values in 2022- meaning that the relationship could only be explored between seven countries in 2022. The initial pivoting of the data completed at the beginning of the data cleaning and preparation process did address much of this issue, however it's effects are evident in the visualisations presented.