GeoStat: Visualizing Country-Level Data for Deeper Insights

Student: Yazeed Mshayekh — TA: Mariam Elmasry

Tech for Jobs, Jan 10, 2024

Contents

1 | Project Description

1.1 | Problem Statement

The global landscape of countries presents a variety of challenges and opportunities across sectors such as economy, politics, and education. Understanding and analyzing key indicators like population, GDP, life expectancy, literacy rates, natural resources, and university rankings is crucial for policymakers, businesses, and researchers [1]. By exploring these variables, organizations can make informed decisions regarding investments, international relations, and sustainable development. This project aims to conduct an exploratory data analysis (EDA) of country-level data alongside university rankings to uncover patterns, correlations, and insights that can aid in understanding global dynamics [2].

The specific problem to be solved in this project is identifying and analyzing key patterns, correlations, and trends within country-level data and university rankings to understand global dynamics. Despite the availability of diverse datasets on various aspects of countries, there is a lack of comprehensive exploration that highlights how socio-economic, political, environmental, and educational factors interact across nations. By conducting an exploratory data analysis (EDA), the goal is to uncover hidden insights in areas like GDP, population, literacy rates, natural resources, and university performance. These insights can support better decision-making in global business strategies, policy formulation, and sustainable development practices [3].

1.2 | Possible Impact of Your Analysis

Exploring country-level data in conjunction with university rankings can have significant impacts across various domains, such as:

- **Guiding Strategic Decision-Making**: By analyzing global trends, socio-economic indicators, and academic performance, businesses and policymakers can identify high-growth markets, allocate resources more efficiently, and tailor strategies to specific regions. Understanding the link between a country's economic development and its universities' rankings can help prioritize investment in education and research [4].
- **Supporting Sustainable Development**: Understanding the interplay between environmental, economic, and social factors at the country level, as well as the academic contributions of universities, can help create policies that promote sustainability, innovation, and reduce inequality. Strong universities often contribute to national development through research, technology transfer, and human capital development [5].
- **Enhancing Investment and Trade Opportunities**: Identifying key patterns in GDP, literacy rates, natural resources, and university rankings can assist investors and trade organizations in spotting emerging markets and assessing the potential for economic growth. Universities in high-ranking countries may also be focal points for international collaborations and technological advancements, influencing global investment trends [6].

1.3 | Dataset(s)

The **"Countries of the World 2023"** dataset and the **"Global University Rankings Dataset 2023"** dataset from Kaggle provide valuable insights into global development, economic conditions, and higher education performance across countries. Both datasets are publicly available, offering a wealth of information that can be used for educational, research, and commercial purposes under open licenses, in accordance with Kaggle's terms of use.

- The "Countries of the World 2023" dataset includes a range of variables such as:
  - `Country`: Name of the country.
  - `Land Area (Km2)`: Total land area of the country in square kilometers.
  - `Population`: The country's total population.
  - `GDP`: Gross Domestic Product, the total value of goods and services produced in the country.
  - `Life Expectancy`: Average number of years a newborn is expected to live.

- `Gasoline_Price`: Price of gasoline per liter in local currency.
- `Agricultural Land (%)`: Percentage of land area used for agricultural purposes.
- `Unemployment Rate`: Percentage of the workforce unemployed but seeking work.
- `CPI`: Consumer Price Index, a measure of inflation and purchasing power.
- `CO2 Emissions`: Carbon dioxide emissions in tons.

- The "Global University Rankings Dataset 2023" dataset includes a range of variables such as:
  - `Rank`: The ranking position of the university in the global rankings.
  - `University Name`: The name of the university, identifying each institution uniquely.
  - `Location`: The geographical location of the university, indicating the country or region where it is situated.
  - `Number of Students`: The total number of students enrolled in the university.
  - `Number of Students per Staff`: The ratio of the total number of students to the total number of academic staff members, providing an indication of the student-to-faculty ratio.
  - `International Student`: The proportion of international students studying at the university, offering insights into its global appeal and diversity.
  - `Female - Male Ratio`: The gender distribution among the university's student body, presenting the ratio of female students to male students.

By combining these datasets, you can analyze the relationship between a country's socio-economic indicators (e.g., GDP, population, literacy rate) and the academic performance of universities within that country. This combination enables a deeper understanding of how national factors like economic stability, literacy, and public spending correlate with the performance and global ranking of universities. It also helps in identifying global trends in higher education and supports decision-making for students, policymakers, and businesses in both the education and economic sectors.

2 | Project Scoping Document

2.1 | Business Problem

The challenge in addressing global dynamics stems from the lack of a unified approach that combines key socio-economic indicators, environmental factors, and educational performance data to provide actionable insights for businesses and policymakers. Despite having access to diverse datasets, the analysis of how these variables influence each other and drive global trends remains limited. This lack of a comprehensive understanding makes it difficult for businesses to target high-growth markets, for policymakers to design effective and sustainable development strategies, and for governments to prioritize investments in areas like education and infrastructure. Additionally, universities, which are key drivers of innovation and socio-economic development, are often overlooked when evaluating a country's global standing, despite their pivotal role in national and international progress [7].

2.2 | Business Impact

The analysis of country-level data and university rankings provides several strategic advantages:

- **Enhanced Market Forecasting**: By identifying emerging economic and educational trends, businesses can more accurately forecast market changes and adapt their strategies to dynamic global conditions.
- **Policy Alignment with Global Trends**: Policymakers can leverage insights from the analysis to align national policies with broader global trends, improving the effectiveness of interventions in areas such as education, workforce development, and innovation.
- **Identification of Collaborative Opportunities**: By examining the intersection of university performance and national development, businesses can pinpoint opportunities for collaboration with academic institutions, enhancing innovation through joint research and technological advancements.
- **Data-Driven Investment**: Investors can use the insights to make data-driven decisions, identifying regions with the most promising socio-economic and educational conditions, thereby increasing the likelihood of high returns on investment.

This comprehensive analysis not only improves decision-making across various sectors but also empowers stakeholders to make proactive, informed decisions that contribute to both economic and social growth.

2.3 | Dataset(s)

Primary datasets

- **"Countries of the World 2023"**: This dataset provides comprehensive socio-economic and political data for over 190 countries. It includes key indicators such as GDP, population, life expectancy, literacy rate, unemployment rate, and natural resources, offering a broad view of a country's economic and social landscape.
- **"Global University Rankings Dataset 2023"**: This dataset contains rankings of universities worldwide, with indicators like teaching quality, research output, citations, industry income, and international outlook. It helps assess the role of educational institutions in national development and innovation.

Future Dataset Needs

- **Economic Indicators**: Additional datasets may be required that focus specifically on macroeconomic factors such as inflation rates, income inequality, and trade statistics for a more granular understanding of each country's economic standing.
- **Education & Research Data**: Further data on research spending, graduate employment rates, and university-industry collaborations may be explored to provide a more complete view of the impact of higher education on national economic performance.

2.4 | Methods

The Key Variables

- **Global Country Information Dataset 2023:**
  - **Country:** Name of the country.
  - **Density (P/Km2):** Population density (persons per square kilometer).
  - **Abbreviation:** Country's abbreviation or code.
  - **Agricultural Land (%):** Percentage of land area used for agricultural purposes.
  - **Land Area (Km2):** Total land area of the country in square kilometers.
  - **Armed Forces Size:** Size of the country's armed forces.
  - **Birth Rate:** Number of births per 1,000 population per year.
  - **Calling Code:** International calling code for the country.
  - **Capital/Major City:** Name of the capital or major city.
  - **CO2 Emissions:** Carbon dioxide emissions in tons.
  - **CPI:** Consumer Price Index, a measure of inflation and purchasing power.
  - **CPI Change (%):** Percentage change in CPI from the previous year.
  - **Currency_Code:** Currency code used in the country.
  - **Fertility Rate:** Average number of children born to a woman during her lifetime.
  - **Forested Area (%):** Percentage of land covered by forests.
  - **Gasoline_Price:** Price of gasoline per liter in local currency.
  - **GDP:** Gross Domestic Product, the total value of goods and services produced in the country.
  - **Gross Primary Education Enrollment (%):** Gross enrollment ratio for primary education.
  - **Gross Tertiary Education Enrollment (%):** Gross enrollment ratio for tertiary education.
  - **Infant Mortality:** Number of deaths per 1,000 live births before age one.
  - **Largest City:** Name of the largest city.
  - **Life Expectancy:** Average number of years a newborn is expected to live.
  - **Maternal Mortality Ratio:** Number of maternal deaths per 100,000 live births.
  - **Minimum Wage:** Minimum wage level in local currency.
  - **Official Language:** Official language(s) spoken in the country.
  - **Out of Pocket Health Expenditure (%):** Percentage of total health expenditure paid out-of-pocket.
  - **Physicians per Thousand:** Number of physicians per thousand people.
  - **Population:** Total population of the country.
  - **Labor Force Participation (%):** Percentage of the population that is part of the labor force.
  - **Tax Revenue (%):** Tax revenue as a percentage of GDP.
  - **Total Tax Rate:** Overall tax burden as a percentage of commercial profits.
  - **Unemployment Rate:** Percentage of the labor force that is unemployed.
  - **Urban Population:** Percentage of the population living in urban areas.
  - **Latitude:** Latitude coordinate of the country.
  - **Longitude:** Longitude coordinate of the country.
- **Global University Rankings Dataset 2023:**
  - **Rank:** The ranking position of the university globally.
  - **University Name:** Name of the university.
  - **Location:** The country or region where the university is located.
  - **Number of Students:** Total number of students enrolled.
  - **Number of Students per Staff:** Ratio of students to academic staff.
  - **International Student:** Percentage of international students.
  - **Female - Male Ratio:** Ratio of female students to male students.

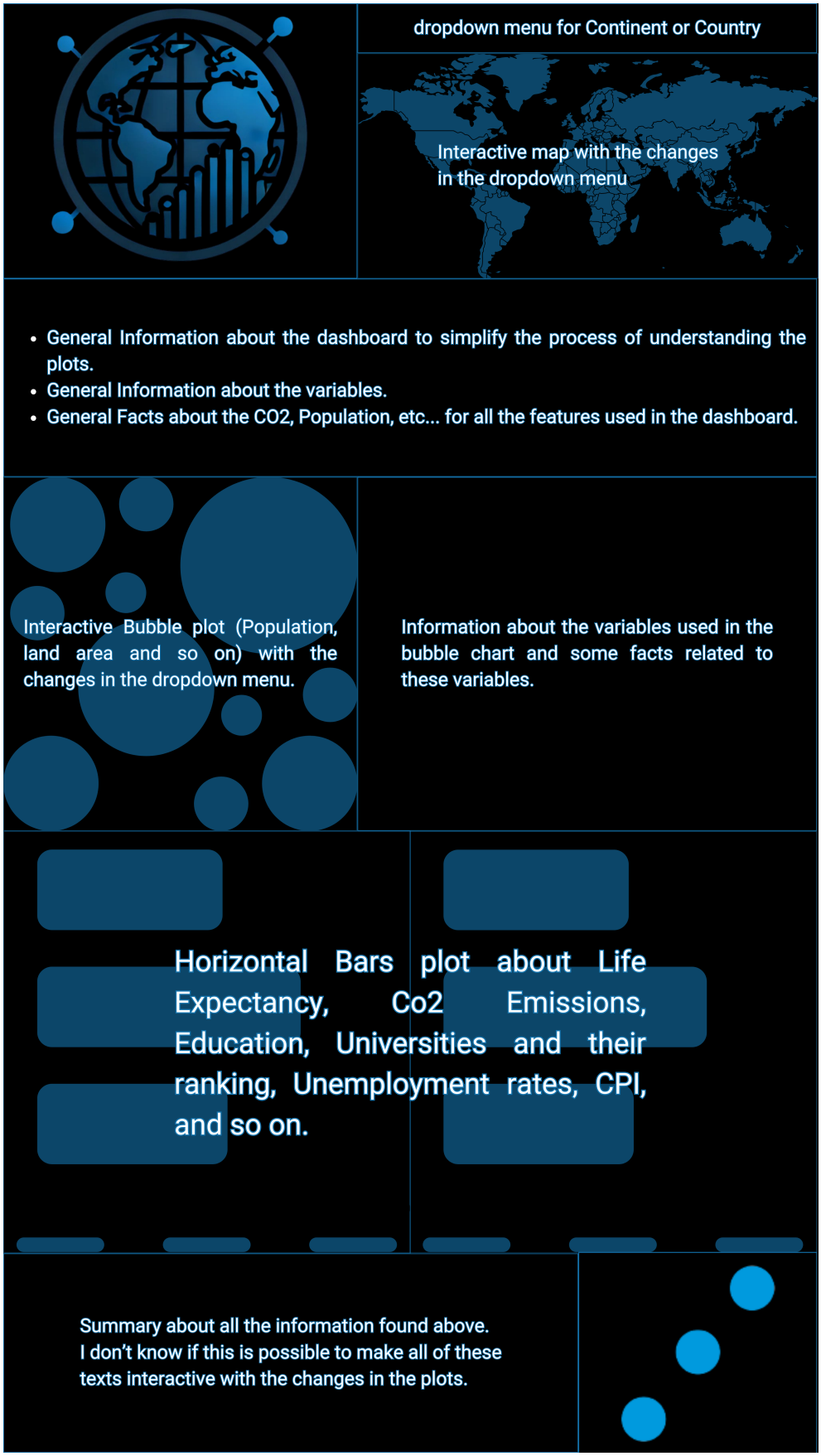
Analysis Plan

The analysis will focus on identifying correlations and trends between socio-economic factors (such as GDP, life expectancy, and unemployment rate) and university rankings. Here's a plan for analyzing the data:

- **Descriptive Analysis:**
  - Begin by exploring basic statistics such as the mean, median, and standard deviation for variables like GDP, life expectancy, and university rank.
  - Visualize distributions of key variables such as GDP, CO2 emissions, and university rank using histograms and box plots.
- **Correlation Analysis:**
  - Use Pearson's or Spearman's correlation coefficients to understand relationships between continuous variables, such as GDP and life expectancy, or CPI and university rankings.
  - Investigate correlations between educational variables (e.g., Gross Primary Education Enrollment, Gross Tertiary Education Enrollment) and economic indicators like GDP and labor force participation.
- **Comparative Analysis:**
  - Compare the university rankings across different regions or income categories (e.g., low-income vs. high-income countries).
  - Perform t-tests or ANOVA to check for significant differences in university rankings based on factors like the size of the armed forces, CO2 emissions, or minimum wage.
- **Clustering and Grouping:**
  - Apply clustering techniques (e.g., K-means clustering) to group countries based on similar socio-economic profiles and analyze if these groups correlate with university rankings.
  - Investigate if countries with high university rankings tend to share common characteristics in terms of economic performance, education enrollment rates, and other variables.
- **Geospatial Analysis:**
  - Use latitude and longitude data to map countries and their respective university rankings, creating a geographical heatmap to see if certain regions show consistent patterns in educational performance and socio-economic conditions.
- **Tools and Techniques:**
  - This analysis will utilize a combination of tools such as **Python**, for statistical modeling and data manipulation, **Tableau**, for interactive data visualization and dashboard creation, and possibly **Excel** for initial data exploration and basic analysis.

By combining these analyses and tools, you can uncover valuable insights into how socio-economic and educational factors are related to university rankings and overall global dynamics. The results could inform strategies for businesses, policymakers, and educational institutions to target high-growth markets, improve global collaboration, and support sustainable development.

2.5 | Dashboard



2.6 | Milestones

To achieve success in this project, the goal will be to focus on thorough data analysis, clear reporting, insightful visualizations, and interactive dashboards. Success will be driven by a step-by-step approach to uncover valuable insights and effectively communicate those insights.

- **First Milestone – Exploratory Data Analysis (EDA):**
  - The first milestone will focus on Exploratory Data Analysis (EDA), which involves thoroughly examining the datasets to understand their structure, distributions, and key relationships between variables.
  - **Key tasks include:**
    - **Data cleaning and preprocessing:** Ensuring the data is free from errors, dealing with missing values, and transforming variables as needed.
    - **Descriptive statistics:** Analyzing basic statistical measures like mean, median, and standard deviation to understand the distribution of key variables.
    - **Visualizations:** Generating visualizations like histograms, scatter plots, and box plots to explore the relationships between variables such as GDP, life expectancy, and university rankings.
- **Mid Milestones – Deeper Analysis and Insight Generation:**
  - Once EDA is complete, the next milestone will involve performing deeper correlation analysis to identify key patterns, trends, and relationships in the data.
  - The focus will be on uncovering correlations between socio-economic indicators (such as GDP and unemployment rates) and educational variables (such as university rankings and enrollment rates).
  - **Comparative analysis** will be done to examine differences between countries, regions, and income categories.
  - Tools like Python will be used for statistical analysis (such as Pearson or Spearman correlation) and for creating visualizations in libraries like Matplotlib or Seaborn.
- **Final Milestone – Reporting, Dashboards, and Visualizations:**
  - The final milestone will center around synthesizing the analysis into a clear, informative report and creating interactive dashboards.
  - **Reports** will summarize key findings, highlight trends and correlations, and offer actionable insights. The focus will be on presenting data in an understandable way that can inform decision-making.
  - **Dashboards** created in Tableau will allow users to interact with the data and explore different variables and trends. These dashboards will present dynamic visualizations to showcase relationships between socio-economic factors and university rankings.
  - The final report and dashboards will be packaged as a comprehensive deliverable, effectively communicating the insights from the data analysis.

2.7 | Timeline

Timeline	
Week	Tasks
Week 1	Project Description, Project Scoping
Week 2	Data Curation
Week 3	Exploratory Data Analysis
Week 4	Datafolio
Week 5	Dashboard
Week 6	Final Report

Table 2.1: Project Timeline

3 | Data Curation

3.1 | Data Sourcing

Global Country Information Dataset 2023

- **File Name:** world-data-2023.csv
- **Description:** This comprehensive dataset provides a wealth of information about all countries worldwide, covering a wide range of indicators and attributes. It encompasses demographic statistics, economic indicators, environmental factors, healthcare metrics, education statistics, and much more. With every country represented, this dataset offers a complete global perspective on various aspects of nations, enabling in-depth analyses and cross-country comparisons.
- **Dataset Details:** 195 Rows and 35 Columns
- **Size:** 49.2KB
- **Source:** kaggle

Global University Rankings Dataset 2023

- **File Name:** world-university-rank.csv
- **Description:** Discover the Global University Rankings Dataset 2023, a comprehensive and insightful compilation of the world's top universities. Uncover the rankings, metrics, and key performance indicators of renowned academic institutions worldwide.
- **Dataset Details:** 2345 Rows and 7 Columns
- **Size:** 171.28 KB
- **Source:** Kaggle

3.2 | Data Profiling

Global Country Information Dataset 2023 Analysis

Preliminary Data Analysis

- **Check duplicates:** Inspect the DataFrame for any duplicates to gain a general insight.
- **Rows and Columns:** The dataset contains 195 rows and 35 columns.
- **Null values and data types:** Use the `.info()` function to examine the null values and data types of each column.
- **Null and Missing Values:** Check for missing data.

- **Statistics:** Get statistics of all columns (both numerical and categorical).
- **Unique values:** Check unique values in each column.
- **Correlation Analysis:** Due to the large number of features, the correlation matrix may become difficult to read. Therefore, a selection of key features will be analyzed. These features cover a wide range of economic, social, environmental, and demographic factors of countries.

Selected Features for Analysis

- **Interesting features:**
  - GDP
  - Life expectancy
  - Population
  - Urban population
  - Infant mortality
  - Unemployment rate
  - Tax revenue (%)
  - CO2-Emissions
  - Agricultural Land (%)
  - Fertility Rate
- **Data transformation:** Transform columns with object data type to numeric.

Observations

- **Duplicate Columns:** There are 0 columns that contain duplicates.
- **Missing Data:** The percent of missing data is 4.94%.
- **Missing values by column:**
  - `Country` - 0 missing values
  - `Density (P/Km2)` - 0 missing values
  - `Abbreviation` - 7 missing values
  - `Agricultural Land(%)` - 7 missing values
  - `Land Area(Km2)` - 1 missing value
  - `Armed Forces size` - 24 missing values
  - `Birth Rate` - 6 missing values
  - `Calling Code` - 1 missing value
  - `Capital/Major City` - 3 missing values
  - `CO2-Emissions` - 7 missing values
  - `CPI` - 17 missing values
  - `CPI Change (%)` - 16 missing values
  - `Currency-Code` - 15 missing values
  - `Fertility Rate` - 7 missing values
  - `Forested Area (%)` - 7 missing values
  - `Gasoline Price` - 20 missing values
  - `GDP` - 2 missing values
  - `Gross primary education enrollment (%)` - 7 missing values
  - `Gross tertiary education enrollment (%)` - 12 missing values
  - `Infant mortality` - 6 missing values

- `Largest city` - 6 missing values
- `Life expectancy` - 8 missing values
- `Maternal mortality ratio` - 14 missing values
- `Minimum wage` - 45 missing values
- `Official language` - 1 missing value
- `Out of pocket health expenditure` - 7 missing values
- `Physicians per thousand` - 7 missing values
- `Population` - 1 missing value
- `Population: Labor force participation (%)` - 19 missing values
- `Tax revenue (%)` - 26 missing values
- `Total tax rate` - 12 missing values
- `Unemployment rate` - 19 missing values
- `Urban population` - 5 missing values
- `Latitude` - 1 missing value
- `Longitude` - 1 missing value

- **Mean for Numerical Columns:** Calculate the mean for all numerical columns.
- **Mode for Categorical Columns:** Replace NaN values with the most frequently occurring value (mode).
- **NaN Check:** Verify if there is any remaining NaN value.

Unique Values in Each Column

- `Country`: 195 distinct values
- `Density (P/Km2)`: 137 distinct values
- `Abbreviation`: 188 distinct values
- `Agricultural Land(%)`: 168 distinct values
- `Land Area(Km2)`: 194 distinct values
- `Armed Forces size`: 105 distinct values
- `Birth Rate`: 171 distinct values
- `Calling Code`: 183 distinct values
- `Capital/Major City`: 192 distinct values
- `CO2-Emissions`: 184 distinct values
- `CPI`: 175 distinct values
- `CPI Change (%)`: 86 distinct values
- `Currency-Code`: 133 distinct values
- `Fertility Rate`: 140 distinct values
- `Forested Area (%)`: 161 distinct values
- `Gasoline Price`: 101 distinct values
- `GDP`: 193 distinct values
- `Gross primary education enrollment (%)`: 141 distinct values
- `Gross tertiary education enrollment (%)`: 171 distinct values

- `Infant mortality`: 145 distinct values
- `Largest city`: 188 distinct values
- `Life expectancy`: 135 distinct values
- `Maternal mortality ratio`: 115 distinct values
- `Minimum wage`: 114 distinct values
- `Official language`: 77 distinct values
- `Out of pocket health expenditure`: 160 distinct values
- `Physicians per thousand`: 153 distinct values
- `Population`: 194 distinct values
- `Population: Labor force participation (%)`: 145 distinct values
- `Tax revenue (%)`: 119 distinct values
- `Total tax rate`: 156 distinct values
- `Unemployment rate`: 164 distinct values
- `Urban population`: 190 distinct values
- `Latitude`: 195 distinct values
- `Longitude`: 195 distinct values

Data Transformation

- **Columns to Convert to Float:** The following columns are transformed from strings to numeric (float) data type:
  - 'Density (P/Km2)', 'Agricultural Land (%)', 'Land Area (Km2)', 'Birth Rate', 'CO2-Emissions', 'Forested Area (%)', 'CPI', 'CPI Change (%)', 'Fertility Rate', 'Gasoline Price', 'GDP', 'Gross primary education enrollment (%)', 'Armed Forces size', 'Gross tertiary education enrollment (%)', 'Infant mortality', 'Life expectancy', 'Maternal mortality ratio', 'Minimum wage', 'Out of pocket health expenditure', 'Physicians per thousand', 'Population', 'Population: Labor force participation (%)', 'Tax revenue (%)', 'Total tax rate', 'Unemployment rate', 'Urban population'.
- **Transformation Process:**
  - Convert values to string.
  - Remove commas, dollar signs, and percentage signs.
  - Convert the cleaned strings to float.

Correlation Analysis

- **Highly Correlated Pairs:**
  - CO2-Emissions / GDP: 0.92 (highly positive correlation)
  - CO2-Emissions / Urban population: 0.93 (highly positive correlation)
  - Urban population / Population: 0.95 (highly positive correlation)
  - Infant mortality / Life expectancy: -0.93 (highly negative correlation)

Global University Rankings Dataset 2023 Analysis

Preliminary Data Analysis

- **Check duplicates:** Inspect the DataFrame for any duplicates to gain a general insight.
  - **Number of duplicate rows:** 0
- **Rows and Columns:** The dataset contains 2345 rows and 7 columns.
- **Null values and data types:** Use the `.info()` function to examine the null values and data types of each column.
- **Missing values:**
  - Rank: 0 missing values
  - University name: 0 missing values
  - Location: 111 missing values
  - Number of Students: 0 missing values
  - Number of students per staff: 1 missing value
  - International Students: 0 missing values
  - Female : Male ratio: 90 missing values
- **Drop null values:** Drop null values where necessary.
- **Statistics:** Generate statistics for all columns (both numerical and categorical).
- **Unique values:** Check unique values in each column.

Correlation Analysis

- **Correlation Matrix:**

Correlation Matrix					
Variable	**Number of Students**	**Number of students per staff**	**International Students**	**Female**	**Male**
Number of Students	1.000	0.312	0.023	0.113	-0.113
Number of students per staff	0.312	1.000	-0.078	0.002	-0.002
International Students	0.023	-0.078	1.000	0.140	-0.140
Female	0.113	0.002	0.140	1.000	-1.000
Male	-0.113	-0.002	-0.140	-1.000	1.000

Table 3.1: Correlation Analysis Between Key Variables

Data Transformation

- **Convert object to numeric:**
  - `Number of Students`: `int64`
  - `Number of students per staff`: `float64`
  - `International Students`: object (requires processing)
    - Replace the '%' symbol and convert to the appropriate format:

```
df['International Student'] = df['International Student'].str.replace('%', ' ')
```

  - `Number of Students`: Remove commas and convert to integer:

```
df['Number of Students'] = df['Number of Students'].str.replace(',', '').astype(int)
```

- **Split 'Female : Male Ratio'**: Create new columns for 'Female' and 'Male' based on the 'Female : Male Ratio' column.
- **Drop the original 'Female : Male Ratio' column** after the split.

3.3 | Data Wrangling

Feature	Type	Description
`country`	STRING	The name of the country.
`density p km2`	FLOAT	The population density of the country, measured in people per square kilometer.
`abbreviation`	STRING	The country abbreviation (two-letter ISO code).
`agricultural land percent`	FLOAT	The percentage of land used for agricultural purposes.
`land area km2`	FLOAT	The total land area of the country in square kilometers.
`armed forces size`	FLOAT	The size of the armed forces in the country (in number).
`birth rate`	FLOAT	The birth rate in the country, measured in births per 1,000 people per year.
`calling code`	INTEGER	The international calling code for the country.
`capital major city`	STRING	The capital or major city of the country.
`co2 emissions`	FLOAT	The CO2 emissions of the country, measured in metric tons.
`cpi`	FLOAT	The Consumer Price Index (CPI), which measures the changes in prices of goods and services.
`cpi change percent`	FLOAT	The percentage change in the Consumer Price Index (CPI) from the previous period.
`currency code`	STRING	The official currency code of the country.
`fertility rate`	FLOAT	The fertility rate of the country, representing the number of children born per woman.
`forested area percent`	FLOAT	The percentage of the country's area covered by forests.
`gasoline price`	FLOAT	The price of gasoline in the country, measured in USD per liter or gallon.
`gdp`	FLOAT	The Gross Domestic Product (GDP) of the country, measured in USD.
`gross primary education enrollment percent`	FLOAT	The percentage of primary school-aged children enrolled in school.
`gross tertiary education enrollment percent`	FLOAT	The percentage of tertiary (university) education enrollment in the country.
`infant mortality`	FLOAT	The number of deaths of infants under one year old, per 1,000 live births.
`largest city`	STRING	The largest city in the country by population.
`life expectancy`	FLOAT	The average life expectancy of a person in the country.
`maternal mortality ratio`	FLOAT	The maternal mortality ratio, which measures deaths due to pregnancy-related causes, per 100,000 live births.
`minimum wage`	FLOAT	The minimum wage set by the country, measured in USD.
`official language`	STRING	The official language(s) of the country.
`out of pocket health expenditure`	FLOAT	The percentage of health expenditure that is paid directly by the individuals in the country.

physicians per thousand	FLOAT	The number of physicians per 1,000 people in the country.
population	FLOAT	The total population of the country.
population labor force participation percent	FLOAT	The percentage of the working-age population that is part of the labor force.
tax revenue percent	FLOAT	The total tax revenue of the country, as a percentage of GDP.
total tax rate	FLOAT	The total tax rate as a percentage of income, including all forms of taxes.
unemployment rate	FLOAT	The percentage of the total labor force that is unemployed.
urban population	FLOAT	The population of people living in urban areas.
latitude	FLOAT	The geographic latitude of the country's central point.
longitude	FLOAT	The geographic longitude of the country's central point.

Table 3.2: Dataset Column Information for Global Country Dataset 2023

Field	Type	Description
rank	STRING	The ranking position of the university.
university name	STRING	The name of the university.
location location	STRING	The location or country where the university is based.
number of studnet	STRING	The total number of students enrolled at the university.
number of student per staffs	FLOAT	The number of students per staff member at the university.
international student	STRING	The percentage of international students at the university.
female male ratio	STRING	The gender ratio of female to male students at the university.

Table 3.3: Dataset Column Information for Global University Rankings 2023

4 | Exploratory Data Analysis (EDA)

Here is the notebooks.

1. Analysis #1 - Country-Level Overview

Key Indicators Examined

- **GDP:** Countries with a GDP above $1 trillion tend to show higher life expectancy (average of 80 years) and education enrollment rates (average of 90% for primary and secondary).
- **Life Expectancy:**
  - Highest: Monaco – 89 years
  - Lowest: Lesotho – 50 years
- **GDP per capita:**
  - Highest GDP per capita: Luxembourg – $116,000
  - Lowest GDP per capita: Burundi – $264
- **Education Enrollment:**
  - Highest enrollment rate for primary education: Russia – 100%
  - Lowest enrollment rate for primary education: Ethiopia – 50%

2. Analysis #2 - Education and Health Indicators Distribution

Education Enrollment

- Primary Education Enrollment: On average, 90% of countries have above 80% enrollment in primary education.
- Tertiary Education Enrollment: Countries like Australia (98%) and United States (95%) show near-complete tertiary education enrollment.

Infant Mortality

- Lowest: Japan – 2 deaths per 1,000 live births
- Highest: Nigeria – 100 deaths per 1,000 live births

Life Expectancy vs. GDP

- Countries with a GDP per capita over $20,000 show an average life expectancy of 82 years, whereas countries with GDP per capita below $5,000 have an average life expectancy of 62 years.

3. Analysis #3 - University Rankings and Distribution

Top 5 Universities by Ranking

- 1st: University of Oxford (UK)
- 2nd: Harvard University (USA)
- 3rd: University of Cambridge (UK)
- 4th: Stanford University (USA)
- 5th: Massachusetts Institute of Technology (USA)

Student-to-Staff Ratio

- Highest: Harvard University – 9.6 students per staff
- Lowest: University of Oxford – 10.6 students per staff

International Students

- Harvard University: 42% of students are international.
- Stanford University: 24% of students are international.

Top Universities Region-wise

- North America dominates with 60% of the top 50 universities being located in the US and Canada.
- Europe follows with 30%, especially United Kingdom and Germany.

Gender Ratios at Top Universities

- Harvard University: Gender ratio is 50% female and 50% male.
- University of Oxford: Gender ratio is 48% female and 52% male.
- Stanford University: Gender ratio is 46% female and 54% male.

4. Analysis #4 - Regional Breakdown of Rankings

Regional Distribution of Top Universities

- North America: 40% of universities in the top 100 are from the United States, with Harvard, Stanford, and MIT leading the charge.
- Europe: The United Kingdom contributes 12% to the top 100 universities, with University of Oxford and University of Cambridge consistently ranking in the top 5.
- Asia: National University of Singapore (NUS) ranked 11th, showing significant growth in the region.

5. Conclusion with Numbers and Facts

Country-Level Data

- There is a strong correlation between GDP and life expectancy: Countries with a GDP per capita of $20,000 or more exhibit an average life expectancy of 82 years, compared to 62 years in countries with GDP per capita below $5,000.
- Urbanization correlates with higher education enrollment: Germany and France, with high urban populations, report 99% enrollment in primary education.

University Rankings Data

- Top-ranking universities consistently show a student-to-staff ratio below 12 students per staff, indicating that higher rankings often correlate with more efficient staff management.
- International student percentages: Top universities like Harvard (42%) and University of Oxford (39%) show significant diversity, suggesting international appeal.
- Gender ratio: Harvard has a 50:50 male-to-female student ratio, demonstrating efforts to maintain gender balance.

5 | References

[1] Sheldon Smith. *Human systems ecology: studies in the integration of political economy, adaptation, and socionatural regions*. Routledge, 2019.

[2] Nishi Doshi, Samhitha Gundam, and Bhaskar Chaudhury. Strategizing university rank improvement using interpretable machine learning and data visualization. *arXiv preprint arXiv:2110.09050*, 2021.

[3] Arif Eser Guzel, Unal Arslan, and Ali Acaravci. The impact of economic, social, and political globalization and democracy on life expectancy in low-income countries: are sustainable development goals contradictory? *Environment, Development and Sustainability*, 23(9):13508–13525, 2021.

[4] Zoljargal Dembereldorj, Garmaa Dangaasuren, and Davaa Jagdag. Relationships between university performances and economic growth. *International Journal of Higher Education*, 7(4):123–132, 2018.

[5] Marta Addis. The key role of universities in sustainable development: The human dimension among the goals of 2030 agenda–a comparison between italy and spain. In *Sustainability in Higher Education: Strategies, Performance and Future Challenges*, pages 349–361. Springer, 2024.

[6] Oleg Bazaluk, Sheikh Abdul Kader, Nurul Mohammad Zayed, Rupok Chowdhury, Md Zahirul Islam, Vitalii S Nitsenko, and Hanna Bratus. Determinant on economic growth in developing country: A special case regarding turkey and bangladesh. *Journal of the Knowledge Economy*, pages 1–25, 2024.

[7] Maia Chankseliani, Ikboljon Qoraboyev, and Dilbar Gimranova. Higher education contributing to local, national, and global development: new empirical and conceptual insights. *Higher Education*, 81(1):109–127, 2021.