# An Exploratory Analysis of Economics and Education

## 1.0 Introduction

The following is an exploratory research project within the fields of Economics and Education. The focus of this project is to showcase the principles of data science and how they can be used in a methodical approach. This project is split into three main parts: data collection and cleaning, data analysis, and data visualization. The subject of this project has two main focuses: economic and educational metrics. The economic data pertains to GDP and GDP per capita, while the educational data contains various markers for educational quality. More information on the data sources can be found in section 2.0.

Three main research goals are formulated for this project, each of which contains two or more research questions:

1. What are some initial findings present in the individual data sources?
    a. These questions are answered using individual data sources.
        i. Which countries have experienced the highest GDP growth? The highest per capita growth?
        ii. What are the min, max, mean, median, and standard deviation of GDP and GDP per capita in 2021, 2022, and 2023?
        iii. What were the average GDP and per capita growth between 2021-2023?
        iv. Which countries have the highest primary education enrollment? The highest tertiary education enrollment?
        v. Which countries have the lowest unemployment rates?

2. Using all data sources in unison, can we find connections between them?
    a. These questions look for correlations between the content of the three sources, answered using an aggregate dataset.
        i. Are there any correlations between education markers, and economic status?
        ii. Which of the given factors (Unemployment rate, university enrollment, literacy rates, etc.), if any, best predict GDP or GDP per capita growth?

3. How can we use visualizations to further understand the data?
    a. These questions make use of both the individual and aggregate datasets to present meaningful visualizations.
        i. How does economic growth (GDP and GDP per capita), since 2021, compare between continents? (2 Visualizations)
        ii. What is the current global GDP distribution between continents?
        iii. Is there a correlation between tertiary education enrollment and GDP per capita?

## 2.0 Data Collection and Wrangling

The data for this project was collected from three different sources. The first of which is an education dataset from Kaggle. This dataset contains 202 rows, with 29 columns. However, most of the columns contained many missing values. Therefore, condensation and wrangling of the dataset led to the final set containing 202 rows, and 5 columns (6 when including country labels). Each row represents a country, with the columns representing birth rate, average youth literacy rate (YLR), gross primary education enrollment (PEE), gross tertiary education enrollment (TEE), and unemployment rate.

The second and third datasets collected were from Wikipedia using the Requests and BeautifulSoup libraries. These datasets contained nominal GDP and GDP per capita information for the years 2021, 2022, and 2023. After scraping the data from their respective wikitables, and converting them to Pandas dataframes, we attain two separate dataframes, with 181 and 194 rows (countries) respectively. There are 4 columns in each of these datasets (5 when including country labels), representing continent, and the GDPs (and per Capita) of 2021, 2022, and 2023. These two datasets were combined by country, making a joint dataset of 180 rows (countries). In addition to the existing columns, two new columns were created using the existing data: 'Percentage growth from 2021 - 2023' (GDP), and 'Per Capita Percentage growth from 2021 - 2023'. The calculation for these columns is simply the difference between GDP (and per capita) in 2023 and 2021 divided by GDP (and per capita) in 2023.

Two new datasets were created for analysis. The first is a combined set of the GDP (Combined) and education datasets. This 'Combined Dataset' is used for analysis of correlations and regression models. This dataset has 167 rows (countries), and 10 columns. It is important to note that certain countries were not represented in all of the datasets, which is why this dataset contains fewer rows. The second dataset, 'Continent Dataset', is essentially a condensed version of the combined GDP dataset. This dataset is used to answer questions on global distribution of GDP, and continental GDP growth. For each continent, the total nominal GDP is computed, along with the average nominal GDP per capita, and the average GDP / GDP per capita growths.

### 2.1 Data Collection Challenges and Resolution

Initially, the entire education dataset from Kaggle was meant to be used. This would have meant an additional 22 columns of data to be used for analysis. The main challenge encountered was the plethora of missing data points within the dataset. To resolve this matter, a condensation of the dataset was necessary. Rather than omitting rows that contained missing values, the columns that contained many missing values were removed. More strategies for future work can be found in section 4.

The second challenge was encountered in both Wikipedia sources. A similar issue of missing data points, due to the nature of web-scraping, led to greater difficulty in downloading the dataset. To overcome this challenge, a conditional statement defining a specific number of data points per row was used to ensure a complete dataset. This led to the loss of 32 and 28 countries from the GDP and GDP per capita data sources, respectively. Strategies for improvement in future work can be found in section 4.

## 3.0 Statistical Analysis

To answer the questions regarding the initial insights, techniques of dataset sorting and basic statistical analysis were used. First, analysis was done on the combined GDP dataset, with the following results:

- The top 5 countries experiencing GDP growth were Guyana, Liberia, Armenia, Seychelles, and Georgia respectively.
- The top 5 countries experiencing per capita growth were Guyana, Ethiopia, Seychelles, Liberia, and the Maldives.
- The average global GDP growth is 11.97%.
- The average per capita growth is 10.9%.

**Table 1.0 Statistical measures of GDP and GDP per capita in 2021, 2022, and 2023.**

| Measure | 2023 GDP (USD Million) | 2022 GDP (USD Million) | 2021 GDP (USD Million) | 2023 GDP per capita (USD) | 2022 GDP per capita (USD) | 2021 GDP per capita (USD) |
|---|---|---|---|---|---|---|
| Min | 63.00 | 60.00 | 60.00 | 246.00 | 238.00 | 311.00 |
| Max | 26949643.00 | 25462700.00 | 23315081.00 | 135605.00 | 126426.00 | 133745.00 |
| Mean | 574725.52 | 549899.84 | 529850.74 | 17922.42 | 16899.77 | 16101.19 |
| Median | 46732.00 | 44975.00 | 40433.00 | 7220.00 | 6790.00 | 6312.50 |
| Standard Deviation | 2465800.68 | 2378780.62 | 2252764.90 | 23858.63 | 22846.69 | 22142.13 |

Similar techniques were used to analyze the education dataset, providing the following results:

- The top 5 countries with primary education enrollment are Malawi, Madagascar, Nepal, Gabon, and Rwanda respectively.
- The top 5 countries with tertiary education enrollment are Greece, Australia, Grenada, South Korea, and Argentina respectively.
- The top 5 countries with the lowest unemployment rates are Qatar, Niger, the Solomon Islands, Laos, and Cambodia respectively.

When looking for a correlation between the variables (GDP, GDP per capita, GDP growth (%), GDP per capita growth (%), Birth rate, PEE, TEE, YLR, and Unemployment rate) a simple correlation matrix, and subsequent heatmap were used, leading to the following results: The highest correlation was found between GDP growth and GDP per capita growth (0.93), then Birth rate and TEE (-0.62), followed by GDP per capita and TEE (0.52), as well as Birth rate and GPD per capita (-0.52). A heatmap depicting the correlation matrix can be found in the Appendix.

To dive further into the correlation between the variables mentioned above, multiple regression models were fit. Four separate models were fit using different dependent variables: GDP, GDP per capita, GDP growth, and GDP per capita growth. All four models utilized the same five predictors: Birth rate, PEE, TEE, YLR, and Unemployment rate. By looking at the adjusted R-squared value, the best of these models is the one that predicts GDP per capita: $R^2_{Adj} = 0.592$. In this model, all of the predictors were found to be statistically significant, except the Unemployment rate. This led to the removal of this Unemployment rate as a predictor and re-fitting of the regression model. The new model's goodness-of-fit decreases to: $R^2_{Adj} = 0.585$.

However, all predictors in this model are statistically significant. Moreover, the AIC and BIC (Scoring criterion used for model selection) values for both models are essentially equal.

### 3.1 Visualizations

Following these analyses, visualizations are used to further build on the prior results. Figure 1 shows the average growth (GDP and GDP per capita) in each continent through a barplot. It is evident that the Americas have seen, on average, the greatest growth since 2021.

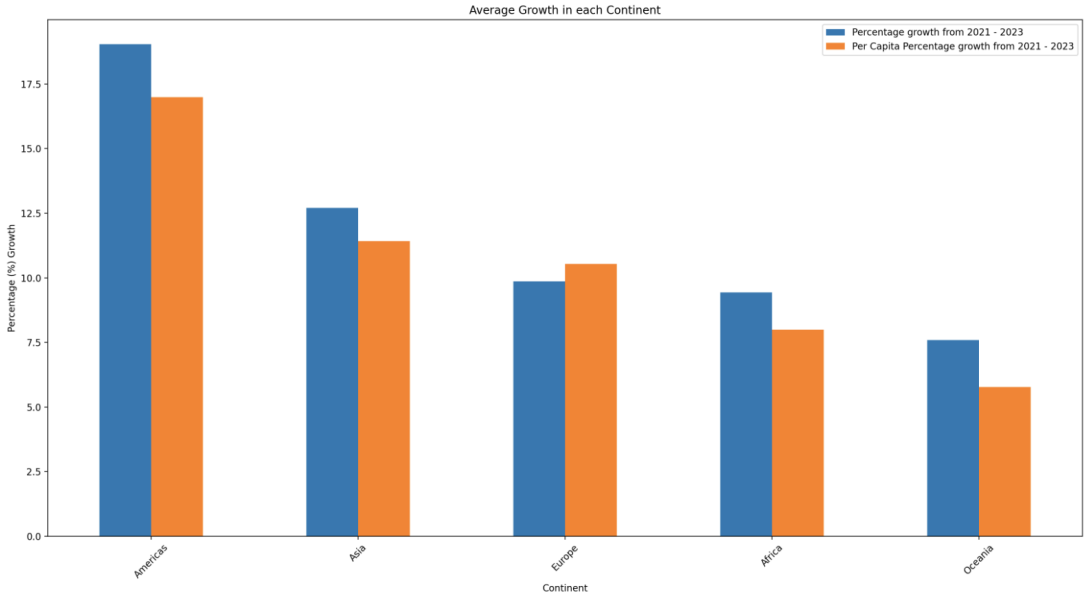


**Figure 1. Barplot showcasing average growth (GDP and GDP per capita) within each continent.**

Figure 2 shows the global distribution of GDP across the five continents. The greatest portion of GDP can be found in Asia (36.3%), followed by the Americas (34.4%), Europe (24.6%), Africa (2.8%), and Oceania (1.9%).

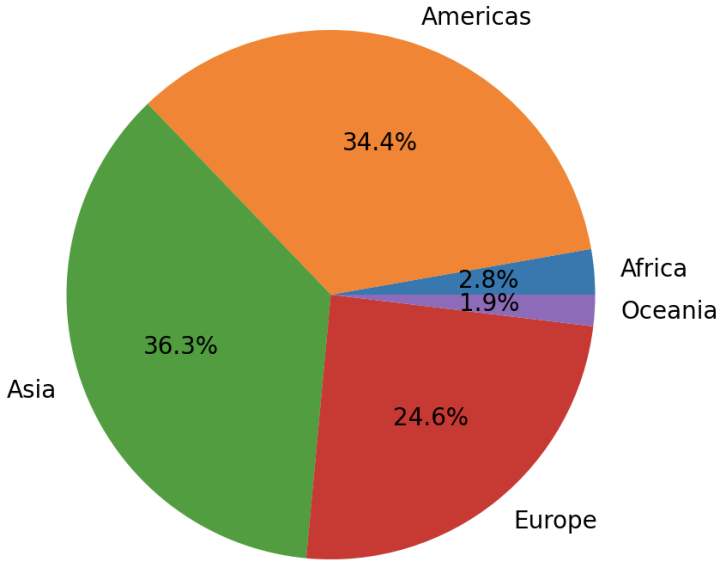


**Figure 2. Pie chart displaying the distribution of global GDP in 2023.**

To visualize the correlation between GDP per capita and TEE, figure 3 shows a scatter plot of the variables, with a best-fit line representing a simple linear regression, predicting GDP per capita using TEE.

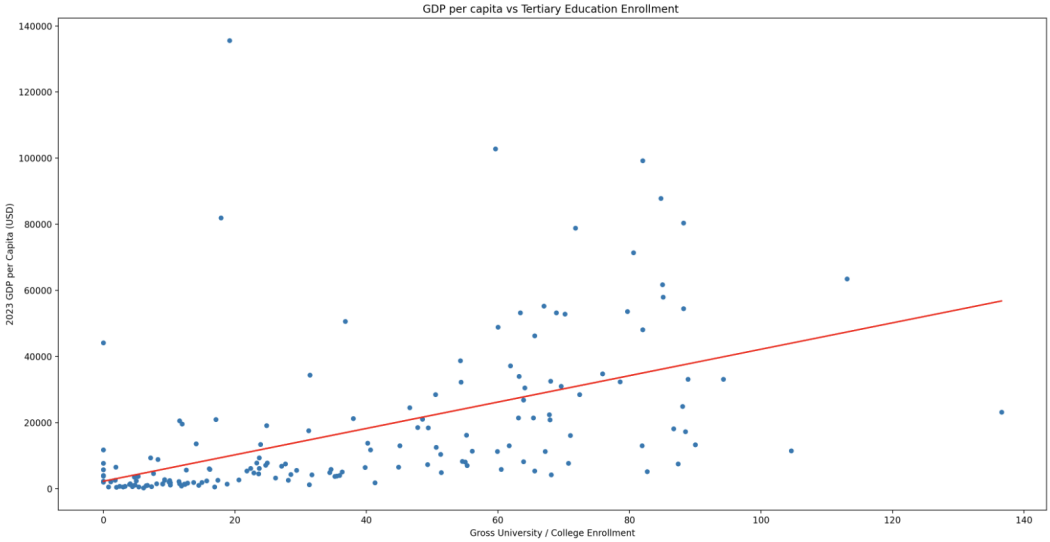


**Figure 3. Scatterplot of GDP per capita vs University enrollment, with regression line.**

Finally, figure 4 displays an animation showing continental GDP over three years. Although the average GDP growth in each continent is positive, this figure shows that the Americas and Europe are the only continents that see significant GDP growth. While Asia, Africa, and Oceania do not appear to experience the same level of growth. This figure shows the difference between growth in actual GDP across continents, versus the average growth within each continent (seen in Figure 1).

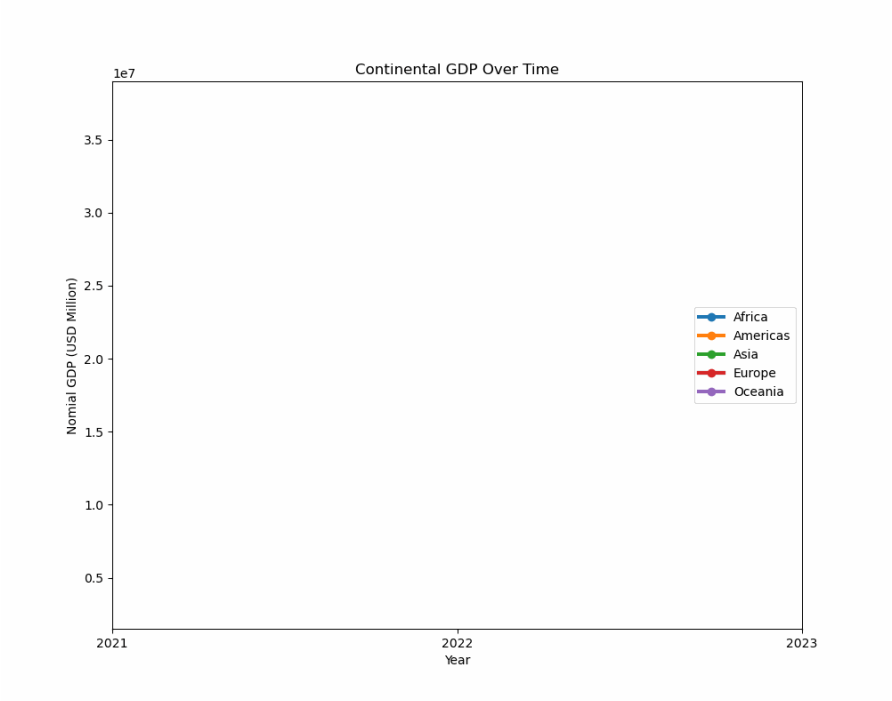


**Figure 4. Animation of continental GDP between 2021 and 2023.**

## 4.0 Findings

These findings can be separated into two categories. The first of which is global growth since 2021. It is evident through this analysis and these figures that the economies of the world have been growing since 2021. This is clear in Figure 1, which shows that, on average, all continents have experienced both GDP and per capita growth. Further supported by the animation in Figure 4, which shows increases in global GDP since 2021. The second category looks for findings regarding the correlation between educational data and economic growth. Regression analyses show that birth rate, PEE, TEE, and YLR are statistically significant predictors of GDP per capita. However, these predictors are not statistically significant estimators of GDP, GDP growth, or GDP per capita growth. The regression function developed through this analysis could be used for the prediction of a nation's GDP per capita given the data of the predictors were to be provided.

### 4.1 Future Work

Given more time, this project would benefit from two main improvements. The first improvement would be an expansion of the data collection process. Collecting data on more countries would allow for a broader scope. Moreover, collecting more educational data, among other data, for a wider variety of predictors, could lead to more precise regression models, which could be used for the prediction of economic markers (GDP, GDP per capita, etc.). The collection of data preceding 2021 could also prove beneficial. This would lead to a better understanding of trends involving GDP and per capita growth. The second improvement would be to develop more complex regression models through the inclusion of second-order and/or interaction terms. This could potentially lead to more precise and accurate models, aiding in the prediction of economic markers.

## 5.0 Appendix

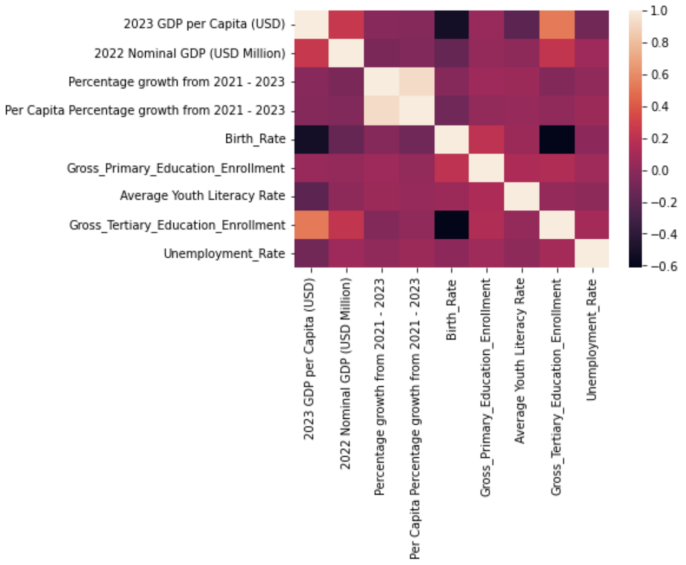


**Figure 5. Heatmap of the Correlation Matrix**