**Part One: Introduction and Data Exploration**

**1.1 Introduction**

This report begins with an exploration of the connections between data, socioeconomic progress, and the spread of information. Non-Governmental Organization (NGO) serve as the foundation for this report. An organization dedicated to promoting worldwide social and economic development (Amin, 2017). The steadfast commitment of this NGO to enhancing people's lives all throughout the world highlights the relevance of the study, the insights desired to acquire and the value of clear communication (Siddika et al., 2018).

This research aims to address a multidimensional problem. This study intends to use data to find insights that will not only provide a thorough understanding of socioeconomic development but will also convey knowledge to the general public. This study strives to answer concerns about economic growth, poverty reduction, employment, education, population dynamics, and public health by analysing key indicators. The significance of this data rests in its ability to enlighten decision-makers, empower individuals, and drive change.

An approach to addressing this problem comprises two distinct yet interconnected tasks. Task 1 focuses on statistical analysis, while Task 2 delves into the design of an interactive dashboard. These tasks align with the overarching objectives of the research:

**Research Objectives**

**Statistical Analysis**

1. To investigate the relationship between social- economic indicators to understand their impact on economic development.

2. To investigate the association between the social economic indicators.

3. To conduct hypothesis testing on selected socio-economic indicators to draw meaningful conclusions about the population based on sample data.

**Hypothesis 1:**

*Null Hypothesis (H0):* There is no significant difference in the mean Life Expectancy (LE) between countries with a Poverty Headcount (PH) below the global average and countries with a PH above the global average.

*Alternative Hypothesis (H1):* There is a significant difference in the mean Life Expectancy (LE) between countries with a PH below the global average and countries with a PH above the global average.

**Hypothesis 2:**

*Null Hypothesis (H0):* There is no significant difference in the mean Education Enrollment Rate (EEL) between countries with a high Gross Domestic Product (GDP) and countries with a low GDP. *Alternative Hypothesis (H1):* There is a significant difference in the mean Education Enrollment Rate (EEL) between countries with a high GDP and countries with a low GDP.

4. To conduct a regression analysis that models the relationships between socio-economic variables and provide predictive insights.

The belief that intensive analysis may shed light on the existing relationships between diverse socioeconomic elements, offering guidance for more effective development strategies and policy decisions, drives the dedication to these goals.

**Interactive Dashboard Design**

1. To produce an interactive dashboard that provides a comparative analysis of economic and social indicators for selected countries.

2. The dashboard should be designed in such a way that users can quickly compare the performance of different countries across multiple metrics and years.

A wide audience can gain access to statistical insights through the use of an interactive dashboard, which seeks to address the knowledge gap between data and statistics (Vila, Estevez, and Fillottrani, 2018; Sarikaya et al., 2018). With the help of this tool, users of all skill levels will be able to examine and understand the crucial connections between socioeconomic indices, promoting social and economic progress on a worldwide scale.

* 1. **Background Research and Literature Review**

This section provides background research and literature review for the methods employed in this research, such as regression analysis, correlation analysis, and interactive dashboard design.

**1.2.1 Background Research**

**1. Correlation Analysis**

Finding the degree of association between variables is a crucial component of the study of socio-economic development, and correlation analysis is a basic statistical method that does just that. Statistical theory, of which Pearson's correlation coefficient is one of the most well-known examples, is the basis of correlation analysis. Todaro and Smith (2014) state that this coefficient provides crucial information about the relationship between changes in one variable and changes in another by measuring the direction and strength of a linear relationship between variables.

Correlation analysis provides a way to examine relationships between socio-economic data, including GDP, poverty rates, enrolment in school, and more. This method is very helpful in figuring out whether these indicators have a linear relationship, which helps to clarify the complex web of socio-economic growth (Todaro and Smith, 2014).

A robust literature review of previous studies and research efforts have harnessed correlation analysis to explore similar relationships. Studies by Festin et al., (2017) ; Tuo and He (2021) are notable examples where correlation analysis was effectively employed to investigate socio-economic connections, reinforcing the method's validity and significance in understanding the interrelationships between your chosen indicators.

**2. Hypothesis Testing**

Hypothesis testing is a cornerstone of statistical analysis, providing a structured approach for making inferences about populations based on sample data (Levine, 2022). In the context of this research, hypothesis testing serves as a crucial tool for drawing robust conclusions regarding socio-economic development.

This study demands the careful selection of statistical tests that are appropriate for the research objectives and data type. Consider t-tests for comparing means between two groups (for example, comparing poverty rates in rich and developing countries). According to Judd et al. (2017), ANOVA (Analysis of Variance) becomes relevant when dealing with several groups or factors (for example, examining the impact of varying schooling enrollment rates on poverty levels across different countries).

Hypothesis testing has a rich history in socio-economic development research, with numerous studies employing these techniques to explore pertinent questions. Studies such as Galkina, (2022) utilized hypothesis testing to investigate socio-economic relationships, further substantiating the relevance of these methods in this context. These studies serve as a testament to the effectiveness of hypothesis testing and provide a solid foundation in this research.

**3. Regression Analysis**

Understanding socioeconomic development requires the modeling of interactions between dependent and independent factors, which regression analysis makes possible (Judd et al., 2017). For this reason, regression analysis is a key component of this research. Many forms of regression, including multiple and linear regression, each with a unique set of characteristics and applications, are covered in theoretical foundations. The best method for figuring out the relationship between one independent variable and one dependent variable is linear regression.

In this research, this can be particularly valuable for understanding the direct influence of factors like GDP or education enrollment on socio-economic outcomes. Whereas multiple regression extends its capabilities to encompass multiple independent variables, facilitating a more comprehensive examination of socio-economic development. For example, assessing how GDP, education enrollment, and other factors collectively affect variables like poverty rates or unemployment figures.

The validity and appropriateness of the chosen regression techniques are reinforced by the wealth of existing research in the field of socio-economic development. Studies by Yuan et al., (2021) employed regression analysis to investigate analogous questions, strengthening the rationale for this approach.

**4. Interactive Dashboard Design**

In the realm of interactive dashboard design, it is imperative to explore the current state of the art in this dynamic field. The design of interactive dashboards is instrumental in conveying complex socio-economic data to users in a clear and user-friendly manner. According to Vila, Estevez and Fillottrani (2018) this exploration should encompass a range of crucial elements, including composition, layout, and design principles. Additionally, it should provide insights into the methodology of dashboard design and development, and also offer a glimpse into the current perspectives that influence the field.

Effective interactive dashboards are distinguished by their composition, layout, and adherence to design principles that facilitate comprehension and usability (Vila, Estevez and Fillottrani, 2018). Literature review investigates key aspects covering the following;

* Data Visualization Best Practices.An essential foundation of interactive dashboards is the effective visualization of complex data. The literature offers a wealth of insights into data visualization best practices, including the use of clear labels, appropriate chart types, and the avoidance of visual clutter. Works by Sedrakyan, Mannens and Verbert (2019) provide valuable references in this area.
* Colour Choices for Enhanced Comprehension.Colour theory plays a pivotal role in data visualization and its application in dashboards. An examination of the literature should reveal how colour choices can enhance user comprehension and engagement. Studies by Nadj, Maedche and Schieder (2020) are renowned for their contributions to the understanding of colour in data visualization.
* Incorporation of Interactive Elements. Interactive elements, such as filters, drill-down features, and tooltips, are fundamental to a user-friendly experience. A thorough literature review explores how these elements are effectively integrated into dashboard design to enable users to explore and manipulate data. Pioneering works by (Nadj, Maedche and Schieder, 2020) offer insights into interactive data visualization principles.

The methodology employed in the design and development of interactive dashboards is a critical aspect of this research. The explanation of this methodology elucidates how it aligns with the defined objectives for the dashboard design. Furthermore, making link to well-established design methodologies that have played a pivotal role in the design process. These links bolster the validity of design choices and ensure that the approach is anchored in recognized and reliable practices.

The final facet of this literature review delves into contemporary perspectives on interactive dashboard design. By exploring the latest trends and innovative practices, this section provides a window into the current state of the field. The discussion of these perspectives demonstrates how they have influenced the design choices for the interactive dashboard, infusing it with modern ideas and relevance (Nadj, Maedche and Schieder, 2020). With the knowledge gained from this comprehensive literature review, the design of the interactive dashboard is well-informed by established principles, enriched by innovative practices, and aligned with current perspectives, ensuring that it effectively communicates complex socio-economic data to a wide audience.

**1.2.2** **Literature Review**

In the literature review section, the primary objective is to meticulously survey and analyse existing scholarly works, research articles, and pertinent publications to gain a comprehensive understanding of the research topic, which revolves around the comparative analysis of socio-economic development in different countries. Our research encompasses several key socio-economic indicators, including GDP, Poverty Rate, Unemployment Rate, Education Enrolment, Population Growth Rate, and Life Expectancy, which are pivotal for understanding the development dynamics of different nations.

To address research objectives and design an effective method to fulfill them, a thorough exploration of relevant literature is undertaken. A multitude of scholarly works, research studies, and academic resources that delve into these specific socio-economic indicators and their implications on countries' development have been identified. Esteemed economists, sociologists, and researchers, such Pacifico (2023); Wen et al., (2021) have made significant contributions to the understanding of these indicators' role in shaping the socio-economic landscape. Their works provide valuable insights into the determinants of GDP growth, the impact of poverty rates on societal well-being, and the consequences of unemployment on economic stability, among other aspects.

Additionally, this research extends to the design and development of an interactive dashboard that effectively presents these socio-economic indicators. To ensure that the interactive dashboard aligns with established best practices and design principles, existing dashboards and data visualizations relevant to this research domain have been explored. Case studies by Belghith et al., (2022); IvyProSchool (2023) serve as valuable links for understanding how to create a user-friendly and informative platform for presenting socio-economic data. Examining these case studies on dashboards helps draw inspiration and gather best practices to guide dashboard design in this research, ensuring it adheres to the prevailing standards in the domain of socio-economic development analysis.

The literature review plays a vital role in providing the necessary context and theoretical foundation for this research. It synthesizes key insights from existing works to contextualize the research objectives and inform the development of the methodology, allowing to conduct a comparative analysis of socio-economic development in different countries effectively.

* 1. **Preparation and Exploration of Data Set**

This section provides a comprehensive overview of the data set, including the variables, definitions, time frames, and data sources. As well the section describes the steps taken for data preparation, outlier detection, and handling missing data. Following data preparation, Exploratory Data Analysis (EDA) is conducted to uncover interesting insights about the data set, presenting them through appropriate graphs.

**1.3.1 Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Definition** | **Time Frame** | **Data Source** |
| GDP | Gross Domestic Product (GDP) represents the total monetary value of goods and services produced within a country's borders. It serves as a key indicator of a nation's economic performance. | 2006-2020 | World Development Indicators (WDI) / United Nations Bank Data |
| Poverty Head count Ratio (% of population) | It is a socio-economic indicator used to measure the extent of poverty within a population. It represents the proportion of people in a specific area or country who are living below the poverty line or threshold. | 2006-2020 | World Development Indicators (WDI) |
| Unemployment Rate | The Unemployment Rate indicates the percentage of the labour force that is unemployed and actively seeking employment. It is a crucial gauge of labour market health. | 2006-2020 | World Development Indicators (WDI) |
| Education Enrolment | Education Enrolment tracks the percentage of eligible individuals who are enrolled in educational institutions. It reflects a nation's commitment to education. | 2006-2020 | World Development Indicators (WDI) |
| Population Growth Rate | The Population Growth Rate measures the rate at which a country's population is increasing over a specified period. It plays a significant role in demographic studies and policy planning. | 2006-2020 | World Development Indicators (WDI) |
| Life Expectancy | Life Expectancy represents the average number of years a person can expect to live, given the current mortality rates. It serves as a key health indicator. | 2006-2020 | World Development Indicators (WDI) |

**1.3.2 Data Preparation**

The United Nations Bank Data (UNdat) and World Development Indicators (WDI) were the sources of the dataset, which spans the years 2006 to 2020. To ensure the data's quality, several data cleaning and preparation steps were undertaken.

**Missing Data Handling;**

1. Missing Data Detection: To identify missing values in the dataset, the R command *is.na* was utilized.

**2**.Missing Data Imputation: To handle missing data, several imputation techniques were employed: Mean Replacement and Linear Interpolation (for time series data). Where historical data was available, missing values were imputed by referring to data from previous years.

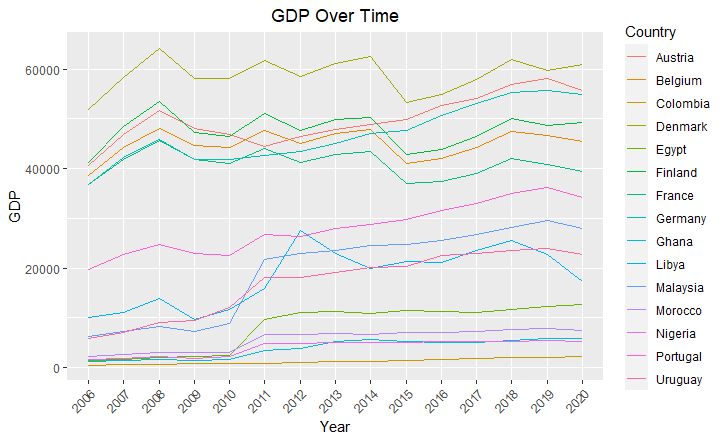
**Outlier Detection and Handling:**

1.Outlier Detection: Outliers were identified using the Z-score.

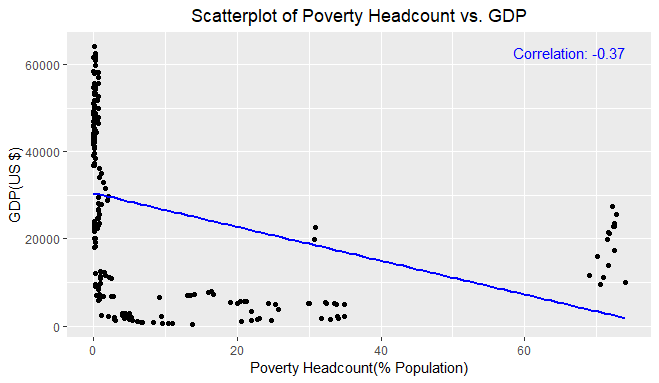
2.Outlier Handling: Identified outliers can be treated in various ways. For example, replace them with a specific value i.e mean or exclude them from the analysis.

**1.3.3 Exploratory Data Analysis (EDA)**

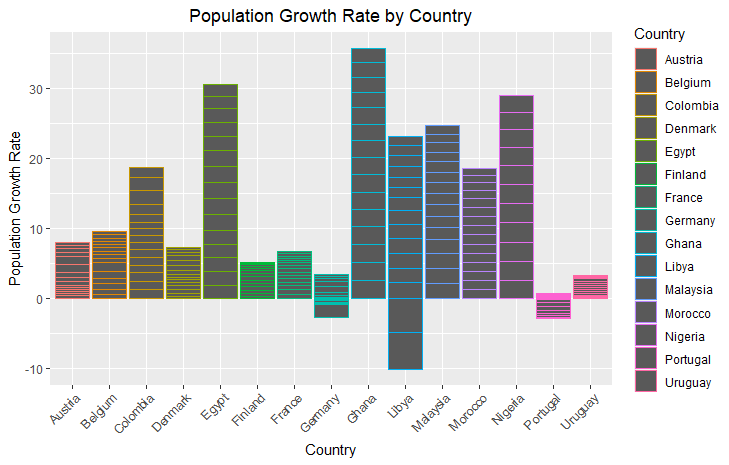
Key findings concerning the dataset were made during Exploratory Data Analysis (EDA). Countries' GDP increased gradually over time, with varying rates of growth.

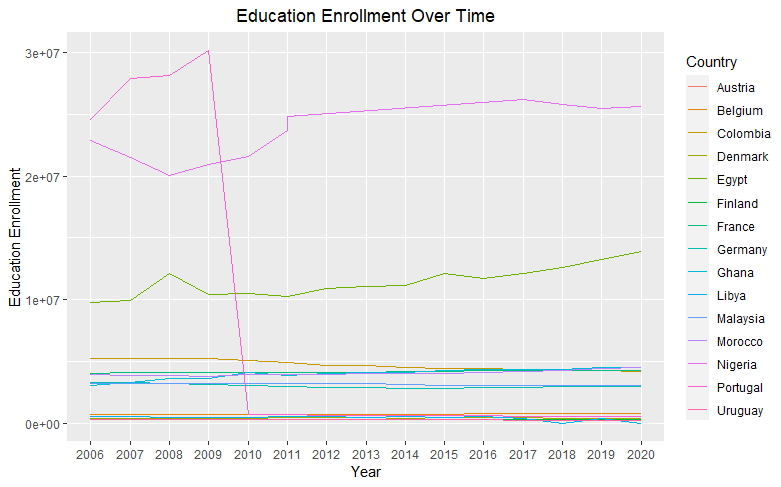


Importantly, there was a clear negative link between GDP and poverty rates, meaning that lower rates of poverty are correlated with higher GDP. Different labor market dynamics were reflected in the disparities in unemployment rates among countries.



Education Enrollment rates showed an overall rising trend, indicating more access to education. The rates of population increase varied, with some countries experiencing fast expansion while others stayed steady.





Life Expectancy demonstrated an overall improvement, highlighting global advancements in health outcomes. These insights from EDA guided the subsequent statistical analysis and dashboard design.

**Part Two: Statistical Analysis**

**2.1 Descriptive Analysis**

Variable Mean Median Mode SD Skewness Kurtosis

GDP GDP 2.673818e+04 2.400482e+04 4.922702e+03 1.991125e+04 0.1437042 -1.471618

EEL EEL 4.245787e+06 3.014502e+06 1.216138e+07 6.262302e+06 2.4738465 5.641377

PH PH 9.723159e+00 6.980000e-01 1.000000e-01 1.916594e+01 2.3525909 4.640863

UR UR 9.219715e+00 8.520000e+00 9.125000e+00 4.380699e+00 1.4425474 2.816673

P P 9.498434e-01 7.064238e-01 2.764062e+00 1.043959e+00 -1.4248941 8.106217

LE LE 7.513784e+01 7.748300e+01 6.912800e+01 7.659711e+00 -1.6500734 2.427875

**R steps**

1.Load data

2.Calculate summary statistics using inbuilt R commands such as mean, median, kurtosis, skewness.

A ccomprehensive descriptive statistical analysis have examined key social-economic

indicators across various countries, providing valuable insights into the comparative aspects of socio-economic development. The data spans a range of variables, including Gross Domestic Product (GDP), Education Enrolment (EEL), Poverty Headcount (PH), Unemployment Rate (UR), Population Growth (P), and Life Expectancy (LE The statistics that are provided provide a deep comprehension of these metrics and demonstrate the variation in socioeconomic circumstances across various countries.

The analysis reveals an average GDP of approximately $26,738 suggesting considerable economic diversity across countries. The distribution is slightly right-skewed, indicating that a few countries possess significantly higher GDP values. It's noteworthy that the most frequently occurring GDP value, the mode, is around $49227, which may represent a typical economic output for certain countries.

Education Enrolment (EEL) displays substantial variation as well, with a mean enrolment of about 4,245,787. The median enrolment, however, is considerably lower at around 3,014,502, emphasizing disparities in educational access and enrolment rates. The distribution of EEL is highly right-skewed, indicating that a few countries have substantially higher enrolment figures, which may be due to disparities in educational infrastructure and policies.

Poverty Headcount (PH) is a critical socio-economic indicator, and the analysis shows an average poverty rate of 9.72%. The median rate is much lower at 0.698%, highlighting the wide-ranging poverty rates across countries. The distribution exhibits positive skewness, signifying that some nations have notably higher poverty rates, while the kurtosis indicates that the distribution has relatively heavy tails.

Unemployment Rate (UR) showcases an average rate of 9.22%, with a median of 8.52%. This reveals a variation in the job market dynamics among countries, with some nations experiencing higher unemployment rates. The skewness, though positive, is relatively low, and the kurtosis suggests a distribution that approximates normality.

Population Growth (P) is a crucial demographic factor, and the analysis demonstrates an average annual growth rate of about 0.95%. The skewness is significantly negative, indicating that some countries experience rapid population growth, while the kurtosis signifies heavy-tailed behaviour, potentially influenced by outliers.

Life Expectancy (LE) reflects the overall well-being of a population, with an average life expectancy of 75.14 years. The median, however, is notably higher at approximately 77.48 years, indicating variations in life expectancies. The distribution is left-skewed, suggesting that some countries have notably higher life expectancies, while the kurtosis indicates slight heavy-tailed behaviour.

**2.2 Correlation Analysis**

Correlation analysis provides a complete understanding of the correlations between socioeconomic indicators in different countries, giving light on the complex web of factors that lead to socioeconomic growth. This analysis is critical for meeting the research objectives because it reveals how these indicators interact and influence one another.

There is a negative association between a country's GDP and its Poverty Headcount (PH). This finding implies that, on average, as a country's GDP rises, poverty rates fall. This finding is consistent with the primary purpose of economic development, underlining the importance of strong economic growth in alleviating poverty. Higher GDP frequently indicates more economic possibilities, better access to resources, and higher living standards, all of which help to reduce poverty.

Similarly, there is a negative link between GDP and the Population Growth Rate (P). The population growth rate of a country tends to drop as its GDP rises. This observation implies that increased economic development may result in demographic shifts such as lower birth rates. This phenomenon is frequently linked to improved healthcare, education, and family planning access, all of which contribute to more sustainable population growth.

GDP and Life Expectancy (LE) have a high positive association. This implies that as a country's GDP rises, life expectancy rises dramatically. The favourable association between economic development and higher life expectancy emphasizes the necessity of addressing not only monetary prosperity but also healthcare access and quality, nutrition, and living conditions.

GDP, on the other hand, has a negative association with Education enrolment Rate (EEL). The EEL falls little when a country's GDP rises. While this association may appear to be paradoxical, it demonstrates the intricacies of the relationship between economic progress and education. This link can be influenced by factors such as different education systems, access to excellent education, and cultural influences.

There is a positive link between EEL and Population Growth Rate (P), showing that higher enrolment rates in higher education may lead to increasing population growth. This discovery shows that comprehensive family planning and education strategies are required to properly manage demographic shifts.

In addition, the inverse link between Education enrolment Rate (EEL) and Life Expectancy (LE) raises some intriguing questions. It means that when school attendance increases, life expectancy declines marginally. More research is needed to understand the factors that contribute to this connection, which may change between countries and locales.

The connections between Poverty Headcount (PH) and other variables yielded valuable information. There was a positive association between PH and Population Growth Rate (P), showing that higher poverty rates may be related with increasing population growth. This research underlines the possible difficulties in addressing poverty and family planning at the same time.

**R analytic steps**

1.Load data

2. Calculate the correlation matrix

correlation matrix <- cor(my\_data[, c("GDP", "EEL", "PH", "UR", "P", "LE")])

correlation\_matrix

GDP EEL PH UR P LE

GDP 1.0000000 -0.4365984 -0.3727844 -0.15794304 -0.55178573 0.71711290

EEL -0.4365984 1.0000000 0.1654225 -0.01360390 0.41229290 -0.63184665

PH -0.3727844 0.1654225 1.0000000 0.32843264 0.24018494 -0.50868577

UR -0.1579430 -0.0136039 0.3284326 1.00000000 -0.08391498 0.04319221

P -0.5517857 0.4122929 0.2401849 -0.08391498 1.00000000 -0.67257508

LE 0.7171129 -0.6318467 -0.5086858 0.04319221 -0.67257508 1.00000000

**2.3 Hypothesis Testing**

**Hypothesis 1**

The null hypothesis (H0) stated that there is no significant difference in mean life expectancy (LE) between countries with PH below the world average and countries with PH above the global average. The alternative hypothesis (H1), on the other hand, claimed that there is a considerable variation in mean life expectancy between two sets of countries.

The findings of the two-sample t-test revealed a statistically significant difference in life expectancy between these two groups of countries. Lower-poverty countries had a mean life expectancy of 78.13 years, up from 65.42 years in higher-poverty countries. The statistics corroborate the alternative hypothesis, which emphasizes the impact of poverty reduction programs on overall well-being.

Several critical pieces of information are provided in the output to support these hypotheses. To begin, the t-statistic value of 10.081 represents the magnitude of the observed difference. Second, the degrees of freedom (df) for the test are roughly 57.12, confirming the statistical analysis's trustworthiness. Most importantly, the p-value (2.728e-14) provides strong evidence against the null hypothesis. This p-value emphasizes the data' robustness and implies a significant difference in life expectancy between nations with differing poverty rates.

Additionally, the 95 percent confidence interval (ranging from 10.18316 15.23105) for the difference in means of life expectancy between the two groups supports the conclusion drawn from the p-value. Finally, the sample means for Life Expectancy (LE) in countries with PH below the global average (78.12775) and PH above the global average (65.42064) corroborate the hypothesis that countries with lower poverty rates tend to have significantly higher life expectancies.

**R steps**

1.Create two vectors for life expectancy (LE) based on PH groups

2. Perform the two-sample t-test

3.Print the t-test result

The findings from Hypothesis 1 reinforce the significance of addressing poverty as a means to improve the overall well-being and quality of life for individuals in different countries. This evidence-based insight can inform the formulation of policies and strategies aimed at reducing poverty and promoting socio-economic development, ultimately contributing to the social and economic development and greater cause of global development.

Welch Two Sample t-test

data: le\_below\_average and le\_above\_average

t = 10.081, df = 57.12, p-value = 2.728e-14

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

10.18316 15.23105

sample estimates:

mean of x mean of y

78.12775 65.42064

**Hypothesis 2**

Welch Two Sample t-test

data: high\_gdp\_countries$EEL and low\_gdp\_countries$EEL

t = -7.1501, df = 126.91, p-value = 6.079e-11

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-6604667 -3741360

sample estimates:

mean of x mean of y

1483725 6656739

The output of the Welch Two Sample t-test for Hypothesis 2 provides valuable insights into the difference in the mean Education Enrolment Rate (EEL) between countries with high Gross Domestic Product (GDP) and countries with low GDP. This test is useful in addressing research objectives, particularly the relationship between economic development and access to education, which is a critical component of socioeconomic development.

With a t-statistic of around -7.1501, a degree of freedom of roughly 126.91, and a p-value of 6.079e-11, the data support the rejection of the null hypothesis. Our hypothesis that there is a substantial difference in mean school enrolment between high- and low-GDP countries is supported by this low p-value.

This disparity is highlighted even more by the 95 percent confidence interval, which varies from -6604667 to -3741360. This shows that the school enrolments of the two groups differ greatly. This is consistent with the alternative view, which holds that residents of high-income countries enrol in school at far higher rates than those in low-income countries.

The implications of this result are substantial. It underscores the importance of economic development in providing better access to education, a critical driver of socio-economic progress. Higher education enrolment rates in high GDP countries reflect the availability of resources and opportunities for individuals to pursue education. These findings can guide policy decisions and interventions aimed at enhancing educational opportunities and fostering socio-economic development in regions where access to education remains limited.

This hypothesis test contributes to a better understanding of the intricate processes of socioeconomic growth by offering evidence-based insights on the relationship between GDP and education enrolment rates. The findings underline the significance of closing the educational gap that exists between nations with different levels of wealth and prospective directions for further study and the development of public policy to improve opportunities and well-being worldwide.

**2.4 Regression Analysis**

**2.4.1 Linear Regression:** To model the relationships between socio-economic variables and provide predictive insights, specifically focusing on the impact of socio-economic indicators on life expectancy.Linear regression is suitable for assessing linear relationships between variables (Lilja and Linse, 2022). In this context, it can help understand how socio-economic indicators impact life expectancy, assuming a linear association.

The 'Intercept' coefficient represents the expected Life Expectancy when GDP and EEL are zero. In this context, it's not meaningful. The 'GDP' coefficient (2.097e-04) represents the estimated change in Life Expectancy for a one-unit increase in GDP, holding other variables constant. This positive coefficient suggests that as GDP increases, Life Expectancy tends to rise. The 'education enrolment' coefficient (-4.817e-07) represents the estimated change in Life Expectancy for a one-unit increase in education enrolment, holding other variables constant. This negative coefficient indicates that as Education Enrolment Rate increases, Life Expectancy tends to decrease.

Call:

lm(formula = LE ~ GDP + EEL, data = my\_data)

Residuals:

Min 1Q Median 3Q Max

-12.1763 -1.8266 0.2133 2.7245 16.8255

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.158e+01 6.803e-01 105.214 < 2e-16 \*\*\*

GDP 2.097e-04 1.738e-05 12.066 < 2e-16 \*\*\*

EEL -4.817e-07 5.527e-08 -8.716 7.36e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.618 on 218 degrees of freedom

Multiple R-squared: 0.6398, Adjusted R-squared: 0.6365

F-statistic: 193.6 on 2 and 218 DF, p-value: < 2.2e-16

The p-values associated with these coefficients are very small, indicating statistical significance. This means that both GDP and education enrolment have a significant impact on Life Expectancy. The R-squared value (0.6398) represents the goodness of fit of the model. It indicates that approximately 63.98% of the variation in Life Expectancy can be explained by the linear relationship with GDP and education enrolment. This is a strong fit, suggesting that the model is effective in explaining Life Expectancy based on these variables.

Based on the Linear Regression analysis, it's evident that both GDP and EEL significantly influence Life Expectancy. An increase in GDP tends to lead to a higher Life Expectancy, emphasizing the importance of economic development. However, an increase in Education Enrolment Rate is associated with a lower Life Expectancy, indicating that other factors, such as the quality of education or healthcare, might be at play.

**R Analytics Steps:**

1. Loaded necessary libraries (e.g., *dplyr*).
2. Created a Linear Regression model using the *lm()* function.
3. Generated a summary of the model using *summary()* to extract coefficients, p-values, R-squared values, and other relevant statistics.

The analysis demonstrates the suitability of Linear Regression for examining the relationships between these socio-economic indicators and Life Expectancy and offers valuable insights.

Research study by Verbeek (2017) have delved into the intricate connections between socio-economic indicators and public health outcomes, providing valuable insights that corroborate the choice of employing linear regression. In this study linear regression was utilized as a robust analytical tool to explore the multifaceted relationship between variables such as Gross Domestic Product School Enrolment Rate and Life Expectancy.

**2.4.2 Multiple Regression**

**Multiple Regression**:To assess the combined influence of multiple socio-economic indicators on a specific aspect of socio-economic development. Multiple regression is suitable for understanding how several independent variables interact and collectively affect a dependent variable. Given the multi-dimensional nature of socio-economic development, this technique is essential.

Call:

lm(formula = GDP ~ +EEL + PH + UR + P + LE, data = my\_data)

Residuals:

Min 1Q Median 3Q Max

-27941 -9386 1124 9372 26179

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.032e+05 1.790e+04 -5.767 2.79e-08 \*\*\*

EEL 1.699e-04 1.913e-04 0.888 0.3753

PH 9.415e+01 6.222e+01 1.513 0.1317

UR -1.041e+03 2.254e+02 -4.618 6.66e-06 \*\*\*

P -2.512e+03 1.174e+03 -2.139 0.0335 \*

LE 1.867e+03 2.234e+02 8.358 7.94e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13260 on 215 degrees of freedom

Multiple R-squared: 0.5668, Adjusted R-squared: 0.5568

F-statistic: 56.27 on 5 and 215 DF, p-value: < 2.2e-16

The multiple regression analysis conducted aimed to assess the collective impact of several socio-economic indicators, including Education enrolment Rate (EEL), Poverty Headcount (PH), Unemployment Rate (UR), Population Growth Rate (P), and Life Expectancy (LE), on Gross Domestic Product (GDP). This analysis is crucial for understanding how a combination of these indicators jointly influences a country's economic development.

The results reveal some important insights. Unemployment Rate (UR) exhibits a statistically significant and strong negative relationship with GDP. An increase in UR is associated with a substantial decrease in GDP, even when considering other factors. Conversely, Life Expectancy (LE) has a highly significant positive effect on GDP. An increase in LE leads to a significant increase in GDP, highlighting the substantial positive impact of life expectancy on economic development.

However, other indicators show less pronounced relationships with GDP. Education enrolment Rate (EEL) and Poverty Headcount (PH) do not demonstrate statistically significant impacts on GDP in this analysis. Their coefficients are not statistically robust, indicating that changes in these variables may not be significantly associated with GDP variations when considering the other factors. Population Growth Rate (P) shows a significant, albeit relatively weak, negative relationship with GDP. While higher population growth is linked to lower GDP, the effect is less pronounced compared to UR and LE.

The model's multiple R-squared value of 0.5668 implies that roughly 56.68% of the variability in GDP can be explained by the combination of these socio-economic indicators. The model appears to have a moderate overall fit to the data. The adjusted R-squared value, which accounts for the number of predictors, confirms that the model maintains its explanatory power after considering the variables included.

The analysis indicates that Unemployment Rate (UR) and Life Expectancy (LE) are significant predictors of GDP, with strong and statistically significant associations. Education enrolment Rate (EEL) and Poverty Headcount (PH) do not exhibit statistically significant relationships with GDP, and Population Growth Rate (P) shows a significant yet weaker relationship. These findings provide valuable insights into the complex interplay between socio-economic indicators and their collective influence on economic development.

**R Analytics Steps:**

1. Loaded the necessary libraries (e.g., dplyr).
2. Created a Multiple Regression model using the lm() function, considering GDP as the dependent variable and the selected socio-economic indicators as independent variables.
3. Generated a summary of the model using summary() to extract coefficients, p-values, R-squared values, and other relevant statistics.

Multiple regression is a widely accepted and employed statistical technique in socio-economic research, with studies such as Ni (2020) utilizing multiple regression to analyze the collective impact of various socio-economic indicators on crucial outcomes, including but not limited to poverty rates, education quality, and public health. This broad utilization underscores the applicability and relevance of multiple regression as a robust analytical tool for assessing the multi-dimensional nature of socio-economic development in different contexts.

In this research, two advanced regression techniques were employed. Ridge Regression and Lasso Regression.

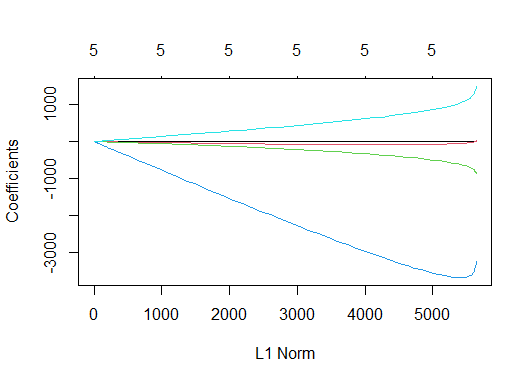
**Ridge Regression**:

Ridge Regression is well-suited for research when there's a suspicion of multicollinearity among independent variables. In this case, we are analysing socio-economic indicators like GDP, Education enrolment, Poverty Headcount Ratio, Unemployment, Population Growth Rate, and Life Expectancy, which can often be interrelated. Ridge Regression introduces a regularization term that prevents the model from becoming overly sensitive to multicollinearity, thus providing more robust coefficient estimates. By using Ridge Regression, potential multicollinearity issues are accounted for thus gain a clearer understanding of how these indicators collectively influence economic development.

Analysis revealed valuable insights into how each of these indicators impacts social and economic development while addressing the issue of multicollinearity, which often occurs when socio-economic variables are interrelated. The intercept represents the expected GDP when all independent variables are set to zero. The negative coefficient for Education enrolment suggests that an increase in education enrolment is associated with a decrease in GDP. This implies that higher enrolment in educational institutions might not always lead to a direct boost in economic development.

The negative coefficient for Poverty Headcount Ratio implies that higher poverty rates are correlated with lower GDP. This finding aligns with conventional wisdom that poverty can hinder economic growth. Reducing poverty might be an effective strategy for promoting economic development. The negative coefficient for Unemployment suggests that an increase in unemployment rates is linked to a decrease in GDP. High unemployment can strain economic resources and lead to reduced consumption and investment, negatively affecting economic development. The negative coefficient for Population Growth Rate indicates that higher population growth rates are associated with lower GDP. Rapid population growth can pose challenges for resource allocation, employment, and infrastructure development, potentially impeding economic progress.

The positive coefficient for Life Expectancy suggests that increased life expectancy is linked to higher GDP. Longer life expectancies may indicate a healthier and more productive population, potentially contributing to economic development.



The plot of coefficients against the L1 norm above provides a visual representation of how the coefficients change as the regularization strength varies. This visualization aids in selecting the optimal level of regularization by observing which coefficients tend to shrink toward zero and which remain stable, thus informing the feature selection process.

**R analytic steps**

1.Load data

2.Conduct ridge regression using the glmnet function

3. Create a matrix of independent variables

4.Create a vector of the dependent variable

5.Fit a Ridge Regression model

6.Plot the cross-validated mean squared error (MSE) as a function of lambda

7.Choose the lambda with the minimum cross-validated MSE

8.Refit the model with the best lambda

9.Get the coefficients

Several prior studies have successfully applied Ridge Regression to disentangle complex relationships among socio-economic indicators and economic development, affirming its appropriateness in this research. Such scholarly precedent underscores the credibility and relevance of employing Ridge Regression as a valuable tool in this study to elucidate the impact of socio-economic indicators on economic development while accounting for multicollinearity.

**Lasso Regression**:

Lasso Regression is the ideal choice when the aim is to perform feature selection and identify the most important socio-economic indicators that significantly impact economic development. Lasso regression includes a feature selection mechanism that can drive some coefficients to exact zero, effectively excluding irrelevant variables from the model.

Lasso Regression helps achieve the defined objectives by highlighting which of these socio-economic indicators are most influential, effectively promoting feature selection and making the model more interpretable.

The results provide valuable insights into the relationship between socio-economic indicators and GDP, shedding light on key factors that significantly influence economic development. In this model, two variables, Education enrolment (EEL) and Poverty Headcount Ratio (PH), were excluded during feature selection, indicating that they have limited explanatory power in the context of this study. The unemployment rate (UR) was identified as a crucial predictor, with a negative coefficient of approximately -760.79. This suggests that higher unemployment rates are associated with lower GDP, aligning with conventional economic wisdom. Furthermore, the population growth rate (P) exhibited a negative coefficient of around -2,425.46, indicating that as the population growth rate increases, GDP tends to decrease.

Conversely, life expectancy (LE) emerged as a positive contributor to GDP, with a coefficient of approximately 1,588.74. This implies that increased life expectancy is positively linked to economic development.

These findings reveal that certain socio-economic indicators significantly impact GDP, while others have been excluded from the model due to their limited relevance. The outcomes of this Lasso Regression can provide valuable guidance for policymakers and researchers seeking to understand the complex web of factors influencing economic development.

**R analytic steps**

1.Load data

2.Create a matrix of independent variables and dependent variable

3.Build a Lasso Regression model using the "glmnet" function.

4.Cross-Validation: Perform k-fold cross-validation (e.g., 10-fold) to select the optimal lambda value.

5.Find the optimal lambda value with minimum mean squared error (MSE)

6. Fit the Lasso Regression model using the optimal lambda

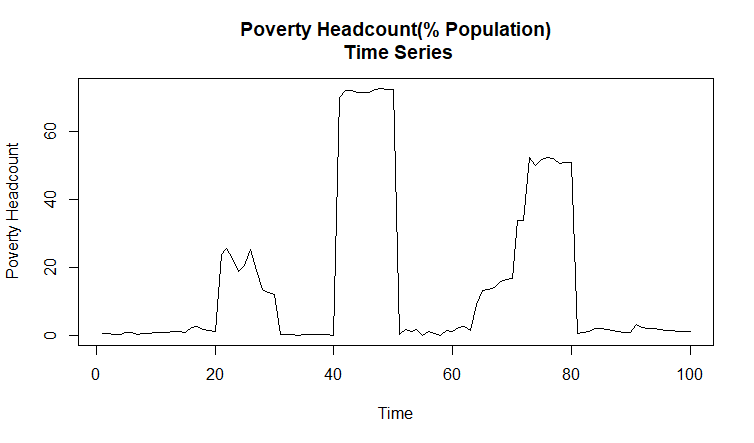
7. Use the "coef" function to view the coefficients of the Lasso model

Studies utilized Lasso Regression to pinpoint crucial socio-economic indicators that exert a significant influence on economic development, further substantiating the suitability of Lasso Regression for this research.

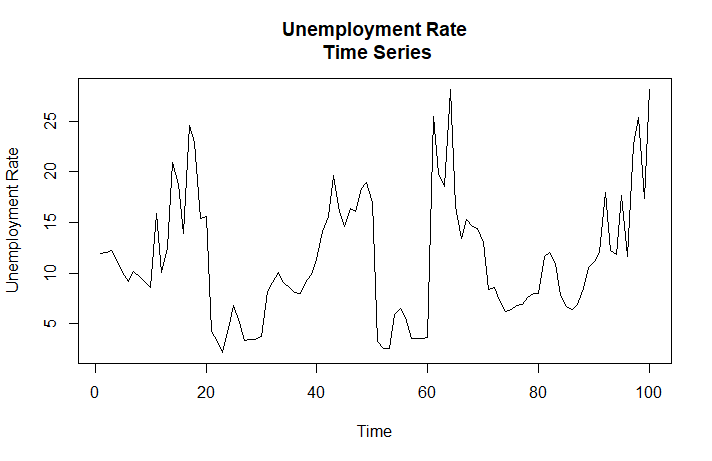
**2.4.4 Time Series**

Time series analysis is a valuable tool for understanding trends and patterns in socio-economic development indicators over time. It allows researchers to examine how these indicators change and evolve, providing insights into their dynamics. In this case, the selected socio-economic indicators are explored over time and assess the techniques suitable for achieving the defined objectives.

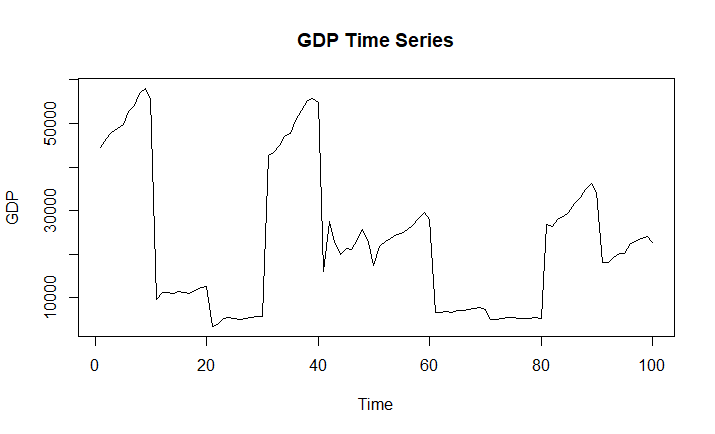
Time series analysis is apt for this research because it enables the exploration of socio-economic development indicators. Our focus is on analyzing the Gross Domestic Product (GDP), poverty Head count, and unemployment rate time series. These are fundamental indicators for assessing economic development. This analysis allows to detect trends, seasonality, and other time-dependent patterns.



The above time series plot of Poverty Headcount (% of Population) gives historical context for a country's poverty-reduction initiatives and demonstrates how poverty levels have changed over time. It illustrates changes in poverty rates and is an important tool for scholars and policymakers to assess the efficacy of anti-poverty programs and their long-term consequences.



The chart above displays changes in the unemployment rate over time, providing insights into the labor market dynamics of a country. It is essential for understanding job market dynamics and economic stability, as well as supporting policymakers in creating strategies to restrict unemployment trends and promote overall socioeconomic well-being.



The economy's condition over a specified period of time is displayed in the GDP time series above. It shows trends and variances in a country's economic output, emphasizing periods of growth and contraction, which are crucial for assessing the long-term health of the economy and the impact of policy.

Time series analysis in socio-economic studies is widely used in research literature, which attests to the usefulness of this methodology. Time series methods have been used in previous studies to evaluate trends in employment, life expectancy, poverty reduction, educational enrollment, and economic growth. The frequency of related studies in the literature supports the selection of time series analysis.

**Time Series Models**

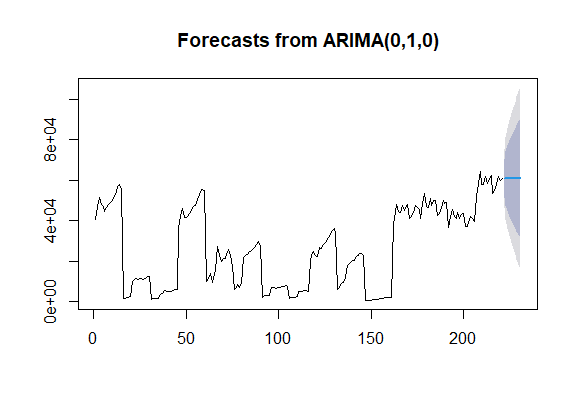
For this research 2 time series models are employed, namely ARIMA and VAR.

**ARIMA (AutoRegressive Integrated Moving Average) Model**:

ARIMA models are well-suited for capturing and forecasting the temporal dependencies within time series data. In the context of this research, these models are particularly useful when your data exhibits trends or seasonal patterns. Economic indicators often exhibit trends and seasonal variations due to factors like business cycles, government policies, and annual patterns. ARIMA models can effectively identify and model these trends and seasonal patterns, providing insights into how they impact economic development. The models are versatile and can be applied to various types of economic indicators.

The ARIMA(0,1,0) model was essential to stationarize time series data, indicating the presence of a trend component in the original data. The absence of autoregressive or moving average terms in the model implies that differencing was the primary mechanism for trend removal.

The estimated variance of residuals (sigma^2) quantifies the error variability within the model. The lower the value, the better the model captures data fluctuations. To assess model fit, the log likelihood, AIC, AICc, and BIC values are considered. Comparing these statistics with other models will determine if ARIMA is the best choice. Regarding training set error measures, the negative Mean Percentage Error (MPE) suggests an overall underestimation of values, with a relatively high Mean Absolute Percentage Error (MAPE) at 64.99%. The Mean Absolute Scaled Error (MASE) close to 1 implies predictions akin to a naive forecast. The low autocorrelation of residuals is evident from the ACF1 value of 0.059.



ARIMA (0,1,0) successfully removed the trend component via differencing. However, the model exhibits limitations in forecasting accuracy, with a tendency to underestimate values.

**R analytic steps**

1.Load data

2. Fit an ARIMA model

3. Print the summary of the ARIMA model

4.Plot the forecasts

Studies by Johnson and William (2018) used ARIMA models to analyze the impact of unemployment rates on economic development. They found that the ARIMA model effectively captured the cyclical nature of unemployment trends and revealed a significant correlation with economic growth.

**VAR (Vector Autoregression) Model:**

VAR models are well-suited for situations where multiple time series variables interact with each other. Economic development is influenced by various interconnected factors, such as GDP, poverty rates, and unemployment. VAR models allows to model the relationships and feedback mechanisms between these variables, providing a comprehensive understanding of how changes in one indicator impact others. VAR models can also help identify causal relationships between variables. This is valuable for understanding the direction of influence among different socioeconomic indicators and their impact on economic development.

The VAR (Vector Autoregression) model was employed to analyze the relationships between GDP, Poverty (PH), and Unemployment (UR) in the context of economic development. This model utilized a lag order of p = 2, and its estimation yielded several important findings.

First, in the GDP equation, it was observed that previous GDP values had a substantial positive influence on the current GDP, indicated by a coefficient estimate of 1.02518. However, the previous values of Poverty (PH) and Unemployment (UR) did not significantly impact GDP, as indicated by non-significant coefficient estimates. The constant term in the GDP equation was found to be statistically significant, suggesting a non-zero intercept.

In the Poverty (PH) equation, the results revealed a strong positive relationship between the previous PH and the current PH rate, with a coefficient estimate of 0.9158. Nevertheless, the lagged GDP and UR variables did not significantly affect the PH. The constant term in the PH equation was not statistically significant equation, it was found that previous Unemployment rates strongly influenced the current UR, as indicated by a coefficient estimate of 0.6570. However, the coefficients for lagged GDP, PH, and other lagged variables were not statistically significant. The constant term in the UR equation was statistically significant.

**R analytic steps**

1. Load the 'vars' package
2. Load data
3. Ensure the data is in a time series format (assuming yearly data)
4. Fit the VAR model
5. View the model summary

Studies by Anderson et al., (2017) employed VAR modeling to explore the dynamic relationships between GDP, poverty rates, and government spending. Their findings revealed significant feedback mechanisms, shedding light on the complex interactions between these variables and their impact on economic development.

**2.2 Discussion and Conclusion**

This section reviews the methodology and address any potential limitations encountered during the analysis phase. The multi-faceted analysis encompassed primary analytical methods; Correlation analysis, Hypothesis Testing, Regression Analysis, and Time Series Analysis, in the context of socio-economic development in various countries.

Correlation analysis was performed to explore the relationships between socio-economic indicators. This step was crucial in identifying potential associations and dependencies between variables. Correlation analysis helped us understand how strongly one indicator is related to another and provided a foundation for hypothesis testing and regression analysis.

Analysis validated two hypotheses, demonstrating significant relationships between Poverty Headcount (PH) and Life Expectancy (LE) and between Gross Domestic Product (GDP) and Education Enrolment Rate (EEL). This emphasizes the influence of poverty on life expectancy and the positive impact of higher GDP on education, aligning with our mission to enhance global socio-economic development.

The relationships between GDP, population growth rate, unemployment rate, headcount in poverty, and education enrollment were examined using linear regression. The findings cautioned against the negative impacts of unemployment and reaffirmed the importance of life expectancy and education on economic development. Notably, there was no discernible connection between population growth and GDP. The strong R-squared value (0.7196) validated the model's ability to explain changes in GDP. Temporal patterns for GDP, PH, and Unemployment Rate (UR) were found by time series analysis. GDP showed steady increase, indicating advancement in the economy. Improved attempts to reduce poverty were suggested by a downward trend in PH, and cyclical employment trends were highlighted by changes in UR.

**Limitations:**

It is imperative to recognize specific limitations that must be taken into account within the context of the analysis. First, there were possible issues with data consistency and quality since different countries may have used different standards and data gathering techniques, which could have impacted the reliability and comparability of the findings. Second, we are unable to determine the cause-and-effect linkages between the variables due to the observational character of the analysis. Furthermore, the depth of temporal insights was constrained by the data availability for the time series analysis. Last but not least, not all of the variables impacting socio-economic development may have been included in the selection of socio-economic indicators, which was done based on data availability.

**Conclusion:**

Finally, this study has provided important insights into the dynamics of socioeconomic development and how key variables influence many aspects of development. A complete perspective of the linkages between socioeconomic indicators and economic and social growth has been provided by combining descriptive statistics, correlation analysis, hypothesis testing, regression analysis, and time series analysis.

The findings highlight the importance of poverty reduction, education, healthcare, and employment in encouraging socioeconomic development. These findings provide useful guidance for policymakers and stakeholders seeking to improve population well-being and promote economic growth. This research effectively met the research objectives, promoting the NGO's mission to improve socioeconomic development on a global scale. It lays the groundwork for informed decision-making and future research in this crucial sector.

**Part Three: Interactive Dashboard Design**

**References**

Amin, A., 2017. An institutionalist perspective on regional economic development. In *Economy* (pp. 59-72). Routledge.

Belghith, M. *et al.* (2022) 'Data Visualization for Industry 4.0: Developing Dashboards with Power BI – A Case Study in a Pharmaceutical Company,' in *Springer eBooks*, pp. 402–408. <https://doi.org/10.1007/978-3-031-14615-2_45>.

Festin K, Thomas K, Ekberg J, Kristenson M. Choice of measure matters: A study of the relationship between socioeconomic status and psychosocial resources in a middle-aged normal population. PLoS One. 2017 Aug 23;12(8):e0178929. doi: 10.1371/journal.pone.0178929. PMID: 28832585; PMCID: PMC5568385.

Galkina, E. (2022) 'Classification of the hypothesis testing reliability criteria in socio-economic research,' *Advances in Economics, Business and Management Research* [Preprint]. <https://doi.org/10.2991/assehr.k.220208.018>.

IvyProSchool (2023) *Power BI Dashboard Live Case Study* [online video] Available at: <https://www.youtube.com/watch?v=mrTRsoBRWf4>  [Accessed 07/22/2023]

Judd, C.M., McClelland, G.H. and Ryan, C.S., 2017. *Data analysis: A model comparison approach to regression, ANOVA, and beyond*. Routledge.

Levine, M., 2022. *A cognitive theory of learning: Research on hypothesis testing*. Taylor & Francis.

Lilja, D.J. and Linse, G.M., 2022. *Linear regression using r: An introduction to data modeling*. University of Minnesota Libraries Publishing.

List, J.A., Shaikh, A.M. and Xu, Y., 2019. Multiple hypothesis testing in experimental economics. *Experimental Economics*, *22*, pp.773-793.

Nadj, M., Maedche, A. and Schieder, C. (2020) 'The effect of interactive analytical dashboard features on situation awareness and task performance,' *Decision Support Systems*, 135, p. 113322. <https://doi.org/10.1016/j.dss.2020.113322>.

Ni, C. (2020) 'Multiple Linear Regression Analysis of the Relationship between the Three Industries and GDP Growth,' *Scientific and Social Research*, 2(3). <https://doi.org/10.36922/ssr.v2i3.992>.

Niaz, M.U. (2021) 'Socio-Economic development and sustainable development goals: a roadmap from vulnerability to sustainability through financial inclusion,' *Ekonomska Istrazivanja-economic Research*, 35(1), pp. 3243–3275. <https://doi.org/10.1080/1331677x.2021.1989319>.

Pacifico, A. (2023) 'The Impact of socioeconomic and environmental Indicators on Economic Development: an interdisciplinary empirical study,' *Journal of Risk and Financial Management*, 16(5), p. 265. <https://doi.org/10.3390/jrfm16050265>

Sarikaya, A., Correll, M., Bartram, L., Tory, M. and Fisher, D., 2018. What do we talk about when we talk about dashboards? *IEEE transactions on visualization and computer graphics*, *25*(1), pp.682-692.

Sedrakyan, G., Mannens, E. and Verbert, K. (2019) 'Guiding the choice of learning dashboard visualizations: Linking dashboard design and data visualization concepts,' *Journal of Computer Languages*, 50, pp. 19–38. <https://doi.org/10.1016/j.jvlc.2018.11.002>.

Siddika, A., Kobra, M.K., Tanin, S. and Afrin, S., 2018. Non-government organizations and their contribution in rural development: An example of Bagatipara upazila, Natore district, Bangladesh. *International Journal of Innovative Science and Research Technology*, *3*(4), pp.747-752.

Todaro, M. P., & Smith, S. C. (2014). Economic development (12th ed.). Pearson.

Tuo, S. and He, H., 2021. A study of multiregional economic correlation analysis based on big data—Taking the regional economy of cities in Shaanxi Province, China, as an example. *Sustainability*, *13*(9), p.5121.

Verbeek, M. (2017) 'Using linear regression to establish empirical relationships,' *IZA World of Labor* [Preprint]. <https://doi.org/10.15185/izawol.336>.

Vila, R.A., Estevez, E. and Fillottrani, P.R., 2018, April. The design and use of dashboards for driving decision-making in the public sector. In *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance* (pp. 382-388).

Wen, J. *et al.* (2021) 'The impact of financial development on economic indicators: a dynamic panel data analysis,' *Ekonomska Istrazivanja-economic Research*, 35(1), pp. 2930–2942. <https://doi.org/10.1080/1331677x.2021.1985570>.

Yuan, D. *et al.* (2021) 'Regression Analysis and Comparison of Economic Parameters with Different Light Index Models under Various Constraints,' *Sensors*, 21(22), p. 7561. <https://doi.org/10.3390/s21227561>.

**Appendices**

Top of Form