




# Tk Tk report

-  Quick Submit
-  Quick Submit
-  Wright College

---

## Document Details

### Submission ID

trn:oid:::1:2751949581

### Submission Date

Nov 15, 2023, 12:20 AM CST

### Download Date

Nov 15, 2023, 12:25 AM CST

### File Name

6620505\_1760161999\_AppliedStatisticsandDataVisual.docx

### File Size

487.9 KB

58 Pages

9,516 Words

58,790 Characters

How much of this submission has been generated by AI?

**\*2%**

of qualifying text in this submission has been determined to be generated by AI.

\* Low scores have a higher likelihood of false positives.

**Caution: Percentage may not indicate academic misconduct. Review required.**

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

## Frequently Asked Questions

### What does the percentage mean?

The percentage shown in the AI writing detection indicator and in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was generated by AI.

Our testing has found that there is a higher incidence of false positives when the percentage is less than 20. In order to reduce the likelihood of misinterpretation, the AI indicator will display an asterisk for percentages less than 20 to call attention to the fact that the score is less reliable.

However, the final decision on whether any misconduct has occurred rests with the reviewer/instructor. They should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in greater detail according to their school's policies.



### How does Turnitin's indicator address false positives?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be AI-generated will be highlighted blue on the submission text.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

### What does 'qualifying text' mean?

Sometimes false positives (incorrectly flagging human-written text as AI-generated), can include lists without a lot of structural variation, text that literally repeats itself, or text that has been paraphrased without developing new ideas. If our indicator shows a higher amount of AI writing in such text, we advise you to take that into consideration when looking at the percentage indicated.

In a longer document with a mix of authentic writing and AI generated text, it can be difficult to exactly determine where the AI writing begins and original writing ends, but our model should give you a reliable guide to start conversations with the submitting student.

### Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify both human and AI-generated text) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

# IMPACT OF SOCIO-ECONOMIC INDICATORS ON ECONOMIC DEVELOPMENT AMONG DIFFERENT COUNTRIES

By (Name)

The Name of the Class (Course)

Professor (Tutor)

The Name of the School (University)

The City and State Where it is Located

The Date

## Table of Contents

<b>Section 1</b>	6
Introduction and data exploration	6
1.1 Introduction	6
Statistical Analysis Objectives	6
Hypothesis 1:	7
Interactive Dashboard Design Objectives	8
1.2 Background Research and Literature Review	8
<b>1.2.1 Research Background</b>	8
<b>I. Correlation</b>	8
II. Hypothesis Testing	9
III. Regression	9
IV. Interactive Dashboard Design	10
<b>1.2.2 Literature Review</b>	12
1.2 Preparation and Data Set Exploration	13
<b>1.3.1 Data Dictionary</b>	13
<b>1.3.2 Data Preparation</b>	14
Handling missing data	14
Outlier Detection and Handling:	16
<b>1.3.3 Exploratory Data Analysis (EDA)</b>	18

<b>Section 2:</b>	21
<b>Statistical Analysis</b>	21
2.1 Descriptive Statistics	21
R steps	21
2.2 Correlation Analysis	22
R analytic steps	24
2.3 Hypothesis Testing	24
<b>Hypothesis 1</b>	24
<b>R steps</b>	25
<b>Hypothesis 2</b>	26
2.4 Regression Analysis	27
<b>2.4.1 Linear Regression</b>	27
<b>R Analytics Steps:</b>	28
2.4.2 Multiple Regression	29
<b>R Analytics Steps:</b>	30
Ridge Regression:	31
<b>R analytic steps</b>	33
Lasso Regression:	34
<b>R analytic steps</b>	35
2.4.4 Time Series	35

Time Series Models .....	39
<b>ARIMA (Autoregressive Integrated Moving Average) Model:</b> .....	39
<b>R analytic steps</b> .....	40
<b>VAR (Vector Autoregression) Model:</b> .....	41
<b>R analytic steps</b> .....	42
2.2 Discussion and Conclusion .....	43
<b>Limitations:</b> .....	44
<b>Conclusion:</b> .....	44
<b>Part Three: Interactive Dashboard Design</b> .....	45
3.1 Introduction.....	45
3.2 Data Visualization Principles:.....	45
3.3 Data Pre-processing .....	46
3.4 Design Rationale.....	47
3.4 Visual Paradigm Selection.....	48
3.5 Conceptual Model.....	48
<b>3.5.1 Integration of Individual Workflows</b> .....	49
<b>3.5.2 Unified Narrative</b> .....	49
<b>3.5.3 Focus Context Principles</b> .....	49
<b>3.5.4 Illustration</b> .....	50
Discussion and Conclusion .....	50

References ..... 52

# Section 1

## Introduction and data exploration

### 1.1 Introduction

This report examines the relationships among data, socioeconomic development, and information dissemination. The basis of this report is Non-Governmental Organizations (NGOs). An organization that works to advance social and economic development globally (Amin, 2017). According to Siddika et al., (2018), the study's significance and the insights sought to understand the importance of clear communication are highlighted by the non-governmental organization's unwavering dedication to improving people's lives worldwide.

The goal of this study is to tackle a multifaceted issue. The goal of this research is to use data to uncover insights that will benefit the general public as well as offer a comprehensive picture of socioeconomic progress. Through the analysis of important variables, this study aims to address issues related to population dynamics, employment, economic growth, poverty reduction, education, and public health. This data is important because it has the potential to empower people, inform decision-makers, and effect change.

Two separate but related tasks are part of a strategy for addressing this issue. While Task 2 focuses on creating an interactive dashboard, Task 1 is primarily concerned with statistical analysis. These tasks correspond with the objectives of the research.

### Statistical Analysis Objectives

- 1.To examine at how social-economic variables relate to one another in order to comprehend how they affect economic development.
2. To investigate into how the social economic indicators are related to one another.



3.To test hypotheses on specific socioeconomic indicators in order to derive relevant population inferences from sample data.

4.To conduct regression study that generates predictions by modeling the relationships between socioeconomic variables.

### **Hypothesis 1**

Null Hypothesis (H0); The Life Expectancy (LE) is not significantly different, between countries with a Poverty Headcount (PH) below the average and countries with a PH above the global average.

Alternative Hypothesis (H1); Life Expectancy is significantly different between countries with a poverty Head count below the global average and countries with a Poverty Head Count above the global average.

### **Hypothesis 2**

Null Hypothesis (H0); The mean Education Enrollment does not differ significantly between countries with low Gross Domestic Product and countries with high GDP.

Alternative Hypothesis (H1); There is a significant difference in the Education Enrollment differ significantly between countries with low GDP and countries, with high GDP.

The commitment to these objectives is motivated by the conviction that thorough analysis may clarify the current connections between various socioeconomic factors, providing direction for more successful development methods and policy choices.

## **Interactive Dashboard Design Objectives**

1. To create an interactive dashboard offering a comparison of social and economic metrics for a chosen group of nations.
2. The dashboard should be structured so that users can quickly compare the performance of different countries across multiple indicators and years.

An interactive dashboard aims to bridge the knowledge gap between data and statistics by providing statistical insights to a broad audience (Vila, Estevez and Fillottrani, 2018; Sarikaya et al., 2018). This tool will enable users to investigate and comprehend the vital relationships between socioeconomic indices, fostering global social and economic advancement.

## **1.2 Background Research and Literature Review**

This section presents background research and literature review related to the techniques used in this study.

### **1.2.1 Research Background**

#### **I. Correlation**

An essential part of studying socio-economic development is figuring out the degree of association between variables, and correlation analysis gives this information. Pearson's correlation coefficient is the foundation of correlation analysis. Todaro and Smith (2014) state that this coefficient provides essential information about the relationship between changes in one variable and changes in another by identifying the direction and strength of a linear relationship between variables.

One method for examining the connections between socioeconomic data is correlation analysis. According to Todaro and Smith (2014), this approach is especially useful for figuring

out whether there is a linear link between these variables, which aids in understanding the complex web of socioeconomic advancement.

Through a careful study of earlier research and studies, correlation analysis has been utilised to investigate linked relationships. Studies by Festin et al.,(2017); Tuo and He (2021), which are notable examples of how correlation analysis was successfully utilised to examine socio-economic links, support the validity of the method and its significance in understanding the interrelationships between your selected variables.

## **II. Hypothesis Testing**

A systematic method for drawing conclusions about populations from sample data is hypothesis testing (Levine, 2022). Hypothesis testing is an essential technique in this research environment for making strong conclusions about socio-economic development. This study necessitates the careful selection of statistical tests suited for the research aims and data type. Consider using t-tests to compare means between two groups.

These hypothesis techniques have been used in many studies to investigate relevant problems. Hypothesis testing was used in studies like Galkina (2022) to look into socioeconomic links, which supports the applicability of these techniques in this situation. These studies offer a strong foundation for this research and attest to the efficacy of hypothesis testing.

## **III. Regression**

Regression analysis enables the modeling of interactions between dependent and independent factors, which is necessary to understand socioeconomic development (Judd et al., 2017). Theoretical foundations span a variety of regression forms. Linear regression is the most

effective technique for determining the association between one independent variable and one dependent variable.

This can be very helpful in this research to comprehend the direct impact of variables such as GDP or enrollment in school on socioeconomic outcomes. In contrast, multiple regression allows for the inclusion of more independent variables, allowing for a more thorough analysis of socioeconomic progress.

Numerous studies on socio-economic development have been conducted, which supports the validity and suitability of the selected regression methodologies. Research conducted by Yuan et al., (2021) utilized regression analysis to examine similar inquiries, bolstering the justification for this methodology.

#### **IV. Interactive Dashboard Design**

The clear and user-friendly communication of complicated socio-economic data to users is largely dependent on the design of interactive dashboards. Vila, Estevez and Fillottrani (2018) state that this investigation ought to cover a number of important components, such as composition, layout, and design principles. It should also give a glimpse into the current viewpoints that are influencing the area and insights into the dashboard design and development approach.

According to Vila, Estevez and Fillottrani (2018), an effective interactive dashboard can be identified by its content, layout, and adherence to design principles that promote comprehension and usability. The following important areas are examined in the literature review:

- **Best Practices for Data Visualization.** The efficient presentation of complicated data is a fundamental component of interactive dashboards. Numerous insights on best practices for data visualization are available in the literature, such as the use of clear labeling, the selection of suitable chart formats, and the avoidance of visual clutter. Sedrakyan, Mannens and Verbert's (2019) works offer helpful connections in this field.
- **Color Selections for Improved Understanding.** When it comes to data visualization and dashboard applications, color theory is essential. Researchers Nadj, Maedche and Schieder, (2020) have made significant additions to our understanding of color in data visualization through their studies.
- **The addition of interactive components.** A user-friendly experience necessitates interactive elements like drill-down features, tooltips, and filters. The effective integration of these components into dashboard design to allow users to study and change data is examined in-depth in a comprehensive literature analysis. The seminal works of Nadj, Maedche and Schieder, (2020) provide insights into the fundamentals of interactive data visualization.

A crucial part of this research is the methods used in the design and development of interactive dashboards. The explanation of this methodology reveals how it matches with the dashboard design objectives. Making connections to well-established design techniques that have played an important part in the design process.

This concludes with some current viewpoints on interactive dashboard design. This part offers a glimpse into the state of the industry by examining the most recent trends and creative approaches. According to Nadj, Maedche and Schieder (2020), the presentation of different

viewpoints shows how they shaped the interactive dashboard's design decisions, giving it contemporary concepts and significance.

### 1.2.2 Literature Review

This section examines previous research to acquire a thorough grasp of the subject of the study. Important socio-economic indicators are covered in this research, such as GDP, Life Expectancy, Unemployment Rate, Poverty Rate, and Enrollment in Education. These indicators are essential for comprehending the dynamics of economic development in various countries.

A comprehensive review of pertinent literature is conducted in order to address research objectives and develop an efficient approach to meet them. Scholars like Pacifico (2023); Wen et al., (2021), have made noteworthy advancements in comprehending the role of these indicators in molding the socio-economic terrain. Their research offers insightful information about a variety of topics, including the factors that influence GDP growth, the effects of poverty rates on social welfare, and the implications of unemployment on economic stability.

The scope of this research is broadened to include the development and deployment of a dashboard that effectively presents these socio indicators. It is imperative to have looked at other dashboards and data visualizations that are pertinent to this field of study in order to make sure the interactive dashboard complies with accepted standards and design values. Notable case studies from IvyProSchool (2023) and Belghith et al. (2022) offer valuable ideas on creating a user-friendly dashboard that presents socioeconomic data.

Examining these dashboard case studies, helps obtain ideas and best practices that directs the creation of dashboard for this research, guaranteeing that it satisfies the current requirements in the field of exploring socio-economic development.

## 1.2 Preparation and Data Set Exploration

A thorough summary of the data set, including the variable names, definitions, metadata, time periods, and data sources, is given in this section. The methods for addressing missing data, detecting outliers, and preparing data are also covered in this section. After the data is prepared, exploratory data analysis (EDA) is carried out to find intriguing insights about the data set, which are then displayed using the relevant graphs.

### 1.3.1 Data Dictionary

Table 1: Data Dictionary

Variable Name	Definition	Time Frame	Data Source	Metadata Code
GDP	Total value of goods and services produced	2006-2020	World Development Indicators	NY.GDP.MKTP.CD
Poverty Headcount Ratio (PH)	Percentage of population below poverty line	2006-2020	World Development Indicators	SI.POV.DDAY
Unemployment Rate (UR)	Percentage of unemployed people in the labour force	2006-2020	World Development Indicators	SL.UEM.TOTL.ZS

Education Enrolment (EEL)	Number of students enrolled in education	2006- 2020	World Development Indicators	SE.PRM.ENRL
Population Growth Rate (P)	Percentage change in population	2006- 2020	World Development Indicators	SP.POP.GROW
Life Expectancy (LE)	Average lifespan of a population	2006- 2020	World Development Indicators, UN Bank Data	SP.DYN.LE00.IN

### 1.3.2 Data Preparation

The dataset used in this analysis was collected from two primary sources: the United Nations Bank Data and the World Development Indicators. The dataset encompasses information covering the duration from 2006 to 2020. To ensure the data's quality, several data cleaning and preparation steps were undertaken.

#### Handling missing data

Missing Data Detection Algorithm:

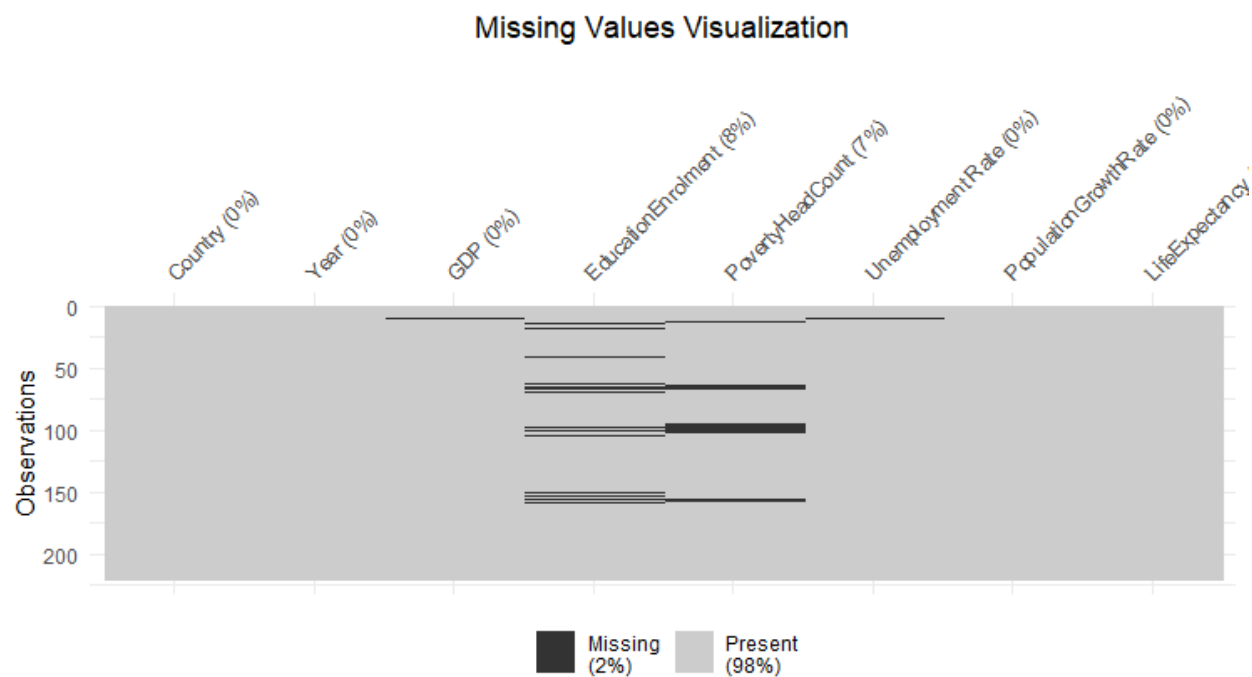
```
missing_counts <- colSums(is.na(MISSING_DATA))
```

```
print(missing_counts)
```

Country	0	Year	0	GDP	EducationEnrolment	1
18	PovertyHeadCount	Unemployment Rate	PopulationGrowthRate	LifeExpe	ctancy	0
0	16	1	0			



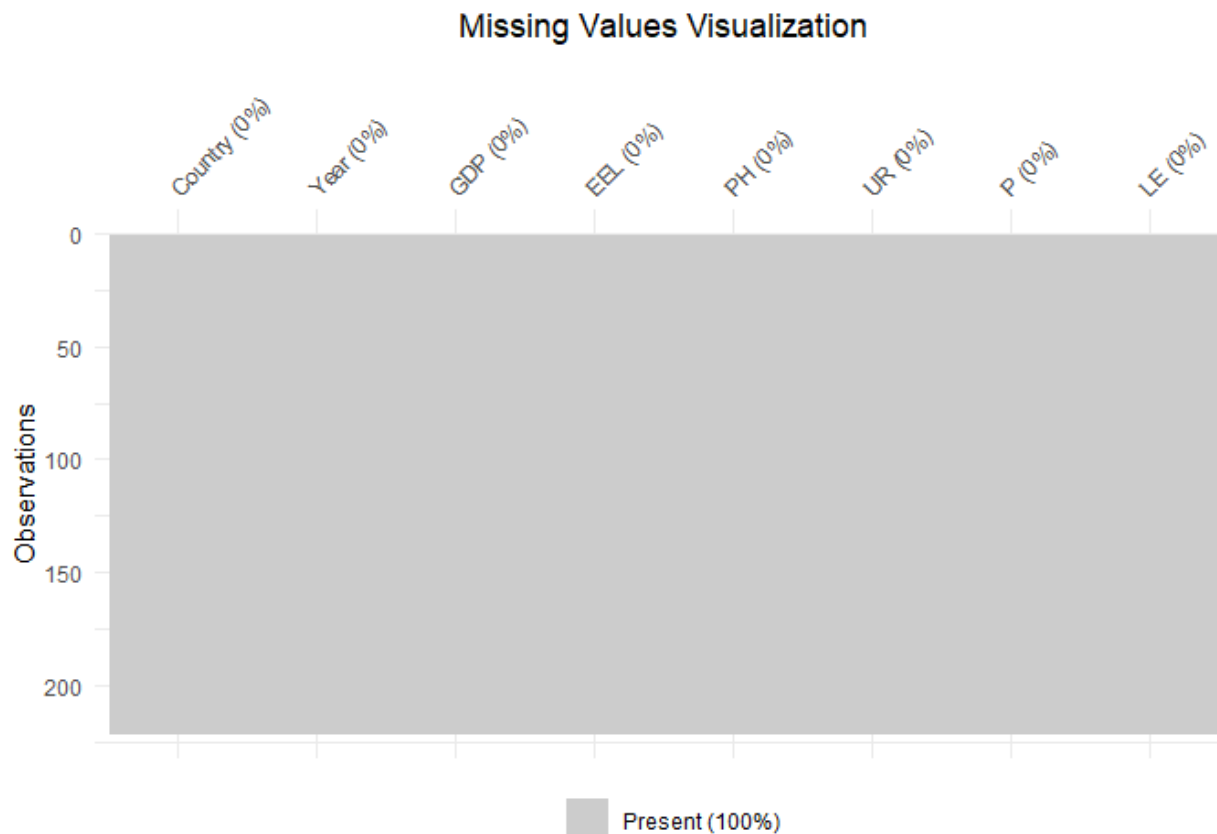
Figure 1: Missing data visualization



The Heatmap above provides a clear visual representation of missing values across different variables and time points.

Where historical data was available, missing values were imputed by referring to data from previous years. Below is a visualization after replacing the missing values.

Figure 2: Missing value visualization after replacement



## Outlier Detection and Handling:

Outlier Detection Algorithm:

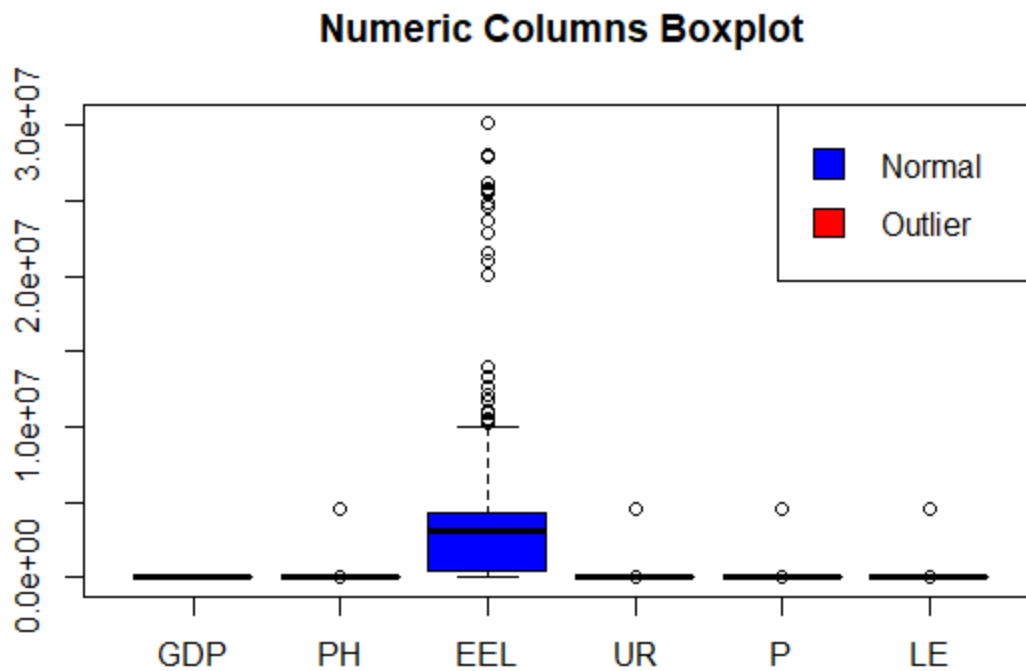
# Detect outliers using Z-score

```
z_scores <- scale(my_data)
```

```
outliers <- abs(z_scores) > 2 #
```

GDP	PH	EEL	UR	P	LE
0	4	10	3	7	5

Figure 3: Outlier visualization



Above is a boxplot showcasing outliers.

## 2.Outlier Handling: Median replacement

# Replace outliers with the median

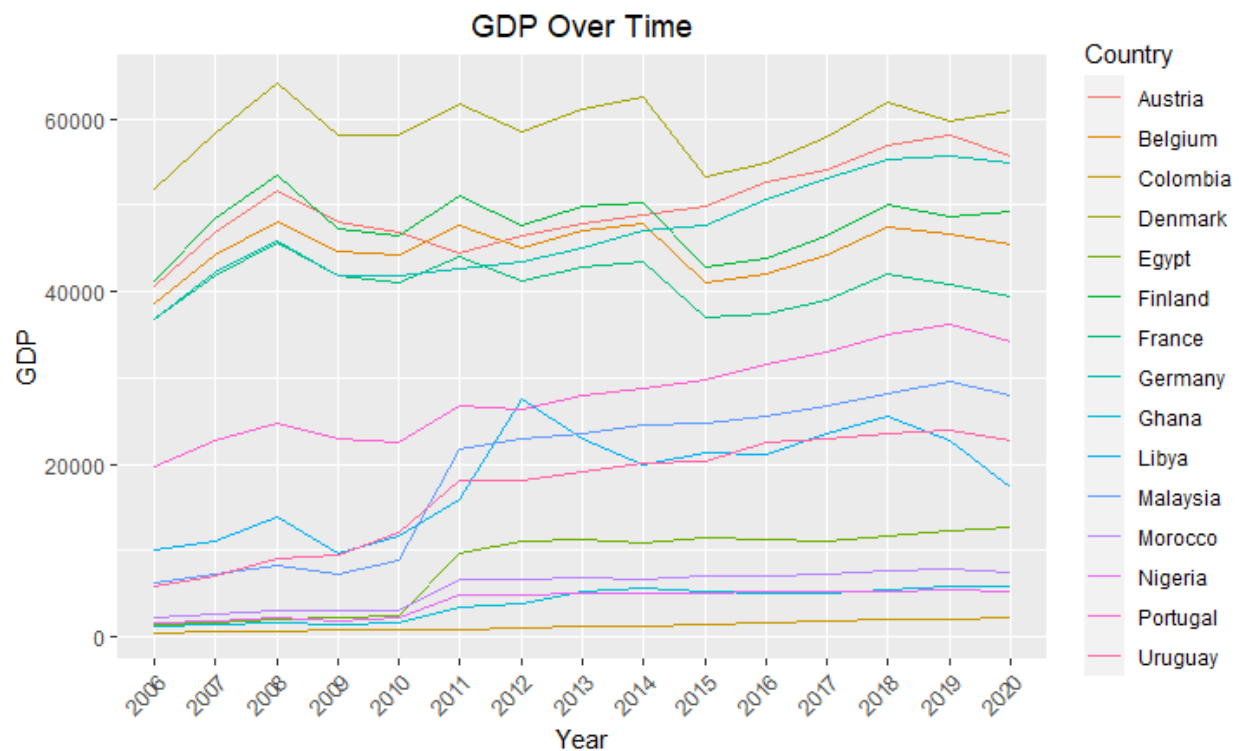
```
dataset[outliers] <- median(my_data, na.rm = TRUE)
```

```
[1] "Handling outliers in column: PH"
[1] "Original median: 1.0118404251491"
[1] "New median: 1.0118404251491"
[1] "Handling outliers in column: UR"
[1] "Original median: 8.57"
[1] "New median: 8.57"
[1] "Handling outliers in column: P"
[1] "Original median: 1.13154130595269"
[1] "New median: 1.13154130595269"
[1] "Handling outliers in column: LE"
[1] "Original median: 76.471"
[1] "New median: 76.471"
```

### 1.3.3 Exploratory Data Analysis (EDA)

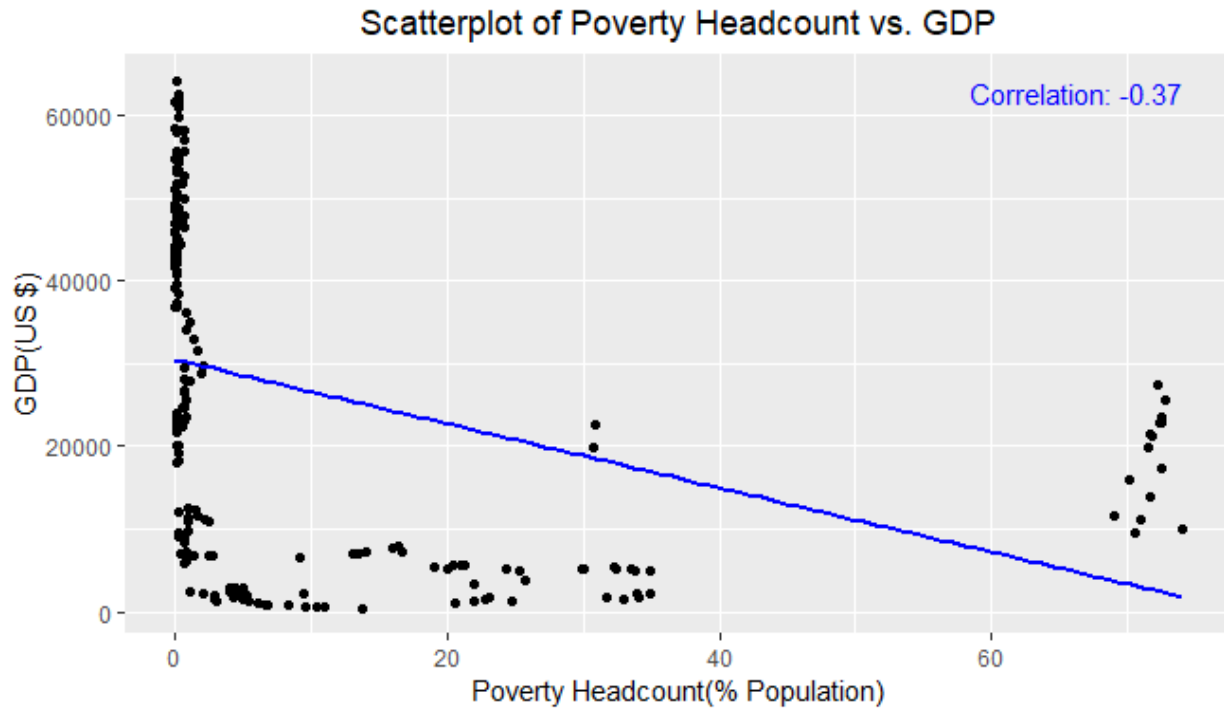
During the process of Exploratory Data Analysis insights about the dataset were uncovered. It was observed that the GDP of countries experienced an increase, over time although at varying rates of growth.

Figure 4: Time series plot of GDP over Time



GDP and the number of people living in poverty were clearly correlated negatively, indicating that higher GDP is associated with lower rates of poverty. The differences in unemployment rates between nations were a reflection of different labour market dynamics.

Figure 5:Scatterplot



There was an overall upward trend in education enrollment rates, suggesting greater accessibility to education. Different countries experienced different rates of population growth; some saw rapid expansion while others observed steady growth.

Figure 6: visualizing population growth by country

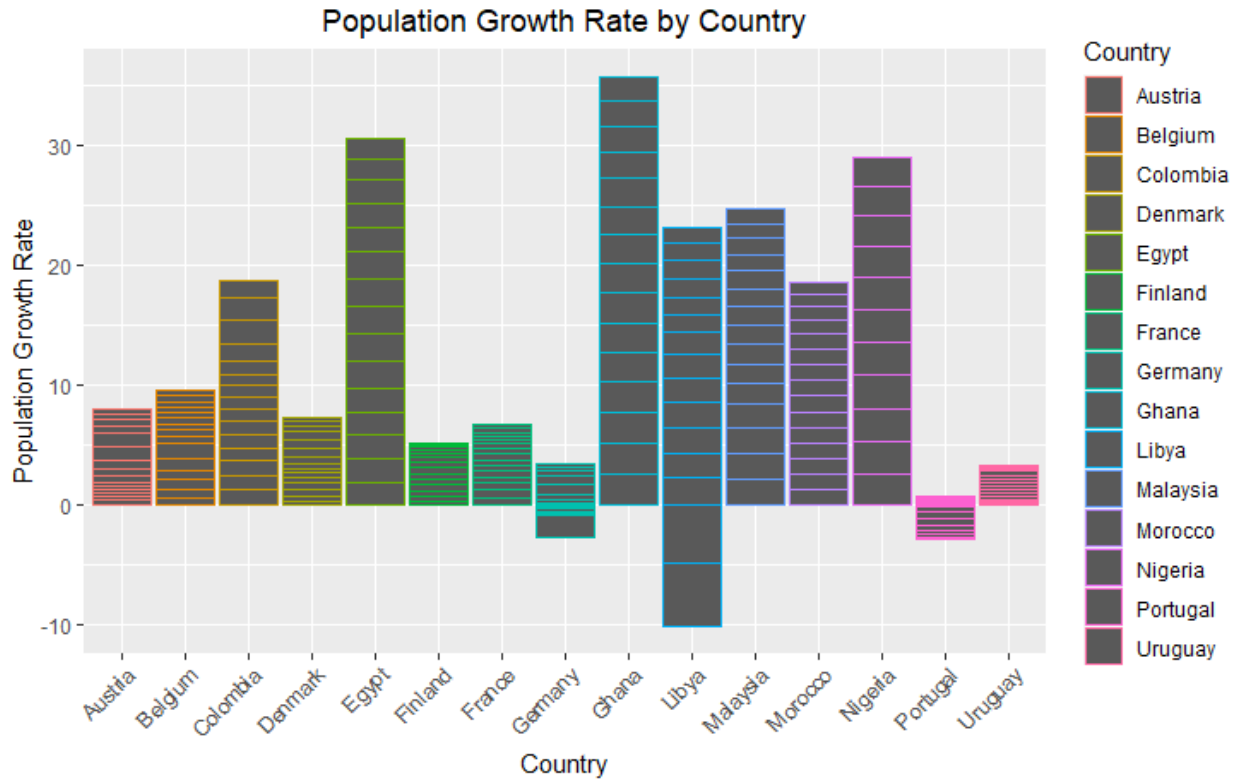
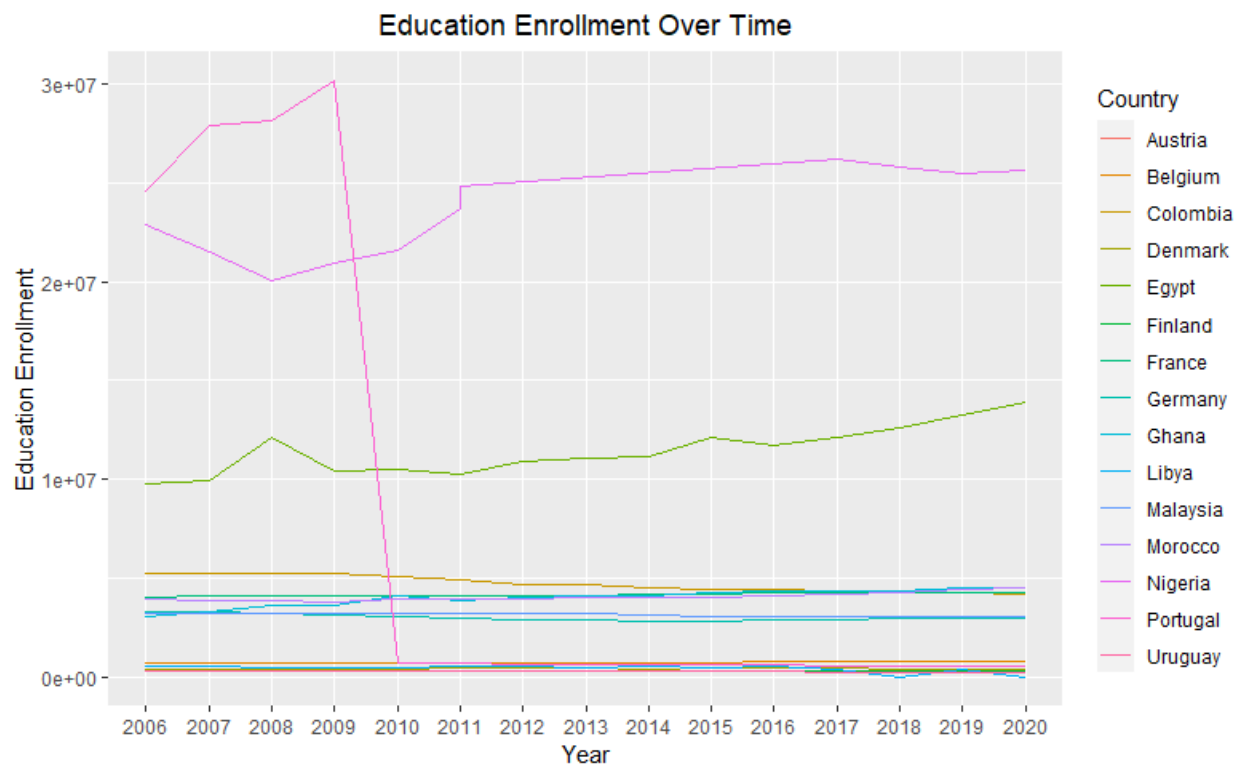


Figure 7: Education Enrollment over Time



Global improvements in health outcomes were highlighted by the general improvement in life expectancy. The dashboard design and statistical analysis that followed were informed by these EDA-gleaned findings.

## Section 2:

### Statistical Analysis

#### 2.1 Descriptive Statistics

Variable		Mean	Median	Mode	SD	Skewness	Kurt
osis							
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: Import or load the dataset into R using appropriate functions like **read.csv()**
2. Calculate Summary Statistics:

Use the `mean()` function to calculate the average value of the variable.

Use the `median()` function to find the middle value of the variable.

Use other packages/functions available for mode calculation.

Use the `sd()` function to compute the standard deviation of the variable's values.

**Skewness and Kurtosis:** Utilize functions from packages like moments (e.g., skewness () and kurtosis ()) to compute skewness and kurtosis, respectively.

In the analysis of socio-economic indicators across diverse countries, distinct patterns and disparities emerge, illuminating essential aspects of economic development. Gross Domestic Product (GDP) statistics reveal a notable variation, with a mean of approximately \$26,738.18 and a median of \$24,004.82, indicating differing economic statuses among nations. Education Enrolment (EEL) data illustrates significant variability, showcasing a mean of around 4.25 million and a median of approximately 3.01 million. Poverty Headcount (PH) statistics exhibit a wide discrepancy, with a mean of about 9.72% and a significantly lower median of 0.70%, highlighting concentrated disparities in poverty rates across the countries.

Furthermore, Unemployment Rate (UR) analysis portrays a moderate distribution, with a mean unemployment rate of approximately 9.22% and a slight concentration around 9.13%. Population Growth Rate (P) statistics reflect substantial variability, indicating outliers influencing the distribution, while Life Expectancy (LE) figures demonstrate diverse spans among nations. These insights into socio-economic indicators emphasize the multifaceted nature of economic development, showcasing disparities, distributions, and outliers crucial for understanding the diverse economic landscapes across different countries.

## **2.2 Correlation Analysis**

The correlation analysis unveils significant relationships among various socio-economic indicators, shedding light on their interconnectedness and potential implications for economic development:



There exists a moderate negative correlation (-0.44) between Gross Domestic Product and Education Enrolment . This negative correlation suggests that in the analysed dataset, as GDP increases, there tends to be a decrease in education enrolment rates. This finding might imply that in some regions or countries with higher GDP, there could be a tendency for reduced enrolment rates, which could signal issues related to education accessibility or priorities in resource allocation.

A moderate negative correlation (-0.37) between GDP and Poverty Headcount (PH) indicates that higher GDP is associated with lower poverty headcount rates. It suggests a tendency where higher economic output might contribute to reduced poverty rates. This correlation can be crucial for policymakers as it highlights the potential impact of economic growth on poverty alleviation.

A weak negative correlation (-0.16) between GDP and Unemployment Rate suggests a limited relationship in the dataset. This finding might indicate that higher GDP doesn't strongly correlate with lower unemployment rates, hinting at complexities within the economy where economic growth doesn't directly translate to reduced unemployment.

A strong negative correlation (-0.55) between GDP and Population Growth Rate indicates that higher GDP is associated with lower population growth rates. This correlation suggests a pattern where countries with higher economic output might experience slower population growth.

A strong positive correlation (0.72) between GDP and Life Expectancy showcases a noteworthy relationship where higher GDP is associated with increased life expectancy. This correlation highlights the potential positive impact of economic development on the health and longevity of a population.

These findings provide insights into the interdependencies among socio-economic indicators, emphasizing the complex relationships crucial for policymakers and organizations aiming to foster economic development and improve societal well-being.

## 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 formulated hypotheses sought to discern if there exists a notable disparity in the mean life expectancy between nations with Poverty Head counts figures below and above the global average. The null hypothesis (H0) posited that no substantial difference exists in LE means, while the alternative hypothesis (H1) asserted the presence of a significant difference in LE means based on PH categorization.

```
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
```

Upon conducting the Welch Two Sample t-test, the results yielded a remarkably low p-value of  $2.728 \times 10^{-14}$ , significantly deviating from the customary significance level of 0.05. This outcome compellingly rejects the null hypothesis, providing compelling evidence in support of the alternative hypothesis. The calculated t-value of 10.081 further reinforces this conclusion, emphasizing a substantial and noteworthy difference in the mean life expectancies between the two groups.

Countries with a Poverty Headcount below the global average demonstrated a considerably higher mean life expectancy of approximately 78.13 years, in stark contrast to countries with a Ppopulation growth rate above the global average, which exhibited a notably lower mean life expectancy of approximately 65.42 years. The 95% confidence interval for the difference in means (10.18 to 15.23) underscores the robustness of this distinction, suggesting that the actual difference in life expectancy means likely falls within this range.

## 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

These findings bear substantial implications, highlighting the influential role of poverty levels on life expectancy across nations. The observed disparity underscores the critical importance of socio-economic factors, particularly poverty alleviation efforts, in positively influencing life expectancy rates. The statistical evidence robustly supports the notion that lower

poverty rates are associated with higher life expectancies, emphasizing the significance of socio-economic policies in enhancing the overall well-being and longevity of populations worldwide.

## Hypothesis 2

Formulating the hypotheses aimed to discern if a significant difference exists in the mean Education Enrolment Rate between countries categorized by their GDP levels. The null hypothesis (H0) stipulated that no noteworthy variance exists in EEL means, while the alternative hypothesis (H1) proposed a significant difference in EEL means based on GDP classification.

```
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 Welch Two Sample t-test conducted yielded compelling results, showcasing a remarkably low p-value of 6.079e-11, significantly diverging from the conventional significance level of 0.05. This exceptionally low p-value decisively rejects the null hypothesis, providing robust evidence in favour of the alternative hypothesis. The derived t-value of -7.1501 further fortifies this conclusion, highlighting a substantial and meaningful difference in the mean Education Enrolment Rates between countries characterized by their GDP levels.

Countries with higher GDP exhibited a significantly lower mean Education Enrolment Rate, estimated at approximately 1,483,725, in stark contrast to countries with lower GDPs, which demonstrated a notably higher mean Education Enrolment Rate, approximately 6,656,739. The 95% confidence interval for the difference in means (-6,604,667 to -3,741,360) emphasizes

the robustness of this disparity, suggesting that the actual difference in Education Enrolment Rates between these groups likely falls within this range.

These findings bear profound implications, emphasizing a significant correlation between economic prosperity and education enrolment rates across nations. The statistical evidence strongly supports the notion that higher GDP levels are associated with lower education enrolment rates, accentuating the pivotal role of economic factors in shaping educational accessibility and participation levels among diverse countries. This insight underscores the significance of policies aiming to bridge this gap, ensuring equitable education opportunities despite economic differences, thereby fostering a more inclusive and educated global society.

## 2.4 Regression Analysis

### 2.4.1 Linear Regression

According to Lilja and Linse (2022) linear regression analysis is apt for examining the relationship between variables, particularly when studying how one variable (dependent) responds to changes in other variables (independent). In this context, using linear regression to analyze Life Expectancy concerning socio-economic indicators like Gross Domestic Product, Poverty Headcount, Education Enrollment is appropriate. This method's suitability lies in its ability to assume a linear relationship, provide interpretable coefficients for variable effects, identify significant predictors impacting LE, and allow evaluation of the overall model fit and assumptions' validity.

The linear regression model results suggest a substantial relationship between Life Expectancy and the socio-economic indicators of Gross Domestic Product and Education Enrollment. The intercept coefficient indicates that with GDP and EEL held constant, the

expected Life Expectancy is approximately 71.58 years. Both GDP and EEL exhibit statistically significant associations with Life Expectancy, as denoted by their respective coefficients: 2.097e-04 and -4.817e-07. A one-unit increase in GDP is linked to an estimated increase of 0.0002097 in Life Expectancy, while a unit increase in EEL is associated with a decrease of -4.817e-07 in Life Expectancy.

```
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
```

These findings suggest higher GDP and lower Education Enrollment correspond to increased Life Expectancy, although the latter association is marginally negative. The model's high Multiple R-squared value of 0.6398 indicates that approximately 63.98% of the variability in Life Expectancy can be explained by GDP and EEL, validating the model's goodness-of-fit. These results underscore the significance of socio-economic factors in influencing Life Expectancy levels.

## R Analytics Steps:

1. Define the linear regression model using 'lm' function

```
model <- lm(LE ~ GDP + EEL, data = my_data)
```

## 2. View summary statistics of the regression model

```
summary(model)
```

Studies such that of Verbeek (2017) might explore how economic factors (GDP, poverty rates), educational attainment, or healthcare access relate to life expectancy within specific populations or across countries. Similar research might use linear regression to understand the impact and significance of various factors on life expectancy levels.

### 2.4.2 Multiple Regression

Multiple regression is suitable for examining the relationship between socio-economic indicators and Gross Domestic Product (GDP). It allows simultaneous assessment of multiple variables' impact on GDP, identification of significant contributors, understanding complex relationships, and provides predictive insights. This technique aligns with the objective of analyzing the association between socio-economic indicators and economic development, using GDP as a representative measure.

```
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
```

Upon scrutinizing the coefficients, it's evident that Education Enrolment (EEL) and Poverty Headcount (PH) lack statistical significance in influencing GDP, with coefficients of (p

= 0.3753) and ( $p = 0.1317$ ), respectively. Conversely, Unemployment Rate (UR) exhibited a noteworthy negative impact on GDP, as evidenced by its statistically significant coefficient ( $p < 0.001$ ). An increase in UR corresponds to a decrease in GDP. Similarly, Population Growth (P) demonstrates a significant negative relationship with GDP ( $p = 0.0335$ ), implying reduced GDP levels with higher population growth. However, Life Expectancy (LE) displayed a highly significant positive association with GDP ( $p < 0.001$ ), indicating that higher LE correlates with increased GDP levels.

Assessing the model fit, the analysis elucidated that approximately 56.68% of the GDP variability is explained by the socio-economic indicators (Multiple R-squared = 0.5668), signifying a moderately fitting model. Moreover, the overall model proved statistically significant ( $p < 0.001$ ), suggesting that at least one independent variable significantly relates to GDP.

Unemployment Rate, Population Growth, and Life Expectancy emerged as influential factors impacting GDP variations, while Education Enrolment and Poverty Headcount did not exhibit significant impacts in this study. This comprehensive analysis provides invaluable insights into how specific socio-economic factors collectively contribute to GDP changes, highlighting their complex interplay and their crucial role in social and economic development.

## R Analytics Steps:

- 1.Loading Required Libraries
- 2.Loading Data:
- 3.Creating a multiple regression model

```
model <- lm(GDP ~ EEL + PH + UR + P + LE, data = my_data)
```



#### 4. Generating a summary of the fitted regression model summary(model)

Research in the literature commonly employs multiple regression to study relationships between various socio-economic indicators and economic development proxies like Gross Domestic Product (GDP). Studies by Ni (2020) investigate how factors such as life expectancy, poverty rates, education levels, and healthcare access relate to Gross Domestic Product, aiming to understand their collective impact on economic growth. Similar research uses multiple regression to identify significant predictors of GDP and comprehend their combined influence on economic development, aligning with the current analysis on the relationship between socio-economic indicators and GDP.

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 (Bager et al., 2017; Jermia et al., 2020). This case focuses on 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.

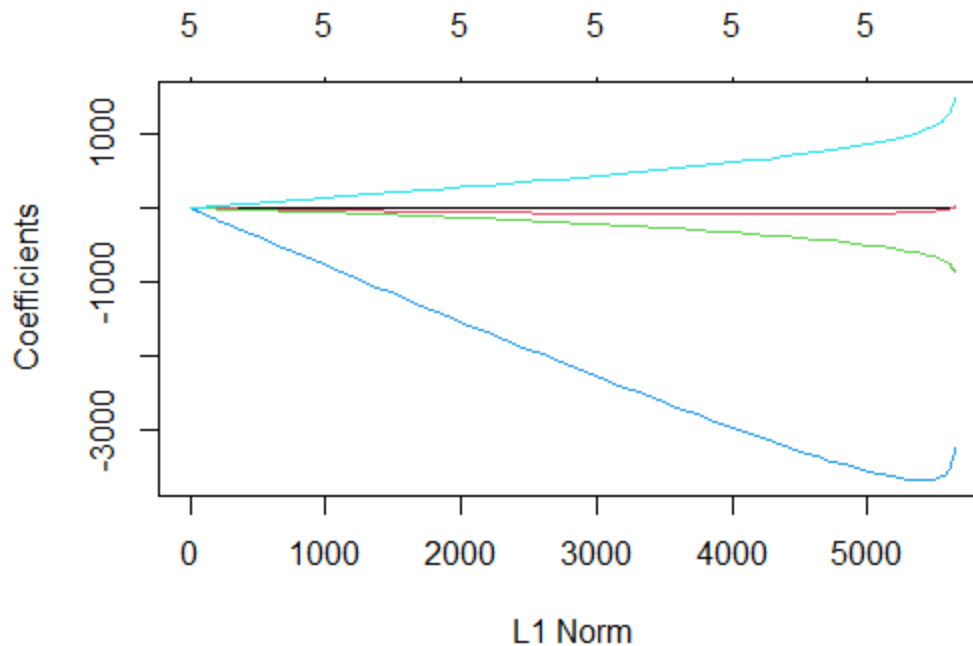
Analysis revealed valuable insights into how each of these indicators impacts social and economic development while addressing the issue of multicollinearity. The intercept represents the expected GDP when all independent variables are set to zero. The negative coefficient for Education enrolment suggests 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. 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.

Figure 8: Ridge Regression



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

Studies by Castro and López (2022) 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 in performing 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 (Altalbany, 2021).

The results provided 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.

### **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

Chen and Wu (2023) 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.

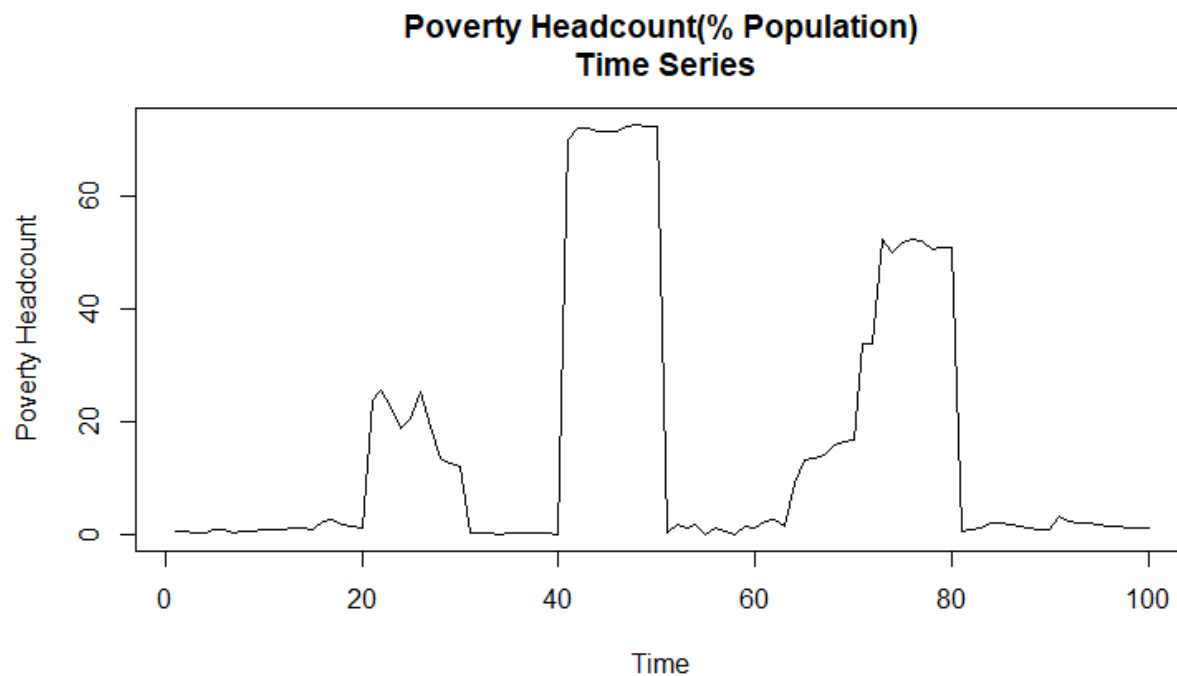
### **2.4.4 Time Series**

Understanding trends and patterns in socioeconomic development indicators across time can be achieved through the use of time series analysis. It offers insights into the dynamics of

these indicators by enabling researchers to look at how they alter and develop (Zhang et al., 2017). In this instance, the chosen socioeconomic indicators are investigated over time in order to evaluate the methods that are appropriate for accomplishing the stated goals.

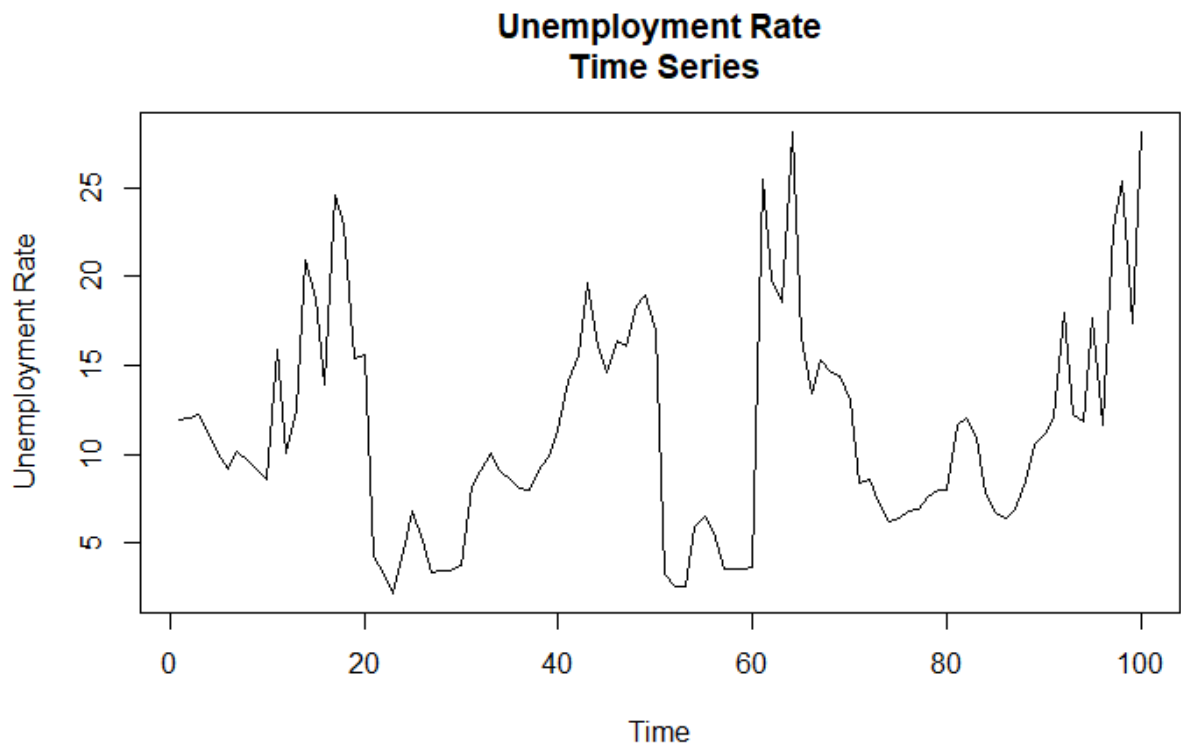
The time series analysis of the GDP, the head count of the poor, and the unemployment rate is the main focus. These are the core metrics used to evaluate economic development. Trends, seasonality, and other time-dependent patterns can be found with this technique.

Figure 9: Time series plot of Poverty Headcount



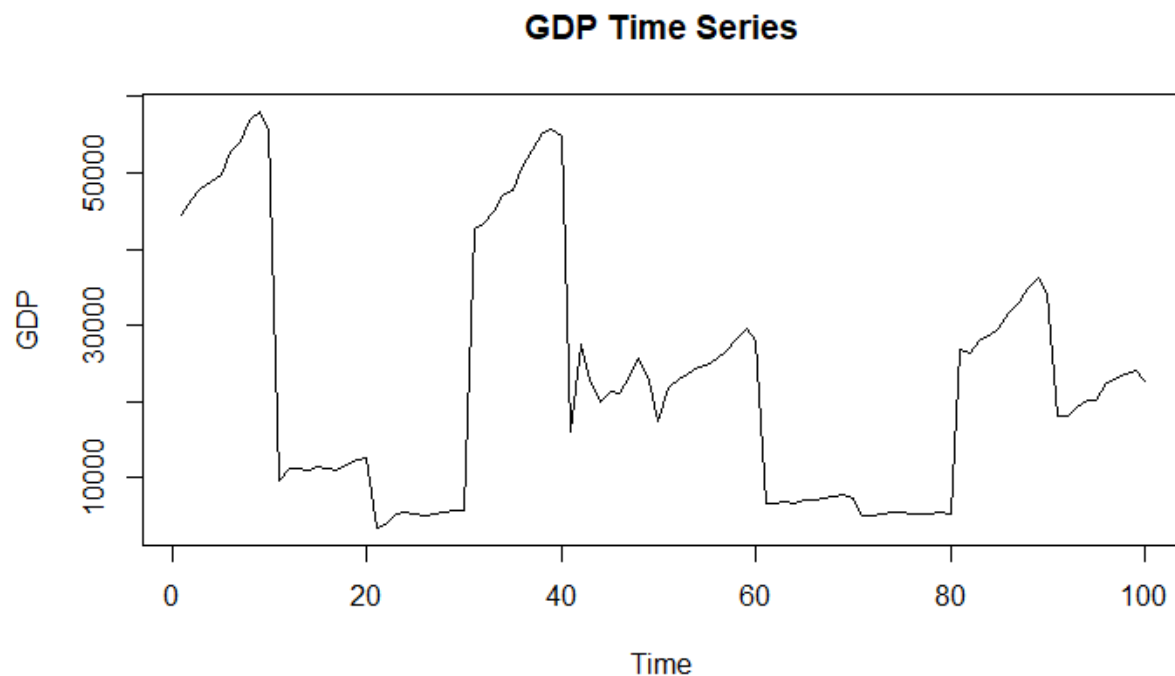
The graph above shows how poverty levels have evolved over time providing historical context for a nation's efforts to reduce poverty. It is a crucial tool for researchers to evaluate the effectiveness of anti-poverty initiatives and their long-term effects since it shows shifts in the poverty rate.

Figure 10: Time series plot of unemployment rate



The above chart shows how the unemployment rate has changed over time, giving information on a nation's labour market characteristics. It is necessary to comprehend the dynamics of the labour market and the state of the economy.

Figure 11: visualizing GDP overTime



The GDP time series above shows the state of the economy over a certain time period. It highlights periods of expansion and contraction, which are critical for evaluating the long-term health of the economy and the effects of policies. It displays trends and variances in a nation's economic output.

The research literature frequently uses time series analysis in socioeconomic studies, demonstrating the value of this methodology. Ren et al., (2019) have evaluated trends in socio-



economic indices in previous studies. Time series analysis was chosen since it is often used in relevant investigations in the literature.

## Time Series Models

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

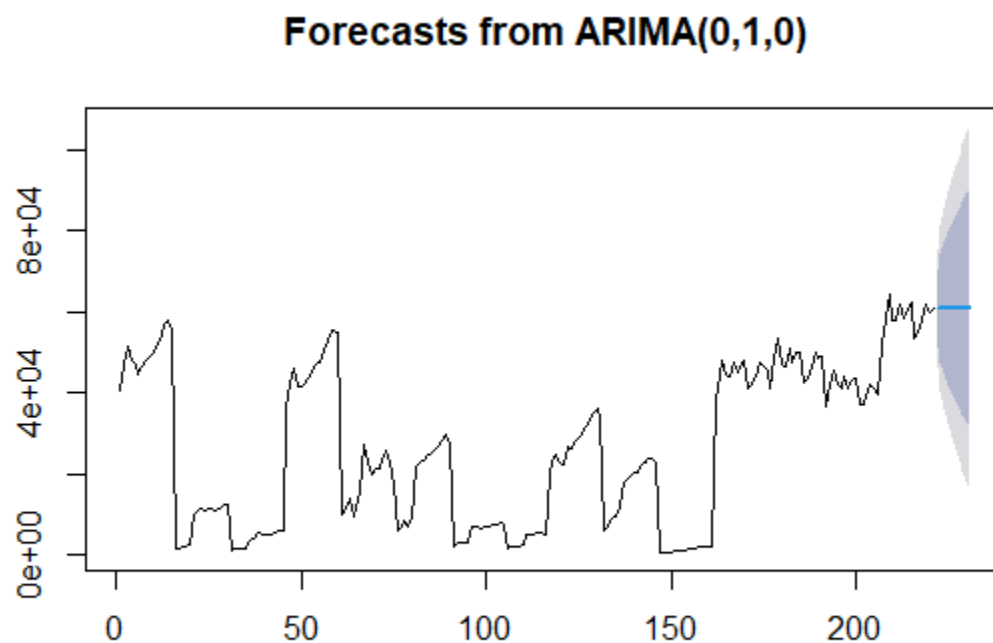
### **ARIMA Model:**

ARIMA( Autoregressive Integrated Moving Average)models are well-suited for capturing and forecasting the temporal dependencies within time series data (Vafin, 2020; Dai and Chen 2019). For this research, these models are particularly useful when the data exhibits trends or seasonal patterns. ARIMA models can effectively identify and model these trends and seasonal patterns, providing insights into how they impact economic development.

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 suggests an overall underestimation of values, with a relatively high Mean Absolute Percentage Error at 64.99%. The Mean Absolute Scaled Error close to 1 implies predictions akin to a naive forecast. The low autocorrelation of residuals is evident from the ACF1 value of 0.059.

Figure 12: visualizing ARIMA model



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 Adenomon, 2017; Sahib and Ibrahim (2022) 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 ideal for scenarios in which many time series variables interact with one another. GDP, poverty rates, and unemployment are all interrelated factors that determine economic progress. VAR models may describe the linkages and feedback mechanisms between different variables, allowing for a thorough understanding of how changes in one indication affect others. VAR models also aid in the identification of causal links between variables.

The study utilized the VAR (Vector Autoregression) model to examine the connections, between GDP, Poverty (PH) and Unemployment (UR) in relation to development. By utilizing a lag order of  $p = 2$  several important findings were obtained.

To begin with in the GDP equation it was observed that previous values of GDP had a impact on the current GDP as shown by a coefficient estimate of 1.02518. However the previous values of Poverty (PH) and Unemployment (UR) did not have an influence on GDP as indicated by significant coefficient estimates. The constant term in the GDP equation was found to be statistically significant suggesting a nonzero intercept.

Moving on to the Poverty (PH) equation the results revealed a relationship between past levels of poverty and current poverty rates with a coefficient estimate of 0.9158. However lagged values of GDP and UR did not have an effect on poverty rates. The constant term in the PH equation was not statistically significant.

In regards to Unemployment (UR) it was found that past unemployment rates had an impact on rates as indicated by a coefficient estimate of 0.6570. However, there were no coefficients for lagged values of GDP or PH or other lagged variables in relation to unemployment rates. 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 Onodugo (2017); Obalade et al., (2019) 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 delves over the technique and addresses any possible issues that arose during the analysis stage. In the framework of the socioeconomic growth of different nations, the multifaceted analysis included the main analytical techniques of correlation analysis, hypothesis testing, regression analysis, and time series analysis.

Correlation analysis played a role in examining the relationships between socio indicators. It allowed us to uncover associations and dependencies between variables. By understanding the strength of these relationships' correlation analysis laid the groundwork, for hypothesis testing. Regression analysis.

Two hypotheses were supported by the analysis, which revealed substantial connections between PH and LE as well as between GDP and EEL. This highlights the impact of poverty on LE as well as the good impact of higher GDP on education, which aligns with the aim to improve global socioeconomic development.

The study employed linear regression analysis to investigate the correlations among selected social- economic indicators. The results emphasised life expectancy and education as critical components of economic development while warning against the detrimental effects of unemployment. Interestingly, population growth and GDP did not appear to be correlated. The model's capacity to explain changes in GDP was validated by the R-squared value of 0.5668%. Through time series analysis, the GDP, PH, and Unemployment Rate showed temporal trends. The economy was growing as evidenced by the GDP's constant increase. A decline in PH indicated better efforts to combat poverty, but variations in Unemployment rate indicated cyclical employment patterns.

## Limitations:

It is critical to acknowledge particular constraints that need to be considered in the context of the study. First, there might have been problems with data consistency and quality since various nations may have utilised various standards and methods for collecting data, which might have affected the validity and comparability of the results. Second, because the research was observational in nature, it was challenging to ascertain the cause-and-effect relationships between the variables. Additionally, the availability of data for the time series analysis limited the depth of temporal insights. Finally, given that the selection of socio-economic indicators was based on the availability of data, it is possible that not all of the elements influencing socio-economic development were taken into consideration.

## Conclusion:

Finally, this research has shed light on how numerous aspects of development are influenced by critical variables and the dynamics of socioeconomic development. Descriptive statistics, correlation analysis, hypothesis testing, regression analysis, and time series analysis have all been combined to provide a comprehensive view of the relationships between socioeconomic indicators and economic and social growth.

The results emphasize how crucial it is to address poverty, education, healthcare and employment to foster progress. These findings offer insights, for policymakers and stakeholders who aim to enhance the wellbeing of the population and stimulate growth. This study successfully achieved its research goals aligning with the NGOs mission to advance development worldwide. It establishes a foundation, for making decisions and conducting further research in this critical field.

## **Part Three: Interactive Dashboard Design**

### **3.1 Introduction**

This section presents the goals of the dashboard project. The primary objectives are to develop a user platform that allows for comparison of economic and social indicators, in selected countries. Users will be able to evaluate and contrast countries across various metrics and time periods.

The dashboard aims to facilitate the examination of social indicators between nations ensuring accessibility for users with levels of data analysis expertise. It will offer a design that allows users to compare metrics simultaneously providing a comprehensive understanding of each country's performance. Additionally, the dashboard will enable the analysis of data over years aiding users in identifying trends and advancements in these countries. The intended audience includes policymakers and researchers seeking insights, into socio conditions.

### **3.2 Data Visualization Principles:**

The dashboard design is informed by fundamental data visualization principles that enhance the clarity, effectiveness, and user-friendliness of the platform (Lanning, 2021). Clarity was prioritized in visual representations, following the guidance of Edward Tufte's work on data visualization (Tufte, 2018). Tufte emphasizes that simplicity in design, while minimizing chartjunk, aids in conveying information effectively.

Appropriate visual encoding methods for data variables were employed, as recommended by (Few, 2013). For instance, use bar charts for comparing quantities and line charts for tracking trends over time. Consistency in design elements, such as colour schemes, labels, and scales, is

based on the principles outlined in "The Visual Display of Quantitative Information" (Tuft, 2018). Consistency across different visualizations aids users in making connections and understanding the data seamlessly.

Colour usage is influenced by the guidelines proposed in "Show Me the Numbers" by (Few, 2013). Colour is employed not only for aesthetic appeal but to convey information, and we avoid using too many colours that can be confusing. The incorporation of interactivity aligns with best practices in modern data visualization (Iliinsky, 2018). Features like filters, drill-through options, and tooltips enhance the user experience by allowing users to explore the data interactively.

### 3.3 Data Pre-processing

Data pre-processing is a crucial aspect of creating an informative dashboard. Key steps taken to prepare the data for our interactive dashboard:

**Data Collection:** Relevant datasets were gathered from sources such World Development Indicators (WDI) and United Nations Data Bank (UNdat), encompassing economic and social indicators for various countries over multiple years.

**Data Cleaning:** To ensure data consistency and reliability missing values, inconsistencies, and standardized units of measurement were addressed.

**Data Integration:** Data was unified from different sources and years into a single comprehensive dataset for dashboard use.

**Data Transformation:** Derived metrics were calculated, normalized data for consistent comparison, and improved usability.



**Data Aggregation:** Data was aggregated at various levels to provide summary information for the dashboard.

**Data Export:** The data was prepared for the dashboard development tool, structured for efficient querying and visualization. These steps set the foundation for an accurate, consistent, and meaningful dashboard showcasing economic and social indicators of selected countries.

### 3.4 Design Rationale

The design of the interactive dashboard is underpinned by key principles that prioritize user-friendliness, data visualization best practices, and effective communication of socio-economic data. To ensure clear navigation for users' Single screen layout was opted to provide a structured narrative. A carefully selected colour palette and visual consistency enhance the user experience, while a variety of charts were chosen to match the nature of the data. Interactivity is a vital component, allowing users to customize their analysis with drop-down menus and sliders.

Data attributes are seamlessly mapped to visual forms, enabling users to make quick comparisons and explore trends. Guided user support, including tooltips and information buttons, ensures users can effectively navigate and interpret the data. References to data visualization experts like Edward Tufte and Stephen Few have guided our design choices, emphasizing clarity, simplicity, and user-centric design. In essence, our dashboard is designed to empower a broad audience, including policymakers and researchers, in making data-driven decisions and disseminating knowledge.

### 3.4 Visual Paradigm Selection

A combination of visual paradigms that best serve the objectives of the dashboard were chosen, to enable users perform comparative analysis of economic and social indicators for selected countries. The key visual paradigms employed include:

**Bar Charts and Line Charts:** Bar charts were used to compare social economic indicators making them suitable for visualizing metrics for specific countries. Line charts were chosen for metrics such as GDP, education enrolment, and life expectancy. This choice aligns with best practices for illustrating trends over time. Line charts offer a clear representation of how these indicators change year by year, enabling users to grasp long-term progress and patterns effectively. These paradigms are rooted in the principles of data visualization put forth by (Tufte, 2013); (Few, 2018).

**Choropleth maps** are geospatial representations that utilize colour gradients to depict data variations across countries on a map. These maps are valuable for displaying geospatial patterns in economic and social data. Works by Maity and Maity (2021) have influenced the choice.

The selections of line charts and bar charts were made with the intention to provide a clear, intuitive, and consistent representation of the selected socio-economic indicators, ensuring that the dashboard will be both informative and user-friendly

### 3.5 Conceptual Model

A crucial aspect of effective dashboard design is the creation of a coherent and unified conceptual model that seamlessly integrates individual analysis workflows. This section will elaborate on the development of our conceptual model, highlighting how it merges distinct data

visualization elements into a single view that empowers users to gain meaningful insights through the principles of focus and context.

### 3.5.1 Integration of Individual Workflows

The dashboard design is founded on the integration of individual analysis workflows. Each socio-economic indicator has its unique storyline to tell, but the challenge lies in presenting these narratives in a manner that facilitates easy interpretation and comparison. To address this challenge, a screenshot layout that accommodates the representations of different indicators without overwhelming the users is incorporated. This approach ensures that the user can focus on specific details while maintaining a contextual understanding of the overall socio-economic landscape.

### 3.5.2 Unified Narrative

The main idea of the model is to create a story that links various workflows allowing for a comprehensive understanding of the data. This story aims to balance focus and context enabling users to switch between views of specific countries and years as well as a broader view that encompasses the entire dataset.

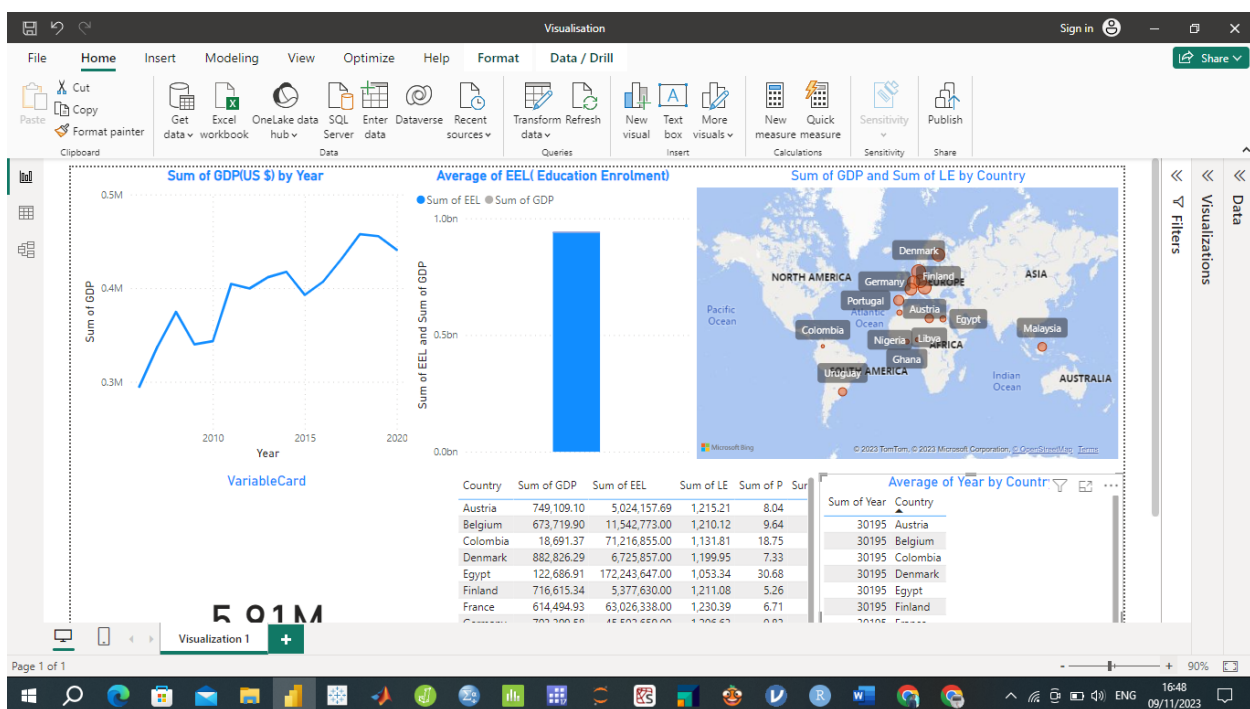
### 3.5.3 Focus Context Principles

The conceptual model is driven by the principles of focus and context, a technique frequently applied in data visualization to support comprehensive data exploration. It enables users to delve into specific aspects (focus) while retaining an awareness of the broader dataset context.

### 3.5.4 Illustration

To provide a visual representation of the conceptual model, is a screenshot that demonstrate how different elements interconnect within the dashboard. This will aid in understanding the flow and functionality of the dashboard, emphasizing how users can traverse between detailed insights and the overall socio-economic landscape seamlessly.

The following Screenshot illustrates the core components of our conceptual model:



The conceptual model serves as the foundation for the interactive dashboard design, effectively guiding users in their exploration of socio-economic indicators for selected countries.

### Discussion and Conclusion

This section critically evaluates the approaches and proposal formed throughout the development of interactive dashboard. Key aspects, including methodology, individual digital workflows, and the final composited dashboard are reviewed.

The Methodology for developing the dashboard consisted of several crucial stages. First was extensive data collection from reputable sources including World Development Indicators (WDI) and United Nations Data Bank (UNdat). The data cleaning process ensured high data integrity. Integration and transformation were essential for creating a unified dataset effectively utilized in the dashboard.

The project involved individual digital workflows for each socio-economic indicator, allowing users to explore and compare indicators across different countries and years. The use of line charts for showing the sum of each indicator over time provided a clear visual representation. The accompanying bar graphs offered a side-by-side comparison of these indicators. The maps allowed users to visualize the distribution of indicators across countries.

The composited dashboard effectively brings together these individual workflows into a single, coherent visual representation, such that Users can easily switch between indicators, and the "focus+context" principle is applied through the integration of various visualizations, including line charts, bar graphs, and maps. This provides a meaningful and comprehensive view of the socio-economic indicators of the selected countries.

In summary the proposed solution effectively achieves the goals outlined in the briefing document. The dashboard offers a user platform, for comparing social indicators across different countries and years. Users can easily. Compare how countries perform in terms of GDP, education enrolment, poverty levels, unemployment rates, population growth and life expectancy. The design of the dashboard follows principles of data visualization to ensure clarity, effectiveness and ease of use. It adheres to practices by utilizing chart types and colour coding for better comprehension. Moreover, it incorporates a "focus+context" approach by integrating visualizations that provide both an overview and detailed insights, into data points.

The dashboard caters to a wide audience, including policymakers, researchers, and stakeholders interested in socio-economic conditions. By providing a valuable tool for data-driven decision-making and knowledge dissemination, the dashboard fulfills its objectives effectively. It offers a comprehensive view of socio-economic indicators, empowering users to make informed comparisons and gain insights into global trends and disparities.

## References

- Abonazel, M.R. and Abd-Elftah, A.I., 2019. Forecasting Egyptian GDP using ARIMA models. *Reports on Economics and Finance*, 5(1), pp.35-47.
- Adenomon, M.O., 2017. Modelling and forecasting unemployment rates in Nigeria using ARIMA model. *FUW Trends in Science and Technology Journal*, 2(1B), pp.525-531.
- Altalbany, S., 2021. Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: A simulation study. *Journal of Applied Economics and Business Studies*, 5(1), pp.131-142.
- Amin, A., 2017. An institutionalist perspective on regional economic development. In *Economy* (pp. 59-72). Routledge.
- Bager, A., Roman, M., Algedih, M. and Mohammed, B., 2017. Addressing multicollinearity in regression models: a ridge regression application.
- 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](https://doi.org/10.1007/978-3-031-14615-2_45).

Castro, D.A. and López, M.J. (2022) 'Predicting socioeconomic indicators using transfer learning on imagery data: an application in Brazil,' *GeoJournal*, 88(1), pp. 1081–1102.

<https://doi.org/10.1007/s10708-022-10618-3>.

Chen, J. and Wu, J. (2023) 'The prediction of Chongqing's GDP based on the LASSO method and chaotic whale group algorithm–back propagation neural network–ARIMA model,' *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-42258-z>.

Dai, J. and Chen, S., 2019, October. The application of ARIMA model in forecasting population data. In *Journal of Physics: Conference Series* (Vol. 1324, No. 1, p. 012100). IOP Publishing.

*Edward Tufte The Visual Display Of Quantitative Information* (2013).

<https://ebookpromotions.online/edward-tufte-the-visual-display-of-quantitative-information.pdf>.

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.

Few, S. (2013). Data Visualization for Human Perception. In: Soegaard, M., & Dam, R. F. (Eds.), *The Encyclopedia of Human-Computer Interaction*, 2nd Ed. Aarhus, Denmark: *The Interaction Design Foundation*.

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>.

Iliinsky, N. (2018) *Designing data visualizations*.

<https://www.oreilly.com/library/view/designing-data-visualizations/9781449314774/ch04.html>.

IvyProSchool (2023) *Power BI Dashboard Live Case Study* [online video] Available at:

<https://www.youtube.com/watch?v=mrTRsoBRWf4> [Accessed 07/22/2023]

Jeremia, N.E., Nurrohmah, S. and Fithriani, I., 2020. Robust Ridge regression to solve a multicollinearity and outlier. In *Journal of Physics: Conference Series* (Vol. 1442, No. 1, p. 012030). IOP Publishing.

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

Lanning, K. (2021) *6 principles of data visualization / Data science for the liberal arts*.

<https://kevinlanning.github.io/DataSciLibArts/principles-of-data-visualization.html>.

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

Li, B. *et al.* (2020) 'Estimation of Regional Economic Development Indicator from Transportation Network Analytics,' *Scientific Reports*, 10(1).

<https://doi.org/10.1038/s41598-020-59505-2>.

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.



Maity, S.K. and Maity, S.K., 2021. Mapping Techniques of Geographical Data. *Essential Graphical Techniques in Geography*, pp.193-305.

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,' *Ekonomika Istraživanja-economic Research*, 35(1), pp. 3243–3275. <https://doi.org/10.1080/1331677x.2021.1989319>.

Obalade, A.A., Ebiwonjumi, A. and Adaramola, A.O., 2019. VAR MODELLING OF DYNAMICS OF POVERTY, UNEMPLOYMENT, LITERACY AND PER CAPITA INCOME IN NIGERIA. *Folia Oeconomica Stetinensia*, 19(1).

Onodugo, V.A., Obi, K.O., Anowor, O.F., Nwonye, N.G. and Ofoegbu, G.N., 2017. Does public spending affect unemployment in an emerging market. *Risk governance & control: Financial markets & institutions*, 7(1), pp.32-40.

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>

- Ren, Y. *et al.* (2019) 'Predicting socio-economic levels of urban regions via offline and online indicators,' *PLOS ONE*, 14(7), p. e0219058.  
<https://doi.org/10.1371/journal.pone.0219058>.
- Sahib, N.S.M.A.J. and Ibrahim, N., 2022. Economic Growth Analysis and Forecasting towards unemployment rate using Arima Model. *Enhanced Knowledge in Sciences and Technology*, 2(1), pp.433-442.
- 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.
- Vafin, A., 2020. Forecasting macroeconomic indicators for seven major economies using the ARIMA model. *Sage Science Economic Reviews*, 3(1), pp.1-16.

- 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,' *Ekonomika 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>.
- Zhang, J., Zhao, Z., Xue, Y., Chen, Z., Ma, X. and Zhou, Q., 2017. Time series analysis. *Handbook of Medical Statistics*, 269.

