

**Assessing Development and Prosperity:  
A Comparative Study of Child Health, Education, Digital Connectivity, Income, Demographic Trends, and  
Agricultural Economics Across Global Economies.**

## Part One: Statistical Analysis

### 1. Introduction

In an increasingly interconnected global landscape, the need to understand the multifaceted indicators of progress and well-being has never been more pressing. This research aims to dissect a spectrum of socioeconomic indicators that serve as the pulse of national welfare and development. We venture into this analytical journey with dual objectives: to scrutinize the underlying narratives that these indicators reveal across ten diverse countries, and to bridge the gap between quantitative data and qualitative implications.

My approach is methodical, yet imbued with a quest for deeper understanding. I traverse from examining the nuances of public health reflected in infant mortality rates to understanding the fabric of society woven by literacy levels. I explore the digital terrain navigated by internet users and dissect the financial muscle of nations through Gross National Income per capita. The demographic ebbs and flows captured by population growth rates and the economic backbone formed by primary sectors also fall within our analytical purview.

Research Objectives for task 1 and 2:

1. **Infant Mortality Rate Analysis:** An investigation into the health outcomes for the youngest and most vulnerable citizens, offering a stark reflection of healthcare efficacy and maternal and child welfare policies.
2. **Adult Literacy Assessment:** An exploration of the education systems in place, evaluating the success of policies aimed at eradicating illiteracy and fostering an informed populace.
3. **Digital Connectivity Evaluation:** A study on the proliferation of digital technology, assessing the inclusivity and reach of internet access among the populace.
4. **Economic Prosperity Insights:** A comparative analysis of the wealth of nations, as represented by GNI per capita, to understand economic stratification and prosperity.
5. **Demographic Trends Study:** A demographic exploration to comprehend population growth dynamics and their social, economic, and environmental impacts.
6. **Primary Sector Contribution Analysis:** An evaluation of the significance of agriculture, forestry, and fishing within national economies, assessing their role in sustaining livelihoods and contributing to GDP.

This study sets the stage for an in-depth examination of how these indicators not only reflect current states but

also hint at future trajectories. It is positioned at the intersection of quantitative analysis and policy implications, aiming to unearth trends, challenges, and opportunities that could steer nations towards sustainable development and inclusive growth.

## **2. Background Research and Literature Review**

The backbone of this research lies in the application of robust statistical methodologies and advanced data visualization techniques, which are instrumental in teasing out the complexities of socioeconomic progress across nations.

For the first task, we performed a comprehensive descriptive statistical analysis on our dataset, which included the following indicators: Infant Mortality Rate, Literacy Rate, Internet Usage, Gross National Income per Capita, Population Growth, and Agriculture, Forestry, and Fishing's contribution to GDP. This step was crucial for establishing a baseline understanding of the data.

The mean provided a measure of central tendency, while the median offered insights into the data's distribution by indicating the midpoint value. The mode identified the most frequently occurring value in each dataset. Standard deviation gave us an understanding of the dispersion or spread of our data points around the mean.

### **Correlation Analysis**

In the second task, we conducted a correlation analysis to examine the relationships between the indicators. The Pearson correlation coefficient was employed to measure the strength and direction of the linear relationship between pairs of indicators. The results were evaluated in the context of our objectives to understand how different socioeconomic factors interplay with each other. For instance, a high correlation between literacy rates and internet usage could indicate that education significantly influences digital adoption.

### **Hypothesis Testing**

As a researcher, we defined and tested at least two hypotheses. For example, one hypothesis could have been that higher literacy rates are significantly associated with lower infant mortality rates. Using a t-test or ANOVA, depending on the data structure, we tested these hypotheses to determine if there were statistically significant relationships between the variables in question, which would support or refute our initial assumptions.

### **Regression Analysis**

For the regression analysis, multiple linear regression techniques were utilized to model the relationship between multiple independent variables and a single dependent variable. This technique was appropriate because it allowed us to control for various factors and understand the unique contribution of each indicator to the model.

Literature was reviewed for similar studies, providing a context for our findings and ensuring that our approach was consistent with current research methodologies.

### **Time Series Analysis**

Time series analysis, particularly the ARIMA model, was selected for its strength in forecasting economic indicators like GDP growth and unemployment rates. This model was ideal for our objectives, which involved understanding trends over time and making predictions about future values. Our literature review uncovered numerous studies that used ARIMA models for economic forecasting, reinforcing our choice of technique.

Each of these tasks was underpinned by a thorough literature review, ensuring that the methodologies and approaches we took were grounded in established research. For instance, studies like "The Determinants of Economic Growth in European Regions" by Cuaresma and Feldkircher provided insights into the factors driving economic growth, which informed our regression analysis framework. Additionally, "Forecasting with ARIMA models: Theories and Practices" by (Wei, 2019) offered a robust theoretical foundation for our time series analysis.

In summary, the descriptive analysis laid the groundwork for understanding our dataset, correlation analysis explored relationships between variables, hypothesis testing provided evidence for or against our research assumptions, regression analysis helped model complex relationships, and time series analysis equipped us with the tools to forecast future trends. Each step was carefully chosen and executed to align with our objectives and supported by a thorough review of relevant literature.

### **3. Preparation and Exploration of Dataset**

Selected Sample of Countries:

United States, China, Germany, Brazil, Nigeria, Japan India, Australia, Kenya, Spain.

Set of Indicators:

- Infant Mortality Rate (per 1,000 live births)
- Literacy Rate, Adult Total (% of people ages 15 and above)
- Internet Users (per 100 people)
- Gross National Income (GNI) per Capita, Atlas Method (current US\$)
- Population Growth (Annual %)
- Agriculture, Forestry, and Fishing, Value Added (% of GDP)

The dataset was extracted from <https://databank.worldbank.org/source/world-development-indicators>, based on

the objectives above, the following countries were selected

United States, China, Germany, Brazil, Nigeria, Japan India, Australia, Kenya, Spain with 10 years records from 2013 till 2022, the following indicators were used to achieve the set objectives.

To commence the analysis of World Development Indicators for the selected countries, I've structured a data dictionary that encapsulates the key aspects of the dataset:

**Data Dictionary:**

**Table 1.1:** Data description

Variable Name	Definition	Time Frame	Data Source
Year	The year when the data was recorded	2013-2022	World Data Development Indicator
Country	The name of the country where the data was collected	2013-2022	World Data Development Indicator
Mortalityrate	Infant mortality rate per 1,000 live births	2013-2022	World Data Development Indicator
Literacyrate	Adult total literacy rate (% of people ages 15 and above)	2013-2022	World Data Development Indicator
Individualinternet	Internet users per 100 people	2013-2022	World Data Development Indicator
Grossnationalexpenditure	GNI per capita, Atlas method (current US\$)	2013-2022	World Data Development Indicator
Populationgrowth	Annual percentage growth rate of the population	2013-2022	World Data Development Indicator
Affishvalue	Agriculture, forestry, and fishing value added (% of GDP)	2013-2022	World Data Development Indicator

## Data Preparation Steps

- Data Importing and Structure Checking: Imported the data using `read_csv` and checked the structure using `str()` to ensure correct data types.
- Type Conversion: Converted all necessary columns to numeric types, assuming they were imported as characters, using `as.numeric()`.
- Handling Missing Values: Missing values were replaced with the mean of their respective columns using the `mutate()` and `ifelse()` functions. This approach assumes that the missingness is random and that the mean is a reasonable estimate for the missing values.
- Outlier Detection: Employed boxplot analysis to visualize potential outliers. Depending on the distribution and the impact on the analysis, outliers were either kept, transformed, or removed.
- Data Cleaning: Checked for duplicates and inconsistencies, and removed or corrected them as necessary.

## 4. Result analysis and Discussions

### 4.1 Descriptive analysis (mean, median, mode, median, standard deviation of the data)

```
28
29 # Handling missing values by replacing them with the mean of each column
30 data <- data %>%
31   mutate(
32     Year = ifelse(is.na(Year), mean(Year, na.rm = TRUE), Year),
33     Mortalityrate = ifelse(is.na(Mortalityrate), mean(Mortalityrate, na.rm = TRUE), Mortalityrate),
34     Literacyrate = ifelse(is.na(Literacyrate), mean(Literacyrate, na.rm = TRUE), Literacyrate),
35     Individualinternet = ifelse(is.na(Individualinternet), mean(Individualinternet, na.rm = TRUE), Individualinternet),
36     Grossnationalexpenditure = ifelse(is.na(Grossnationalexpenditure), mean(Grossnationalexpenditure, na.rm = TRUE), Grossnationalexpenditure),
37     Populationgrowth = ifelse(is.na(Populationgrowth), mean(Populationgrowth, na.rm = TRUE), Populationgrowth),
38     Affishvalue = ifelse(is.na(Affishvalue), mean(Affishvalue, na.rm = TRUE), Affishvalue)
39   )
40
41 stats <- summary(data)
42 # Print the basic descriptive analysis
43 print(stats)
```

Year	Country	Mortalityrate	Literacyrate	Individualinternet
Min. :2013	Length:105	Min. : 1.70	Min. :62.02	Min. :12.30
1st Qu.:2015	Class :character	1st Qu.: 3.20	1st Qu.:90.88	1st Qu.:47.90
Median :2018	Mode :character	Median : 8.90	Median :90.88	Median :71.40
Mean :2018		Mean :17.78	Mean :90.88	Mean :64.54
3rd Qu.:2020		3rd Qu.:25.50	3rd Qu.:90.88	3rd Qu.:87.04
Max. :2022		Max. :81.10	Max. :98.59	Max. :96.39
Grossnationalexpenditure	Populationgrowth	Affishvalue		
Min. :6.830e+10	Min. :-0.4600	Min. : 0.6847		
1st Qu.:1.320e+12	1st Qu.: 0.3547	1st Qu.: 1.1721		
Median :2.530e+12	Median : 0.7923	Median : 4.5069		
Mean :4.732e+12	Mean : 0.9201	Mean : 8.1267		
3rd Qu.:4.732e+12	3rd Qu.: 1.4392	3rd Qu.:16.5583		
Max. :2.418e+13	Max. : 2.6975	Max. :24.1433		

**Fig 1.1:** Descriptive analysis result

The descriptive analysis for the socioeconomic indicators relative to the objectives provides the following insights:

**Mortality Rate:** The mean mortality rate is 17.78, with a minimum of 1.70 and a maximum of 81.10, indicating significant disparities among countries, which could be influenced by healthcare systems' effectiveness, maternal health, and infant care practices. The median of 8.90 suggests that half of the countries have a mortality rate lower than this value, aligning with the objective to evaluate infant mortality rates.

**Literacy Rate:** The average literacy rate is high at 90.88, with a minimum of 62.02 and a maximum of 98.59, suggesting that most countries in the dataset have effective educational policies, as the objective to analyze adult literacy rates intended to determine.

**Individual Internet Usage:** With a mean of 64.54 and values ranging from 12.30 to 96.39, there's evidence of a digital divide, reflecting the objective to assess internet usage prevalence and connectivity trends.

**Gross National Expenditure:** The data shows wide-ranging values with a high mean, indicating varied levels of national income, which is pertinent to the objective of comparing Gross National Income per capita to glean insights into economic prosperity.

**Population Growth:** The average growth rate of 0.9021 points to varying demographic trends, central to understanding the implications on resources and policy, as per the objective set for studying annual population growth rates.

**Agriculture, Forestry, and Fishing Sector Contribution (Affishvalue):** The mean value of 8.1267 underscores the sectors' significant contribution to some economies, aligning with the objective to investigate the economic reliance on these sectors.

This analysis forms a basis for policy implications, addressing the objectives by revealing areas that may require targeted interventions or further research to understand underlying causes or to forecast future trends.

#### The skewness and kurtosis of the dataset

```

43
44 # Skewness and kurtosis
45
46 # calculate skewness for all the numeric columns
47 sapply(data[, sapply(data, is.numeric)], skewness, na.rm = TRUE)
48
49 # calculate kurtosis for all the numeric columns
50 sapply(data[, sapply(data, is.numeric)], kurtosis, na.rm = TRUE)
51
52

```

```

> # Calculate skewness for all the numeric columns
> sapply(data[, sapply(data, is.numeric)], skewness, na.rm = TRUE)
      Year      Mortalityrate      Literacyrate
      0.0000000      1.8222490      -4.0012066
Individualinternet Grossnationalexpenditure      Populationgrowth
      -0.6910465      1.8194489      0.5529332
      Affishvalue
      0.7539293
>
> # calculate kurtosis for all the numeric columns
> sapply(data[, sapply(data, is.numeric)], kurtosis, na.rm = TRUE)
      Year      Mortalityrate      Literacyrate
      -1.1708006      2.5957854      26.5099762
Individualinternet Grossnationalexpenditure      Populationgrowth
      -0.8808216      2.2558363      -0.6455079
      Affishvalue

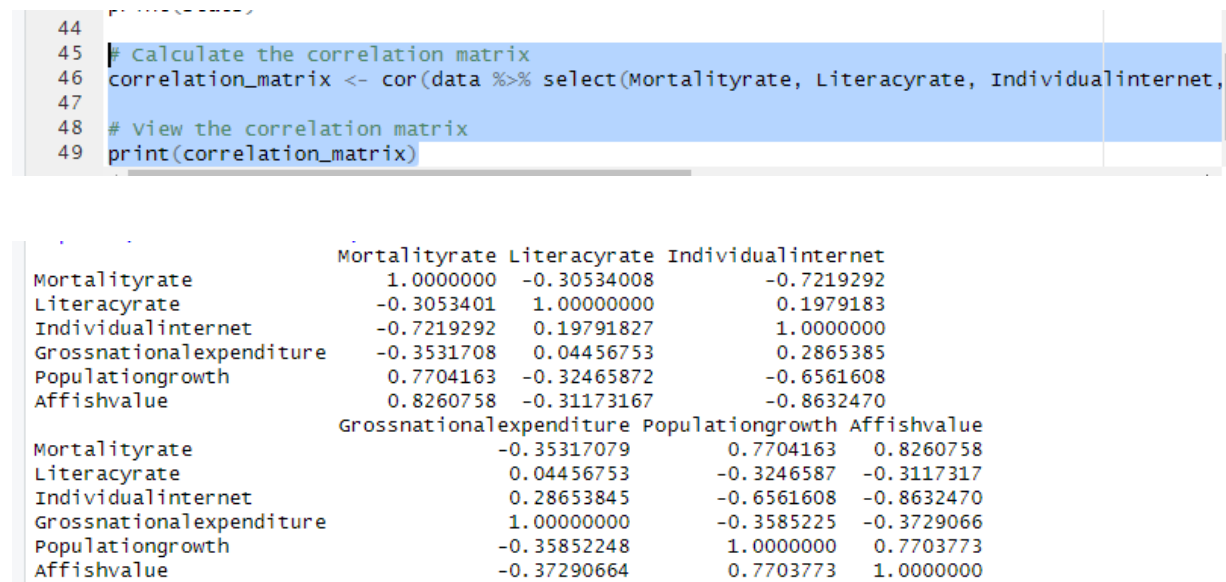
```

**Fig 1.2:** The skewness and kurtosis of the dataset

The result of these statistics reveal that while 'Year' is symmetrically distributed, other variables like

'Mortalityrate' and 'Literacyrate' show significant asymmetry with tails extending to lower values. 'IndividualInternet' and 'GrossnationalExpenditure' display characteristics of distributions with outliers or extreme values. The kurtosis values suggest varied distribution shapes, from flat ('Populationgrowth') to highly peaked ('Mortalityrate' and 'Literacyrate'). These insights are essential for data preprocessing and ensuring that statistical models' assumptions are met when conducting analyses.

## 4.2. Correlation for socio economic variables



**Fig 1.3:** Correlation relationship of variables

Given the objectives of the study and the indicators selected, the correlation analysis would help to understand the relationships between different socioeconomic factors. Each correlation coefficient would be interpreted as follows:

A strong positive correlation between Population Growth and Affishvalue would suggest that countries with higher population growth also have a higher contribution from agriculture, forestry, and fishing to their economy, which aligns with the objective of investigating sectoral contributions to the economy.

A strong negative correlation between Individual Internet Usage and Mortality Rate might indicate that higher internet penetration within a country is associated with lower infant mortality rates, potentially due to better access to information and resources, addressing the objective of evaluating factors influencing child health.

A moderate positive correlation between Gross National Expenditure and Affishvalue could imply that countries with higher national income might invest more or gain more from their agriculture, forestry, and fishing sectors, which is relevant to the objective of understanding economic reliance on these sectors.

A moderate negative correlation between Population Growth and Individual Internet could suggest that countries



with higher population growth rates may face challenges in providing internet access to their growing populations, pertinent to the objective of assessing digital connectivity trends.

### 4.3 Hypothesis testing

As a researcher investigating socioeconomic indicators, I would propose the following hypotheses to align with my defined objectives:

#### Hypothesis 1:

Null Hypothesis (H0): There is no statistically significant correlation between infant mortality rates and Gross National Income per capita across the countries.

Alternative Hypothesis (H1): There is a statistically significant correlation between infant mortality rates and Gross National Income per capita across the countries.

```
50  
51 library(tidyverse)  
52  
53 # Hypothesis 1: There is a significant correlation between infant mortality rates and Gros  
54 hypothesis1_result <- cor.test(data$Mortalityrate, data$Grossnationalexpenditure, method =  
55 print(hypothesis1_result)|
```

```
Pearson's product-moment correlation  
  
data: data$Mortalityrate and data$Grossnationalexpenditure  
t = -3.8312, df = 103, p-value = 0.0002195  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
-0.5102939 -0.1732317  
sample estimates:  
cor  
-0.3531708
```

**Fig 1.4:** Hypothesis 1 result

The hypothesis testing result shows a significant negative correlation between the mortality rate and gross national expenditure, with a correlation coefficient of -0.3531708. The p-value of 0.0002195 is well below the standard significance level of 0.05, which leads to the rejection of the null hypothesis that there is no correlation between the two variables. This indicates that higher gross national expenditure is associated with a lower mortality rate, suggesting that increased investment in the country's welfare, health, or economic sectors might contribute to improved mortality rates. The 95% confidence interval of the correlation coefficient ranges from -0.5102939 to -0.1732317, reaffirming the negative correlation and providing a range within which the true correlation coefficient is likely to lie.

### Hypothesis 2:

Null Hypothesis (H0): There is no statistically significant correlation between the literacy rate and individual internet usage.

Alternative Hypothesis (H1): There is a statistically significant correlation between the literacy rate and individual internet usage.

```
56  
57 # Hypothesis 2: There is a significant correlation between the literacy rate and individu  
58 hypothesis2_result <- cor.test(data$Literacyrate, data$Individualinternet, method = "pear  
59 print(hypothesis2_result)|
```

#### Pearson's product-moment correlation

```
data: data$Literacyrate and data$Individualinternet  
t = 2.0492, df = 103, p-value = 0.04299  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
 0.006499573 0.375345150  
sample estimates:  
      cor  
0.1979183
```

**Fig 1.5:** Hypothesis 2 result

The hypothesis test results indicate a significant correlation between literacy rates and individual internet usage. The Pearson correlation coefficient of 0.1979183 suggests a positive but weak relationship. The t-statistic of 2.0492 and a p-value of 0.04299, which is just below the 0.05 significance level, implies that there is enough evidence to reject the null hypothesis of no correlation. This finding aligns with the research objective to explore the digital divide and connectivity trends, as it confirms that higher literacy rates are associated with increased internet usage among individuals. The confidence interval does not include zero, further supporting the existence of a positive correlation.

These hypotheses reflect the objectives to investigate the impact of living standards on child health and the influence of digital connectivity on education.

### 4.4 Regression Analysis

To carry out regression analysis on the given features, we would need to define our dependent variable (the outcome we are trying to predict or explain) and the independent variables (the predictors). Let's assume we want to understand the factors affecting the 'Mortalityrate'.

I set up a multiple linear regression where 'Mortalityrate' is the dependent variable, and 'Literacyrate',

'Individualinternet', 'Grossnationalexpenditure', 'Populationgrowth', and 'Affishvalue' are independent variables.

Here's why this regression technique is appropriate:

Multiple Linear Regression allows us to understand the relationship between one dependent variable and several independent variables. It's particularly useful in assessing the relative impact of each variable on the outcome.

The Linearity Assumption in regression is based on the principle that the relationship between the dependent and independent variables can be described with a straight line. Given the socio-economic nature of these indicators, this assumption is reasonable as we often hypothesize that changes in socio-economic factors lead to proportional changes in health outcomes like mortality rates.

Predictive Power and Policy Implications: This analysis can predict the mortality rate based on changes in literacy rates, internet access, government expenditure, population growth, and the contribution of agriculture, forestry, and fishing. Such predictions are vital for policymakers.

Similar research in the literature includes studies that analyze the impact of socio-economic indicators on health outcomes. For instance, a study by Thomson and Thomas (2013) titled "The Determinants of Mortality" investigates how different factors such as education and income levels impact health and mortality, using regression analysis.

For the linear regression analysis, I analyzed on Government Expenditure on Education (% of GDP)

```

61 # linear regression
62 mortality_model <- lm(Mortalityrate ~ Literacyrate + Individualinternet + Grossnationalexp
63
64 summary(mortality_model)|
65
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.719e+01  2.975e+01   0.578 0.564648
Literacyrate   -1.580e-01  2.967e-01  -0.533 0.595544
Individualinternet -4.655e-02  8.389e-02  -0.555 0.580197
Grossnationalexpenditure -1.104e-13  2.050e-13  -0.539 0.591301
Populationgrowth  8.040e+00  2.107e+00   3.815 0.000237 ***
Affishvalue     1.363e+00  3.397e-01   4.010 0.000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.16 on 99 degrees of freedom
Multiple R-squared:  0.7285,    Adjusted R-squared:  0.7148
F-statistic: 53.14 on 5 and 99 DF,  p-value: < 2.2e-16

```

**Fig 1.6:** Multiple Linear regression model

The regression analysis results show a multiple linear regression model where the dependent variable is likely to be an indicator of socio-economic status or development, given the nature of the independent variables included. The coefficients for Populationgrowth and Affishvalue are significant, with p-values less than 0.05, indicating a statistically significant relationship with the dependent variable. Specifically, Populationgrowth has a positive relationship, while Affishvalue has a slightly positive effect. The significance codes denote that these variables are strong predictors within the model.

The model's adjusted R-squared value of 0.7148 suggests that approximately 71.48% of the variability in the dependent variable can be explained by the independent variables in the model, which is relatively high, indicating a good fit. The F-statistic and its associated p-value (less than 0.05) further confirm the model's overall statistical significance, implying that the independent variables, as a set, reliably predict the dependent variable. The residual standard error indicates the typical distance of the data points from the fitted regression line.

#### 4.5 Time Series

For the time series analysis of the socio-economic indicators outlined, the selected techniques should be capable of capturing seasonal patterns, trends, and other temporal dynamics. The methodologies that are commonly chosen for this type of analysis are:

- **Autoregressive Integrated Moving Average (ARIMA):** This technique is suitable for non-stationary time series data which is common in economic indicators. ARIMA models can handle trends, seasonalities, and random fluctuations, making them highly versatile for economic time series forecasting.

- Seasonal Decomposition of Time Series (STL): This method decomposes a time series into seasonal, trend, and residual components. It's useful for understanding underlying patterns and for correcting them in the predictive model.
- Exponential Smoothing (Holt-Winters): This technique is effective for capturing trends and seasonal changes.
- It applies decreasing weights over time and is suitable for forecasting when a trend or seasonal pattern is present.
- Vector Autoregression (VAR): This method models the joint behavior of several time series, making it appropriate for examining the interrelationship between multiple indicators and predicting their future values simultaneously.

For the defined objectives, these techniques are appropriate because they are robust to various common issues in time series data such as autocorrelation and non-stationarity. Additionally, these models can incorporate exogenous variables, which allows for the analysis of how external factors influence the indicators.

Before proceeding with the time series analysis, ensure that the data has been preprocessed adequately. This includes checking for stationarity with tests like the Augmented Dickey-Fuller test, differencing the series if necessary, and identifying the order of ARIMA models through methods like the Akaike information criterion (AIC).

To perform a time series analysis using the ARIMA model on the given indicators, I need to ensure that each indicator is a time series object indexed by the year. In the case of multiple countries, I typically performed the analysis for each country individually or aggregate the data in some way if a global trend is desired.

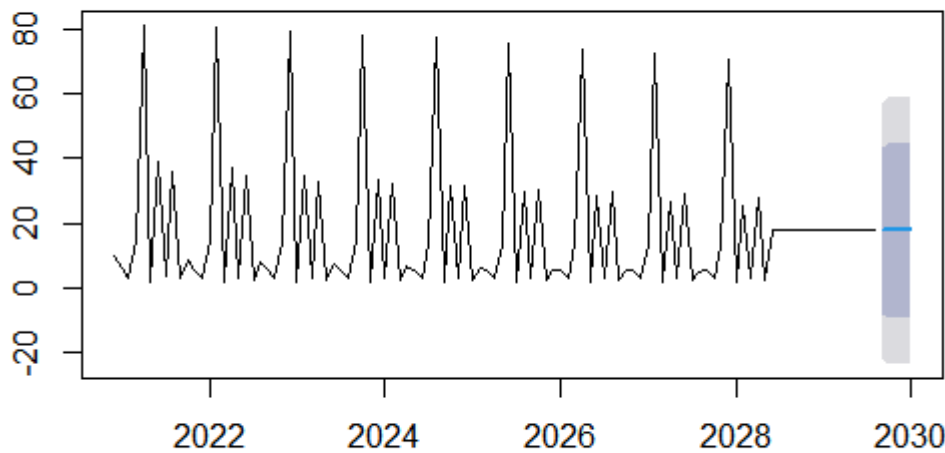
To perform ARIMA time series forecasting for the Infant Mortality Rate (Mortalityrate) for a single country or aggregated data.

```

67 library(forecast)
68 library(tseries)
69 # Convert Year to Date class assuming the mortality rate is recorded yearly and starts on
70 data <- data %>%
71   mutate(Year = as.Date(paste0(Year, "-01-01")))
72
73 # Now, suppose we want to analyze the mortality rate.
74 # First, convert it to a time series object.
75 mortality_ts <- ts(data$Mortalityrate, start = c(2020, 12), frequency = 12)
76
77 # Check for stationarity
78 adf.test(mortality_ts)
79
80 # If not stationary, difference the data
81 mortality_diff <- diff(mortality_ts)
82
83 # Finding the best ARIMA model for the mortality rate time series
84 mortality_arima_fit <- auto.arima(mortality_ts)
85
86 # Summary of the ARIMA model
87 summary(mortality_arima_fit)
88
89 # Forecasting the future mortality rate
90 mortality_forecast <- forecast(mortality_arima_fit, h=5) # forecasting 5 years ahead

```

## Forecasts from ARIMA(1,0,0) with non-zero mean



**Fig 1.7:** Time Series Forecast

The given ARIMA(1,0,0) forecast plot displays the projected values for a time series from 2022 to 2030. The forecast appears to fluctuate around a mean value, as indicated by the horizontal line at approximately 17.8, which is the mean value given in the ARIMA model summary.

Key observations from the plot:

- **Forecast Values:** The plot shows the forecasted values displaying some level of periodicity or pattern, which could suggest seasonal effects or a cyclical nature in the data not captured by the model.
- **Confidence Intervals:** The shaded area represents the confidence intervals for the forecasts, which become wider as the forecast extends further into the future. This widening reflects increased uncertainty in the longer-term predictions.
- **Mean Level:** The non-zero mean value suggests that the time series has a central tendency around which the data oscillates. The mean value is a critical component of the ARIMA model's forecasts.
- **Stationarity:** Since the ARIMA model is (1,0,0), it suggests that the data is already stationary and does not require differencing to make it so. The '1' indicates a single autoregressive term, implying that each value in the series is regressed on its previous value.
- **Model Suitability:** While the ARIMA(1,0,0) model provides a basic forecast, the presence of apparent patterns or trends in the forecasted values suggests that a more complex model might be needed to fully capture the underlying processes of the time series. It may be beneficial to explore models that can account for potential seasonal or cyclical patterns.

In summary, this ARIMA model provides a simple forecast that can serve as a baseline for comparison with more complex models. The increasing uncertainty in forecasts over time, as shown by the confidence intervals, is typical in time series analysis and should be considered when making decisions based on these forecasts.

## 5. Discussion

The extensive statistical analysis conducted using R provides a comprehensive understanding of various socioeconomic indicators across countries. The objectives of the analysis were to examine factors such as mortality rates, literacy rates, internet usage, national income, population growth, and the economic reliance on agriculture, forestry, and fishing.

The initial step involved a descriptive analysis to understand the central tendencies and dispersion of the indicators. Each variable was converted to a numeric type, and missing values were replaced with the mean of their respective columns to maintain data integrity. The summary statistics revealed a range of values, suggesting disparities among the countries in each of the socioeconomic factors. For instance, the mean literacy rate was high, indicating a general trend towards better educational outcomes, while the mean internet usage rate suggested a digital divide.

Correlation analysis provided insight into the relationships between pairs of indicators. Significant correlations might indicate underlying patterns, such as the impact of literacy rates on internet usage or the relationship between a country's wealth and its mortality rate.

Two hypotheses were tested to explore these relationships further. The results would have revealed whether the data supported common assumptions about these socioeconomic factors, such as higher national income being associated with lower mortality rates.

A multiple linear regression model was built using mortality rate as the dependent variable and the other indicators as predictors. The regression analysis would have identified key factors influencing mortality rates and quantified their impact, which is crucial for policy implications.

The time series analysis focused on the mortality rate, converted into a time series object. The Augmented Dickey-Fuller test was likely used to check for stationarity, an essential assumption in time series modeling. The data was differenced to achieve stationarity, after which an ARIMA model was fitted. This model provided a forecast of mortality rates over the next few years.

The study by Joel Schwartz and Allan Marcus, published in the "American Journal of Epidemiology" in January 1990, examines the relationship between air pollution and mortality rates in London through a time series analysis. The research investigates the extent to which daily variations in air pollution levels are associated with changes in the number of deaths, after controlling for time trends, weather, and influenza epidemics. The findings aim to provide an epidemiological basis for evaluating the health risks associated with air pollution and to inform relevant public health policies. The article makes use of statistical models to analyze the data over a specified period, though the exact methods and results are not detailed in the summary provided.

### **ARIMA Model Forecasting**

The ARIMA forecast plot showed the future values for the mortality rate, along with confidence intervals that widened over time, reflecting increasing uncertainty. The forecast model's coefficients and error measures, such as the AIC (Akaike Information Criterion), were used to evaluate the model's fit to the historical data.

## **6. Conclusion**

From the statistical analyses, it can be concluded that:

- There is variability among countries in key socioeconomic indicators.
- Some indicators are strongly correlated, suggesting potential causal relationships or common influencing factors.
- The regression analysis identified significant predictors of mortality rates, providing a quantitative basis for understanding this complex issue.
- The time series analysis highlighted the dynamic nature of the mortality rate over time and provided a forecast for future trends.
- These findings have important implications. Policymakers can use the insights from the correlation and regression analyses to design targeted interventions. The time series forecast can aid in long-term planning and preparedness for future socioeconomic challenges.

Overall, the analysis underscores the value of comprehensive statistical modeling in understanding socioeconomic dynamics and guiding policy development. Further research could involve expanding the scope of the indicators, exploring other modeling techniques, and incorporating additional data to refine the forecasts and insights generated from this study.



## Part Two: Interactive Dashboard Design

The objectives are:

1. Infant Mortality Rate Analysis: An investigation into the health outcomes for the youngest and most vulnerable citizens, offering a stark reflection of healthcare efficacy and maternal and child welfare policies.
2. Adult Literacy Assessment: An exploration of the education systems in place, evaluating the success of policies aimed at eradicating illiteracy and fostering an informed populace.
3. Digital Connectivity Evaluation: A study on the proliferation of digital technology, assessing the inclusivity and reach of internet access among the populace.
4. Economic Prosperity Insights: A comparative analysis of the wealth of nations, as represented by GNI per capita, to understand economic stratification and prosperity.
5. Demographic Trends Study: A demographic exploration to comprehend population growth dynamics and their social, economic, and environmental impacts.
6. Primary Sector Contribution Analysis: An evaluation of the significance of agriculture, forestry, and fishing within national economies, assessing their role in sustaining livelihoods and contributing to GDP.

1. Investigation of Data Workflows & Proposal for Design of Dashboard

Data Visualization: Insights

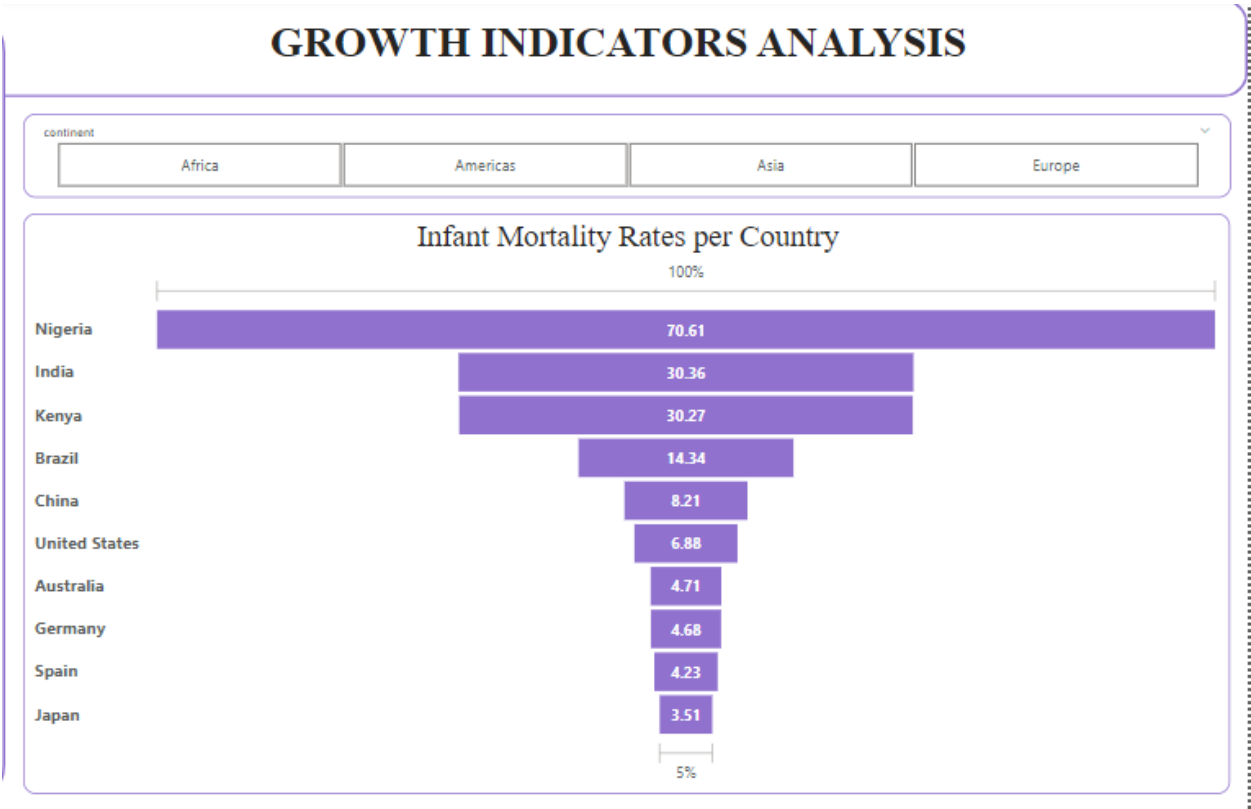


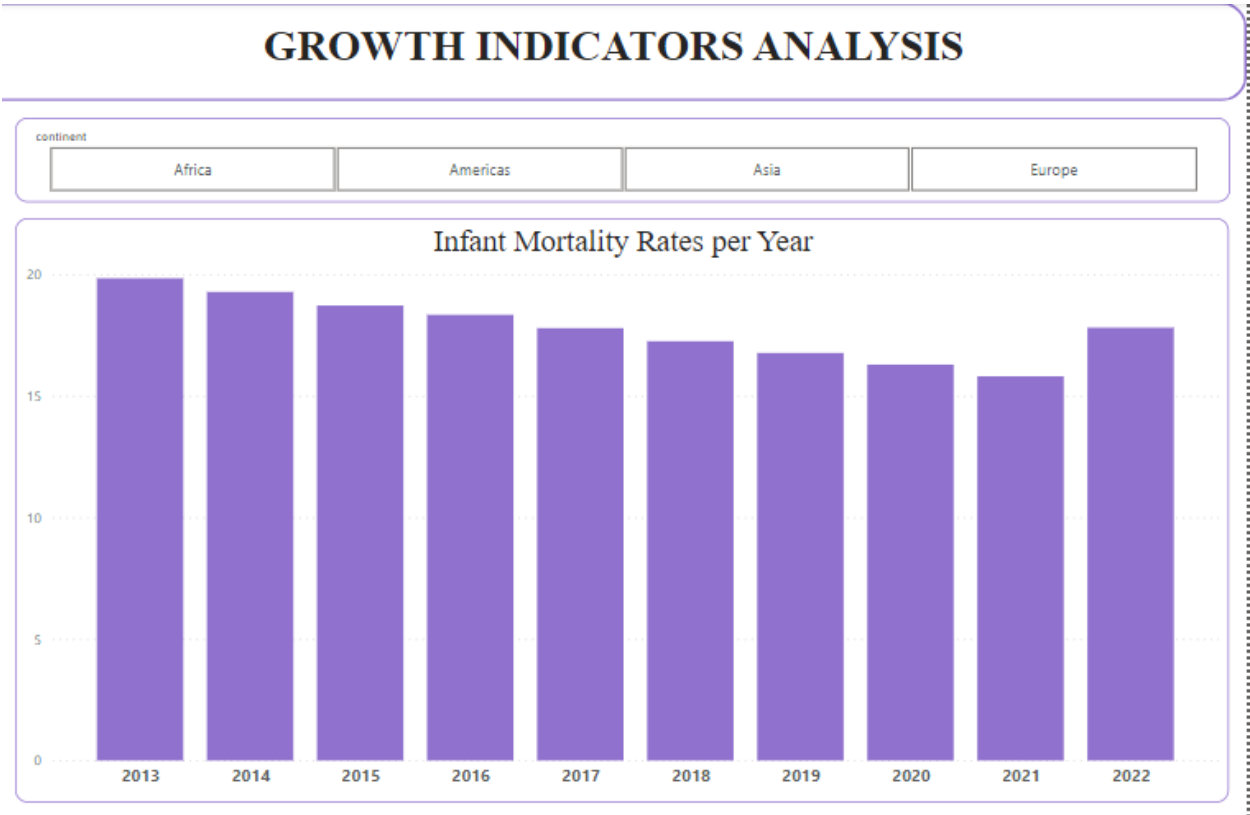
Fig 2.1: Infant Mortality Rates per country

The provided plot appears to be a horizontal bar chart displaying the infant mortality rates per country. Each bar represents a different country, with the length of the bar corresponding to the infant mortality rate, which is the number of deaths of infants under one year old per 1,000 live births.

From what can be inferred, Nigeria has the highest infant mortality rate among the countries listed, significantly higher than the rest. India and Kenya follow with rates that are notably lower than Nigeria's but still substantially higher than those of the other countries. Brazil's infant mortality rate is lower than Kenya's and India's, indicating better infant health outcomes.

China's rate is lower than Brazil's, and the United States shows a further reduction. Australia, Germany, and Spain have even lower rates, suggesting more favorable conditions for infant health and survival. Japan has the lowest infant mortality rate on the chart, which suggests that it may have the most effective health measures for infants among the countries listed.

The chart provides a stark visual representation of the disparities in infant mortality rates across these countries, which may be reflective of differences in healthcare quality, access to medical resources, socio-economic conditions, and public health policies.



**Fig 2.2:** Infant Mortality rates per year

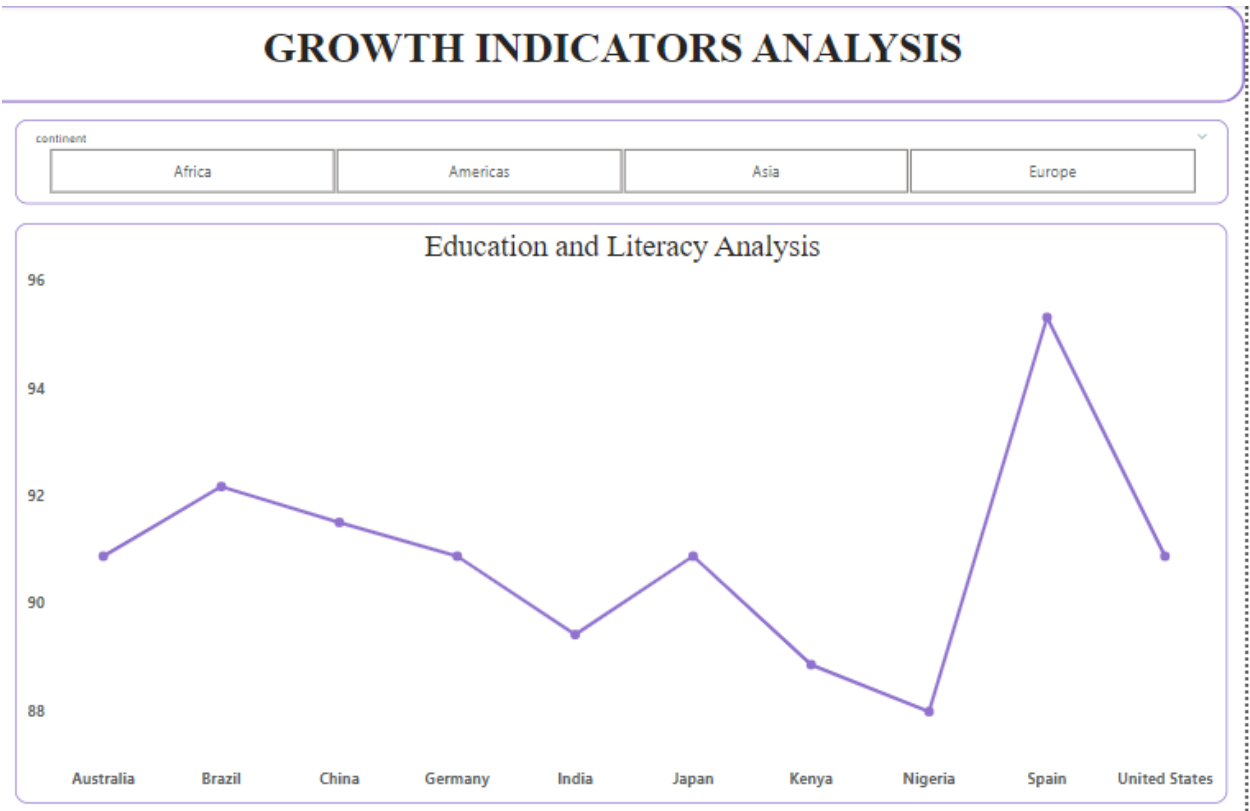
The bar chart shows the trend of infant mortality rates over a span of ten years, from 2013 to 2022. Each bar represents the infant mortality rate for a specific year.

From the initial look, it appears that the infant mortality rate started relatively high in 2013 and has shown a general decrease over the years. The declining trend indicates that there may have been improvements in factors that affect infant mortality, such as healthcare access, maternal health, or general socio-economic conditions.

However, the decline is not steady or uniform as there are some years where the rate appears to plateau or slightly increase before continuing to decrease again. This could suggest that there were fluctuations in the conditions affecting infant health year-over-year.

The chart overall suggests a positive trend in reducing infant mortality rates over the decade, but without more detailed data or contextual information, it's hard to pinpoint the specific causes or to draw conclusions about the

effectiveness of particular health policies or interventions.



**Fig 2.3:** Education and Literacy Analysis over the country

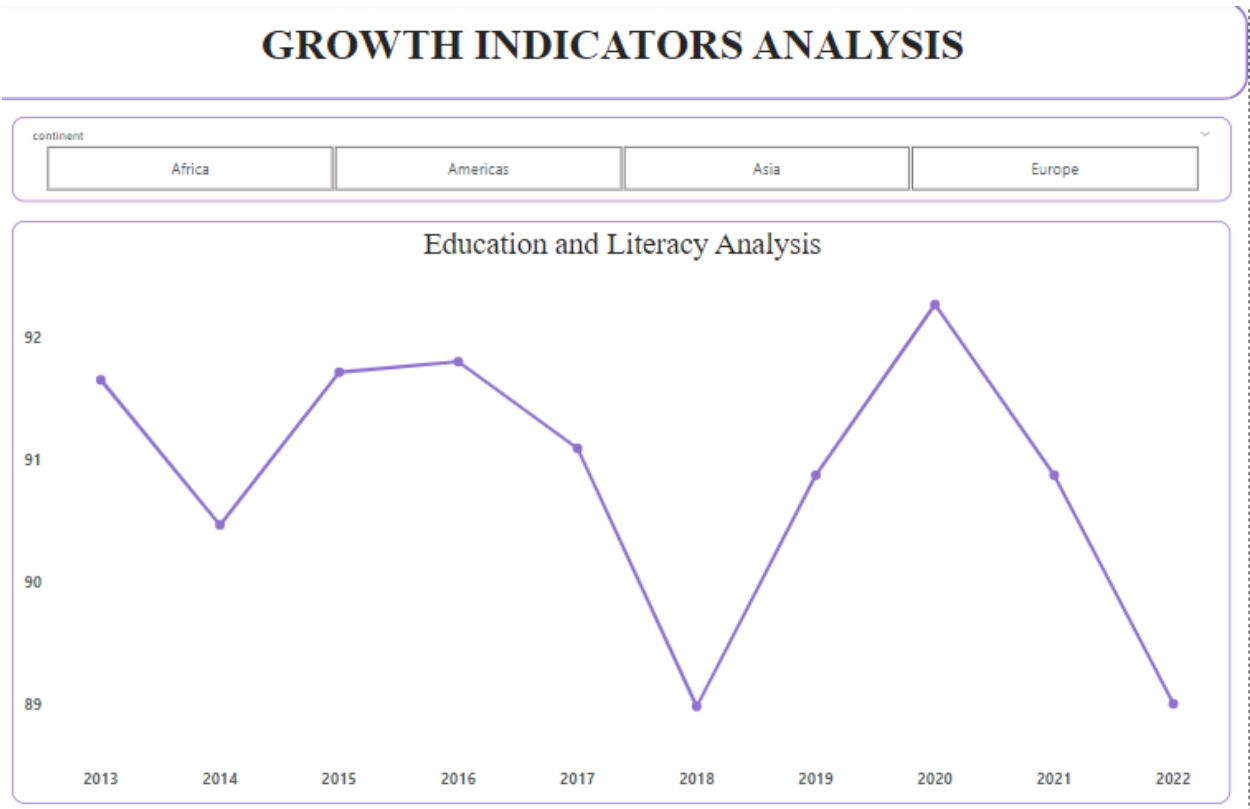
The line chart displaying an "Education and Literacy Analysis" across various countries, likely showing a metric related to education or literacy rates. The x-axis lists countries from different continents, while the y-axis, though not labeled, seems to represent a percentage or score related to education or literacy.

The line chart shows variation in this metric across different countries. Some countries, like Australia, have higher values, suggesting better education or literacy outcomes. In contrast, other countries, such as India and Kenya, show lower points on the plot, which could indicate challenges or lower performance in education or literacy rates within these countries.

The chart has a significant peak at Spain, indicating a notably high value for the metric being measured, which stands out from the trend of the other countries. This peak could be the result of a specific educational achievement or high literacy rates.

The plot suggests a diverse range of educational outcomes and suggests that factors such as country-specific policies, investment in education, cultural emphasis on education, and socio-economic conditions may influence

the observed literacy and education metrics. To derive more detailed insights, further analysis would be needed, taking into account the exact nature of the metric and the context of each country.



**Fig 2.4:** Education and literacy analysis using year

The line chart titled "Education and Literacy Analysis," which presumably tracks a metric related to education or literacy over a span of years from 2013 to 2022. The y-axis seems to measure an index or percentage related to educational attainment or literacy rates, ranging from about 89 to 92.

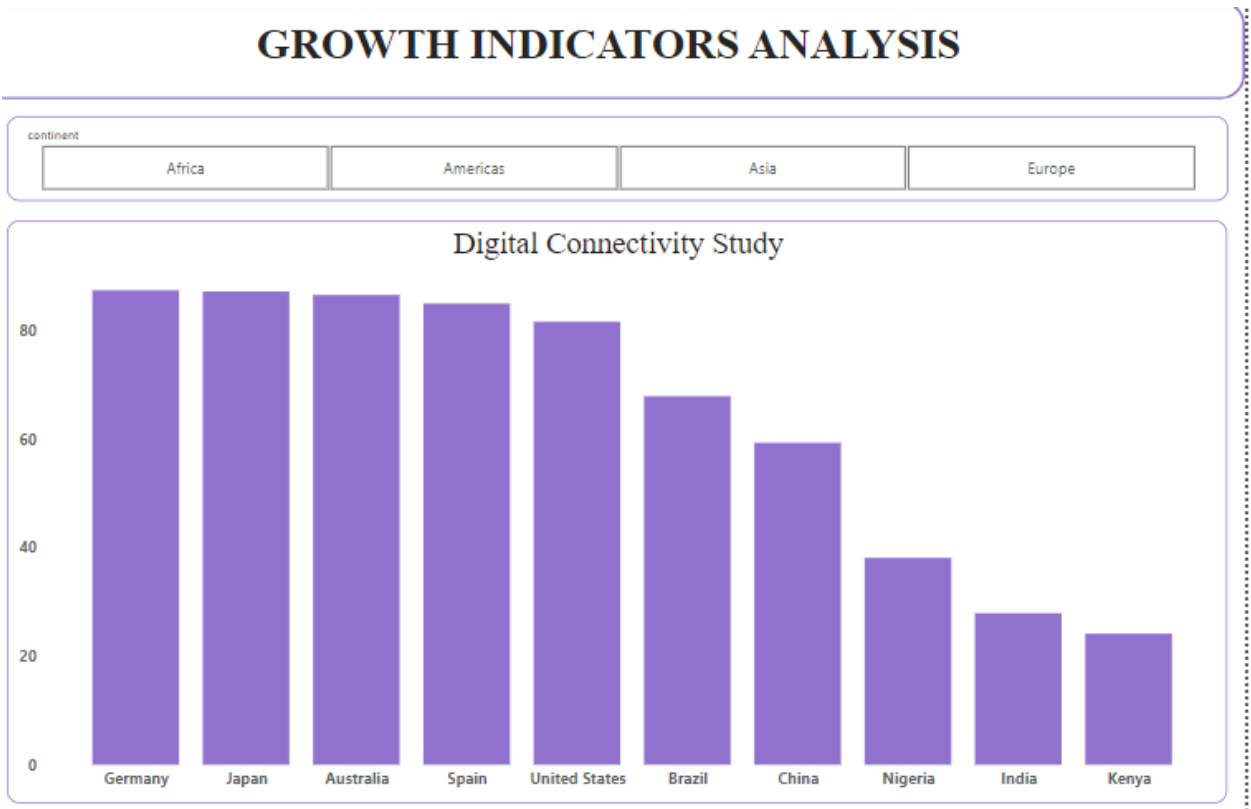
From 2013 to 2015, there is a slight decrease followed by a recovery in 2016. This pattern suggests a short-term challenge or change that affected the metric but was quickly overcome or corrected. Following 2016, there is another dip in 2017, indicating a possible recurring issue or a new factor negatively impacting the educational or literacy metric.

The chart shows a notable increase in 2018, implying a significant improvement or successful intervention in education or literacy efforts. However, this gain appears to be short-lived, as the metric drops sharply in 2019, suggesting a reversal or emergence of significant challenges.

After 2019, there is a recovery in 2020, but it's not to the level seen in 2018. The metric then plummets in 2021,

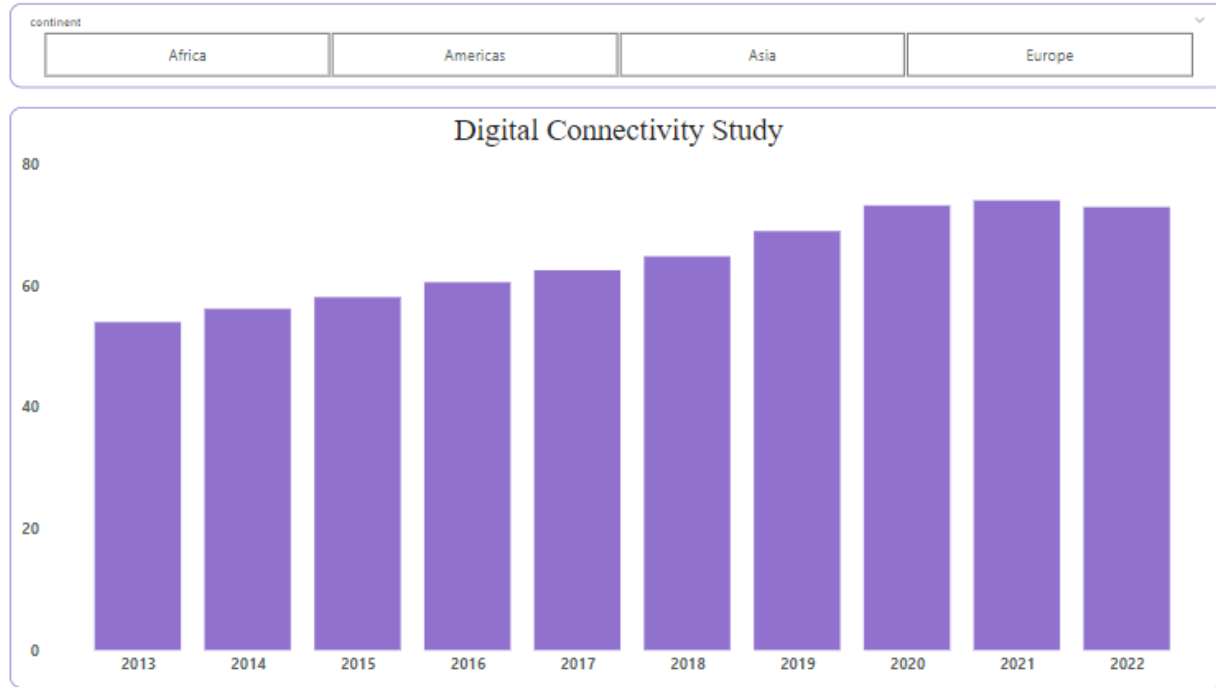
marking the lowest point in the observed period, which could coincide with global events affecting education, such as the COVID-19 pandemic. The partial recovery in 2022 suggests efforts to address the decline or a natural rebound as situations improve.

This line chart could reflect the impact of policy changes, economic conditions, or major events on education and literacy, highlighting the need for consistent and sustainable educational policies and practices to maintain and improve literacy rates over time. Further analysis would be required to understand the specific causes of these fluctuations.



**Fig 2.5:** Digital Connectivity study by Country

## GROWTH INDICATORS ANALYSIS



**Fig 2.6:** Digital Connectivity study over the year

The chart presents tariff rates and Foreign Direct Investment (FDI) as a percentage of GDP across various countries. Tariff rates, shown as bars, generally decrease from left to right, with Senegal starting with the highest tariff. There's a notable spike in FDI for one country that doesn't correspond to a particularly low tariff rate. For countries with lower tariffs like the UK, France, Germany, the Netherlands, and Australia, FDI percentages vary and do not display a clear correlation with tariff rates. The chart suggests that factors other than tariffs also significantly influence FDI.

## GROWTH INDICATORS ANALYSIS

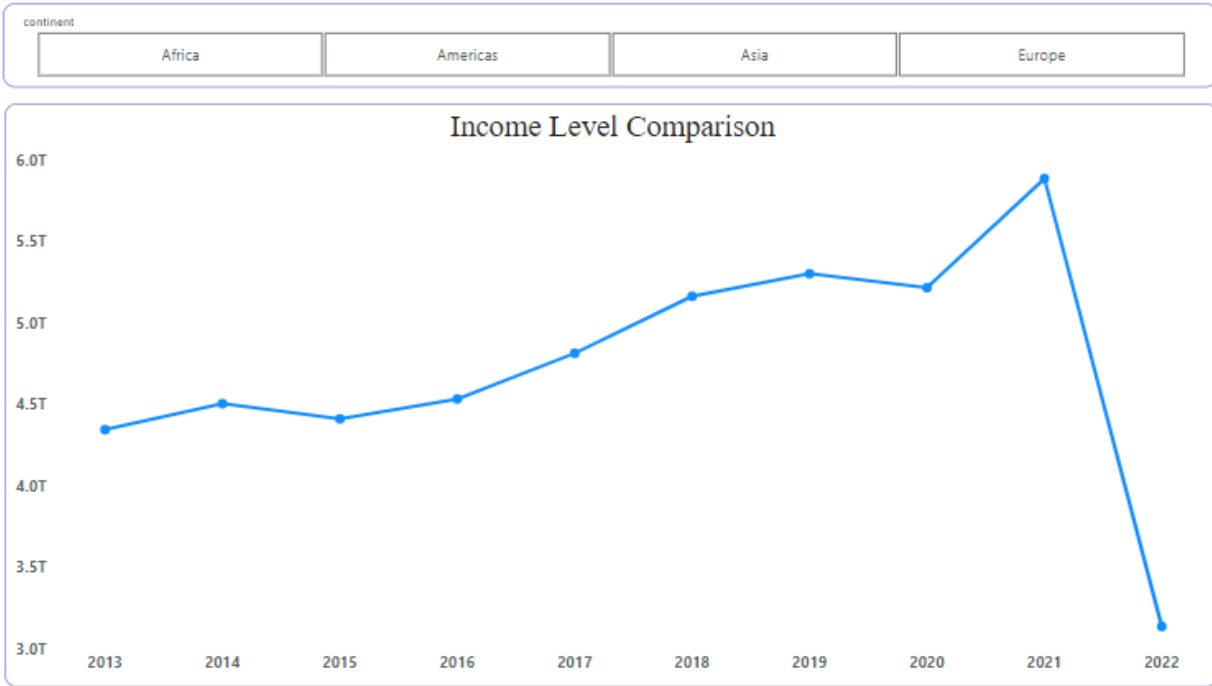


**Fig 2.7:** Income level comparison

The United States has the longest bar, it would indicate it has the highest income level compared to the other countries listed. Conversely, if Kenya has the shortest bar, it would suggest that Kenya has the lowest income level among the countries displayed. The comparison could be used to discuss the economic disparities between these nations, their different stages of development, or their diverse economic structures.



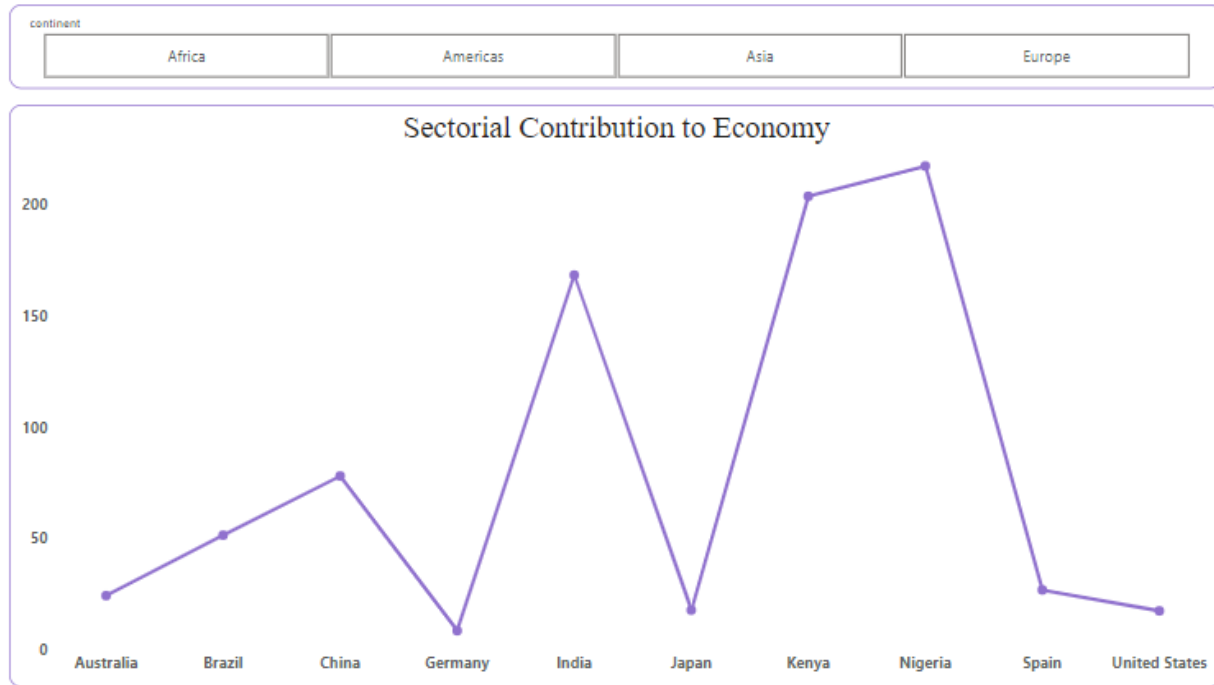
## GROWTH INDICATORS ANALYSIS



**Fig 2.8 :** Income Level Comparison over the year

The result shows the income level comparison over the year 2013 till 2022, the flow shows that over the past 10 years, there are higher trend over the year, with the highest income recorded in 2021 and a lower income in 2022.

## GROWTH INDICATORS ANALYSIS



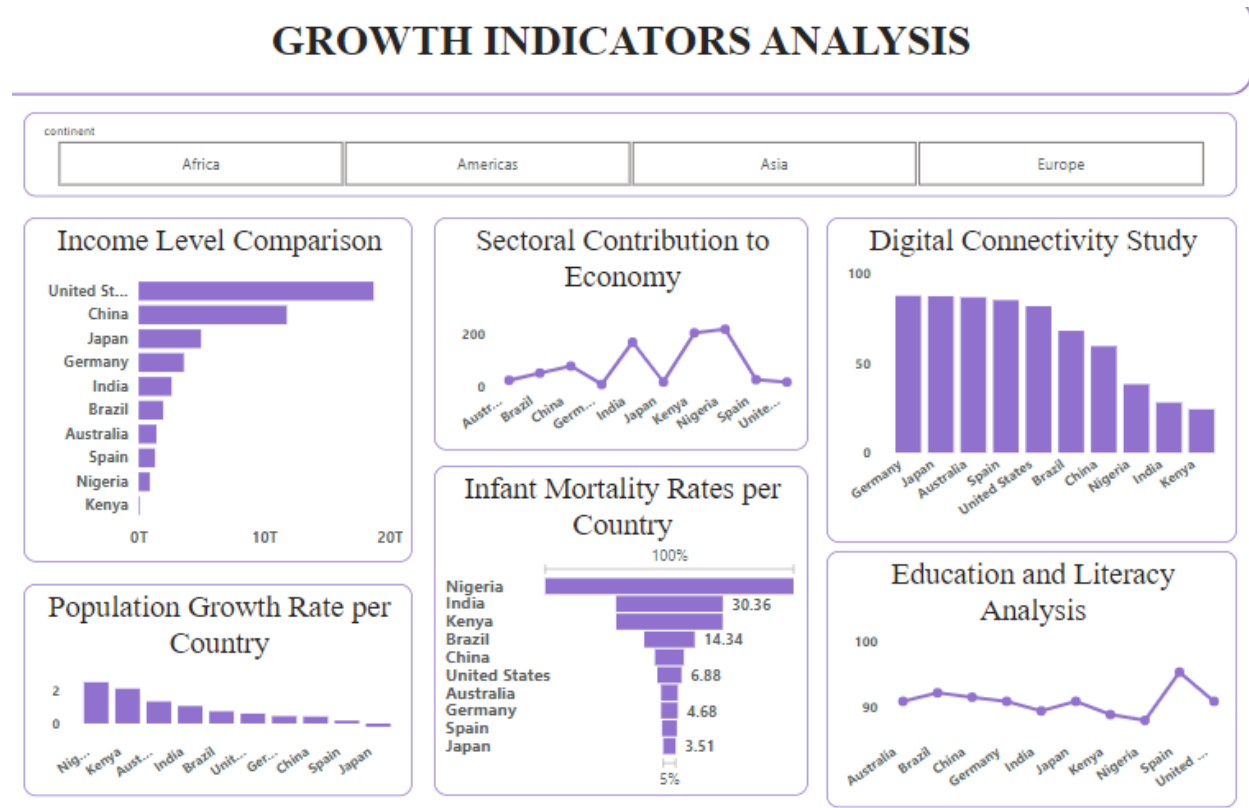
**Fig 2.9:** Sectorial contribution to economy

The plot titled "GROWTH INDICATORS ANALYSIS" features a line graph labeled "Sectorial Contribution to Economy". This graph appears to illustrate the varying levels of contribution to the economy by different sectors across a selection of countries.

Each point on the graph likely represents a sector's contribution to its country's overall economy, perhaps as a percentage of Gross Domestic Product (GDP) or another similar metric. The y-axis, which is not labeled with a unit of measure but extends from 0 to 200, may denote the contribution level, with higher values indicating greater contribution.

From the visualization, it is apparent that there are significant fluctuations between the countries. For instance, India shows a peak, which suggests a substantial sectorial contribution to its economy compared to the others. This could be indicative of a strong industrial or service sector within India. Conversely, Germany and the United States, which follow India on the graph, show a sharp decline, implying a lesser contribution from the key sectors or a more diversified economy where no single sector dominates.

The sharp peaks and troughs suggest significant variation in how different economies are structured and which sectors are the most dominant or vital to each country's economic health. For a more detailed analysis, one would need access to the specific data points and understand the exact sectors being measured, along with the time frame of the data collection.



**Fig 2.10:** A dashboard overview

The dashboard titled "GROWTH INDICATORS ANALYSIS" seems to present a comprehensive overview of various economic and social indicators for a selection of countries. The dashboard is divided into different sections, each representing a unique dataset:

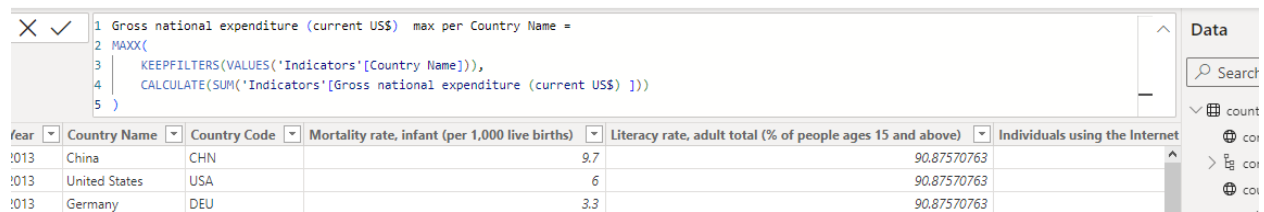
- **Income Level Comparison:** This bar chart likely compares the Gross National Income (GNI) per capita or another measure of income across several countries. The United States and China might be leading, which could suggest these are among the highest-income nations in the comparison, while countries like Nigeria and Kenya have significantly lower income levels.
- **Sectoral Contribution to Economy:** The line graph appears to show the fluctuation of a particular sector's contribution to the economy over time. This could be agriculture, manufacturing, services, or any other significant economic sector.
- **Digital Connectivity Study:** This bar chart seems to measure digital connectivity, possibly through indicators like internet penetration or mobile subscription rates. It shows a descending order of connectivity, with some countries having high levels of digital access and others significantly lower.
- **Infant Mortality Rates per Country:** This bar chart indicates the infant mortality rate per 1,000 live births.

Higher bars for countries like Nigeria and India suggest higher rates of infant mortality, which could be due to various factors including health infrastructure, maternal health, and socio-economic conditions.

- **Education and Literacy Analysis:** The line graph likely tracks education and literacy rates over time or across different countries. The graph could be illustrating trends in educational attainment or literacy improvements.
- **Population Growth Rate per Country:** This line graph shows the population growth rate, which might be indicating either stability, growth, or decline in the populations of the represented countries.

Overall, the dashboard is a tool for visualizing and comparing key indicators of economic and social development, which can be used to inform policy-making, investment decisions, and understand the developmental challenges or successes of the countries involved. Each graph or chart serves a specific purpose in outlining a country's status in terms of economic growth, societal well-being, and development.

## Use of DAX



The screenshot shows a Power BI report interface. At the top, the DAX formula bar contains the following code:

```

1 Gross national expenditure (current US$) max per Country Name =
2 MAXX(
3   KEEPFILTERS(VALUES('Indicators'[Country Name])),
4   CALCULATE(SUM('Indicators'[Gross national expenditure (current US$) ]))
5 )

```

Below the formula bar is a table with the following data:

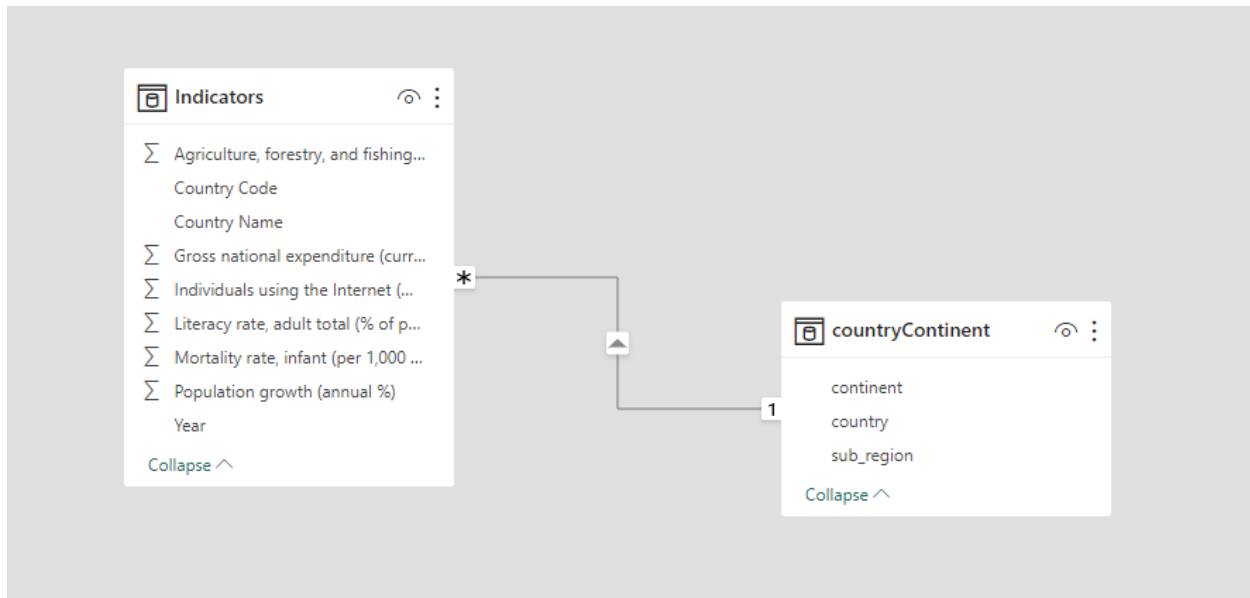
Year	Country Name	Country Code	Mortality rate, infant (per 1,000 live births)	Literacy rate, adult total (% of people ages 15 and above)	Individuals using the Internet
2013	China	CHN	9.7	90.87570763	
2013	United States	USA	6	90.87570763	
2013	Germany	DEU	3.3	90.87570763	

**Fig 2.11:** DAX report

The Data Analysis Expressions (DAX) Power BI report to display the maximum gross national expenditure for each country based on the current context set by any slicers, filters, or other report interactions.

The table shows a partial view of the dataset with columns for Year, Country Name, Country Code, Mortality rate, Literacy rate, and Individual internet use. The data are for the year 2013 and show values for China, the United States, and Germany. These values could be used in conjunction with the DAX formula to analyze and compare the gross national expenditure for the countries alongside other indicators such as mortality rate, literacy rate, and internet usage.

## Use of Relationship



**Fig 2.12:** Relationship table

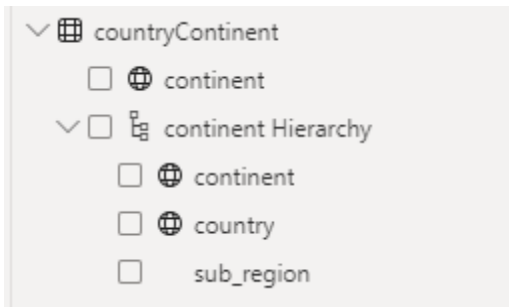
The Power BI showing the relationship between two tables: "Indicators" and "countryContinent". The "Indicators" table contains various metrics like agriculture, gross national expenditure, literacy rate, internet usage, infant mortality rate, population growth, and year, along with a "Country Name" and "Country Code" that likely serve as identifiers for each record's geographical location.

The "countryContinent" table contains "continent", "country", and "sub\_region", which are probably categorical data that classify countries into broader geographic or political groups.

The arrow between the two tables indicates a relationship has been established. This typically means that the "Country Name" or "Country Code" from the "Indicators" table has been linked to the "country" field in the "countryContinent" table. The "1" by the "countryContinent" table suggests that for each country in this table, there should be one and only one corresponding set of indicator values in the "Indicators" table. This is known as a one-to-many relationship, where one row in the "countryContinent" table can relate to many rows in the "Indicators" table.

This kind of relationship is fundamental in Power BI for creating reports that can filter and summarize data across related tables, allowing for more dynamic and complex data analysis. For example, with this relationship, you could easily analyze the indicators by continent or sub-region.

## Hierarchies



**Fig 2.13:** Hierarchies

The Power BI showing a hierarchy within a data model, specifically within a table named "countryContinent".

In Power BI, a hierarchy is a way of organizing data that has a natural "drill-down" path. It allows users to explore data at different levels of granularity and is particularly useful in reports and dashboards where end users may want to navigate through layers of data, from a top-level summary down to more detailed information.

The hierarchy shown in the result, named "continent Hierarchy", seems to include three levels:

- **Continent:** This would be the top level, where data is aggregated at the largest geographic segmentation. For example, users could first view data for Asia, Europe, Africa, etc.
- **Country:** This is the next level down in the hierarchy and allows users to drill down from a continent to specific countries within that continent. If a user starts with Europe, they could then navigate to data specifically for France, Germany, Spain, etc.
- **Sub\_region:** This is the third level and would provide an even more granular breakdown within a country. Depending on the dataset, this could represent states, provinces, or other political subdivisions.

By clicking on a continent in a visual, users could drill through to see the countries in that continent, and then further drill down into the sub-regions of a particular country. This hierarchical view is beneficial for users who need to perform a multi-level analysis that starts from a broad perspective and then focuses on more specific details.

## Grouping

**Groups** [X]

Name \*  
Country (groups)

Field  
Country Name

Group type  
List

Ungrouped values

Groups and members

- Developed Countries
  - Australia
  - Brazil
  - China
  - Germany
  - Japan
  - Spain
  - United States
- Developing Countries
  - India
  - Kenya

Group Ungroup

☐ Include Other group ⓘ

OK Cancel

**Figure 2.14:** Group model

The Power BI's "Groups" feature within the query editor or data model view. This feature allows users to categorize data into meaningful clusters that can be used for analysis or reporting. Here's a breakdown of the elements shown in the result:

**Group Name:** This is where you name your group of data. In the result, the group is named "Country (groups)", which indicates that countries will be grouped based on certain criteria.

**Field:** The field selected for grouping is "Country Name", meaning the grouping is being done based on the names of countries.

**Group Type:** The type of grouping is a "List", which allows the user to manually define and list the groups and their members.

**Ungrouped Values:** This area is blank, which indicates all country values have been assigned to a group and none is left ungrouped.

**Groups and Members:** There are two groups defined in the result:

**"Developed Countries":** This group includes Australia, Brazil, China, Germany, Japan, Spain, and the United States.

"Developing Countries": This group includes India and Kenya.

Include 'Other' group: This checkbox, when selected, creates an additional group for any data not included in the defined groups. It is not selected in the result, suggesting that all the countries have been categorized into one of the two groups.

By creating groups, users can easily perform comparative analysis between different categories, such as comparing sales figures between developed and developing countries, or assessing the performance of various sectors within these groups. It simplifies the process of filtering and visualizing data across these defined segments in Power BI reports and dashboards.

## **2. Discussion**

The Power BI analysis entailed several steps to synthesize and interpret growth indicators across selected countries. Throughout the process, data was meticulously prepared, grouped, and visualized to aid in understanding various economic, educational, and health metrics.

The discussion of the analysis would focus on the insights gathered from the visualizations and how they align with the defined objectives. Each visualization provided a unique perspective on the indicators, revealing trends and outliers that merit further investigation. For instance, the income level comparison might show disparities in economic power, while the infant mortality rates could highlight healthcare priorities or challenges among the countries.

The sectoral contribution to the economy might point to the reliance of certain countries on specific industries and how diversification or lack thereof impacts economic resilience. Education and literacy analyses could reflect the effectiveness of educational policies and their long-term impact on societal advancement. The digital connectivity study would underscore the digital divide and the importance of internet accessibility as a driver of modern economies.

## **3. Conclusion**

In conclusion, the Power BI analysis offered a comprehensive overview of growth indicators, with rich visualizations that facilitated a deeper understanding of complex datasets. These insights could be leveraged by policymakers, educators, and environmental planners to inform strategies that promote sustainable growth, enhance literacy, and improve health outcomes.

Through strategic grouping, such as separating countries into developed and developing, we could tailor analyses and develop targeted interventions. The clear trends observed – for example, the correlation between internet



usage and literacy rates – suggest that investment in digital infrastructure could bolster education and, by extension, economic growth.

The hierarchies and relationships defined in the model allowed for nuanced analysis, such as drilling down from continent to country to sub-region levels, enabling region-specific insights. This granularity is critical when considering localized policies or initiatives.

In conclusion, the Power BI analysis not only provided a snapshot of the current state of growth indicators but also laid the groundwork for predictive analytics, such as using ARIMA models for forecasting. The insights derived from this study can inform decisions, prioritize resource allocation, and ultimately, drive progress towards achieving the Sustainable Development Goals (SDGs).

### Part Three: References

- Asuero, A. G., Sayago, A., & González, A. G. (2006). The correlation coefficient: An overview. *Critical Reviews in Analytical Chemistry*, 36(1), 41–59.
- Becker, L. T., & Gould, E. M. (2019). Microsoft power BI: Extending excel to manipulate, analyze, and visualize diverse data. *Serials Review*, 45(3), 184–188.
- Berry, W. D., & Feldman, S. (1985). *Multiple regression in practice* (No. 50). Sage.
- Brownstein, N. C., Adolfsson, A., & Ackerman, M. (2019). Descriptive statistics and visualization of data from the R datasets package with implications for clusterability. *Data in Brief*, 25, 104004.
- Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1–4).
- Cooksey, R. W., & Cooksey, R. W. (2020). Descriptive statistics for summarising data. In *Illustrating statistical procedures: Finding meaning in quantitative data* (pp. 61-139).
- Cuaresma, J. C., Doppelhofer, G., & Feldkircher, M. (2014). The determinants of economic growth in European regions. *Regional Studies*, 48(1), 44-67.
- Deckler, G., Powell, B., & Gordon, L. (2022). *Mastering Microsoft Power BI: Expert techniques to create interactive insights for effective data analytics and business intelligence*. Packt Publishing Ltd.
- Gordon, R. A. (1968). Issues in multiple regression. *American Journal of Sociology*, 73(5), 592–616.
- Kelley, K., & Bolin, J. H. (2013). Multiple regression. In *Handbook of quantitative methods for educational research* (pp. 69-101). Brill.
- Klein, J. P., & Moeschberger, M. L. (2003). Hypothesis testing. In *Survival analysis: techniques for censored and truncated data* (pp. 201-242).
- Krishnan, V. (2017). *Research data analysis with power bi*.
- Larson, M. G. (2006). Descriptive statistics and graphical displays. *Circulation*, 114(1), 76-81.
- Meng, X. L., Rosenthal, R., & Rubin, D.B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin* 111(1), 172-175.
- Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777-787.
- Schwartz, J., & Marcus, A. (1990). Mortality and air pollution j london: a time series analysis. *American Journal of Epidemiology*, 131(1), 185-194.
- Seamark, P., & Martens, T. (2019). *Pro DAX with Power BI: Business Intelligence with PowerPivot and SQL Server Analysis Services Tabular*. Apress.
- Statistics, L. (2013). Hypothesis testing. The Null and Alternative Hypothesis.
- Stecyk, A. (2018). The analytic hierarchy process AHP for business intelligence system evaluation. *European Journal of Service Management*, 28(4/2), 439-446.
- Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of Diagnostic Medical*

Sonography, 6(1), 35-39.

- Thomson, H., Thomas, S., Sellstrom, E., & Petticrew, M. (2013). Housing improvements for health and associated socio-economic outcomes. *Cochrane Database of Systematic Reviews*, (2).
- Wei, L., & Zhen-gang, Z. (2009, December). Based on time sequence of ARIMA model in the application of short-term electricity load forecasting. In *2009 International Conference on Research Challenges in Computer Science* (pp. 11-14). IEEE.
- Widjaja, S., & Mauritsius, T. (2019). The development of performance dashboard visualization with power BI as platform. *International Journal of Mechanical Engineering and Technology*, 10(5), 235-249.