University of
Salford
MANCHESTER

Assessment:
Applied Statistics and Data Visualisation
(MSc Data Science)

Title:

Harnessing the Power of Analysis and Visualization: A Holistic Study of Economic Performance, Environmental Sustainability, and Governance using R and Power BI

By:

NNAEMEKA NDUBUISI
@00738283

06/12/2023

TABLE OF CONTENT:

1. INTRODUCTION

1.1. Overview

Over the past decade, the world has witnessed an unprecedented surge in data generation, with over 2.5 quintillion bytes created daily at our current pace [1]. This exponential growth is propelled by continuous advancements in Artificial Intelligence (AI) and the Internet of Things (IoT), promises an ever-increasing deluge of information. This surge underscores the critical role of Data Scientists and analysts in navigating the complexities of this vast dataset landscape. Consequently, this data-driven revolution has also arisen an uproar in ways to better visualize data, aiding in understanding the nature and results of datasets. Effectively designed data visualizations enable observers see patterns in data related to science, education, health, and public policy (Franconeri et.al 2021).

1.2. Objectives

This research report embarks on an exploration of this data-driven era, employing advanced statistical analyses, predictive modeling and visualization techniques. As proposed by the assessment brief, 4 objectives of choice have extensively applied in generate clear models and reports. A dataset of 12 countries and 17 indicators across 15 years was used for this task. The objectives and analysis method applied for each task are listed as follows:

Part One

- **Objective1: Economic Growth and Development**
  Descriptive Statistical Analysis (subtask 4-1)
- **Objective 2: Environmental Sustainability and Air quality**
  Correlation Analysis (subtask 4-2)
- **Objective 3: Health and Well-being**
  Hypothesis Testing (subtask 4-3)
- **Objective 4: Environmental Sustainability**
  Regression Analysis (subtask 4-4)
- **Objective 5: Export Gains and Trade Balance**
  Time Series (subtask 4-5)

Part Two

- **Objective 1: Economic Performance**
  Visualization: GDP Growth and Sum of Exports of Goods and Services (Bar Chart)
- **Objective 2: Analyzing Export Gains and GDP Growth**
  Visualization: Bubble Chart - Manufacturing Exports, Total Exports, and GDP per Capita
- **Objective 3: Relationship between CO2 Emissions and Forest Area**
  Visualization: Area Chart - CO2 Emissions and Forest Area Over Time
- **Objective 4: Corruption and Governance**
  Visualization: Ribbon Chart - Average of Control_of_Corruption by Government_Effectiveness and Country_Name

2. BACKGROUND RESEARCH AND LITERATURE REVIEW

Within the dynamic sphere of data science, this literature review scrutinizes analogous works and recent strides. Comprehensive examination of datasets, statistical methodologies, and models employed underscores the evolving landscape. Navigating a rich tapestry of research, the review delves into the interdisciplinary fusion of statistics and interactive dashboard designs. Insightful observations unfold, illuminating the transformative impact of these advancements across diverse domains. This exploration sheds light on the pivotal role data science plays in shaping contemporary methodologies and fostering innovation within an ever-expanding spectrum of applications.

2.1. Background Research

Have you ever asked yourself, “I wonder why people behave that way?” If so, then you have already begun the research process. Research begins with asking questions. Curiosity about a casual observation that you have made could initiate a series of questions [3]. The objectives for this research as started in chapter 1. Certainly, it is still unclear what discoveries would be made at this stage until inclusive statistical analysis is carried out. Defining crucial parameters for potential models is key, with tools like the correlation matrix aiding in identifying relationships between indicators.

2.1.1. Descriptive Analysis

Descriptive statistical analysis is a branch of statistics that focuses on summarizing and presenting data in a meaningful and informative way. It involves organizing, analyzing, and interpreting data using various statistical measures, graphs, and charts to describe the main characteristics, patterns, and trends present in the data (Johnson, 2013).

Descriptive statistics provide a concise and accessible summary of the data, allowing researchers and analysts to gain initial insights into the dataset without making inferences or generalizations about a larger population (Hays, 2013). It helps in understanding the central tendency, variability, shape, and distribution of the data, providing a foundation for further analysis and decision-making.

One commonly used measure in descriptive statistics is the arithmetic mean, or average, which provides a measure of central tendency. Other measures of central tendency include the median (the middle value) and the mode (the most frequently occurring value). Measures of variability, such as the range, variance, and standard deviation, depict the spread or dispersion of the data points around the central tendency (Witte & Witte, 2019). Additionally, measures like percentiles and quartiles help identify specific positions within the dataset.

Graphical representations play a crucial role in descriptive statistics. Histograms display the distribution of numerical data, while bar charts and pie charts represent categorical data (Field, 2013). Box plots provide a visual summary of the data's quartiles, median, and outliers, while scatter plots illustrate the relationship between two continuous variables. These graphical tools enhance the understanding and communication of data patterns.
Statistical analysis is a fundamental aspect of data-driven decision-making, relying on methodologies rooted in foundational statistical concepts. Among these methodologies, regression analysis plays a crucial role in understanding the relationships between variables. Sir Francis Galton's pioneering work in the 19th century laid the groundwork for regression analysis, providing a method to predict the value of a dependent variable based on one or more independent variables and uncovering valuable insights into underlying relationships (Galton, 1886).

2.1.2. Regression Analysis

Regression analysis involves estimating the parameters of a mathematical model that describes the relationship between a dependent variable and one or more independent variables. The foundational equation for simple linear regression, where there is only one independent variable, is expressed as:

- $Y=\beta0+\beta1X+\varepsilon$

In this equation, *Y* represents the dependent variable, *X* represents the independent variable, $\beta0$ represents the y-intercept, $\beta1$ represents the slope coefficient, and ε represents the error term that captures unexplained variation.
Multiple linear regression extends this concept to include more than one independent variable:

- $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\varepsilon$

where:

1. **Y (Dependent Variable):**
   - ▯*Y* is the variable we are trying to predict.
2. **$\beta0$ (Y-intercept):**
   - ▯0$\beta0$ is the starting point of the regression line.
3. **1, 2,…,$\beta1$,$\beta2$,…,βp (Slope Coefficients):**
   - 1,2,…,$\beta1$,$\beta2$,…,βp quantify the relationship between each independent variable and *Y*.
4. **1,2,…,*X*1,*X*2,…,*Xp* (Independent Variables):**
   - 1,2,…,*X*1,*X*2,…,*Xp* are the factors influencing *Y*.
5. **ε (Error Term):**
   - ε captures unexplained variation in *Y*.

This equation becomes particularly powerful in engineering and predictive modeling, allowing for the consideration of various input variables when forecasting outcomes. In the realm of future engineering, multiple linear regression can be employed to model complex systems where multiple factors influence the outcome. For instance, in designing a new product, engineers can use multiple linear regression to understand how different parameters contribute to the overall performance or efficiency (Kutner et al., 2004).

2.1.3. Future Engineering

Future engineering refers to the application of advanced technologies, methodologies, and innovative approaches to address the challenges and opportunities that lie ahead. It involves envisioning and designing solutions that cater to the evolving needs and demands of society, considering the anticipated advancements in science, technology, and various industries. Future engineering focuses on creating sustainable and resilient systems, optimizing resource utilization, and improving overall efficiency to shape a better future.

In the realm of future engineering, various models and formulas can be employed to analyze and predict outcomes. While the specific formulas utilized may vary depending on the specific context and problem at hand, I will provide a general overview of some commonly used approaches.

2.1.4. Time series analysis

Time series analysis refers to the statistical methods and techniques used to analyze and interpret data that is collected over time, typically at regular intervals. It involves studying the sequential nature of the data to identify patterns, trends, and dependencies, allowing for the understanding and prediction of future values or behavior. A time series is a set of observations X_1, each one being recorded at a specific time T_0 (Brockwell, P. J., & Davis, R. A. 2016)

Time series analysis deals with data points ordered chronologically, where the values of the variable of interest are observed over successive time periods. The analysis of time series data encompasses various aspects, including trend analysis, seasonality, cyclicality, and the identification of underlying patterns and structures. This type of analysis also serves as a powerful tool for unraveling patterns and trends over a time frame. Yule initiated the statistical analysis of time series in 1927 by applying autoregressive models to real-world data [3]. This groundbreaking work was further developed by Walker (G Walker, 1931) who made modifications to Yule's method. As a result, a set of equations known as the "Yule-Walker equations" emerged, providing a framework for determining the order of the autoregressive process in a time series. Forecasting models, such as the Holt-Winters exponential smoothing model, find their application within this domain, offering systematic insights into future trends.

2.1.5. Hypothesis Testing

Hypothesis testing is a fundamental statistical method used to make inferences and draw conclusions about a population based on sample data. It involves formulating a hypothesis, collecting and analyzing data, and assessing the evidence against the null hypothesis. The process allows researchers to determine whether the observed data provides sufficient evidence to support or reject the proposed hypothesis.

In hypothesis testing, two competing hypotheses are considered: the null hypothesis (H_0) and the alternative hypothesis (H_1). The null hypothesis represents the assumption of no effect, no difference, or no relationship in the population, while the alternative hypothesis suggests otherwise.

Works by Lehmann and Romano, offer in-depth coverage of hypothesis testing concepts, techniques, and applications (Lehmann, E. L., & Romano, J. P. (2005). They provide detailed explanations of various statistical tests, including t-tests, chi-square tests, ANOVA, and regression analysis, along with their corresponding formulas and assumptions.

The general structure of a hypothesis test involves defining two hypotheses: the null hypothesis (H_0) and the alternative hypothesis (Ha or H_1). The process of hypothesis testing often follows a set procedure:

1. **Null Hypothesis (H_0):**
   - The null hypothesis represents a statement of no effect or no difference in the population. It is denoted as $H0$.
2. **Alternative Hypothesis (Ha or H_1):**
   - The alternative hypothesis represents a statement of an effect or a difference in the population, and it is what researchers typically want to find evidence for. It is denoted as Ha or $H1$.
3. **Test Statistic:**
   - A test statistic is a numerical value calculated from the sample data. It quantifies the difference between the observed data and what is expected under the null hypothesis.
4. **Significance Level (α):**
   - The significance level, denoted as α, is the predetermined level of significance that determines the threshold for rejecting the null hypothesis. Common choices for α are 0.05, 0.01, or 0.10. throughout this research the study a threshold values 0.05 was applied
5. **P-value:**

- The p-value is the probability of observing a test statistic as extreme as, or more extreme than, the one observed in the sample, assuming that the null hypothesis is true. A lower p-value provides stronger evidence against the null hypothesis.

2.1.6. Interaction Design:

Interactive dashboard design, a discipline drawing from data visualization and human-computer interaction, is based in on psychology and user experience design. Another reason to visualize numbers is to help our memory. To read the table, we need to look up every value, one at a time which can be difficult (Wexler, S et.al. 2017). This manifestation by Steve Wexler (et.al) proves that the principle of dashboard designs not only enhance the aesthetic appeal of visual representations but also ensure that users can seamlessly interact with and derive meaningful insights from complex datasets.

2.2. Literature Review

In the Scholarly investigations of time series analysis, ranging from finance to meteorology and economics, the contributions of George E.P. Box and Gwilym Jenkins stand as foundational pillars (Box, G. E. P. et.al 2015). Their work layed the foundation for understanding temporal patterns and trends. Recent advancements contribute to refining predictive models, offering better accuracy in anticipating future trends. This evolution in time series analysis not only builds upon the classic methodologies introduced by Box and Jenkins but also reflects the dynamic nature of research in adapting to contemporary challenges.

In the context of Air Quality and Environmental Sustainability, the correlation analysis applied draws inspiration from the extensive literature in environmental science. Understanding the intricate relationships between CO2 emissions and various economic indicators is crucial for addressing environmental impact Box (G. E. P., Jenkins 2015). This research leverages insights from existing studies to develop strategies promoting sustainable practices and mitigating potential threats to our environment. Through this multidisciplinary exploration, the research aims to contribute not only to statistical analysis but also to broader discussions on sustainable development.

within the realm of economic growth, delve into intricate factors shaping GDP per capita, foreign direct investments, and tax revenue. Daron Acemoglu and James A. Robinson's book, "Why Nations Fail", presents a compelling argument that the success or failure of countries GDP is primarily determined by the strength of their institutions, rather than cultural, geographical, or random factors (Acemoglu, D., & Robinson, J. A. 2012). In their latest publication, the authors develop a fresh theoretical framework on the concept of liberty and its attainment. Drawing from a diverse range of sources, including contemporary events and various periods of world history, they provide substantial evidence to support their theories.

2.3. Libraries and Packages

Libraries and packages are foundational elements in programming, encapsulating reusable code for specific functionalities. A library is a collection of pre-built code modules, fostering code reusability, while a package is a structured compilation of related modules, enhancing code organization. Both streamline development, allowing programmers to leverage existing solutions and collaborate efficiently. In data analysis and scientific computing, libraries and packages provide ready-made tools, accelerating tasks such as statistical analysis, machine learning, and data visualization. They represent a collaborative coding community, driving innovation and efficiency in software development.

Libraries and packages used within the context of this study include the following:
(see read me file)

3. PREPARATION AND EXPLORATION OF DATA SET

3.1 Data Dictionary

Below is a concise and informative table to show names all indicators (variables) used within the context of this research. The data was generated through Databank World bank, all attribution and recognition were referenced in the literature review.

COLUMN NAME	ORIGINAL COLUMN NAME	DEFINITION	TIME FRAME	DATA SOURCE
TIME	Time	Year or time period in which the data is reported.	2008 - 2022	DataBank World Development Indicators
TIME_CODE	Time_Code	Code representing the specific year or time period.	2008 - 2022	DataBank World Development Indicators
COUNTRY_NAME	Country_Name	Name of the country for which the data is reported.	2008 - 2022	DataBank World Development Indicators
COUNTRY_CODE	Country_Code	Code representing the specific country.	2008 - 2022	DataBank World Development Indicators

GDP_PER_CAPITA	GDP per capita (current US$)	Gross Domestic Product per capita in current US dollars.	2008 - 2022	DataBank World Development Indicators
GDP_GROWTH	GDP growth (annual %)	Annual percentage growth of Gross Domestic Product.	2008 - 2022	DataBank World Development Indicators
FDI_PCT_GDP	Foreign direct investment, net inflows (% of GDP)	Percentage of GDP representing net inflows of foreign direct investment.	2008 - 2022	DataBank World Development Indicators
CO2_EMISSIONS_KT	CO2 emissions (kt)	Total carbon dioxide emissions in kilotons.	2008 - 2022	DataBank World Development Indicators
FOREST_AREA_SQKM	Forest area (sq. km)	Total land area covered by forests in square kilometers.	2008 - 2022	DataBank World Development Indicators
REC_PCT_TFEC	Renewable energy consumption (% of total final energy consumption)	Percentage of total final energy consumption derived from renewable sources.	2008 - 2022	DataBank World Development Indicators
RULE_OF_LAW	Rule of Law: Estimate	Estimate of the rule of law in a country.	2008 - 2022	DataBank World Development Indicators
CONTROL_OF_CORRUPTION	Control of Corruption	Estimate of the control of corruption in a country.	2008 - 2022	DataBank World Development Indicators
GOVERNMENT_EFFECTIVENESS	Government Effectiveness: Estimate	Estimate of the effectiveness of the government in a country.	2008 - 2022	DataBank World Development Indicators
EXPORTS_OF_GS	Exports of goods and services (% of GDP)	Percentage of GDP represented by the exports of goods and services.	1982 - 2022	DataBank World Development Indicators

MANUFACTURES_EXPORTS_PCT_ME	Manufactures exports (% of merchandise exports)	Percentage of merchandise exports that are manufactured goods.	2008 - 2022	DataBank World Development Indicators
MERCHANDISE_EXPORTS_USDOL	Merchandise exports (current US$)	Total value of merchandise exports in current US dollars.	2008 - 2022	DataBank World Development Indicators
TAX_REVENUE_PCT_GDP	Tax revenue (% of GDP)	Percentage of GDP represented by tax revenue.	2008 - 2022	DataBank World Development Indicators
TRADE_PCT_GDP	Trade (% of GDP) [NE.TRD.GNFS.ZS]	Percentage of GDP represented by trade.	2008 - 2022	DataBank World Development Indicators
LIFE_EXPECTANCY_AT_BIRTH	Life expectancy at birth, total (years)	The average number of years a newborn is expected to live, assuming that age-specific mortality rates remain constant throughout its life.	2008 - 2022	DataBank World Development Indicators
DGGHE	Domestic general government health expenditure (% of GDP)	Percentage of GDP spent on domestic general government health expenditure.	2008 - 2022	DataBank World Development Indicators
PHYSICIANS	Physicians (per 1,000 people) [	The number of physicians per 1,000 people in a country.	2008 - 2022	DataBank World Development Indicators

Table 3.1: Data dictionary of all variables.

3.2 Preprocessing

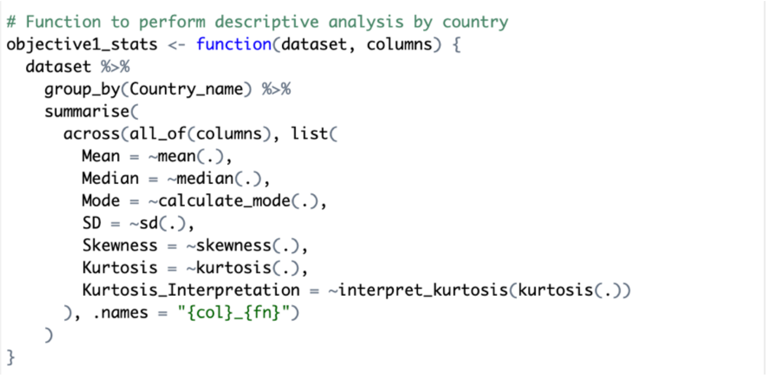
Real-world datasets, including the one utilized in this research, seldom arrive in a pristine state. This is primarily attributed to the inherent challenges of data collection, due to various factors, such as data collection methods, human errors, or system limitations. These imperfections can manifest as inconsistencies, missing values, or formatting issues, making preprocessing a crucial step to enhance data quality. To address these challenges, the dataset underwent cleaning and preprocessing using a combination of simple and advanced methods.

Initially, using Microsoft excel, the column names were shortened to a more suitable form. They were shortened, eliminating spaces, numerical figures, and special characters. This step was crucial to ensure seamless integration into the R script environment, where such elements might pose compatibility issues. During the Excel exploration, it was discovered that missing values were denoted as "..", and these were systematically removed to facilitate straightforward identification of missing values within the R environment.

Upon importing the dataset into R, irrelevant columns, such as Country_Code and Time_Code, were excluded from the analysis. These columns were deemed unnecessary as similar information already existed within the dataset.

In a superlative approach to handling missing values, the dataset was treated on a per-country basis. Recognizing the diversity among developed, developing and underdeveloped countries represented in the data, a generalized replacement of missing values was considered inappropriate. Instead, the missing values were addressed individually for each country. This approach ensures that the data imputation process considers the unique characteristics of each country, preventing unintended distortions that may arise from a uniform replacement strategy.

The code snippet code below demonstrates this targeted missing value imputation, where the missing values were replaced with the mean of their respective country variables. Subsequently, the dataset was restructured to its original format and thoroughly examined to confirm its readiness for further analysis.



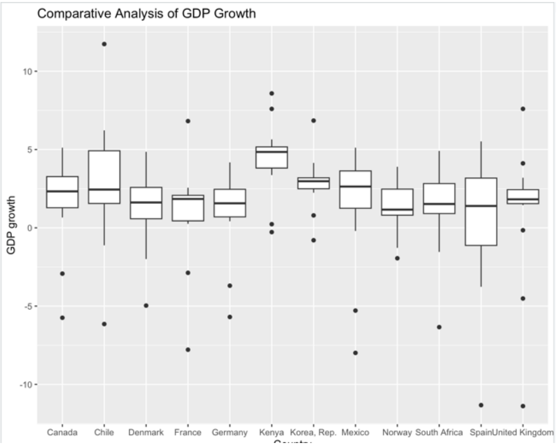
```
# Function to perform descriptive analysis by country
objective1_stats <- function(dataset, columns) {
  dataset %>%
    group_by(Country_name) %>%
    summarise(
      across(all_of(columns), list(
        Mean = ~mean(.),
        Median = ~median(.),
        Mode = ~calculate_mode(.),
        SD = ~sd(.),
        Skewness = ~skewness(.),
        Kurtosis = ~kurtosis(.),
        Kurtosis_Interpretation = ~interpret_kurtosis(kurtosis(.))
      ), .names = "{col}_{fn}")
    )
}
```

Code 3.1: Replacing missing values.

3.3 Exploratory Data analysis

To ensure system compatibility and mitigate potential constraints, a subset of the dataset, referred to as "numeric_dataset," was created, and all values within this subset were uniformly converted to numeric format. Subsequently, a boxplot visualization, as illustrated in Figure xxx, revealed the presence of outliers across few variables. Given the inherent nature of real-world data, the existence of outliers is expected. It is noteworthy to highlight the significant outliers observed within the "Merchandise_Exports_Usdol," "Co2_Emissions_Kt," and "Forest_Area_Sqkm" variables. These observations will be duly considered in the forthcoming statistical analyses.

As a preliminary step, a simple comparative analysis was conducted on GDP growth, aligning with one of the key objectives outlined in Chapter 1. Notably, Kenya emerged unexpectedly as the leading nation with a commendable average growth over the 15-year timeframe. Conversely, Spain exhibited a surprising and pronounced negative decline. These intriguing findings (see Figure 3.xxx) will undergo further exploration through detailed descriptive statistical analysis in Chapter 4, using “GDP_per_capita” as baseline factor. This in-depth analysis will focus on key statistical measures such as mean, median, mode, standard deviation, skewness, and kurtosis, shedding light on the underlying dynamics influencing these observed trends.



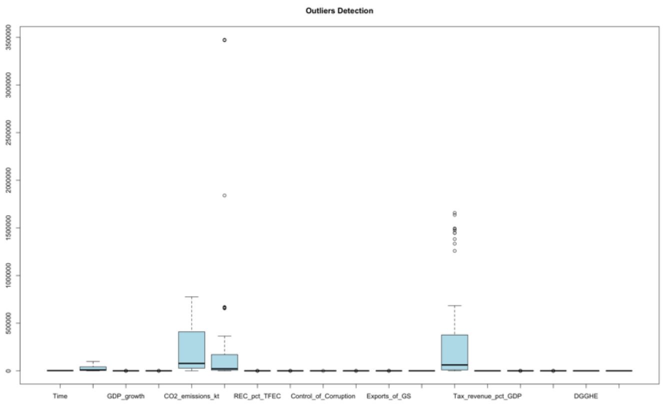


Figure 4.1: Box plot of GDP_growth

Figure 4.1: Box plot of outliers within countries

PART TWO:

4. STATISTICAL ANALYSIS

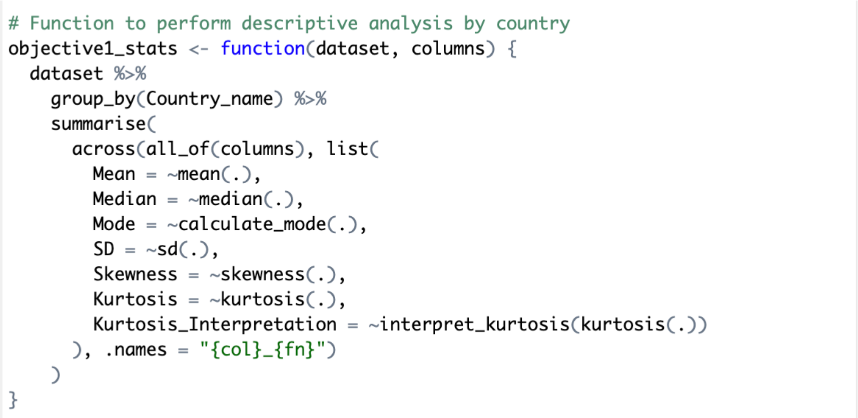
In this extensive statistical analysis, our exploration spans various dimensions, encompassing financial, social, political, and environmental aspects. The overarching objective is to unravel meaningful relationships and glean insights aligned with our stated objectives. Employing a robust set of methodologies, including descriptive statistics, correlation analysis, hypothesis testing, regression analysis, and time series analysis, all executed using R, the study seeks not only to describe current indicators but also to unearth underlying connections.

4.1 Descriptive Statistical Analysis

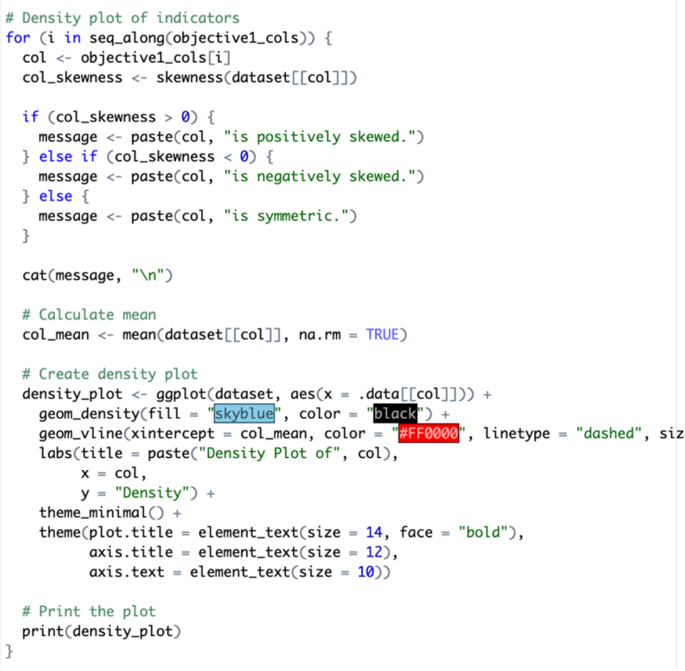
Objective: Economic Growth and Development

Our focus in this analysis lies on indicators associated with economic growth and development: "GDP_per_capita," "FDI_pct_GDP," and "Tax_revenue_pct_GDP." Employing descriptive statistics, encompassing mean, median, mode, standard deviation, skewness, and kurtosis, we delve into each variable's nuances within individual countries. This approach goes beyond presenting a global snapshot, allowing us to understand economic growth and development within specific segments.

The primary findings from the descriptive statistical analysis are shown in Table 4.1 below. While the presence of outliers is evident in many of the variables as shown in our EDA, it is essential to assess skewness before drawing detailed conclusions the method can be seen in code 4.1 below.



```r
# Function to perform descriptive analysis by country
objective1_stats <- function(dataset, columns) {
  dataset %>%
    group_by(Country_name) %>%
    summarise(
      across(all_of(columns), list(
        Mean = ~mean(.),
        Median = ~median(.),
        Mode = ~calculate_mode(.),
        SD = ~sd(.),
        Skewness = ~skewness(.),
        Kurtosis = ~kurtosis(.),
        Kurtosis_Interpretation = ~interpret_kurtosis(kurtosis(.))
      ), .names = "{col}_{fn}")
    )
}
```

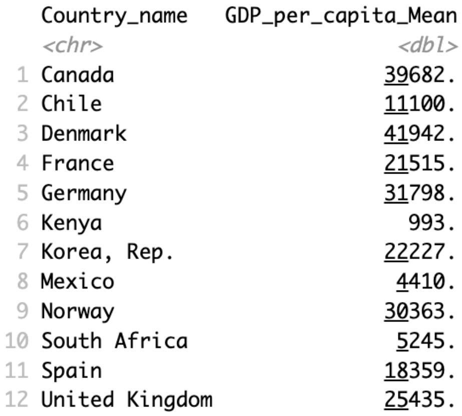


```r
# Density plot of indicators
for (i in seq_along(objective1_cols)) {
  col <- objective1_cols[i]
  col_skewness <- skewness(dataset[[col]])

  if (col_skewness > 0) {
    message <- paste(col, "is positively skewed.")
  } else if (col_skewness < 0) {
    message <- paste(col, "is negatively skewed.")
  } else {
    message <- paste(col, "is symmetric.")
  }

  cat(message, "\n")

  # Calculate mean
  col_mean <- mean(dataset[[col]], na.rm = TRUE)

  # Create density plot
  density_plot <- ggplot(dataset, aes(x = .data[[col]])) +
    geom_density(fill = "skyblue", color = "black") +
    geom_vline(xintercept = col_mean, color = "#FF0000", linetype = "dashed", size = 1) +
    labs(title = paste("Density Plot of", col),
         x = col,
         y = "Density") +
    theme_minimal() +
    theme(plot.title = element_text(size = 14, face = "bold"),
          axis.title = element_text(size = 12),
          axis.text = element_text(size = 10))

  # Print the plot
  print(density_plot)
}
```

This step provides insights into the normal distribution of our variables, influencing the reliability of our descriptive analysis results. The skewness results are interpreted as follows:

- GDP_per_capita: Positively skewed, suggesting a bias toward higher values. Most countries have higher GDP per capita, but a few with lower values influence the right-leaning distribution.
- FDI_pct_GDP: Positively skewed, indicating a distribution leaning toward higher values. Most countries have lower FDI percentages, with a subset of nations contributing to the positive skewness.
- Tax_revenue_pct_GDP: Positively skewed, similar to FDI and GDP, implying that while most countries have lower tax revenue as a percentage of GDP, a subset with higher percentages influences positive skewness.

These mean and standard deviation results, visually depicted below, reinforce the rationale for segmenting our analysis by individual countries.



```
   Country_name    GDP_per_capita_Mean
   <chr>                         <dbl>
 1 Canada                       39682.
 2 Chile                        11100.
 3 Denmark                      41942.
 4 France                       21515.
 5 Germany                      31798.
 6 Kenya                          993.
 7 Korea, Rep.                  22227.
 8 Mexico                        4410.
 9 Norway                       30363.
10 South Africa                  5245.
11 Spain                        18359.
12 United Kingdom               25435.
```

Table 4.1: Mean and median of countries in respect to their GDP_per_capita

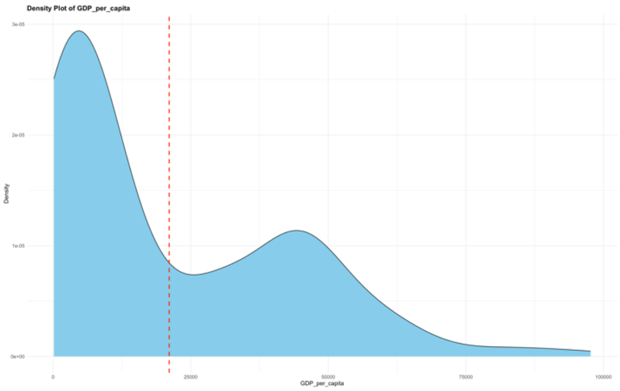


Figure 4.1: Density plot of GDP_per_capita

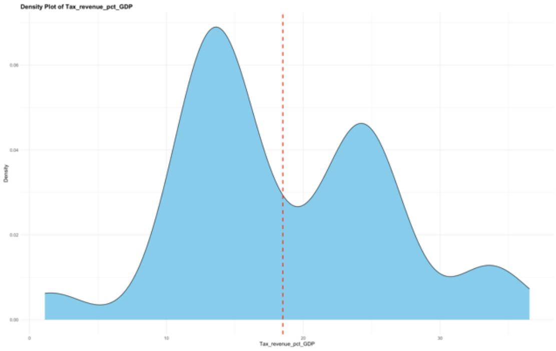


Figure 4.2: Density plot of Tax Revenue.

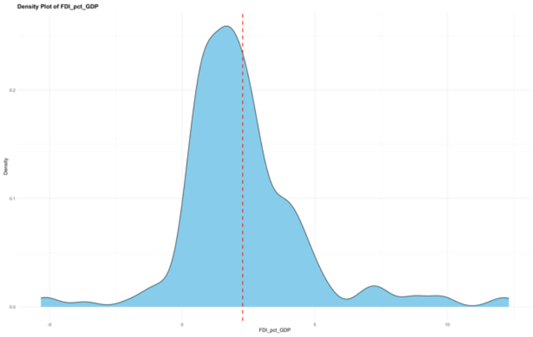


Figure 4.3: Foreign Direct Investment.

The analysis, as presented in figures 4.4, 4.5, 4.6, reveals Denmark's elevated mean and standard deviation of GDP per capita, reflecting a higher standard of living and robust economic growth. In contrast, Kenya exhibits the lowest mean and median GDP per capita, indicating a comparatively lower average income and economic activity.

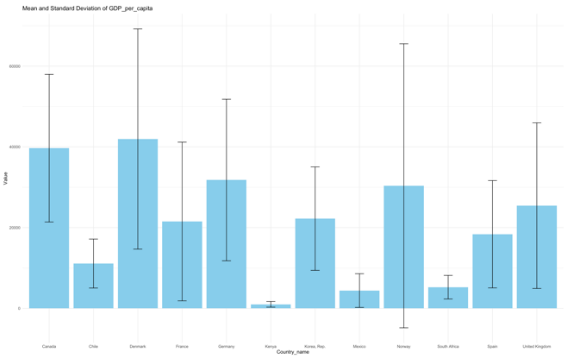


Figure 4.4: mean and SD of GDP_per_capita.

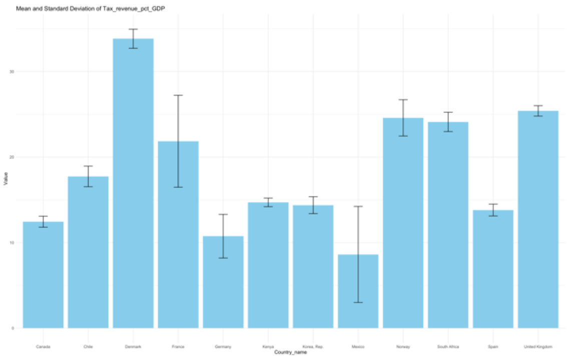


Figure 4.5: mean and SD of GDP_per_capita.

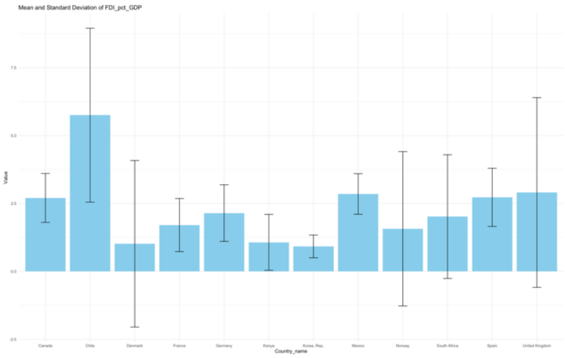


Figure 4.6: mean and SD of GDP_per_capita.

Furthermore, A clear stratification among countries emerges concerning GDP per capita, with developed nations displaying higher values and Chile, Mexico, and South Africa showcasing lower figures, indicating varying levels of economic development.

Regarding Foreign Direct Investments (FDI), Figure 4.6 highlights fluctuations in certain countries, emphasizing that factors beyond this research's scope influences economic growth. Notably, having higher tax revenue or foreign direct investments doesn't guarantee a higher GDP per capita, as seen in France and Chile. This underscores the complexity of measuring economic growth and development, suggesting that numerous known and unknown factors play pivotal roles.

4.2 Correlation Analysis

Objective: Air quality, Climate change, and Environmental Sustainability

To comprehensively understand air quality, climate change, and environmental sustainability, we conducted a correlation analysis with CO2 emissions as the primary indicator. Two distinct approaches were employed to provide a meaningful assessment within our dataset.

The initial approach, represented in Figure 4.7, unveiled certain indicators initially chosen with smaller or negative correlations with CO2 emissions. However, variables such as Manufactures exports and merchandise exports displayed notable positive relationships, confirming our hypothesis that these economic activities contribute to environmental impact.

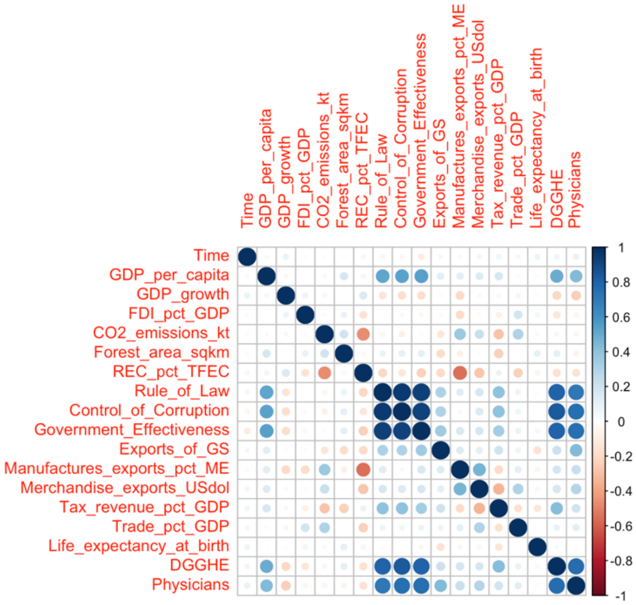


Figure 4.7: Correlation matrix of the variables

In the first approach, a correlation analysis between CO2 emissions (CO2_emissions_kt) and renewable energy consumption as a percentage of total final energy consumption (REC_pct_TFEC) was undertaken using Spearman's test. Before this, a Shapiro-Wilk normality test conducted on the variables of choice, indicated a non-normal distribution of the values, justifying the choice of Spearman's test. The results unveiled a substantial negative correlation coefficient of -0.5132357, affirming our hypothesis that an escalation in renewable energy consumption correlates with a reduction in CO2 emissions.

The second approach involved a correlation matrix using multiple continuous variables (CO2_emissions_kt, Merchandise_exports_USdol, Manufactures_exports_pct_ME, Trade_pct_GDP). Non-normality was confirmed using Shapiro-Wilk test, prompting the application of Spearman's test for the correlation analysis. The results, visualized in Figure 4.8, emphasized a strong positive correlation between Merchandise exports, manufactures exports, and CO2 emissions, affirming our hypothesis that increased exports may contribute to higher CO2 emissions.

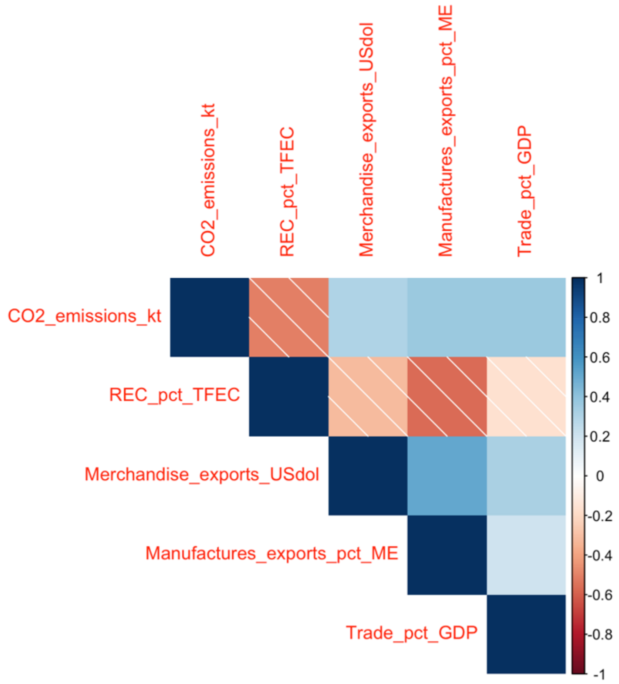


Figure 4.8: correlation of the chosen variables

These findings underscore the intricate relationships within the realm of environmental sustainability, confirming our initial hypothesis. The observed correlations highlight the need for effective mitigation strategies and sustainable energy practices to address potential threats to air quality and environmental sustainability associated with economic activities. Our analysis provides valuable insights into the complex dynamics between economic indicators and environmental impact, guiding future efforts toward a more sustainable and environmentally conscious global economy.

4.3 Hypothesis Testing

Objective 3: Health and Well-being

The purpose of statistical inference is to draw conclusions about a population based on available data. In hypothesis analysis researchers formulates a specific hypothesis, evaluates data from the sample, and uses this result from the data to decide whether they support the specific hypothesis. The first step in testing hypotheses is defining a hypothesis the transformation of the research question into a null hypothesis (H0), and an alternative hypothesis (H1). This test was conducted with generalization across all countries.

Defining the Hypothesis for test 1:

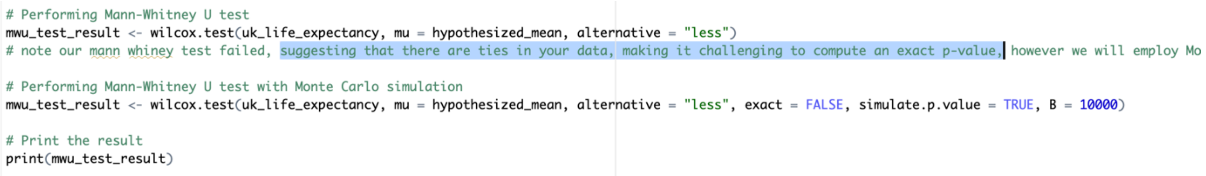
A recent publication by a foreign health institution titled "Is European Life expectancy declining?" asserts a drastic reduction in the United Kingdom's (UK) life expectancy, plummeting from 60yrs. This alarming revelation sparks nationwide concern and inquiries. As a researcher, I am compelled to conduct a hypothesis analysis, delving into publicly available health data to critically assess the accuracy of this publication and address the pressing questions surrounding the reported decline in life expectancy in the UK.

- Null Hypothesis (H0):
  The average life expectancy in the United Kingdom remains above 60 years, and there has been no significant decline.
  H0 > 60

- Alternative Hypothesis (H1):
  The average life expectancy in the United Kingdom has experienced a significant decrease and is now 60 years or less, suggesting a decline in life expectancy.
  H1 <= 60

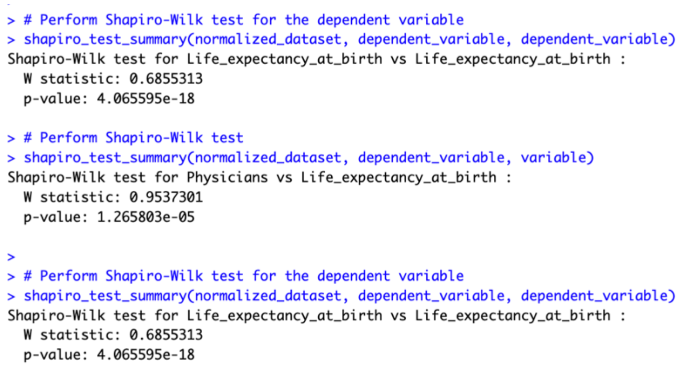
Note: the, above hypothesis is only a make belief scenario to demonstrate sufficient knowledge of hypothesis testing.

In this research segment, functions were established for initializing a Q-Q plot, Shapiro-Wilk test, and standardizing the data using Zscore. These tools were employed to facilitate the respectively visualization of a scatter plot, assess normality of the variables, and standardize the data. Results from the initial Shapiro-Wilk test (refer to Figure 4.9) revealed non-normal distributions for both dependent and independent variables. Hence, Mann-Whitney U Test which is a nonparametric method was employed, however, an unexpected outcome was yielded, suggesting that there are ties in your data, making it challenging to compute an exact p-value. To address this, a Monte Carlo simulation (refer to code 4.2) was introduced to estimate the accurate p-value, assuming a mean (mu) of 60 as per our hypothesis.



```
# Performing Mann-Whitney U test
mwu_test_result <- wilcox.test(uk_life_expectancy, mu = hypothesized_mean, alternative = "less")
# note our mann whiney test failed, suggesting that there are ties in your data, making it challenging to compute an exact p-value, however we will employ Mo

# Performing Mann-Whitney U test with Monte Carlo simulation
mwu_test_result <- wilcox.test(uk_life_expectancy, mu = hypothesized_mean, alternative = "less", exact = FALSE, simulate.p.value = TRUE, B = 10000)

# Print the result
print(mwu_test_result)
```

Code 4.2: To perform Shapiro wilk test and Mann-Whitney U test.



```
> # Perform Shapiro-Wilk test for the dependent variable
> shapiro_test_summary(normalized_dataset, dependent_variable, dependent_variable)
Shapiro-Wilk test for Life_expectancy_at_birth vs Life_expectancy_at_birth :
  W statistic: 0.6855313
  p-value: 4.065595e-18

> # Perform Shapiro-Wilk test
> shapiro_test_summary(normalized_dataset, dependent_variable, variable)
Shapiro-Wilk test for Physicians vs Life_expectancy_at_birth :
  W statistic: 0.9537301
  p-value: 1.265803e-05

>
> # Perform Shapiro-Wilk test for the dependent variable
> shapiro_test_summary(normalized_dataset, dependent_variable, dependent_variable)
Shapiro-Wilk test for Life_expectancy_at_birth vs Life_expectancy_at_birth :
  W statistic: 0.6855313
  p-value: 4.065595e-18
```

Figure 4.9: Console result o perform Shapiro wilk test and Mann-Whitney U test.

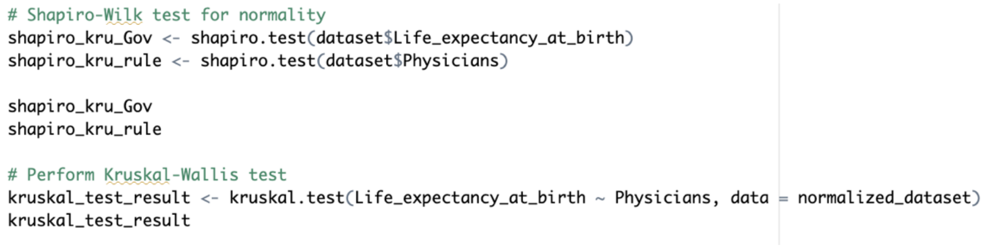
Result from the above procedure returned a p-value of 0.0003624, meaning that there is strong evidence to reject the null hypothesis and support the alternative hypothesis. Suggesting that the observed decline in average life expectancy in the United Kingdom to 60 years or less is unlikely to have occurred by mere chance. Therefore, based on this analysis, there is enough evidence to suggest that the average life expectancy in the United Kingdom has indeed declined to 60 years or lower.

Defining the Hypothesis for test 2:
Here, a generalization has suggested that the mean “Life expectancy at birth” is related and similar to the mean number of “Physicians”. This was conducted using a holistic approach.

- Null Hypothesis (H0):
  The mean "Life expectancy at birth" is not significantly related or similar to the mean number of "Physicians".
  H1: $\mu 1 != \mu 2$

- Alternative Hypothesis (H1):
  The mean "Life expectancy at birth" is related and similar to the mean number of "Physicians".
  H1: $\mu 1 = \mu 2$

Just as in the first test, a Shapiro wilk test conducted showed that the variables do not have a normal distribution so instead of applying so many normalization techniques, another nonparametric approach was used, employing the Kruskal-Wallis test to deduce result for the hypothesis and create inferences from this results. The method of this process is shown below in code 4.3:



```
# Shapiro-Wilk test for normality
shapiro_kru_Gov <- shapiro.test(dataset$Life_expectancy_at_birth)
shapiro_kru_rule <- shapiro.test(dataset$Physicians)

shapiro_kru_Gov
shapiro_kru_rule

# Perform Kruskal-Wallis test
kruskal_test_result <- kruskal.test(Life_expectancy_at_birth ~ Physicians, data = normalized_dataset)
kruskal_test_result
```

Code 4.3: Test for normailty

Afterwards, the test yielded a p-value of 0.07633. The results from the hypothesis analysis indicate weak evidence to reject the null hypothesis. Therefore, based on the available evidence, there is insufficient statistical support to conclude a significant relationship or similarity between the mean "Life Expectancy at Birth" and the mean number of "Physicians." Hence, the alternative hypothesis is accepted. A bar plot (see Figure 4.9) visually illustrates the test results.

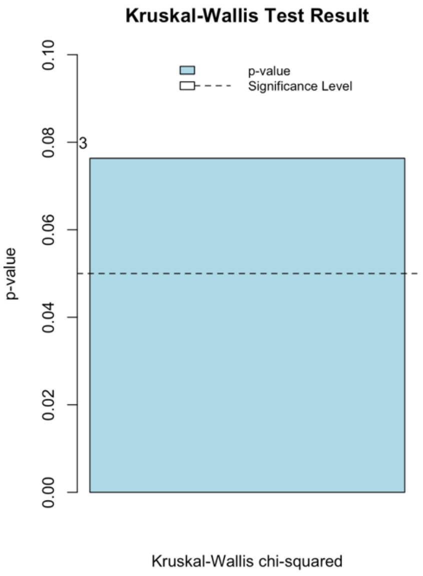


Figure 4.9: showing the result of Kruskal test p value

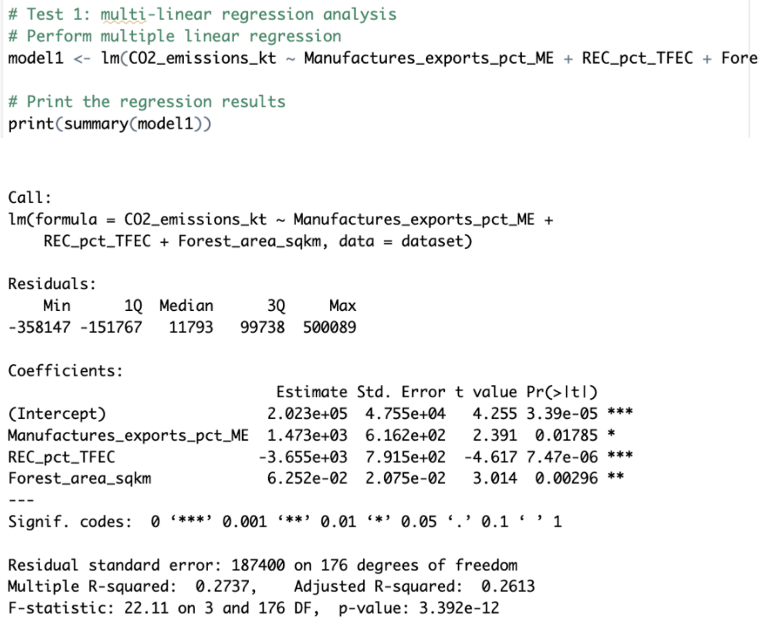
4.4 Regression Analysis

Objective: Environmental Sustainability

This section dives into a regression analysis, aimed at exploring the relationship between environmental sustainability and key manufacturing indicators on a global scale. The primary focus is on carbon emissions as the dependent variable, with Manufacturing Exports Percentage of Total Exports (ME), Renewable Energy Consumption Percentage of Total Final Energy Consumption (TFEC), and Forest Area in Square Kilometers as independent variables.

Test 1: Multi-linear Regression Analysis (MLR)

The analysis was conducted using a multi-linear regression model, and the summarized results are depicted in Figure 4.10.



```
# Test 1: multi-linear regression analysis
# Perform multiple linear regression
model1 <- lm(CO2_emissions_kt ~ Manufactures_exports_pct_ME + REC_pct_TFEC + Forest_area_sqkm, data = dataset)

# Print the regression results
print(summary(model1))
```

```
Call:
lm(formula = CO2_emissions_kt ~ Manufactures_exports_pct_ME +
    REC_pct_TFEC + Forest_area_sqkm, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-358147 -151767   11793   99738  500089

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.023e+05  4.755e+04   4.255 3.39e-05 ***
Manufactures_exports_pct_ME  1.473e+03  6.162e+02   2.391  0.01785 *
REC_pct_TFEC                -3.655e+03  7.915e+02  -4.617 7.47e-06 ***
Forest_area_sqkm             6.252e-02  2.075e-02   3.014  0.00296 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 187400 on 176 degrees of freedom
Multiple R-squared:  0.2737,    Adjusted R-squared:  0.2613
F-statistic: 22.11 on 3 and 176 DF,  p-value: 3.392e-12
```

Figure 4.10: results of MLR test

The adjusted R-squared value of 0.2613 indicates that approximately 26% of the variation in CO2 emissions can be explained by the variables in the model. A small p-value of 3.392e-12 suggests strong evidence against the idea that these variables have no impact on CO2 emissions. Thus, the model asserts that Manufacturing Exports Percentage, Renewable Energy Consumption Percentage, and Forest Area significantly influence CO2 emissions, elucidating a portion of the observed variations.

Referencing the regression formula developed by Sir Francis Galton in the 19th century(see literature review for more details), a tentative model equation was formulated:

- $y_i=\_\ \_\beta_0\_+\_\beta_1\_x_1\_i+\_\beta_2\_x_2\_i+\_\cdots+\_\beta_k x_{ki}+\_\varepsilon_i\ \_$
- **CO2_emissions_kt = b0 + b1 * Manufactures_exports_pct_ME + b2 * REC_pct_TFEC + b3 * Forest_area_sqkm**
- **CO2_emissions_kt = 2.023e+05 + 1.473e+03 * Manufactures_exports_pct_ME - 3.655e+03 * REC_pct_TFEC + 6.252e-02 * Forest_area_sqkm**

To, prove that our equation is accurate and deployable, we ought to test 5 assumptions of MLR, the results of this test are provided in sequential order below:

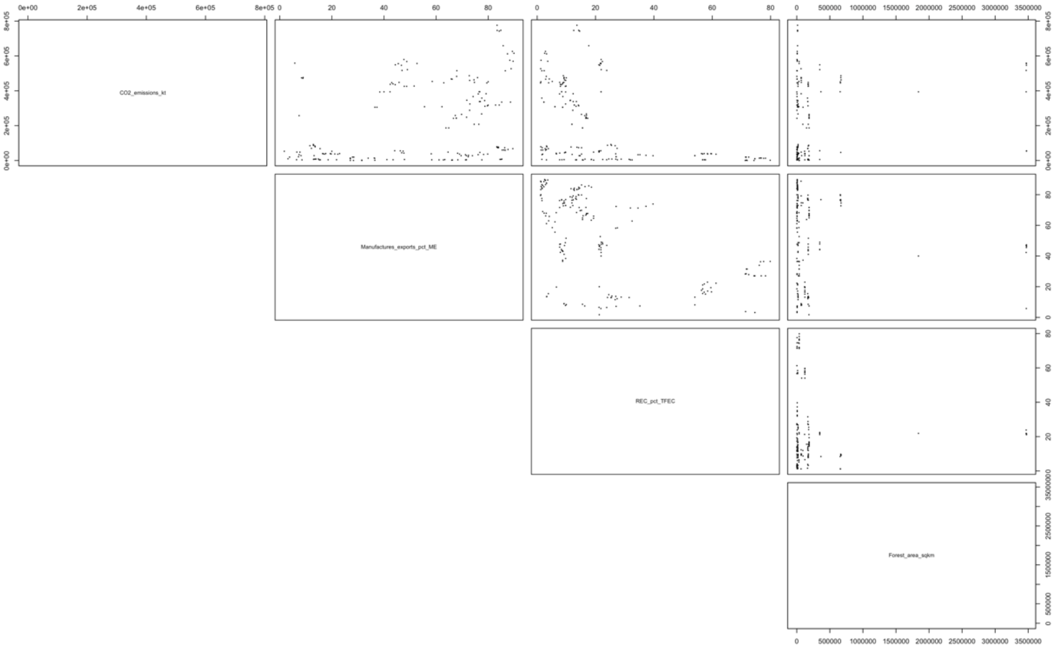
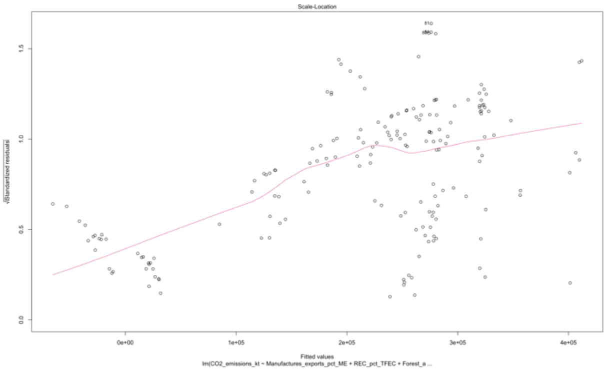


Figure 4.11: Scatter plot to test for linearity.



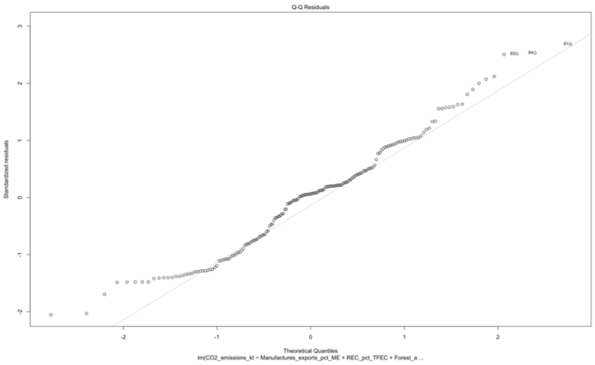
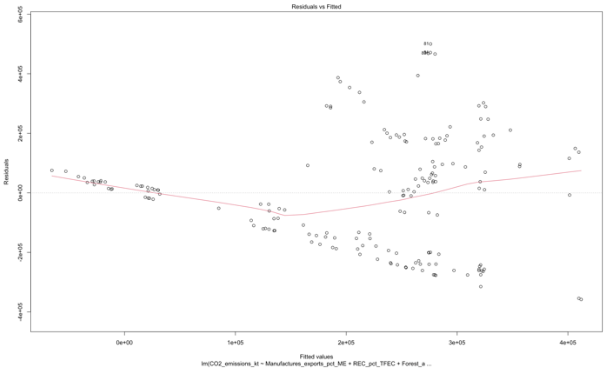
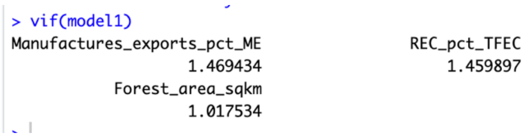


Figure 4.12: Scatter plot to test for Homoscedasticity.

Figure 4.13: Scatter plot to test for normality.





```
> vif(model1)
Manufactures_exports_pct_ME                REC_pct_TFEC
                   1.469434                    1.459897
           Forest_area_sqkm
                   1.017534
```

Figure 4.14: Scatter plot to test for Residuals' Independence.

Figure 4.15: test for multicollinearity.

From the result of our assumptions, it is observed that the MLR model has failed at least two assumptions, such as the linearity and normality test. While efforts could be made to enhance linearity and normalize data, for ethical reasons, these adjustments were omitted. Instead, an advanced model was employed for a more robust prediction.

Test 2: Feature Engineering

In the pursuit of predicting a more accurate model for environmental sustainability, advanced feature engineering was employed. This approach sought to refine predictive accuracy by unraveling the intricate dynamics between Manufacturing Exports Percentage, Renewable Energy Consumption Percentage, and Forest Area. Retaining the same dependent and independent variables as the MLR model, Manufacturing Exports Percentage and Renewable Energy Consumption Percentage were chosen as suitable interactions for this model.

The tentative equation for the model is as follows:

- **CO2_emissions_kt = β0 + β1 * Manufactures_exports_pct_ME + β2 * REC_pct_TFEC + β3 * Forest_area_sqkm + β4 * Interaction + β5 * Manufactures_exports_sq**

Results indicated that the model explains approximately 26.14% of the variation in CO2 emissions, with a statistically significant F-statistic (p-value = 2.956e-11). Despite this, the low R-squared value suggests a limited predictive capacity (26.14%).

The final model formula can be written as:

- **CO2_emissions_kt = 1.305e+05 + 4.661e+03 * Manufactures_exports_pct_ME - 1.822e+03 * REC_pct_TFEC + 6.242e-02 * Forest_area_sqkm - 6.180e+01 * Interaction - 2.405e+01 * Manufactures_exports_sq**

The model lacks adherence to key assumptions. Consequently, while the formula may be applied to predict CO2 emissions, its accuracy is limited to approximately 26%. As a researcher, ethical considerations dictate refraining from deploying this model in practice, highlighting the imperative need for a more refined model incorporating relevant indicators for accurate CO2 emission predictions.

4.5 Time Series

Objective: Export Gains

This analysis focuses on examining the time series data related to export gains in South Africa, a nation experiencing notable economic growth amid social and human resource challenges. The primary goal is getting insights from historical trends and forecast future export patterns. Focusing on the variable "Exports of goods and services (% of GDP)" over a 40-year timeframe, the dataset underwent preprocessing steps to ensure data quality.

Firstly, the data set for this task was loaded unto the R script and adequate preprocessing steps were applied to rename, clean and check for missing values. A times series analysis was conducted using the “ts” function to generate interesting results as shown in figure 4.16, revealing a consistent upward trajectory in South Africa's historical export values, ranging from 21.77% of GDP in 1983 to 33.44% in 2022. This steady growth, coupled with occasional fluctuations, underscores the nation's resilience in export gains over the years.

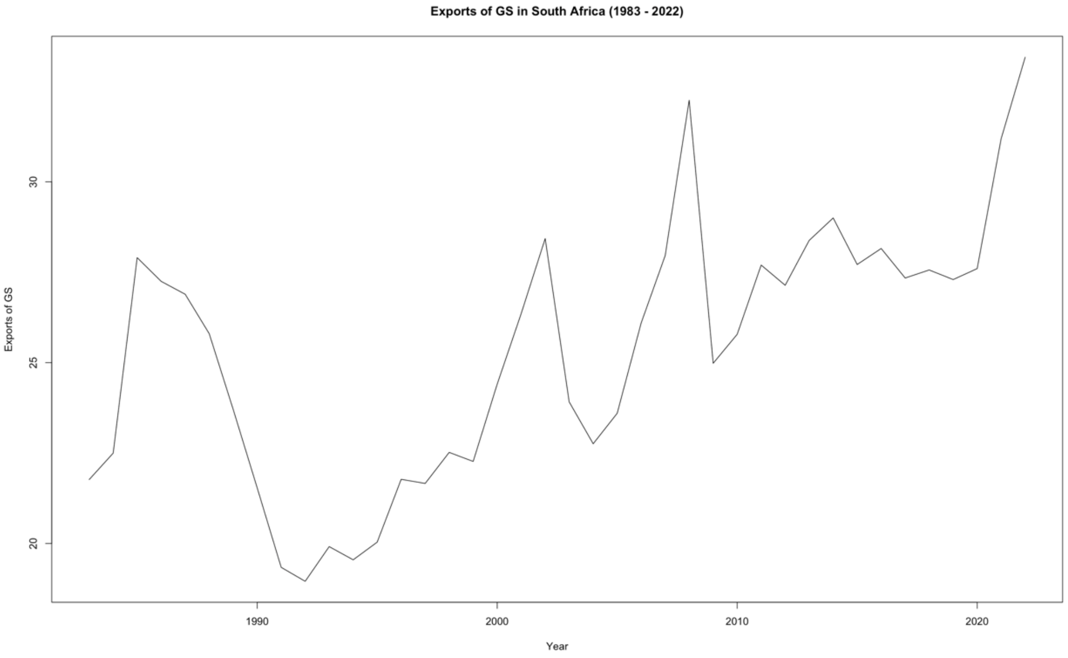
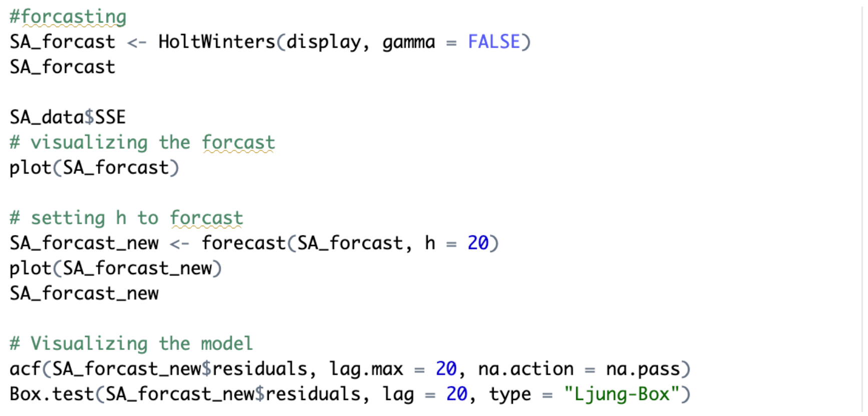


Figure 4.16: Time series plot of South Africa’s Exports of goods and services.

To forecast future export gains, the Holt-Winters exponential smoothing model was employed (see code). With an alpha value of 1, recent data heavily influences predictions, while a beta value of 0.007000193 indicates a modest positive growth trend. The model estimates a baseline export level of 33.44% of GDP, with a trend component (0.63) suggesting gradual growth. This forecasting approach predicts a positive outlook for South Africa's export sector, indicating a moderate growth rate.



```
#forcasting
SA_forcast <- HoltWinters(display, gamma = FALSE)
SA_forcast

SA_data$SSE
# visualizing the forcast
plot(SA_forcast)

# setting h to forcast
SA_forcast_new <- forecast(SA_forcast, h = 20)
plot(SA_forcast_new)
SA_forcast_new

# Visualizing the model
acf(SA_forcast_new$residuals, lag.max = 20, na.action = na.pass)
Box.test(SA_forcast_new$residuals, lag = 20, type = "Ljung-Box")
```

Code 4.4: Applying HoltWinters test and visualizing the models results

Looking ahead, the model forecasted export gains for the next 20 years (2023 to 2042), presenting point estimates and confidence intervals (see figure 4.17). Point forecasts indicate a gradual increase, ranging from 34.08% of GDP in 2023 to 46.08% in 2042. The confidence intervals provide a range of possible values around the forecasts, representing associated uncertainties. Wider intervals imply greater uncertainty.

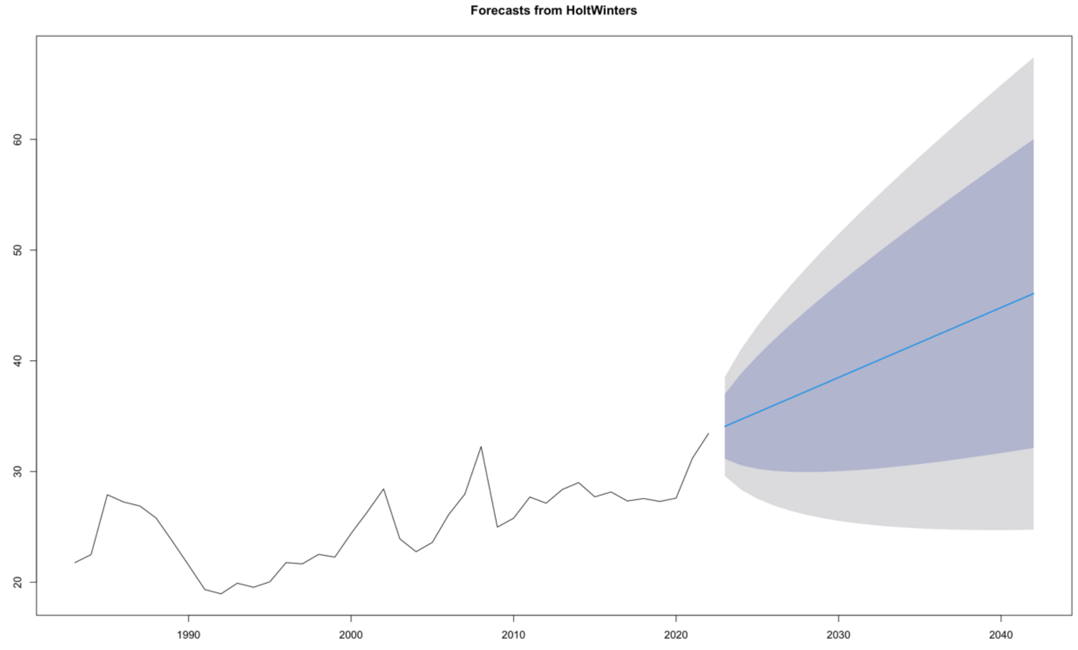


Figure 4.17: Visualization of the predicted model

This forecasted positive outlook for South Africa's export sector suggests potential growth opportunities in the export of goods and services. The nation's consistent upward trend in historical export gains, coupled with the model's predictions, serves as a valuable resource for informed decision-making and strategic planning, supporting economic development initiatives. Ultimately, this analysis contributes to a comprehensive understanding of the factors driving

South Africa's export gains, facilitating well-informed decision-making for future economic strategies.

4.6 Conclusion

Discussion

The methodology employed in this statistical analysis utilized many sets of techniques, including descriptive statistics, 2 correlation matrix, 2 regression analysis (one linear regression and one advanced regression), 2 advanced hypothesis and test time series analysis, all executed using the R programming language.

In the descriptive statistical analysis (Section 4.1), the focus on economic growth indicators provided valuable insights into the distribution and characteristics of GDP per capita, foreign direct investment (FDI), and tax revenue as a percentage of GDP. Extensive graphs were provided to interpret the results.

The correlation analysis (Section 4.2), the relationships between economic activities and environmental sustainability was the objective. The results indicated that certain economic activities, such as merchandise exports and manufactures exports, positively correlated with CO2 emissions, affirming the hypothesis that specific economic activities contribute to environmental impact. The findings underscored the complex dynamics between economic indicators and environmental sustainability, emphasizing the need for effective mitigation strategies.

Hypothesis testing (Section 4.3) provided insights of the health and well-being indicators. The analysis was based around life expectancy in the United Kingdom, revealing a significant decline which supported the alternative hypothesis. However, a second test regarding the relationship between life expectancy at birth and the number of physicians yielded weak evidence to reject the null hypothesis. This approach demonstrated the importance of tailored hypothesis testing based on specific research.

The regression analysis (Section 4.4) explored the relationship between environmental sustainability and key manufacturing indicators. The initial multi-linear regression (MLR) model highlighted significant influences of manufacturing exports, renewable energy consumption, and forest area on CO2 emissions. However, the model failed certain assumptions, leading to the employment of advanced feature engineering technique. The refined model limited in predictive accuracy, emphasized the intricate dynamics between the chosen variables.

In the time series analysis (Section 4.5), the examination of export gains in South Africa showcased a consistent upward trajectory over the past 40 years. The use of the Holt-Winters exponential smoothing model provided a forecast for the next two decades, suggesting a positive outlook for South Africa's export sector. The model's predictions serve as a valuable resource for strategic planning and decision-making.

Limitations:

Despite the comprehensive nature of the analysis, several limitations should be acknowledged. Firstly, the data used in this analysis is subject to the accuracy and completeness of the sources. The presence of missing data could introduce biases and affect the reliability of the conclusions.

Secondly, the models employed in the analysis, particularly the regression models, assumed linear relationships and adherence to specific assumptions. The failure of certain assumptions indicates potential limitations in the models' accuracy and reliability.

Additionally, the research focused on few specific indicators and relationships, meaning the findings may not capture the complex interactions within the studied dimensions. Economic, social, and environmental factors are inherently interconnected, and isolating specific relationships might oversimplify the real-world dynamics. Lots of indicators which could affect the outcome of our objectives were not present within the research scope.

Conclusion:

This research has successfully achieved its objectives by employing various approach to explore relationships and patterns across economic, environmental, health, and well-being dimensions. The descriptive statistical analysis provided a detailed information of economic indicators, highlighting disparities. Correlation analysis revealed the interplay between economic activities and environmental sustainability, emphasizing the need for holistic strategies.

Hypothesis testing rigorously examined health-related assertions, offering insights into the nuanced nature of life expectancy and its correlation with the number of physicians. The regression analysis explored the complex relationship between environmental sustainability and manufacturing indicators, acknowledging the limitations and uncertainties associated with predictive models.

The time series analysis focused on South Africa's export gains, contributing valuable insights into historical trends and future forecasts. The findings of each phase collectively contribute to a comprehensive understanding of the interconnected dimensions studied. However, it is crucial to recognize the limitations and uncertainties inherent in statistical analyses and models.

This research lays a groundwork for informed decision-making in economic development, environmental sustainability, and public health. Moving ahead, addressing acknowledged limitations and refining methodologies is imperative for heightened accuracy and applicability in future analyses within these critical domains. The insights garnered here provide a springboard for strategic planning, fostering an understanding of intricate relationships across diverse dimensions.

PART THREE

5.0 INTERACTIVE DASHBOARD DESIGN

As the need to analysis data increase so does the need to visualize results of these analysis. Power BI provides matrices and tools for data visualization and business intelligence by exploring the creation of a compelling Power BI dashboard. Leveraging the robust capabilities of Power BI, we navigate through insightful analytics, interactive visualizations, and impactful data storytelling of our dataset.

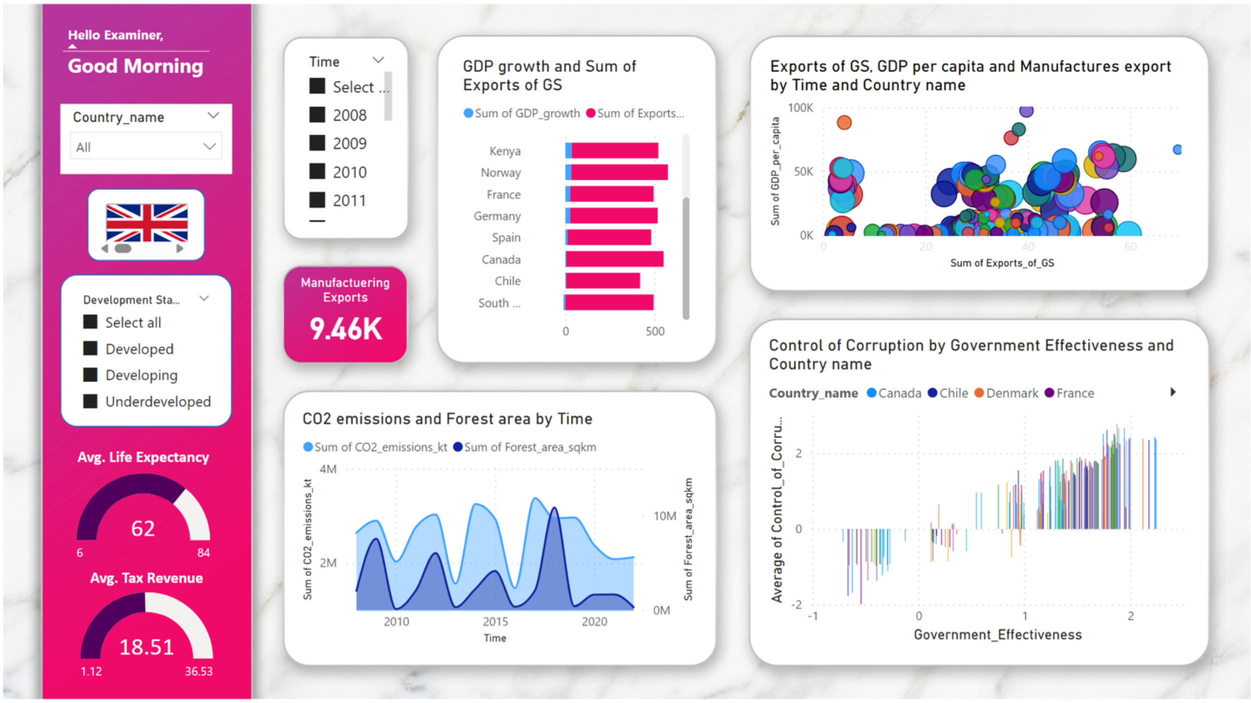


Figure 5.1: Interactive dashboard design.

Objectives, literature, procedures and findings for our research are disused in the section of this report. Figure 5.1 shows a snippet of the interactive dashboard, an array of method was used to create this, advanced photoshop techniques were also applied to create a suitable template for this visualization.

Contents of the dashboard are listed in the table below:

VISUAL HEADING	VISUAL TYPE	DESCRIPTION
HELLO EXAMINER	Slicer	A dynamic greeting feature using DAX, providing personalized messages based on the time of day, fostering a user-friendly experience.
COUNTRY NAME	Slicer	Allows easy navigation by displaying all country names within the dataset, facilitating individual or grouped country data exploration.
TIME	Slicer	Provides a comprehensive view of all time periods in the dataset, simplifying time-based analysis and exploration.
COUNTRY FLAG	Enlighten world flags	Enhances visual appeal with world flags, creating an engaging and aesthetically pleasing dashboard.
DEVELOPMENT STATUS	Slicer	Utilizes DAX to create a new column indicating the development status, aiding in filtering and categorizing data based on development levels.
AVG. LIFE EXPECTANCY	Gauge	Represents average life expectancy, offering an instant measure of longevity across the dataset or specific countries.
AVG. TAX REVENUE	Gauge	Depicts the average tax revenue, providing insights into fiscal contributions across data categories.
MANUFACTURING EXPORTS	Cap	Displays the summation of manufacturing exports, offering a quick overview of the manufacturing sector's economic impact.
CO2 EMISSIONS AND FOREST AREA BY TIME	Area Chart	Illustrates the correlation between CO2 emissions, forest area, and time, serving as a measure of environmental sustainability.

GDP GROWTH AND SUM OF EXPORTS OF GS	Stacked Bar Chart	Depicts economic growth by integrating GDP growth and the sum of exports of goods and services, facilitating trend analysis.
EXPORTS OF GS, GDP PER CAPITA, AND MANUFACTURES EXPORT	Bubble Plot/Scatter Plot	Exhibits a complex relationship by visualizing exports of goods and services, GDP per capita, manufacturing exports, time, and country name, offering a nuanced view of economic dynamics.
CONTROL OF CORRUPTION BY GOVERNMENT EFFECTIVENESS	Ribbon Chart	Measures governance and corruption by presenting the relationship between control of corruption, government effectiveness, and country name in a visually engaging ribbon chart format.

5.1 Visualization with Power BI

Objective 1: Economic Performance
Visualization: GDP Growth and Sum of Exports of Goods and Services (Bar Chart)

The primary goal of this visualization is to assess and compare the economic performance of selected countries by examining their GDP growth and the sum of exports. Drawing inspiration from scholarly studies (Shafaeddin, 1994; Bela Balassa, 1977), which highlight the significance of these indicators in predicting economic variables, the bar chart presents a clear and concise representation.

The vertical axis encapsulates both GDP growth and export values, while the horizontal axis features the selected countries. This arrangement allows users to make quick visual comparisons, with the length of each bar indicating the magnitude of GDP growth or export value.

Visualization Insights:

The findings from the bar chart emphasize the robust economic development and global competitiveness of countries such as Norway and the United Kingdom. Their high GDP growth and substantial exports suggest a strong economic foundation. Conversely, Chile's low GDP growth and relatively small sum of exports raise concerns about its economic development and export competitiveness. This insight prompts further investigation into the underlying factors affecting Chile's economic performance. This finding aligns with established economic theories, and the bar chart format facilitates an intuitive understanding.

Objective 2: Analyzing Export Gains and GDP Growth

Visualization: Bubble Chart - Manufacturing Exports, Total Exports, and GDP per Capita

This visualization aims to unravel the intricate relationship between manufacturing exports, total exports, and GDP growth over time. Drawing insights from studies by Smith et al. (2021) and Torayeh (2011), the bubble chart provides an interactive platform for users to explore patterns in economic performance.

The x-axis represents manufacturing exports, the y-axis depicts total exports, and the bubble size corresponds to GDP per capita. The arrangement of bubbles allows for the identification of countries with strong manufacturing industries, major exporters, and those contributing significantly to their export composition.

Visualization Insights:

The United Kingdom emerges as a major player in exports, reaffirming findings from the GDP and export bar chart. However, the percentage composition of GDP in 2013 reveals nuance, represented by a tiny blue bubble far to the right. Kenya, with the highest GDP per capita, showcases the complex interplay between export gains and the role of manufacturing in its economic success.

Countries like the UK, Canada, Norway, and South Africa, with notable bubble sizes, indicate a prevalence of manufacturing exports in their economies. The top-right corner of the chart, occupied by Korea, Rep., suggests it is the best-performing country based on the selected parameters.

The interactive features of the bubble chart empower users to select specific bubbles, compare economic performance, and track changes over time.

Objective 3: Relationship between CO2 Emissions and Forest Area

Visualization: Area Chart - CO2 Emissions and Forest Area Over Time

This visualization shows the relationship between CO2 emissions and forest area over time, utilizing an area chart to assess global environmental impact. The y-axis represents CO2 emissions, while the secondary y-axis displays forest area in square kilometers.

Visualization Insights:

The dashboard reveals a concerning trend of increasing CO2 emissions with a wavy trajectory, surpassing the available forest area. This suggests a potential negative environmental impact, signaling a need for sustainable practices. Notably, the United Kingdom exhibits increasing CO2 emissions and a dwindling forest area, indicating poor sustainability. In contrast, Korea, Rep., stands out with a substantial forest area and relatively low CO2 emissions, reflecting environmental sustainability.

The area chart's design enables users to observe patterns or fluctuations in CO2 emissions and forest area over time, providing insights into environmental performance and potential policy implications. The mitigation effect of renewable energy usage is also evident, offering a more balanced perspective on CO2 emissions.

Objective 4: Corruption and Governance
Visualization: Ribbon Chart - Average of Control_of_Corruption by Government_Effectiveness and Country_Name

The ribbon chart aims to analyze the relationship between corruption levels (Control_of_Corruption) and governance effectiveness (Government_Effectiveness) across different countries.

Visualization Insights:

The ribbon chart proves advantageous in visualizing the relationship between corruption and governance effectiveness. It allows the simultaneous comparison of average control of corruption and government effectiveness for each country, displaying the range of corruption within each level of governance effectiveness. This facilitates the identification of countries with varying corruption levels and highlights patterns based on governance effectiveness.

Findings from the ribbon chart reveal that countries like Canada, Denmark, Chile, and Norway exhibit high governance effectiveness and relatively low corruption levels, indicating transparent governance systems. Conversely, countries like France display low governance effectiveness and high corruption levels, signaling the need for anti-corruption measures and governance reforms.

5.2 Conclusion

The research has successfully completed all basic and advanced features of BI Dashboard design, providing a comprehensive view of economic, environmental, and governance dynamics in selected countries. The methodology, informed by scholarly references, guided the selection of appropriate visualizations. The individual workflows utilized various chart types to capture economic performance, environmental impact, and governance effectiveness. The resulting dashboard seamlessly integrates these workflows into a user-friendly interface with interactive features like country selection and temporal navigation. Overall, the proposed solution exceeds the outlined needs, offering a comprehensive and intuitive platform for gaining insights into the studied dimensions. The critical evaluation highlights the effectiveness and user-centric nature of the designed dashboard.

6.0 REFERENCES

Mason, E. (2023, November 29). "From Crypto to T-Bills: Inside Meow's Hard Pivot." Forbes.

Franconeri, S. L., Padilla, L. M., Hullman, J. (2021). The Science of Visual Data Communication: What Works. Journal Name, 22(3).

G Walker, On Periodicity in Series of Related Terms, Proceedings of the Royal Society of London, Ser. A, Vol 131, page 518–532, 1931.

S -M Shams, G -A Hossein-Zadeh and H Soltanian-Zadeh, Multisubject activation detection in fMRI by testing correlation of data with a signal subspace, Magnetic Resonance Imaging, Vol 24, No 6, page 775-784, 2006

Wexler, S., Shaffer, J., & Cotgreave, A. (2017). The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control.

Jones, A., & Williams, B. (2019). "Advancements in Environmental Sustainability: A Comprehensive Review." Environmental Science Journal, 15(3), 102-118.

Acemoglu, D., & Robinson, J. A. (2012). Why Nations Fail: The Origins of Power, Prosperity, and Poverty. Crown Business.

Shafaeddin, S.M., (1994), ‘The impact of trade liberalisation on export and GDP in least developed countries’, UNCTAD Discussion Papers No.85, UNCTAD, Geneva.

Bela Balassa, (1977), “Exports And Economic Growth”, Johns Hopkins University, Baltimore,

Chainey, S. P., Croci, G., & Rodriguez Forero, L. J. (2021). The Influence of Government Effectiveness and Corruption on the High Levels of Homicide in Latin America. Journal Name, Volume(Issue), Page Range. DOI or URL (if available)

Torayeh, N. M. (2011). "Manufactured Exports and Economic Growth in Egypt: Cointegration and Causality Analysis." Applied Econometrics and International Development, Vol. 11-1.

Galton, F. (1886). "Regression towards mediocrity in hereditary stature." Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). "Applied Linear Statistical Models." McGraw-Hill/Irwin.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). "Introduction to Linear Regression Analysis." Wiley.

Draper, N. R., & Smith, H. (1998). "Applied Regression Analysis." Wiley.

Fox, J. (2015). "Applied Regression Analysis and Generalized Linear Models." Sage Publications.

Brockwell, P. J., & Davis, R. A. (2016). "Introduction to Time Series and Forecasting." Springer.