Introduction and Data Preparation

1. Introduction

This research and analysis is aimed to dive into health systems and dynamic practices in various countries; from highincome to developing countries, primarily highlighting the interplay between population and its age ranges, fiscal allocation (government, private and external), mortality, and health system equity and efficacy. This cut-across objective is to identify patterns, disparity, and correlation as well as provide a zoomed-in perspective of global health systems to inform both policies and best practices as well as data scientists focused on economics.

In the analysis, a mix of qualitative as well as quantitative, and domain knowledge is employed as the strategy for approaching the problem, to show a holistic position of these versatile health systems and how the indicators we will discuss all interplay across varying socio-economic contexts.

This analytical research is driven down into tasks that sum up to the holistic approach. These tasks have been listed to include the objective for each task, yet aligned with the central objective. They are listed below:

1. Comprehensive Descriptive Analysis:

1a. The objective is to show a holistic statistical summa of key health and population indicators across different countries. These calculations to determine measures- like median, mode, standard deviation, skewness, and kurtosis are used to understand the distribution and properties of data related to health expenditure, population age structures, and mortality.

In addition, boxplots for key indicators were also explored. The boxplots and histograms provide a comprehensive overview of critical metrics.

1. Correlation Analysis:

2a. The objective here is to identify and quantify the direction the strength and direction of relationships between various health and population variables. This involves analyzing how factors like age dependency ratios, health expenditure (private and government), and physician indicators are interrelated, using statistical methods such as Spearman's rank correlation, Confidence Interval, etc.

1. Hypothesis Testing:

3a. Testing hypotheses from different statistical statements was the objective of this analysis, such as the impact of age dependency on health expenditure as well as the viability of gender and locationbased impacts. This task utilizes statistical tests like T-tests and Mann-Whitney U tests to accept or reject these hypotheses.

1. Regression Analysis:

4a. This objective is aimed at exploring the impact of population and economic variables on health indicators while examining how this impact may influence mortality and life expectancy. This includes using regression techniques, such as backward stepwise regression, to determine the most significant predictors and understand their influence.

1. Time Series Analysis:

4a. This objective is aimed at analyzing the trends in primary indicators like life expectancy, mortality, population in different age ranges, and health expenditure. This task aims to understand how these variables evolve over the years. This includes using regression techniques, such as ARIMA and ETS models.

Background Research and Literature Review

Background Research

Descriptive Statistical Analysis

The basics of any data-driven analytic research is Descriptive statistics, offering a foundational understanding between datasets. As stated by (Kaur et al., 2018), descriptive statistics such as mean, median, mode, standard deviation, skewness, and kurtosis provide condensable insights into the central tendency, spread, and shape of the data distribution. In our analysis, these statistical tools were applied to healthcare data across nations to grasp the central trends in our objectives, aligning with best practices.

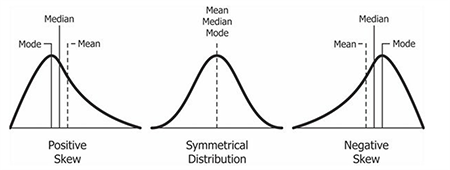


Figure 1: Descriptive Analysis (Danielle & ’Chad, 2023)

Mean, Median and mode are the most common measures of central tendency.

Mean: The mean is also referred to as “Average”. It is calculated by summation of the list of numbers divided by the number count of the numbers

Median: This is the middle number. It is the middle number when the sets of numbers to be calculated have been sorted from smallest to largest.

Further descriptive analysis can be used using:

Standard Deviation: In simple terms, this measures how varied a group of values are from the mean. In that when the values are close to the mean, the standard deviation is said to be low but if it spreads out over a wider range, then the standard deviation is said to be high.

Skewness: The level of asymmetry when normally distributed data is viewed in a symmetrical bell-shaped curve is called Skewness. Depending on what side the curve is titled, Skewness can either be negative (when titled towards the left) or positive (when titled towards the right), (MacGillivray, 1986)

Kurtosis: The degree of clustering at the tail or peak of a frequency distribution is the Kurtosis. When Kurtosis is positive, they are called “Platyjurtic” distribution and “Lepokurtic” distribution when they are negative.(Kim, 2013)

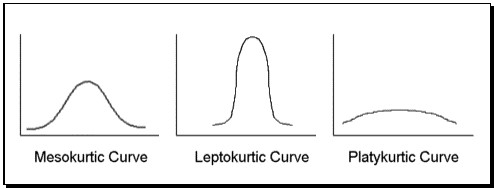
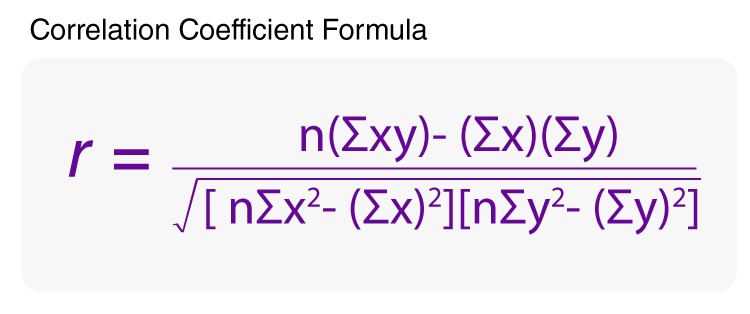


Figure 2: Kurtosis Curves (Saul, 2023)

Correlation Analysis

Correlation analysis is used to measure and understand the strength and direction of the relationship between variables. It is a measure of association between variables(Schober et al., 2018). Correlations may be negative or positive and are measured from +1 to -1, with 1 indicating a total correlation (i.e. one variable increases as the other does so), and vice versa for -1. Our study explores the Spearman correlation coefficient to evaluate correlations between different relevant indicators. This provided a good base to identify potential relationships such as between age dependency ratios and health expenditures, thus guiding further in-depth analysis.

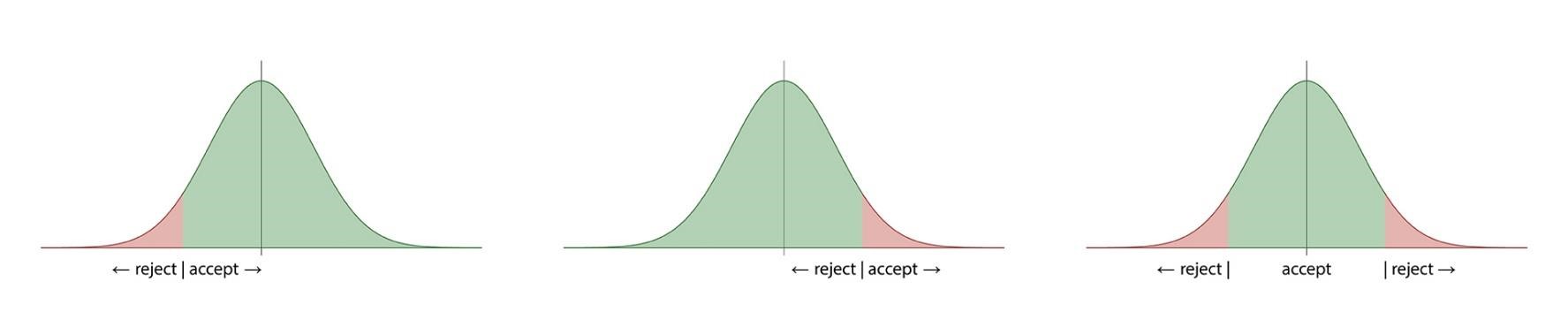


***Figure 3: Correlation Coefficient Formula***

Hypothesis Testing

Hypothesis testing is a critical component of inferential statistics used to make decisions or inferences about population data- it is used to determine if the assumption made by the statistician can be accepted or rejected (i.e. determining its validity). So how does this work? -The significance level. The null hypotheses would be accepted in the case where the significance value of the test is greater than the level that was initially defined. For the alternate hypotheses to be accepted, the inverse is the case this study utilized various hypothesis testing techniques, including T-tests, Mann-Whitney U tests, and Wilcoxon tests to investigate specific research hypotheses related to gender disparities in health outcomes and the impact of out-of-pocket expenditure on healthcare (Fay & Proschan, 2010) Figure 4: Hypothesis Curves (Steven, Walker n.d.,2023)

Regression Analysis

Regression analysis, as detailed by (van Oordt, 2015) is game-changing in predicting and forecasting outcomes based on independent variables. This study applied both linear and polynomial regression models to explore how population shifts and financial dynamics impact health system outcomes. The inclusion of polynomial terms, as suggested by (*Understanding Polynomial Regression Model - Analytics Vidhya*, n.d.) allowed for a more detailed understanding of non-linear relationships, a methodological approach that has been gradually adopted in healthcare research.

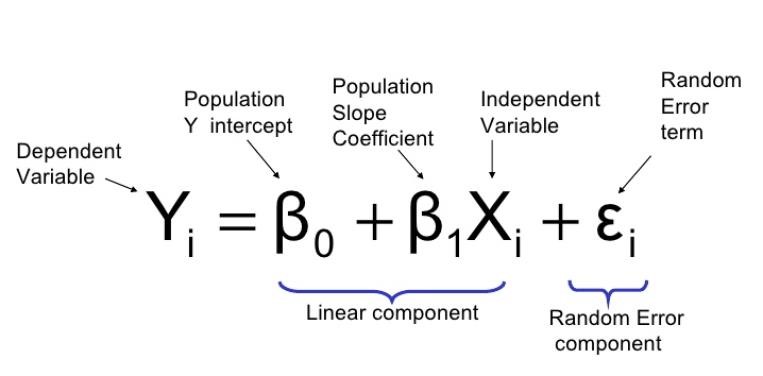


Figure 5: (How Are Logistic Regression & Ordinary Least Squares Regression (Linear Regression) Related? Why the “Regression” in Logistic? | by Rakshith Vasudev | Towards Data Science, n.d.)

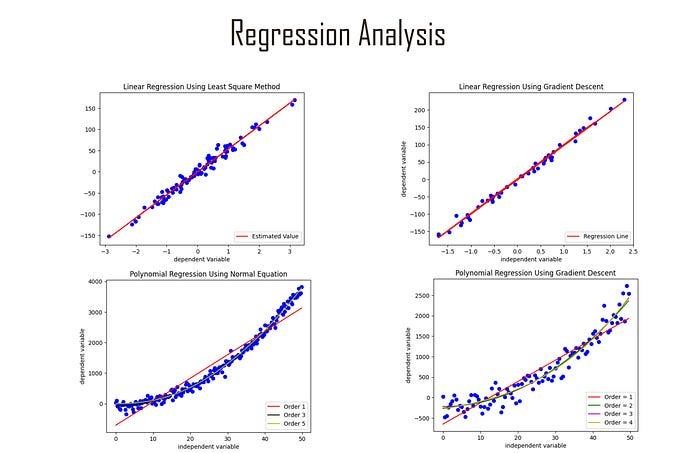


Figure 6: (Regression Analysis. Regression Analysis Models Explained… | by Anas BRITAL | Medium, n.d.)

Visualization of regression analysis serves as a tool for selecting the most appropriate model for any analysis based on the underlying relationship between variables.

Time Series Analysis

Time series analysis is an important statistical technique that focuses on analyzing data collected over defined time intervals. The data collected over a defined and regular timeline is crucial for identifying underlying trends, seasonal dynamics, and cyclical patterns within the data. In the context of our health expenditure analysis, this approach has proved meaningful in understanding how health-related metrics such as expenditure per capita evolve and what future trends may actually look like.

For this analytical research, the Autoregressive Integrated Moving Average (ARIMA) model was employed, which is mostly effective for understanding and forecasting time-based data when the dataset exhibits non-stationarity. It marries methods for differencing data to induce stationarity. The ARIMA model's efficiency in handling a variety of time series patterns is quite detailed in the work by (*Box-Jenkins Model: Definition, Uses, Timeframes, and Forecasting*, n.d.) This model has been selected for its adaptability in modeling complex data-behaviors in time series data, which is often encountered in health expenditure trends like ours. Complementing the ARIMA model, the Exponential Smoothing State Space Model (ETS), as discussed by (Shi, 2022) was used to capture the level, trend, and seasonality in the health expenditure data. The ETS model applies exponential smoothing to forecast future data points, which makes it particularly suitable for the complexity of the dataset of interest that may show non-linear patterns not easily captured by ARIMA.

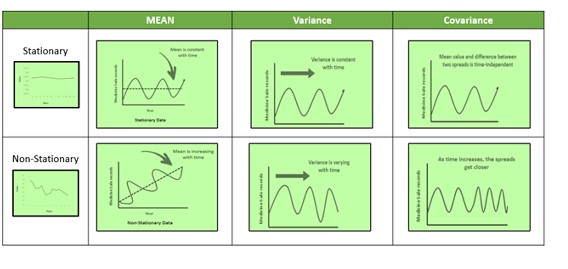


Figure7: (Time Series Analysis and Forecasting | Data-Driven Insights, n.d.)

Studies by (Box-Jenkins Model: Definition, Uses, Timeframes, and Forecasting, n.d.; Otu et al., 2014; Shi, 2022) highlight the importance of using robust statistical tools to accurately forecast time series data. The analytical research utilization of ARIMA and ETS models provided a dual-lens perspective, ensuring a comprehensive analysis of the time series data.

Ultimately, these high-plot analyses provided important insights into the health expenditure trajectory of the nations under evaluation. This may not only the historical patterns and fluctuations but may also offer reliable projections for future periods.

# Integration of Analyses in Research

The combination of these statistical methodologies supplies a holistic and robust analysis of the healthcare data. This integrated approach is in line with modern statistical practices in health economics and policy research, where multiple methods are employed to gain a complete picture of the data, as advocated by researchers like (Barbu, 2023) This multifaceted and versatile approach ensures that findings are accurate, and reflective of the complex realities of global health systems.

Task 2:

Dashboard Design Overview in Data Visualization

In various ecosystems, like healthcare analytics, the design and development of interactive dashboards have become a very important and game-changing aspect of data visualization. Dashboards present a concise way to present highlevel data for easier decision-making- for both technical and non-technical members of a business.

Effective dashboards should be easy to interact with as well as provide functional insights for the user. The approach of usability and functionality whilst providing a good UX aligns with the principles outlined by (Matheus et al., 2020), highlighting the importance of clear, effective, and honest data communication.

# A viable tool for Dashboard creation: Power BI

Power BI by Microsoft offers an encompassing platform for building interactive and dynamic dashboards. Power BI boasts of the ability to handle large and complex datasets as well as various kinds of data residing in different data stores. PBI (Power BI) possesses the DAX feature (Data Analysis Expression), making it even more suitable for complex data such as is being used in this case. Its integration with other analytics and project management tools makes it a top player in the analytics industry. Its ease of use also comes as an advantage.

# Incorporation of Advanced Features in Dashboard Design

For this analytical research, the use of advanced features including DAX measures and data model relationships was employed to enhance the analytical capacity of the dashboard and offer all encompassing user experience, while offering the opportunity for users to offer their own customers breathtaking user experience within their dashbaord as detailed in this article: (*The Importance of User Experience in Reporting and Power BI - A Must-Read Guide!*, n.d.). The application of these advanced features in the dashboard is consistent with best practice according to leaders in this domain -@@@@- as they emphasize the importance of leveraging on the wholesome potential of power BI for more efficient analyses.

# Design Principles

Dashboard design should focus on clarity, minimalism and relevance. There are best practice dos and don’ts worthy of noting as emphasized by \_@@@@-Insert dashboard design image--. Dashboards must satisfy both user experience and user’s pre-determined function (i.e the business aim of the dashboard)

# Data Processing and Preparation for PBI Analysis

Data processing and preparation was explained by (*Power BI Data Preparation in 5 Steps*, n.d.) is a pertinent step for dashboard dev. Its primarily involves data cleaning, transformation and integration to ensure data accuracy and consistency.

# PBI; Features and Application in Healthcare Data Analytics

The processes of evaluating policy impacts, monitoring healthcare indicators as well as aiding strategic planning as been made more efficient with error- probability as low as ever doo due to the increasing use of interactive dashboards in healthcare.

The dashboard in this study aligns with these utilizations, whilst aiding a well-rounded examination of healthcare expenditure, population, and service dynamics across the countries of our interest.

In summary, the inclusion of PBI dashboard in this analytical research really offers the required balance to the entire research, as it stands at the meeting point of health care analysis and advanced visualization techniques. It epitomizes the principles of data presentation that is effective and interactive analysis; emphasizing its importance in pushing data-driven decisions in the complex domain of global health

Literature Review

# Health Expenditure and Demographic Influences

As the proportion of senior individuals within a demographic increases, the demand of all facets of health care services also increases, thereby elevating the health expenditure across board.

The relationship between health expenditure and population changes is a very enlightening analytical discuss in the field of health economics according to studies of (Beylik et al., 2022)

(Zhang et al., 2015) highlights the role of changing demographic structures including age dependency ratio in shaping the patterns of expenditure. Their study infers that countries will face increased healthcare cost when they have a higher dependency ratio within the entire population. This may be attributed to a larger non- population who will require health services

# Gender Imbalance in Health Outcomes

Gender imbalance in health outcomes has been highlighted in this study (Milner et al., 2021)recognized the genderbased differences in the occurrence of certain decisions and healthcare utilization trends, further highlighting the importance of gender consideration in healthcare policy planning. In addition, during our study as well as highlighted research, (Danielsen et al., 2022)there seems to be a significant difference in the life expectancy between genders across various demographics. Some of these imbalances have been attributed to biological factors and differential access to healthcare.

# Healthcare Services Efficacy

Healthcare services efficacy vis a vis fiscal allocation is an area of sensitive study (Mosadeghrad, 2014) Research from various schools of though,t (Papanicolas et al., 2018) have analyzed the correlation between high spending and better healthcare. Higher spending does not always mean better spending. Paradoxically, the efficient allocation of resources may prove more effective as the importance of quality over quantity is even more emphasized.

# Out-of-Pocket Expenditure Impact

The impact of out-of-pocket expenditure can lead to pulling households below the poverty line. Increasing significant financial strain in particularly low to middle-income countries. Studies suggests a really strong correlation between out-of-pocket expenditure and increased risk of financial catastrophe. (Sirag & Mohamed Nor, 2021)

# Comparative Analysis Across Countries

Research comparing different health systems and expenditures across different countries delivers relevant insights into global health transformations. (Basu et al., 2012)highlights the difference of healthcare models across various countries whilst inquiring into the versatile approach to funding, delivery of healthcare service and organization and how they affect the overall health outcomes.

# Literature Review Conclusion

This literature review spotlights the complex interaction between demographic factors, healthcare expenditure, and service effectiveness. It also points out the immense need for continuous adaptation to the ever-changing economic realities.

This current analytical research aims to contribute to this river of knowledge by providing a comprehensive analysis of these dynamics using statistical techniques and intuitive dashboard design. This analysis also seeks to draw the attention of data collectors to investigate optimizing data collection processes for low-income countries as the available data disparity is extremely wide in the different socio-economic climates.

3. Preparation and Exploration of Data Set

The dataset used to bring this analysis alive was sourced from [World Bank Source(](https://databank.worldbank.org/source/world-development-indicators)*World Development Indicators | DataBank*, n.d.). A total of forty-five (45) indicators and twenty-three (23) countries- spanning from high to middleincome countries as aggregated in the World Bank Data were selected for this analysis. The datasets analyzed were datasets available from 2010-2021.

For the Power BI, in addition to the primary dataset highlighted above, a similar secondary dataset was included in the model used for the analytical visualization. This secondary dataset is also within the theme of discussion, and it carried a total of four (4) related indicators with the same countries and the same period highlighted above.

|  |  |
| --- | --- |
| **Indicator Name** | **Indicator Definition** |
| Age dependency ratio (% of working-age population) | Age dependency ratio is the ratio of dependents--people younger than 15 or older than 64--to the working-age population--those ages 15-64. |

|  |  |
| --- | --- |
| Age dependency ratio, old (% of working-age population) | Age dependency ratio, old, is the ratio of older dependents-people older than 64--to the working-age population--those ages 15-64. |
| Age dependency ratio, young (% of workingage population) | Age dependency ratio, young, is the ratio of younger dependents--people younger than 15--to the working-age population--those ages 15-64. |
| Birth rate, crude (per 1,000 people) | Crude birth rate indicates the number of live births occurring during the year, per 1,000 population |
| Current health expenditure (% of GDP) | Level of current health expenditure expressed as a percentage of GDP. Estimates of current health expenditures include healthcare goods and services consumed during each year. |
| Current health expenditure per capita (current  US$) | Current expenditures on health per capita in current US dollars. Estimates of current health expenditures include healthcare goods and services consumed during each year. |
| Death rate, crude (per 1,000 people) | Crude death rate indicates the number of deaths occurring during the year |
| Domestic general government health expenditure (% of GDP) | Public expenditure on health from domestic sources as a share of the economy as measured by GDP. |
| Domestic general government health expenditure (% of general government expenditure) | Public expenditure on health from domestic sources as a share of total public expenditure. It indicates the priority of the government to spend on health from own domestic public resources. |
| Life expectancy at birth, female (years) | Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. |
| Life expectancy at birth, male (years) | Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. |
| Life expectancy at birth, total (years) | Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. |
| Mortality caused by road traffic injury (per  100,000 population) | Mortality caused by road traffic injury is estimated road traffic fatal injury deaths per 100,000 population. |

|  |  |
| --- | --- |
| Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70 (%) | Mortality from CVD, cancer, diabetes or CRD is the percent of 30-year-old-people who would die before their 70th birthday from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease, assuming that s/he would experience current mortality rates at every age and s/he would not die from any other cause of death (e.g., injuries or HIV/AIDS). |
| Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, female (%) | Mortality from CVD, cancer, diabetes or CRD is the percent of 30-year-old-people who would die before their 70th birthday from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease, assuming that s/he would experience current mortality rates at every age and s/he would not die from any other cause of death (e.g., injuries or HIV/AIDS). |
| Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, male (%) | Mortality from CVD, cancer, diabetes or CRD is the percent of 30-year-old-people who would die before their 70th birthday from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease, assuming that s/he would experience current mortality rates at every age and s/he would not die from any other cause of death (e.g., injuries or HIV/AIDS). |
| Out-of-pocket expenditure (% of current health expenditure) | Share of out-of-pocket payments of total current health expenditures. Out-of-pocket payments are spending on health directly out-of-pocket by households. |
| Physicians (per 1,000 people) | Physicians include generalist and specialist medical practitioners. |
| Population ages 0-14 (% of total population) | Population between the ages 0 to 14 as a percentage of the total population. Population is based on the de facto definition of population. |
| Population ages 0-14, female (% of female population) | Female population between the ages 0 to 14 as a percentage of the total female population. Population is based on the de facto definition of population. |
| Population ages 0-14, male (% of male population) | Male population between the ages 0 to 14 as a percentage of the total male population. Population is based on the de facto definition of population. |
| Population ages 00-04, female (% of female population) | Female population between the ages 0 to 4 as a percentage of the total female population. |
| Population ages 00-04, male (% of male population) | Male population between the ages 0 to 4 as a percentage of the total male population. |

|  |  |
| --- | --- |
| Population ages 10-14, female (% of female population) | Female population between the ages 10 to 14 as a percentage of the total female population. |
| Population ages 10-14, male (% of male population) | Male population between the ages 10 to 14 as a percentage of the total male population. |
| Population ages 15-19, female (% of female population) | Female population between the ages 15 to 19 as a percentage of the total female population. |
| Population ages 15-19, male (% of male population) | Male population between the ages 15 to 19 as a percentage of the total male population. |
| Population ages 15-64 (% of total population) | Total population between the ages 15 to 64 as a percentage of the total population. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. |
| Population ages 15-64, female (% of female population) | Female population between the ages 15 to 64 as a percentage of the total female population. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. |
| Population ages 15-64, male (% of male population) | Male population between the ages 15 to 64 as a percentage of the total male population. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. |
| Population ages 20-24, female (% of female population) | Female population between the ages 20 to 24 as a percentage of the total female population. |
| Population ages 20-24, male (% of male population) | Male population between the ages 20 to 24 as a percentage of the total male population. |
| Population ages 30-34, female (% of female population) | Female population between the ages 30 to 34 as a percentage of the total female population. |
| Population ages 30-34, male (% of male population) | Male population between the ages 30 to 34 as a percentage of the total male population. |
| Population ages 35-39, female (% of female population) | Female population between the ages 35 to 39 as a percentage of the total female population. |
| Population ages 35-39, male (% of male population) | Male population between the ages 35 to 39 as a percentage of the total male population. |
| Population ages 70-74, female (% of female population) | Female population between the ages 70 to 74 as a percentage of the total female population. |
| Population ages 70-74, male (% of male population) | Male population between the ages 70 to 74 as a percentage of the total male population. |
| Population ages 75-79, female (% of female population) | Female population between the ages 75 to 79 as a percentage of the total female population. |
| Population ages 75-79, male (% of male population) | Male population between the ages 75 to 79 as a percentage of the total male population. |

|  |  |
| --- | --- |
| Population ages 80 and above, female (% of female population) | Female population between the ages 80 and above as a percentage of the total female population. |
| Population ages 80 and above, male (% of male population) | Male population between the ages 80 and above as a percentage of the total male population. |
| Population, female (% of total population) | Female population is the percentage of the population that is female. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. |
| Population, male (% of total population) | Male population is the percentage of the population that is male. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. |
| Population, total | Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. |
| **POWER BI ADDITIONAL INDICATORS BELOW** | |
| Risk of catastrophic expenditure for surgical care (% of people at risk) | The proportion of population at risk of catastrophic expenditure when surgical care is required. Catastrophic expenditure is defined as direct out of pocket payments for surgical and anaesthesia care exceeding 10% of total income. |
| Proportion of population pushed below the 60% median consumption poverty line by outof-pocket health expenditure (%) | This indicator shows the fraction of a country’s population experiencing out-of-pocket health impoverishing expenditures, defined as expenditures without which the household they live in would have been above the 60% median consumption but because of the expenditures is below the poverty line. |
| Coverage of social insurance programs (% of population) | Coverage of social insurance programs shows the percentage of population participating in programs that provide old age contributory pensions (including survivors and disability) and social security and health insurance benefits (including occupational injury benefits, paid sick leave, maternity and other social insurance). Estimates include both direct and indirect beneficiaries. |
| Coverage of social safety net programs (% of population) | Coverage of social safety net programs shows the percentage of population participating in cash transfers and last resort programs, noncontributory social pensions, other cash transfers programs (child, family and orphan allowances, birth and death grants, disability benefits, and other allowances), conditional cash transfers, in-kind food transfers (food stamps and vouchers, food rations, supplementary feeding, and emergency food distribution), school feeding, other social assistance programs (housing allowances, scholarships, fee waivers, health subsidies, and other social assistance) and public works programs (cash for work and food for work). Estimates include both direct and indirect beneficiaries. |

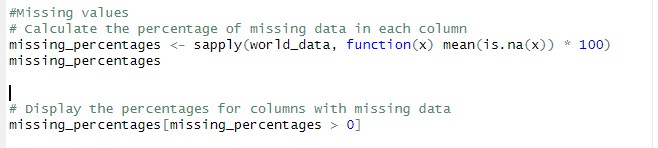
Missing Data

On missing data, some percentages of missing data were present within the dataset. They are outlined in the table below:

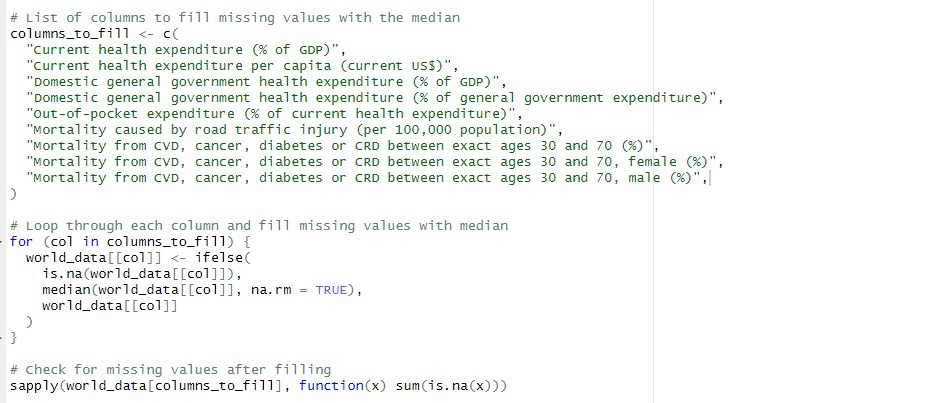
|  |  |
| --- | --- |
| **Indicators with Missing Values** | **% of**  **Missing**  **Values** |
| Current health expenditure (% of GDP) | 6.88 |
| Current health expenditure per capita (current US$) | 6.88 |
| Domestic general government health expenditure (% of GDP) | 7.61 |
| Domestic general government health expenditure (% of general government expenditure) | 7.61 |
| Mortality caused by road traffic injury (per 100,000 population) | 16.67 |
| Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and  70 (%) | 16.67 |
| Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and  70, female (%) | 16.67 |
| Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and  70, male (%) | 16.67 |
| Out-of-pocket expenditure (% of current health expenditure) | 7.61 |
| Physicians (per 1,000 people) | 41.30 |

Two different methods were used to fill the missing values. These methods were chosen based on the objectives and type of data:

* Median Imputation- This method was used for columns with a lower percentage of missing values. It is also a robust measure against outliers. E.g., Domestic general government health expenditure (% of GDP)
* Predictive Model Imputation- This method was used for columns with higher percentages. E.g., Physicians (per 1,000 people)



Code\_screenshot\_1: Code to check the percentage of missing values



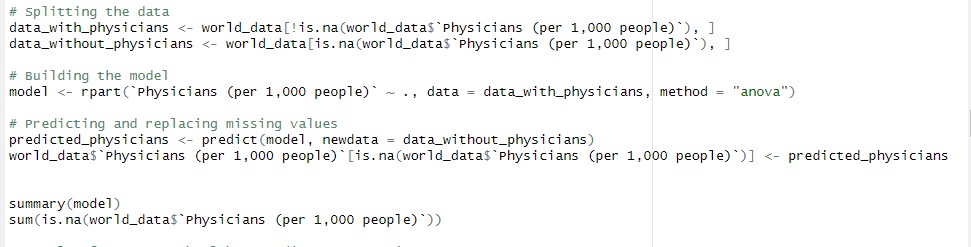
Code\_screenshot\_2: Columns to be filled using median imputation

Code\_screenshot\_

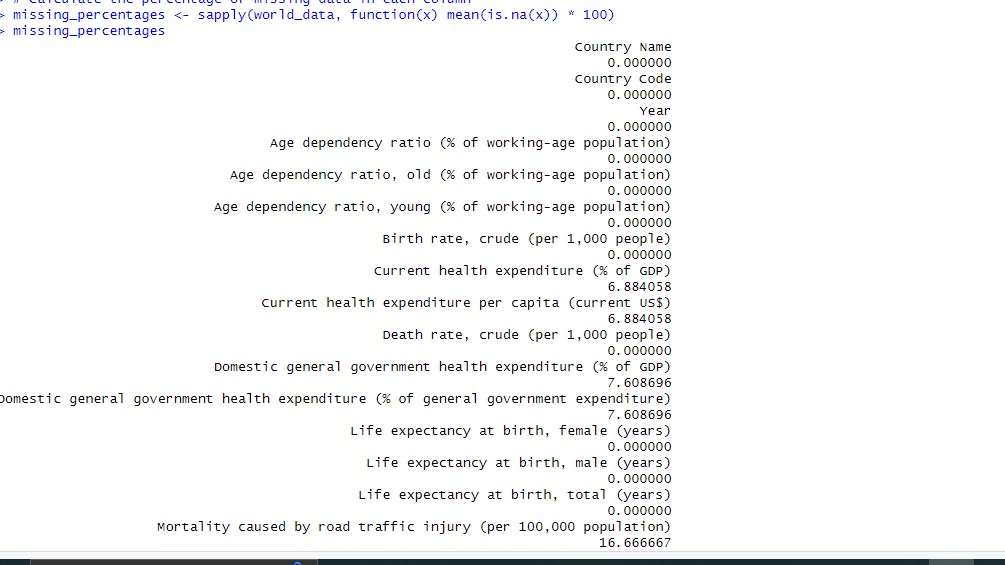
3

:

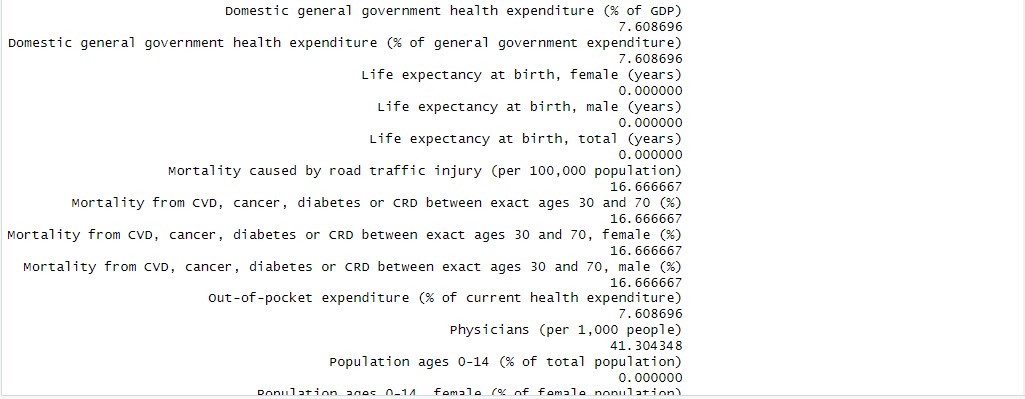
Predictive Model Imputation



RESULTS:



Code\_Output\_1: The percentages of missing columns



Code\_Output\_2: The percentages of missing columns(contd)



Code\_Output\_3: Output after all imputations are successfully done

Part Two: Statistical Analysis

1. Statistical Analysis

1.1 Descriptive Statistical Analysis

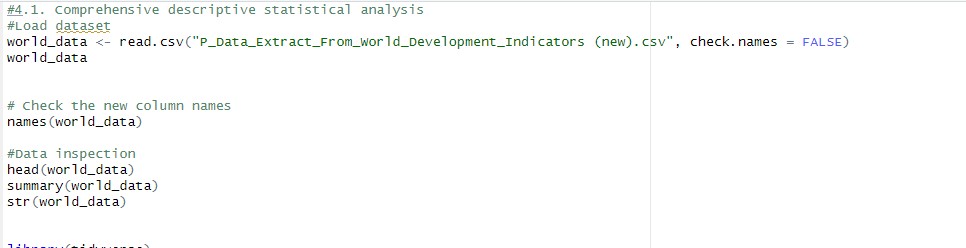
This section provides a descriptive overview of the health-related metrics essentially relevant to the stated objective. In this section, the mean, mode, standard deviation, skewness, and kurtosis of the dataset will be explored. Additionally, plots to physically explore the data will be shown in this section. These plots are a critical step in gaining insights into the data structure, outliers, and relationships between the variables and the distributions. 1.1i Analytically Exploring the Data

1

. Load the dataset from

the

local machine



1. Inspect the dataset

This provides a quick snapshot of the data enabling the analyst to understand the dataset structure. This inspection also generates statistical summary per column and provide a concise view of the data frame. These inspections could spot potential problems within the data like outliers, skewed distributions, etc.



1. Statistical Measures

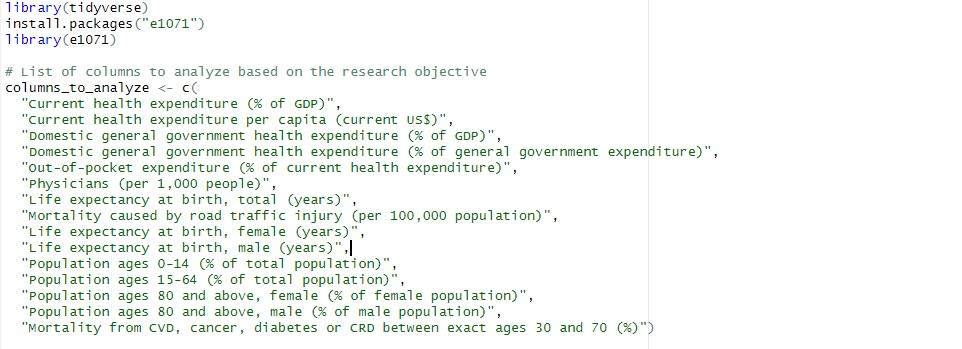
These stats measures aims to deliver insights into the distribution and behavior of the primary indicators.

The ‘calculate\_stats’ function is employed in this code to compute various statistics column per column.

The following tendencies have been measured: a. Mean- Calculates the value of average

* 1. Median-Computes the overall middle value
  2. Mode-Computes most frequently occurring value in the dataset
  3. Standard\_Deviation-Measures the variation amount of a set of values
  4. Skewness-Assesses the leaning asymmetry of the data distribution from a normal distribution.
  5. Kurtosis- Indicates how heavily the tail of a distribution differs from the tail of a normal distribution.
  6. Range-The difference between the maximum and minimum value
  7. IQR (Interquartile Range)- Calculates the spread of the middle 50% of the values.
  8. Variance-Measures how far each value in the set is from the mean.
  9. Quantiles-Computes the data into equal quarters

Each of these measures are pertinent for the understanding the shape, dispersion, tendency of the data ensuring best-practice quality data.



1.1ii Visually Exploring the Data

In this visual EDA, boxplots, histograms,scatterplots and multiplots are leveraged.

* + - Understanding Distribution and Detection of Outliers:

The variables plotted against each other aid in identifying outliers and the spread of data, which are compasses in unveiling the behavior of health expenditure and life expectancy across the dataset. Current health expenditure per capita and Life expectancy at birth are box-plotted against each other to reveal the central tendency and dispersion of the variables.

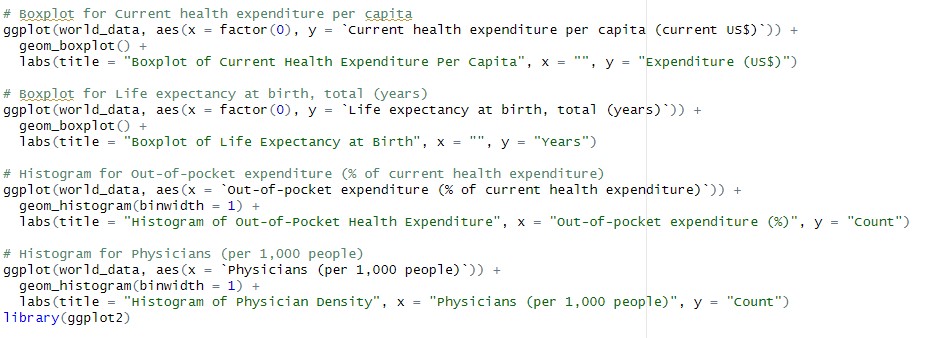
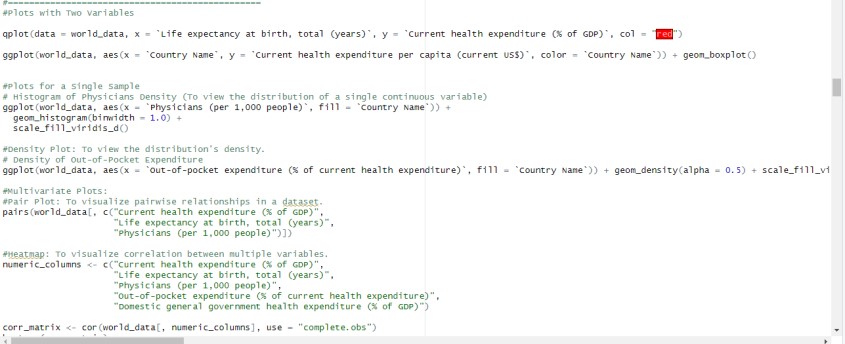
* + - Examining Data Skewness and Modality

The frequency of distribution for Out-of-pocket expenditure and Physicians (per 1,000 people) are illustrated by histograms. These plots are expected to spot any deviance in pattern that

could be data entry errors that could prompt further investigations. They are also essential to observe skewness, modality, etc.

* + - Visualize Overall Data Structure

Multivariate plots like pair plots and heatmaps proffers a deeper view of couple-like relationships and correlations among multiple variables. The are essentially handy for spotting which variables are moving together, which can be indicative of underlying patterns relevants to the objective



1.1a Descriptive Statistical Analysis Results

1. Analytically Exploring the Data

1a. Loaded dataset



1

b.

Data Inspection

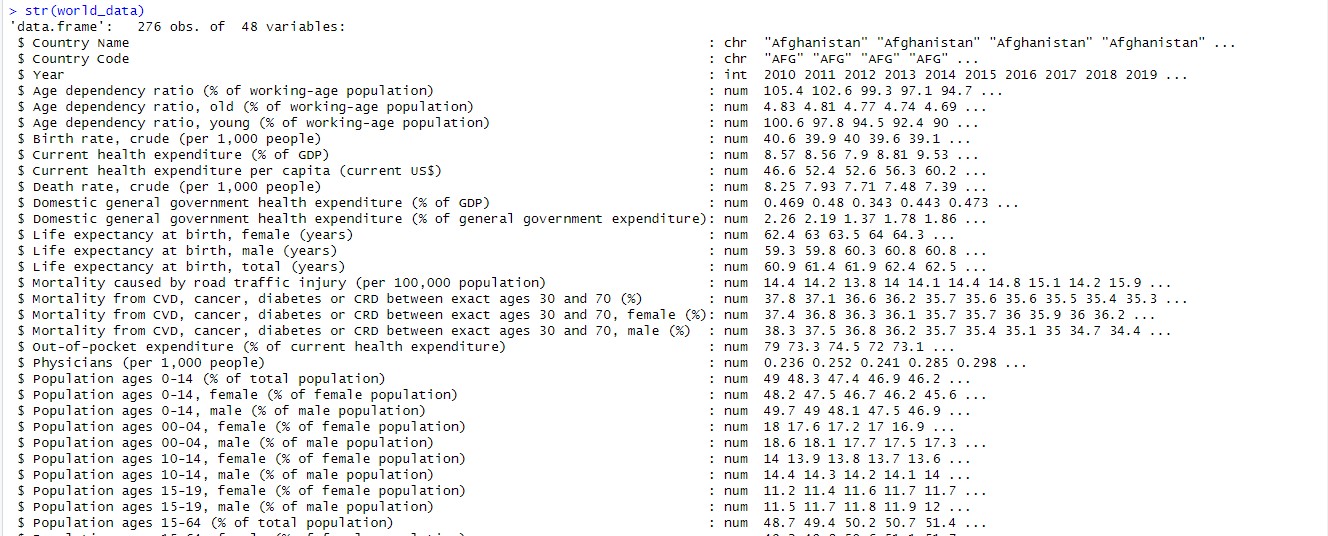


1

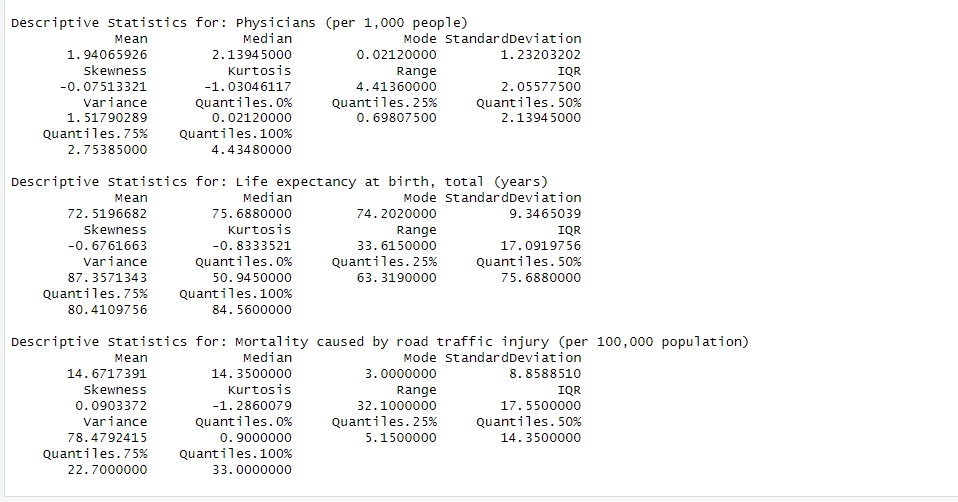
c.

Statistical

Measure



1.

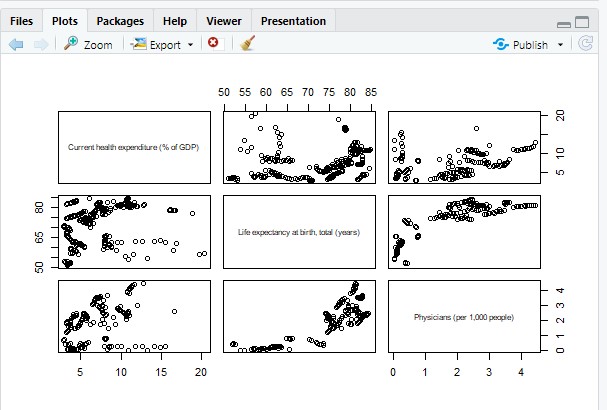
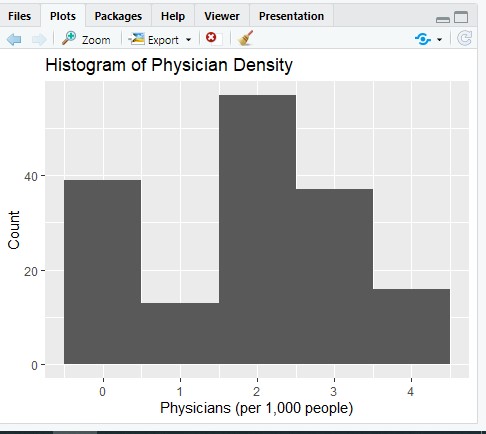
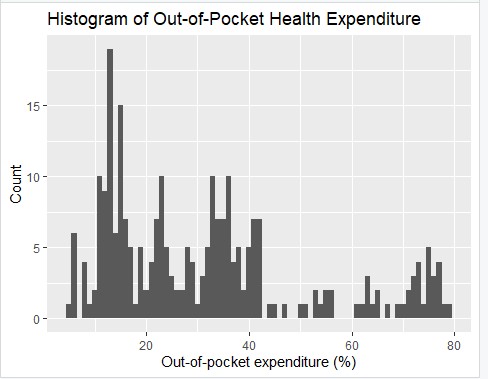
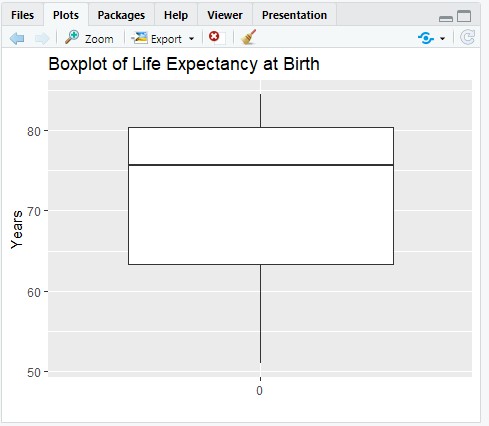
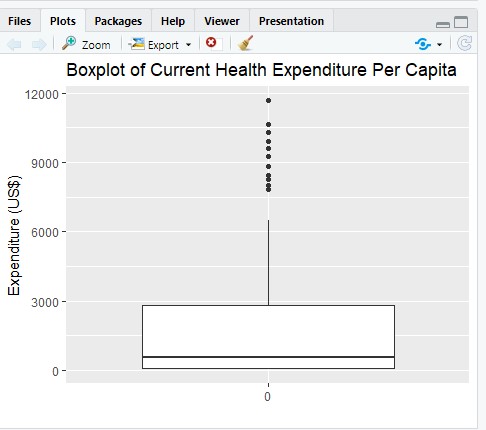


2

a.

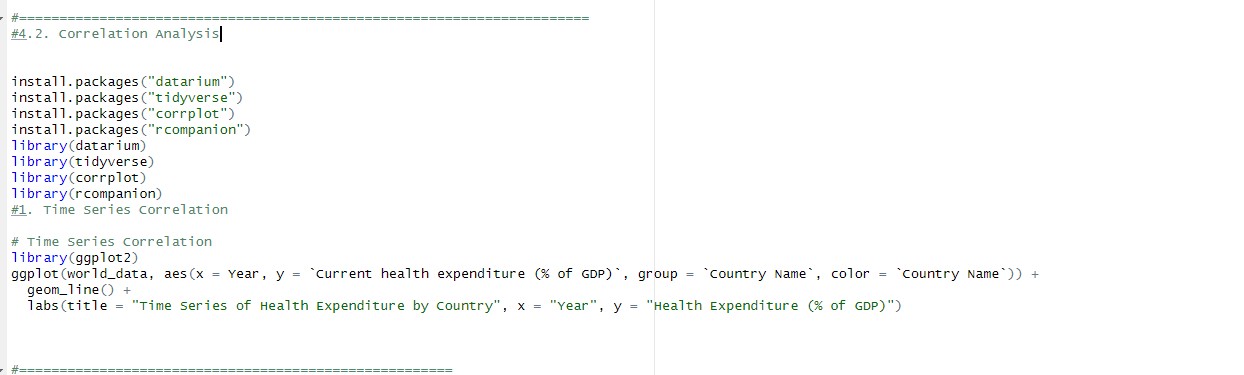
Visually

Exploring the Data



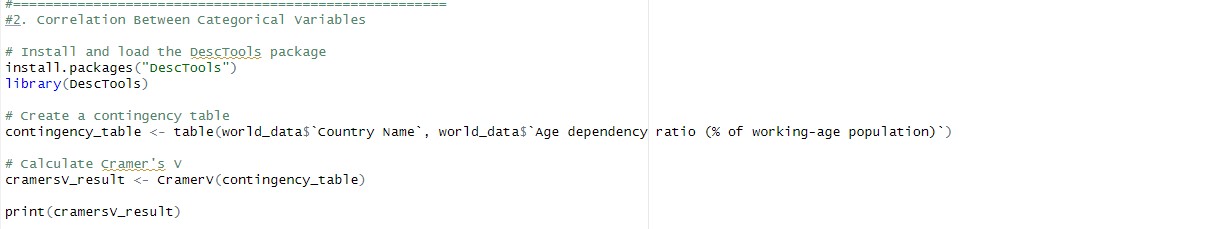
2.1 Correlation Analysis

In this section of the analysis, the aim is to highlight the correlations between the variables in this study based on the objective already stated at the beginning. It is important to note the complexity and versatility of the dataset which is normal for health economics datasets, on that basis, we will explore multiple ways to compute the correlation amongst variables. DescTools, corrplot, RColorBrewer, and lme4 libraries were used in this analysis to provide an encompassing toolkit to ensure the viability of the analysis.



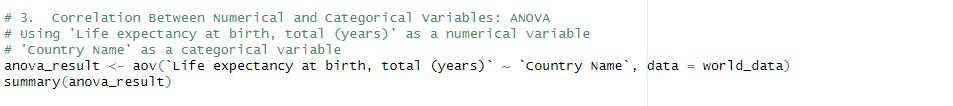
# a. Cramer’s V-Test

This test assesses the relationship and association strength between two categorical variables. In this case, it is used to understand the relation between Country Name and the Age dependency ratio.



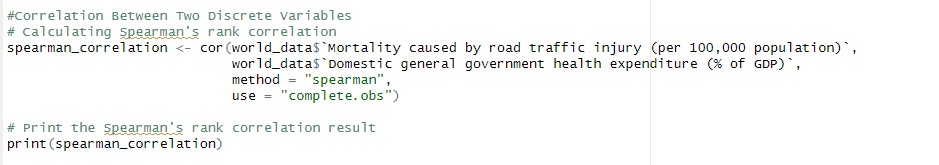
# b. ANOVA Test

This test evaluates the possibility of any statistical significant difference in Life expectancy at birth, total (years) across different countries.



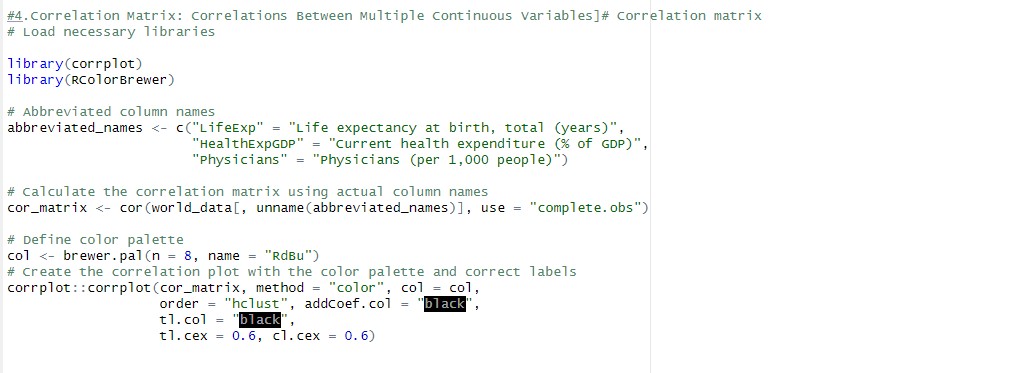
# c. Spearman's Rank Correlation

During the course of analysis, it was discovered that the dataset exhibited non-parametric qualities hence the use of a non-parametric test. This test measures the strength and association direction between two ranked variables- Mortality caused by road traffic injury and Domestic general government health expenditure.



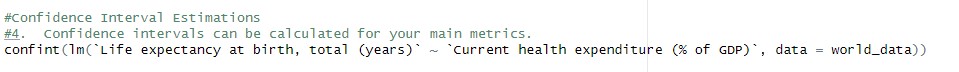
# d. Correlation Matrix

This is a visual representation and it provides and indepth perspective of the relationships between multiple continuous variables such as physician density, health expenditure and life expectancy.



# e. Confidence Interval Estimation

This method estimates the range of values within which the true population parameter resides within a certain level of confidence- for essential metrics like Life Expectancy at Birth.

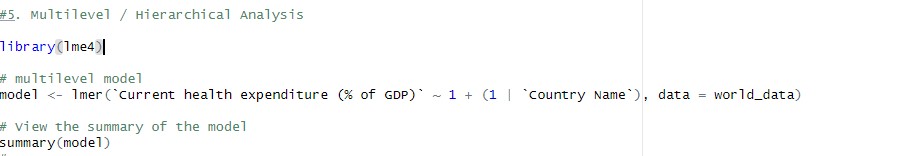


# f. Multilevel/Hierarchical Analysis

This analysis determines how indicators like Current health expenditure (% of GDP) varies within and across countries using linear mixed model. The fixed effects provide the overall health expenditure while the random effects provides the due variance to country differences.

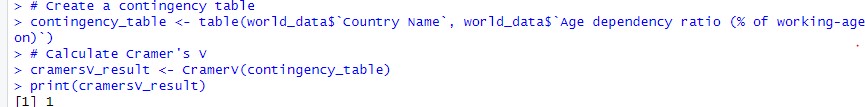
2.2

Correlation Analysis Results



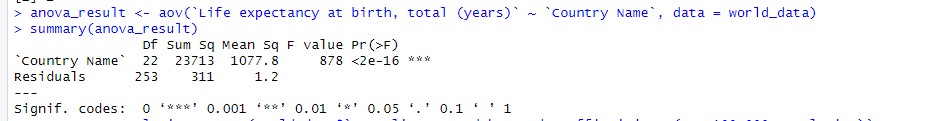
# a. Cramer’s V-Test

The Cramer V statistic is 1. This indicates a perfect association between these two categorical variable which is quite unlikely in most real-world scenario.



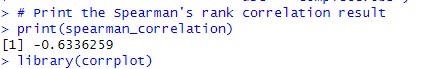
# b. ANOVA Test

The low p-value (<2e-16) in the ANOVA test shows that there are indeed significant difference between Life expectancy at birth, total (years) and different countries. Suggesting that the variable of life expectancy varies significantly by country.



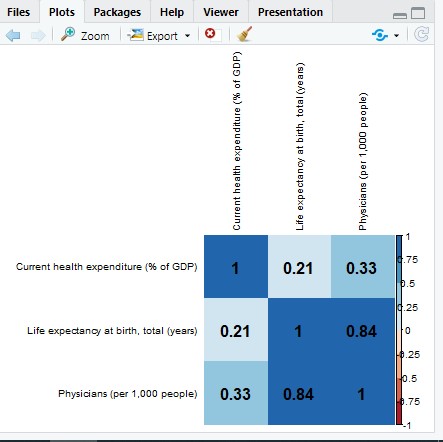
# c. Spearman's Rank Correlation

The Spearman correlation coefficient of -0.6336259 implies a fairly strong negative relationship between Domestic general government health expenditure (% of GDP) and Mortality caused by road traffic injury (per 100,000 population). This suggests that as one variable increases, the other tends to decrease (i.e. as Domestic general government health expenditure increases, Mortality caused by road traffic injury would decrease)



# d. Correlation Matrix

The result of this correlation matrix shows the strongest relationship is between Life expectancy at birth, total (years) and Physicians (per 1,000 people) with a correlation score of 0.84. It is important to note that the correlation coefficient values are between -1 to +1



# e. Confidence Interval Estimation

The confidence interval for the regression model predicting the concerned variable is highlighted as follows:

* The intercept has a confidence interval of 65.23,70.28, inferring that if the expenditure were to fall to null, the life expectancy at birth would also fall within this range with a 95% confidence.
* The result of the slope of Current health expenditure (% of GDP) has a confidence interval of

0.34,0.94/0.34,0.94. This would infer that for each one percent increase in the health expenditure of GDP, the life expectancy would be expected to increase by an amount within this range, assuming all the other variables remain constant.

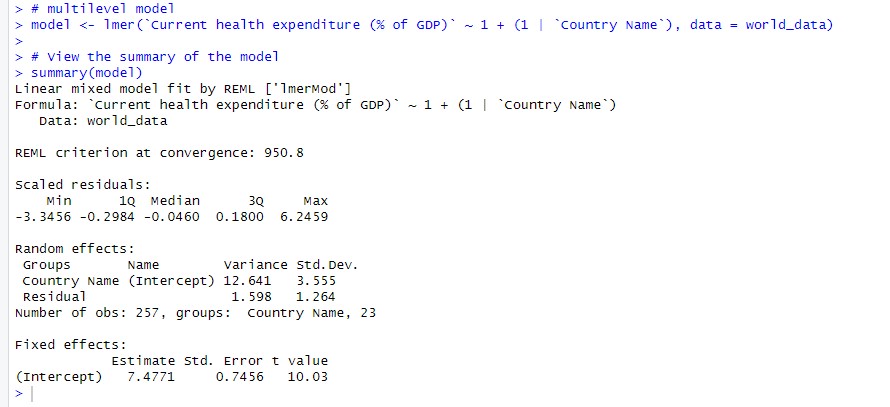
* The summary of the linear mixed model shows:
  + The touch-point for Current health expenditure (% of GDP) is estimated at 7.4771 with standard error of 0.7456. The t-value of 10.03 is highly significant, inferring that the average health expenditure has a strong positive effect on the outcome variable measured across the dataset.
  + The random effects show a variance of 12.641 and standard deviation of 3.555 for the intercept across the different countries. This implies significant deviation in health expenditure among different countries. - The residual variance within each country is 1.598, with a standard deviation of 1.264, pointing to the variability of health expenditure within a country.

These results shows analytical evidence that higher health expenditure possesses strong association with increased life expectancy by birth and reduced mortality.



# f. Multilevel/Hierarchical Analysis

This analysis determines how indicators like Current health expenditure (% of GDP) varies within and across countries using linear mixed model. The fixed effects provide the overall health expenditure while the random effects provides the due variance to country differences.



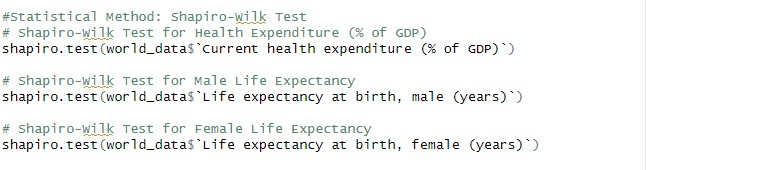
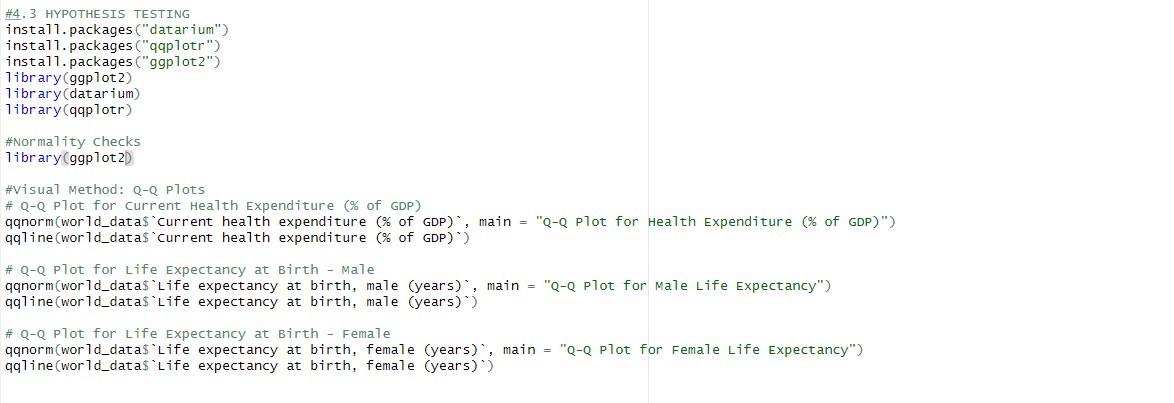
3.1 Hypothesis Testing

The hypothesis testing was methodologically analyzed and executed as discussed below. Qqplotr, ggplot2 and datarium libraries were the major library used in this analysis:

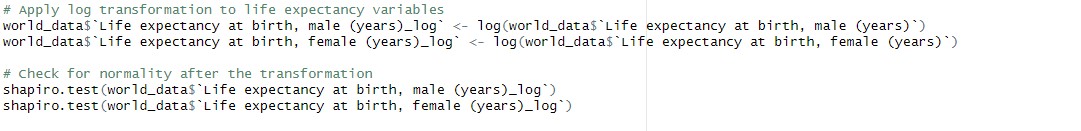
a. Normality Check (Visual and Statistical): Normality check is a statistical technique used to modify or scale data to a standard range making it an improved version of itself by improving the comparability and performance of various variables.

A normality check is essential to allow the analyst to know how to procced with the analysis. For the visibility check, Q-Q plots would be used over the variables essential to the objective of this hypothesis testing.

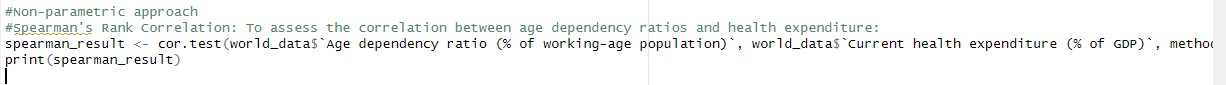
For the statistical check, Shapiro-Wilk Test was employed as the determining test



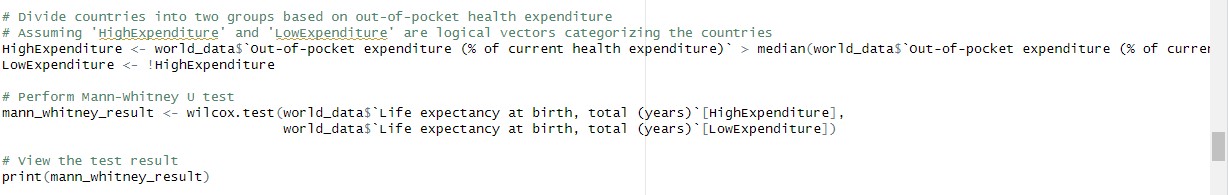
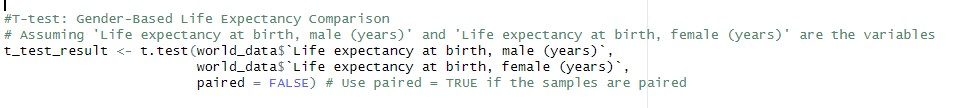
1. Data Normalization: In the case where the data deviates from normality, as in the case of our data, log transformations were employed to moderate skewness, striving to achieve distributions to better satisfy the conditions for parametric testing.



1. Spearman’s Correlation: This correlation coefficient was employed to compute the strength and direction of association between distributed variables that are not normally distributed or ordinal data such as age dependency ratios and health expenditure, by passing the compulsion for normal distribution prerequisites. As opposed to Pearson's correlation which assumes normality, Spearman's correlation analyses monotonic relationships between two continuous or ordinal variables without the assumption of normal distribution.

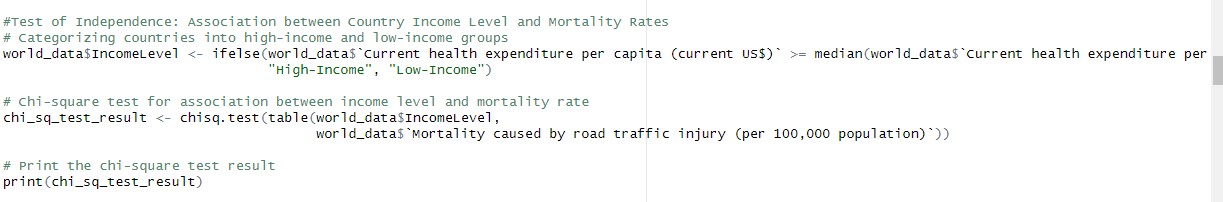


1. Wilcoxon and Mann-Whitney U Tests: These non-parametric tests were employed to serve as rich alternatives to t-tests. These kinds of tests are best-fit for datasets that could not be normalized or essentially non-parametric. The tests were utilized to compare life expectancy between genders and different health expenditure levels.



1. Chi-Square Test of Independence:

The chi-square test was leveraged to show significant associations or independence. This test of independence is used to evaluate the relationship between categorical variables, such as financial levels of countries and mortality rates.



3.2 Hypothesis Testing Results

a. For the visibility check; from the output of all the Q- Q Plots the data points do not lie on or close to the diagonal line indicating the data’s deviation from normality.

-For the statistical check; the Shapiro-Wilk Test was employed as the determining test. This test provides the W statistic and p-value.

-For the Current health expenditure (% of GDP) variable, the W statistic is 0.91023, and the p-value is low

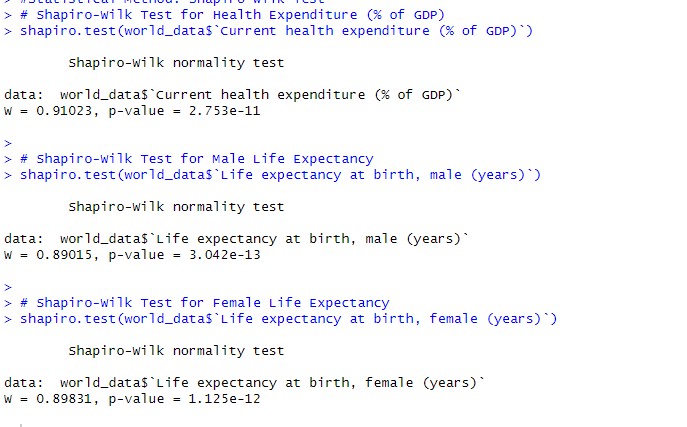
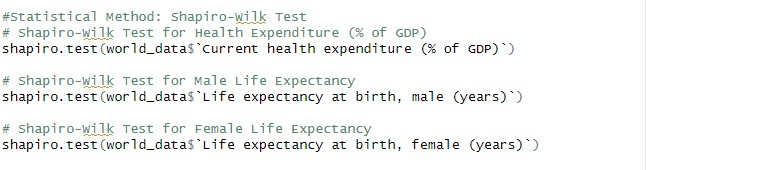
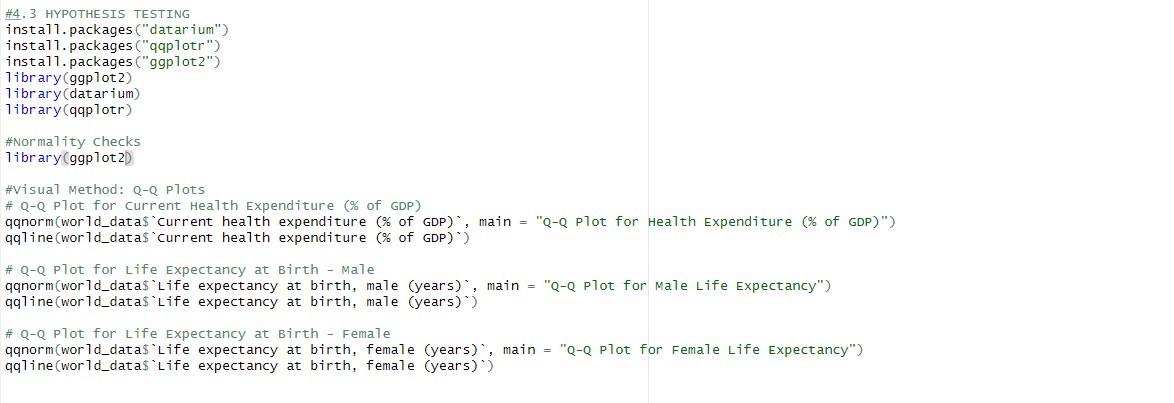
2.753e-11, suggesting that the data has a significant deviation from a normal distribution.

-The Life expectancy at birth, male (years) variable yields a W statistic of 0.89015 with an even lower p-value

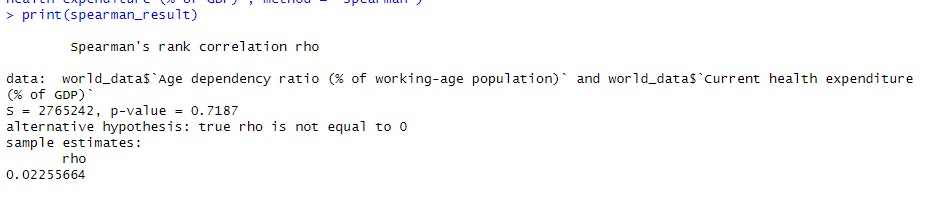
3.042e-13, also suggesting a significant deviation from normality.

- Life expectancy at birth, female (years) variable has a W statistic of 0.89831 and a p-value of 1.125e-12, confirming a similar significant deviation from a normal distribution for female life expectancy data.

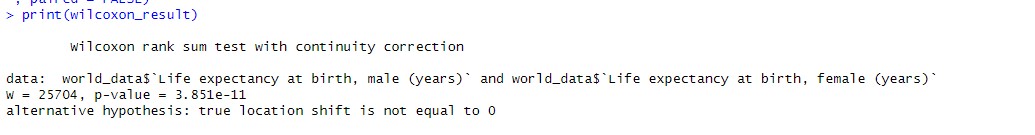
In all three cases, the p-values fall below the common alpha level of 0.05 leading to the rejection of the null hypothesis that the data are from a normally distributed population.



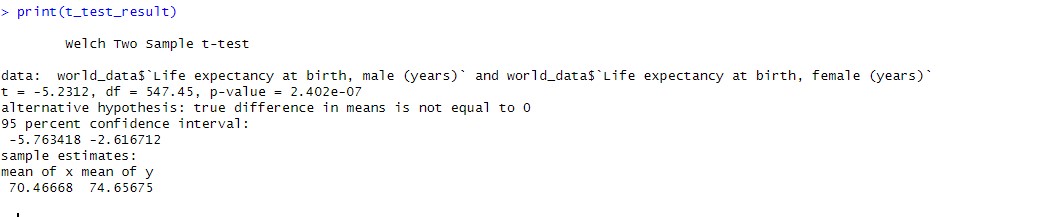
b. **Spearman’s Correlation:** The output of this test shows a weak positive correlation determined by the rho value; where rho = 0.02255664, between the age dependency ratio and current health expenditure (% of GDP), with a high p-value of 0.7187- implying that there is no significant relationship between the two variables.



1. Wilcoxon Rank Sum Test: The resultant value for this test yields a p-value of 3.851e-11, pointing towards the significant difference life expectancy of males and females. This is suggestive of the fact that gender plays a differential role in life expectancy when studied within countries.

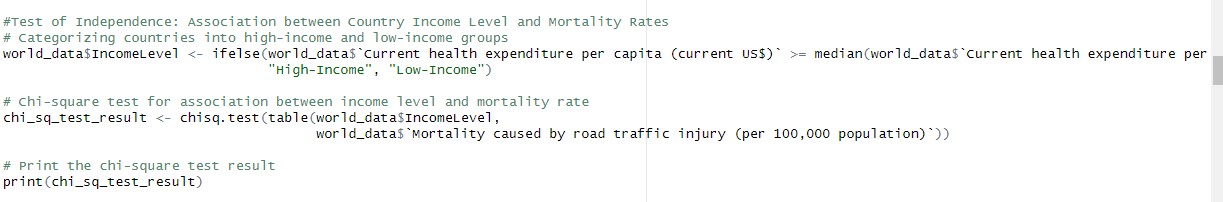


1. The Welch Two Sample t-test: The output for this test yields a p-value of 2.402e-07,reconfirming an important difference in life expectancy between males and females. This results also infers that females have a high life expectancy



1. The Mann-Whitney U test: This result indicates a high significant p-value < 2.2e-16 inferring a noticeable difference between the two categorized groups of countries (categorized by their out-of-pocket health expenditure).

These results heavily suggest that life expectancy at birth significantly varies between countries categorized.



1. Chi-Square Test of Independence: The results of this test yields a p-value of 0.00379. This indicates that there is a statistically significant relationship between the income level of countries and mortality rates caused by road traffic injuries. Suggestive of the fact that the economic status of a country can exhibit relationships with the mortality rates caused by road traffic injuries

3.3 The Hypotheses, Rejected or Accepted?

* + - 1. Hypothesis on the Impact of Out-of-Pocket Health Expenditure on Life Expectancy

Null Hypothesis (H0): There is an absent significant difference in life expectancy at birth between countries with high and low out-of-pocket health expenditures.

Alternative Hypothesis (H1): There is a significant difference in life expectancy at birth between countries with high and low out-of-pocket health expenditures.

Result: The Mann-Whitney U test displayed a highly significant p-value < 2.2e-16, inferring to reject the null hypothesis (H0). We accept the alternative hypothesis that there is a significant difference in life expectancy at birth between countries with high and low out-of-pocket health expenditures.

* + - 1. Hypothesis on Gender-Based Differences in Life Expectancy:

Null Hypothesis (H0): There is an absent difference in life expectancy between males and females within the countries evaluated.

Alternative Hypothesis (H1): There is a significant difference in life expectancy between males and females within the countries evaluated.

Result: The Wilcoxon rank sum test yielded a significant p-value 3.851e-11, leading us to reject the null hypothesis (H0). We accept the alternative hypothesis that there is a significant difference in life expectancy between genders within the countries studied.

* + - 1. Hypothesis on the Association Between Country Income Level and Mortality Rates:

Null Hypothesis (H0): There is no association between a country's income level and its mortality rates caused by road traffic injuries.

Alternative Hypothesis (H1): There is an association between a country's income level and its mortality rates caused by road traffic injuries.

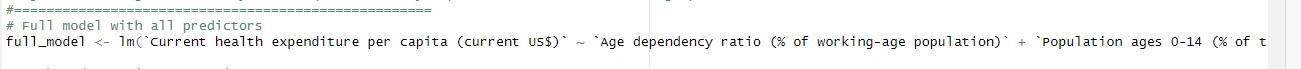
Result: Pearson's Chi-squared test yielded a p-value of 0.00379, hence we reject the null hypothesis (H0). We accept the alternative hypothesis that there is a statistically significant association between the country's income level and mortality rates caused by road traffic injuries.

4.1 Regression Analysis

The analysis is targeted at exploring the population impact and economic variables on health indicators whilst evaluating how this impact may influence mortality and life expectancy.

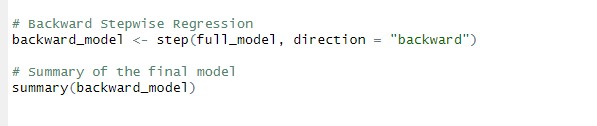
* + 1. First, Model Setup

This is a requirement to ensure a full model incorporating all potential predictors that are influencers like mortality and life expectancy.



* + 1. Backward Stepwise Regression

Given the complexity of the dataset and objectives, a Backward Stepwise Regression is recommended. This method begins with all potential predictor variables and systematically removes the least significant variables one per time. It's competent for datasets with multiple variables, it supports the identification of the most important predictors whilst avoiding fitting.

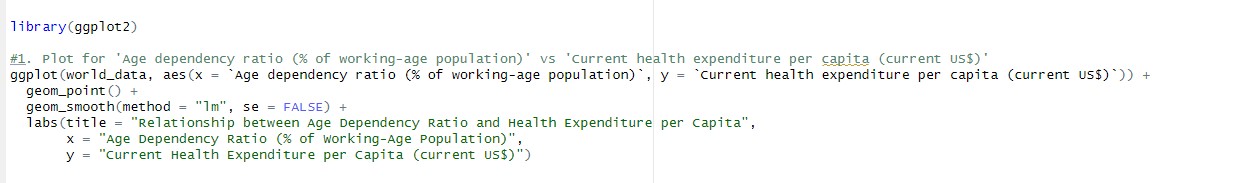


* + 1. Final Model Evaluation

On conclusion of the regression step, the resultant model is analyzed to validate the influence of the retained predictor variable.

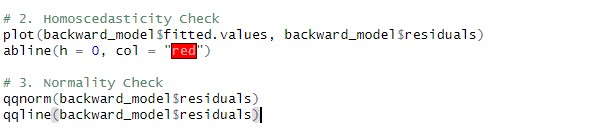
* + 1. Visual Exploration

Visualizations with regression lines are used to evaluate the relationships between variables like age dependency and health expenditure, and any other viable predictor.



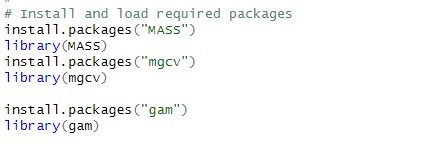
* + 1. Model Assumption Checks

To ensure the model’s viability, diagnostic checks are pertinent. These tests homoscedasticity, normality, independence, and multicollinearity are computed to verify that the model’s assumptions are valid.

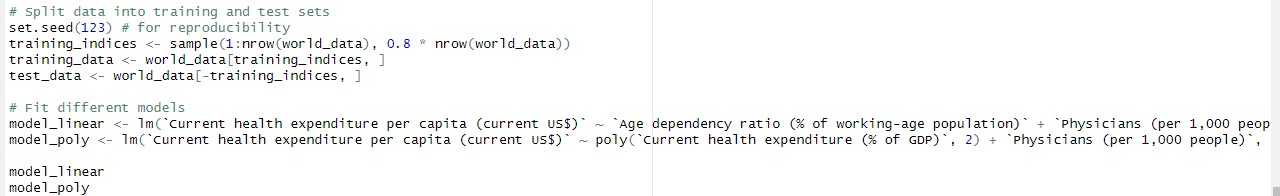


1. Model Fitting on Training Data (Empolying the use of alternate models due to the peculiarity of our data)

g.

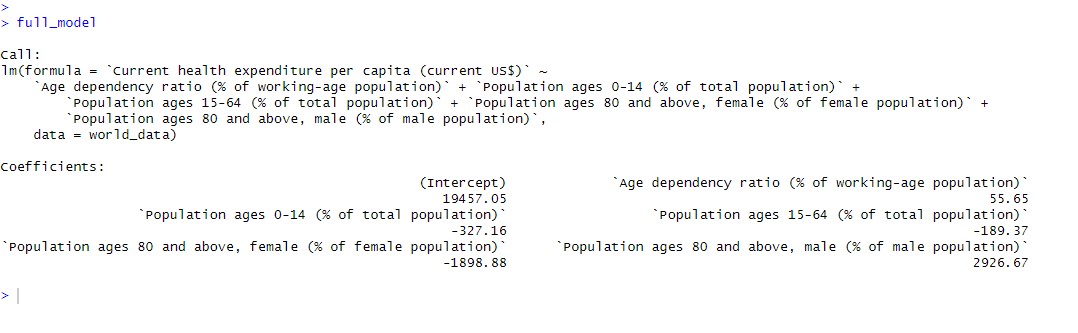


The data is divided into training and test sets to validate the predictive capabilities of the model. Models like linear, polynomial, etc. are fitted on the training data. Fitting the model on training data is an important step to model comparison and selection to find the best model for health outcomes prediction, based on demographs and economic variables.



4.2 Regression Analysis Results

a. The Full Model Results:



-The intercept is sitting high at 19457.05. This indicates the base level of current health expenditure per capita when all other variables are zero.

-The age dependency ratio has a positive coefficient of 55.65, inferring that when there is an increase in age dependency ratio, there is also a rise in health expenditure per capital- a spotted relationship. Therefore, this validates the expectation of additional healthcare costs that will arise with a higher proportion of dependents (young and old) compared to the working-age population.

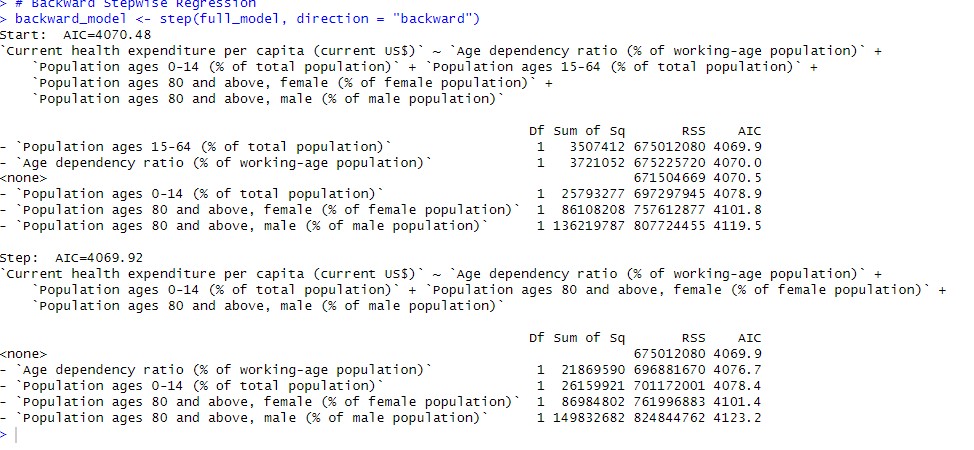
-The population ages 0-14 has an output of a negative coefficient -327.16, inferring that when the percentage of younger population is higher, there is viable correlation with lower health expenditure per capita. Therefore, it may be viable to suggest that countries having a younger population may have lower healthcare costs per individual, possibly due to lower healthcare utilization or different healthcare needs.

-Similarly, the population ages 15-64 coefficient resultant output is seen to be negative -189.37, on this, an inference would be analytically drawn that defines a higher proportion of working-age individuals is associated with lower per capita health expenditure, likely due to this group being generally healthier.

-For the population ages 80 and above, the coefficients are at obvious variance from the population age ranges examined above . The negative coefficient for females -1898.88 and the positive one for males 2926.67 might infer varying patterns of healthcare expenditure for the oldest individuals in the population, potentially reflecting gender differences in health conditions or treatment approaches in senior life.

This model can be informative for policymakers and health economists as it underscores the impact of demographic factors on health expenditure. The significant coefficients imply that demographic shifts towards an older population may drive higher health expenditure per capita, which is a crucial consideration for health system sustainability and planning.

1. The Backward Model Results:



The backward method is focused on optimizing the regression model by removing the least significant predictors one after the other. This process is guided by the Akaike Information Criterion (AIC). Lower AIC points to a better model.

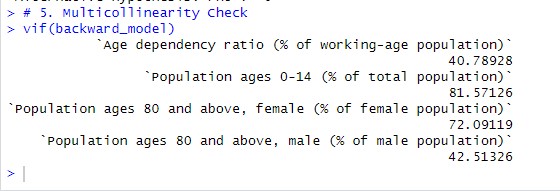
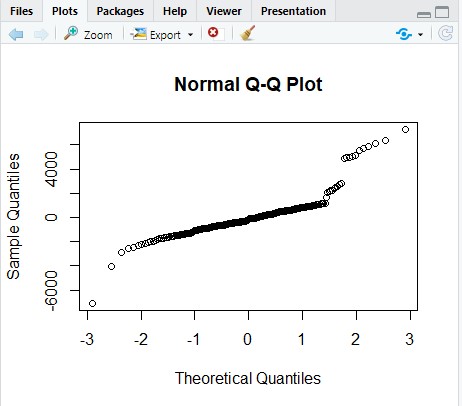
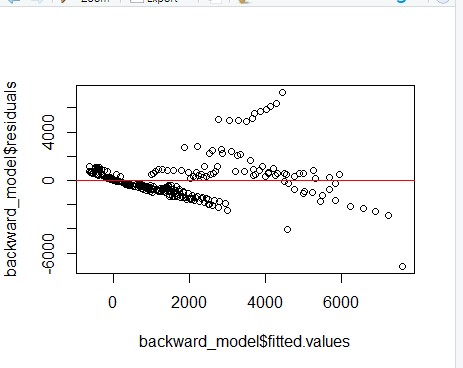
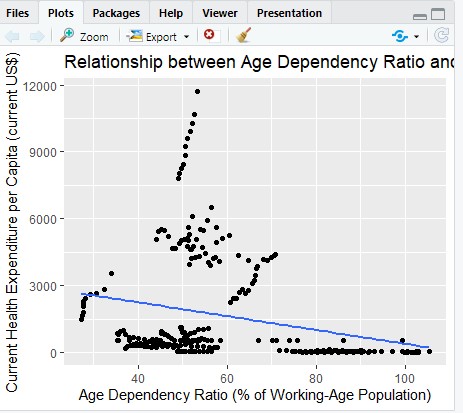
-The beginning step of the backward model selection began with a full model including all variables. It suggests removing the variable Population ages 15-64 (% of total population) first as it has the least impact on the RSS (Residual Sum of Squares) and would result in the lowest AIC if removed (4069.9 vs. the current 4070.5).

-On removal of Population ages 15-64 (% of total population), the next step evaluated the remaining variables. At this point, none of the remaining variables were removed, as dispalyed by <none> under the Df (Degrees of Freedom) column. The AIC at this step is slightly lower (4069.9) than the initial model, suggesting a better model fit after removing Population ages 15-64 (% of total population).

-T Variables such as Age dependency ratio, Population ages 0-14, Population ages 80 and above, female, and Population ages 80 and above, male were considered to be like removed by the process. The proposed next steps of removal would lead to increased AIC values; this is not viable. It shows that these variables should remain in the model as their presence lowers the AIC.

-The final model includes Age dependency ratio (% of working-age population), Population ages 0-14 (% of total population), Population ages 80 and above, female (% of female population), and Population ages 80 and above, male (% of male population). This model is considered the best among the tested models based on the AIC criteria.

1. Model Assumption Check & Plots:



1. The Linear Model Results:

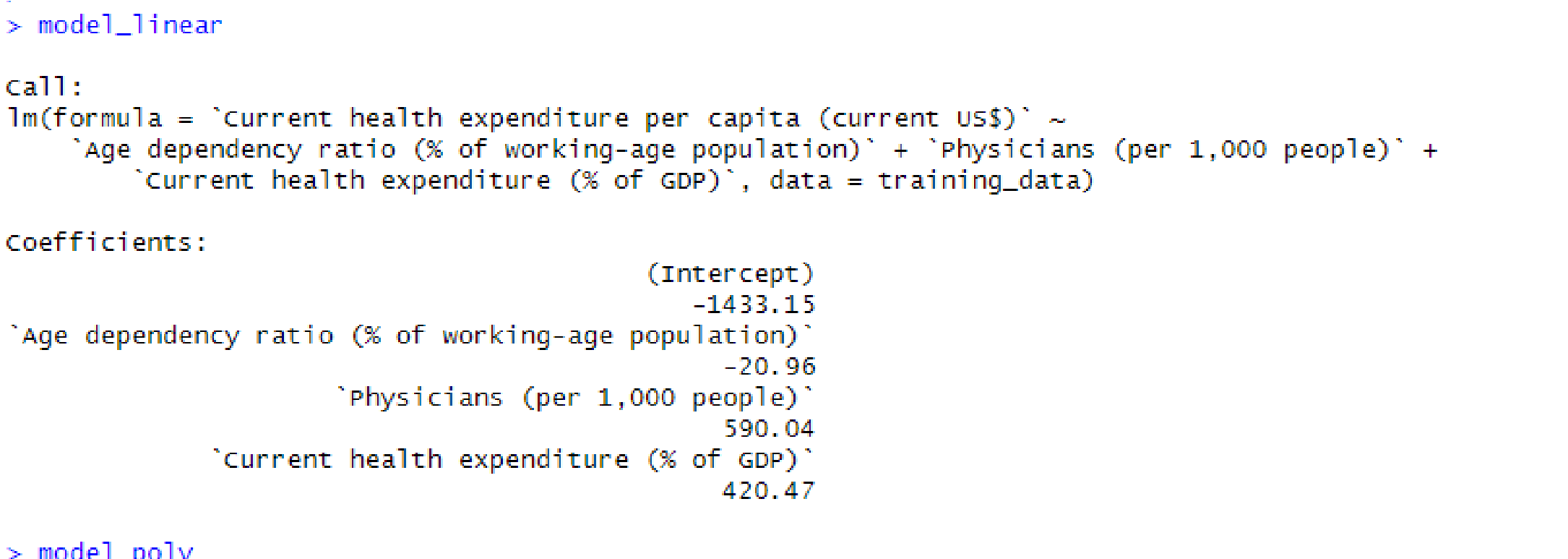
* This model predicts Current health expenditure per capita (current US$) using Age dependency ratio (% of working-age population), Physicians (per 1,000 people), and Current health expenditure (% of GDP).

-The intercept is estimated to be -1433.15, however kind of a summary of the model including the standard errors and t-values is unavailable, hence, we cannot assess the statistical significance of this value.

* As the age dependency ratio increases, the health expenditure per capita tilts to decrease when the coefficient for the Age dependency ratio is negative -20.96, when all else is equal.

* A higher density of physicians is associated with higher health expenditure per capita when the coefficient for Physicians (per 1,000 people) is positive at 590.04.

* As the percentage of GDP spent on health increases, so does the health expenditure per capita. This is valid when the coefficient for Current health expenditure (% of GDP) is positive 420.47



1. The Polynomial Model Results:

-This model includes a polynomial (squared) term for Current health expenditure (% of GDP) to capture nonlinear effects.

-The intercept is near zero (57.03), which is not as relevant for polynomial models because the interpretation is not as straightforward as for linear models.

- A strong non-linear relationship with health expenditure per capita is observed as the first polynomial term coefficient is very large 22387.26.

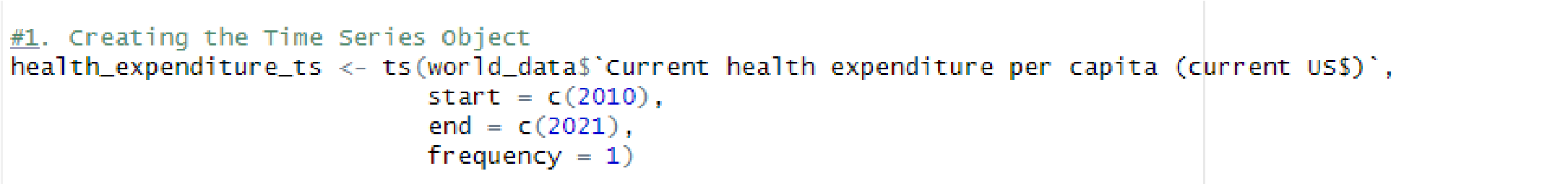
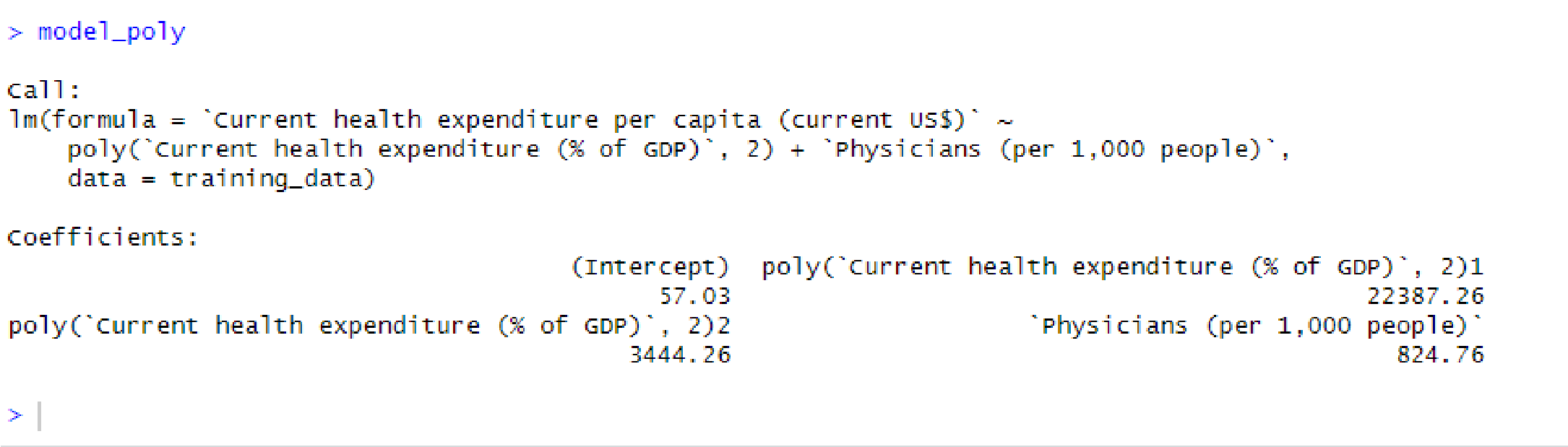
-The second polynomial term has a lower but still substantial coefficient 3444.26, furthermore indicating the non-linear relationship.

-Physician density is strongly associated with health expenditure per capita as the Physicians (per 1,000 people) coefficient is also positive 824.76. This is consistent with the linear model.

5.1 Time Series Analysis

This objective is aimed at analyzing the trends in primary indicators like life expectancy, mortality, population in different age ranges, and health expenditure. This task aims to understand how these variables evolve over the years.

This includes using regression techniques, such as ARIMA and ETS models.



* + 1. Creating the Time Series Object:

Using the ts() function, specifying the start (2010), end (2021), and frequency (annual), time series object for health expenditure per capita was created. The main aim is to model the time-series data to analyze trends and patterns over the years.

* + 1. Plotting the Time Series:

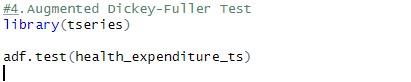
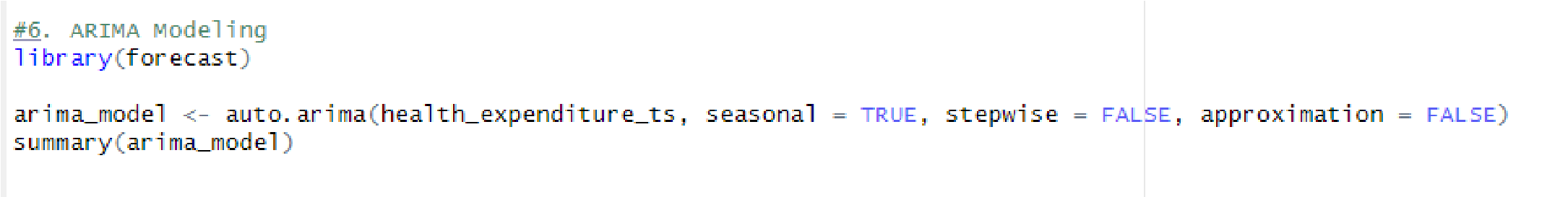
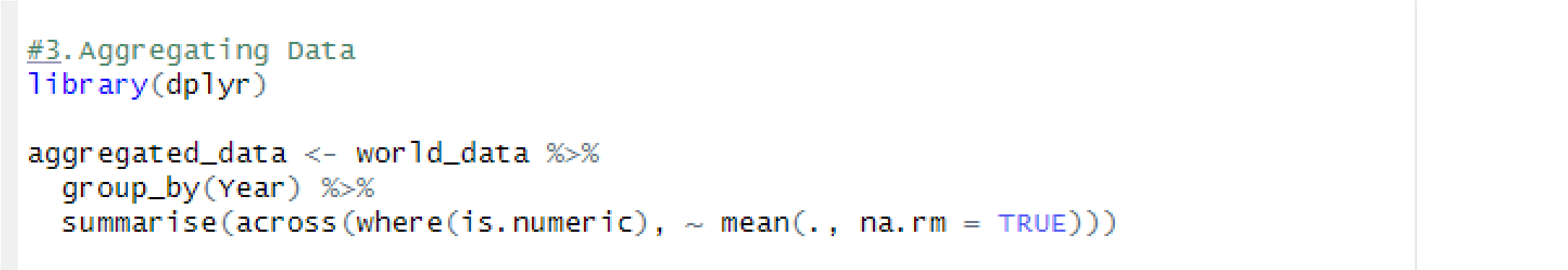
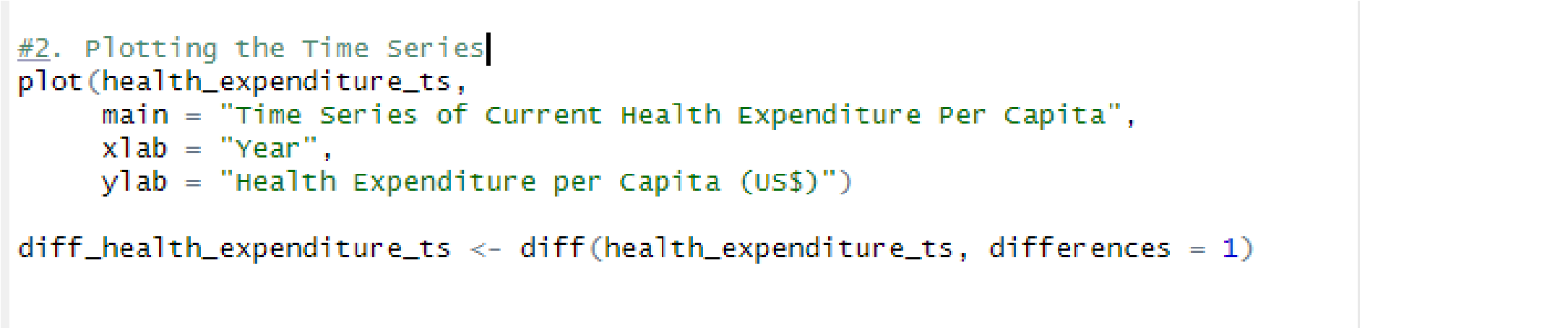
The time series of health expenditure per capita was plotted, providing a visual representation of how this metric have grown over time.

* + 1. Data Aggregation:

Data was aggregated yearly using the dplyr package, summarizing number-based columns to understand annual trends across all metrics.

* + 1. Augmented Dickey-Fuller Test (ADF Test):

The ADF test was used to check the stationarity of the time series, which is a prerequisite for plenty time series forecasting methods.



* + 1. ARIMA Modeling:

An ARIMA model was fitted to the time series using auto.arima() from the forecast package. This model is suitable for non-seasonal data and captures trends and patterns over time.The ARIMA model's summary was inspected to understand its fit and features

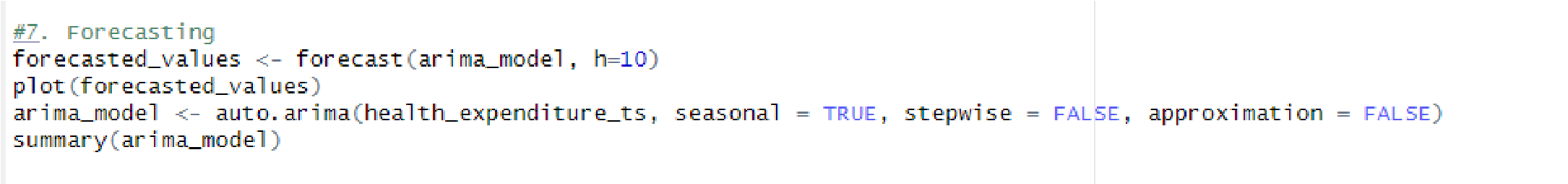
.

1. Forecasting:

Future values of health expenditure per capita were forecasted using the ARIMA model, providing insights into expected trends.

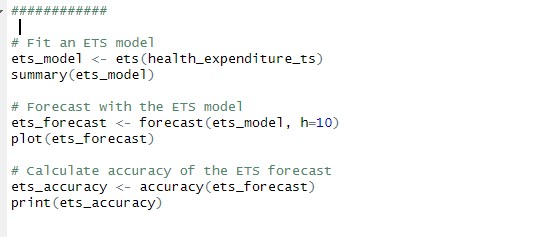
1. Model Evaluation:

The accuracy of the ARIMA model was assessed using various metrics like AIC (Akaike Information Criterion), which balances model complexity and goodness of fit. These metrics help evaluate the model's performance in predicting health expenditure.



1. ETS Modeling:

An ETS (Error, Trend, Seasonality) model was also fitted to the time series as an alternative to the ARIMA model. The ETS model forecasts were plotted and their accuracy was assessed, offering a comparison to the ARIMA model's performance.

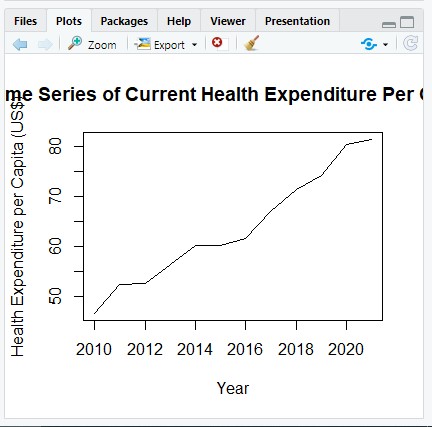
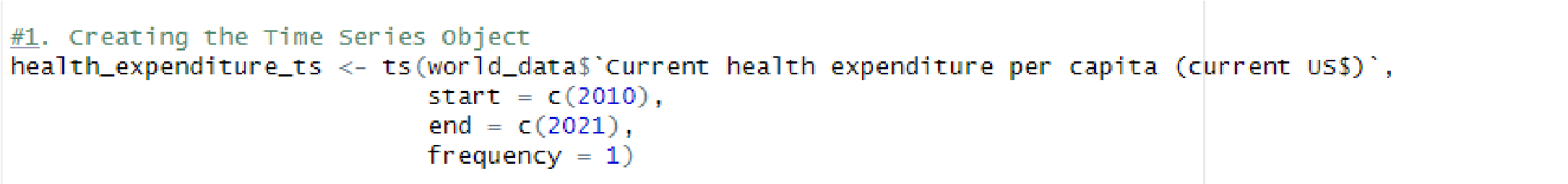


5.2 Time Series Results

* + 1. Creating the Time Series Object:

* + 1. Plotting the Time Series:

The plotted time series graph showing how expenditure per capita metric has evolved over time.



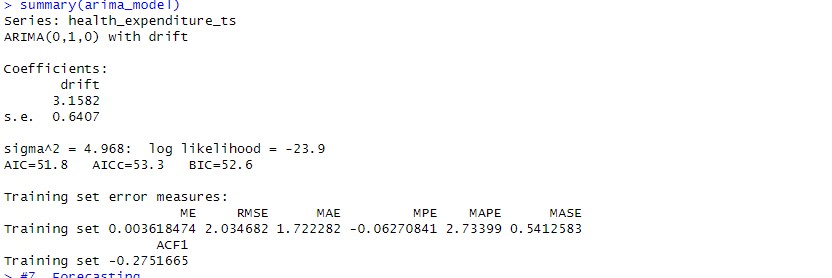
* + 1. Augmented Dickey-Fuller Test (ADF Test):

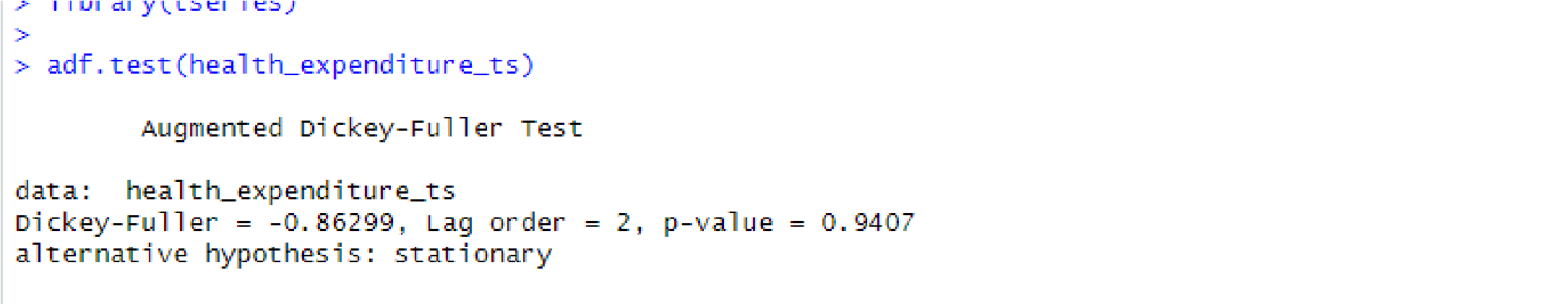
-The high p-value (i.e p-value = 0.940) indicates that the time series is non-stationary, meaning it has trends or patterns that persist over time. This could reflect underlying factors like economic growth, policy changes, or demographic shifts influencing health expenditure.

* + 1. ARIMA Modeling:

-The chosen ARIMA(0,1,0) model with drift indicates that differencing the series once makes it stationary.

The drift coefficient (3.1582) suggests a positive trend in health expenditure over time.

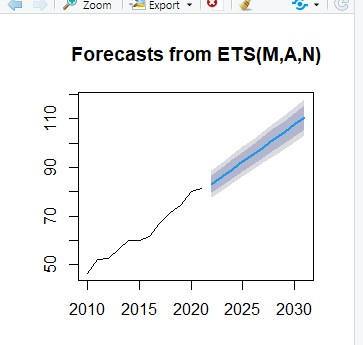
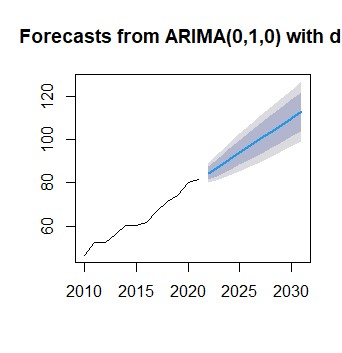
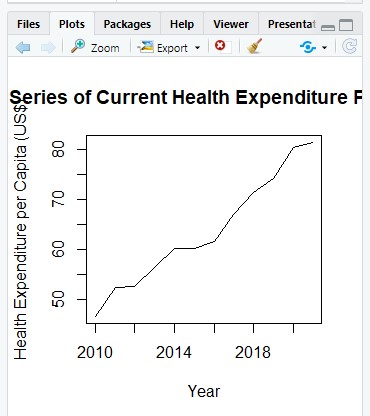




-The low RMSE and MAE values in the ARIMA model imply good model fit and forecasting accuracy on the training set.

- The negative ACF1 value (-0.2751665) hints at some autocorrelation in the residuals, suggesting that the model might be improved by including additional explanatory variables or different configurations.

1. Forecast Plots:

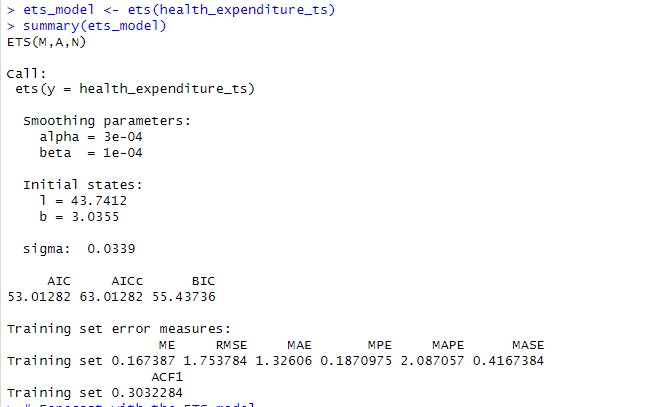


1. ETS Modeling:

The ETS(M,A,N) model suggests that a multiplicative error, additive trend, and no seasonal component best represent the series.

The model's smoothing parameters alpha = 3e-04, beta = 1e-04 imply minimal adjustments to the level and trend, indicating a stable trend over time.

The AIC and BIC values are slightly higher than those in the ARIMA model, which might suggest a slightly less efficient fit to the data.



1. Forecasting Accuracy:

The ETS model shows a slightly better forecasting accuracy compared to the ARIMA model, as evidenced by lower RMSE and MAE values. The positive ME value indicates the model's forecasts are, on average, slightly higher than actual values.

* + Both models confirm a consistent trend of increasing health expenditure per capita over time.
  + The presence of autocorrelation in both models’ residuals suggests additional patterns in the data that the models have not captured, which could be addressed in future model refinements.

6. Discussion and Conclusion

The research methodology implemented in this study involved a multifaceted approach, integrating descriptive statistics, correlation analysis, hypothesis testing, regression analysis, and time series analysis to investigate the interplay between health service outcomes, demographics, and fiscal allocations across various countries.

-Descriptive Statistics: The initial phase focused on providing a foundational understanding of the dataset. Measures like mean, median, mode, standard deviation, skewness, and kurtosis offered an overview of the central tendencies and variability within the data. This phase was crucial for identifying trends and anomalies in the data, setting the stage for more complex analyses.

-Correlation Analysis: This phase involved assessing relationships between variables. Techniques like Spearman's rank correlation, ANOVA, and Cramer's V tests were employed to understand the strengths and directions of associations between various demographic and health-related variables. The use of a combination of parametric and non-parametric methods ensured robustness, especially in dealing with non-normally distributed data.

-Hypothesis Testing: Several hypotheses were formulated and tested to examine specific relationships and differences, such as the impact of age dependency on health expenditure and gender differences in life expectancy. Methods like the t-test, Mann-Whitney U test, and Wilcoxon Signed-Rank test provided statistically significant insights, although the non-normal distribution of some data required careful interpretation and the application of non-parametric tests.

-Regression Analysis: Backward stepwise regression was utilized to explore complex relationships between multiple predictors and health expenditure per capita. This method effectively identified the most impactful factors while controlling for overfitting. The regression models highlighted significant predictors such as age dependency ratios and specific age groups.

-Time Series Analysis: ARIMA and ETS models were employed to analyze trends in health expenditure over time. These models identified non-stationarity in the data and provided reliable forecasts. However, the presence of autocorrelation in the residuals indicated room for improvement in the models.

6.1 Limitations:

-Data Quality: Missing values and potential biases in the data collection process might have affected the analysis.

While measures were taken to address missing data, the quality and completeness of the data remain a concern.

-Model Assumptions: The assumptions underlying regression and time series models, such as linearity and normality, may not fully hold true for the complex, real-world data used in this study.

-Generalizability: The findings are based on the specific dataset and time frame analyzed, which may limit the generalizability of the results to other contexts or periods.

6.2 Conclusion

The research successfully addressed the objectives set at the outset, providing a comprehensive understanding of how demographic shifts and fiscal allocations impact health service outcomes. Key findings include:

-Demographic Influence: Age dependency ratios and specific age groups significantly predict health expenditure per capita. This insight is vital for policymakers, indicating the need for health systems to adapt to changing demographic structures, especially in countries with aging populations.

-Economic Factors: The correlation and regression analyses revealed significant relationships between health expenditure and economic indicators like GDP. The results emphasize the importance of economic considerations in health policy planning.

-Gender Differences: The hypothesis tests confirmed significant differences in life expectancy between genders, highlighting gender-based disparities in health outcomes.

-Trends Over Time: The time series analysis showed a consistent increase in health expenditure per capita over time, aiding in forecasting future trends and assisting in strategic healthcare planning.

This research contributes valuable insights into the dynamics of global health systems, demonstrating the intricate connections between demographics, economics, and health outcomes. The methodologies and findings can guide informed decision-making and policy formulation in the field of public health and economics.

Interactive Dashboard Design

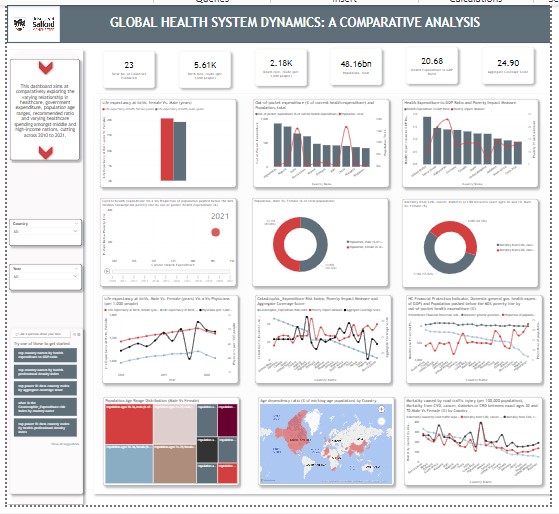
1.1 Introduction and Dashboard Title

This dashboard is titled “GLOBAL HEALTH SYSTEM DYNAMICS: A COMPARATIVE ANALYSIS”. It is aimed

at comparatively exploring the varying relationship in healthcare, government expenditure, population age ranges, recommended ratio, and varying healthcare spending amongst middle and high-income countries spanning from 2010 to 2021.

In this task, a single and simple screen PBI dashboard will be built in alignment with the highlighted objectives as detailed in the literature review. This interactive dashboard has been built with user ease and experience at forefront.

The usefulness of this dashboard spans through various users- from the government to private stakeholders. The intuitive nature of the dashboard will ensure the proposed function for the dashboard is being fulfilled, offering a clear and concise view of the trends within the health care systems across nations. The user will have complete access to filters, drill-downs, large view, etc., within the dashboard.



*Figure 1:* Overview of the Global Health System Dynamics Dashboard

1.2 Dashboard Composition

Visual Paradigm Selection

For the dashboard, we chose a combination of bar charts, line graphs, scatter plots, and heatmaps. Each visual paradigm was selected to best represent the underlying data structure and to facilitate easy comprehension of trends and relationships within the data.

Data to Visual Mapping

The primary step in this dashboard design involved mapping complex health and demographic data into a visually comprehensible format. Key metrics were focused one; like health expenditure, life expectancy, and physician density, transforming them into engaging visual elements. This conversion was pivotal in making the data actionable and insightful.

As per best practice, the schema for this dashboard was drafted before the actual dashboard building to ensure constructive and critical utilization of available space as this dashboard is restricted to a single view only.

**Visually**, this dashboard is split into 3 broad sections:

Section A: The ‘Quick Glance’ section. This section holds the dashboard title and the KPIs from our dataset.

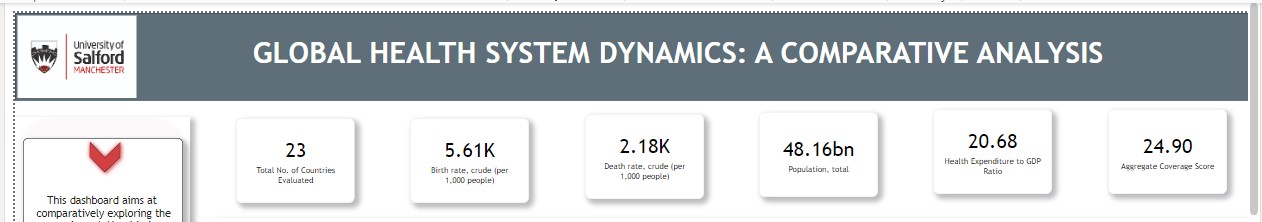
It holds the total countries being evaluated, the total birth and death crude rates across the entire population evaluated, the Average Social Coverage Score, and the Health Expenditure to GDP Ratio. This section allows the user a quick, informative glance; giving them an overview idea of the metric interaction.

*Figure*

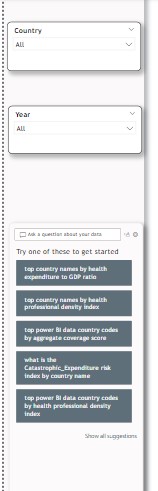
*2*

*:*

*A cross section of Section A*



Section B: This section holds the filters, the objective and the Q & A add-on. These options available are characteristics of why the dashboard may be referred to as intuitive and interactive. The user is able to filter within countries, through timelines. The filter feature has enabled multiple selection, allowing the users to target datapoints of choice. In addition, the Q&A is a feature allows the user explore the data in their own words. For ease, it has been prefilled with FAQs.



*Fig 3: A cross section of Section B*

Section C: This section holds the comparative metrics. In this section comparative analysis are visualized. Some of the metrics compared includes, Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, Male Vs. Female (%), Health Expenditure to GDP Ratio and Poverty Impact Measure,etc. The main story is told here.

*Fig 4: A cross section of Section C*



Conceptual Model Development

**Structurally**, The dashboard was developed with a layered approach:

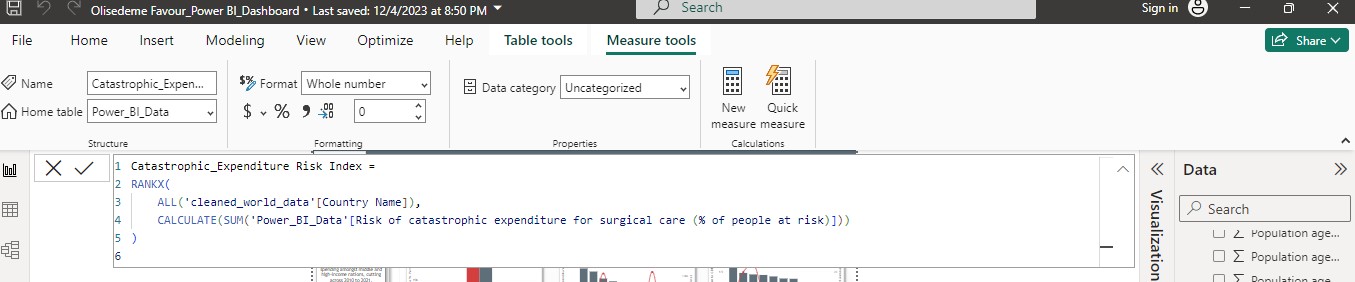
Overview Layer: Provides a high-level summary of key metrics across different countries.

Interactive Layer: Allows users to drill down into specific metrics, offering a dynamic exploration of data based on user interaction.

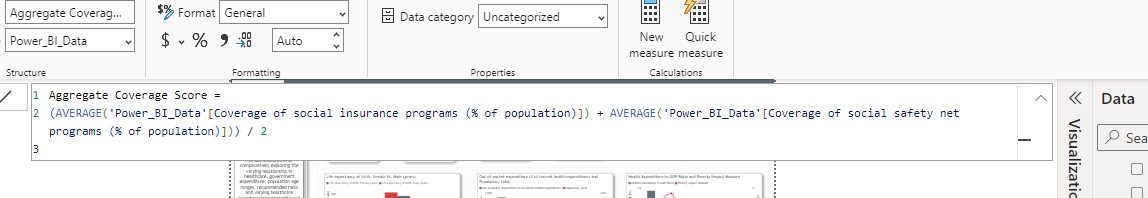
Drilldown Layer: Gives in-depth information on selected data points, such as country-specific health expenditure and demographic details.

1.3 Integration of Advanced Techniques: DAX Measures, Models

DAX Measures: Custom measures were created using Data Analysis Expressions (DAX) for complex calculations and data transformations.



*Fig 5: A cross section of an edit section of a DAX measure (1)*



*Fig 6: A cross section of an edit section of a DAX measure (2)*

Health Expenditure to GDP Ratio: This measure contextualizes health spending by relating it directly to the countries economic size, providing a GDP-adjusted perspective of investment in healthcare.

Health Professional Density Index: A representation of the number of healthcare professionals per capita, this index is crucial for understanding the healthcare workforce's capability to serve the population.

Healthcare Financial Protection Indicator: This measure calculates the proportion of healthcare funding not borne outof-pocket, reflecting the financial protection afforded to the populace by the healthcare system.

Aggregate Coverage Score: This average score combines social insurance and safety net program coverage, offering a single metric for social healthcare protection.

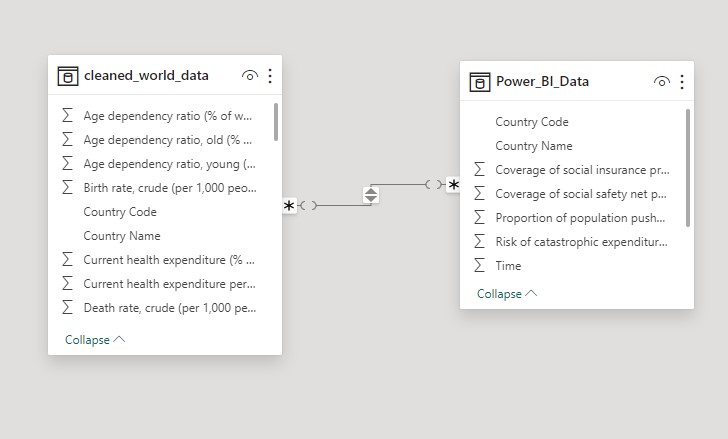
Catastrophic Expenditure Risk Index: This index ranks countries by the risk of incurring financially catastrophic healthcare expenses, highlighting areas where policy intervention may be necessary.

Poverty Impact Measure: This calculation determines the proportion of the population at risk of falling into poverty due to healthcare expenses, underscoring the socio-economic impact of health-related spending.

Data Model Relationships

Utilizing Power BI’s relationship management features, we linked different data sets based on common attributes like country codes.

The primary dataset is enriched with a secondary dataset containing four key indicators: Risk of catastrophic expenditure for surgical care, Proportion of population below poverty due to health expenditure, Coverage of social insurance programs, and Coverage of social safety net programs. This integration offers a multi-dimensional view of global health, tying economic factors to health outcomes.



*Fig 7: A cross section of a relationship model diagram between the primary and secondary dataset*

Hierarchies, Grouping, and Binning

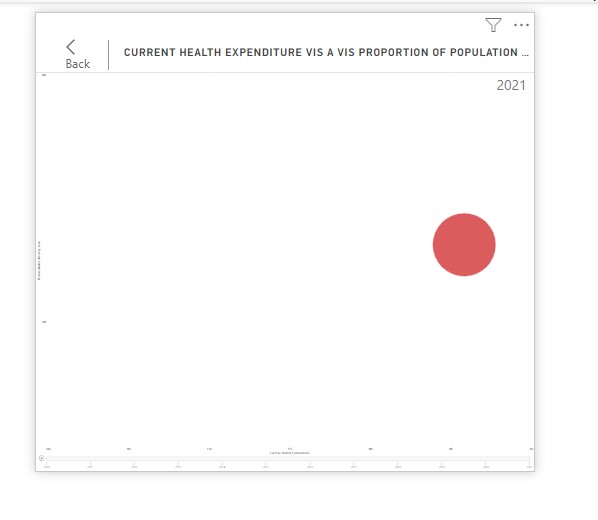
Data was organized into hierarchies and groups for efficient analysis and visualization, such as grouping countries by income levels or health expenditure brackets.

1.4 Selected Charts/Graphs

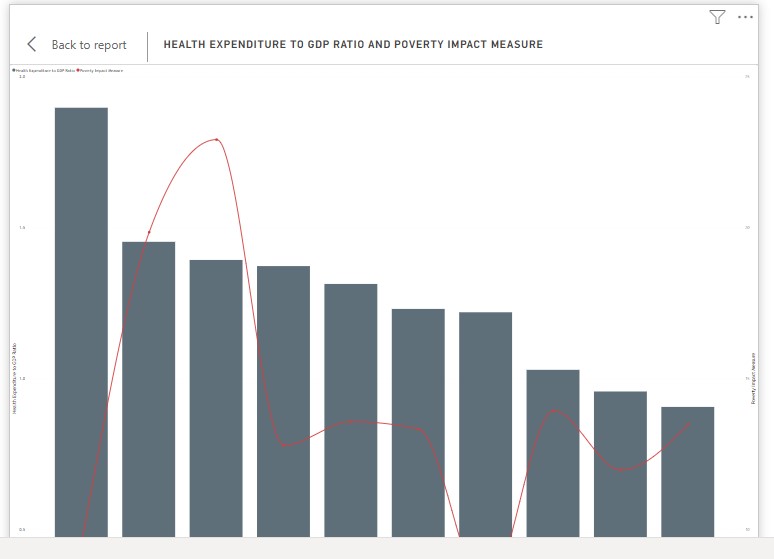
Why and how are the different charts chosen? Charts and bars are selected based on the type of data and the objective of that particular visualization

Current health expenditure (% of GDP) Indicator:

The ‘Current Health Expenditure (% of GDP” is an essential metric within our dataset as it’s interaction with other indicators provides relevant insights into understanding the complex health systems. The figures below will show a few examples of interactions of this essential metric with other metrics.



*Fig 8: [This visualization plays when “Play” is clicked to show comparative analysis of Current Health Expenditure (% of GDP) vis a vis population over time period]*

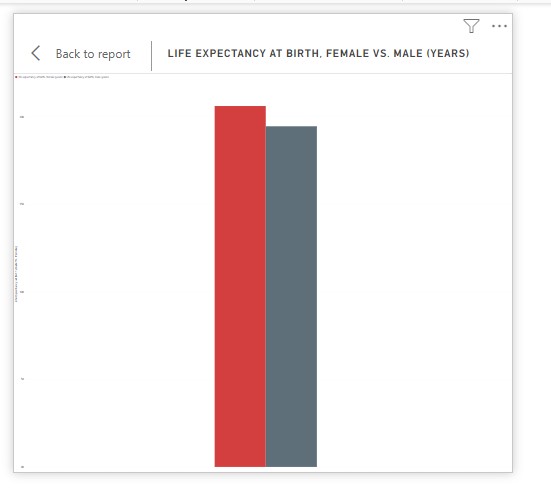


*Fig 9: [This visualization shows the health expenditure ratio and the poverty impact measure. The Current Health*

*Expenditure (% of GDP) was used in a dax measure to calculate the health expenditure ratio]*

Country and Gender Indicators:

These metrics are also primary indicators within this dataset. They are determinants in validating the analytical hypotheses throughout the analytical research.



*Fig 10: [This visualization shows the validation of the ‘Life Expectancy at Birth’ by gender differentiation]*

1.5 Discussion and Conclusion

The datasets utilized in this analysis was sourced from the World Bank data repository. The dashboard design process involved critical decision-making to ensure that the final product was not only visually appealing but also functionally robust. Challenges such as ensuring data integrity and selecting the most effective visual paradigms were encountered in building this dashboard. Despite these challenges, The dashboard effectively meets the required research objectives by providing a holistic view of dynamic health systems in various countries.

1.5a Critical Evaluation of Design Approach

Our design approach was anchored in best practices of data visualization and user interface design. While the dashboard effectively conveys complex data in an understandable format, there is always room for improvement, particularly in terms of user interactivity and real-time data integration.

A time series object for health expenditure per capita was created The final dashboard is a comprehensive tool that encapsulates a wide range of health metrics into a single, interactive platform. It serves as a valuable resource for policymakers, researchers, and public health professionals, offering insights into the intricate relationships between various health indicators and demographic factors. The dashboard stands as a testament to the power of effective data visualization in transforming raw data into meaningful information.