VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

CSE3040

Exploratory Data Analysis

Project Report

On

Exploring Health Trends Across the Globe

By

Tarlana Vidya

23MIA1176

Abstract:

This project presents a comprehensive Exploratory Data Analysis (EDA) on a global health statistics dataset, aiming to uncover critical patterns, relationships, and disparities in healthcare systems worldwide. The analysis is aligned with Sustainable Development Goal (SDG) 3: Ensure healthy lives and promote well-being for all at all ages.

The dataset includes both numerical and categorical data, allowing for a multidimensional exploration. Numerical variables are analysed using statistical measures such as mean, median, and standard deviation, while categorical data is assessed for unique value counts and frequency distributions. Univariate analysis explores individual variables to understand distributions and trends using histograms and bar charts, while bivariate analysis investigates relationships between factors like healthcare access and life expectancy using scatter plots and grouped bar charts.

The preprocessing stage ensures the dataset's quality and integrity. The missing values are treated using techniques like regression which we learned in class. Outlier detection, using methods like boxplots and statistical capping, ensures anomalies don't skew analysis. Correlation analysis, supported by correlation matrices and heatmaps, helps identify strong relationships among variables for better insights.

From the analysis, we derive key inferences: proper healthcare infrastructure and access strongly correlate with improved health outcomes, such as higher life expectancy and lower treatment costs. Furthermore, socio-economic conditions and healthcare spending play a crucial role in determining a country's overall health index. These insights can support evidence-based policy decisions and highlight areas needing improvement to achieve global health equity.

Overall, this project serves as a data-driven foundation for understanding and addressing global health challenges through visual and statistical exploration, contributing meaningfully to the achievement of SDG Goal 3.

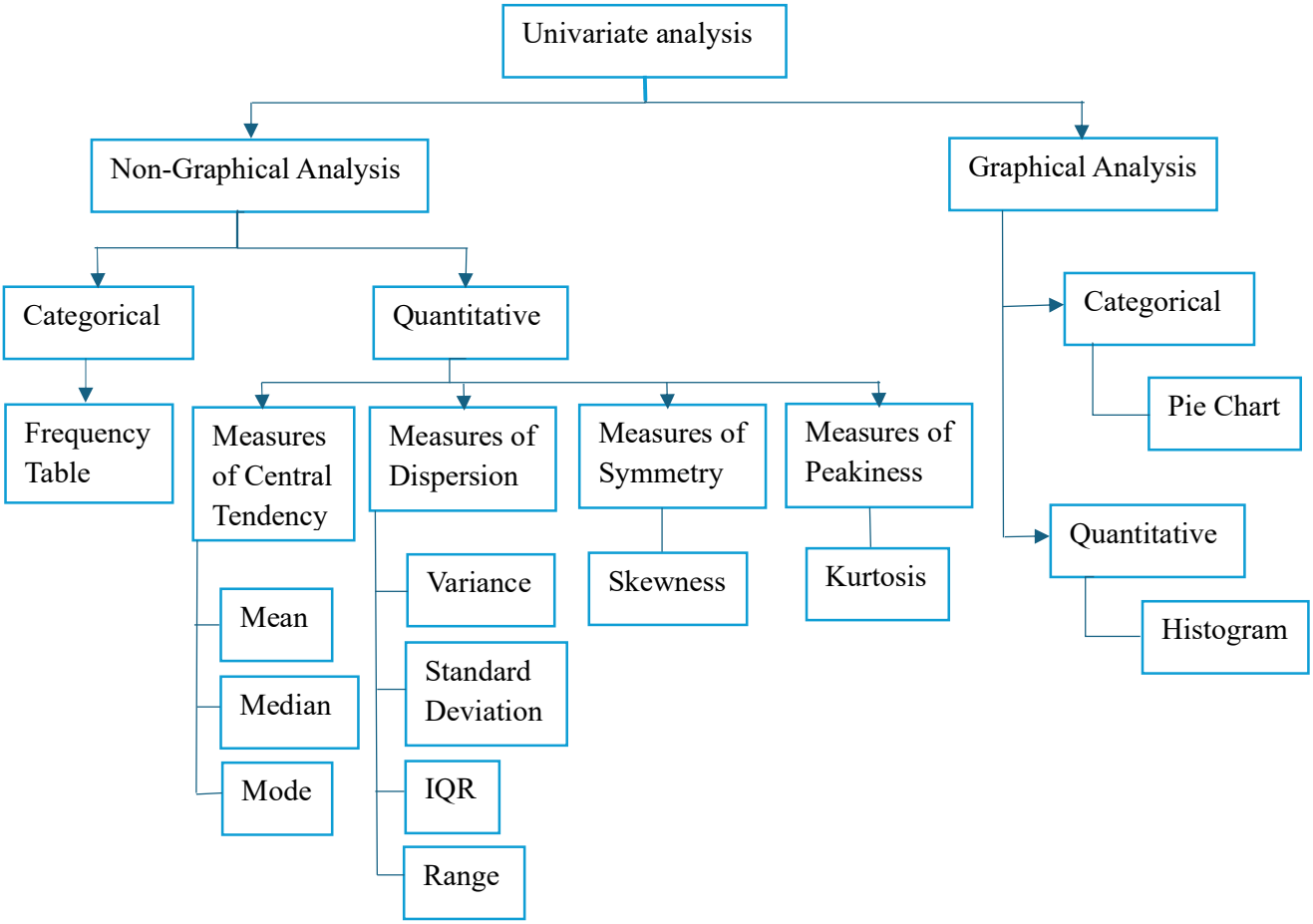
Objectives:

- To explore and analyse global health data using EDA techniques.
- To understand the statistical properties and relationships between healthcare variables.
- To identify missing values and apply imputation methods.
- To detect and handle outliers using advanced statistical and ML-based methods.
- To visualize data for better interpretation and communication.
- To align findings with the broader context of SDG Goal 3.

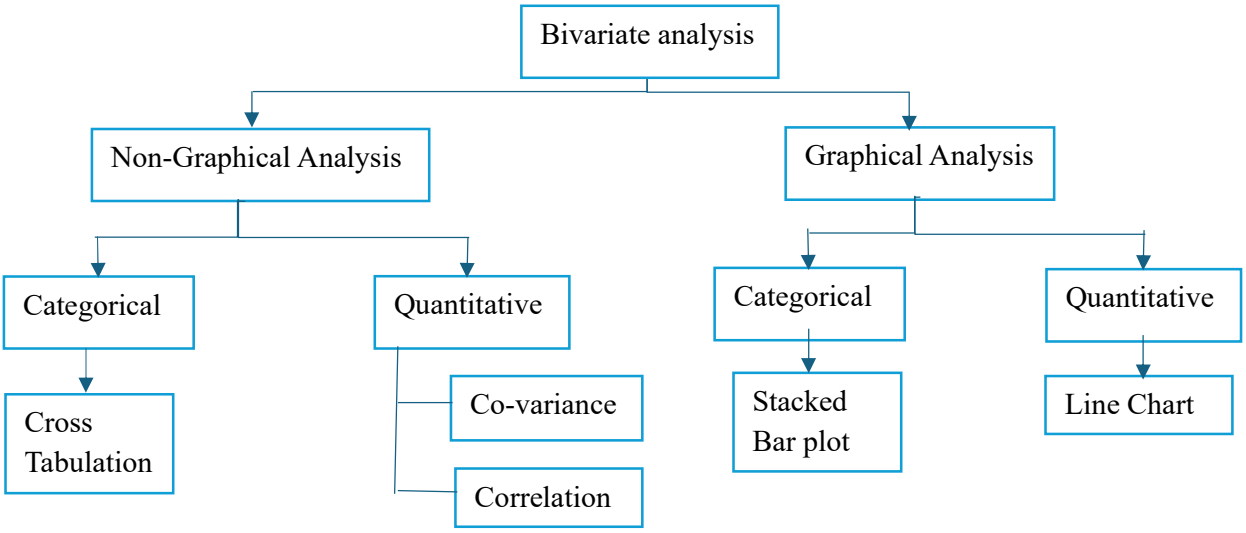
Proposed Methodology:

The project follows a step-by-step EDA pipeline:

- Data Type Identification: Understanding variable types (categorical/quantitative).
- Basic Metrics: Shape of the dataset, Summary statistics, head and the tail of the dataset.
- Univariate Analysis:



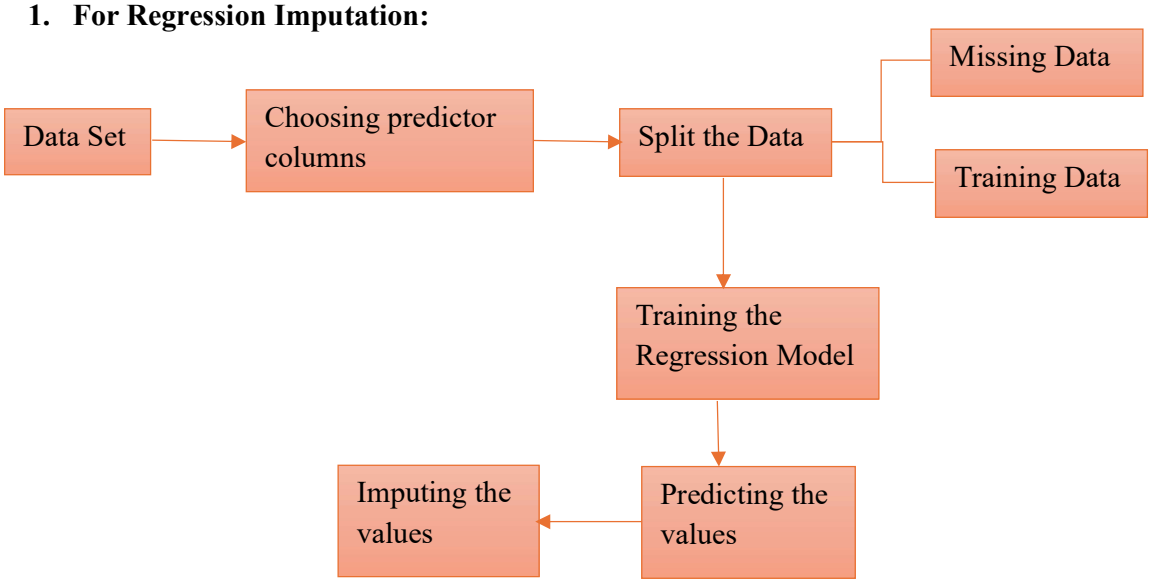
- Multivariate (Bivariate) Analysis:



- Missing Value Treatment: Linear regression imputation for key columns.
- Outlier Detection: Using Grub's Test, Chi-Square, LOF, KDE, and Convex Hull.
- Clustering: Using K-Means, hierarchical clustering, and dendrograms.
- Dimensionality Reduction: PCA for simplifying analysis.
- Transformation: Normalization and log transformation for skewed data.

System Architecture:

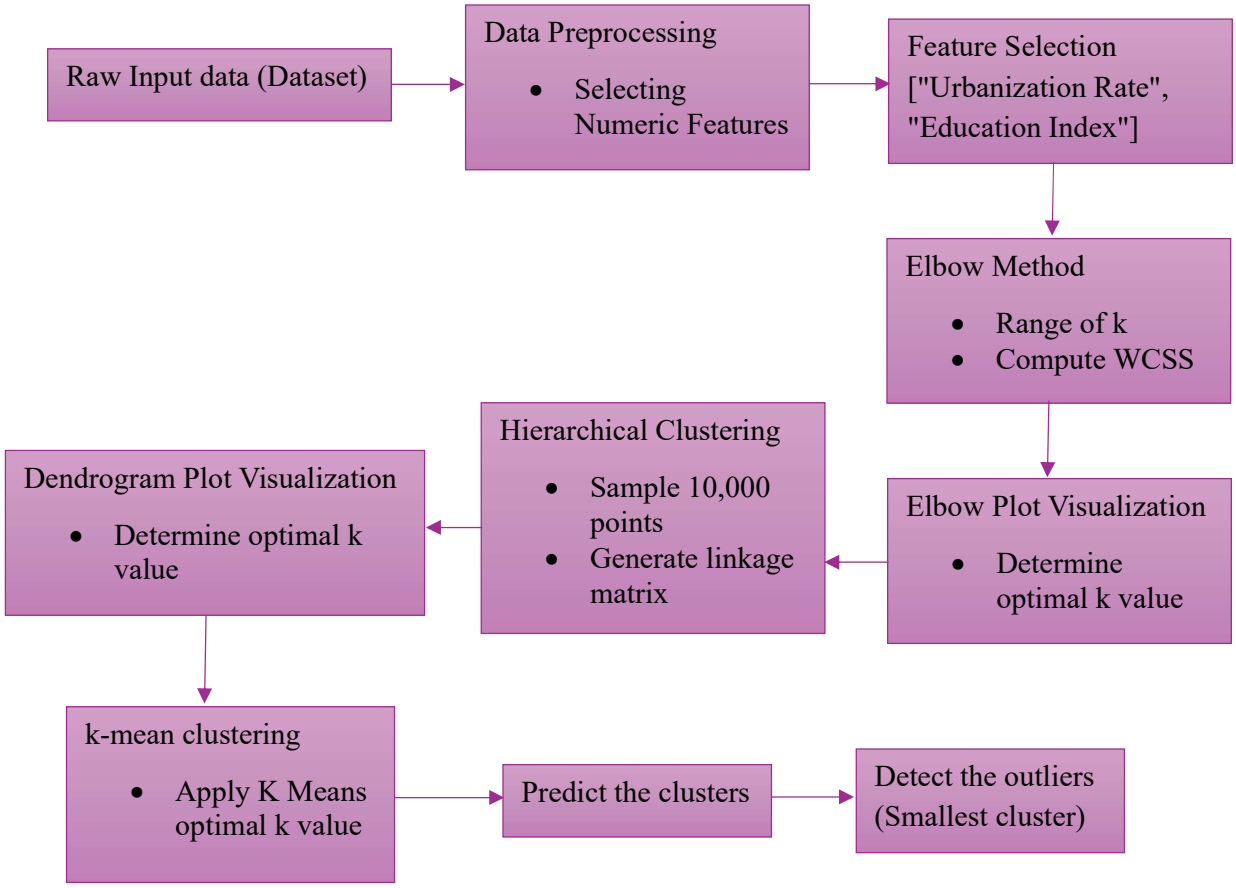
1. **For Regression Imputation:**



Regression Imputation is a method used to fill in missing data by predicting values based on existing patterns in a dataset. The system architecture follows a structured approach: First, predictor columns (features) are selected—these are attributes that could help estimate

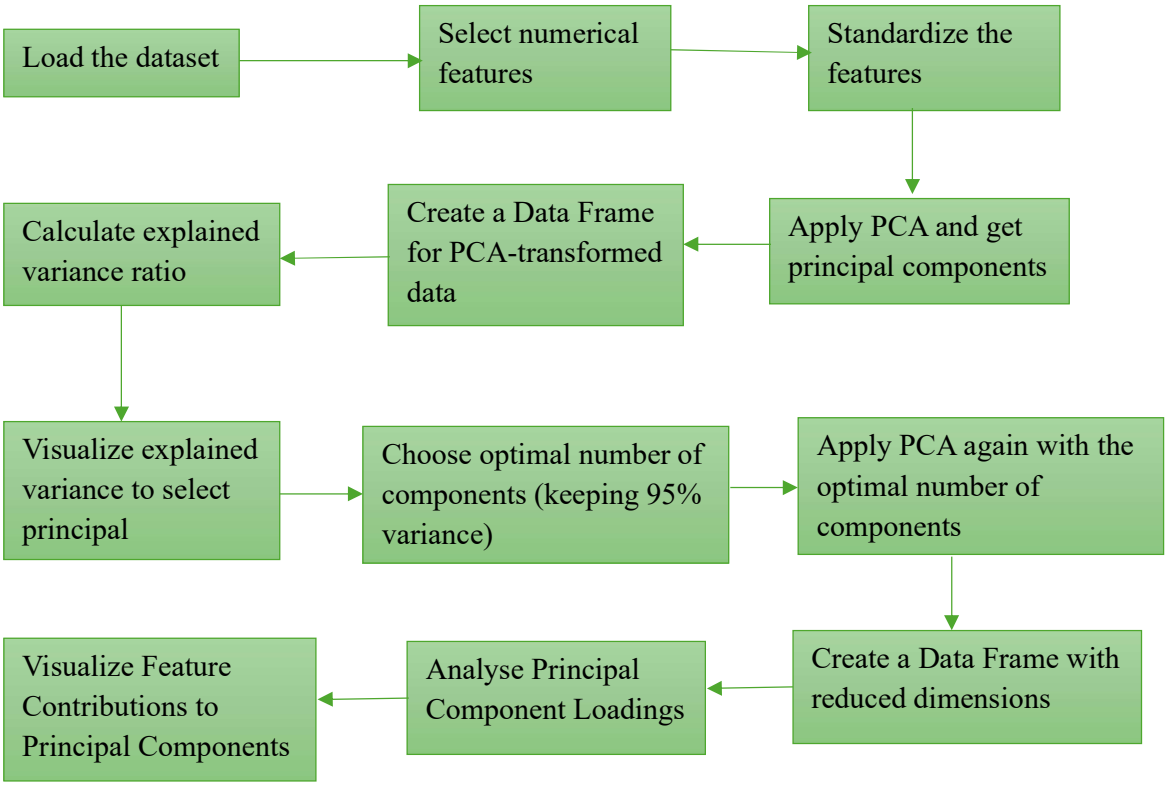
missing values. The dataset is then split into two parts: known values (for training) and missing values (to be predicted). A linear regression model is trained using the known data, learning relationships between features and the target variable (hospital beds per 1000). Finally, the trained model estimates missing values based on available predictor columns, effectively completing the dataset. The attached diagram visually represents this process, showing how the dataset progresses through selection, training, prediction, and imputation.

2. **For K-means Clustering**



K-means clustering is used for detecting outliers in a dataset by grouping data points into clusters based on similarity. The process starts with selecting relevant features (e.g., urbanization rate and education index). The Elbow Method determines the optimal number of clusters by analysing how variance decreases as the number of clusters increases. A hierarchical clustering dendrogram helps validate this selection visually. Once an optimal number of clusters (K=3) is chosen, K-means clustering is applied to group data points. Outliers are identified as the smallest cluster, where fewer data points exist, indicating anomalies. The attached diagram visually represents this step-by-step approach, making the clustering process and outlier detection clear.

3. **For PCA:**



The system architecture of Principal Component Analysis (PCA) follows a structured process to reduce dimensionality while retaining essential data patterns. It begins with loading the dataset and selecting numerical features for analysis. These features are then standardized to ensure consistency, preventing differences in scale from influencing the results. PCA is applied to generate principal components, which are new variables that summarize the dataset's variance using linear combinations of the original features. The explained variance ratio is calculated to determine how much information each principal component retains, helping to decide the optimal number of components. A visualization of the variance distribution guides the selection of components that preserve a significant portion—typically 95%—of the dataset's structure. Finally, PCA is reapplied with the chosen number of components, and the dataset is transformed into a lower-dimensional format. Feature contributions to principal components are analysed to understand their impact, enabling data-driven insights for clustering, visualization, and predictive modelling. This streamlined approach ensures efficiency while minimizing information loss in high-dimensional datasets.

Module Description:

About the dataset:

This dataset contains global health statistics, focusing on diseases, treatments, and healthcare infrastructure across various countries and years. It includes information such as disease

prevalence, mortality rates, recovery rates, healthcare access, treatment costs, and socio-economic factors. With over 1 million rows and 22 columns, the dataset is well-structured and comprehensive, making it suitable for health research, policy planning, and machine learning applications.

Column description:

🌍 **Country:** The nation where health data is recorded.

🗓️ **Year:** The specific year the data was collected.

🦠 **Disease Name:** The health condition being tracked.

📁 **Disease Category:** The type of disease (e.g., Infectious, Non-Communicable).

📊 **Prevalence Rate (%):** The percentage of the population affected by the disease.

📈 **Incidence Rate (%):** The percentage of new cases or diagnoses.

💀 **Mortality Rate (%):** The percentage of affected individuals who die from the disease.

👶 **Age Group:** The age range most affected by the disease.

🚻 **Gender:** The gender(s) impacted (Male, Female, or Both).

👥 **Population Affected:** The total number of individuals suffering from the disease.

🏥 **Healthcare Access (%):** The percentage of the population with healthcare access.

🩺 **Doctors per 1000:** The number of doctors available per 1000 people.

🛏️ **Hospital Beds per 1000:** The number of hospital beds per 1000 people.

💊 **Treatment Type:** The primary treatment method for the disease (e.g., Medication, Surgery).

💵 **Average Treatment Cost (USD):** The typical cost of treating the disease in USD.

💉 **Availability of Vaccines/Treatment:** Whether vaccines or treatments are accessible.

🏥 **Recovery Rate (%):** The percentage of individuals who recover from the disease.

📉 **DALYs (Disability-Adjusted Life Years):** A measure of the overall disease burden.

📈 **Improvement in 5 Years (%):** The progress in disease outcomes over the past five years.

💰 **Per Capita Income (USD):** The average income per person in the country.

🎓 **Education Index:** The country's average level of education.

🏙️ **Urbanization Rate (%):** The percentage of the population living in urban areas.

Steps in Data Exploration:

- **Identification of variables and data types:**
  This module identifies and classifies the variables in the dataset to understand their types and roles. It separates columns into categorical and numerical using data types, then further categorizes them as nominal, ordinal, interval, or ratio based on their

characteristics. This helps in selecting appropriate analysis techniques for each type of data.

- **Analysing the basic metrics:**
  This module analyses the basic structure and content of the dataset by displaying the number of rows and columns, generating summary statistics for numerical columns, and showing the first and last few rows. This provides a quick overview of the dataset's size, distribution, and data values.

- **Non-Graphical Univariate Analysis:**
  This module performs non-graphical univariate analysis to understand the distribution of individual variables. For categorical data, it uses a frequency table to count occurrences of each category. For numerical data, it calculates central tendency (mean, median, mode), dispersion (variance, standard deviation, range, interquartile range), symmetry (skewness), and peakedness (kurtosis) to summarize and describe data patterns.

- **Graphical Univariate Analysis:**
  This module uses graphical univariate analysis to visualize the distribution of individual variables. Categorical data is represented using pie charts to show the proportion of each category, while quantitative data is visualized with histograms to display the frequency distribution of numerical values, making it easier to interpret patterns and trends.

- **Bivariate Analysis:**
  - **Non-Graphical Bivariate Analysis:**
    This module examines the relationship between two variables without visual representation. For categorical variables, it uses cross-tabulation to display the frequency distribution of combinations of categories, helping to identify potential associations. For numerical variables, it calculates covariance and correlation to measure the strength and direction of linear relationships between variables.
  - **Graphical Bivariate Analysis:**
    This part visualizes the relationships between pairs of variables. For categorical data, a stacked barplot is used to compare the distribution of one category within another, making trends easier to spot. For numerical data, line charts help track changes or patterns over a continuous variable, highlighting trends and relationships between two quantitative variables.

- **Missing value treatment:**
  This module addresses missing data in the dataset using Linear Regression Imputation. It predicts and fills missing values of a variable by modelling it as a linear function of other relevant variables, ensuring minimal loss of information and maintaining data consistency for accurate analysis.
- **Outlier Analysis and Treatment:**
  - **Outlier Analysis:**
    Outlier analysis focuses on identifying data points that significantly differ from the rest of the dataset, as they can distort results and lead to inaccurate conclusions. A variety of methods are employed to detect these anomalies. Statistical methods include Grub’s Test, which is used to detect outliers in

univariate data, and the Chi-Square Test for identifying outliers in multivariate contexts. Graphical methods such as Box Plots help visually identify univariate outliers, while Hexbin Plots are useful for analysing the density of bivariate data and spotting regions with sparse points.
In addition, several advanced techniques are used. The Depth-Based Method, like the Convex Hull approach, identifies points that lie at the boundary of the data distribution. The Distance-Based Method, such as Local Outlier Factor (LOF), evaluates the local density of each point to find deviations from the surrounding neighbourhood. The Density-Based Method, particularly Kernel Density Estimation (KDE), estimates the probability density function to detect low-density regions as potential outliers.
Clustering-Based Methods are also applied. First, the Elbow Method is used to determine the optimal number of clusters. Then, Hierarchical Clustering is performed and visualized through a dendrogram to understand the natural groupings within the data. Finally, K-Means Clustering is applied using the optimal number of clusters to isolate outliers that do not belong to any cluster effectively. For categorical data, the Frequent Pattern Outlier Factor (FPOF) is utilized to detect unusual or infrequent category combinations that deviate from common patterns.

- **Outlier Treatment:**
  Once potential outliers are identified, treatment involves assessing their impact and deciding whether to retain, modify, or remove them. In this analysis, the Interquartile Range (IQR) method was used to check for outliers in numerical variables. This method defines outliers as values that fall below the first quartile (Q1) minus 1.5 times the IQR or above the third quartile (Q3) plus 1.5 times the IQR. After applying this method, it was found that no numerical columns contained outliers. Therefore, no further treatment or adjustments were necessary, ensuring that the dataset remains intact and reliable for further analysis.

- **Correlation Analysis:**
  This module explores the strength and direction of relationships between numerical variables using a correlation matrix. The matrix displays correlation coefficients ranging from -1 to 1, where values close to 1 indicate a strong positive relationship, values close to -1 indicate a strong negative relationship, and values near 0 suggest no linear relationship. This helps in identifying patterns, multicollinearity, and selecting relevant features for modelling.

- **Dimensionality Reduction:**
  This module applies Principal Component Analysis (PCA) to reduce the number of features in the dataset while retaining most of the original information. PCA transforms the original variables into a new set of uncorrelated variables called principal components, which capture the maximum variance in the data. This technique helps simplify complex datasets, improve model performance, and reduce computational cost by eliminating redundant or less important features.

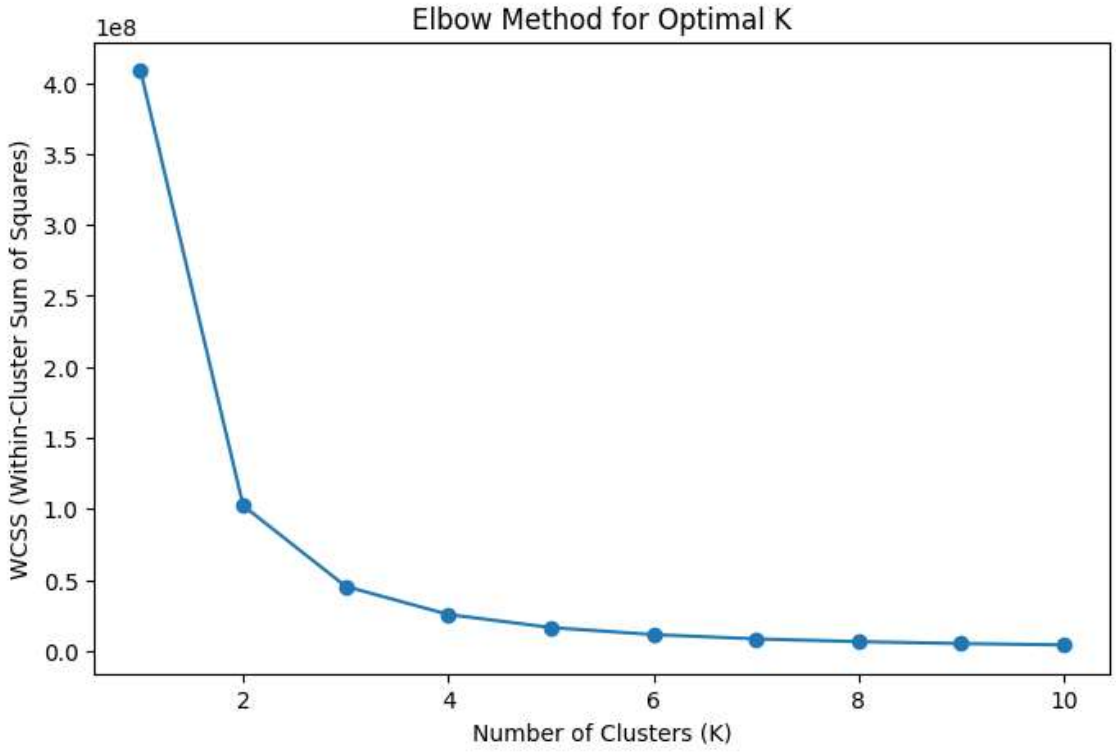
- **Variable transformations:**
  This module enhances data quality and model readiness by applying various transformation techniques to prepare the dataset for analysis. It begins with a log transformation on the "Population Affected" column to reduce skewness and

normalize large values. Standardization is then applied to all numerical columns using StandardScaler, ensuring each variable has a mean of 0 and a standard deviation of 1, which is crucial for many machine learning algorithms.
Next, categorical variables are transformed: One-Hot Encoding is used for "Disease Category" and "Country" to convert them into binary vectors, while Label Encoding is applied to "Age Group" to convert ordinal categories into numerical labels.
Additionally, feature binning is performed on "Urbanization Rate (%)", categorizing it into "Low", "Medium", and "High" levels for easier interpretation and potential model improvement. These transformations collectively ensure the dataset is well-structured and suitable for advanced analysis or modelling.

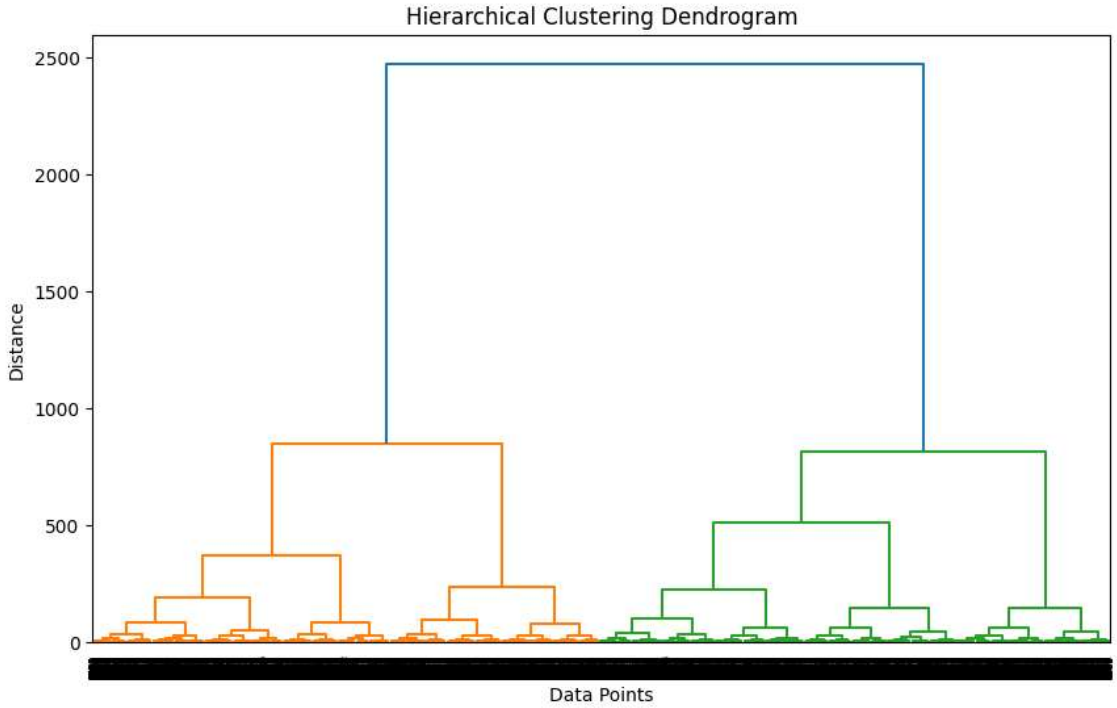
Results:

- **K means Clustering**

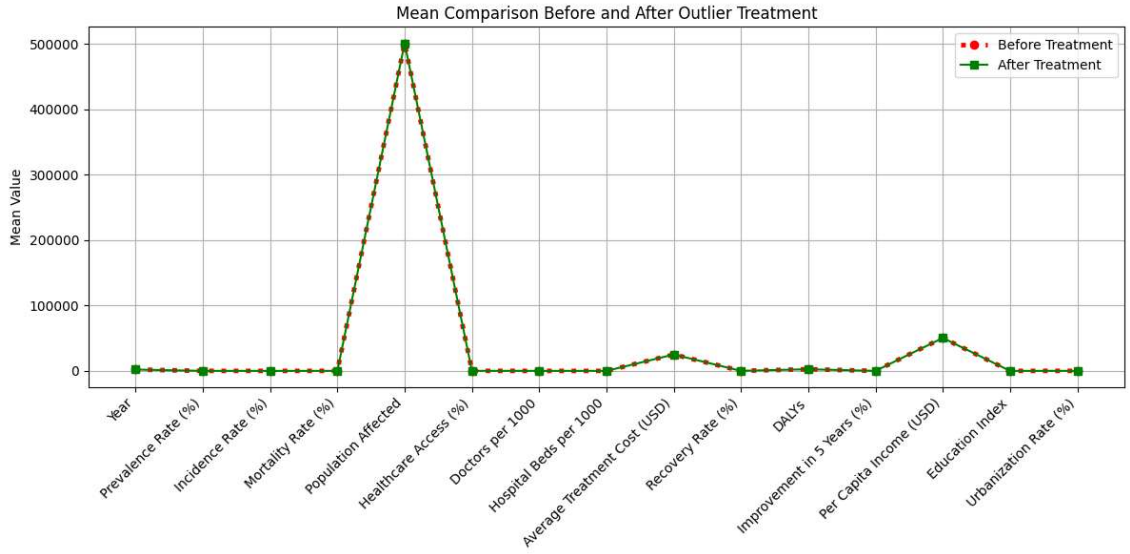
Elbow Method



Dendrogram

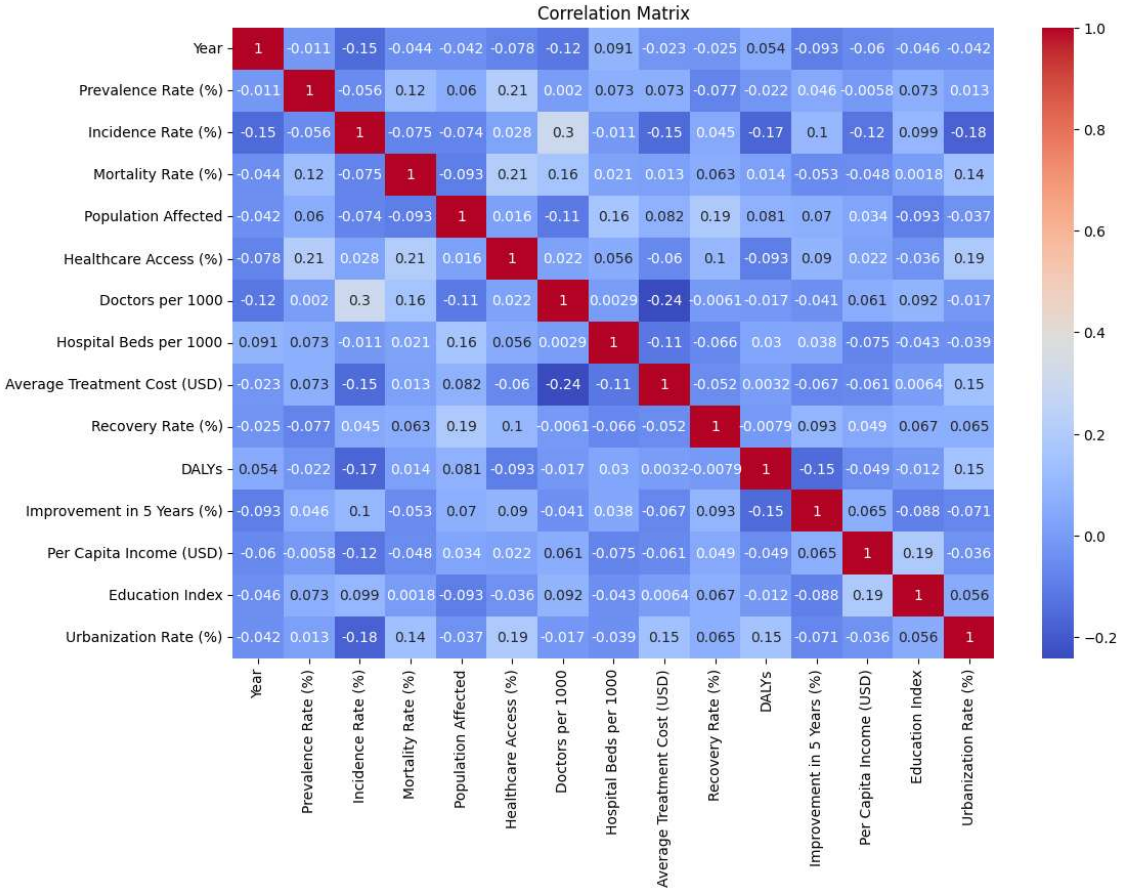


- **Outlier Treatment**



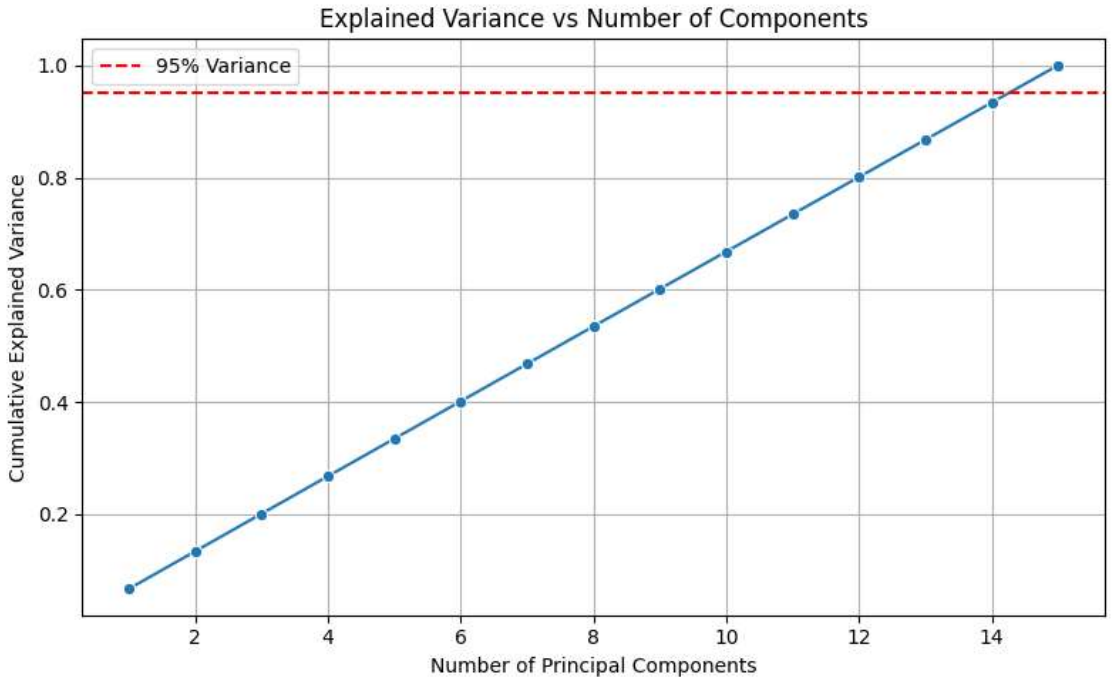
- **Correlation Analysis**

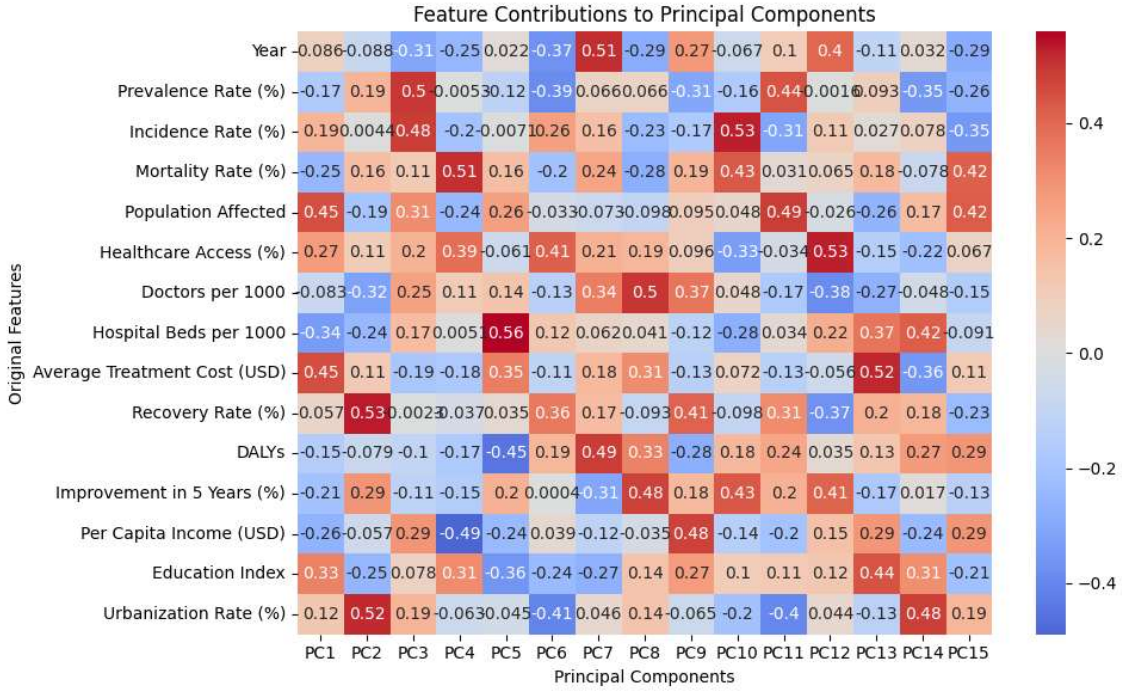
Correlation Matrix



- **Dimensionality Reduction:**

Principal Component Analysis (PCA)





Conclusion:

The exploratory data analysis (EDA) project provided valuable insights into the dataset by systematically exploring its structure, relationships, and underlying patterns. By identifying variables and data types, we established a clear understanding of the dataset's composition, ensuring proper selection of transformation techniques. Analysing basic metrics helped uncover initial trends, distributions, and potential inconsistencies in the data. Through univariate analysis, both graphical and non-graphical, key statistical properties of individual variables were examined, revealing skewness, variability, and general data distribution.

A bivariate analysis further allowed us to study relationships between different variables, identifying significant dependencies that influence health outcomes. Variable transformations were applied to improve data consistency, making features more suitable for analysis. Missing value treatment and outlier detection ensured data integrity by addressing gaps and anomalies that could skew results. Correlation analysis played a crucial role in understanding variable interactions, helping to identify redundant features and guiding dimensionality reduction.

The final step of dimensionality reduction streamlined the dataset by retaining the most critical components, enabling a more efficient and interpretable dataset. These explorations reinforced the significance of data preprocessing and feature refinement in producing reliable and actionable insights. The findings from this project demonstrate how structured, well-explored data can aid in decision-making, research, and policy planning by improving the accuracy and effectiveness of subsequent modelling techniques. Future analyses could expand on these techniques by incorporating time-series trends or clustering approaches to reveal further insights into global health patterns.

References:

GitHub Link: https://github.com/Vidya-7777/EDA-Project

Dataset Link: https://www.kaggle.com/datasets/malaiarasugraj/global-health-statistics

PPT Link: https://www.canva.com/design/DAGVh3IqXy0/DJleuedDfMp-8xY_-MUfbQ/edit?utm_content=DAGVh3IqXy0&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

- Class Notes and the material uploaded in LMS
- https://link.springer.com/chapter/10.1007/978-3-031-54111-7_8?utm_
- https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0201950&utm_
- https://www.sciencedirect.com/science/article/abs/pii/S1053482216300353?utm_
- https://arxiv.org/abs/1811.12199?utm_
- https://www.researchgate.net/publication/348871497_Exploratory_Data_Analysis_Based_on_Remote_Health_Care_Monitoring_System_by_Using_IoT