VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

Exploring Health Trends Across the Globe using Exploratory Data Analytics

Tarlana Vidya | Dr. Asnath Victy Phamila Y | School of Computer Science and Engineering

Motivation/ Introduction

In today's data-driven world, uncovering hidden patterns and insights from raw data is crucial for informed decision-making. This project leverages EDA techniques to deeply understand a healthcare dataset—identifying trends, relationships, and anomalies. By transforming and refining the data, we aim to prepare a strong foundation for predictive modelling and meaningful interpretations that can aid research, policy planning, and healthcare improvements.

Scope of the Project

This project focuses on performing comprehensive EDA to understand the structure, trends, and relationships within a healthcare dataset. It includes variable identification, missing value and outlier treatment, correlation analysis, and dimensionality reduction. The scope also extends to transforming variables for better model compatibility, laying a strong foundation for future predictive analysis and decision-making in public health research.

Methodology

Variable Identification & Classification: Classify variables into categorical and numerical types for appropriate analysis.

Basic Metrics Analysis: Evaluate the structure, summary statistics, and initial trends of the dataset.

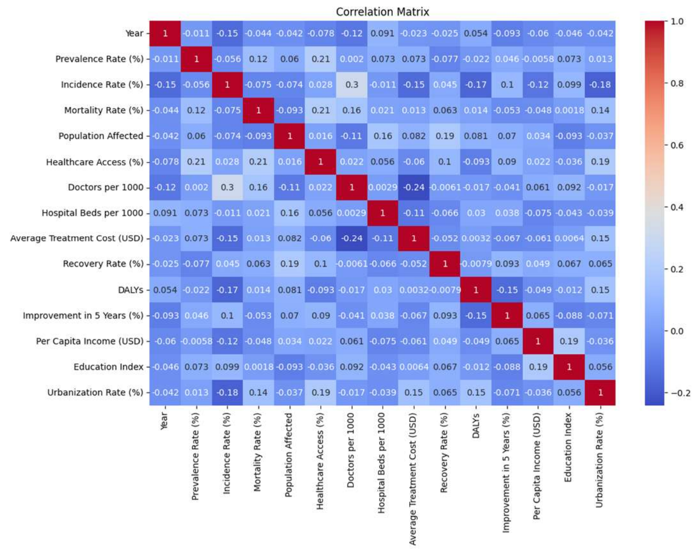
Univariate Analysis: Perform both graphical and non-graphical analyses to examine the distribution and key properties of individual variables.

Bivariate Analysis: Analyse relationships between pairs of variables using both graphical and non-graphical techniques.

Missing Data Treatment: Use Linear Regression Imputation to handle missing values.

Outlier Detection & Treatment: Apply statistical and clustering techniques to identify and manage outliers.

Correlation Analysis: Examine correlations between variables to identify potential relationships and reduce multicollinearity.

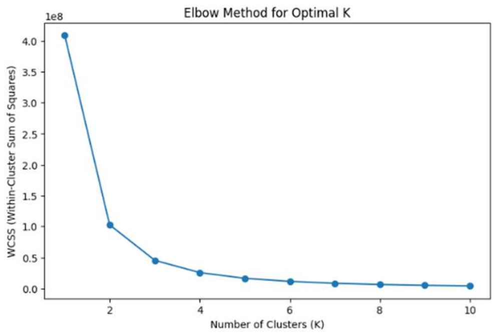


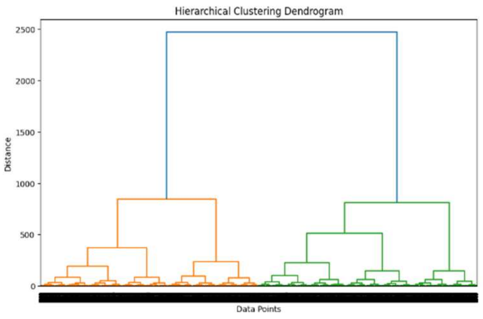
Dimensionality Reduction: Use Principal Component Analysis to reduce the number of features while retaining important information.

Data Transformation: Standardize, encode, and bin variables to prepare the dataset for modelling.

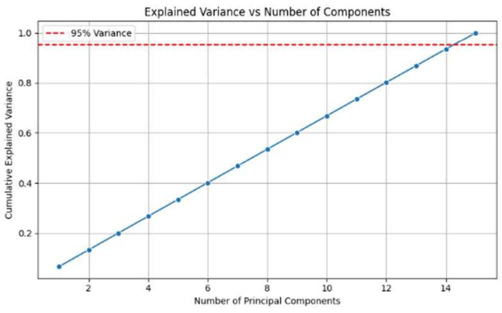
Results

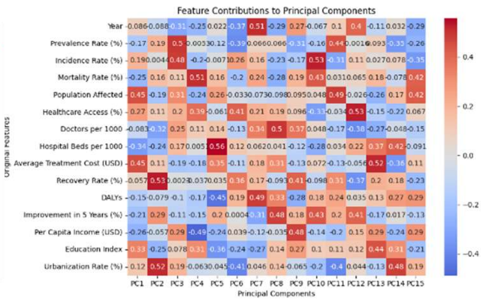
K-Means Clustering for Outlier Detection: K-Means effectively grouped data into optimal clusters, allowing the identification of data points that did not fit into any cluster as potential outliers.





PCA for Dimensionality Reduction: PCA reduced the dataset to key components, retaining maximum variance while simplifying the feature space for efficient analysis.





Conclusion and Summary

This project demonstrated the power of Exploratory Data Analysis in uncovering hidden patterns, identifying inconsistencies, and preparing data for advanced modelling. By thoroughly exploring data types, handling missing values and outliers, and applying transformations and dimensionality reduction techniques like PCA, we enhanced the quality and structure of the dataset. The insights gained provide a solid foundation for informed decision-making and future predictive modelling, highlighting the importance of a well-executed EDA in any data science workflow.

Contact Details

tarlana.vidya2023@vitstudent.ac.in

Acknowledgments/ 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/DAGVh3lqXy0/DJleuedDfMp-8xY_-MUfbQ/edit?utm_content=DAGVh3lqXy0&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton