Report On Visualization Assignment -Lab2(b)

Dataset Topic: Life Expectancy Analysis

Source of Dataset: Kaggle

Link of Dataset: https://www.kaggle.com/code/yashgupta261100/life-expectancy-analysis/input?select=Life+Expectancy+Data.csv

Attributes:

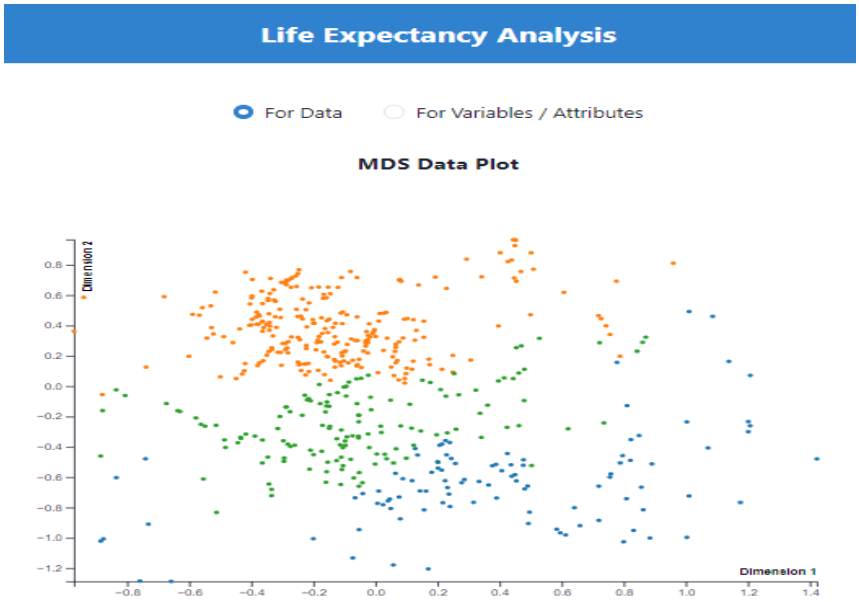
1. **Country:** Name of the country.
2. **Year:** Year of observation.
3. **Status:** Development status of the country (Developed or Developing).
4. **Life Expectancy:** The average number of years a person is expected to live.
5. **Adult Mortality:** Probability of dying between 15 and 60 years per 1000 population.
6. **Infant Deaths:** Number of infant deaths per 1000 population.
7. **Alcohol:** Alcohol consumption measured in liters per capita.
8. **Hepatitis B:** Hepatitis B immunization coverage among 1-year-olds (%).
9. **BMI:** Average Body Mass Index of the entire population.
10. **Polio:** Polio immunization coverage among 1-year-olds (%).
11. **Total Expenditure:** Skipped here.
12. **Diphtheria:** Diphtheria immunization coverage among 1-year-olds (%).
13. **HIV/AIDS:** Deaths per 1000 live births due to HIV/AIDS (0-4 years).
14. **Human Development Groups:** Human Development status of the country (Low, Medium, High, Very High).
15. **Population:** Skipped here.
16. **Income Composition of Resources:** Skipped here.
17. **Schooling:** Number of years of Schooling.

The purpose of this report is to analyze and visualize a life expectancy analysis dataset using Multidimensional Scaling (MDS) plots and Parallel Coordinates Plot (PCP).

1.1 MDS Plot of Data

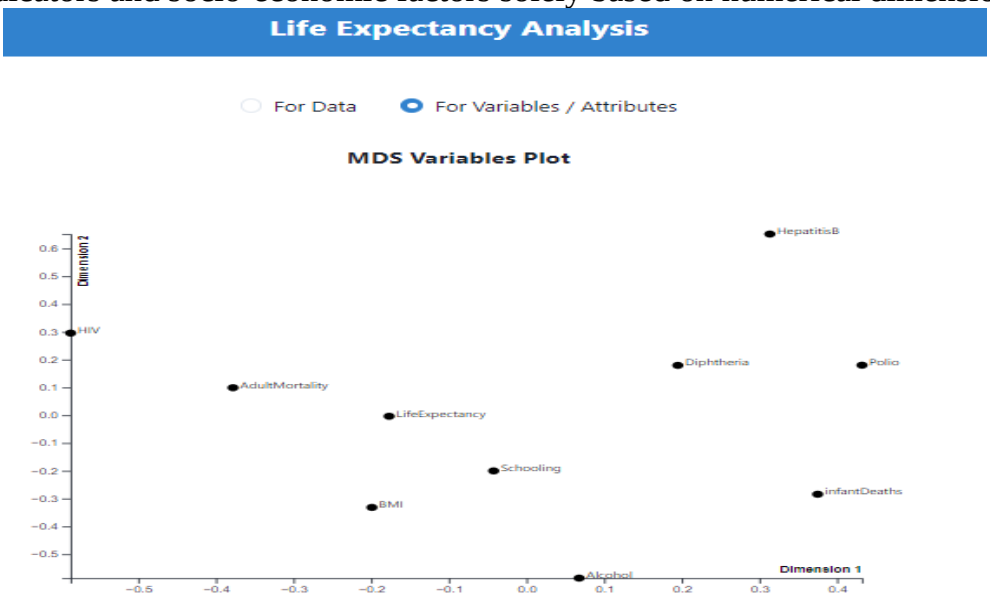
To construct the MDS plot of the data, we focused solely on numerical data dimensions, excluding categorical variables such as country, year, status, and human development groups. This allowed us to visualize the underlying structure of the dataset in a two-dimensional space based solely on numerical attributes such as life expectancy, adult mortality, infant deaths, alcohol consumption, and others. The scatterplot colored the data points by cluster ID, highlighting similarities and differences in health indicators and development status among countries. Notably, developed countries tended to cluster

together, while developing nations formed distinct clusters, revealing disparities in health outcomes and socio-economic indicators solely based on numerical dimensions.

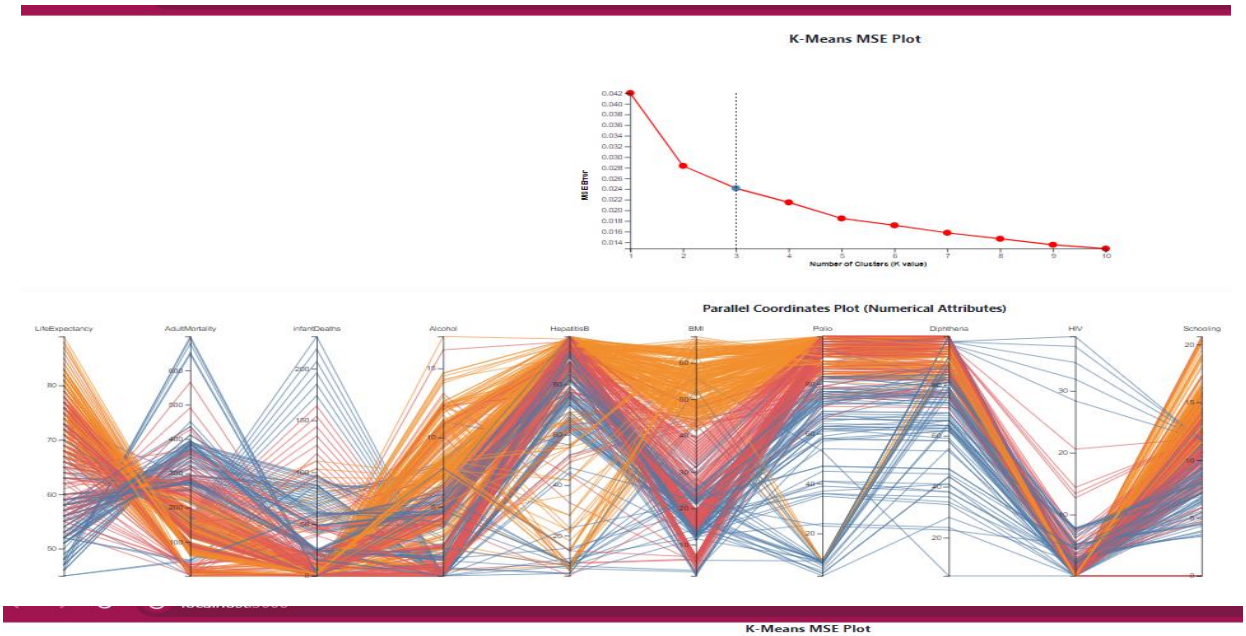


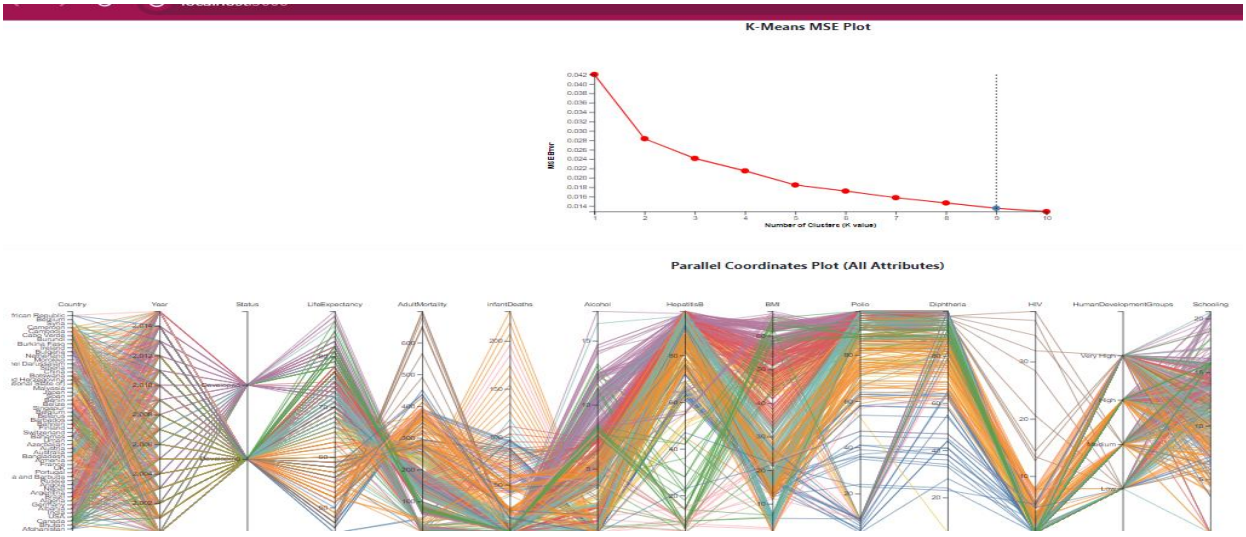
1.2 MDS Plot of Variables

For the MDS plot of variables, we focused exclusively on numerical variables to visualize their relationships in a two-dimensional space. Categorical variables such as country, year, status, and human development groups were excluded from this analysis. By calculating the correlations between numerical variables and computing the dissimilarity matrix using the (1-|correlation|) distance metric, we identified clusters of variables that exhibited strong associations with each other. The scatterplot allowed us to observe patterns of correlation among numerical variables, providing insights into the interrelationships between different health indicators and socio-economic factors solely based on numerical dimensions.

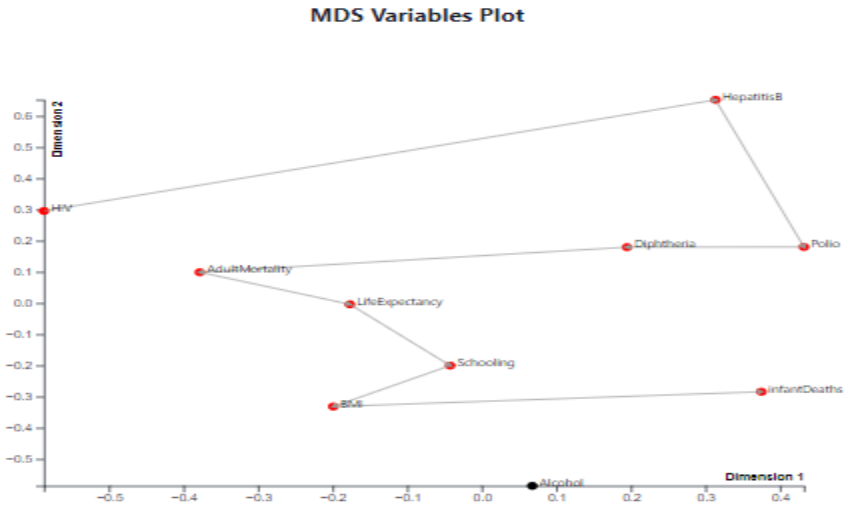


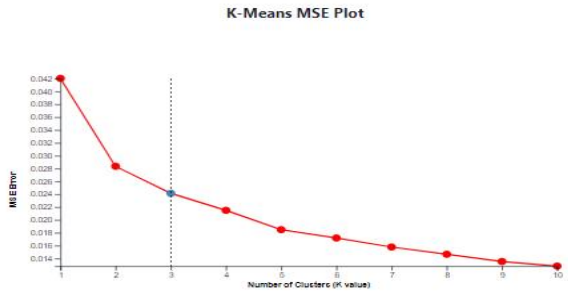
2.1 Parallel Coordinates Plot: The parallel coordinates plot provides a comprehensive visualization of all dimensions of the dataset, including both numerical and categorical variables. Each line represents a country, and the position of the line on each axis corresponds to the value of the respective variable for that country. The lines are colored based on the development status of countries, allowing for the identification of clusters and patterns. By interactively rearranging axes based on user input, meaningful axes ordering can be determined, reflecting the underlying relationships between variables. For example, arranging axes based on correlations observed in the MDS plot of variables can help in highlighting clusters of variables that are strongly associated with each other. Coloring the polylines by cluster ID further enhances the visualization, enabling the identification of distinct clusters of countries with similar characteristics. This facilitates the comparison of health outcomes and development status between different clusters and provides valuable insights for targeted interventions and policy formulation.

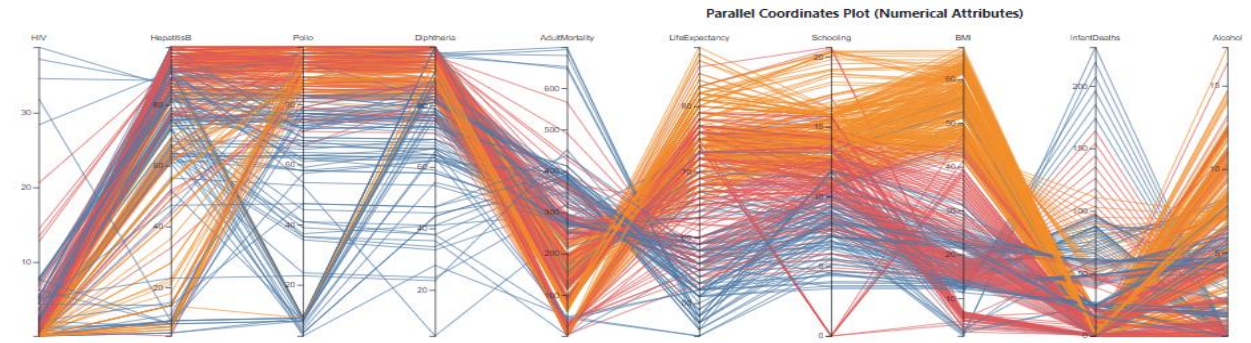




3.1 PCP Axes Ordering from Correlations: This task aims to establish a meaningful axis ordering for the parallel coordinates plot (PCP) by leveraging correlations observed in the MDS plot of numerical variables. This process focuses solely on numerical values and employs user interaction to guide axis arrangement. Users click on points in sequence, indicating their preference for specific variables. The PCP then organizes axes based on these preferences, aligning them according to the correlations observed in the MDS plot. This approach enhances the interpretability of the PCP, enabling users to identify clusters of variables with similar characteristics more efficiently. Ultimately, it facilitates deeper exploration and understanding of the dataset, aiding in informed decision-making and policy formulation in the realm of global health and development.







4. Additional observations of Implementation

Upon analyzing the life expectancy dataset, several intriguing observations emerge from the MDS plots and parallel coordinates plot (PCP):

- **Clusters in MDS Plot of Data:** Developed countries cluster together, separate from developing nations, indicating disparities in health outcomes and socio-economic indicators.
- **Variable Associations in MDS Plot:** Certain clusters of variables exhibit strong associations, such as immunization coverage variables clustering together and variables related to mortality rates forming another cluster.
- **User Interaction in PCP Axis Ordering:** The implementation of user interaction enhances flexibility, allowing users to focus on specific variables of interest and uncover insights tailored to their preferences.
- **Insights from PCP Visualization:** Color coding the polylines by cluster ID facilitates the identification of distinct patterns and clusters within the data, aiding comparisons across different development statuses. These implementation strategies contribute to a more insightful and interactive exploration of the life expectancy dataset, enabling a deeper understanding of the factors influencing health outcomes and development status across countries.