Project Documentation: Clustering Analysis of Global Causes of Death

1. Project Overview

- Data mining project analyzing global causes of death using clustering techniques.
- Uncovers patterns and country groupings based on health statistics.
- Combines Python-based preprocessing and an interactive web dashboard.

2. Dataset Description

- Countries and territories.
- Death rates for major causes (cardiovascular, cancer, diabetes, accidents, etc.).
- Time span from 2015 to 2022.
- Year–country records; each row represents a country in a given year.

3. Tools and Technologies Used

Backend (Data Analysis)

- Python
- Pandas for data manipulation
- Scikit■learn for normalization and K■Means clustering
- Matplotlib / Seaborn for preliminary plots

Frontend (Visualization Dashboard)

- HTML, CSS, JavaScript
- Plotly.js for interactive charts
- D3.js for dynamic data manipulation
- Responsive UI with year and country filters

4. Data Mining Techniques Applied

Data Preprocessing

- Filter by year range and country.
- Z■score normalization of death■rate features.
- Handle missing values (if any).

Clustering

- K■Means with 3 clusters (default).
- Euclidean distance metric.

- Iterative refinement until convergence or 100 iterations.
- PCA approximation to 2D for visualization.

5. Dashboard Features

Controls

- Upload CSV file.
- Filter by country and year range.
- Run clustering analysis.
- Download clustered results.

Visualizations

- Cluster Scatter Plot (2■D PCA).
- Bar Chart of average causes by cluster.
- Choropleth map of clusters by country.
- Stats cards: record count, year range, clusters.
- Data table (first 100 rows).

6. Output and Interpretation

- Dataset labeled by cluster ID per country■year.
- Identify health patterns across countries.
- Inform public■health policy.
- Compare temporal health trends.

7. Requirements to Run the Code

Jupyter Notebook Environment

- Python 3.8+ environment.
- Install libraries: pandas, numpy, matplotlib, seaborn, scikit■learn
- Launch with `jupyter notebook` and open `data_mining_2.ipynb`.