Student Name: Aniruddha Joshi

10 questions and answers:

1. What specific socio-economic and health indicators were used to cluster the countries?

Below specific socio-economic were used in project

* 1. child\_mort: Child mortality rate
  2. exports: Exports as a percentage of GDP
  3. health: Health spending
  4. imports: Imports as a percentage of GDP
  5. income: Net income per person
  6. inflation: Annual inflation rate
  7. life\_expec: Average life expectancy
  8. total\_fer: Total fertility rate
  9. gdpp: GDP per capita

1. How was the missing data handled?

Missing data was handled by filling the gaps with the average values for each respective column.

1. The dataset has only 167 records, which is a relatively small sample size. How might this small sample size affect the accuracy?

The small sample size potentially limit the accuracy of the clustering model.

With fewer data points, the model may not capture all scenarios leading to less precise clusters.

1. What methods were used to determine the optimal number of clusters?

Elbow method: Identifying the "elbow" point where adding more clusters doesn't significantly reduce the sum of squares.

Silhouette analysis: Measuring how similar an object is to its own cluster compared to other clusters.

1. What visualization were used and why those were selected?

Scatter plots: To explore correlations between variables.

Histograms: To understand the distribution of individual variables.

Heatmap: To visualize the correlation matrix, showing relationships between all pairs of variables.

World maps: To visualize child mortality and GDP geographically.

1. Why k=4 selected? What were the challenges in selecting K?

k=4 was selected because both the elbow method and silhouette analysis indicated it as the optimal number of clusters.

Challenges in selecting k: The "elbow" in the elbow was ambiguous, making it challenging to pinpoint the value of k.

1. How were the results of the clustering model evaluated?

The results were evaluated by using Silhouette analysis.

1. What are the specific limitations of the project?

Small sample size (167 records): May not fully capture the diversity of countries.

Limited number of variables (10 columns): May not fully represent the complexity of socio-economic and health needs.

1. What are the next steps in the project?

Expanding the dataset with more countries and variables to improve model accuracy and applying the clustering model to other socio-economic and health datasets.

1. How can the results of this clustering be used to inform policy decisions?

Identifying countries with similar socio-economic and health profiles allows for tailoring interventions to specific needs and prioritizing funding allocation based on cluster characteristics, focusing on regions with the greatest needs.