Student Name: Aniruddha Joshi

Data Set: <https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data/data>

**Milestone 1 - Proposal**

**Project Topic:** Clustering Countries for Targeted Socio-Economic and Health Interventions

**Business Problem:** Many countries face significant challenges in meeting the socio-economic and health needs of their populations. Limited resources necessitate a targeted approach to development and funding. This project aims to categorize countries based on key indicators to identify specific needs and prioritize interventions in areas such as infrastructure, health, and education.

**Datasets:**

**Unsupervised Learning on Country Data (Kaggle):** is the reference data set selected for the project. The dataset contains around 167 records of country information such as child mortality, exports, health, import, income, inflation, life expectancy, gdp etc.

1. **Link:** <https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data/data>
2. **Format:** the data format is CSV file
3. **Contents:** upto 10 variables, such as child mortality, GDP per capita, life expectancy, literacy rate, access to clean water

**Methods:**

1. **Data Exploration:**
   1. By using Pandas, NumPy and seaborn(sns) in Python to clean the dataset, remove invalid/incomplete rows, fillup the missing data with mean/averages. Converting the categorical data if any. If its categorical data then converting it to one-hot encoding as required.
   2. Visualizing the data by using Matplotlib, plotly and/or Seaborn to identify any relationships. Different visualizations shall be created such as scatter plots to explore correlations, histograms to understand distributions, box plots to compare groups.
2. **Feature Engineering:**
   1. As we have more than 10 variables, a feature engineering may not be required. However, some columns that are unrelated may have to be reviewed based on the cross-reference matrix review.
3. **Machine Learning Modeling:**
   1. Training the model using machine learning algorithms such as k-means clustering to categorize the data. Optimal number of clusters to be evaluated by using elbow method and silhouette analysis.
   2. Scikit-learn or the TensorFlow libraries may be used.
4. **Model Evaluation and Selection:**
   1. Evaluating the model performance using metrics like accuracy and loss.

**Ethical Considerations:**

1. **Data Bias**: Acknowledge that the data may reflect historical inequalities or biases in data collection methods. Explain how you will try to mitigate these biases in your analysis.
2. **Representation**: Ensure that the dataset includes a diverse representation of countries from different regions and income levels to avoid reinforcing existing inequalities.

**Challenges/Issues:**

1. **Data Cleaning and feature:** As there are more around 167 data records based on country the clustering algorithm may not work as expected. It may need more data.
2. **Model Selection:** testing and finding the accuracy of the model on the low volume of data may be challenging.

**References:**

1. (Kokkula, n.d.)Unsupervised Learning on Country Data