1. **PROJECT OVERVIEW**

**Project Name:-** Government Revenue(Percentage of GDP) by Global Economics World Bank(Kaggle Dataset)

**Context:-**

In an increasingly competitive Global-Economical landscape, understanding Geographical implications on Economical behaviour is essential for improving retention and increasing GDP. The project aims to enable data-driven Global-Economical analysis through the use of machine learning.

**Problem Statement:-**

Current Global-Economical GDP behaviour results in low engagement and poor performance in GDP for many countries. There is a need to improvise few countries effectively to enable Government Revenue(Percentage of GDP) significant.

**Proposed Solution:-**

Apply supervised machine learning techniques to identify countries Government Revenue(Percentage of GDP) based on their behaviour and demographics.

1. **PROJECT GOALS**

**Purpose:-**

To enhance percentage of GDP effectively through intelligent segmentation

**Project Focus:-**

Supervised learning to discover grouping of countries for more effective targeting to improvise Government Revenue(Percentage of GDP) with world bank data.

**Specific Goals:-**

* Detail level exploratory analysis on world bank data
* Segment world bank data using different clustering models
* Understand behavioural pattern of each segment
* Develop actionable strategies for each segment

**Expected Outcome:-**

* + - Improved percentage of GDP
    - Reduced Debt, Unemployment Rate
    - Higher GDP growth Annually

1. **PROJECT DESCRIPTION**
   1. **PROJECT OBJECTIVE AND SCOPE**

**Objective:-**

To develop a clustering model that identifies natural grouping withing countries based for actionable Global\_Economical Insights

**Scope:-**

Includes data collection, analysis, model development, evaluation and dashboard integration. Excludes external vendor integration or real time deployment

* 1. **DATA DESCRIPTION:-**

**Data Sources:-**

* Kaggle Dataset Global Economical World Bank Data(Country profile and demographics)
* Behavioral analytics

**Key Features:-**

* GDP related Attributes except “GDP Revenue”
* Interest Rate, Inflation, Unemployment Rate

**3.3 Exploratory Data Analysis:-**

EDA is critical step in any machine learning project. It involves summarizing the main characteristics of a dataset often visually to uncover patterns, spot anomalies and form hypothesis before model building.

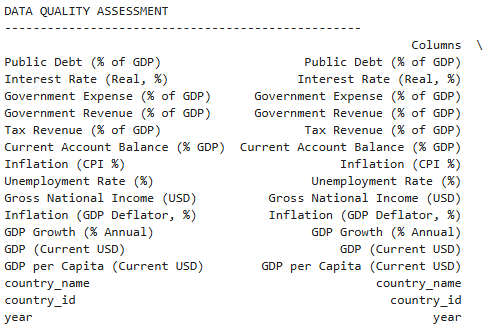
* + 1. **Understand the dataset:-**

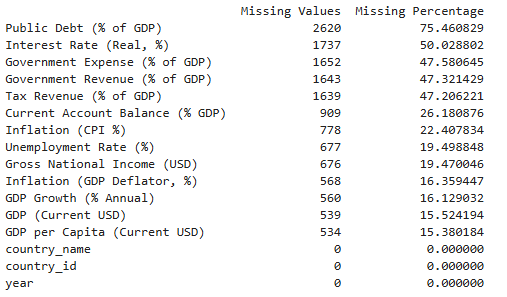
1. Load dataset using pandas, numpy or similar tools
2. Check shape:- number of rows and columns
3. Use .info() and .describe() to review datatypes, missing values and basic stats.
   * 1. **Check for Missing or NULL Values**
4. Visualize with seaborn.heatmap() or check with .isnull().sum()

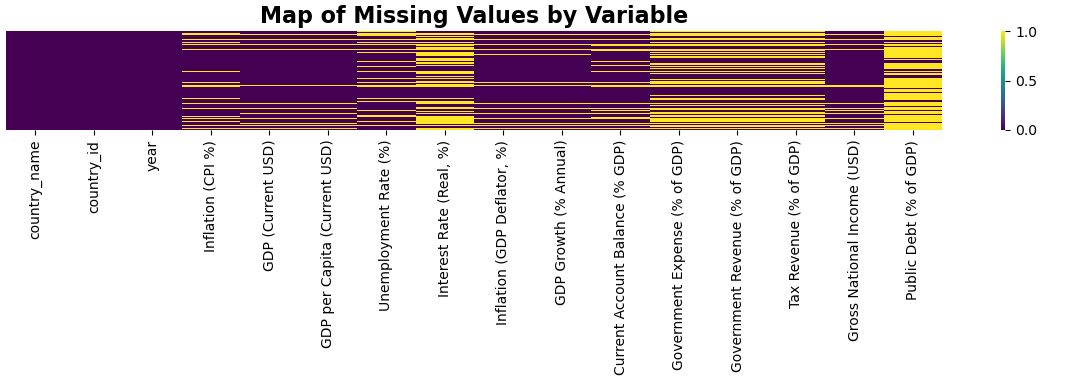
**Result**:- Dataset does not contain any Null value

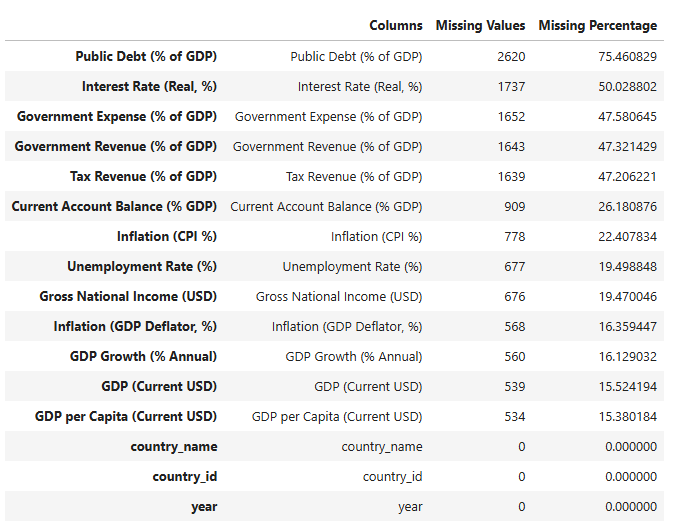
**To be more specific with detailed analysis:-**

**Data Quality Assessment:-**

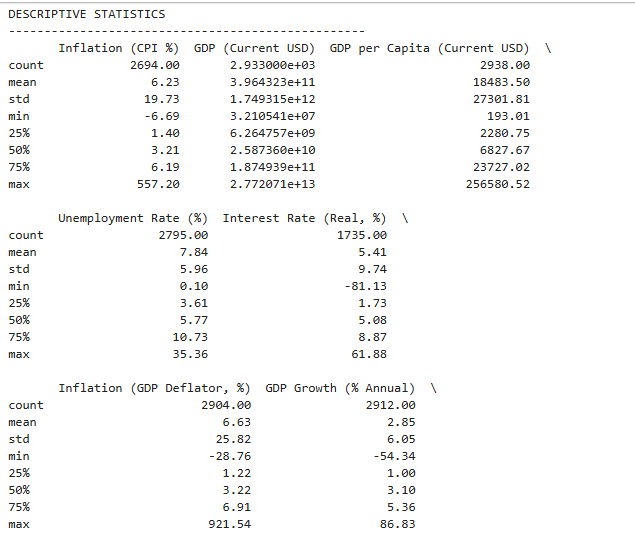


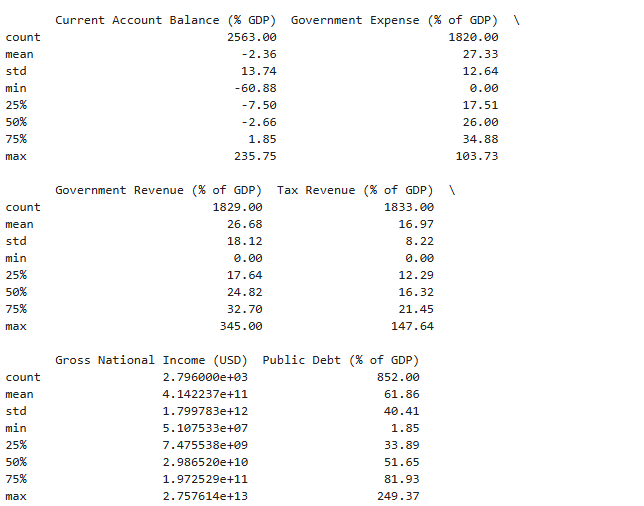


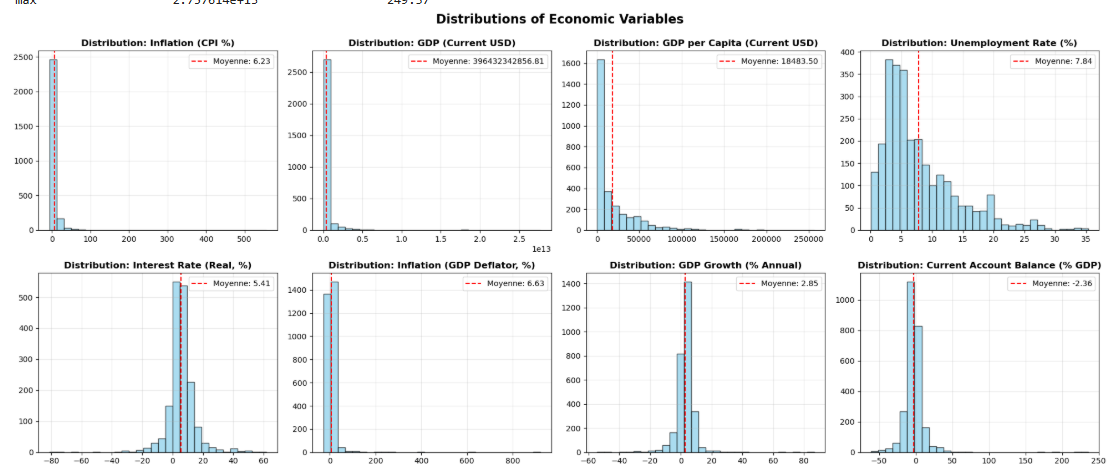


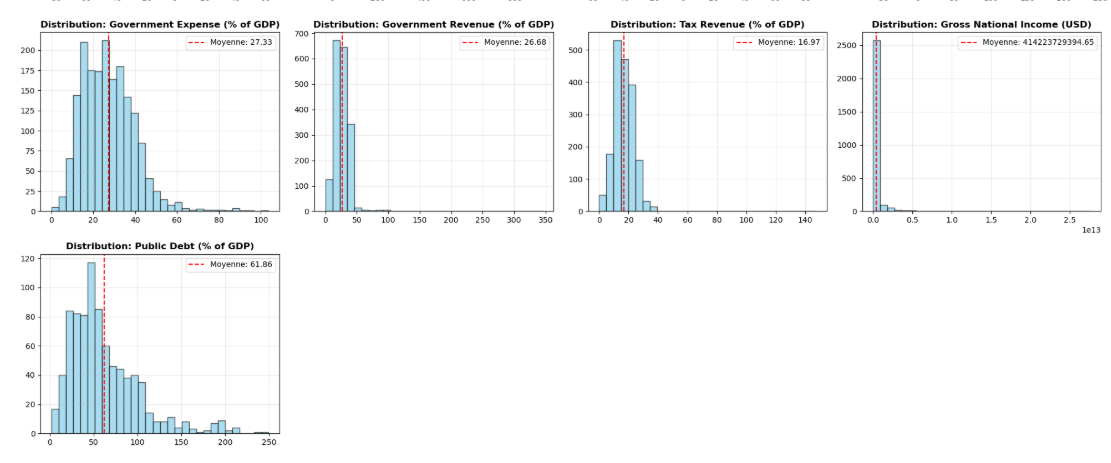


**In-Depth Descriptive Analysis:-**

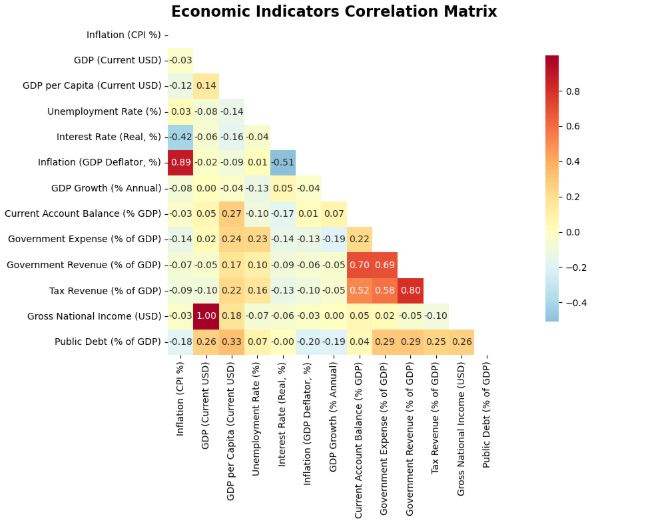
****

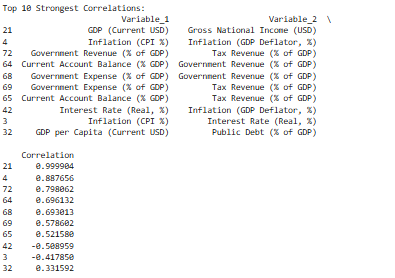
****

****

****

**CORRELATION ANALYSIS:-**

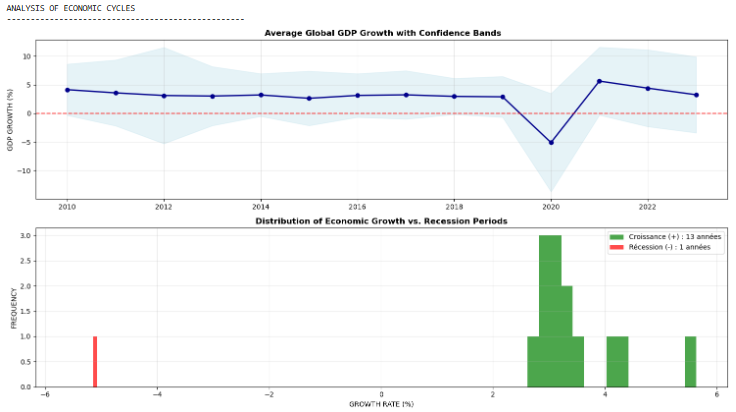
****

****

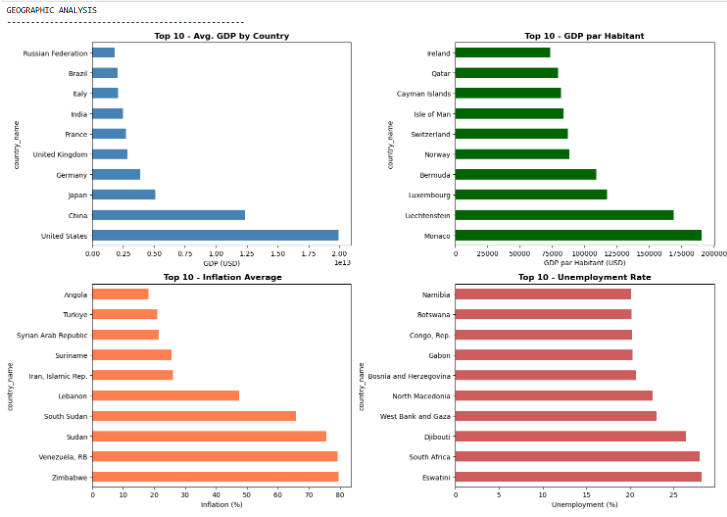
* + 1. **TEMPORAL ANALYSIS**



**Economic Cycle Analysis:-**



**GEOGRAPHIC ANALYSIS:-**

****

**Key Insights from Analysis listed above**

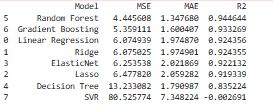
* **GDP and related metrics** (GDP per capita, GNI) have strong correlations with each other, suggesting they are interconnected in their economic performance.
* **Inflation and its different forms** (CPI, GDP deflator) show strong positive correlations, particularly between CPI and GDP deflator.
* **Unemployment** tends to be inversely related to inflation, which is a classic observation from economic theory.
* **Government revenue and expenses** are closely tied, and revenue is largely driven by taxes.
* **Public debt** seems to have a modest relationship with GDP-related metrics, indicating that debt levels increase in economies with larger or growing GDPs.

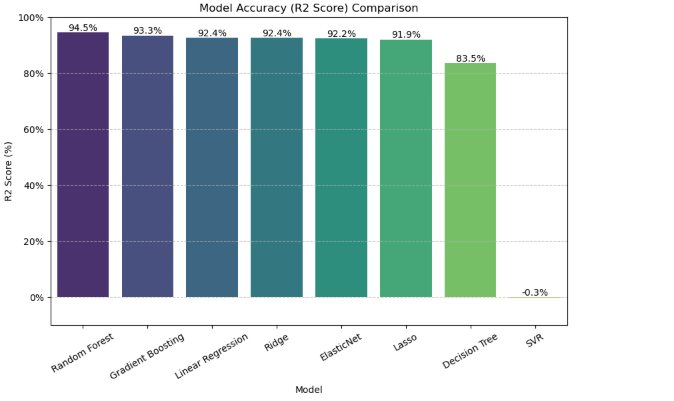
**3.4 DATA PREPARATION AND CLEANING:-**

In order to proceed for data preparation and cleaning few steps to be followed mentioned in DUTTA Proposal.pdf

**3.5 MODEL TRAINING & EVALUATION:-**

For Model Training & Evaluation, we need to walkthrough different kind of model analysis and the output is as follows:-





**3.6 USER INTERFACE INTEGRATION:-**

* Deploy the Trained Model: Need to make the model accessible for web application. We can use frameworks and services designed for model deployment, such as:
  + - FastAPI or Flask (Python): Lightweight frameworks for building APIs that serve ML models.
    - TensorFlow Serving: A dedicated system for deploying TensorFlow models in production.
    - Cloud-based ML services: Google Cloud AI, AWS SageMaker, or Azure ML provide tools for easy integration.
* Build the User Interface: Need to create a front-end interface using frameworks like React, Vue, or Angular, allowing users to interact with the ML model.
* Connect the UI to the Backend: Need to integrate the UI with the deployed model through APIs or other communication mechanisms.

1. **CAPSTONE COMPLEXITY**

* Medium to High Complexity due to integration of Supervised ML, data engineering and dashboard development
* Requires proficiency in Python, data visualization and analytical insights

1. **SOFTWARE AND TOOLS**

* Python (pandas, scikit-learn, matplotlib, seaborn)
* Tableau (dashboard visualization)
* Google Drive (optional for cloud compute and storage)
* GitHub (Version Control)

1. **PROJECT COMPLETION PLAN**

**Week1:** Data Collection and cleaning

**Week2-3:** Exploratory Data Analysisand Feature Engineering

**Week4-5:** Model Training and Evaluation

**Week6:** Integration into Tableau Dashboard

**Week7:** Final Reporting and Presentation

1. **PRESENTATION PLAN**

* Introduction and Problem Statement
* Data and Methodology
* Insights from Clustering
* Actionable Recommendations
* Live Demo of Dashboard

1. **RESOURCES**

* **Team:** DataScientist
* **Hardware:** Personal computer or cloud VM
* **Software:** Python, Tableau, Microsoft Office
* **Budget:** Approx. $100 for cloud compute and visualization licenses if required