# Exploratory Data Analysis on UN Sustainable Development Goals (SDG) Dataset

**Name:** Sukalp Warhekar
**PRN:** 22070521118
**Course:** Machine Learning

## 1. Introduction

The United Nations Sustainable Development Goals (SDG) dataset provides an extensive collection of global indicators tracking progress toward the 17 SDGs. This dataset includes multiple country-level and time-series observations across various developmental dimensions, such as poverty reduction, health, education, environment, and economic growth.

The primary objective of this study is to **perform Exploratory Data Analysis (EDA)** to understand the dataset's structure, identify patterns and trends, and prepare it for **future machine learning applications**, such as predictive modeling and clustering of countries based on SDG performance.

## 2. Dataset Overview

- **Source:** United Nations SDG Official Data Portal
- **Number of columns/features:** *37*
- **Key Features:**
  - GeoAreaName – Country/Region name
  - TimePeriod – Year of the observation
  - Indicator – SDG indicator description
  - SeriesCode – Unique code for each indicator series
  - OBS_VALUE – Observed value for the indicator

A preliminary review revealed:

- Missing values in some indicators
- Varying scales across different SDGs
- A combination of categorical (country, indicator) and numerical (value) data

## 3. ETL (Extract, Transform, Load) & Data Cleaning

The following preprocessing and cleaning steps were performed:

1. **Extract:** Loaded the dataset into a Jupyter notebook using Pandas.

2. **Transform:**
    - Removed duplicate rows
    - Dropped irrelevant columns that did not add analytical value
    - Converted TimePeriod to integer for easier time-series analysis
    - Renamed columns for readability
3. **Handle Missing Values:**
    - Checked the proportion of null values
    - Dropped rows with excessive missing data
4. **Load:** Created a cleaned CSV ready for EDA and ML tasks.

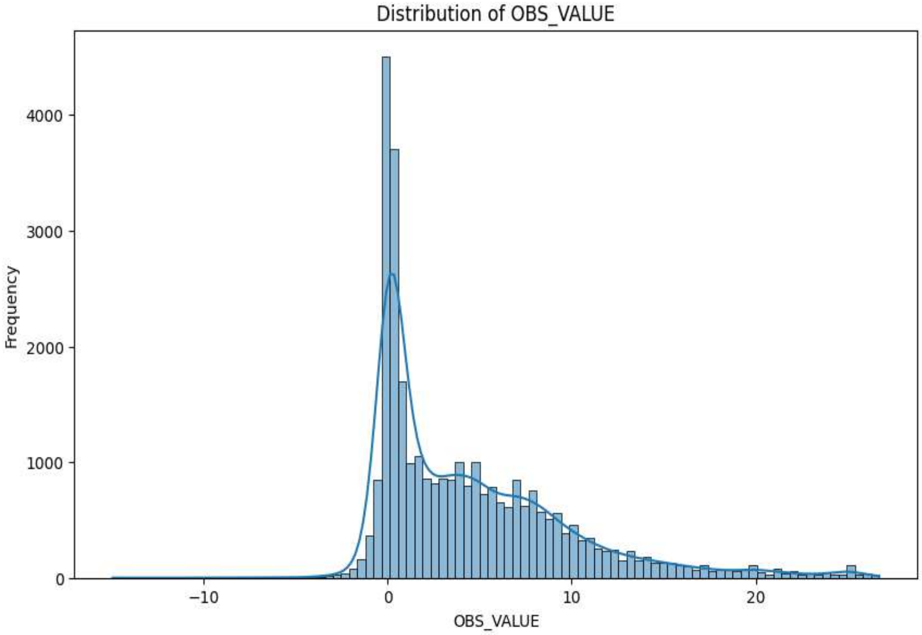
**Observation:** After cleaning, the dataset became well-structured, with consistent formats across time periods and indicators.

## 4. Exploratory Data Analysis (EDA)

We performed EDA to identify **trends, distributions, and relationships** within the dataset.

### 4.1 Univariate Analysis

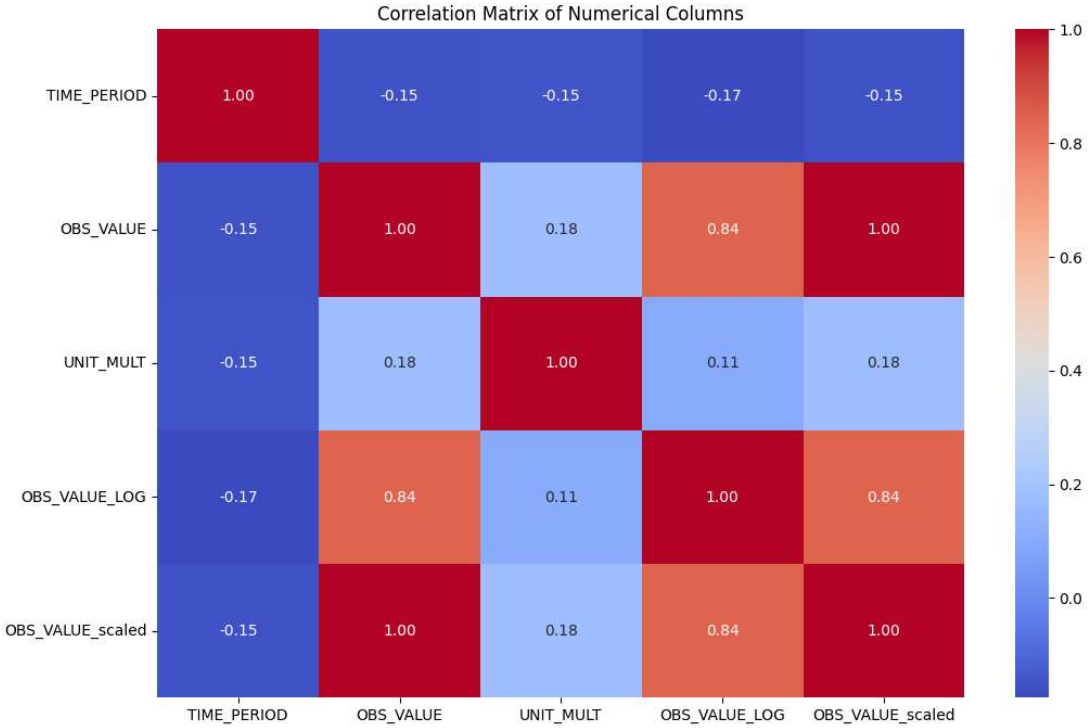
- Focused on the OBS_VALUE distribution to check the spread of indicator values.



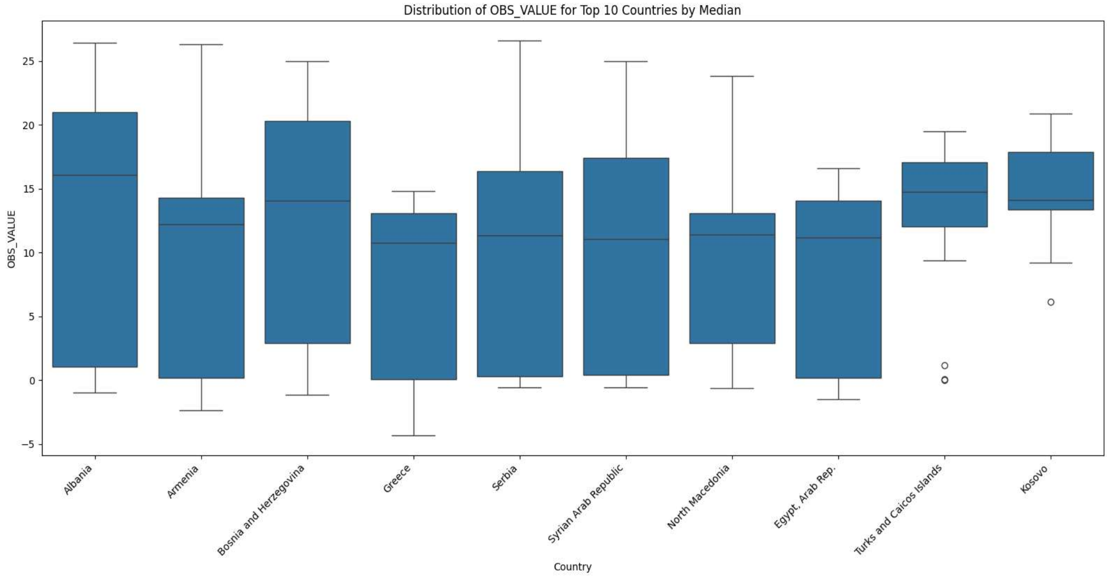
**Insight:**
The histogram reveals that most SDG indicator values are concentrated within a certain range, with a few extreme outliers. This reflects that some countries or indicators have exceptionally high measurements compared to the global average.

### 4.2 <u>Multivariate Analysis</u>

We explored relationships between indicators using correlation and compared distributions among countries.

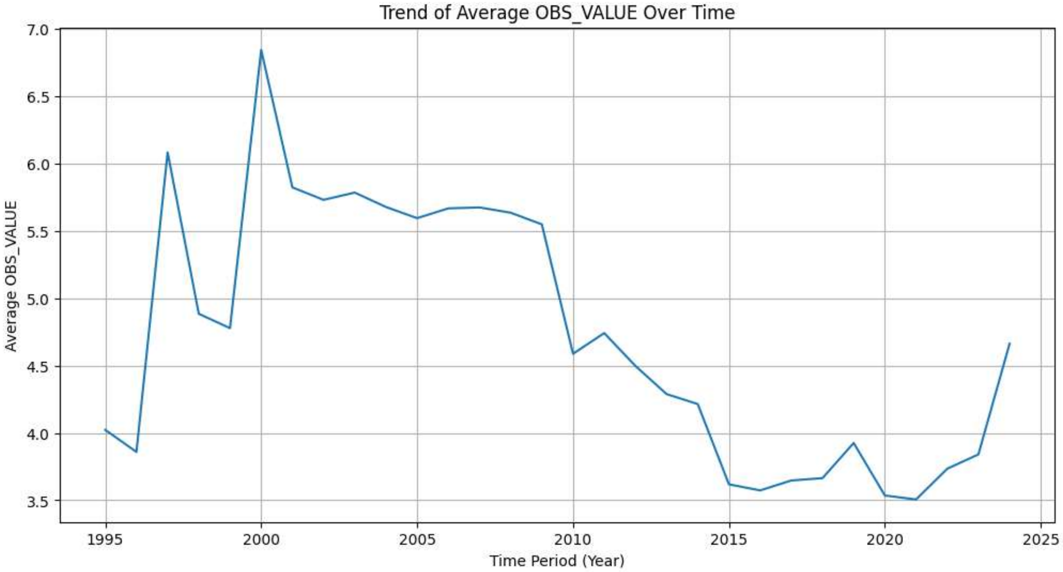


**Insight:**
The correlation heatmap highlights clusters of related SDG indicators. Strong correlations suggest interdependence between indicators, e.g., **economic growth metrics often correlate with energy and infrastructure indicators**.



**Insight:**
The boxplot shows variability in SDG indicator values among leading countries. Countries with larger boxes exhibit greater fluctuations in indicator performance, signaling potential disparities.

### 4.3 Time Series / Trend Analysis



**Insight:**
The line plot indicates **overall trends in global SDG performance over time**. Certain years show spikes or drops, possibly influenced by global events such as economic slowdowns or health crises.

### 4.4 Key Insights

- Indicator values are **unevenly distributed**, with a few extreme outliers.
- **Strong correlations** exist between some SDG indicators, which can guide feature selection for ML.
- **Temporal trends** reveal consistent improvement in some indicators, while others remain stagnant.
- Country-level variability suggests potential for **clustering and benchmarking** in ML.

## 5. Planned Machine Learning Algorithms

Based on the EDA findings, the following ML techniques are planned:

1. **Regression:**
   - Predict specific SDG indicator values based on other related indicators.
2. **Classification:**
   - Categorize countries into **High / Medium / Low performance tiers**.
3. **Clustering:**
   - Group countries with **similar SDG profiles** to identify development patterns.
4. **Time Series Forecasting:**
   - Predict **future SDG performance trends** using historical data.

## 6. Recommendations for ML Modeling

- **Feature Engineering:** Focus on indicators with strong correlations and consistent time coverage.
- **Normalization:** Standardize features to handle varying scales across SDG indicators.
- **Handling Missing Data:** Use imputation methods for minor gaps instead of dropping valuable observations.
- **Dimensionality Reduction:** Apply PCA if needed to simplify high-dimensional indicator space.

## 7. Conclusion

The EDA of the UN SDG dataset provided **meaningful insights into global development trends**. We identified outliers, correlations, and variability across countries and time. This dataset is now **well-prepared for machine learning tasks**, including regression, clustering, and forecasting to aid in sustainable development planning.