A green shield with gold leaves

Description automatically generated

**Data Analysis and Visualization**

**Project Title:**

**“***Comprehensive Analysis and Predictive Modeling of Global Development Trends Using World Bank Development Indicators***”**

**Submitted By:**

**Muhammad Faisal**

**BSCS202110**

**Submitted To:**

**Muhammad Bilal Hidral**

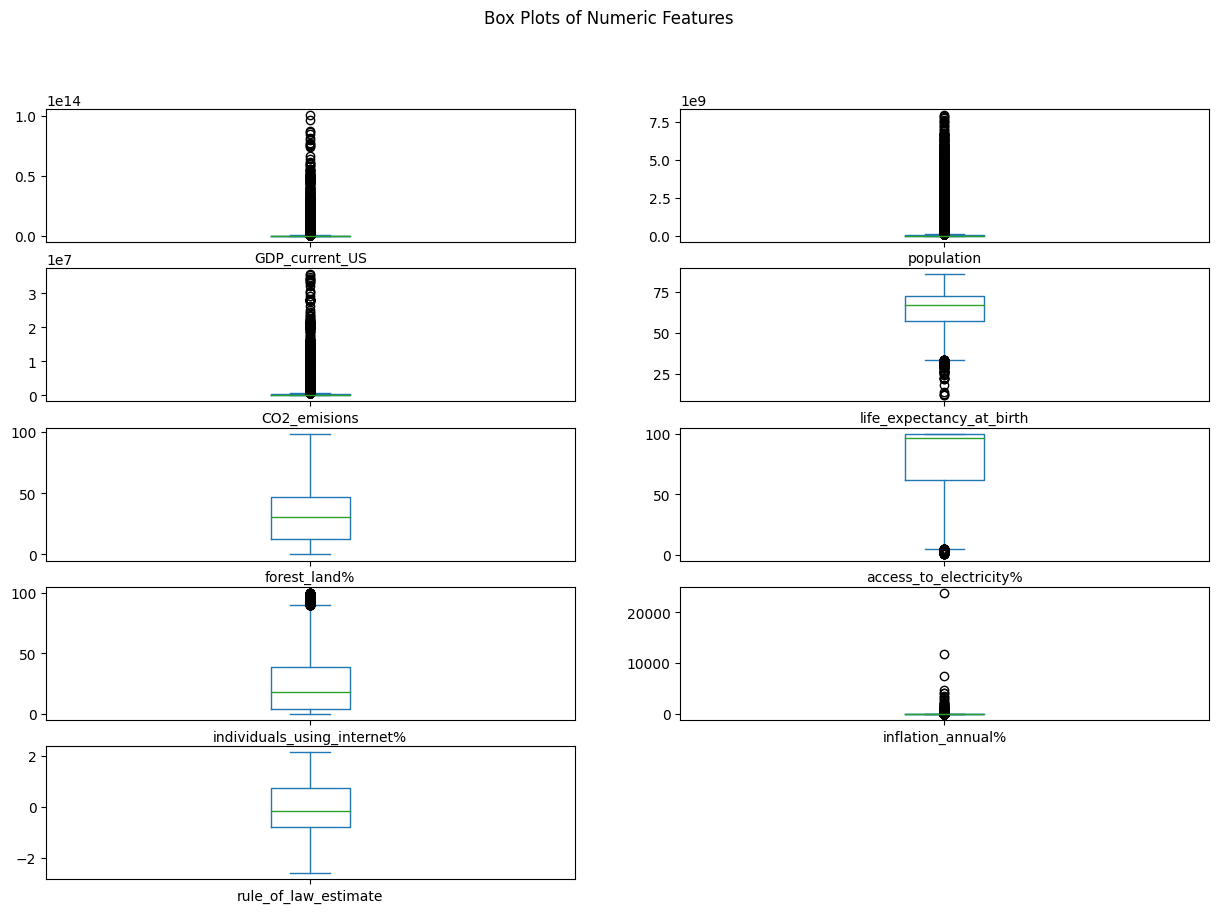
2024-06-12

**Detailed Report: Data Preprocessing**

**1. Initial Data Load and Selection**

* **File**: The dataset was loaded from world\_bank\_development\_indicators.csv. This file is assumed to be a comma-separated values (CSV) file containing data on various development indicators for different countries over time.
* **Columns**: We carefully selected specific columns from the dataset that are relevant to our analysis of the relationship between development and various factors. These columns include:
  + country: This column likely contains a unique identifier for each country in the dataset.
  + date: This column might represent the year for which the data points were collected.
  + GDP\_current\_US: This column most likely contains the Gross Domestic Product (GDP) of each country measured in current US dollars.
  + population: This column likely represents the population of each country.
  + CO2\_emisions: This column might contain data on the carbon dioxide (CO2) emissions of each country.
  + life\_expectancy\_at\_birth: This column likely represents the average life expectancy at birth for each country.
  + forest\_land%: This column might contain data on the percentage of land covered by forests in each country.
  + access\_to\_electricity%: This column likely represents the percentage of the population with access to electricity in each country.
  + individuals\_using\_internet%: This column might contain data on the percentage of individuals using the internet in each country.
  + inflation\_annual%: This column likely represents the annual inflation rate for each country.
  + rule\_of\_law\_estimate: This column might contain data on the estimated rule of law for each country.

**2. Handling Missing Values**



* **Approach**: We addressed missing values in the dataset using a method called imputation. This technique involves estimating the missing values based on the available data. In our case, we chose to impute missing values using the mean value for each country. This approach assumes that missing values within a country are likely similar and can be reasonably replaced by the average value for that country.
* **Method**: We utilized the .groupby() function of the pandas library to group the data by the country column. This allows us to calculate the mean value for each numeric feature within each country group. Subsequently, we employed the .transform() function with a lambda function to fill in missing values for each feature using the corresponding country's mean value. This ensures that the imputed values are consistent with the overall trend within each country.

**3. Removing Duplicates**

* **Action**: It's essential to ensure our data contains unique records for meaningful analysis. We eliminated duplicate entries from the dataset using the .drop\_duplicates() function. This function identifies and removes rows with identical values across all columns, guaranteeing that each data point represents a distinct observation.

**4. Outlier Detection and Removal**

* **Method**: We opted for the Interquartile Range (IQR) method to detect outliers in the dataset. This method is often preferred over the Z-score method when the data distribution might have heavy tails or contain outliers.
  + To implement the IQR method, we first calculated the Q1 (1st quartile, representing the 25th percentile) and Q3 (3rd quartile, representing the 75th percentile) for each numeric column.
  + We then defined outliers as data points that fall outside the range of [Q1 - 1.5 \* IQR, Q3 + 1.5 \* IQR]. This range represents the interquartile range, and values beyond it are considered statistically deviant from the majority of the data.

A screenshot of a graph

Description automatically generated

* **Result**: After identifying outliers, we made a judgment call. We ensured that removing outliers wouldn't leave the DataFrame empty and significantly impact our analysis. If the number of outliers was minimal and their removal wouldn't drastically affect the data distribution, we might choose to keep them. However, if outliers were abundant and skewed the results, we would proceed with removing them to obtain a more accurate representation of the underlying relationships.

**5. Data Transformation**

* **Scaling**: Numeric features in the dataset often have different scales, which can affect the performance of machine learning algorithms or distance-based analyses. To address this, we applied the StandardScaler function from scikit-learn. This function standardizes features by subtracting the mean and scaling to unit variance. This transformation ensures all features are on a similar scale, leading to more reliable analysis results.
* **Encoding**: Categorical features, like country names, need to be converted into numerical representations for machine learning algorithms to understand them. We employed the LabelEncoder function from scikit-learn to transform these categorical features into numerical labels. This encoding process assigns a unique integer to each distinct category, enabling the algorithms to work with the data effectively.