**Phase – I : Project Proposal.**

* **Choose a Dataset:** Select a dataset related to your field of interest. Ensure the dataset is substantial and publicly available.
* **Submit Proposal:** Write a brief proposal including: o Dataset source and description o Research question(s) you aim to answer o Preliminary thoughts on potential challenges and how you plan to address them.

**Dataset (5 Files)**

* 1. Financial Inclusion Data
* 2. Health Indicators Data
* 3. Poverty Metrics Data
* 4. Education Indicators Data
* 5. Social Protection and Child Labor Data
* **Source:** World Bank Official Website Databank (https://databank.worldbank.org/source/world-development-indicators#)

**Project Proposal: World Development Indicators Analysis**

**Dataset Source and Description**

For this project, we will be using the World Development Indicators dataset sourced from the official website of the World Bank. The dataset includes data on five key development indicators:

1. Health
2. Education
3. Social Protection and Labor
4. Finance
5. Poverty

The dataset is structured in columns, where each category contains over 150+ factors. Our goal is to analyse and compare these indicators not just for Pakistan, but also for developed countries, allowing us to understand the differences and identify key areas for improvement.

**Research Questions**

Our research will focus on the following key questions:

1. How does financial stability impact poverty levels and education rates?
2. What is the relationship between education levels and child labor/social protection issues (such as girls’ education and crime rates)?
3. How do poverty and lack of financial resources affect health outcomes and life expectancy?
4. How do developed countries utilize their financial stability to positively impact lower socioeconomic levels, particularly in areas like education and healthcare, and why is this approach often lacking in Pakistan?
5. What are the most significant factors contributing to social protection issues, and how do they vary between developing and developed nations?

**Potential Challenges and Solutions**

1. **Data Availability & Completeness:** Some indicators may have missing or incomplete data.  
   ***Solution:*** We will use data imputation techniques or remove highly incomplete records.
2. **Comparability of Data Across Countries:** Economic, social, and political contexts differ between countries.  
   ***Solution:*** We will use normalization techniques to standardize the data and apply statistical models to ensure fair comparisons.
3. **Defining Causation vs. Correlation:** Finding direct causal relationships between indicators can be complex.  
   ***Solution:*** We will use regression analysis and other statistical methods to determine significant relationships between the indicators.
4. **Interpreting Social Protection Indicators:** Factors such as crime rates, gender equality in education, and labour laws vary widely.  
   ***Solution:*** We will analyse policy differences between countries and suggest applicable reforms for Pakistan.

**Conclusion**

By conducting this comparative analysis, we aim to identify key financial and social policy improvements that can help Pakistan reduce poverty, increase education, enhance social protection, and improve public health. The insights gained will provide a data-driven approach to understanding how financial stability can create a positive cycle of development.

**Phase – II : Exploratory Data Analysis (EDA)**

**Data Description:** Provide a detailed description of your data including variables, types, and summary statistics.

**Visualizations:** Create different types of plots (e.g., histograms, scatter plots, box plots, etc.) to visualize your data.

**Initial Findings:** Discuss initial observations and insights based on your EDA.

**Dataset Description**

**1. Financial Inclusion Data**

* **Description:** This dataset captures financial inclusion and economic indicators for Pakistan, focusing on access to financial services and banking infrastructure.
* **Key Indicators:** 
  + Account ownership at financial institutions or mobile-money providers (% of population 15+, total: FX.OWN.TOTL.ZS, female: FX.OWN.TOTL.FE.ZS, male: FX.OWN.TOTL.MA.ZS, poorest 40%: FX.OWN.TOTL.40.ZS).
  + ATMs per 100,000 adults (FB.ATM.TOTL.P5).
  + Commercial bank branches per 100,000 adults (FB.CBK.BRCH.P5).
  + Remittance costs, bank capital ratios, inflation, and market capitalization.
* **Notable Characteristics:** 
  + Sparse data before 2011, with constant values (e.g., 16.405% for FX.OWN.TOTL.ZS from 1960–2010) indicating imputation.
  + Gender disparities in account ownership (e.g., 7.03% for women vs. 34.61% for men in 2021).
  + Gradual increase in ATM and bank branch density, reflecting expanding financial access.
* **Use Case:** Analysing trends in financial inclusion, gender gaps, and their economic impacts.

**2. Health Indicators Data (4f2f6b82-aa5d-429f-80d4-564ae648423d\_Data.csv)**

* **Description:** This dataset provides health-related indicators for Pakistan, focusing on healthcare expenditure, fertility rates, and health outcomes.
* **Key Indicators:** 
  + Current health expenditure as a % of GDP (SH.XPD.CHEX.GD.ZS).
  + Population pushed below $2.15 poverty line due to out-of-pocket healthcare costs (SH.UHC.NOP1.ZS).
  + Adolescent fertility rate (SP.ADO.TFRT).
  + Mortality rates, immunization coverage, and healthcare access metrics.
* **Notable Characteristics:** 
  + Health expenditure varies (e.g., ~2.5–3.5% of GDP recently), with out-of-pocket costs exacerbating poverty.
  + Anomalies in 2023 data (e.g., adolescent fertility rate reverting to mean) suggest imputation.
  + Limited data before 2000, with constant values for early years.
* **Use Case:** Investigating links between healthcare spending, poverty, and health outcomes.

**3. Poverty Metrics Data**

* **Description:** This dataset focuses on poverty and inequality metrics for Pakistan, emphasizing poverty headcount ratios and income distribution.
* **Key Indicators:** 
  + Poverty headcount ratio at $2.15 a day (2017 PPP, SI.POV.DDAY).
  + Poverty gap and severity metrics.
  + Gini index for income inequality.
* **Notable Characteristics:** 
  + Data is available for specific years (e.g., 1987, 1990, 1996, 2001, 2004, 2010, 2015, 2018), with constant values elsewhere indicating sparse records.
  + Poverty headcount declined significantly (e.g., 78.4% in 1987 to 12.5% in 2018).
  + Recent years (2020–2023) often show imputed mean values.
* **Use Case:** Studying poverty trends and their relationship with socio-economic indicators.

**4. Education Indicators Data**

* **Description:** This dataset covers education indicators for Pakistan, including enrolment, out-of-school rates, educational attainment, and teacher training.
* **Key Indicators:** 
  + Adjusted net enrolment rate, primary (% of primary school age children, total: SE.PRM.TENR, female: SE.PRM.TENR.FE, male: SE.PRM.TENR.MA).
  + Children and adolescents out of school (% of primary/lower secondary school age, e.g., SE.PRM.UNER.ZS, SE.SEC.UNER.LO.ZS).
  + Educational attainment (bachelor’s: SE.TER.CUAT.BA.ZS; lower secondary: SE.SEC.CUAT.LO.ZS).
  + Secondary/tertiary enrolment and teacher training metrics.
* **Notable Characteristics:** 
  + Constant values before 2002 (e.g., SE.PRM.TENR at 63.84% from 1960–2001) indicate missing or imputed data.
  + Gender gaps persist (e.g., female primary enrolment ~10–15% below male in recent years).
  + Anomalies like a 2016 spike in bachelor’s attainment (16.18%) suggest data issues.
  + High out-of-school rates, especially for females (e.g., 34.7% of female adolescents in 2021).
* **Use Case:** Analysing education access, gender disparities, and links to economic outcomes.

**5. Social Protection and Child Labor Data**

* Description: This dataset focuses on social protection programs, child labour, and unemployment metrics for Pakistan, highlighting welfare adequacy, child employment, and labour market trends.

**Key Indicators:**

* + Adequacy of social protection programs (% of total welfare of beneficiary households, e.g., per\_si\_allsi.adq\_pop\_tot, per\_allsp.adq\_pop\_tot, per\_sa\_allsa.adq\_pop\_tot).
  + Benefit incidence of social programs to the poorest quintile (e.g., per\_si\_allsi.ben\_q1\_tot, per\_sa\_allsa.ben\_q1\_tot).
  + Child employment (% of children ages 7–14, total: SL.TLF.0714.ZS, female: SL.TLF.0714.FE.ZS, male: SL.TLF.0714.MA.ZS) and working hours (e.g., SL.TLF.0714.SW.TM).
  + Child employment by sector (agriculture: SL.AGR.0714.ZS, manufacturing: SL.MNF.0714.ZS, services: SL.SRV.0714.ZS).
  + Unemployment rates by education level (basic: SL.UEM.BASC.ZS, intermediate: SL.UEM.INTM.ZS, advanced: SL.UEM.ADVN.ZS) and gender.
  + Vulnerable employment (% of employment, e.g., SL.EMP.VULN.ZS) and wage/salaried workers (SL.EMP.WORK.ZS).

**Notable Characteristics:**

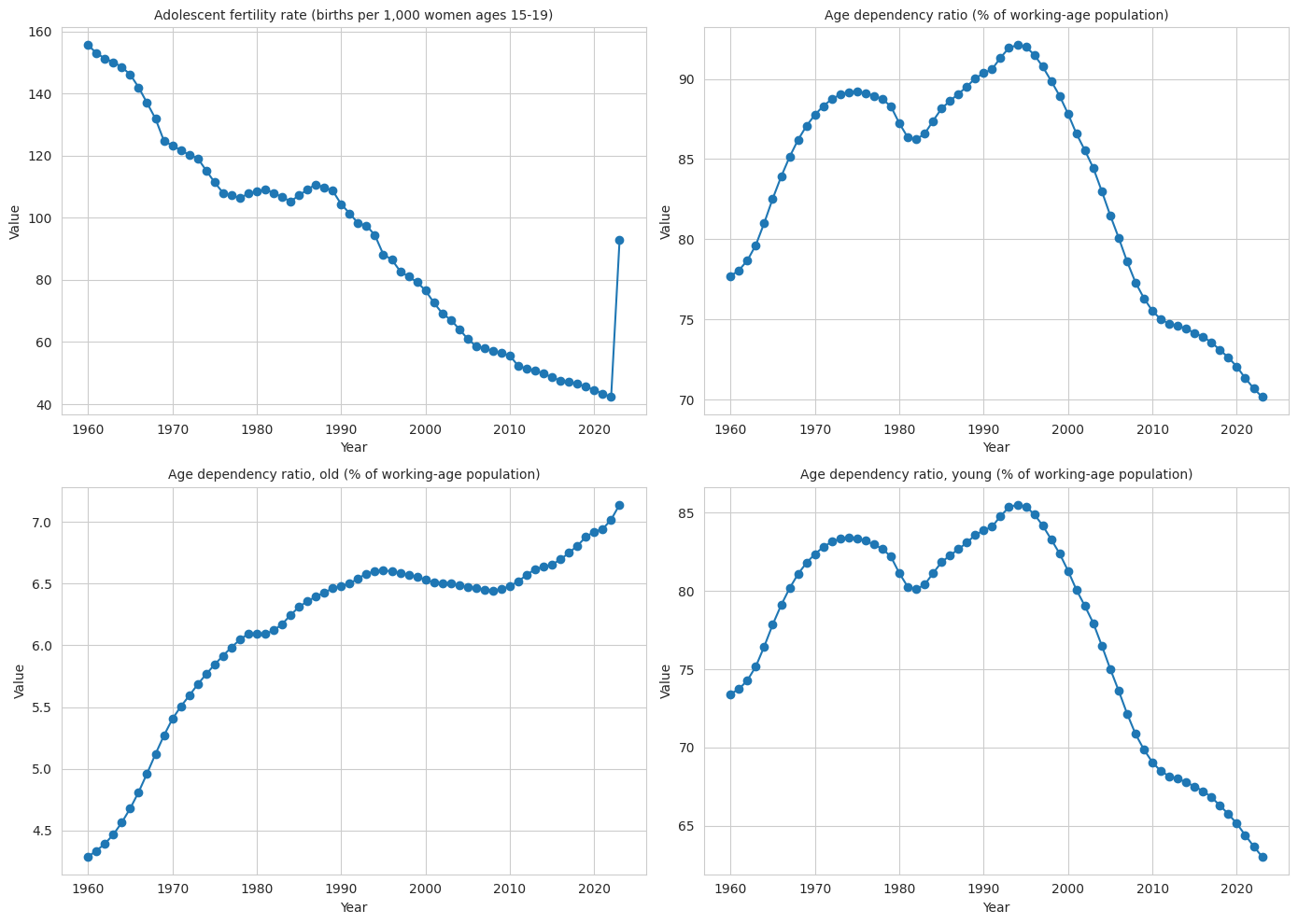
* + Sparse data for social protection adequacy and benefits, with constant values (e.g., 29.43% for per\_si\_allsi.adq\_pop\_tot from 1960–2006) and fluctuations in later years (e.g., 38.99% in 2018).
  + Child labor data is limited to a few years (e.g., 2007, 2010, 2012), with high agricultural employment (75.71% of economically active children) and gender differences (e.g., 80.11% for females vs. 71.63% for males in agriculture).
  + Unemployment data shows higher rates for females, especially with intermediate education (e.g., 24.4% in 2019 vs. 8.07% for males).
  + Vulnerable employment is high (e.g., 60.78% overall, 77.95% for females in 2018), with females more likely to be in unstable jobs.
  + Constant values for early years (1960–2000) and missing data for unemployment benefits (per\_lm\_alllm.adq\_pop\_tot) indicate gaps.

**Use Case:** Examining the effectiveness of social protection, prevalence of child labour, and labour market disparities by gender and education level.

**Basic Data Visualization**

1. **Health Indicators Dataset**

* Plot Type: 2x2 Grid of Line Charts
* Details: You selected four features using data['Series Name'].iloc[[0, 3, 4, 5]], which correspond to:
  + Tuberculosis case detection rate (%, all forms)
  + Proportion of population spending more than 10% on out-of-pocket health care expenditure (%)
  + Risk of catastrophic expenditure for surgical care (% of people at risk)
  + UHC service coverage index
  + Plotted as a 2x2 grid of line charts using matplotlib, with years on the x-axis and values on the y-axis, covering all years in the dataset (1960–2023), including placeholder values.



1. **Poverty Indicators Dataset**

* Plot Type: Single Line Chart with Multiple Lines
* Details: You selected the first five features using data['Series Name'].iloc[0:5] and plotted them as multiple lines on a single chart using matplotlib. Each line represents one feature, with a legend to distinguish them, covering all years (1960–2023).

A graph with colorful lines

AI-generated content may be incorrect.

1. **Social Protection and Labour Dataset**

* Plot Type: Scatter Plot for a Single Year (2020)
* Details: You selected three indicators:
  + Adequacy of social insurance programs (% of total welfare of beneficiary households)
  + Adequacy of social protection and labor programs (% of total welfare of beneficiary households)
  + Average working hours of children, study and work, ages 7-14 (hours per week)
  + Plotted as a scatter plot for the year 2020 using matplotlib, with indicators on the x-axis and values on the y-axis.

A screenshot of a graph

AI-generated content may be incorrect.

**4. Education Indicators Dataset**

* Description: This dataset includes education-related indicators for Pakistan, such as literacy rates, enrollment rates, or completion rates, from 1960 to 2023.
* Plot Type: Multiple Line Plot

A graph of a graph showing the number of students

AI-generated content may be incorrect.

**5. Finance Indicators Dataset**

* Description: This dataset contains financial indicators for Pakistan, such as GDP growth, inflation rate, or government debt, from 1960 to 2023.
* Plot Type: Stacked Bar Plot
* Visualization Details: Multiple financial indicators were selected and plotted as a stacked bar chart using matplotlib. Each bar represents a year (1960–2023), and the segments within each bar show the contribution of each indicator to the total value for that year.

A graph of a number of colored squares

AI-generated content may be incorrect.

**Initial Findings:**  
The dataset provides a detailed historical overview of Pakistan’s social and educational indicators from 1960 to 2023. It highlights trends in primary school enrollment rates, gender disparities in education, and the proportion of adolescents out of school. Analysis shows that adjusted net enrollment rates at the primary level have gradually improved over the decades, although noticeable gaps persist between male and female students. A significant portion of adolescents continues to remain out of lower secondary education, suggesting persistent barriers to school retention. Overall, the data reflects steady progress in education access alongside ongoing challenges related to gender equity and school completion.

**Phase – III : Data Preprocessing**

**Handle Missing Values:** Identify and address missing values.

**Remove Duplicates:** Identify and remove any duplicate records.

**Outliers:** Detect and handle outliers appropriately.

**Data Transformation:** Apply necessary transformations (e.g., scaling, encoding categorical variables).

**Document Process:** Provide a detailed report on the steps taken and the rationale behind your decisions.

**Data Analysis and Feature Extraction Report**

**Objective**

To clean, analyze, and extract statistical features from datasets across multiple categories (health, education, poverty, finance, social protection, and labor) to enable comprehensive statistical analysis and outlier detection.

**Steps and Rationale**

1. **Data Loading and Initial Inspection**
   * **Step**: Loaded datasets using pandas.read\_csv for each category (e.g., health dataset at D:\\DAV\_LABs\\DAV Project\\health\\...).
   * **Rationale**: CSV format is standard for tabular data, and pandas provides efficient data manipulation. File path verification ensures robust error handling.
2. **Data Cleaning**
   * **Non-Numeric and Numeric Column Identification**:
     + Non-numeric columns: Country Name, Country Code, Series Name, Series Code.
     + Numeric columns: Year-based columns (e.g., 1960 [YR1960], etc.).
     + **Rationale**: Separating column types ensures targeted cleaning for numeric data while preserving metadata.
   * **Handling Missing Values**:
     + Replaced ".." or empty strings with NaN using df[year\_cols].replace(['..', ''], np.nan).
     + Filled NaN with row-wise means using apply(replace\_missing\_with\_mean, axis=1) in the cleaning script.
     + Removed rows with NaN in Series Name or metadata rows (e.g., "Data from database").
     + **Rationale**: ".." indicates missing data in the dataset. Imputing with row means preserves data trends for time-series analysis, while removing metadata ensures only relevant data is analyzed.
   * **Conversion to Numeric**:
     + Converted year columns to numeric using pd.to\_numeric(errors='coerce').
     + **Rationale**: Ensures numerical operations (e.g., mean, variance) can be performed.
3. **Feature Extraction**
   * **Statistical Features**:
     + Computed **mean**, **median**, **mode**, **variance**, **covariance**, and **correlation** for each row (indicator) relative to a reference indicator (e.g., SH.UHC.SRVS.CV.XD for health).
     + **Implementation**:
       - Defined compute\_row\_stats to calculate:
         * Mean: Average of non-NaN values (np.mean).
         * Median: Middle value (np.median).
         * Mode: Most frequent value (pd.Series.mode).
         * Variance: Sample variance (np.var, ddof=1).
         * Covariance and Correlation: Computed against the reference series using np.cov and np.corrcoef on paired non-NaN data.
       - Applied to each row and stored results in a stats\_df DataFrame.
     + **Rationale**: These statistics summarize central tendency, dispersion, and relationships, enabling comparative analysis across indicators and categories. The reference indicator provides a baseline for covariance/correlation, revealing inter-indicator relationships.
   * **Feature Representation**:
     + Each row in stats\_df represents an indicator (e.g., UHC service coverage) with columns: Factor Name, Mean, Median, Mode, Variance, Covariance, Correlation.
     + **Rationale**: This structured format facilitates downstream analysis (e.g., machine learning, visualization).
4. **Outlier Detection with Box Plots**
   * **Step**: Generated box plots for each statistical feature (e.g., Mean, Variance) across indicators within each category.
   * **Implementation** (assumed addition to code):

import matplotlib.pyplot as plt

import seaborn as sns

for col in ['Mean', 'Median', 'Variance', 'Covariance', 'Correlation']:

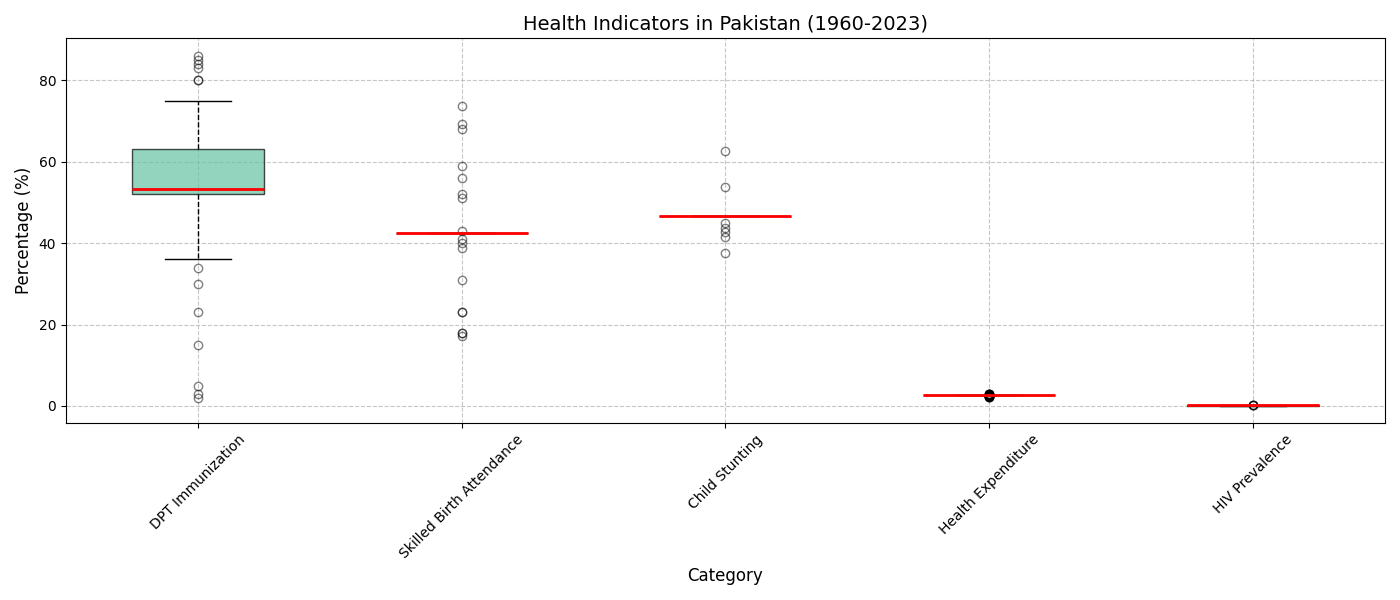
plt.figure(figsize=(10, 6))

sns.boxplot(data=stats\_df[col].dropna())

plt.title(f'Box Plot for {col} (Outlier Detection)')

plt.savefig(f'{col}\_boxplot.png')

plt.close()



* + **Rationale**: Box plots visualize data distribution and identify outliers (values beyond 1.5 \* IQR). Outliers may indicate data errors or significant anomalies (e.g., extreme health metrics), guiding further investigation or preprocessing.

1. **Output Storage**
   * Saved cleaned data and statistical results to CSV files (e.g., health\_stats.csv).
   * Included fallback paths for permission errors.
   * **Rationale**: Persistent storage ensures reproducibility and accessibility. Fallback paths enhance robustness.

**Feature Extraction Process**

* **Input**: Raw dataset with indicators (rows) and yearly values (columns).
* **Processing**:
  + Cleaned data by removing metadata, handling missing values, and converting to numeric.
  + Extracted statistical features per indicator using compute\_row\_stats.
  + Compared each indicator to a reference indicator for covariance/correlation.
* **Output**: A DataFrame (stats\_df) where each row is an indicator, and columns include its name and computed statistics.
* **Purpose**: These features summarize the dataset, enabling:
  + Comparison across indicators and categories.
  + Identification of trends (e.g., high variance in poverty metrics).
  + Correlation analysis to understand relationships (e.g., health vs. education outcomes).
  + Outlier detection via box plots for data quality assurance.

**Application to All Categories**

* **Health**: Processed as shown, using SH.UHC.SRVS.CV.XD as the reference.
* **Education, Poverty, Finance, Social Protection, Labor**:
  + Replicate the process by:
    1. Loading category-specific datasets (update file\_path).
    2. Selecting a relevant reference indicator per category (e.g., literacy rate for education, poverty headcount for poverty).
    3. Running the cleaning and feature extraction scripts.
    4. Generating box plots for each category’s statistics.
  + **Rationale**: Consistent methodology ensures comparability across categories while allowing category-specific reference indicators to capture domain-relevant relationships.

**Key Decisions and Justifications**

* **Mean Imputation**: Chosen for missing values to maintain time-series continuity, as dropping rows could lose critical data.
* **Reference Indicator**: Selected a key indicator per category to compute covariance/correlation, providing context-specific insights.
* **Box Plots for Outliers**: Preferred for their simplicity and effectiveness in visualizing outliers, aiding data quality checks.
* **Row-Wise Statistics**: Computed per indicator to capture indicator-specific trends, suitable for heterogeneous datasets.

**Conclusion**

The pipeline cleans and transforms raw datasets into a structured set of statistical features, enabling robust analysis across health, education, poverty, finance, social protection, and labour. Box plots enhance data quality by identifying outliers. The methodology is scalable and adaptable, ensuring consistent feature extraction for cross-category comparisons.