DATA SCIENCE TOOLBOX: PYTHON PROGRAMMINGPROJECT REPORT

(Project Semester January-April 2025)

Exploratory Analysis of Socio-Economic Indicators using World Bank Data

Submitted by

Akshat V Sahay

Registration No. 12304012

Programme and Section.

Course Code . CSE375

Under the Guidance of

Dr. Mrinalini Rana

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Akshat V Sahay bearing Registration no. 12304012 has completed CSE375 project titled, **"Exploratory Analysis of Socio-Economic Indicators using World Bank Data"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

School of Computer Science
Lovely Professional University
Phagwara, Punjab.

Date: 12-04-2025

DECLARATION

I, Akshat V Sahay, student of DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-0402025 Signature

Registration No. 12304012 Akshat V Sahay

Table of Contents

1. Introduction

In today's data-driven world, insights derived from comprehensive data analysis are crucial for informed decision-making. This project, undertaken as part of the Data Science minor curriculum, aims to explore global development indicators using a dataset sourced from the World Bank. The focus is on conducting a structured and in-depth Exploratory Data Analysis (EDA) to uncover patterns, relationships, and trends among key socio-economic variables.

By examining indicators such as population density, internet usage, GDP per capita, and life expectancy, the project highlights regional and income-based disparities across the globe. The goal is to derive meaningful insights from complex datasets, communicate them effectively through visualizations, and lay the groundwork for data-driven storytelling and policy suggestions.

2. Source of Dataset

The dataset used for this project has been sourced from the **World Bank Open Data Platform**, which provides free and open access to global development data. It includes a wide range of socio-economic indicators across various countries and regions over multiple years.

Dataset Details:

- **Source:** World Bank Open Data
- **File Type:** Excel (.xlsx)
- **Variables Included:**
    - Country Name and Code
    - Region and Income Group
    - Year
    - GDP and GDP per Capita (USD)
    - Birth and Death Rates
    - Life Expectancy Metrics
    - Internet Usage (%)
    - Population Density
    - Infant Mortality Rate

The dataset provides a rich foundation for comparative analysis and visual exploration of global development indicators.

3. EDA Process (Exploratory Data Analysis)

Exploratory Data Analysis (EDA) is a critical first step in the data science workflow. It involves summarizing the main characteristics of the dataset, identifying patterns, detecting anomalies, and forming hypotheses using statistical graphics and visualization tools.

Steps Followed in EDA:

1. **Loading the Data:**
    - Imported the Excel dataset using the pandas library.
    - Inspected the structure of the dataset using df.head(), df.info(), and df.describe().
2. **Data Cleaning:**
    - Checked for missing values and handled them through removal or imputation.
    - Converted data types where necessary for accurate analysis.
    - Removed duplicates and filtered out irrelevant columns for clarity.
3. **Feature Selection:**
    - Identified key variables like GDP per capita, life expectancy, internet usage, and population density for further analysis.
4. **Data Transformation:**
    - Standardized column names for ease of access.
    - Aggregated data by year, region, or income group for meaningful comparisons.
5. **Visualization Tools:**
    - Used matplotlib and seaborn for plotting:
        - Histograms for distribution
        - Boxplots for outlier detection
        - Scatter plots to study relationships
        - Heatmaps for correlation analysis
6. **Initial Insights:**
    - Observed trends in life expectancy with respect to economic indicators.
    - Noted regional and income group variations in internet usage and population density.

4. Analysis on Dataset

4.i Introduction

This section delves into the core analysis of the World Bank dataset, focusing on key socio-economic indicators such as **GDP per capita**, **Life Expectancy**, **Internet Usage**, **Fertility Rate**, **CO_2 Emissions**, and **Infant Mortality Rate**. The objective of this analysis is to extract meaningful trends and relationships among variables, explore regional and income-level disparities, and present the findings through clean visualizations.

The dataset encompasses multiple countries and spans several years, enabling temporal and comparative insights across global, regional, and income-level dimensions.

4.ii General Description

The dataset used in this project was sourced from the World Bank Open Data platform, containing a variety of development indicators across different countries and time periods. The key features of the dataset are:

- Countries: Covering low, middle, and high-income economies across all major global regions.
- Time Period: Yearly data for each indicator across multiple years.
- Indicators: The dataset includes a wide array of development indicators such as:
    - GDP per capita (current US$)
    - Life Expectancy
    - Internet Users (% of population)
    - Fertility Rate
    - CO_2 Emissions (metric tons per capita)
    - Infant Mortality Rate (per 1,000 live births)
    - Income Group and Region classification

Each record in the dataset is associated with a specific country and year, allowing for longitudinal and cross-sectional analysis.

4.iii Specific Requirements, Functions, and Formulas

To perform the analysis, the following tools, libraries, and techniques were employed:

Python Libraries Used:

- pandas: For data manipulation and preprocessing.
- numpy: For numerical operations.
- matplotlib & seaborn: For visualization.
- plotly: For interactive plots.

Data Cleaning:

- Handling missing values using:
    - .isna().sum()

    - .fillna() and .dropna()
- Data type conversions for correct plotting and numerical operations.

Analysis Techniques:

- Grouping and aggregating by Region and Income Group using groupby().
- Correlation Matrix using df.corr() to understand relationships among indicators.
- Trend Analysis using line plots over years.
- Distribution Analysis using boxplots and histograms.

Formulas/Concepts Used:

- Correlation Coefficient:

$$r = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

- Mean, Median, Standard Deviation for descriptive statistics.
- Custom functions for computing normalized values and ranking.

4.iv Analysis Results

1. Income and Development Indicators:
    - High-income countries show significantly higher GDP per capita, greater internet penetration, and lower fertility/infarct mortality rates.
    - Low-income countries lag behind in almost all indicators, especially in internet access and life expectancy.
2. Correlation Observations:
    - Strong negative correlation between GDP per capita and Infant Mortality Rate.
    - Positive correlation between Internet Usage and Life Expectancy.
    - Negative correlation between Fertility Rate and GDP per capita.
3. Regional Disparities:
    - Sub-Saharan Africa exhibits high fertility rates and low life expectancy.
    - Europe & Central Asia have the highest internet usage and life expectancy rates.
4. Temporal Trends:
    - Global Internet Usage has increased significantly over the past decade.
    - Fertility Rate and Infant Mortality Rate are declining trends globally.
    - CO_2 Emissions per capita are relatively stable, with some decrease in high-income regions due to policy interventions.

4.v Visualization

Below are a few key visualizations generated through the notebook:

1. Correlation Heatmap:

```
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
```

2. GDP vs Life Expectancy Scatter Plot:

```
sns.scatterplot(data=df, x='GDP_per_capita', y='Life_Expectancy',
hue='Income_Group')
```

3. Regional Distribution of Internet Usage:

```
sns.boxplot(data=df, x='Region', y='Internet_Users')
plt.xticks(rotation=90)
```

4. Trend of Fertility Rate Over Years:

```
sns.lineplot(data=df_grouped_by_year, x='Year', y='Fertility_Rate')
```

5. CO_2 Emissions by Income Group:

```
sns.barplot(data=df, x='Income_Group', y='CO2_Emissions')
```

5. Conclusion

The analysis of the World Bank development indicators dataset has revealed significant insights into the socio-economic dynamics across countries and over time. By exploring variables such as GDP per capita, Life Expectancy, Fertility Rate, Internet Usage, and CO_2 Emissions, this project has highlighted the disparities and correlations among nations grouped by region and income level.

Key Takeaways:

- Income-Level Impact: High-income countries exhibit better health and development outcomes, such as higher life expectancy and lower infant mortality rates. Low-income countries show lagging metrics across multiple indicators.
- Strong Correlations:
  - GDP per capita positively correlates with life expectancy and internet usage.
  - GDP per capita negatively correlates with fertility and infant mortality rates.
- Temporal Improvement:
  - Global trends show improvements in internet penetration and life expectancy.
  - Fertility rates and infant mortality rates are gradually declining in most regions.
- Regional Variation:
  - Sub-Saharan Africa remains behind in most development indicators.
  - Europe and North America show strong performance in education, technology, and healthcare indicators.

This analysis has underlined the importance of targeted policy intervention and development programs in underperforming regions. By using data-driven insights, global development agencies and governments can make informed decisions to address inequalities and foster sustainable growth.