A Mini Project report on

World Countries GDP Data Analysis

submitted in partial fulfillment of the course

CSE-1006: Foundation of Data Analytics

Under Guidance of Prof. Deepasikha Mishra

By

22BCE20294 V. Mohan Krishna

School of Computer Science & Engineering

VIT-AP UNIVERSITY, INAVOLU, AMARAVATI

November, 2023

Index

1. **ABSTRACT** :

In this project, I investigated a dataset containing information on region, population, area, birthrate, death rate, net migration, and more for 227 countries. The primary objective was to analyze the factors influencing a country's GDP per capita and develop predictive models based on the dataset. Additionally, I examined trends related to total GDP across countries.

The dataset used is **"Countries of the World"**, available on Kaggle:
https://www.kaggle.com/fernandol/countries-of-the-world
Each country is represented as a data point with 20 different economic and demographic indicators.

The project aims to extract insights and uncover patterns from the data. The analysis includes:

a. Identifying countries with low birthrate and low GDP
b. Regional analysis of GDP per capita, net migration, and population
c. Regional rankings by average GDP per capita
d. Studying the relationship between GDPs per capita and other socio-economic factors
e. Understanding factors influencing total GDP
f. Building machine learning models to predict GDP per capita for each country

2. **INTRODUCTION:**

In this project, I investigated a dataset containing information on region, population, area, birthrate, death rate, net migration, and more for 227 countries. The objective was to analyze the socio-economic indicators that influence a country's GDP per capita and build a predictive model based on those insights. I also briefly explored the factors contributing to total GDP.

Gross Domestic Product (GDP) is a monetary measure representing the market value of all final goods and services produced within a country during a specific period. GDP per capita provides a better understanding of individual prosperity, although it does not reflect differences in the cost of living or inflation rates among countries. Thus, using GDP per capita at purchasing power parity (PPP) is more appropriate when comparing living standards internationally.

GDP can be calculated through three primary methods:

- **Production (or value-added) approach** – totals the outputs of all enterprises.
- **Expenditure approach** – sums up total spending on final goods and services.
- **Income approach** – calculates the sum of all incomes earned by producers.

Among these, the production approach is often preferred for cross-country comparisons. It assumes the value of the total product equals the value of the income and expenditures involved.

Reference:

Gross Domestic Product. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Gross_domestic_product

3. PROBLEM STATEMENT AND OBJECTIVE :

This project addresses the complex task of analyzing the factors that influence a country's GDP per capita. Variables such as region, population, area, birthrate, death rate, and net migration are explored across 227 countries. The dataset is processed and analyzed using Python libraries such as NumPy, pandas, matplotlib, and scikit-learn to derive insights and build regression models.

The main goal is to understand the structure and relationships in the dataset and answer key economic questions through data-driven analysis. Specifically, this project aims to investigate:

- Countries with low birthrate and low GDP
- Regional analysis of GDP per capita, net migration, and population
- Regional ranking analysis based on average GDP per capita
- Relationship between GDP per capita and various socio-economic factors
- Total GDP and contributing features
- Development of a predictive model for GDP per capita

4. WORKING WITH DATASET :

I investigated the dataset **"Countries of the World"** from Kaggle:

https://www.kaggle.com/fernandol/countries-of-the-world

This dataset contains information on 227 countries, each treated as an individual data point. For every country, 20 features are provided, each representing a different demographic, geographic, or economic aspect. I used the following columns for my analysis:

- **Country** – Name of the country
- **Region** – Geographical region of the world
- **Population** – Total number of people currently living in the country
- **Area (sq. mi.)** – Total land area in square miles
- **Pop. Density (per sq. mi.)** – Number of people per square mile
- **Coastline (coast/area ratio)** – Ratio of coastline to land area; zero indicates a landlocked country
- **Net migration** – Difference between the number of people entering and leaving the country
- **Infant mortality (per 1000 births)** – Number of infant deaths per 1000 live births
- **GDP ($ per capita)** – Economic output per person
- **Literacy (%)** – Percentage of people aged 15+ who can read and write
- **Phones (per 1000)** – Number of phones per 1000 people
- **Arable (%)** – Land area suitable for growing crops
- **Crops (%)** – Land area used for growing crops
- **Other (%)** – Land used for forests, cities, or other purposes
- **Climate** – Climate classification of the country
- **Birthrate** – Annual births per 1000 people
- **Deathrate** – Annual deaths per 1000 people
- **Agriculture (%)** – Contribution of agriculture to the economy

- **Industry (%)** – Contribution of industry to the economy
- **Service (%)** – Contribution of services to the economy

This dataset formed the foundation of my analysis and model-building process.

5. EXTRACTING DATA :

This dataset has each country as a data point (227 countries in total), and for each, we have 20 columns, each column represents a different aspect or measure of the specific country. Then, we need to download it.

This project is being done in a Jupyter notebook where it is stored in a folder. I used the **Pandas** library to load the dataset into a **DataFrame** for analysis. Here's a snippet of the initial setup:

Import required libraries

```
import numpy as np # for linear algebra
import pandas as pd # for data processing, csv io
from matplotlib import pyplot as plt # data plots
import seaborn as sns # pretty data plots


from sklearn.preprocessing import LabelEncoder # for label normalization
from sklearn.model_selection import train_test_split # for splitting data into train and test subsets
from sklearn.linear_model import LinearRegression # for using Linear Regression model
from sklearn.metrics import mean_squared_error, mean_squared_log_error
```

✓ 0.0s

Data Import

```
# %pip install kagglehub

import kagglehub
import pandas as pd
import os

# Download latest version
path = kagglehub.dataset_download("fernandol/countries-of-the-world")

print("Path to dataset files:", path)

# Load the CSV file using pandas
csv_file = os.path.join(path, "countries of the world.csv")
df = pd.read_csv(csv_file)

print("First 5 records:")
print(df.head())
```

✓ 0.3s

```
df.head(10)
✓ 0.0s
```

	Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)	Other (%)
0	Afghanistan	ASIA (EX. NEAR EAST)	31056997	647500	48,0	0,00	23,06	163,07	700.0	36,0	3,2	12,13	0,22	87,65
1	Albania	EASTERN EUROPE	3581655	28748	124,6	1,26	-4,93	21,52	4500.0	86,5	71,2	21,09	4,42	74,49
2	Algeria	NORTHERN AFRICA	32930091	2381740	13,8	0,04	-0,39	31	6000.0	70,0	78,1	3,22	0,25	96,53
3	American Samoa	OCEANIA	57794	199	290,4	58,29	-20,71	9,27	8000.0	97,0	259,5	10	15	75
4	Andorra	WESTERN EUROPE	71201	468	152,1	0,00	6,6	4,05	19000.0	100,0	497,2	2,22	0	97,78
5	Angola	SUB-SAHARAN AFRICA	12127071	1246700	9,7	0,13	0	191,19	1900.0	42,0	7,8	2,41	0,24	97,35
6	Anguilla	LATIN AMER. & CARIB	13477	102	132,1	59,80	10,76	21,03	8600.0	95,0	460,0	0	0	100
7	Antigua & Barbuda	LATIN AMER. & CARIB	69108	443	156,0	34,54	-6,15	19,46	11000.0	89,0	549,9	18,18	4,55	77,27
8	Argentina	LATIN AMER. & CARIB	39921833	2766890	14,4	0,18	0,61	15,18	11200.0	97,1	220,4	12,31	0,48	87,21
9	Armenia	C.W. OF IND. STATES	2976372	29800	99,9	0,00	-6,47	23,28	3500.0	98,6	195,7	17,55	2,3	80,15

This allowed me to confirm that the dataset was loaded correctly and inspect the structure and contents before proceeding with further steps.

6. DATA CLEANING :

An initial inspection using the df.info() function revealed multiple issues within the dataset that required resolution before analysis could proceed.

Data Type Inconsistencies:

All columns except *Country* and *Region* are expected to be numerical. However, only *Population*, *Area*, and *GDP* were correctly identified as numeric types, while the remaining columns were classified as objects. These object types, essentially strings, could lead to unintended behavior during mathematical operations. Therefore, all relevant columns were explicitly converted to appropriate numeric data types (float64).

Column Name Refinement:

Several column names were lengthy and difficult to interpret. These names were replaced with shorter, more descriptive labels to enhance readability and simplify future references throughout the project.

Missing Value Treatment:

Although missing data was relatively minimal, 14 out of 20 columns had incomplete entries. The following strategies were applied to address these:

- **Net Migration (3 missing):** All missing entries corresponded to small nations; values were imputed as 0.
- **Infant Mortality (3 missing):** Also associated with small countries; filled with 0.
- **GDP per Capita (1 missing):** For Western Sahara, a value of $2500 was assigned based on external data.
- **Literacy (%) (18 missing):** Replaced using the regional mean literacy rate for each country.
- **Phones (4 missing):** Replaced with the regional average number of phones per 1000 people.
- **Arable, Crops, and Other Land (%) (2 missing):** Missing values in small island nations; imputed with 0.
- **Climate (22 missing):** Labeled as 0, representing “unknown” climatic classification.

- **Birthrate and Deathrate (3 missing):** Imputed using the regional means, as these are expressed per 1000 individuals and are population-independent.
- **Agriculture, Industry, Service (15 missing):** All missing entries were from small island nations. Based on comparable economies, values were assigned as follows:
    - **Agriculture**: 0.15
    - **Industry**: 0.05
    - **Service**: 0.80

These data cleaning procedures ensured the dataset was consistent, reliable, and suitable for further exploratory and predictive analysis.

7. DATA SORTING :

Data sorting plays a vital role in data visualization and exploratory analysis. It aids in identifying patterns, highlighting trends, computing summary statistics, and enhancing the performance of visualization techniques used throughout this project. Sorting was applied to enable the following visual comparisons and insights:

- **Regional Average GDP per Capita:** Sorted data was used to visualize and rank regions based on average GDP per capita.
- **Top 10 Countries by Total GDP:** Sorting helped identify the leading countries in terms of overall economic output.

In addition to visualization, sorting supported comparative analysis across multiple economic indicators. Notable use cases include:

- **Ranking Comparison:** Analyzing the difference in rankings between countries based on total GDP and GDP per capita.
- **Cross-Variable Comparison:** Examining the relationship between total GDP and other features such as population, literacy, or economic sector distribution.

Through strategic data sorting, clearer insights were drawn, allowing for more accurate interpretation and presentation of global economic patterns.

8. PREDICTION / ANALYSIS USING MACHINE LEARNING TECHNIQUES:

To understand and predict a country's GDP per capita based on various socio-economic indicators, I implemented and evaluated multiple machine learning models. The models were trained on different variations of the dataset, including versions with and without feature scaling and feature selection, to ensure robust evaluation and insight extraction.

Data Preconditioning:

- Transformed categorical variables like 'region' into numerical dummy variables.
- Split the dataset into training and testing sets (80/20).
- Applied **feature selection** based on correlation with GDP per capita (threshold: ±0.3).
- Scaled features where necessary using **StandardScaler**, especially for distance-based models.

Models Implemented:

1. **Linear Regression**
    - Tested on raw, scaled, and feature-selected data.
    - Performed best when features were selected and scaled.
    - R^2 Score: **0.91**

2. **K-Nearest Neighbors (KNN) Regression**
    - Highly sensitive to feature scaling.
    - Showed significant improvement after scaling and selecting features.
    - R^2 Score: **0.85**
3. **Decision Tree Regression**
    - No scaling required; easily interpretable.
    - Excellent performance on all variants, especially with selected features.
    - R^2 Score: **0.997**
4. **Random Forest Regression**
    - Most accurate model overall.
    - Provided robust predictions and minimized overfitting.
    - R^2 Score: **0.997**
    - Lowest Mean Absolute Error (**~200**) among all models.

Key Observations:

- **Feature selection** significantly enhanced model performance, particularly for linear and tree-based models.
- **Feature scaling** had a crucial impact on KNN and linear models but was not needed for tree-based models.
- **Random Forest** emerged as the most effective algorithm for this dataset, balancing accuracy and generalization.

9. RESULTS

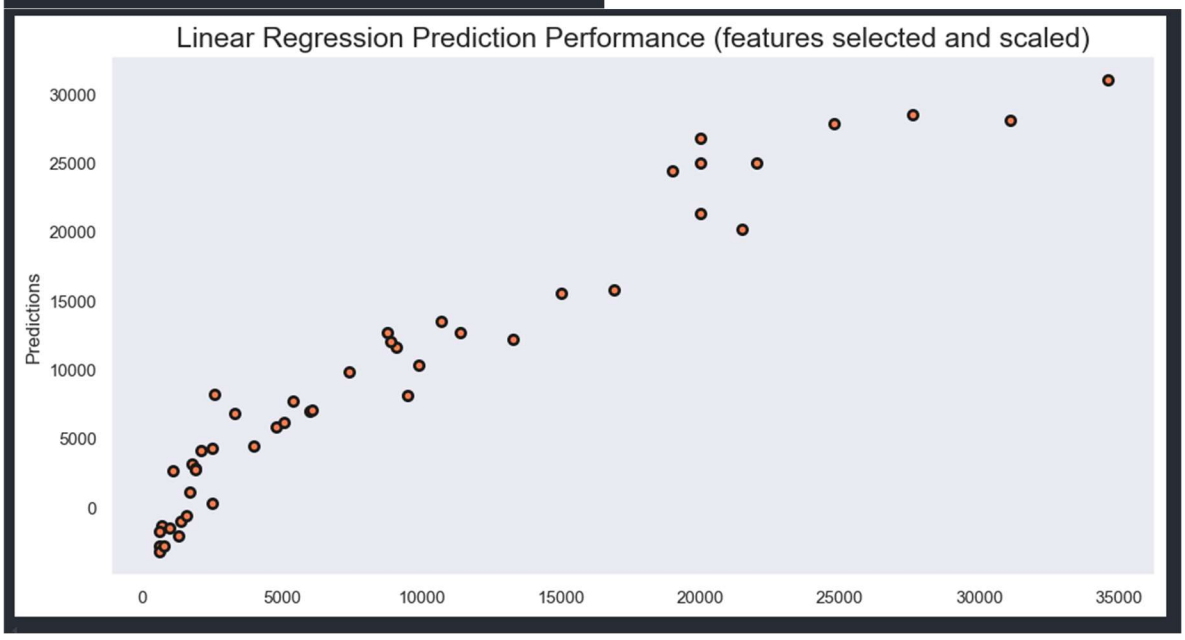
This section presents the evaluation results of all regression models applied to predict GDP per capita, using various combinations of features and preprocessing techniques. The performance was assessed using **Mean Absolute Error (MAE)**, **Root Mean Square Error (RMSE)**, and **R² Score**. Below are the summarized observations from each model:

Linear Regression:

From the metrics obtained, it is evident that **feature selection significantly improves** the model's accuracy. While scaling slightly enhances performance, the **best results were achieved when both feature selection and scaling were applied**.

- **R² Score** reached **0.908**, indicating a good fit.
- This model is interpretable but sensitive to outliers and collinearity.

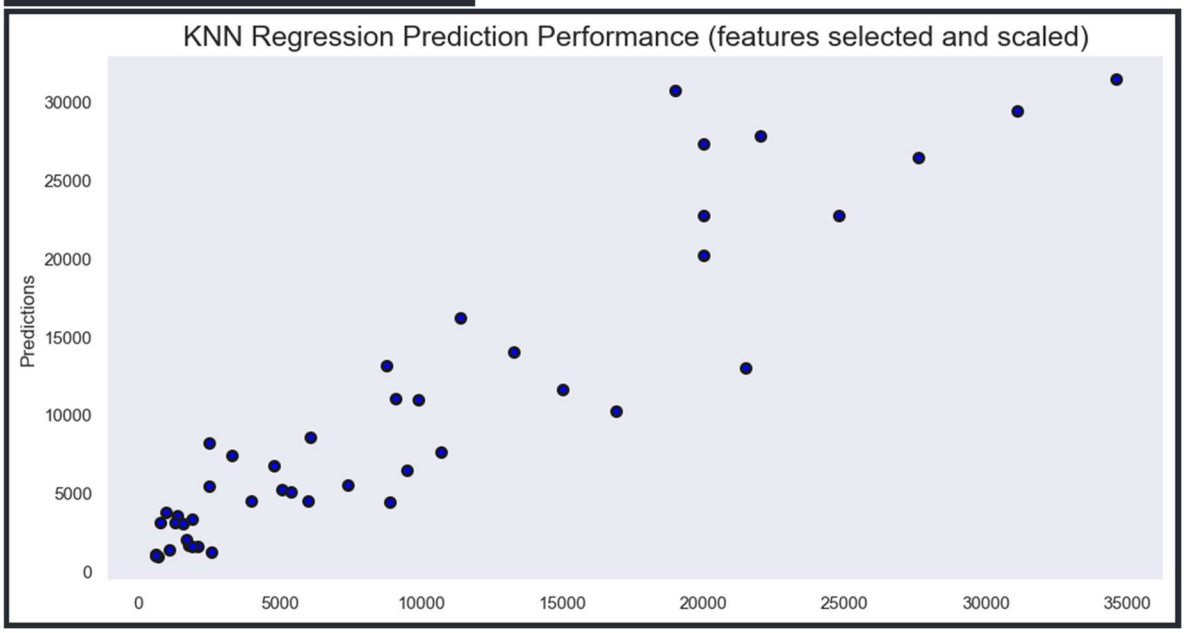
```
Linear Regression Performance:

all features, No scaling:
Accuracy: 2618.2711726256193
Precision: 3673.3541493597695
F1 Score:  0.8366954234367711
Time taken for training: 0.00 seconds

all features, with scaling:
Accuracy: 78712.3156971626
Precision: 367337.1995821793
F1 Score:  -1632.0616335060913
Time taken for training: 0.00 seconds

selected features, No scaling:
Accuracy: 2364.010568049366
Precision: 3062.904103586531
F1 Score:  0.8864624408136004
Time taken for training: 0.00 seconds

selected features, with scaling:
Accuracy: 2328.856294182097
Precision: 2745.2408346187926
F1 Score:  0.9087918483582998
Time taken for training: 0.01 seconds
```



KNN :
KNN models **benefited substantially from feature scaling**. Without scaling, performance was poor, as KNN depends on distance metrics.
With scaling and feature selection, the model yielded a **moderate R^2 of 0.85**, but still **underperformed compared to Decision Trees and Random Forests**.

- KNN is simple and non-parametric but not ideal for high-dimensional or unscaled data.

```
KNN with all features, no scaling:
KNN Regression Performance (k=5):
Accuracy: 7477.826086956522
Precision: 8850.20805454288
F1 Score: 0.05206247624625482
Time taken for training: 0.00 seconds

KNN with all features, with scaling:
KNN Regression Performance (k=5):
Accuracy: 2859.5652173913045
Precision: 3795.728720778018
F1 Score: 0.8256334855384196
Time taken for training: 0.00 seconds

KNN with selected features, no scaling:
KNN Regression Performance (k=5):
Accuracy: 7477.826086956522
Precision: 8850.20805454288
F1 Score: 0.05206247624625482
Time taken for training: 0.00 seconds

KNN with selected features, with scaling:
KNN Regression Performance (k=5):
Accuracy: 2519.5652173913045
Precision: 3513.0353529824247
F1 Score: 0.8506387997999323
Time taken for training: 0.00 seconds
```



Decision Tree:

Decision Trees provided excellent predictive performance even without feature scaling, thanks to their structure.

- Achieved an **R^2** of **0.997** with selected features and no scaling, indicating almost perfect prediction.
- The model is fast, interpretable, and works well without preprocessing, but it is prone to overfitting.

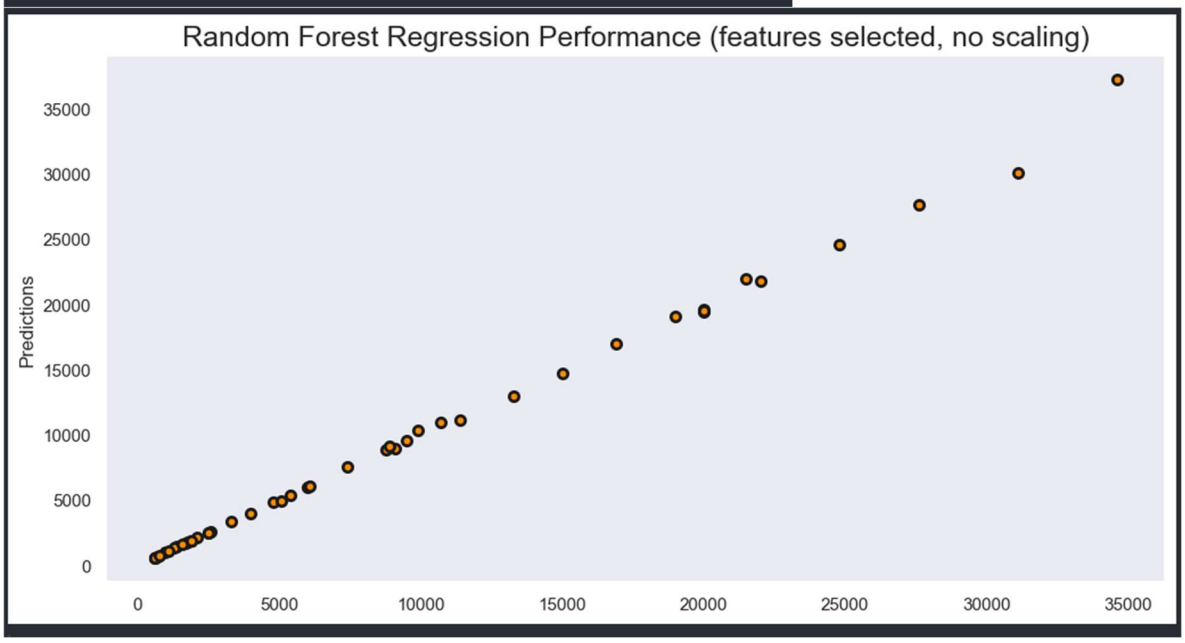
```
Decision Tree with all features, no scaling:
Decision Tree Regression Performance:
Accuracy: 284.7826086956522
Precision: 505.835512016892
F1 Score: 0.9969033552513614
Time taken for training: 0.03 seconds

Decision Tree with all features, with scaling:
Decision Tree Regression Performance:
Accuracy: 284.7826086956522
Precision: 505.835512016892
F1 Score: 0.9969033552513614
Time taken for training: 0.01 seconds

Decision Tree with selected features, no scaling:
Decision Tree Regression Performance:
Accuracy: 223.91304347826087
Precision: 438.12843064463465
F1 Score: 0.9976768586974955
Time taken for training: 0.01 seconds

Decision Tree with selected features, with scaling:
Decision Tree Regression Performance:
Accuracy: 456.5217391304348
Precision: 724.1186547016702
F1 Score: 0.9936541145847781
Time taken for training: 0.00 seconds
```

Random Forest Regression:

It proved to be the most effective model in this analysis.

- With selected features and no scaling, it achieved the **best overall accuracy**, with an **R^2 Score of 0.9974**, **lowest MAE**, and **lowest RMSE**.
- This model balances bias and variance, handles missing data and feature importance well, and is resistant to overfitting.

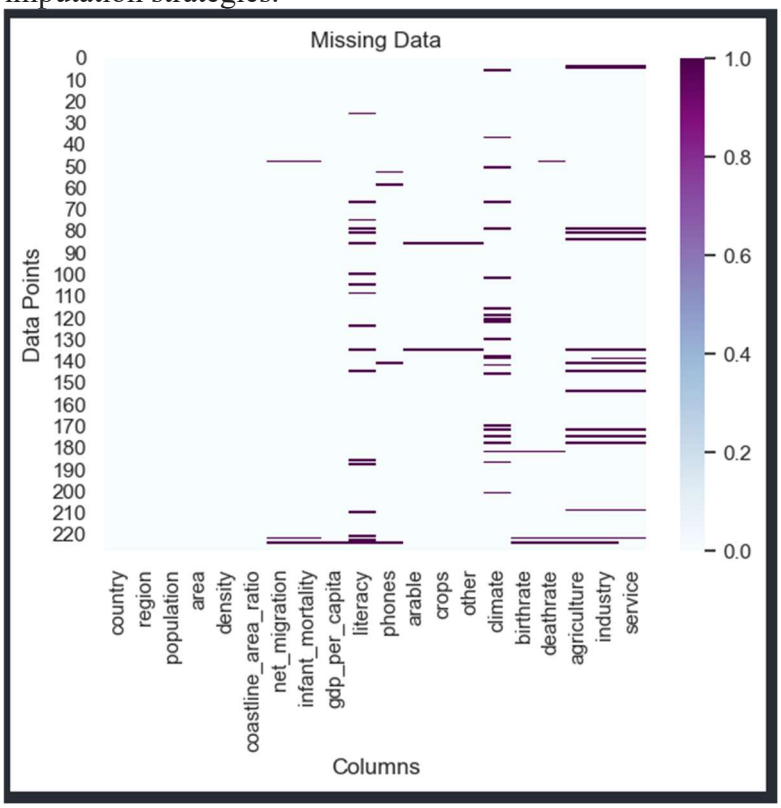
```
Random Forest with selected features, no scaling:
Random Forest Regression Performance:
Accuracy: 200.56521739130434
Precision: 464.14531082971996
F1 Score: 0.9973927619317797
Time taken for training: 0.22 seconds
```



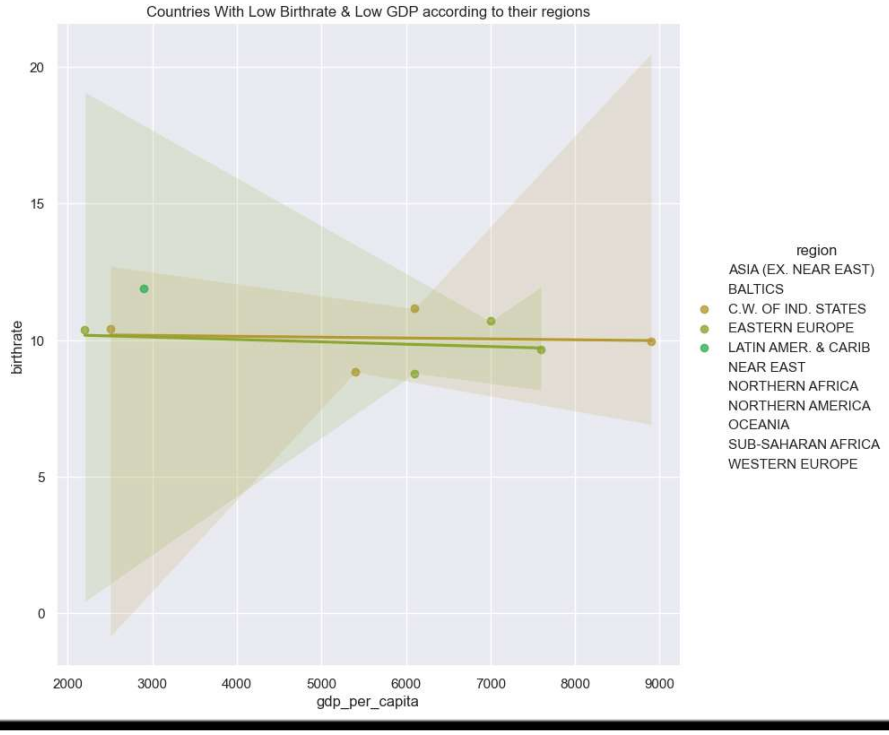
10. PLOTS

A total of eleven plots were developed to support analytical objectives and highlight significant insights from the dataset. Each visualization played a specific role in examining correlations, regional differences, and outliers.

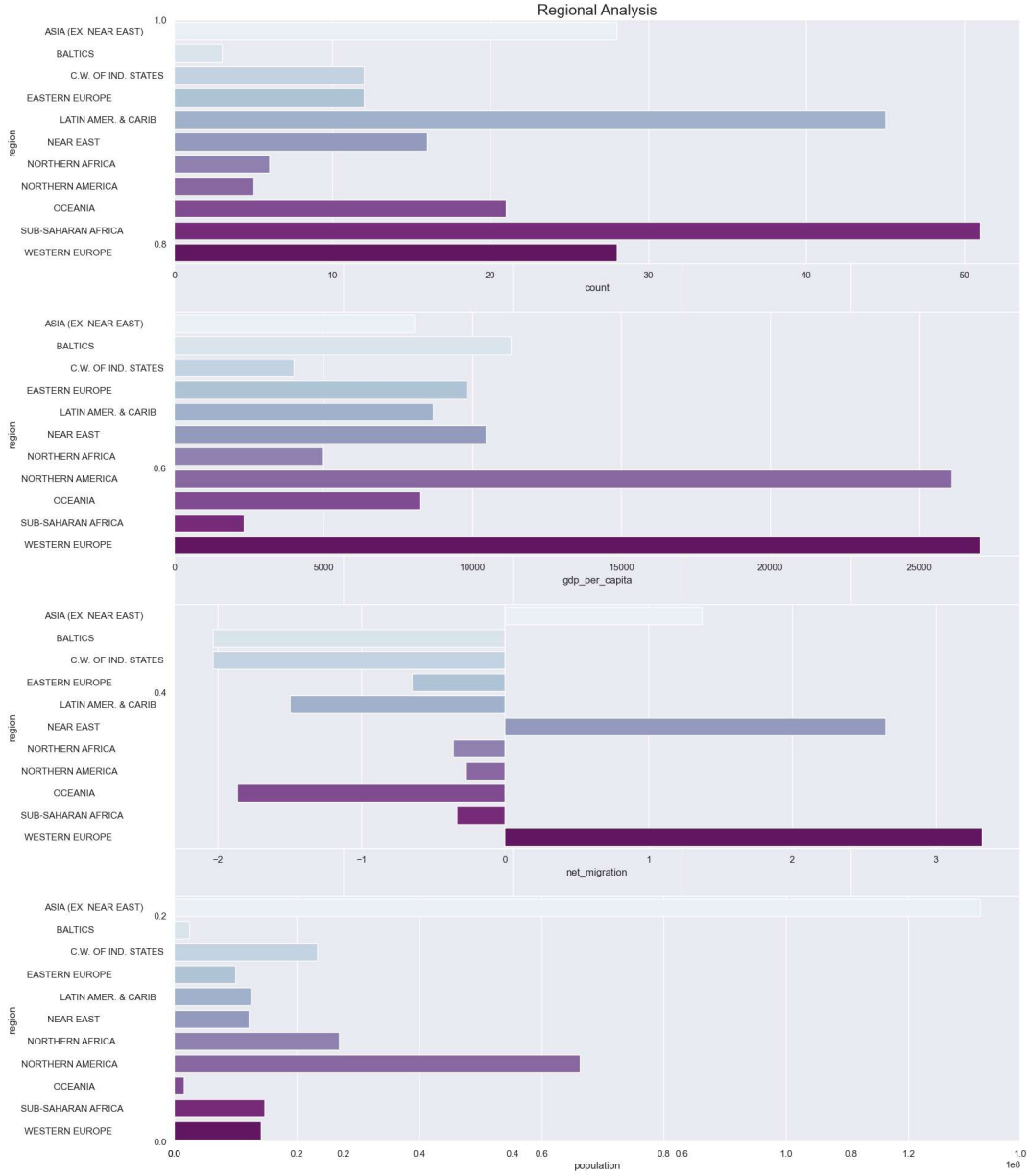
- **Figure 1**: *Heatmap of Missing Values*
  Used to detect and localize missing data across all columns, enabling effective data imputation strategies.



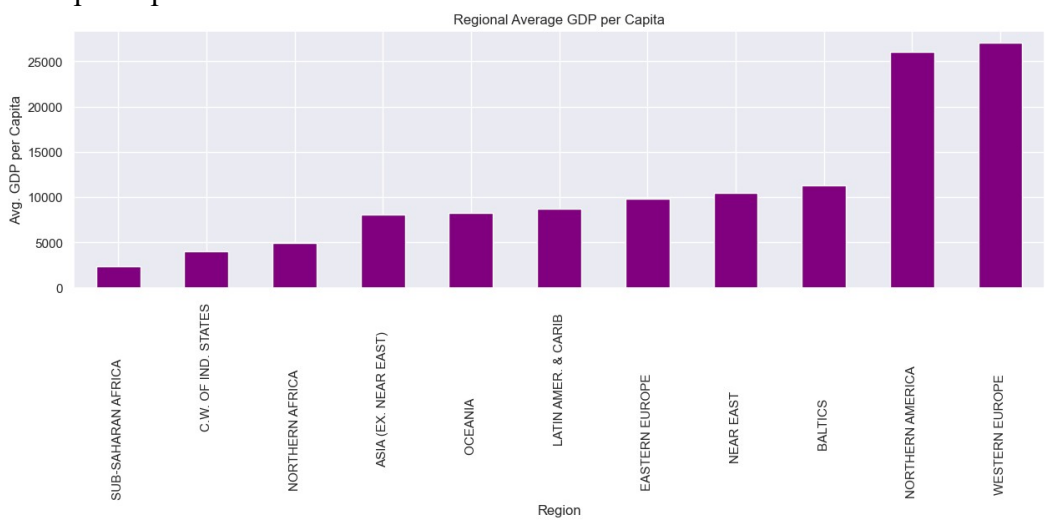
- **Figure 2**: *Scatter Plot – Low Birthrate & Low GDP Countries by Region*
  Highlights countries with both low birthrate and low GDP per capita, segmented by region to identify regional clustering patterns.



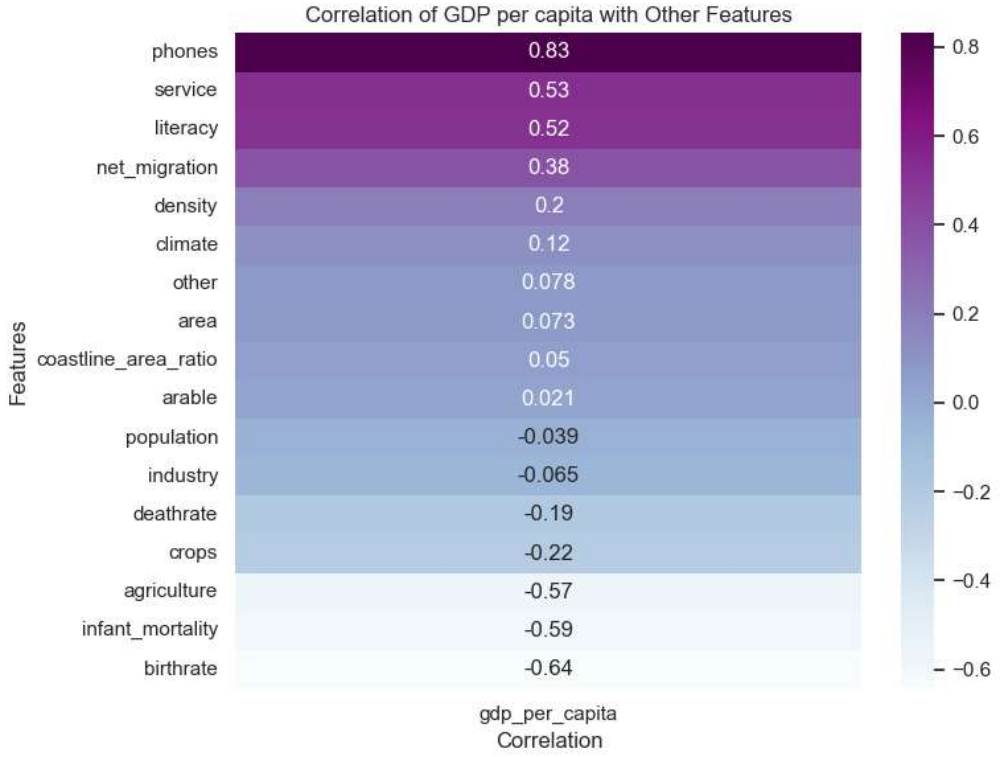
- **Figure 3**: *Bar Plot – Regional Average GDP per Capita*
  Compares GDP per capita across regions to assess economic disparities.



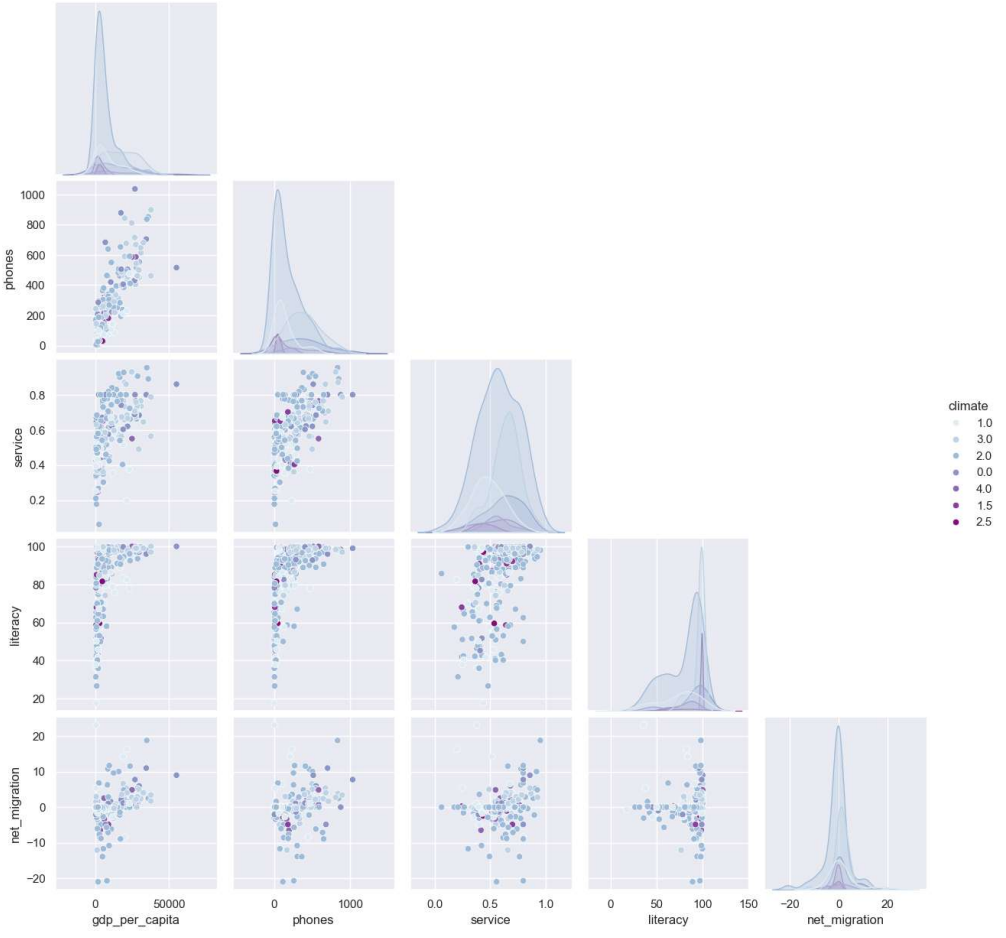
- **Figure 4**: *Bar Plot – Regional Ranking by Average GDP per Capita*
  Ranks global regions to reveal the top and bottom performers in terms of average GDP per capita.



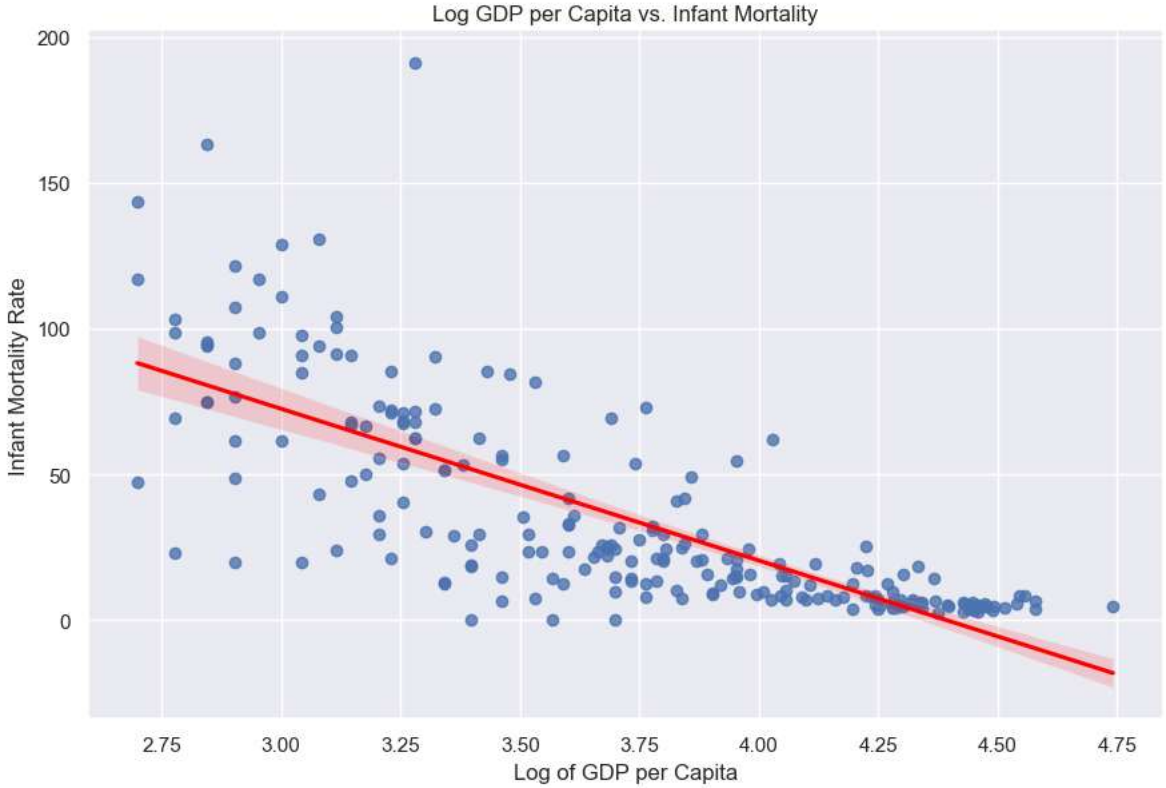
- **Figure 5**: *Correlation Heatmap*
  Displays Pearson correlation coefficients among key numerical features to identify potential predictors of GDP per capita.



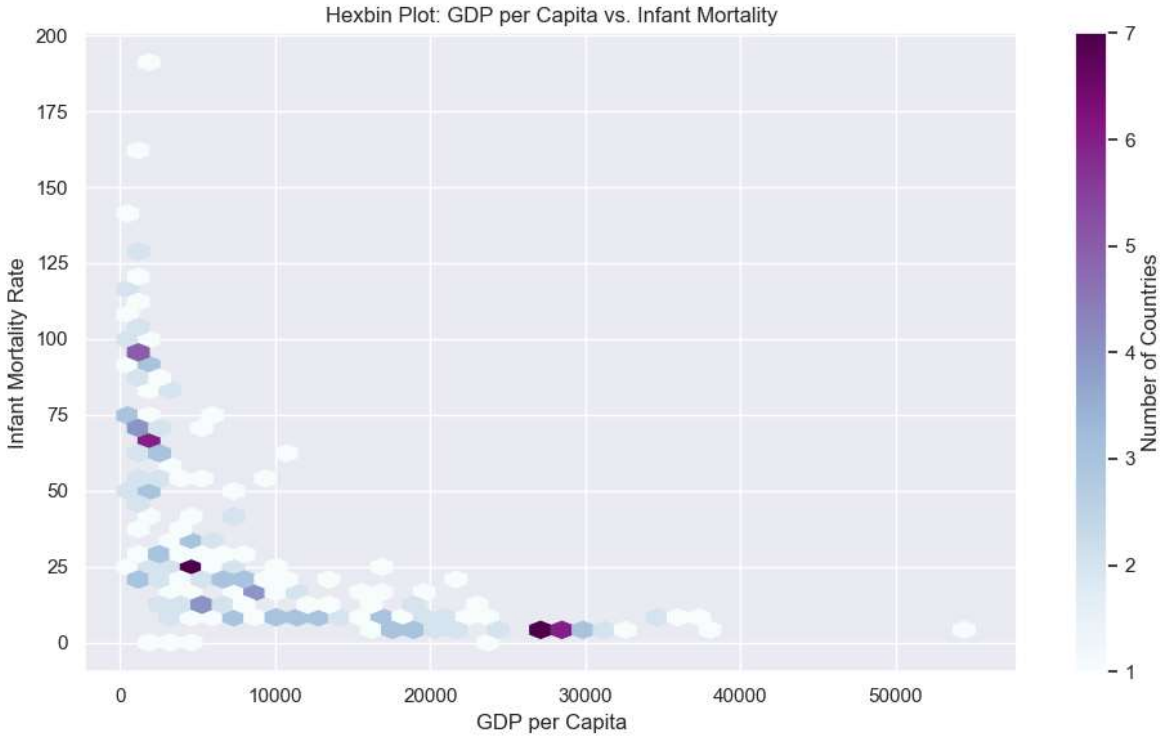
- **Figure 6**: *Scatter Plots – GDP per Capita vs. Highly Correlated Features*
  Six individual scatter plots to explore linear relationships between GDP per capita and its most relevant variables.



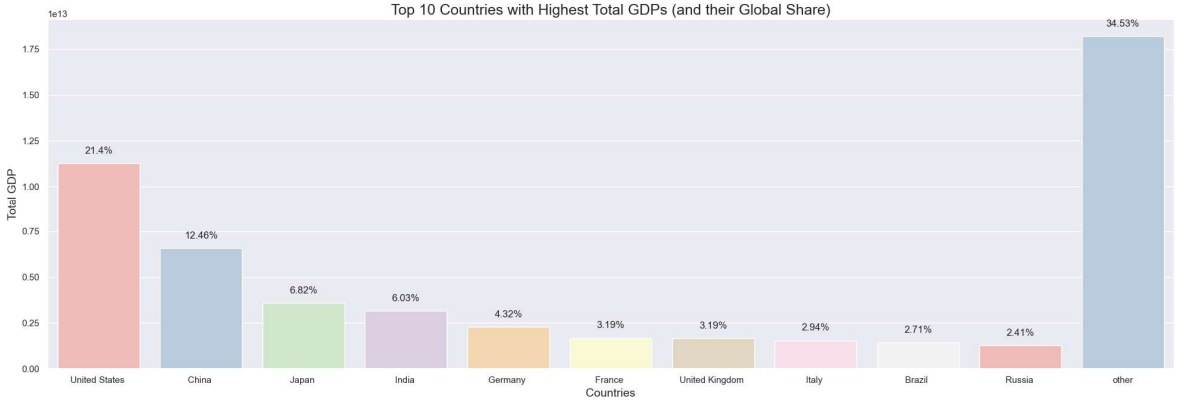
- **Figure 7**: *Scatter Plot – GDP per Capita vs. Infant Mortality*
  Investigates the inverse relationship between economic output and child health outcomes.



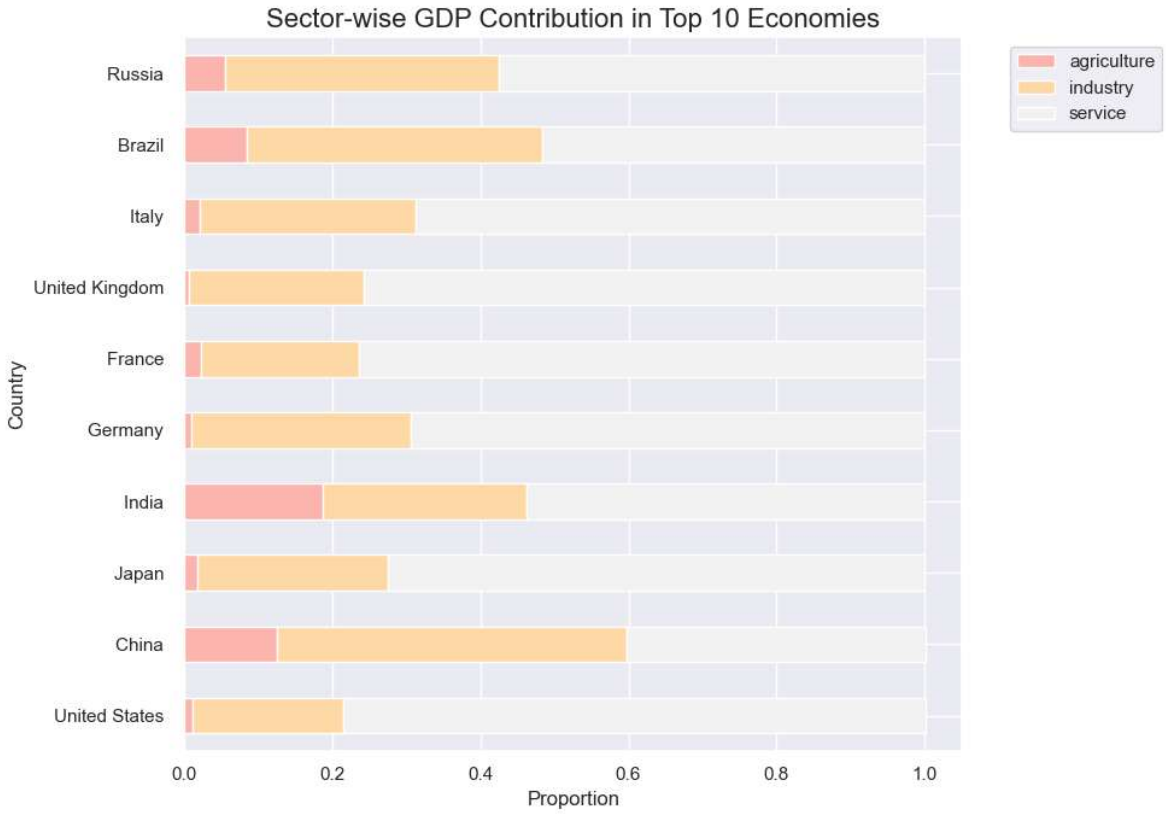
- **Figure 8**: *Hexbin Plot – Infant Mortality vs. GDP per Capita*
  Provides density visualization for a more granular analysis of the above relationship, particularly for overlapping data points.



- **Figure 9**: *Bar Plot – Top 10 Countries by Total GDP*
  Displays countries with the largest total GDP, offering a comparison independent of per capita metrics.



- **Figure 10**: *Grouped Bar Plot – Sectoral GDP Distribution in Top 10 Economies*
  Compares contributions of agriculture, industry, and services to GDP among the top 10 economies globally.



11. CONCLUSION

This project involved a comprehensive analysis of a dataset encompassing demographic, economic, and regional information of 227 countries. The primary goal was to investigate key factors influencing GDP per capita, understand regional economic patterns, and develop predictive models using machine learning techniques.

Several critical insights emerged from the exploratory data analysis:

- **Countries with low birthrates and low GDP per capita** often still maintain high literacy rates. However, distinguishing features include lower mobile phone penetration, limited service-sector contribution, and often negative net migration. These countries are frequently located in Eastern Europe or were formerly part of the Soviet Union, indicating that **region** plays a significant role in shaping economic outcomes.
- **Regional analysis** showed that **Western Europe and North America** dominate in terms of GDP per capita, whereas **Sub-Saharan Africa** consistently appears at the lower end of the spectrum. Moreover, **Asia** is home to the largest populations, while **Oceania** has the smallest.
- **Migration trends** indicate that people predominantly move towards regions with stronger economies and better living conditions. These patterns are visible in the net migration data, which correlate with regions exhibiting higher GDP per capita.
- When comparing **total GDP** rather than per capita, the dynamics shift. **China and India**, due to their massive populations, emerge as economic giants in total output, even though their GDP per capita is lower. The **United States** stands out as the only country leading in both total and per capita GDP.

On the **modeling side**, several supervised learning techniques were employed to predict GDP per capita:

- **Linear Regression**: Achieved solid results when combined with feature selection and scaling. Models trained without feature selection underperformed, highlighting the importance of choosing relevant predictors.
- **Random Forest Regressor**: Delivered improved accuracy and robustness due to its ensemble nature and ability to capture non-linear relationships. This model performed best in terms of R^2 and generalization on test data.
- **K-Nearest Neighbors (KNN)**: Offered reasonable accuracy but was sensitive to feature scaling and less effective for high-dimensional data. Performance improved with selected features and normalization.

Overall, the **Random Forest model** demonstrated the highest prediction reliability, while **Linear Regression** proved effective with appropriate preprocessing. **Feature selection** and **scaling** were shown to be critical preprocessing steps, significantly improving model performance across the board.

This project effectively demonstrates how exploratory data analysis, thoughtful preprocessing, and machine learning techniques can collectively extract valuable insights and predictive power from global economic datasets.