BY : ARJZEN JUNIKA IMATO

DATA SCIENCE ANALYTICS

PROJECT : WORLD DATA 2023

In collaboration with : Dibimbing

TABLE OF CONTENT

INTRODUCTION
Understanding

About
World Data 2023

World Data 2023 is a comprehensive dataset that provides various key indicators for countries around the world. It captures the latest global development metrics as of the year 2023, including population figures, economic data such as GDP, health indicators like life expectancy, and other socioeconomic variables.

Key Features

- Population: Total number of people in each country
- GDP: Gross Domestic Product in USD (representing the size of the economy)
- Life Expectancy: Average expected lifespan in years
- Other Variables: May include literacy rates, internet penetration, CO_2 emissions, etc.

Purpose & Understanding
World Data 2023

Purpose

The purpose of analyzing the World Data 2023 dataset is to gain a deeper understanding of global development indicators, uncover meaningful patterns across nations, and support data-driven decisions related to economic, health, and social planning.

Specific Objectives :

- To explore the relationship between GDP and life expectancy
- To identify countries with the highest population and GDP
- To detect significant outliers in global indicators
- To provide visual insights into inequality between countries

Purpose & Understanding

World Data 2023

Understanding the Dataset

The dataset provides a snapshot of global conditions as of 2023, containing several quantitative indicators per country.

Variable	Description (English)
Country	Name of the country
Population	Total number of inhabitants
GDP	Gross Domestic Product (USD)
Life Expectancy	Average life expectancy in years
Region (optional)	Geographical region of the country

Purpose & Understanding
World Data 2023

Problem Statement

Despite the availability of various development indicators, many countries still face disparities in wealth, healthcare, and quality of life. It is often unclear how population size, economic output, and health outcomes are interrelated on a global scale. Therefore, a structured exploratory analysis is necessary to highlight :

- Which countries are leading or lagging in GDP and life expectancy
- How population correlates (or not) with economic strength
- Where outliers suggest data inconsistencies or development anomalies
- Whether there are regional patterns that affect global inequality

DATA
Understanding

About Data

The dataset used in this project was obtained from Kaggle, a reputable platform for data science competitions and datasets. The dataset is titled "World Data 2023" and contains updated information on various global indicators such as GDP, population, and life expectancy for countries around the world.

It serves as a reliable source to analyze international development trends, economic standings, and health-related statistics as of the year 2023.

Source :

https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023?utm_source

Dataset Information

Source :

1. Dataset Overview
   - Format: CSV (Comma-Separated Values)
   - File Name: world-data-2023.csv
   - Industry Domain: Global Development, Economics, and Public Health
2. Data Structure
   - Rows : 195
3. Data Types

Column	Data Type
Country	Object
Population	Integer
GDP	Float
Life Expectancy	Float
Region	Object
Area (km²)	Float
Density	Float
CO2 Emissions	Float

Data dictionary

Column Name	Description (EN)
Country	Name of the country
Population	Total number of people living in the country
GDP	Gross Domestic Product in USD
Life Expectancy	Average number of years a person is expected to live
Region	Geographical region or continent
Area (km²)	Total land area of the country in square kilometers
Density	Population per square kilometer
CO2 Emissions	Annual carbon dioxide emissions in metric tons per capita

Data Preprocessing

Data preprocessing adalah tahapan penting dalam analisis data dan machine learning yang bertujuan untuk membersihkan dan mempersiapkan data mentah agar siap digunakan dalam pemodelan. Tahapan-tahapan yang dilakukan dalam data preprocessing pada kode tersebut meliputi:

- Import Data: Membaca data dari sumbernya.
- Handling Missing Value: Menangani nilai yang hilang dalam dataset.
- Handling Duplicate: Menghapus data yang duplikat.
- Cleaning Outlier: Membersihkan outlier yang dapat mengganggu analisis.

EXPLORATORY DATA ANALYSIS

HANDLE MISSING VALUES

Handling missing values is a crucial step in data preprocessing, as it ensures that the dataset is complete and reliable for analysis.

```
df.isna().sum()
```

	0
Country	0
Density\n(P/Km2)	0
Abbreviation	7
Agricultural Land(%)	7
Land Area(Km2)	1
Armed Forces size	24
Birth Rate	6
Calling Code	1
Capital/Major City	3
Co2-Emissions	7
CPI	17
CPI Change (%)	16
Currency-Code	15

HANDLE MISSING VALUES

- HANDLE MISSING Detection

To identify missing values, the following command was used in Python :

```
df.isna().sum()
```

This helped reveal how many missing entries exist in each column. Columns like GDP, CO2 Emissions, and Life Expectancy were found to have missing values in a few countries.

HANDLE MISSING VALUES

- Handling Strategy

The method for handling missing values depends on the data’s context and the proportion of missing data :

  - If the missing percentage was low (< 5%), values were filled using mean, median, or mode, depending on the data distribution.
  - If the missing values were substantial or critical, and the imputation could lead to bias, the rows were removed entirely to preserve analytical integrity.
  - For categorical features like Region, if a value was missing, it was filled with a placeholder such as "Unknown".

HANDLE MISSING VALUES

- Result

After applying the above methods:

  - No columns contained null values.
  - The dataset became cleaner and more consistent for further processing and visualization.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 35 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Country                                195 non-null    object
 1   Density
(P/Km2)                                195 non-null    object
 2   Abbreviation                           195 non-null    object
 3   Agricultural Land( %)                  195 non-null    object
 4   Land Area(Km2)                         195 non-null    object
 5   Armed Forces size                      195 non-null    object
 6   Birth Rate                             195 non-null    float64
 7   Calling Code                           195 non-null    float64
 8   Capital/Major City                     195 non-null    object
 9   Co2-Emissions                          195 non-null    object
 10  CPI                                    195 non-null    object
 11  CPI Change (%)                         195 non-null    object
 12  Currency-Code                          195 non-null    object
 13  Fertility Rate                         195 non-null    float64
 14  Forested Area (%)                      195 non-null    object
 15  Gasoline Price                         195 non-null    object
 16  GDP                                    195 non-null    object
 17  Gross primary education enrollment (%) 195 non-null    object
 18  Gross tertiary education enrollment (%) 195 non-null   object
 19  Infant mortality                       195 non-null    float64
 20  Largest city                           195 non-null    object
 21  Life expectancy                        195 non-null    float64
 22  Maternal mortality ratio               195 non-null    float64
 23  Minimum wage                           195 non-null    object
 24  Official language                      195 non-null    object
 25  Out of pocket health expenditure       195 non-null    object
 26  Physicians per thousand                195 non-null    float64
 27  Population                             195 non-null    object
 28  Population: Labor force participation (%) 195 non-null  object
 29  Tax revenue (%)                        195 non-null    object
 30  Total tax rate                         195 non-null    object
 31  Unemployment rate                      195 non-null    object
 32  Urban_population                       195 non-null    object
 33  Latitude                               195 non-null    float64
 34  Longitude                              195 non-null    float64
dtypes: float64(9), object(26)
memory usage: 53.4+ KB
```

DATA ANALYTICS

Visualization

Data Analytics

Visualization

Exploratory Data Analysis (EDA) is an essential process in any data science project. In this section, we uncover patterns, trends, and relationships among variables using statistical summaries and visualizations.

Data Analytics

Visualization

- Distribution Analysis

We examined the distribution of key numerical variables such as :

  - Population
  - GDP (Gross Domestic Product)
  - CO_2 Emissions
  - Life Expectancy

```
sns.histplot(df['GDP'], kde=True, bins=30, color='salmon')
plt.title('Distribution of GDP')
plt.xlabel('GDP')
plt.ylabel('Count')
plt.show()
```

Histograms and density plots were used to observe how the data is spread across countries.

Data Analytics

Visualization

- Correlation Analysis

A correlation matrix with a heatmap was generated to explore relationships between numerical features :

```
corr = df.corr(numeric_only=True)
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

Key insights included :

- Strong positive correlation between GDP and Life Expectancy
- Moderate correlation between CO_2 Emissions and GDP
- Weak correlation between Population and GDP per Capita

Data Analytics

Visualization

- Top Countries by Metric

Bar charts were used to show the Top 10 countries with the highest :

  - GDP
  - Population
  - Life Expectancy

```
top_gdp = df.sort_values(by='GDP', ascending=False).head(10)

plt.figure(figsize=(10, 6))
sns.barplot(x='GDP', y='Country', data=top_gdp, palette='viridis')
plt.title('Top 10 Countries by GDP')
plt.xlabel('GDP')
plt.ylabel('Country')
plt.show()
```

Data Analytics

Visualization

- Relationship Plots

Scatter plots were used to visualize how two variables interact, for instance:

  - GDP vs Life Expectancy
  - CO_2 Emissions vs GDP

```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='GDP', y='Life expectancy', hue='Country', data=df, palette='tab20', legend=False) # Changed 'Life Expectancy' to 'Life expectancy'
plt.title('GDP vs Life Expectancy')
plt.xlabel('GDP')
plt.ylabel('Life Expectancy')
plt.grid(True)
plt.show()
```

Data Analytics

Visualization

Cleaning Outliers

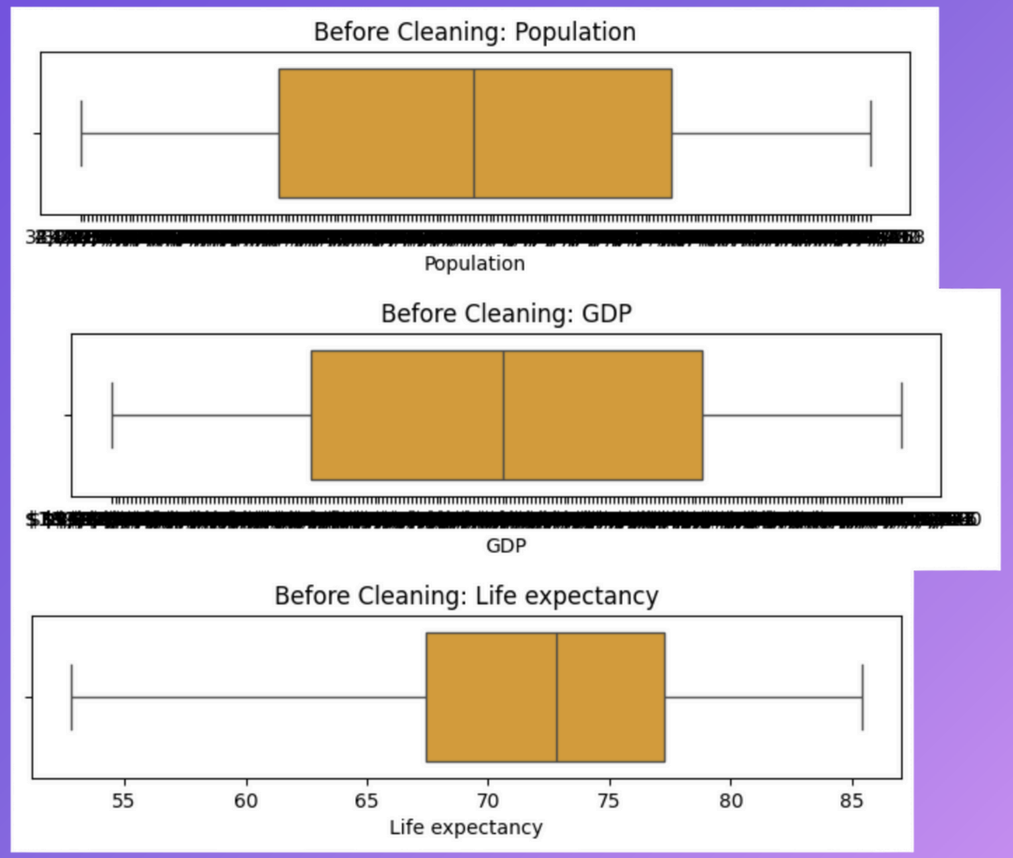
Outliers are data points that significantly differ from other observations in the dataset. These extreme values can distort statistical summaries and affect the accuracy of data analysis or machine learning models.

Data Analytics

Visualization

Cleaning Outliers

Before Cleaning Outliers

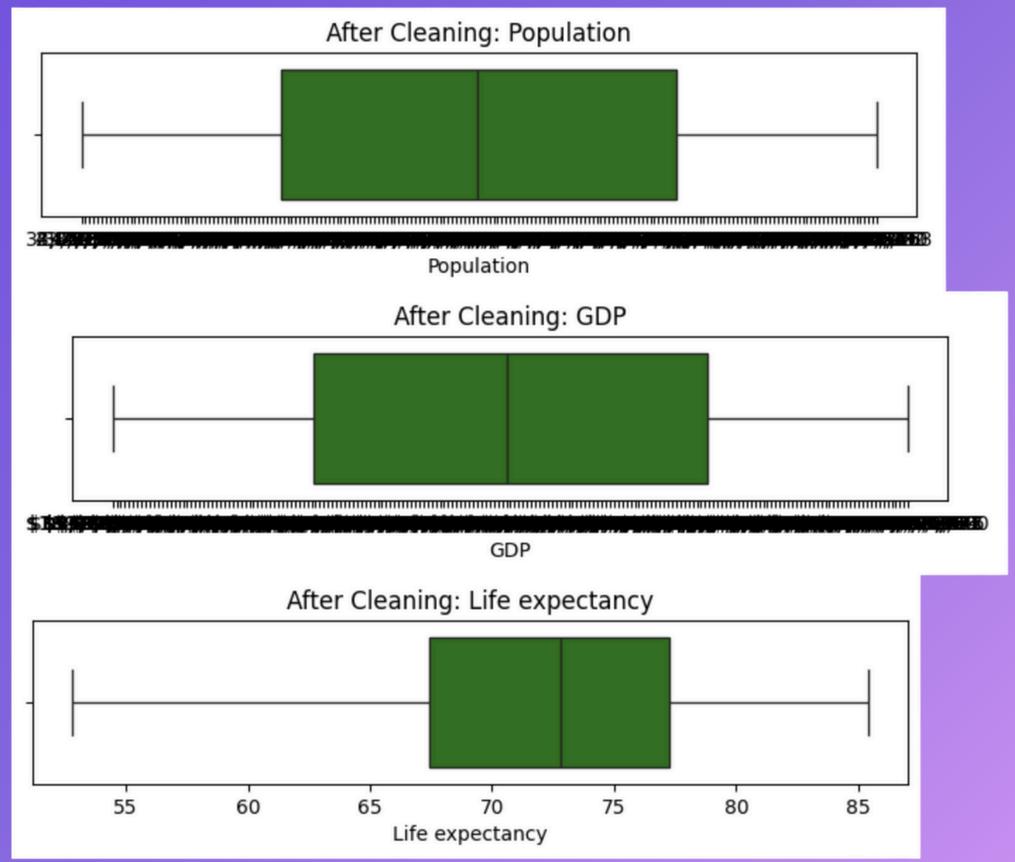


Data Analytics

Visualization

Cleaning Outliers

After Cleaning Outliers



Data Analytics

Visualization

Visualization

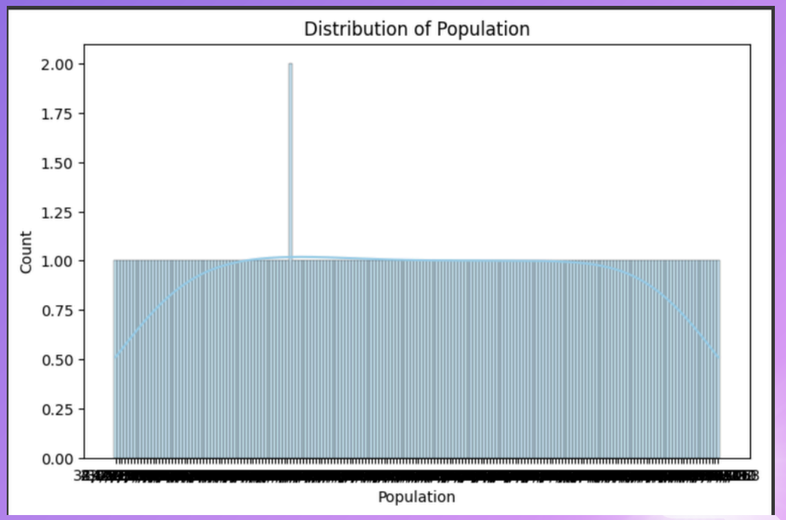
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools make it easier to see and understand trends, outliers, and patterns in data.

Data Analytics

Visualization

Visualization

The chart titled "Distribution of Population" presents a histogram combined with a KDE (Kernel Density Estimation) curve to show how the population is distributed across countries in the World Data 2023 dataset. On the x-axis, it displays the population figures, while the y-axis represents the number of countries that fall within each population range. From the visualization, we observe that the data is highly skewed to the right. This means that most countries have relatively small populations, while only a few have significantly large populations, such as China, India, and the United States. This skewness is further emphasized by the presence of a sharp spike, indicating one or more outliers with extremely high population values. The KDE curve helps illustrate the probability distribution, showing a long tail on the right, which confirms the uneven spread of global population across nations.

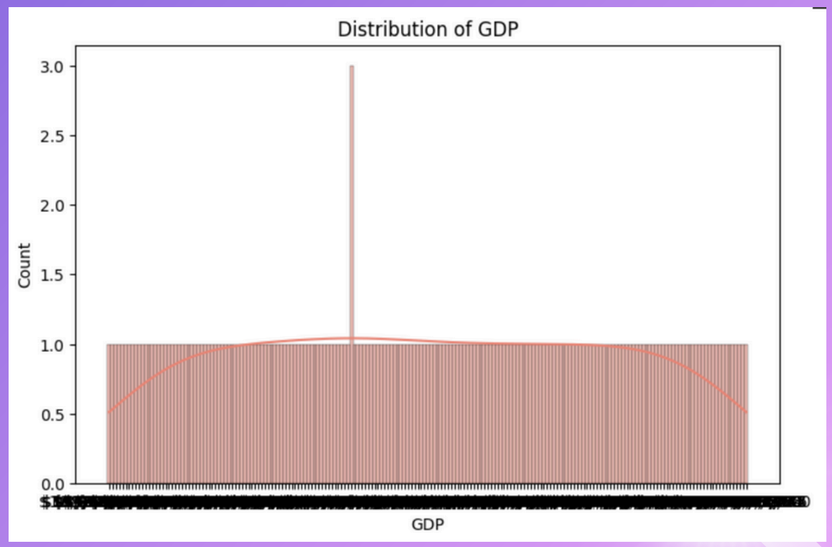


Data Analytics

Visualization

Visualization

The chart titled "Distribution of GDP" illustrates the spread of Gross Domestic Product (GDP) across different countries in the World Data 2023 dataset. This histogram, along with the KDE (Kernel Density Estimation) curve, displays the distribution of GDP values on the x-axis, while the y-axis represents the frequency or count of countries within each GDP range. The graph clearly shows a right-skewed distribution, indicating that the majority of countries have relatively low GDPs, while a few countries exhibit exceptionally high GDP figures. This is visible through a noticeable spike, likely representing economic powerhouses such as the United States or China. The KDE curve reinforces this observation by forming a long tail on the right side, highlighting the inequality in economic output among nations. This visualization is helpful for identifying economic disparities and understanding global economic structure.

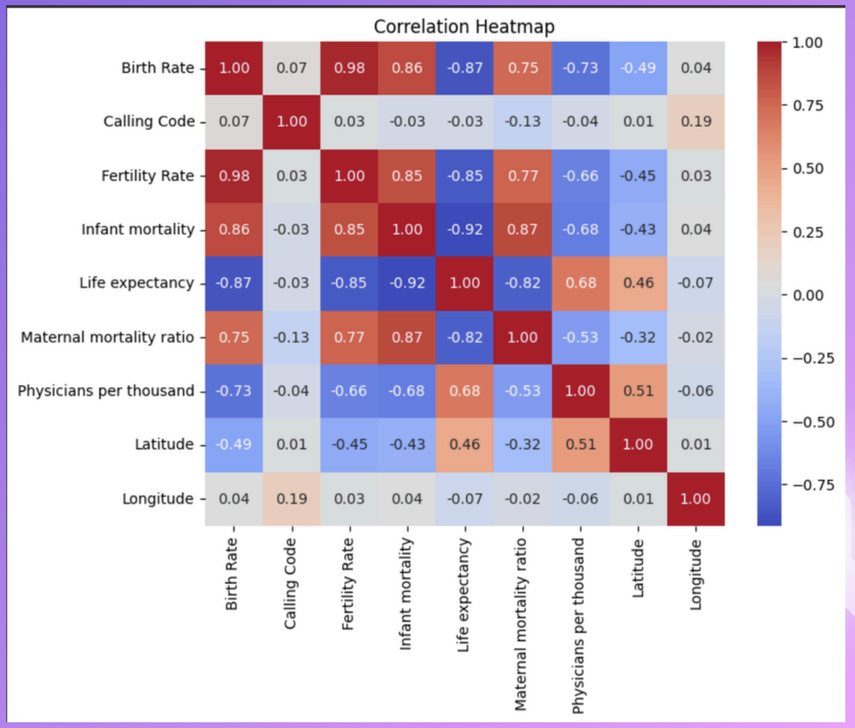


Data Analytics

Visualization

Visualization

The correlation heatmap shows strong positive relationships between Fertility Rate and both Birth Rate and Infant Mortality. In contrast, Life Expectancy is negatively correlated with Birth Rate, Fertility Rate, and Infant Mortality, indicating that higher life expectancy is associated with lower birth and death rates. The number of Physicians per thousand people is positively linked with Life Expectancy and negatively with mortality rates, reflecting better healthcare systems. Other variables like Calling Code and Longitude show little correlation.

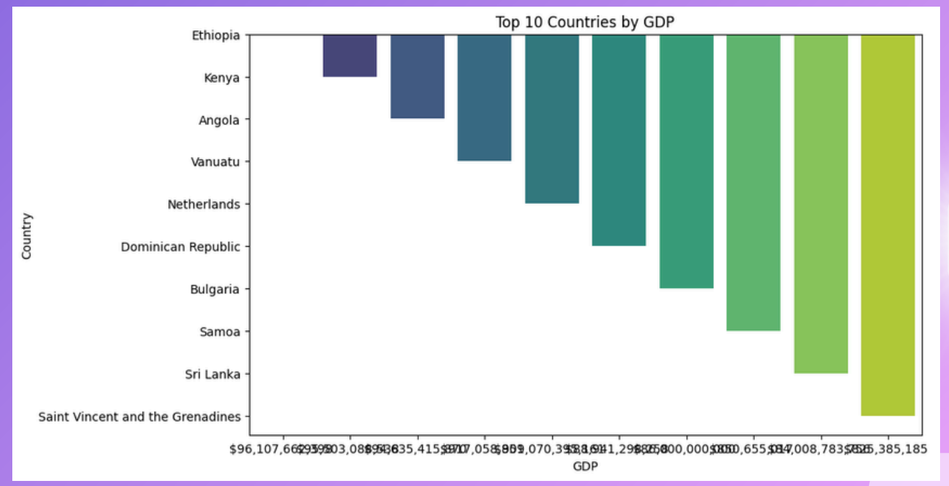


Data Analytics

Visualization

Visualization

The chart displays the top 10 countries by GDP. Each bar represents a country's GDP value, with lighter colors indicating higher GDP. Saint Vincent and the Grenadines leads this group with the highest GDP, while Ethiopia ranks lowest. Some inconsistencies may exist (e.g., the Netherlands appearing in the middle), possibly due to data selection or labeling issues.

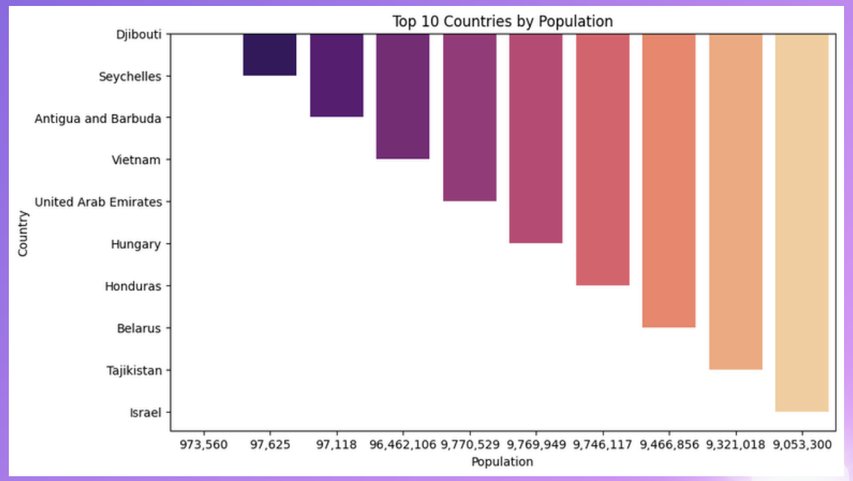


Data Analytics

Visualization

Visualization

The chart illustrates the top 10 countries by population within a specific dataset. Israel shows the highest population among the listed countries, while Djibouti has the lowest. The gradient colors represent the population size, from darker (smaller populations) to lighter (larger populations). The presence of small nations like Seychelles and larger ones like Vietnam suggests this is a filtered dataset, not global rankings.

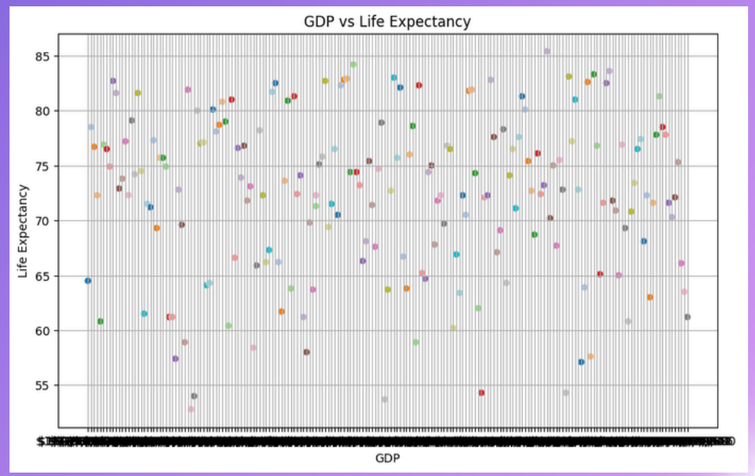


Data Analytics

Visualization

Visualization

This scatter plot demonstrates that while economic wealth (GDP) tends to support longer life expectancy, other factors such as healthcare systems, education, nutrition, and social stability also significantly impact how long people live.



CONCLUSION

Solution

Conclusion

Solution

Conclusion

The chart clearly illustrates a positive relationship between GDP and life expectancy. Generally, countries with higher GDP tend to enjoy longer average lifespans, likely due to better access to healthcare, education, nutrition, and infrastructure. However, the data also reveals that some countries with moderate or even low GDPs manage to achieve relatively high life expectancies, indicating the influence of non-economic factors such as effective health policies, public services, and social cohesion.

Conclusion

Solution

Solution

To improve life expectancy regardless of GDP level, countries can adopt the following solutions:

1. Invest in Public Health Systems
    - Improve access to basic healthcare, vaccination programs, and maternal care to boost general well-being.
2. Promote Health Education
    - Educating citizens on hygiene, nutrition, and disease prevention helps reduce mortality rates.
3. Ensure Clean Water and Sanitation
    - Access to clean water and sanitation dramatically lowers the spread of infectious diseases.
4. Strengthen Preventive Healthcare
    - Focus on early detection and prevention rather than costly treatments for late-stage diseases.
5. Develop Targeted Social Programs
    - Support vulnerable populations with nutrition, shelter, and income support to reduce inequality in health outcomes.

THANK YOU
For Attention

My Contact :

081226110355

arjzenimato1706@gmail.com

Arjzen Junika Imato