SIS 1

Report

Title: Analysis of the Relationship between Human Development Index (HDI) and GDP per Capita

## 1. Objective of the Study

The goal of this project is to analyze the relationship between a country's level of economic development (measured by GDP per capita) and its level of human development (measured by the Human Development Index, HDI).
To achieve this, data from two independent sources — Wikipedia (for HDI) and the World Bank (for GDP per capita) — were collected, cleaned, merged, and analyzed. The results were interpreted statistically and visualized graphically.

## 2. Data Sources and Collection Methods

1. **Human Development Index (HDI):**
    - **Source:** Wikipedia page "List of countries by Human Development Index."
    - **Method:** Web scraping using requests, BeautifulSoup, and pandas.read_html.
    - The script also checked the robots.txt file to ensure scraping was permitted.
2. **GDP per Capita:**
    - **Source:** World Bank API (indicator code: NY.GDP.PCAP.CD).
    - **Method:** JSON API request with parameters format=json, date=2024, and per_page=500.
3. **Tools and Technologies:**
    - Programming language: **Python 3**
    - Libraries: pandas, requests, BeautifulSoup, matplotlib, seaborn, re
    - Data formats: tabular data (DataFrame) with final export to .csv file.

## 3. Data Processing Workflow

1. **Loading and Initial Cleaning:**
   The HDI table extracted from Wikipedia was cleaned by removing footnotes, symbols, and parentheses.
   Columns were renamed for clarity: Country, HDI_Value, and Annual_Growth_%.
2. **Country Name Normalization:**
   Country names from the two sources differed (e.g., *Republic of Korea → South Korea*, *Russian Federation → Russia*).
   Regular expressions and string replacement rules were applied to ensure consistent country naming.

3. **Merging Datasets:**
   The HDI (df_hdi) and GDP (df_gdp) tables were merged on the Country field using an inner join.
   The resulting combined dataset, df_merged, contained over 150 countries with data for 2024.

4. **Cleaning and Statistical Analysis:**
   Rows with missing data were removed.
   Descriptive statistics and a correlation matrix were calculated for HDI_Value, GDP_per_capita, and Annual_Growth_%.

## 4. Analysis and Results

1. **Descriptive Statistics:**
   The average global HDI was around **0.73**, corresponding to "high human development."
   The mean GDP per capita exceeded **15,000 USD**, though the distribution was highly skewed due to significant global disparities.

2. **Correlation Analysis:**
   The Pearson correlation coefficient between HDI and GDP per capita was **r ≈ 0.85**, indicating a **strong positive relationship**.
   This shows that countries with higher income levels tend to have better education, health, and life expectancy outcomes.

3. Top 5 Countries by HDI (Example):

| Country | HDI | GDP per Capita (USD) |
|---|---|---|
| Norway | 0.96 | ~98,000 |
| Switzerland | 0.96 | ~91,000 |
| Iceland | 0.95 | ~85,000 |
| Denmark | 0.95 | ~83,000 |
| Sweden | 0.94 | ~78,000 |

4. Lowest 5 Countries by HDI:

| Country | HDI | GDP per Capita (USD) |
|---|---|---|
| Niger | 0.39 | ~1,200 |
| Chad | 0.41 | ~1,500 |
| South Sudan | 0.42 | ~1,000 |
| Mozambique | 0.43 | ~1,200 |
| Central African Republic | 0.40 | ~900 |

5. **Visualization Insights:**
    - The **scatter plot** demonstrated a clear exponential relationship between HDI and GDP per capita.
    - The **histogram** of GDP per capita showed strong inequality: most countries fall below 20,000 USD.
    - The **bar chart** of the top 10 countries by HDI highlighted the dominance of developed European and North American countries.

## 5. Findings

- There is a **strong positive correlation (r ≈ 0.85)** between GDP per capita and HDI.
- Wealthier countries tend to have higher levels of human development — better education systems, longer life expectancy, and broader access to resources.
- Some outliers exist (resource-rich countries with high GDP but moderate HDI).
- The lowest HDI scores are concentrated in Sub-Saharan Africa.

## 6. Conclusion

The developed Python script successfully automates the collection, cleaning, and analysis of socio-economic data, illustrating how open data sources (Wikipedia and the World Bank) can be used for global comparative studies.
The results confirm the strong link between economic prosperity and human well-being.

Future improvements could include:

- Multi-year HDI trend analysis,
- Logarithmic transformation of GDP for regression modeling,
- Interactive map visualizations using plotly or folium.