09.01.2026

DSA 210 TERM PROJECT REPORT

Ceren Kula

Project Title: Economic Growth, Environmental Impact, and Human Development: A Cross-Country Analysis of CO_2 Emissions, GDP, HDI, Life Expectancy, and Gender Inequality

1. Introduction

This report presents a cross-country data science project that investigates the relationships between **CO_2 emissions, Gross Domestic Product (GDP), Human Development Index (HDI), life expectancy, and gender inequality**. The study is structured to combine economic, environmental, and social indicators in order to evaluate whether economic growth is consistently associated with higher human development outcomes, or whether trade-offs emerge in terms of sustainability and social equality.

At a global level, countries with higher income levels tend to exhibit better performance in education, healthcare, and overall living standards. However, these improvements often coincide with increased environmental pressure, particularly through higher carbon emissions. At the same time, social dimensions such as gender inequality may not improve at the same pace as economic indicators, raising questions about the inclusiveness and long-term sustainability of development. This project aims to empirically examine these patterns using internationally comparable data.

From a methodological perspective, the project follows the main stages of the data science pipeline. Multiple open-source datasets are collected and integrated into a unified cross-country dataset covering the period **2010–2019**, with a focus on constructing a clean **2019 cross-sectional dataset** for analysis. The study applies data cleaning, exploratory data analysis, correlation analysis, hypothesis testing, and basic machine learning techniques to uncover patterns and relationships among the variables.

The main objectives of this project are to analyze the relationship between countries' CO_2 emissions and GDP, examine whether higher GDP levels are associated with improvements in HDI and life expectancy, and assess the link between economic growth and gender inequality. In addition, the study compares economic, social, and environmental indicators across regions using appropriate statistical tests and visualizations.

The report is structured as follows. The first section introduces the research topic, motivation, and objectives of the study. The second section describes the data sources and data preparation process. The third section outlines the data analysis methodology, including exploratory data analysis and hypothesis testing. The fourth section presents the empirical findings, supported by visualizations and machine learning results. Finally, the last section discusses the limitations of the study, suggests directions for future research and conclusion.

1.2 Motivation

I chose this topic because I am genuinely interested in the relationship between **economic growth, sustainability, and life expectancy**, and in understanding how development can be evaluated beyond purely economic measures. While high-income countries often perform better in education, health, and overall living standards, they are also responsible for a large share of global **CO_2 emissions**, which raises important concerns regarding environmental sustainability. This contrast led me to question whether **true development** can occur without increasing environmental costs, and whether countries that prioritize sustainability can still achieve strong economic and social outcomes. In addition, I was interested in examining whether improvements in economic indicators are accompanied by reductions in **gender inequality**, or whether social disparities persist despite economic progress.

Beyond the subject matter itself, this project was also motivated by a desire to **develop stronger analytical and scientific data analysis skills**. Since collecting original, individual-level data was not feasible for this topic, I deliberately chose to work with **international, publicly available datasets** that are widely used in academic and policy-oriented research. This allowed me to learn how real-world scientific data is structured, combined, and analyzed in practice.

Through this project, I aimed to improve my ability to work with **multidimensional datasets**, handle missing and inconsistent data, and apply statistical and exploratory data analysis techniques in a systematic way. The skills gained in this process are transferable and can be applied to other academic and applied data analysis projects, particularly those involving social, economic, and environmental indicators. Overall, this project helps me understand whether key global indicators—**CO_2 emissions, GDP, HDI, life expectancy, and gender inequality**—tend to move together or diverge as countries develop, while also serving as an opportunity to strengthen my capacity to conduct **scientifically grounded data analysis** using real-world data.

---

2. Data Sources

All data used in this project are **publicly available, ethically obtained**, and commonly used in academic and policy-oriented research. The datasets cover economic, environmental, and social development indicators at the country level and were selected to enable a comprehensive cross-country analysis.

The main data sources used in this study are summarized below:

Dataset	**Description**	**Source**
CO_2 Emissions	Annual country-level CO_2 emission data	Kaggle – CO_2 Emissions
GDP (current US$)	Annual GDP values by country in current USD	World Bank – GDP (NY.GDP.MKTP.CD)
Human Development Indicators	HDI, Life Expectancy, Gender Inequality Index (GII)	UNDP / Kaggle

2.1 Data Cleaning and Preparation

The data cleaning and preparation process was conducted in a structured and transparent manner to ensure consistency and analytical reliability across datasets. All raw datasets were first inspected to understand their structure, variable definitions, and reporting formats.

To ensure cross-dataset consistency, all datasets were standardized using **ISO3 country codes** and aligned over the **2010–2019** period. Regional aggregates and non-country entities (such as *World*, *Europe & Central Asia*, and similar groups) were excluded to focus solely on country-level observations.

Cleaning Steps by Dataset

- **CO_2 Emissions Data**
  Only observations corresponding to individual countries were retained. Variable names were standardized (e.g., country name, year, total CO_2 emissions), and the dataset was restricted to the 2010–2019 period. Irrelevant columns were removed to retain only essential indicators.
- **GDP Data (World Bank)**
  GDP data were loaded from World Bank CSV files and converted from wide format to long format to ensure compatibility with other datasets. Only country codes and annual GDP values between 2010 and 2019 were retained.
- **Human Development Indicators (HDI, Life Expectancy, GII)**
  HDI-related indicators were extracted and reshaped into a country–year panel format. Variable names were standardized to maintain consistency across datasets.

Handling Missing Values

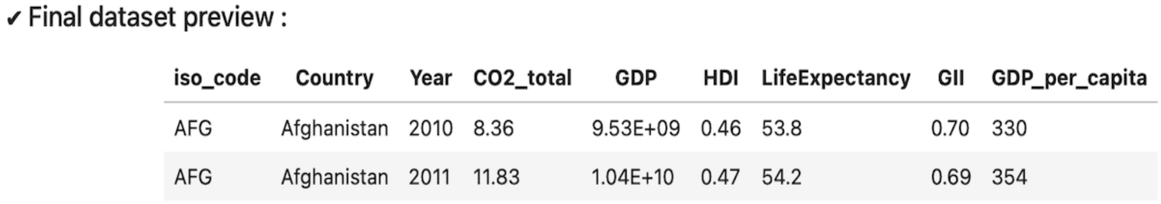
Due to differences in reporting coverage and frequency across countries, missing values—particularly for **HDI**, **Gender Inequality Index (GII)**, and **CO_2 emissions**—were addressed using **linear interpolation** over the 2010–2019 period. This approach assumes gradual changes over time and improves dataset completeness while preserving temporal consistency.

Dataset Merging and Final Selection

All cleaned datasets were merged using **ISO3 country codes and year identifiers**. After merging, observations with missing values in critical variables (CO_2 emissions, GDP, HDI, and Life Expectancy) were removed to ensure the reliability of subsequent analyses.

A unified panel dataset was constructed, from which a **clean 2019 cross-sectional dataset** was extracted for exploratory analysis, hypothesis testing, visualization, and machine learning applications. The final dataset was saved as **master_cross_section.csv** and used consistently throughout the project.

All data processing, cleaning, and merging steps were conducted programmatically using **Python**, ensuring reproducibility and transparency. The datasets were manually downloaded from official online sources, including **Kaggle** and the **World Bank**, and no web scraping or automated data collection was performed.



✔ Final dataset preview :

iso_code	Country	Year	CO2_total	GDP	HDI	LifeExpectancy	GII	GDP_per_capita
AFG	Afghanistan	2010	8.36	9.53E+09	0.46	53.8	0.70	330
AFG	Afghanistan	2011	11.83	1.04E+10	0.47	54.2	0.69	354

3. Data Analysis

The analysis was conducted in multiple stages to ensure transparency and methodological consistency. First, the raw datasets were cleaned and standardized. Second, exploratory data analysis was used to identify patterns and potential relationships among variables. Finally, statistical testing was applied to formally evaluate the proposed hypotheses.

3.1 Methodology

The analysis follows a structured data science pipeline consisting of data cleaning, exploratory data analysis, statistical testing, and visualization. The focus is on identifying relationships and patterns between economic, environmental, and social indicators across countries. To ensure robustness, the main hypothesis tests were conducted using the **2019 cross-sectional dataset**, where complete observations were available for all key variables.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to gain an initial understanding of the data and uncover potential patterns:

- Descriptive statistics (mean, median, variance, and missing values) were generated for **CO_2, GDP, HDI, Life Expectancy, and GII**.
- Variable distributions and pairwise relationships were examined using a **seaborn pairplot**, with kernel density estimates on the diagonal and scatter plots off-diagonal.
- A **correlation heatmap** was created to visualize relationships among the main variables.
- Country-ranked bar charts were produced for the **top 20 CO_2 emitters** and **top 20 countries by HDI**.
- Regional comparisons were conducted using **continent-level boxplots** for CO_2, GDP, HDI, Life Expectancy, and GII.
- **World choropleth maps** were generated using Plotly to visualize global patterns in CO_2 emissions, HDI, GDP, and GII.

3.3 Statistical Testing

Both Pearson and Spearman correlation tests were employed to capture linear as well as monotonic relationships and to reduce sensitivity to non-normality and outliers. To formally test the relationships between variables, both **Pearson** and **Spearman correlation analyses** were applied. These methods were chosen to capture both linear relationships and monotonic associations.

All hypotheses were evaluated using the **clean 2019 cross-sectional dataset**, ensuring that results were not driven by missing data. In interpreting the results, **statistical significance**, **effect size**, and **expected directional relationships** were taken into account.

3.4 Hypotheses

#	Hypothesis	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
1	**CO_2–GDP Relationship**	CO_2 and GDP are not positively related.	Higher GDP is associated with higher CO_2 emissions.
2	**GDP–Life Expectancy Relationship**	GDP has no relationship with Life Expectancy.	Higher GDP increases Life Expectancy.
3	**GDP–Gender Inequality Relationship**	GDP has no effect on GII.	Higher GDP reduces GII.
4	**CO_2–HDI Relationship**	CO_2 emissions have no effect on HDI.	Higher CO_2 emissions reduce HDI.

4. Findings

4.1 Expected Findings

Before conducting the analysis, the following relationships were expected based on existing literature and theoretical reasoning:

- A clear **positive relationship** between **GDP and CO_2 emissions**, as economic activity is generally associated with higher energy consumption.
- A strong **positive relationship** between **GDP and Life Expectancy**, reflecting better healthcare access and living standards in wealthier countries.
- **Lower gender inequality (GII)** in countries with higher GDP levels, as economic development is often linked to social progress.
- A **negative relationship** between **CO_2 emissions and HDI**, under the assumption that environmental degradation may reduce overall human development.

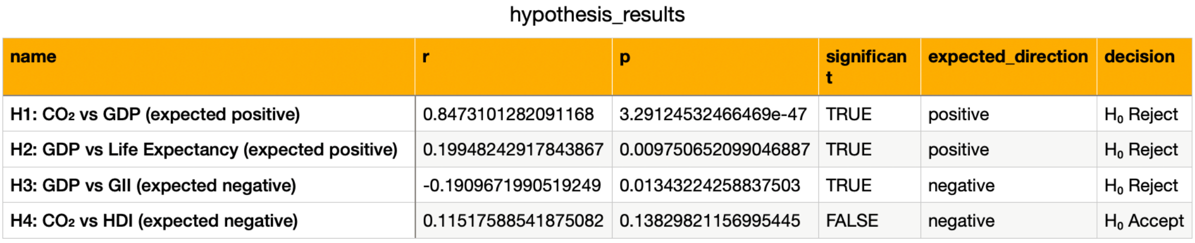
These expectations provided a benchmark against which the empirical results were evaluated.

4.2 Hypothesis Test Results

The hypotheses were tested using both **Pearson** and **Spearman correlation coefficients** on the clean **2019 cross-sectional dataset**. The results indicate that the expectations were **only partially supported** by the data.

- **H1 (CO_2–GDP)** shows a strong and statistically significant positive relationship. Countries with higher GDP levels tend to emit substantially more CO_2.
- **H2 (GDP–Life Expectancy)** is supported, although the strength of the relationship is moderate. Higher GDP is associated with longer life expectancy, but the effect varies across countries.
- **H3 (GDP–GII)** is supported, indicating that higher-income countries generally exhibit lower gender inequality.
- **H4 (CO_2–HDI)** is not supported. Contrary to expectations, CO_2 emissions do not decrease as HDI increases. Many high-HDI countries also display high levels of CO_2 emissions.

Overall, the results suggest that **economic and social development often progress together**, while **environmental sustainability does not necessarily improve alongside human development**.



hypothesis_results

name	r	p	significant	expected_direction	decision
H1: CO_2 vs GDP (expected positive)	0.8473101282091168	3.29124532466469e-47	TRUE	positive	H_0 Reject
H2: GDP vs Life Expectancy (expected positive)	0.19948242917843867	0.009750652099046887	TRUE	positive	H_0 Reject
H3: GDP vs GII (expected negative)	-0.1909671990519249	0.01343224258837503	TRUE	negative	H_0 Reject
H4: CO_2 vs HDI (expected negative)	0.11517588541875082	0.13829821156995445	FALSE	negative	H_0 Accept

4.3 Summary Interpretation

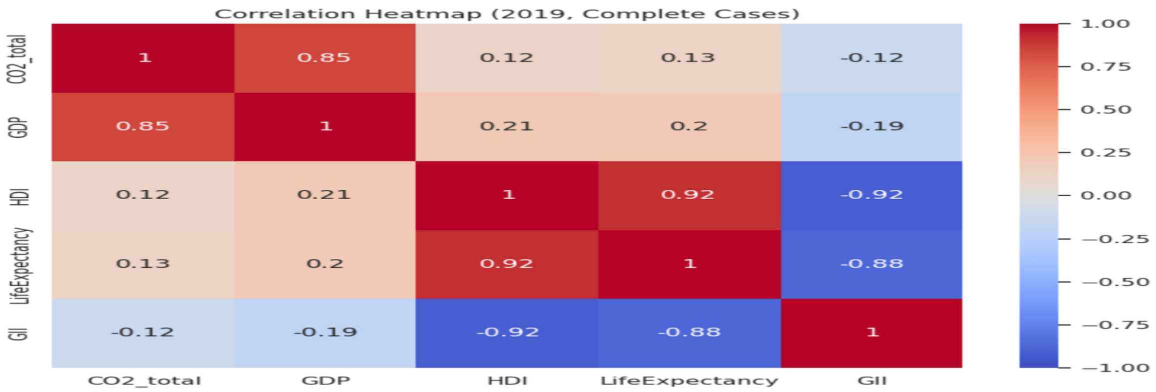
Taken together, the findings indicate that **GDP plays a central role** in shaping both social outcomes and environmental pressure. While higher GDP is associated with better life expectancy and lower gender inequality, it is also strongly linked to increased CO_2 emissions.

The absence of a negative relationship between CO_2 emissions and HDI highlights a key insight of the study: historically, **high human development has been achieved at the cost of higher environmental impact**. This raises concerns about the long-term sustainability of prevailing development models.

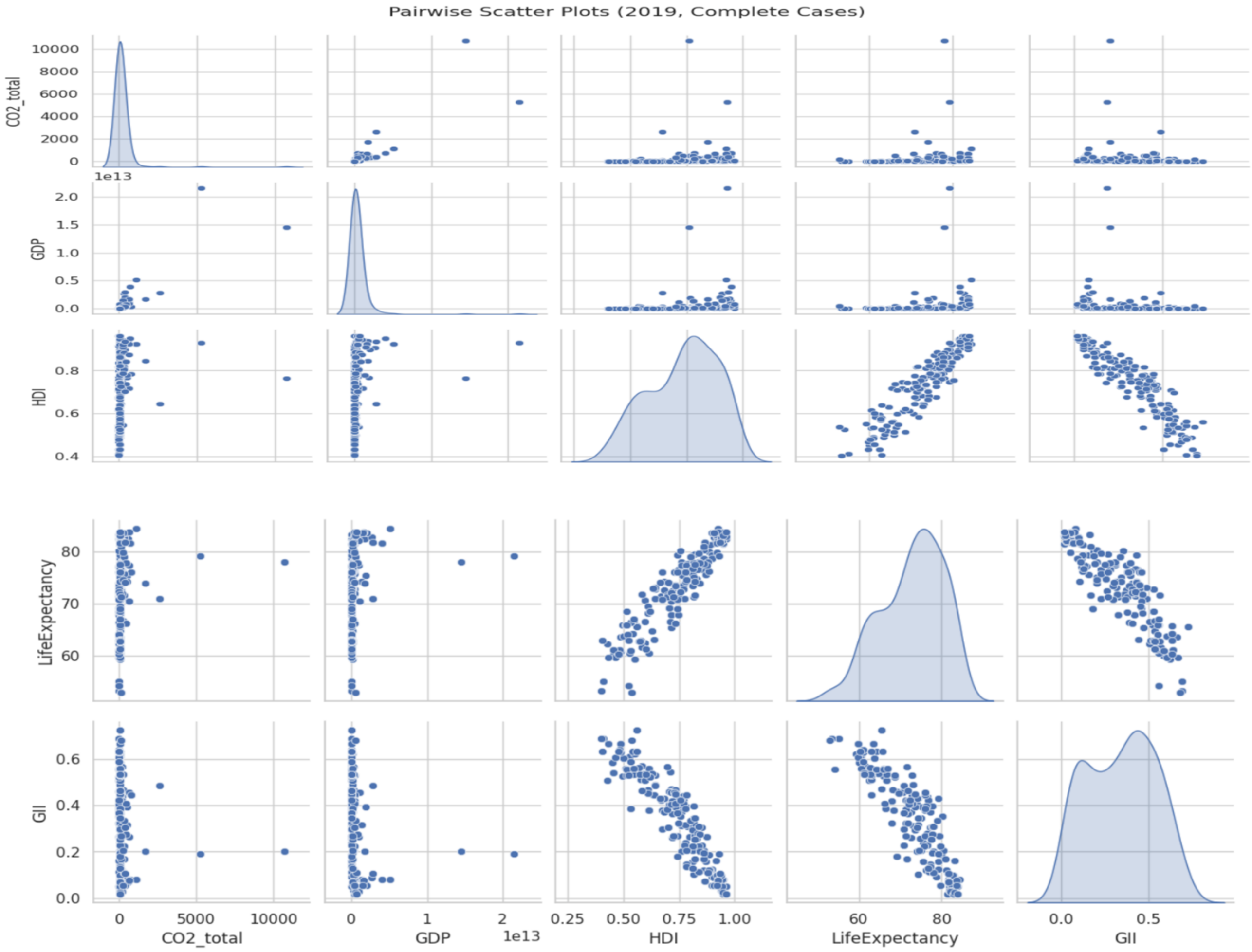
4.4 Visualizations Used

The quantitative findings are supported by a range of visualizations designed to improve interpretability:

- **Correlation heatmap** for CO₂, GDP, HDI, Life Expectancy, and GII

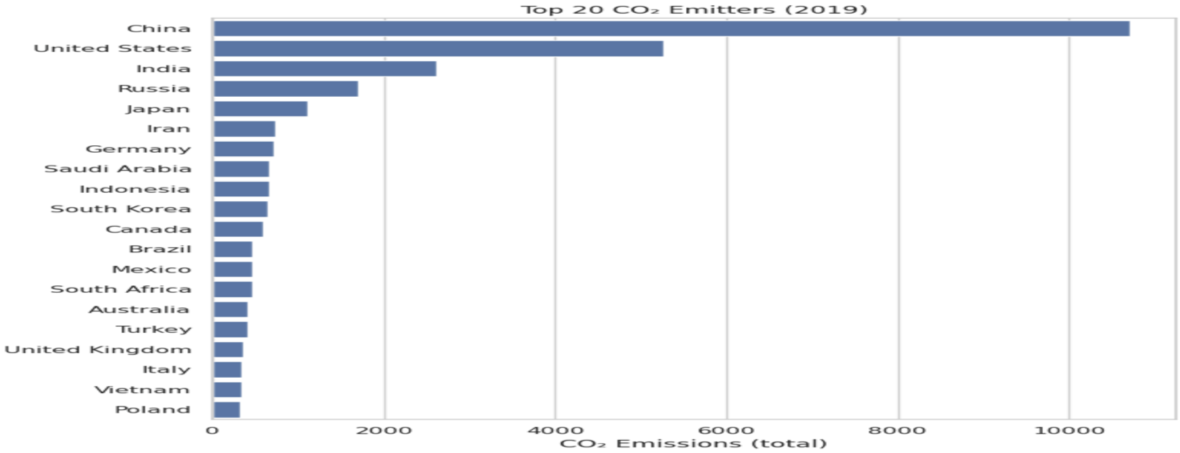


- **Pairwise scatter plots (pairplot)** showing variable relationships and distributions

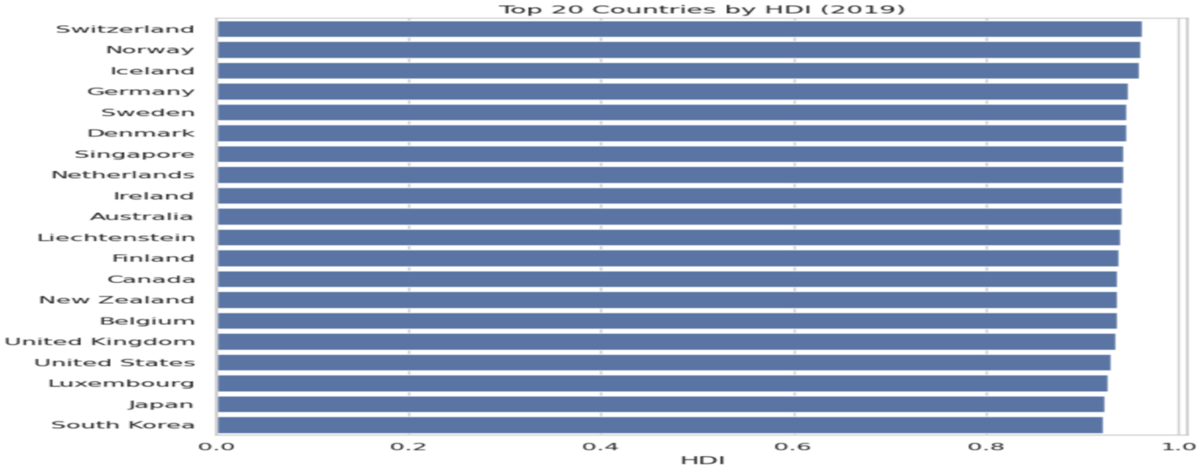


- **Top 20 bar charts**:

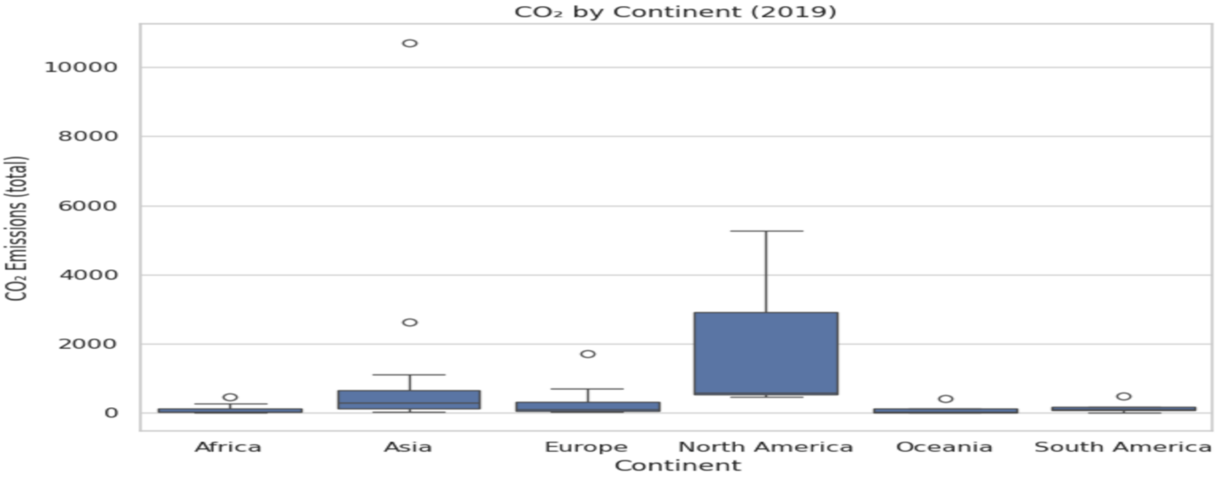
Countries with the highest total CO₂ emission:

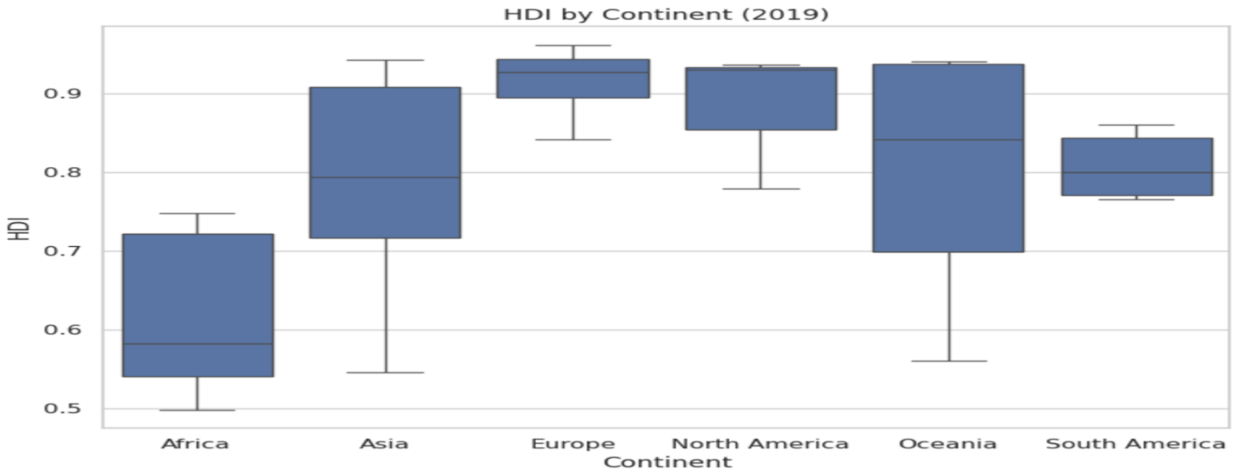


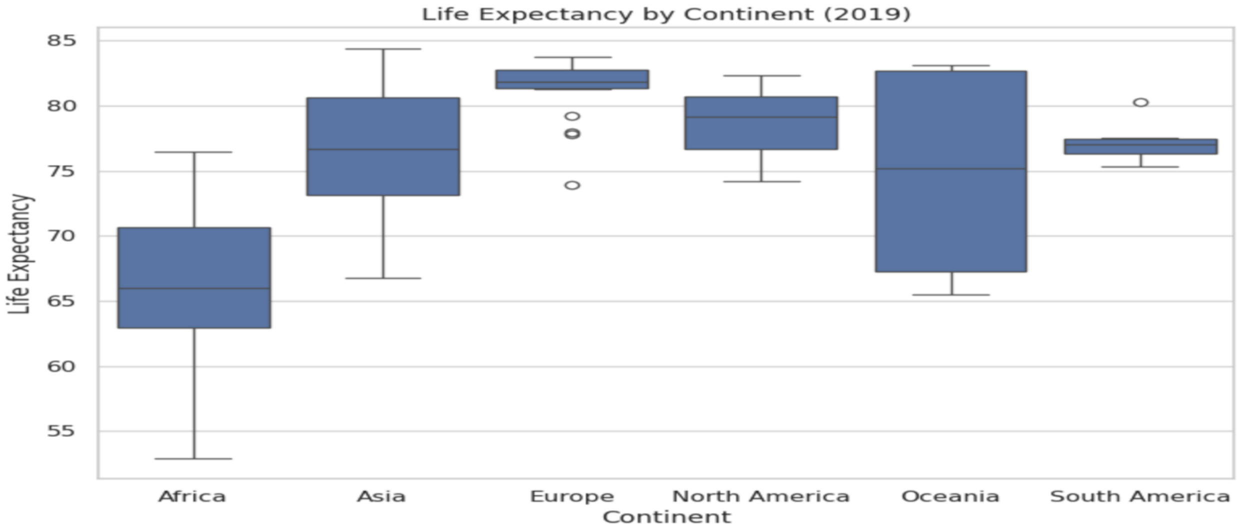
Countries with the highest HDI values:

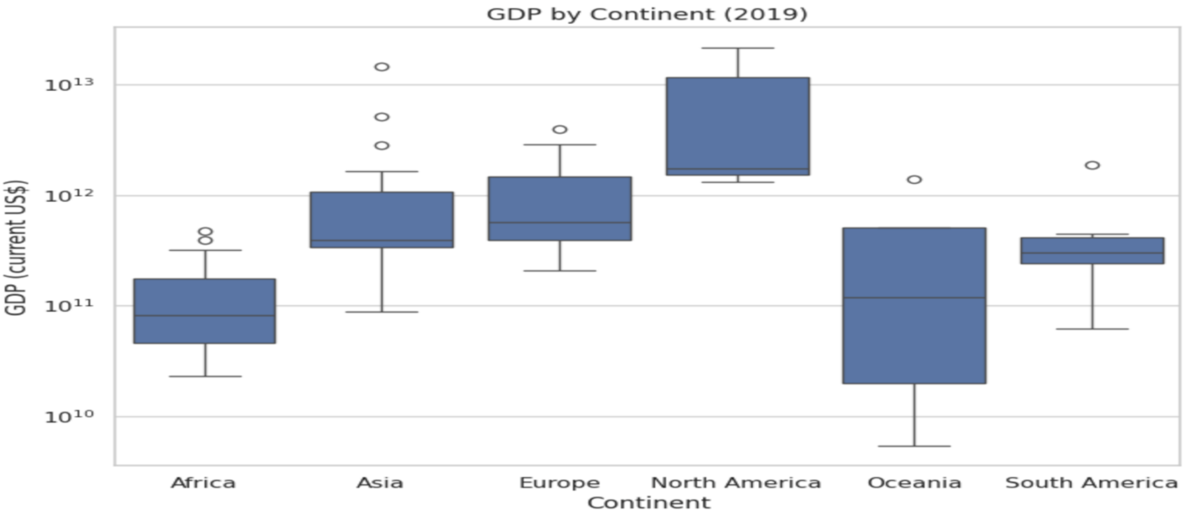


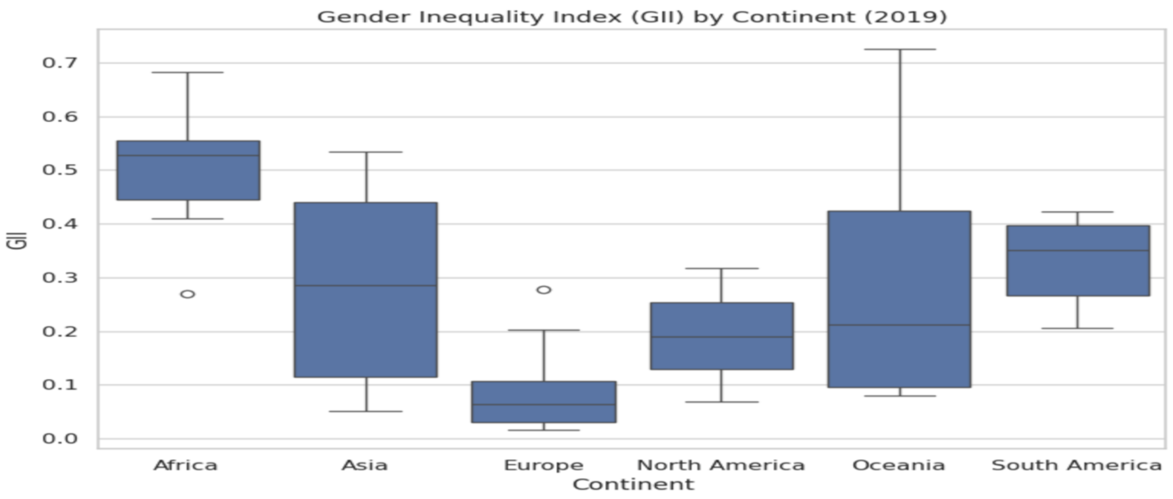
- **Continent-level boxplots** for CO₂, GDP, HDI, Life Expectancy, and GII



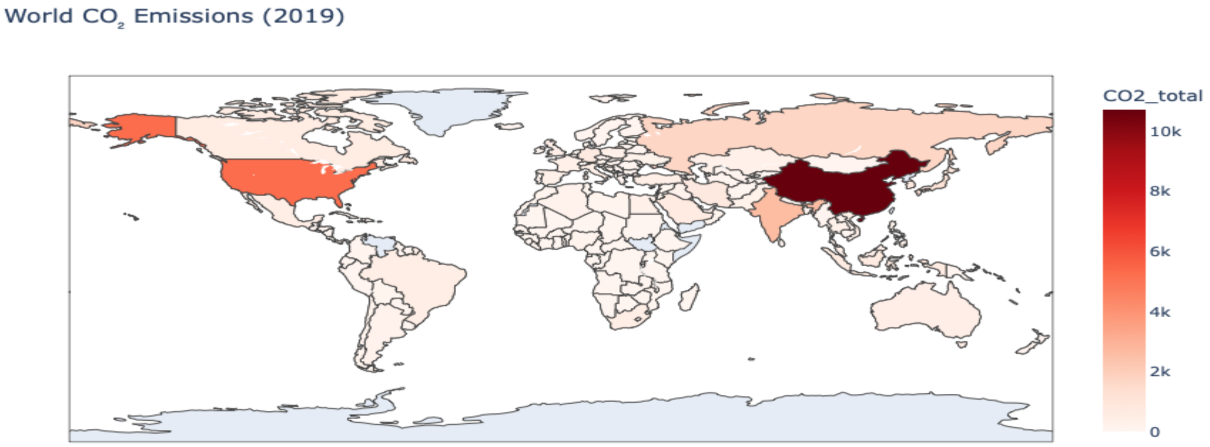

HDI by Continent (2019)
HDI
0.5
0.6
0.7
0.8
0.9
Africa
Asia
Europe
North America
Oceania
South America
Continent

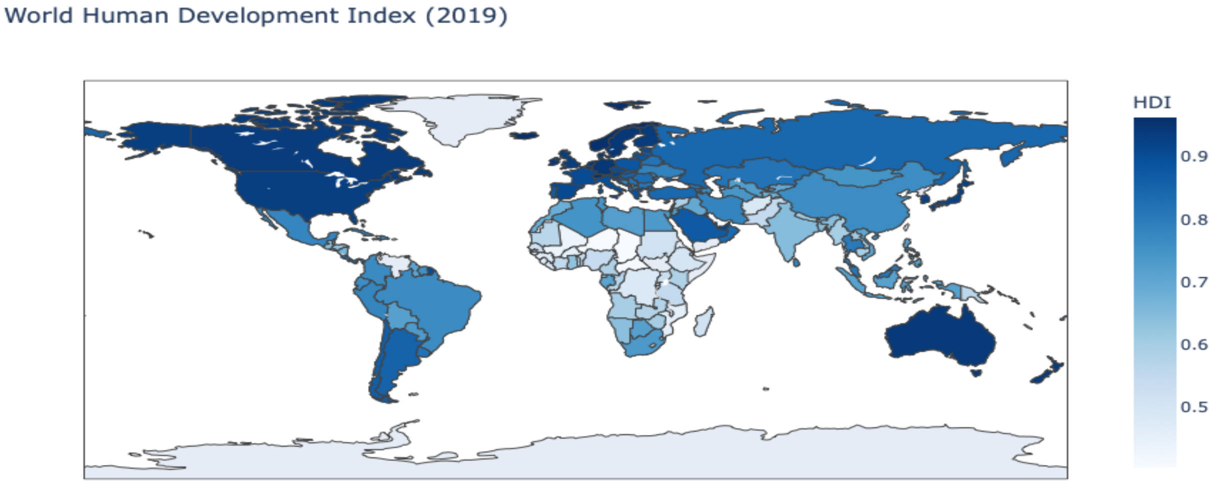

Life Expectancy by Continent (2019)
Life Expectancy
55
60
65
70
75
80
85
Africa
Asia
Europe
North America
Oceania
South America
Continent


GDP by Continent (2019)
GDP (current US$)
10^{10}
10^{11}
10^{12}
10^{13}
Africa
Asia
Europe
North America
Oceania
South America
Continent

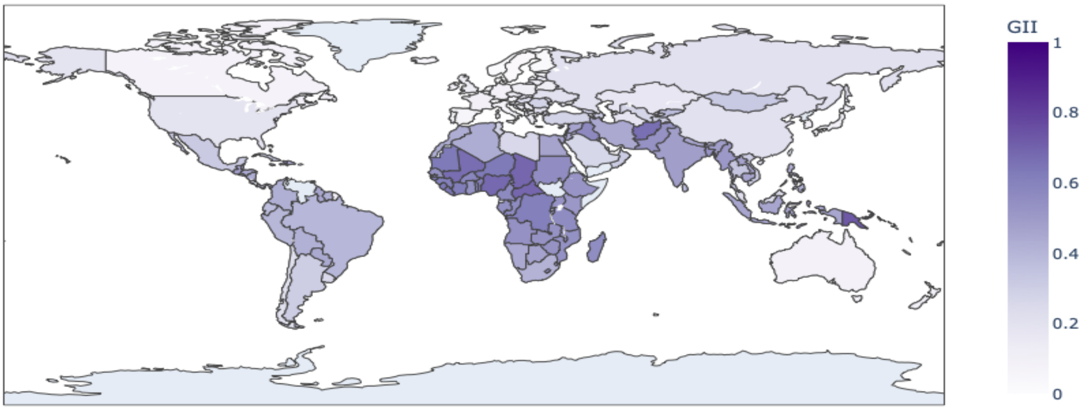


- **World choropleth maps** illustrating the global distribution of CO₂ emissions, HDI, GDP, Life Expectancy, and GII (2019)

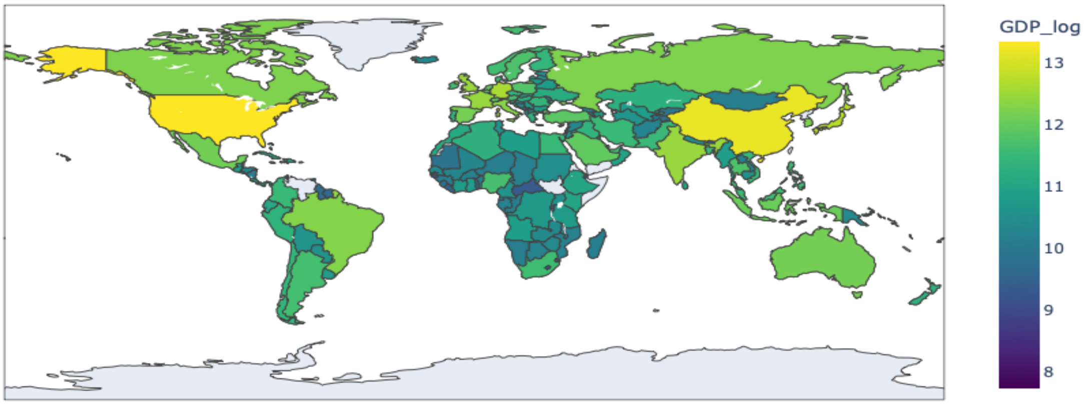




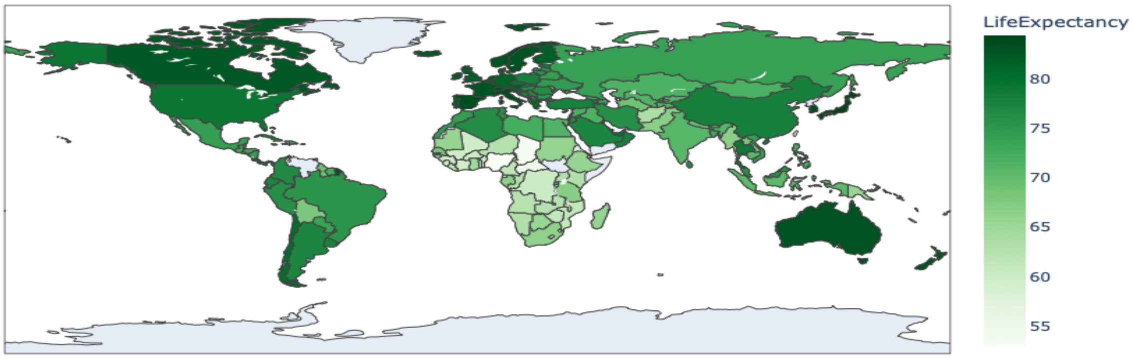
World Gender Inequality Index (2019)



World GDP (log10 scale, 2019)



World Life Expectancy (2019)



4.5 Machine Learning Analysis

In addition to correlation-based hypothesis testing, **supervised machine learning methods** were employed to examine whether economic, environmental, and social indicators can be used to predict **life expectancy** across countries. This approach allows for a more flexible, data-driven evaluation of development outcomes compared to purely correlation-based analysis.

The machine learning analysis was conducted using the **clean 2019 cross-sectional dataset**, ensuring consistency with the earlier hypothesis testing framework. Focusing on a single year eliminates time-related effects and allows the models to capture structural cross-country differences in development.

Target Variable

Life Expectancy was selected as the target variable because it is a comprehensive and widely accepted indicator of human well-being. It reflects health outcomes, living standards, and overall quality of life, making it a suitable outcome variable for assessing human development beyond purely economic measures.

Predictor Variables

The following variables were used as predictors in the models:

- Gross Domestic Product (GDP)
- CO_2 emissions (total)
- Human Development Index (HDI)
- Gender Inequality Index (GII)

Together, these predictors represent different dimensions of development and allow the models to assess how economic progress, social development, and environmental pressure are associated with life expectancy outcomes.

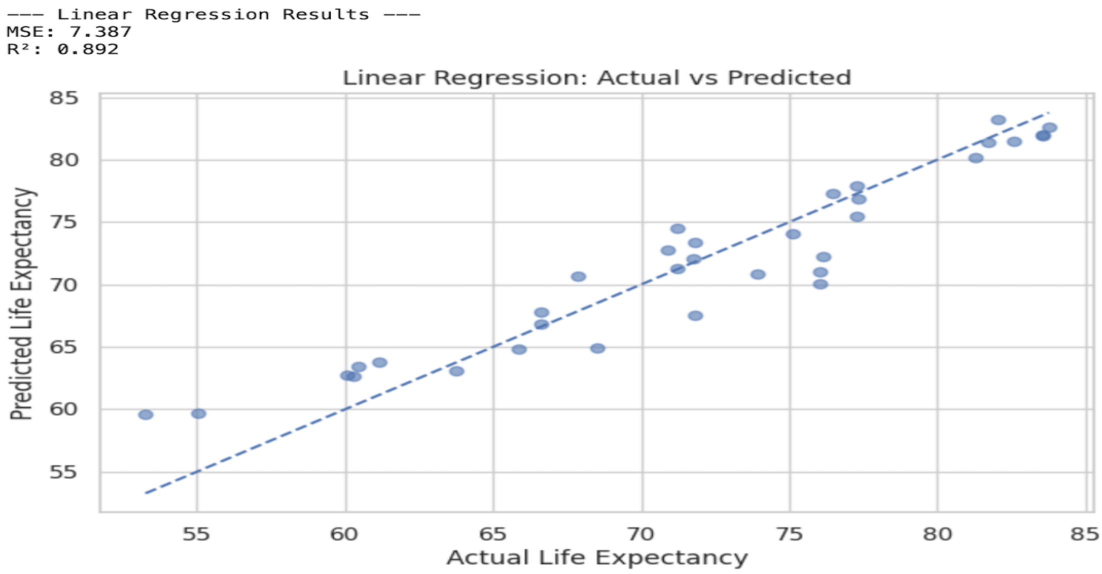
Train–Test Split

The dataset was divided into **training and testing subsets** using an **80%–20% split**. This approach enables model performance to be evaluated on unseen data and reduces the risk of overfitting. Using a single-year cross-sectional dataset avoids temporal dependence and ensures consistency with the earlier hypothesis testing framework.

Linear Regression (Baseline Model)

Linear Regression was employed as a **baseline model** to establish a clear and interpretable reference for predicting life expectancy. By assuming a linear relationship between the predictors and the target variable, this model allows for a transparent assessment of how development indicators are associated with life expectancy on average.

The Linear Regression model demonstrated **strong overall explanatory power**, achieving a high R^2 score. This suggests that a substantial portion of the variation in life expectancy across countries can be explained through linear combinations of economic and social development indicators.

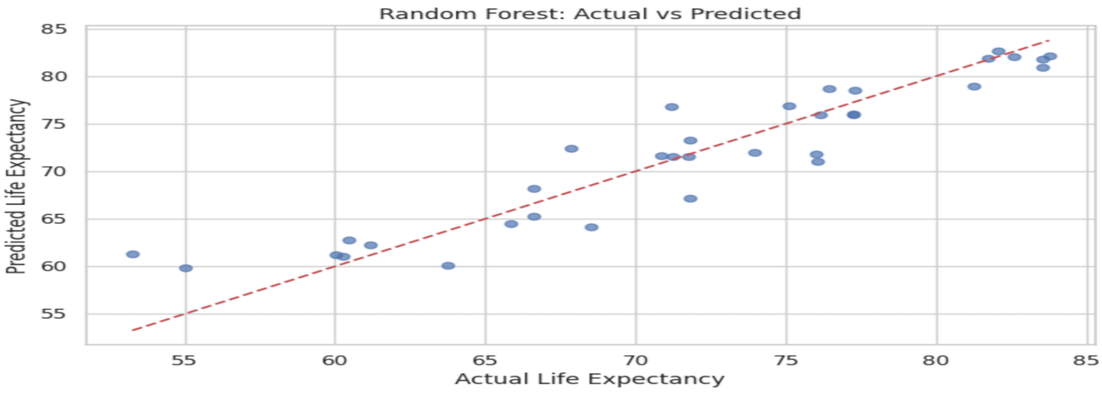


```
--- Linear Regression Results ---
MSE: 7.387
R²: 0.892
```

Random Forest Regressor

To capture potential **non-linear relationships and interactions** among the predictors, a Random Forest Regressor was implemented as a complementary model. Random Forest aggregates predictions from multiple decision trees, improving robustness and predictive performance.

The Random Forest model achieved performance comparable to the Linear Regression model, indicating that while non-linear effects exist, much of the explanatory structure in the data can already be captured by simpler linear relationships.



Model Performance Comparison

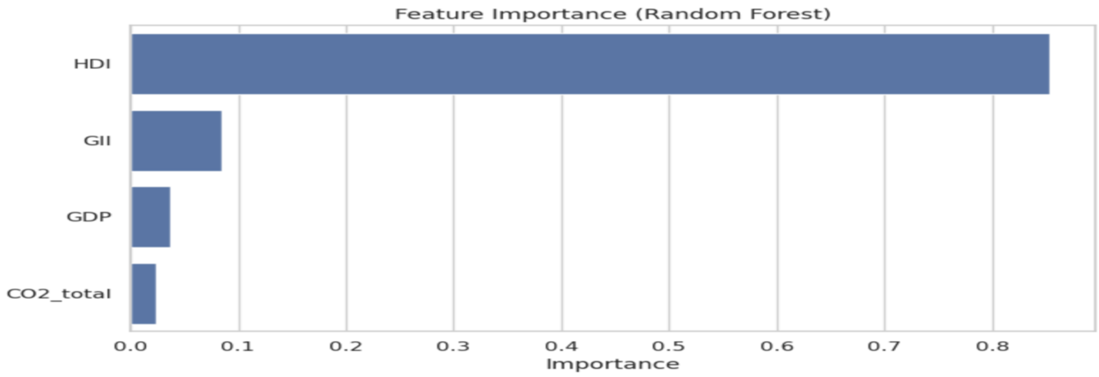
Both models performed well in predicting life expectancy. The Linear Regression model achieved a slightly higher R² score, while the Random Forest model provided additional insights into non-linear patterns and variable importance.

Model	MSE	R^2 Score
Linear Regression	7.387	0.892
Random Forest Regressor	8.515	0.876

Feature Importance Analysis

Feature importance analysis from the Random Forest model reveals that **HDI is the most influential predictor** of life expectancy, followed by **GDP**. In contrast, **CO_2 emissions** and **Gender Inequality Index (GII)** contribute more modestly to predictive performance.

This result reinforces earlier findings from the correlation analysis and highlights the dominant role of **social and human development factors** in explaining differences in life expectancy across countries.



5. Limitations and Future Work

Despite providing meaningful insights, this study has several limitations that should be considered when interpreting the results.

First, the analysis is based on a **cross-sectional dataset for the year 2019**, which restricts the ability to examine dynamic relationships or causal effects over time. While correlation and regression analyses reveal important associations, they do not allow for conclusions about causality or long-term development trajectories.

Second, missing data were addressed using **linear interpolation across the 2010–2019 period**. Although this approach improves data completeness and consistency, it assumes smooth changes over time and may introduce approximation errors, particularly for countries with irregular or sparse reporting.

Third, the study relies on **aggregate country-level indicators**, which may conceal within-country inequalities and regional disparities. For example, high GDP or HDI values do not necessarily reflect equal access to healthcare, education, or economic opportunities across different population groups.

Additionally, CO_2 emissions are measured as **total national emissions**, which does not account for population size, consumption-based emissions, or differences in production structures. This may bias comparisons between large and small countries.

Finally, the machine learning analysis was limited to **basic regression models** and a small set of predictors. While the results are informative, more advanced models or a richer feature set could provide additional insights.

5.1 Future Work

Several extensions could improve and expand this analysis in future research.

First, incorporating a **panel data framework** would allow for the examination of changes over time and enable the use of fixed-effects or random-effects models. This would provide a stronger basis for understanding causal relationships between economic growth, environmental impact, and human development.

Second, future studies could include additional indicators such as **renewable energy usage, education quality, healthcare expenditure, environmental policy measures, or income inequality metrics**. These variables could help explain why countries with similar GDP levels exhibit different development and sustainability outcomes.

Third, alternative environmental measures—such as **CO_2 emissions per capita or consumption-based emissions**—could offer a more nuanced perspective on environmental responsibility and sustainability.

From a modeling perspective, future work could explore **more advanced machine learning techniques**, such as gradient boosting or regularized regression methods, as well as cross-validation strategies to improve generalizability.

Finally, conducting **region-specific or income-group analyses** may help uncover structural differences between countries and provide more targeted policy-relevant insights.

6. Conclusion

This project examined the relationships between **economic growth, environmental impact, and human development** through a cross-country data science framework. By integrating economic, social, and environmental indicators, the study aimed to assess whether higher economic performance consistently translates into improved human development outcomes, or whether trade-offs emerge—particularly in terms of sustainability and social equality.

The findings indicate that **economic growth plays a central but not sufficient role** in human development. While higher GDP levels are associated with longer life expectancy and lower gender inequality, they are also strongly linked to higher CO_2 emissions. This highlights a persistent global pattern in which economic and social progress often coincides with increased environmental pressure. Contrary to initial expectations, higher human development levels (HDI) do not correspond to lower CO_2 emissions, suggesting that historically, many countries have achieved high development outcomes at the cost of environmental sustainability.

The machine learning analysis complements the statistical findings by providing a predictive perspective. Both Linear Regression and Random Forest models demonstrate that **life expectancy is primarily driven by social and human development factors**, particularly HDI, rather than by environmental indicators alone. GDP and gender inequality play secondary roles, while CO_2 emissions exhibit a limited direct association with life expectancy in a cross-sectional context. These results reinforce the robustness of the earlier correlation-based analysis and underline the dominant importance of social investment in explaining cross-country differences in human well-being.

Overall, the project reveals a fundamental tension between **economic growth, social progress, and environmental sustainability**. The results suggest that improving human development outcomes requires more than economic expansion alone and that future development strategies must prioritize inclusive social policies while addressing long-term environmental challenges. From a methodological perspective, this project demonstrates the value of combining exploratory data analysis, statistical testing, and machine learning techniques to gain a comprehensive understanding of complex global development patterns.