University of Exeter

Mini Business Analytics Report:

Progress and Challenges in Achieving SDG Education Targets in the UK

Presented by

Chayutpong Prateepavanitch

No. 740003557

Table of Contents

My GitHub Repository Link: https://github.com/Chayutpong-p/SDG4_Mini_Business_Analytics

Dataset and every other output are also provided in my Repository.

Table of Figures

1. Introduction

1.1 Project Description

This project assesses how the UK is on track towards meeting Sustainable Development Goal (SDG) 4, with a focus on data-driven analysis of educational policies and addressing systemic inequalities. It studies key metrics such as school enrollment rates and government spending on education to identify critical gaps and challenges in the educational system. The results are reported with actionable recommendations aimed at improving access to education and enhancing educational quality. By leveraging business analytics techniques, the findings offer policymakers concrete, evidence-based strategies to scale successful initiatives, close education gaps, and align the UK's education system with global sustainability goals in the long term.

1.2 Aims and Objectives

This report compares how well the UK is in progress to achieve SDG 4 (Quality Education) using objective educational results and analyzing gaps in equity and access. This will include assessing outcome indicators (like enrollment and completion) and pinpointing important influencing factors (like government investment and socioeconomic differences). The results will be followed by practical recommendations for closing these gaps and making access to high-quality education more accessible. It will provide evidence-based policy and programme recommendations that will make a difference to advancing the UK towards SDG 4 with long-term improvements to education.

Key Project Objectives:

No.	Objective	Description
1	Assess SDG 4 Progress	Analyze education indicators' trend, with the focus on equitable access and quality education.
2	Identify Influencing Factors	Look at variables like government spending and socioeconomic inequality and look at their effect on education.
3	Provide Recommendations	Develop specific, evidence-based recommendations for policy adjustments and resource allocation that aim to reduce educational disparities.

Figure 1: Key Objectives Description

1.3 Research Questions

To achieve these objectives, the project addresses the following research questions:

No.	Research Questions
1	How has the UK progressed toward SDG 4 education targets, focusing on access, quality, and equity?
2	How do key factors, such as government expenditure and income inequality impact education performance metrics?
3	What data-driven strategies and policy interventions can optimize the UK's achievement of SDG 4?

Figure 2: Research Questions

1.4 Relevance of the Project

This project supports UNESCO's SDG 4 monitoring framework by providing data-driven insights for addressing disparities and improving education equity and quality.

2. Data Access, Ethics, Security, and Privacy Considerations

2.1 Data Source

The information that went into this project came from open, reliable datasets that are compatible with the SDG 4 targets of measuring education progress. Key sources include:

- Global Database on Intergenerational Mobility (World Bank, n.d.) for understanding long-term impacts of education policies.
- UNESCO SDG 4 Data Hub: (UNESCO Institute for Statistics, n.d.-a) Offers global, regional, and country-level education indicators.
- UNESCO Data Browser: (UNESCO Institute for Statistics, n.d.-b) Provides comprehensive access to historical and comparative education data for monitoring SDG 4.

All datasets chosen are time-series and cross-sectional based which allows me to study the trend and relations in detail.

2.2 Privacy, Security, and Ethical Considerations

Although the datasets used in this project are publicly available, ethical and privacy considerations remain a top priority in handling the data. The following measures are taken to ensure compliance with relevant laws and ethical standards:

1. **Data Privacy**: No identifiable data are stored in the datasets used, therefore fully compatible with the General Data Protection Regulation (GDPR) and other privacy legislation. In concentrating on aggregated only information such as enrollment, completion and government spending, the probability of re-identification is minimal and privacy at the highest level is assured (Podda, 2021).
2. **Ethical Use of Data**: Any data use is according to the terms and conditions of the data source and hence it's always open for the use of data. This work maintains strict reference to sources and method to guarantee data is being handled ethically and held responsible during analysis (DataCamp, 2024).
3. **Avoiding Misrepresentation**: The interpretation of the data is contextualized so that it makes sense. Visualizations and reports are designed with a focus on objectivity and simplicity, so the data is presented honestly and without prejudice (Cipan, 2023).

2.3 Reliability, Integrity, and Validity

To ensure the reliability and integrity of the analysis, the following measures were implemented:

1. **Data Cleaning**: Missing data points were handled using methods like linear interpolation and mean imputation, depending on the extent of missingness, while indicators with over 50% missing data were excluded (Ying et al., 2024).
2. **Cross-Verification**: Data consistency was verified across sources to ensure alignment with established benchmarks and trends (Research Method, 2023).
3. **Validation**: Results were validated through triangulation across datasets to ensure robustness and relevance (Bhandari, 2022).

2.4 Data Structure and Transformation Decisions

The data set in this project was built by combining multiple data files to create a single Excel file using the Excel tool so it would be uniform to analyse. This gave one sheet with 1,552 rows and 7 columns that shows us education related metrics for every time interval. The dataset is for several years and suitable for time-series analyses and temporal comparisons, but with a particular interest in UK education trends.

Field	Description
indicatorId	Unique identifiers for each indicator, representing specific education metrics.
indicator_name	Descriptive names of the indicators.
geoUnit	Geographical unit identifiers, indicating countries.
year	The year corresponds to the recorded data point.
value	Numeric values of indicators, the quantitative backbone of the dataset.
qualifier	Qualitative data providing additional context.
magnitude	Details about the scale of certain indicators.

Figure 3: Field Indicators Description

It has many columns but need to clean it up to make some analysis work. Not achieving the irrelevant or unfinished parts could degrade the insights and bias the findings.

To enhance analytical precision, I will focus on the following key columns:

Field	Purpose
indicator_name	To identify and analyze the specific educational metrics under consideration.
year	To facilitate trend analysis and observe temporal changes in the metrics.
value	To provide the quantitative basis for assessing education outcomes.

Figure 4: Focus Field Indicators

Some columns such as **qualifier, magnitude, geoUnit** aren't used because of data gaps or redundant values within this project's focuses on the United Kingdom. This option provides a tradeoff between data and analysis detail, while keeping the analysis specific and pragmatic.

3. Data Preparation and Transformation

As a part of getting the data ready for analysis, I had to verify and record missing values in order not to bias or contaminate the data. "Na" values in Value column, the missing data i.e 1,038 data. Missing numbers per indicator were determined and were ranked as **Low** (<5%), **Moderate** (5–50%), or **High** (>50%).

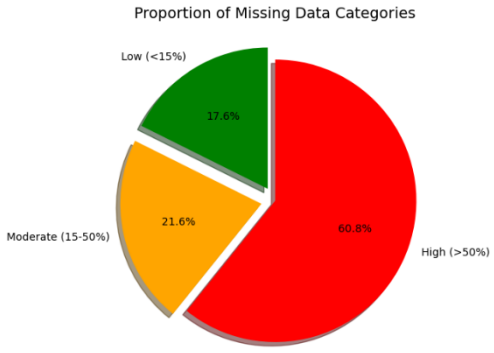


Figure 5: Proportion of Missing Data Categories

Categorizing and Handling Missing Values

I had a specific approach in place for each category to make sure the data was accurate, and the analysis was valid.

Missing Data Percentage	Method	Reason
<5%	Linear Interpolation	Small missing data makes no difference in trends. Linear interpolation maintains continuity using neighboring points , which is great for constant or categorical data.
5–50%	Median Imputation	Small to medium missing values allow imputation. Median imputation is good since it will correct for outliers.
>50%	Indicator Removal	Signs with more than 50% missing data are not trustworthy, imputation carries great bias. Remove them, and the analysis becomes more reliable.

Figure 6: Method of Handling Missing Values

4. Exploratory Data Analysis (EDA)

4.1 Data Understanding

To make sure the dataset is transparent and useful for the analysis goals, I organised into group indicators, types and contextual interpretations.

Categorizing Indicators

The cleaned dataset has **40 indicators** which I have divided into the following **7 categories** for convenience.

Classification by Type

For trend-finding purposes and research questions, indicators were further divided into two types:

1. **Outcome Indicators**: Measures of educational performance or achievements.
2. **Influencing Factors**: Factors that affect the outcomes.

Category	Type	Description
Completion Rates	Outcome Indicators	Metrics related to graduation rates at different education levels.
Enrollment Ratios	Outcome Indicators	Indicators reflecting student enrollment across various demographics.
Out-of-School Rates	Outcome Indicators	Metrics capturing the percentage of children and adolescents not in school.
Participation Rates	Outcome Indicators	Measures of involvement in educational activities or programs.
Government Expenditure	Influencing Factors	Data on financial investments in education.
Legal Guarantees	Influencing Factors	Indicators measuring policy commitments to education.
Socioeconomic Factors	Influencing Factors	Indicators related to income inequality, poverty, and economic status.

Figure 7: Data Categorization & Classification

Contextualizing Trends

That way, progress is evaluated in a way that accounts for the effect that each indicator is supposed to make. The Data Understanding Table describes trends as Positive, stagnant, or declining with "Positive When" column explanation. For instance:

- **Declining** "**Out-of-School Rates**" is **positive**, with fewer children out of school.
- **Increasing** "**Government Expenditure on Education**" is **positive**.
- **Mixed trends** in "**Labor Force Participation Rate**" are assessed based on demographic impacts.

Indicator	Type	Sub-category	Trend Categorization	Positive When
Completion rate, lower secondary education, both sexes (modelled data) (%)	Outcome	Completion Rates	Positive Progress	Increasing
Completion rate, primary education, both sexes (modelled data) (%)	Outcome	Completion Rates	Positive Progress	Increasing
Completion rate, upper secondary education, both sexes (modelled data) (%)	Outcome	Completion Rates	Positive Progress	Increasing
Gross enrolment ratio for tertiary education, both sexes (%)	Outcome	Enrollment Ratios	Positive Progress	Increasing
Proportion of 15- to 24-year-olds enrolled in vocational education, both sexes (%)	Outcome	Enrollment Ratios	Decline	Increasing
Out-of-school rate for adolescents of lower secondary school age, bo	Example	Out-of-School Rates	Decline	Decreasing
Out-of-school rate for children of primary school age, both sexes (%)		Out-of-School Rates	Decline	Decreasing
Out-of-school rate for youth of upper secondary school age, both sexes (%)	Outcome	Out-of-School Rates	Decline	Decreasing
Gini index	Influencing	Socioeconomic Factors	Decline	Decreasing
Income share held by highest 20%	Influencing	Socioeconomic Factors	Stagnation	Decreasing
Income share held by lowest 20%	Influencing	Socioeconomic Factors	Stagnation	Increasing
Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population)	Influencing	Socioeconomic Factors	Decline	Decreasing

Figure 8: Example Data from Data Understanding Table

4.2 Progress Trend Analysis

I examined trends in each indicator through visualisations to explore trend over time. The trends were categorized into three groups: **Positive Progress, Stagnation** and **Decline**.

Categorization of Trends by Sub-Category and Type

Beyond the general trend classification, I also sub-categories indicators to learn more about the dataset and to understand their impact on the trend. In this description we see the ways in which factors contribute or suppress it.

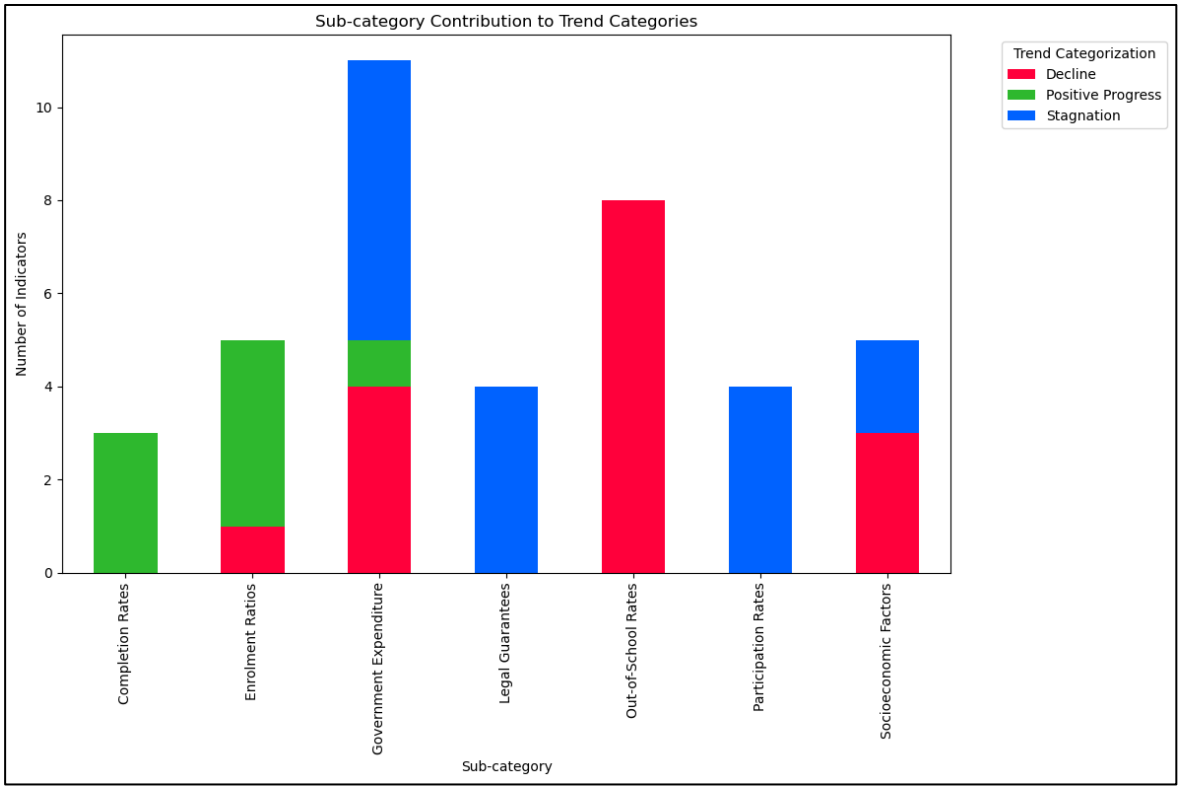


Figure 9: Sub-category Contribution to Trend Categories

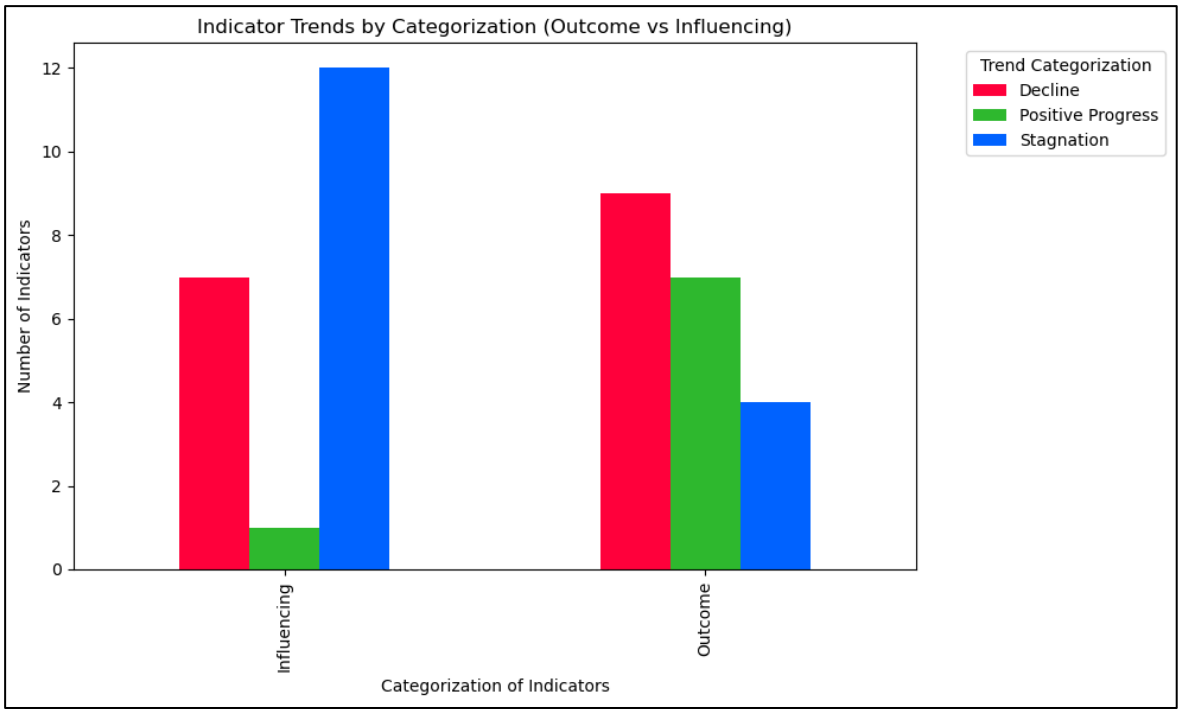


Figure 10: Indicator Trends by Categorization (Outcome VS Influencing)

Insights from Outcome Indicators:

Completion Rates:

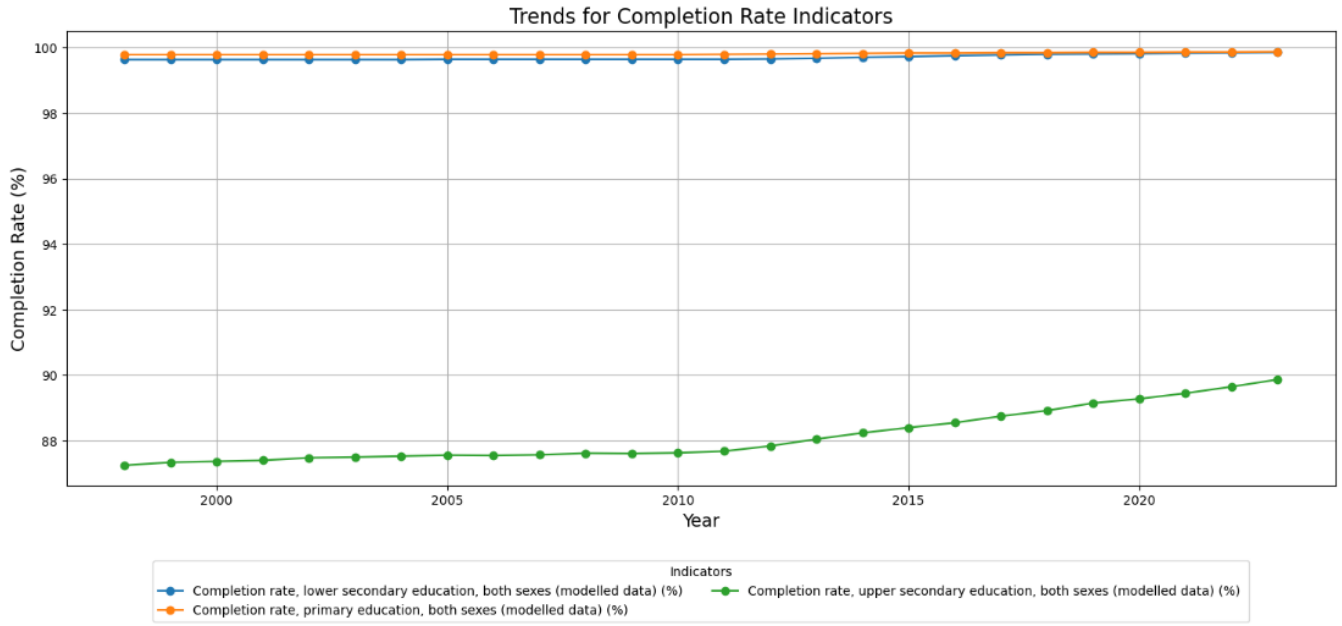


Figure 11: Trends for Completion Rates

- **Statistical Insights**:
    - **Primary and Lower Secondary Completion**: Exceptionally high (**mean ~99.7%**), showing strong retention.
    - **Upper Secondary Completion**: Lower mean (**88.12%, std 0.81%**), indicating room for improvement.
- **Key Findings:**
    - The secondary-to-tertiary transition is a critical bottleneck for the education system.
    - Supporting upper secondary completion could improve tertiary enrollment.

Enrollment Ratios:

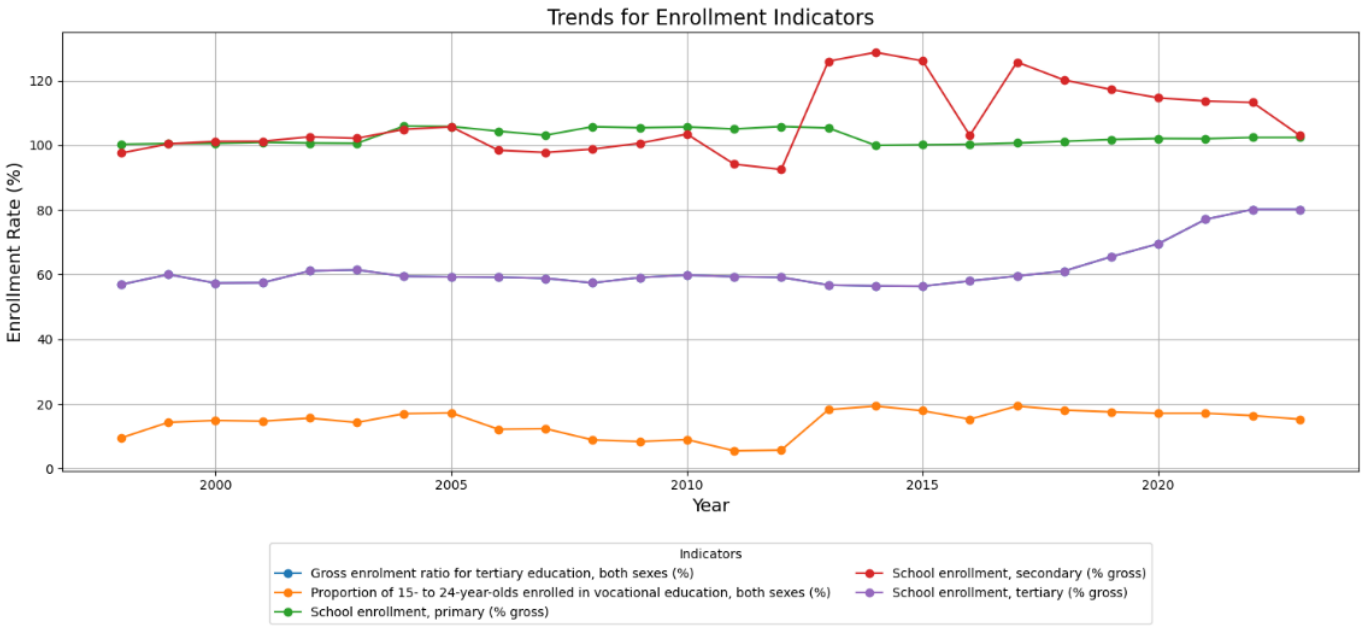


Figure 12: Trends for Enrollment Ratios

- **Statistical Insights**:
    - **Enrollment rates** above 100% in **primary** (**mean 102.65%**) and **secondary** (**mean 107.44%**) suggest repeat or overage enrollments.
    - **Tertiary enrollment** is lower (**mean 61.8%, std 6.98%**), indicating limited higher education access.
- **Key Findings:**
    - Foundational education is strong, but tertiary access requires improvement.
    - Over-enrollment at primary and secondary levels may reflect inefficiencies such as repetition and overage.

Out-of-School Rates:

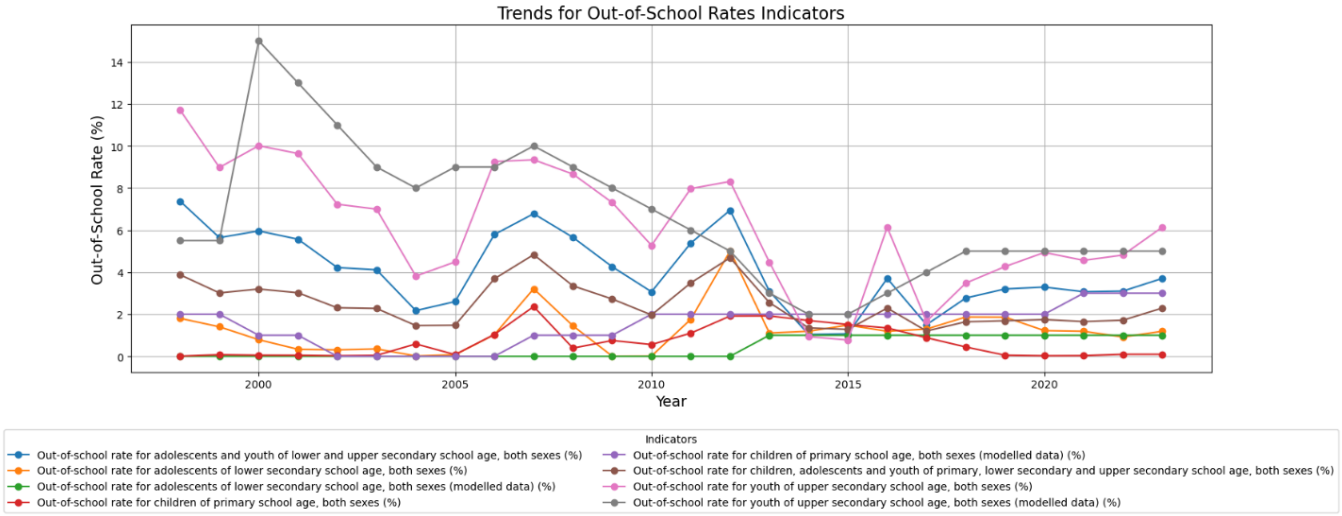


Figure 13: Trends for Out-of-School Rates

- **Statistical Insights**:
    - **Declining trends** show fewer children excluded from education systems, **indicating progress**.
    - **Upper secondary out-of-school rates** are highest (**mean 6.69%, std 3.26%**), highlighting persistent challenges.
    - **Primary out-of-school rates** are lowest (**mean 0.42%**), reflecting strong access and attendance.
    - **Lower secondary rates** are minimal (**mean 1.23%,** tight variation), suggesting effective policies.

Participation Rates:

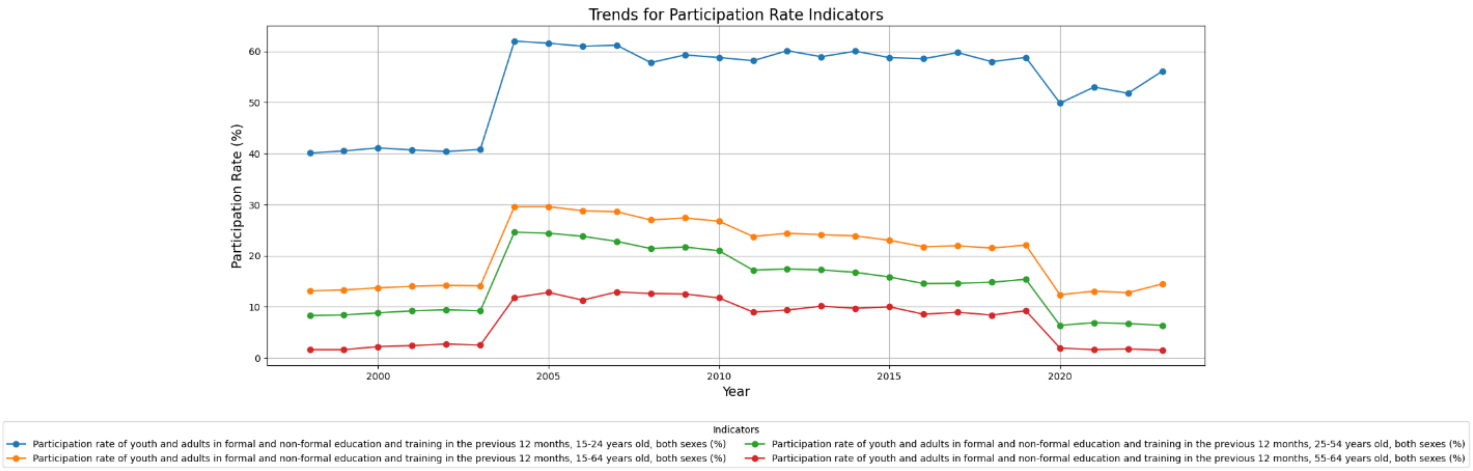


Figure 14: Trends for Participation Rates

- **Statistical Insights**:
  - **Youth participation (ages 15–24)** is highest (**mean 54.13%**), reflecting strong engagement.
  - Participation **decreases with age**:
    - Ages 25–54: Mean 20.74%.
    - Ages 55–64: Mean 14.73%.
    - Above 64: Mean 7.25%.
  - The **15–24 age group** shows the **largest variation** (std 8.06%), indicating inequalities or policy impacts.

Insights from Influencing Factors:

Government Expenditure:

Figure 15: Trends for Government Expenditure

- **Statistical Insights**:
    - **Education as a Budget Priority**: Education constitutes **12.58%** of government spending and **5.05%** of GDP, reflecting steady investment.
    - **Per-Student Funding:** Increases with each education level:
    - **Funding as GDP Percentage**:
        - Funding per tertiary student is the highest (**29.29%** of GDP per capita), while pre-primary students receive the least (**14.69%**).
- **Key Findings:**
    - Pre-primary to tertiary education is funded unevenly, with escalating inequalities in early childhood development.

Legal Guarantees:

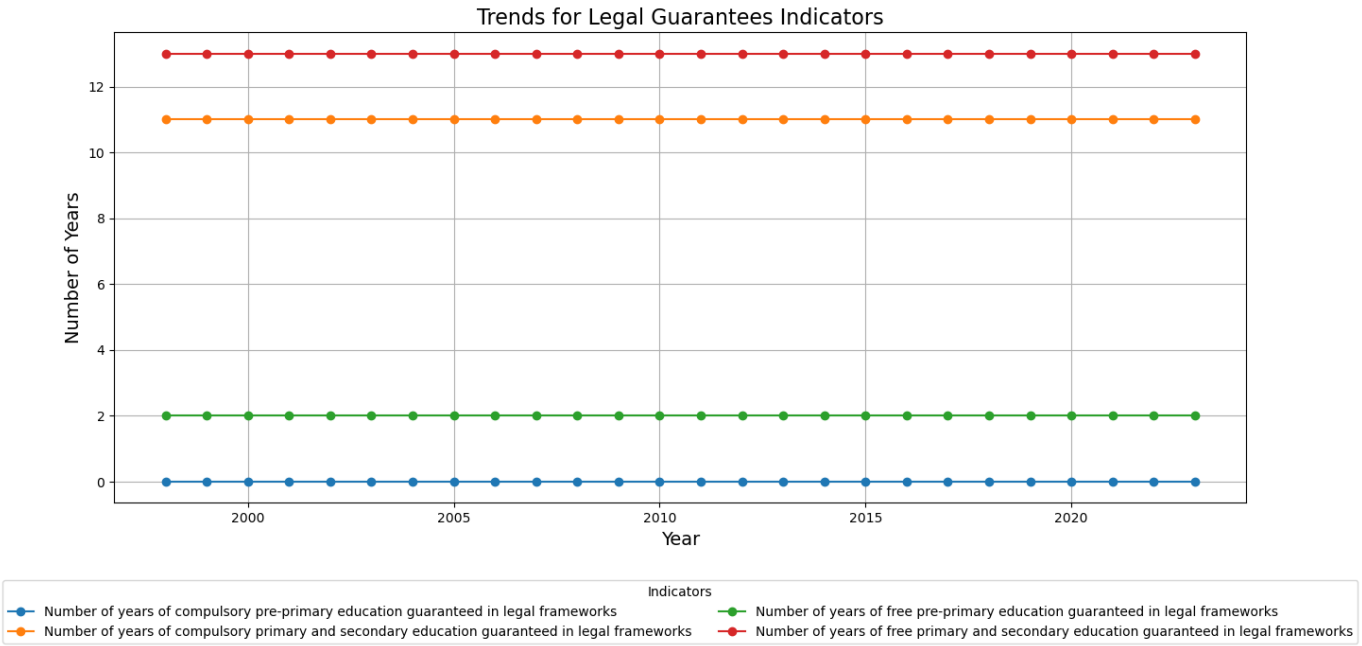


Figure 16: Trends for Legal Guarantees

- **Statistical Insights**:
    - Compulsory years for primary and secondary education is **11 years**, and **13 years are guaranteed as free**.
    - Pre-primary education has **2 free years** but no compulsory requirements.
- **Key Findings:**
    - Education is not compulsory at the pre-primary level, discouraging early learning, especially for poor families.

Socioeconomic Factors:

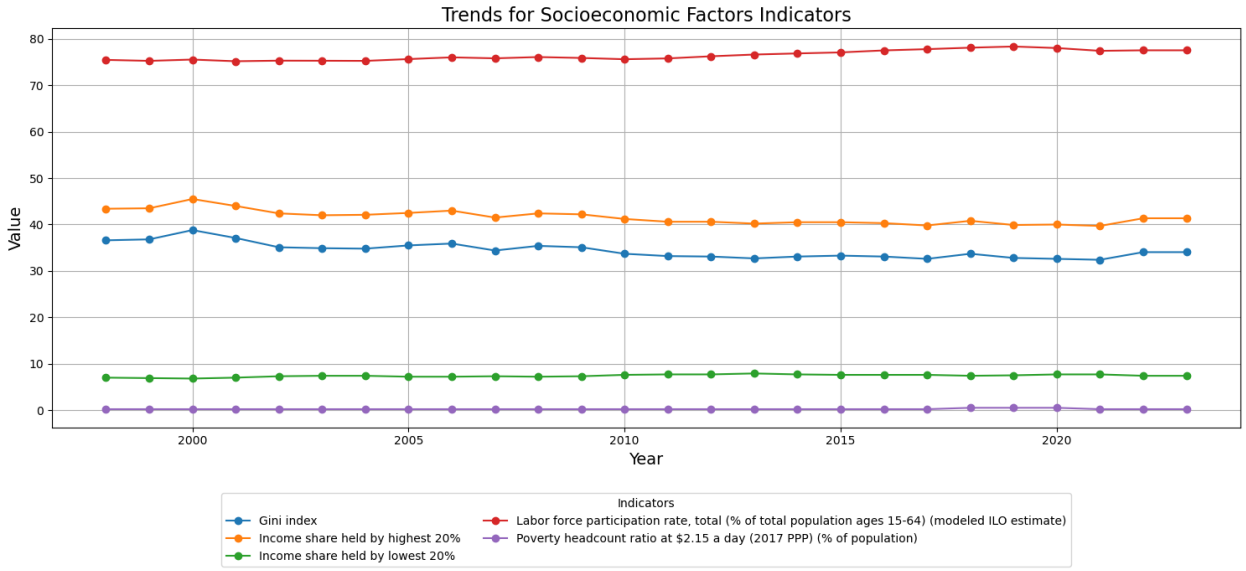


Figure 17: Trends for Socioeconomic Factors

- **Gini Index**:
    - Average of **34.42** indicates moderate income inequality, with low variation (**std 1.65**) suggesting stable distribution.
- **Income Distribution**:
    - The top 20% holds **41.59%** of income, while the lowest 20% holds only **7.4%,** showing significant disparity.
- **Labor Force Participation**:
    - High participation (**mean 76.45%**) shows strong economic activity.
- **Poverty Rate**:
    - Minimal poverty at $2.15/day (**mean 0.23%**) indicates effective socioeconomic policies.

Each indicator's pattern is reviewed for consistency with the result, and a combination of statistical and didactic findings. The next step is to assess the strength and nature of the relationships between key variables, such as government expenditure and socioeconomic factors, and educational performance.

4.3 Correlation Analysis

This correlation matrix shows some key cross-pollination, and I have specifically selected **Government Expenditure, Income Inequality** and **Labour Force Participation** to plot their association with education. Here is a rundown of correlations and recommendations.

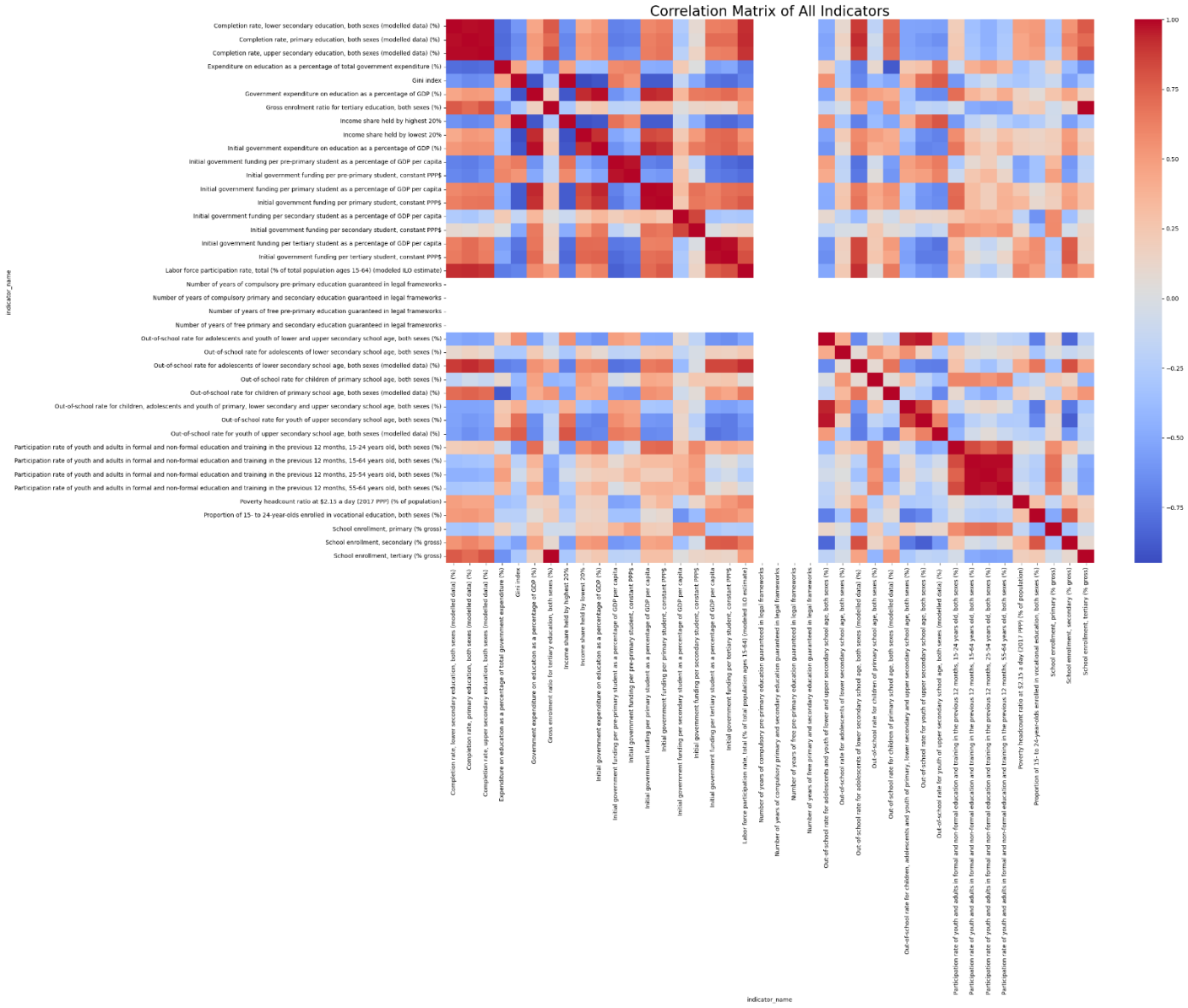


Figure 18: Correlation Matrix

1. Government Expenditure

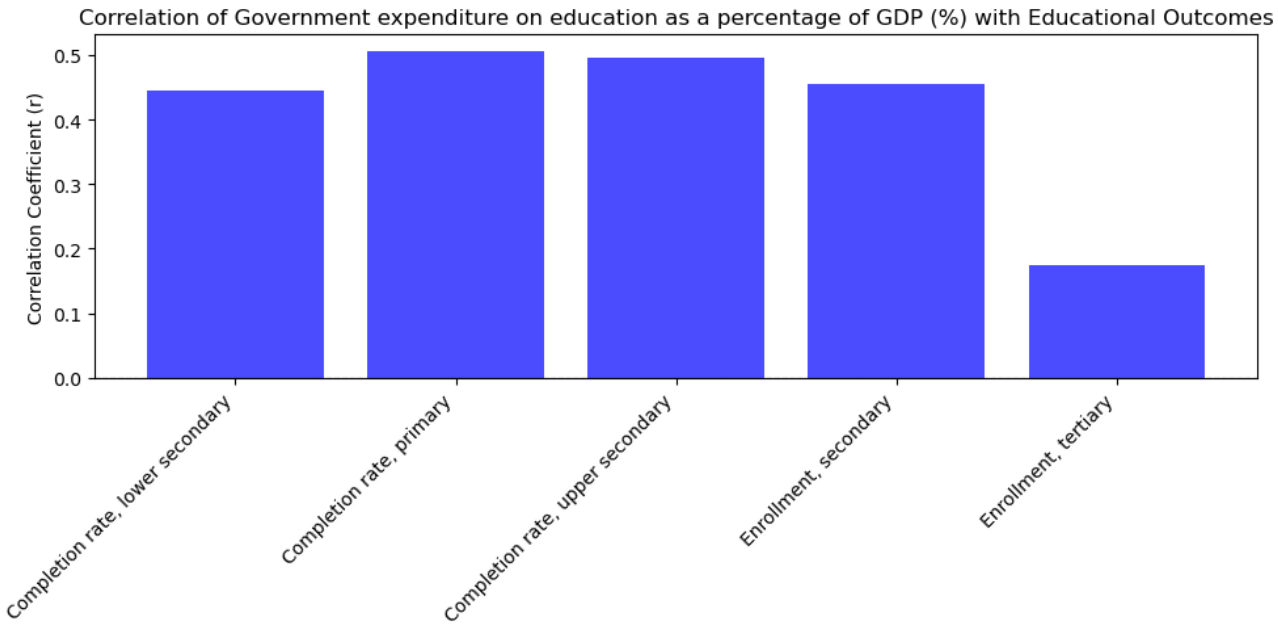


Figure 19: Correlation of Gov. Exp. on Edu. as a % of GDP with Education Outcomes

Key Insights

- **Moderate Positive Correlations:**
  - Lower secondary, primary, and upper secondary completion rates show moderate positive correlations with government expenditure ($r \approx 0.45$, 0.50, and 0.49, respectively).
  - Secondary school enrollment shows a moderate positive correlation ($r \approx 0.45$), while tertiary enrollment has a weaker positive correlation ($r \approx 0.17$).
- **Budget Allocation Impact:**
  - Increased government expenditure modestly enhances primary and secondary education outcomes.

2. Income Inequality

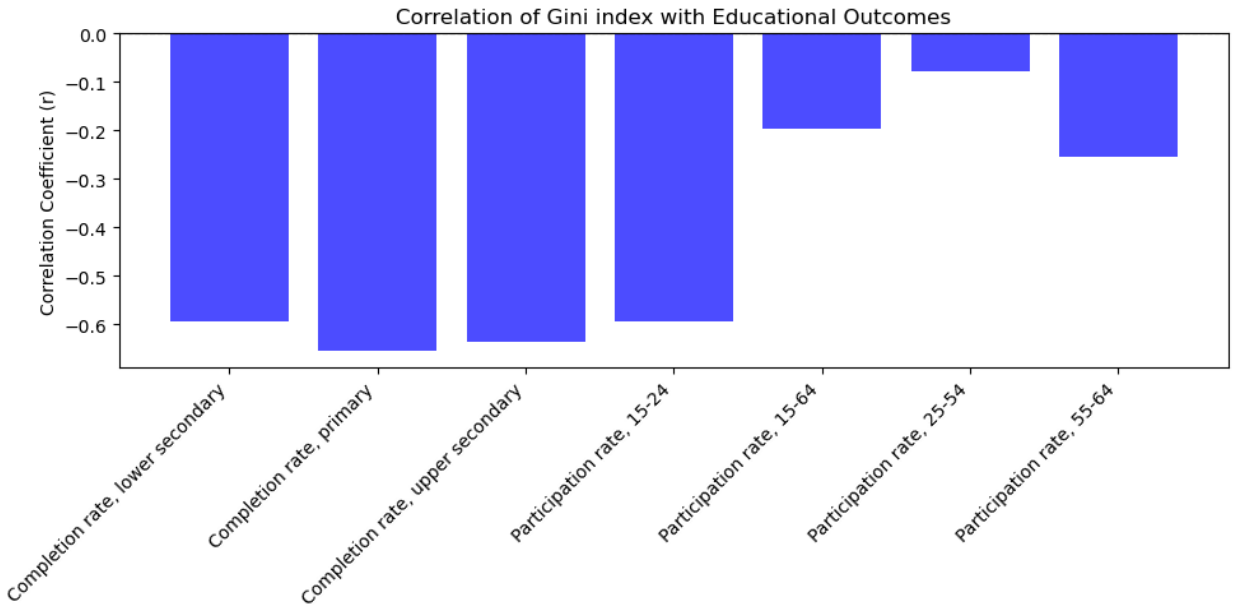


Figure 20: Correlation of Gini Index with Education Outcomes

Key Insights

- **Negative Correlation with Educational Outcomes:**
    - The Gini Index (income inequality) is:
        - Lower secondary, primary, and upper secondary completion rates show strong negative correlations with income inequality ($r \approx -0.59$, -0.65, and -0.63, respectively).
        - Participation rates in youth and adult education are negatively correlated with income inequality, ranging from moderate ($r \approx -0.59$ for youth, 15-24 years old) to weak ($r \approx -0.08$ for older adults, 25-54 years).
- **Positive Impact of Reduced Inequality:**
    - The income proportions of the bottom 20% moderately associate with higher participation and completion rates, particularly for adults (15–64) where the correlation is strongest ($r = 0.839$).

3. Labor Force Participation

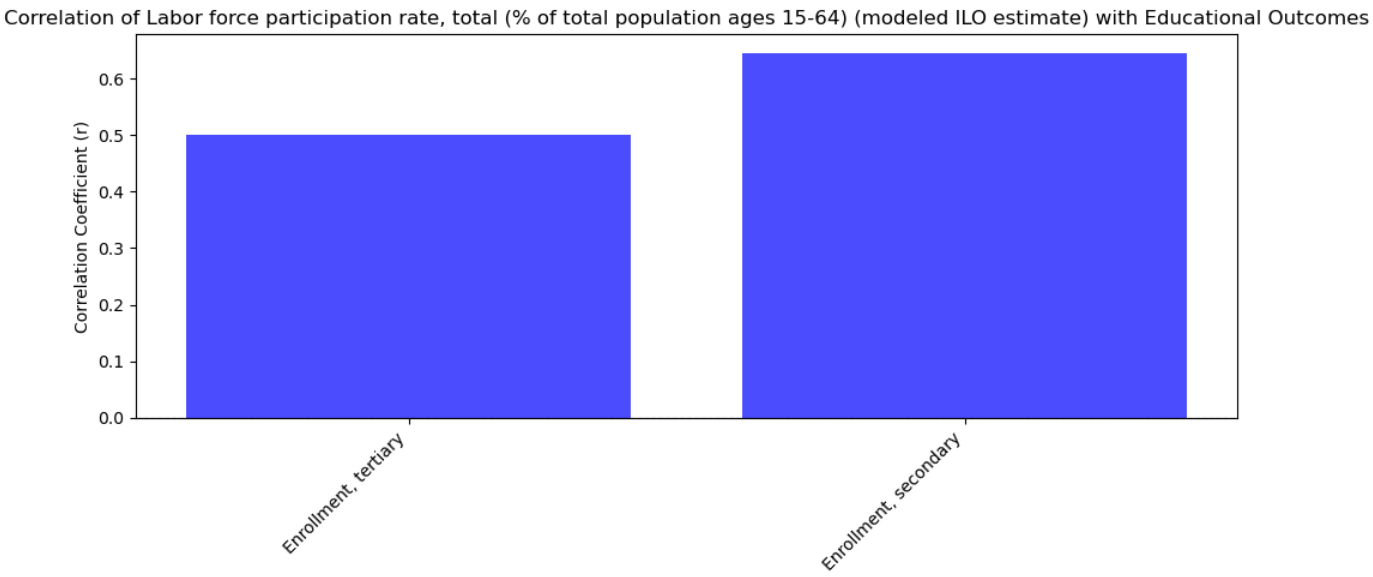


Figure 21: Correlation of Labor Force Participation Rate with Education Outcomes

Key Insights

- **Positive Correlations:**
    - Labor force participation has a moderate positive correlation with tertiary school enrollment ($r \approx 0.50$).
    - Secondary school enrollment also shows a moderate positive correlation ($r \approx 0.65$)

While correlation analysis reveals the strength of relationships, next section, regression analysis, enables us to quantify these effects and determine the extent to which key variables influence educational outcomes.

5. Regression Analysis

5.1 Analysis of Government Expenditure

This section traces how education investment by the government (% of GDP) affects major educational outcomes. There are the completion rates at primary, lower secondary and upper secondary as well as enrollment rates of secondary and tertiary levels. I want to identify trends, needs, and make recommendations to optimize investment in education in line with SDG4.

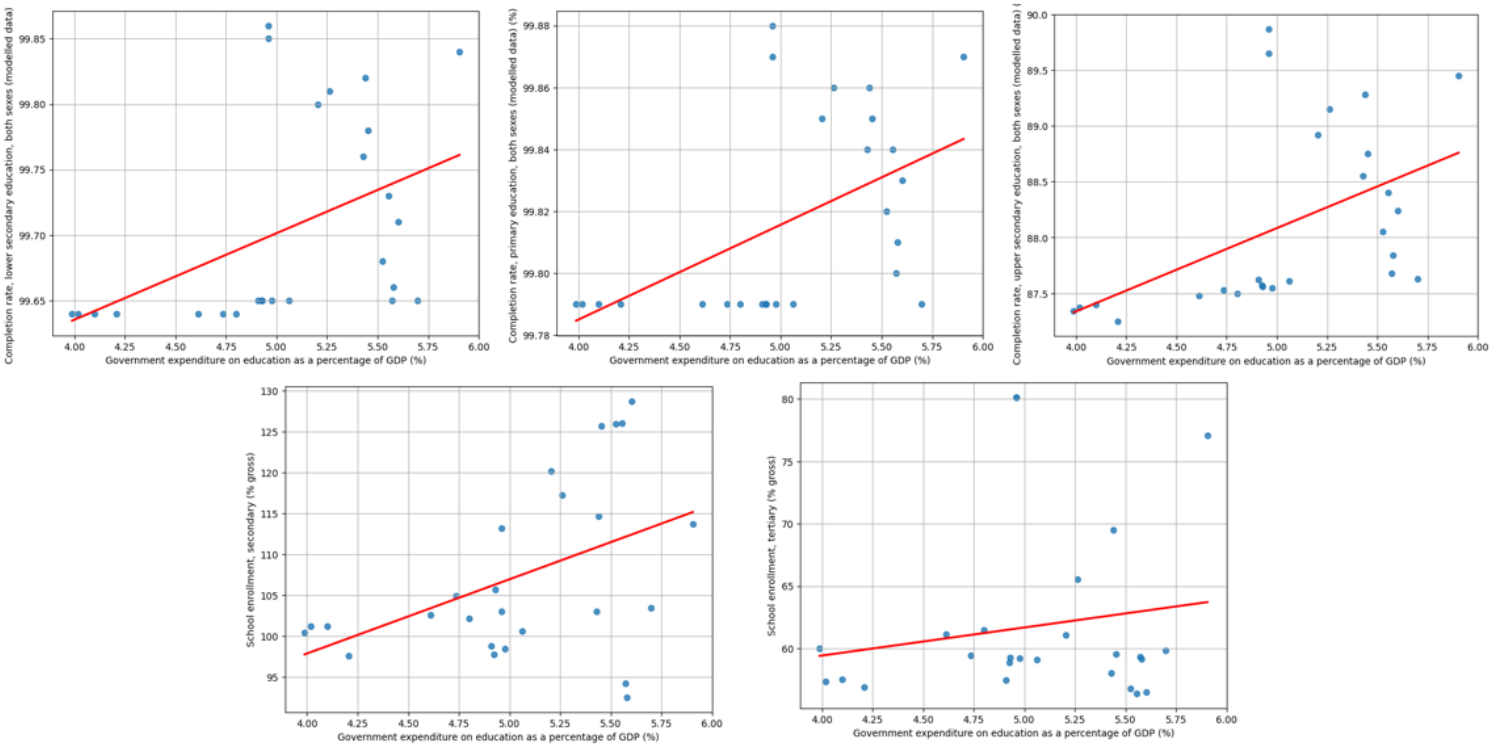


Figure 22: Regression Analysis of Government Expenditure

The regression analysis provides deeper insights into the relationships in the graphs. Key statistics are summarized in the heatmap below:

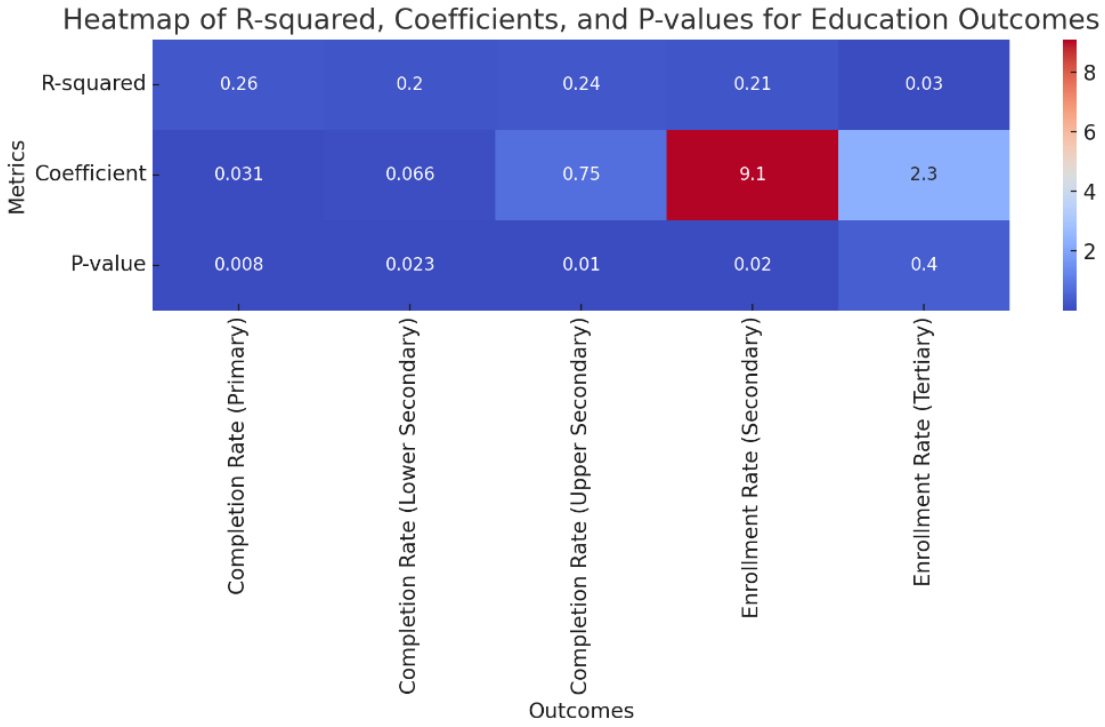


Figure 23: Key Statistics Summary of Government Expenditure

Key Observations:

1. **Completion Rates**:
    - **Primary and Lower Secondary**: Positive trends observed, although state investment not much of a impact as baseline completion rates are already high.
    - **Upper Secondary**: More responsive to higher expenditure, and more room for targeted spending.
2. **Enrollment Rates**:
    - **Secondary**: High responds to government spending; this level of funding is very critical.
    - **Tertiary**: Weaker correlation, pointing to barriers like affordability and accessibility.

5.2 Analysis of Income Inequality

A measure of inequality called the Gini index can tell us a lot about structural barriers to education access and achievement.

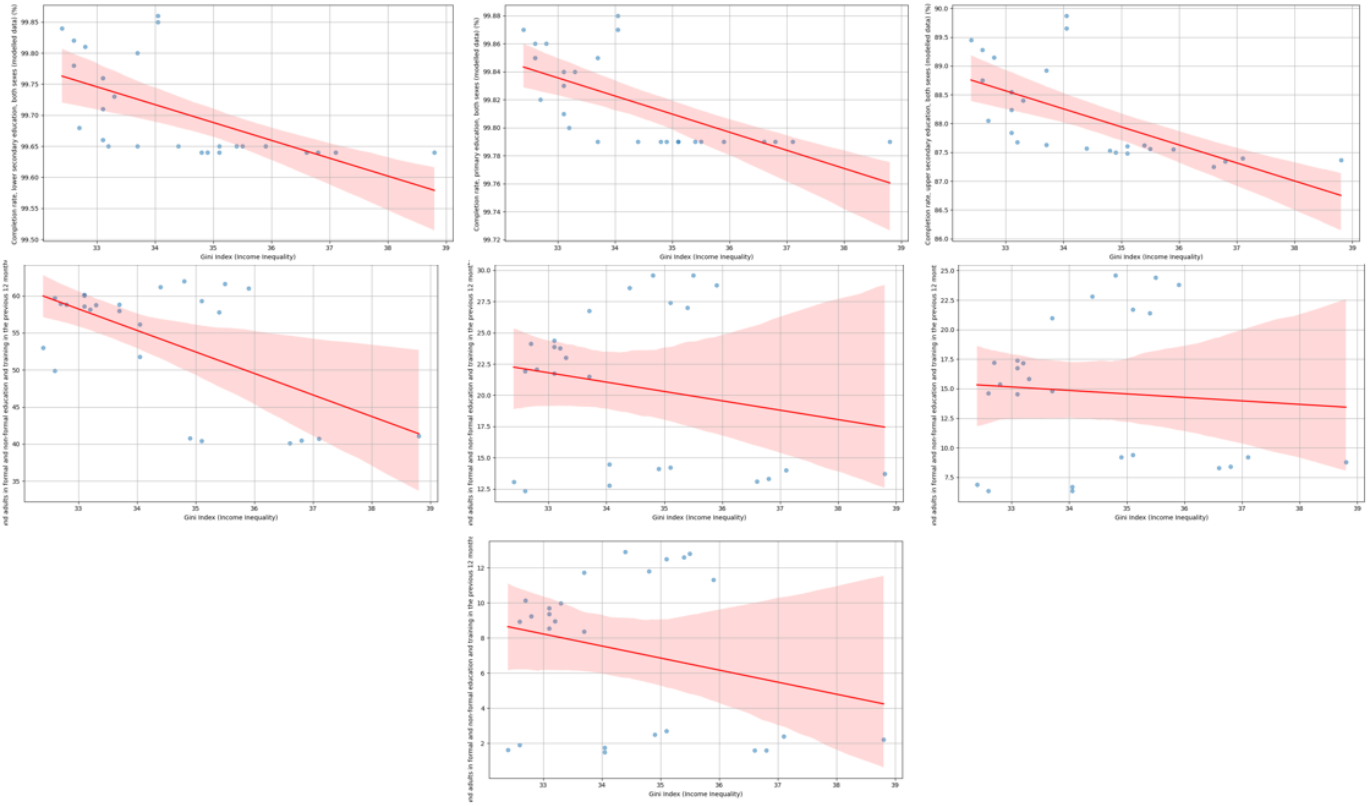


Figure 24: Regression Analysis of Income Inequality

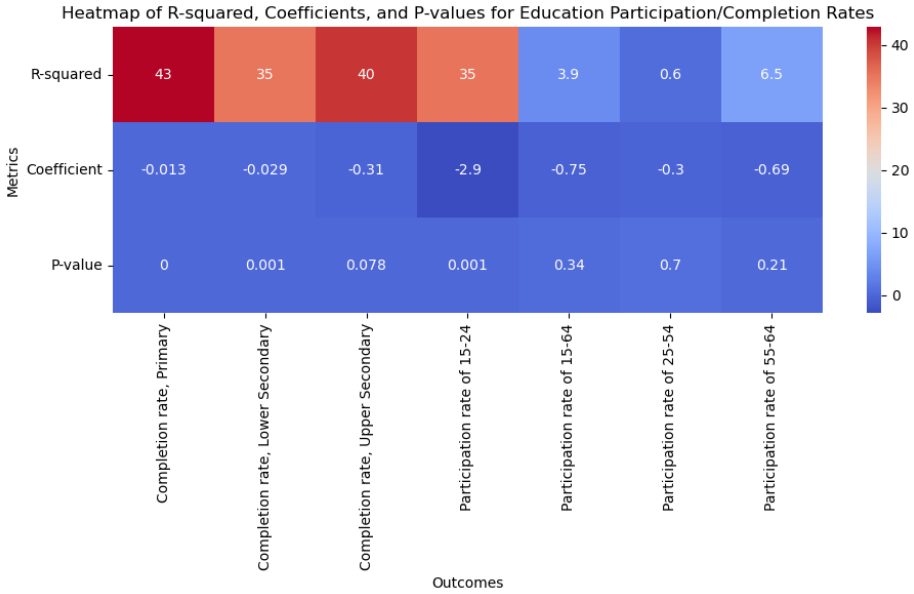


Figure 25: Key Statistics Summary of Income Inequality

	Key Insight
1	Strong negative relationship. Higher income inequality reduces primary education completion rates.
2	Significant negative impact. Income inequality hampers lower secondary completion rates.
3	Most sensitive to income inequality. Reflects systemic inequities at higher levels of education.
4	Income inequality discourages youth participation in education.
5	Weak relationship. Other factors likely play a larger role in adult participation.
6	Minimal impact on this age group's participation rates.
7	Weak relationship. Likely due to focus on other socioeconomic priorities for this age group.

Figure 26: Income Inequality Regression Key Insight

General Trends

1. **Significant Impact on Completion Rates**:
    - Income inequality affects the rate of completion at all levels but most visibly lower secondary schooling because of systemic inequality.
2. **Youth Participation Affected**:
    - A unit increase in the Gini index reduces youth participation by 2.9%.
3. **Weak Impact on Adults**:
    - Adult participation rates (15-64 years) do not correlate very strongly or significantly with income inequality, suggesting the contribution of other socioeconomic variables.

5.3 Analysis of Labor Force Participation

This analysis examines how labor force participation affects secondary and tertiary enrollment. Secondary enrollment depends on household economic stability, while tertiary enrollment faces barriers like cost and accessibility, guiding targeted policy development.

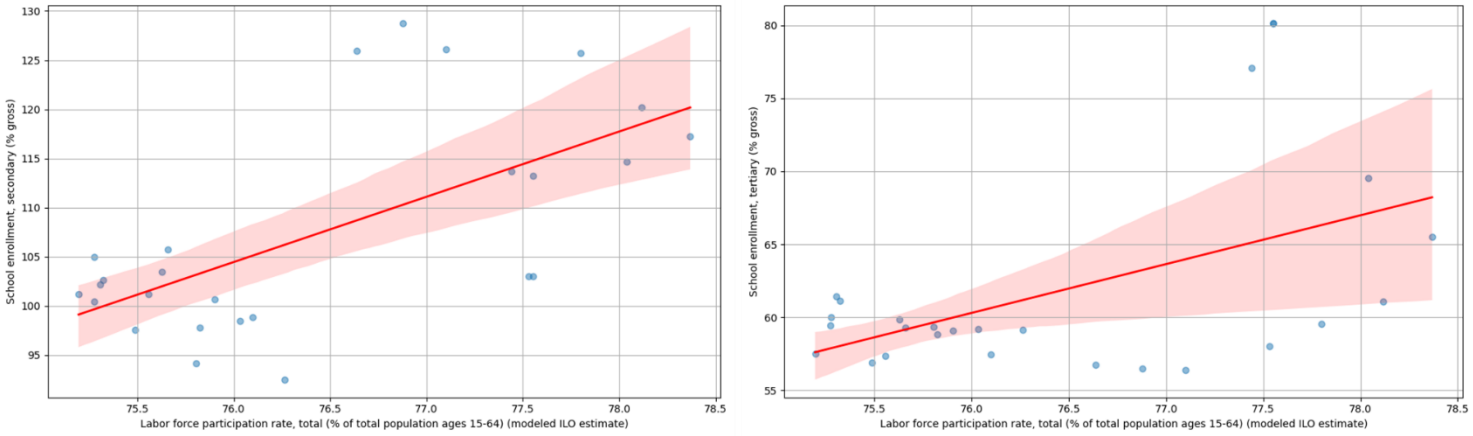


Figure 27: Regression Analysis of Labor Force Participation

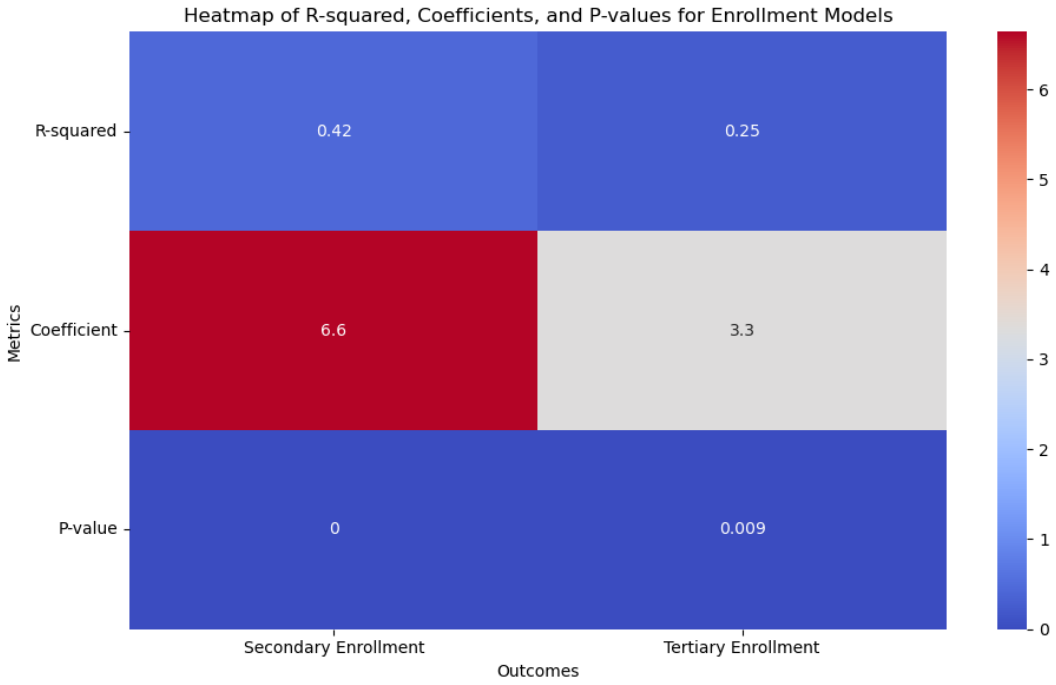


Figure 28: Key Statistics Summary of Labor Force Participation

Key Findings

1. **Strength of Correlation**:
    - **Secondary Enrollment**: Strong positive correlation ($R^2 = 0.417$, coefficient = 6.64); a 1% increase in labor force participation links to a 6.64% rise in enrollment.
    - **Tertiary Enrollment**: Weaker positive correlation ($R^2 = 0.251$, coefficient = 3.35); a 1% increase links to a 3.35% rise in enrollment.
2. **Statistical Significance**:
    - Both models are statistically significant (P-values: secondary = 0.000, tertiary = 0.009).
3. **Intercept Values**:
    - Both models exhibit **negative intercepts**:
        - **Secondary**: -400.13.
        - **Tertiary**: -193.94.
    - Reflect theoretical limitations outside observed labor force participation rates.
4. **Explanatory Power**:
    - **Secondary Enrollment**: Explains 41.7% of variance, so labor force participation is a key factor.
    - **Tertiary Enrollment**: Explains 25.1% of variance; other factors like affordability and accessibility are crucial.

6. Key Findings and Recommendations

Key Findings:

1. **Completion Rates:**
    - Primary and Lower Secondary completion rates are high, indicating strong educational access. However, upper secondary education requires improvement.
2. **Enrollment Ratios:**
    - Tertiary enrollment is lower, reflecting financial and accessibility barriers. Primary and secondary enrollment rates exceed 100%, suggesting overage or repetition.
3. **Out-of-School Rates:**
    - Out-of-school rates have been declining, particularly at the primary and lower secondary levels, indicating positive progress in education access.
4. **Socioeconomic Inequality:**
    - Inequality in incomes remains an impediment to universal access to education, and it negatively affects school completion and participation, especially for poorer students.

Strategic Recommendations:

1. **Policy Recommendations for Access:**
    - **Expand Access to Upper Secondary Education**: More scholarships, mentorship and community outreach programs to minimize the number of students who drop out and achieve at upper secondary.
    - **Address Over-age Enrollment**: Set up policies to focus on grade level, and avoid over-age enrollment, so that children graduate at the correct age.
2. **Policy Recommendations for Quality:**
    - **Increase Investment in Vocational Education**: Address the drop in vocational education enrollment by expanding access and understanding of career paths. Promote public-private collaborations to design programs directly based on labor market demand.
    - **Enhance Teacher Training**: Invest in Teacher Training so that education becomes more high-quality, especially secondary and higher education schools so that teachers can effectively manage different learning needs.

3. **Policy Recommendations for Equity:**
    - **Reduce Socioeconomic Barriers to Higher Education**: To reduce tertiary enrollment, scale up financial aid, scholarships and student loan programs so students of low income are equally accessible.
    - **Targeted Support for Disadvantaged Students**: Adopt affirmative action and funding models to reach out to students with lower socioeconomic status and guarantee them high-quality education at every level.
4. **Promote Lifelong Learning:**
    - **Increase Adult Education Participation**: Encourage a range of flexible learning programs including online, vocational and part-time courses, so that employed adults can continue to learn and develop skills in response to job demands.

7. Conclusion and Reflection

This project rated the UK's attainment of SDG 4 using key education indicators and government spending and socioeconomic inequality. The analysis showed that the UK has made progress on primary and secondary education but there are still gaps – particularly in access to and quality of tertiary education. Government spending did correlate with education, but income inequality prevented equity from happening.

The principal recommendations include extending upper secondary education, expanding vocational training and lowering tuition fees for higher education. These results can be a base for policy to help implement SDG 4 and drive equitable and high-quality education.

Reflection on the Project

I developed a stronger critical thinking approach, especially in using analytics techniques like correlation and regression analysis to assess educational outcomes. The ability to connect the dots between various factors. It allowed me to provide actionable recommendations for addressing educational disparities. One of the project's strengths was effectively categorizing data and presenting it visually to make complex relationships clearer. However, challenges arose in dealing with missing data and ensuring data consistency across sources.

In future projects, I would focus more on improving data cleaning processes and exploring additional regression models to further refine the insights. Ultimately, this project

sharpened my ability to think critically, identify key patterns, and develop recommendations that can influence real-world policy decisions.

8. References

Bhandari, P. (2022, January 3). *A beginner's guide to triangulation in research*. Scribbr; Scribbr. https://www.scribbr.com/methodology/triangulation/

Cepelak, C. (2023, February 2). *An Introduction to Data Ethics: What is the Ethical Use of Data?* Datacamp.com; DataCamp. https://www.datacamp.com/blog/introduction-to-data-ethics

Cipan, V. (2023, May 7). *Ethics and ethical data visualization: A complete guide • viborc.com*. Viborc.com. https://viborc.com/ethics-and-ethical-data-visualization-a-complete-guide/

Global Database on Intergenerational Mobility | Data Catalog. (2022). Worldbank.org. https://datacatalog1.worldbank.org/search/dataset/0050771/Global-Database-on-Intergenerational-Mobility

Hassan, M. (2022, August 28). *Data Verification - Process, Types and Examples*. Research Method. https://researchmethod.net/data-verification/

Marziyeh Afkanpour, Hosseinzadeh, E., & Hamed Tabesh. (2024). Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02310-6

Podda, E. (2021). CONFERENCE OF EUROPEAN STATISTICIANS Expert Meeting on Statistical Data Confidentiality. https://unece.org/sites/default/files/2021-12/SDC2021_Day1_Podda_AD.pdf

UNESCO Institute for Statistics. (n.d.-a). *SDG 4 data*. Retrieved December 15, 2024, from https://sdg4-data.uis.unesco.org/

UNESCO Institute for Statistics. (n.d.-b). *UIS education data browser: SDG 4 monitoring*. Retrieved December 15, 2024, from https://databrowser.uis.unesco.org/browser/EDUCATION/UIS-SDG4Monitoring

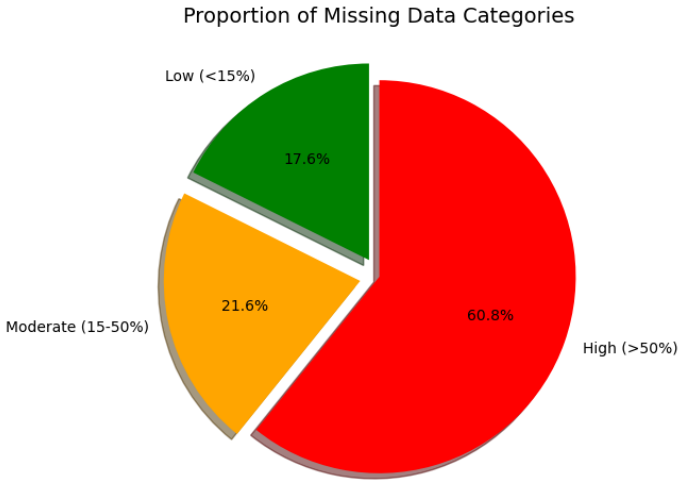
9. Appendix

9.1 Handling Missing Value

9.1.1 Calculate and visualize the percentage of missing values per indicator

```python
# Step 1: Calculate the percentage of missing values per indicator
missing_by_indicator = (
    long_df.groupby('indicator_name')['value']
    .apply(lambda x: x.isnull().mean() * 100)
    .reset_index(name='missing_percentage')
)

# Step 2: Display indicators with their missing percentages
print("Missing Value Percentage by Indicator:")
print(missing_by_indicator)

# Step 3: Visualize the missing data pattern
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 15))
sns.barplot(data=missing_by_indicator, x='missing_percentage', y='indicator_name', orient='h')
plt.title('Missing Data Percentage by Indicator')
plt.xlabel('Percentage of Missing Values')
plt.ylabel('Indicator Name')
plt.show()

# Save the missing value percentages to a CSV file
missing_by_indicator.to_csv('C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in busine
print("File saved as 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business anal
```

9.1.2 Visualize and Handling with my 3 ranks of missing values




```python
# Import necessary libraries
import pandas as pd
import numpy as np

# Load the datasets
low_missing_path = 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/globa
moderate_missing_path = 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/
high_missing_path = 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/glob
long_dataset_path = 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/glob

# Read the data
low_missing = pd.read_csv(low_missing_path)
moderate_missing = pd.read_csv(moderate_missing_path)
high_missing = pd.read_csv(high_missing_path)
long_df = pd.read_csv(long_dataset_path)

# Ensure the 'value' column is numeric, converting non-numeric values to NaN
long_df['value'] = pd.to_numeric(long_df['value'], errors='coerce')

# 1. Linear interpolation for low missing indicators (<5%)
low_missing_indicators = low_missing['indicator_name']
for indicator in low_missing_indicators:
    indicator_data = long_df[long_df['indicator_name'] == indicator]
    long_df.loc[long_df['indicator_name'] == indicator, 'value'] = indicator_data['value'].interpolate(method='linear')

# 2. Median imputation for moderate missing indicators (5-50%)
moderate_missing_indicators = moderate_missing['indicator_name']
for indicator in moderate_missing_indicators:
    median_value = long_df[long_df['indicator_name'] == indicator]['value'].median()
    long_df.loc[(long_df['indicator_name'] == indicator) & (long_df['value'].isnull()), 'value'] = median_value

# 3. Dropping high missing indicators (>50%)
high_missing_indicators = high_missing['indicator_name']
long_df = long_df[~long_df['indicator_name'].isin(high_missing_indicators)]

# Save the cleaned dataset
cleaned_dataset_path = 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/g
long_df.to_csv(cleaned_dataset_path, index=False)

print("Cleaned dataset saved as 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini
```

9.2 Generate and Visualize Time-Series Line Chart

9.1.3 Generate Time-series Line Chart

```python
import pandas as pd
import matplotlib.pyplot as plt
import os
import re

# Load the cleaned dataset
file_path = 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/global education policies
cleaned_df = pd.read_csv(file_path)

# Create the specified folder to save the plots
output_folder = "C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/Mini-Project_Progress
os.makedirs(output_folder, exist_ok=True)

# Get the unique indicators
all_indicators = cleaned_df['indicator_name'].unique()

# Function to sanitize filenames
def sanitize_filename(name):
    # Replace invalid characters with underscores
    return re.sub(r'[^\w\-_\. ]', '_', name)

# Generate and save plots for all indicators
for indicator in all_indicators:
    try:
        # Filter data for the indicator
        indicator_data = cleaned_df[cleaned_df['indicator_name'] == indicator]

        # Check if data exists for the indicator
        if indicator_data.empty:
            print(f"No data for indicator: {indicator}")
            continue

        # Create the plot
        plt.figure(figsize=(10, 6))
        plt.plot(indicator_data['year'], indicator_data['value'], marker='o', linestyle='-', label=indicator)
        plt.title(f"Progress Trend: {indicator}")
        plt.xlabel("Year")
        plt.ylabel("Value")
        plt.grid(True)
        plt.legend()

        # Save the plot with sanitized file name
        file_name = sanitize_filename(indicator) + ".png"
        plot_path = os.path.join(output_folder, file_name)

        # Fallback for overly long file paths
        if len(plot_path) > 255:  # Limit for many operating systems
            file_name = f"Indicator_{hash(indicator)}.png"
            plot_path = os.path.join(output_folder, file_name)

        plt.savefig(plot_path)
        plt.close()  # Close the figure to avoid display overload
        print(f"Saved plot: {file_name}")  # Log saved plots
    except Exception as e:
        # Log the error and ensure the process continues
        print(f"Error processing indicator: {indicator}. Error: {e}")
        continue

# Print the absolute path of the folder where plots are saved
absolute_folder_path = os.path.abspath(output_folder)
print(f"Plots saved in folder: {absolute_folder_path}")
```

9.2.2 Visualize Trend Categorization by Types and Sub-categories

```python
import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
file_path = 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/global education policies
data = pd.read_csv(file_path)

# Group by Categorization and Trend Category to count indicators
trend_summary = data.groupby(['Categorization of Indicators', 'Trend Categorization']).size().unstack(fill_value=0)

# Define custom colors for the trends
colors = {'Positive Progress': '#2EB82E', 'Stagnation': '#0062FF', 'Decline': '#FF003B'}

# Create a bar chart for Outcome vs Influencing grouped by trend categories
ax = trend_summary.plot(kind='bar', stacked=False, figsize=(10, 6), color=[colors[col] for col in trend_summary.columns])
plt.title('Indicator Trends by Categorization (Outcome vs Influencing)')
plt.xlabel('Categorization of Indicators')
plt.ylabel('Number of Indicators')
plt.legend(title='Trend Categorization', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()

# Group by Sub-category and Trend Category to count indicators
subcategory_summary = data.groupby(['Sub-category', 'Trend Categorization']).size().unstack(fill_value=0)

# Create a stacked bar chart for Sub-category vs Trend Categories
ax = subcategory_summary.plot(kind='bar', stacked=True, figsize=(12, 8), color=[colors[col] for col in subcategory_summary.columns])
plt.title('Sub-category Contribution to Trend Categories')
plt.xlabel('Sub-category')
plt.ylabel('Number of Indicators')
plt.legend(title='Trend Categorization', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```

9.3 Create Correlation Matrix and Visualize by Group

9.3.1 Create Correlation Matrix

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
file_path = 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/global
data = pd.read_csv(file_path)

# Pivot the data for correlation analysis (rows: year, columns: indicators)
correlation_data = data.pivot(index='year', columns='indicator_name', values='value')

# Compute the correlation matrix
correlation_matrix = correlation_data.corr()

# Plot the correlation matrix as a heatmap
plt.figure(figsize=(30, 24))  # Further increased figure size for even better readability
sns.heatmap(correlation_matrix, cmap='coolwarm', annot=False, fmt=".2f", cbar=True)
plt.title('Correlation Matrix of All Indicators', fontsize=24)  # Increased font size for title
plt.tight_layout()

# Display the heatmap
plt.show()
```

9.3.2 Visualize Correlation by Group

```python
import pandas as pd
import matplotlib.pyplot as plt

# Load the correlation matrix
file_path = 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/global education policies impact social equity and economic mobility across '
correlation_matrix = pd.read_csv(file_path)

# Define the independent and dependent variables for each case
analysis_vars = {
    "Government expenditure on education as a percentage of GDP (%)": [
        "Completion rate, lower secondary education, both sexes (modelled data) (%) ",
        "Completion rate, primary education, both sexes (modelled data) (%) ",
        "Completion rate, upper secondary education, both sexes (modelled data) (%) ",
        "School enrollment, secondary (% gross)",
        "School enrollment, tertiary (% gross)"
    ],
    "Gini index": [
        "Completion rate, lower secondary education, both sexes (modelled data) (%) ",
        "Completion rate, primary education, both sexes (modelled data) (%) ",
        "Completion rate, upper secondary education, both sexes (modelled data) (%) ",
        "Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, 15-24 years old, both sexes (%)",
        "Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, 15-64 years old, both sexes (%)",
        "Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, 25-54 years old, both sexes (%)",
        "Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, 55-64 years old, both sexes (%)"
    ],
    "Labor force participation rate, total (% of total population ages 15-64) (modeled ILO estimate)": [
        "School enrollment, tertiary (% gross)",
        "School enrollment, secondary (% gross)"
    ]
}

# Simplify the labels
label_mappings = {
    "Completion rate, lower secondary education, both sexes (modelled data) (%) ": "Completion rate, lower secondary",
    "Completion rate, primary education, both sexes (modelled data) (%) ": "Completion rate, primary",
    "Completion rate, upper secondary education, both sexes (modelled data) (%) ": "Completion rate, upper secondary",
    "Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, 15-24 years old, both sexes (%)": "Participation rate, 15-24",
    "Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, 15-64 years old, both sexes (%)": "Participation rate, 15-64",
    "Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, 25-54 years old, both sexes (%)": "Participation rate, 25-54",
    "Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, 55-64 years old, both sexes (%)": "Participation rate, 55-64",
    "School enrollment, secondary (% gross)": "Enrollment, secondary",
    "School enrollment, tertiary (% gross)": "Enrollment, tertiary"
}

(variable) dependent_vars: list[str]

# Extract correlations for each independent variable and its dependent variables
analysis_results = {}
for independent_var, dependent_vars in analysis_vars.items():
    correlations = correlation_matrix.loc[
        correlation_matrix['indicator_name'] == independent_var, dependent_vars
    ]
    analysis_results[independent_var] = correlations.T  # Transpose for readability

# Create separate bar charts with the simplified labels and legends showing full dependent variable names
for independent_var, correlations in analysis_results.items():
    dependent_vars = correlations.index
    values = correlations.values.flatten()

    # Simplify the labels for the x-axis using the label_mappings
    simplified_labels = [label_mappings[label] for label in dependent_vars]

    # Plot for this independent variable
    plt.figure(figsize=(10, 5))
    bars = plt.bar(simplified_labels, values, color='blue', alpha=0.7)
    plt.title(f"Correlation of {independent_var} with Educational Outcomes")
    plt.ylabel("Correlation Coefficient (r)")
    plt.xticks(rotation=45, ha="right")
    plt.axhline(0, color="black", linewidth=0.8, linestyle="--")

    plt.tight_layout()
    plt.show()
```

9.4 Create & Visualize Regression Analysis and Statistic Summary

9.4.1 Create & Visualize Regression Model Analysis

```python
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
file_path = 'C:/Users/User/OneDrive - University of Exeter/PostGradute/Topic in business analytics/Mini Project/global education
dataset = pd.read_csv(file_path)

# Define the relevant indicators
independent_var = "Government expenditure on education as a percentage of GDP (%)"
dependent_vars = [
    "Completion rate, lower secondary education, both sexes (modelled data) (%) ",
    "Completion rate, primary education, both sexes (modelled data) (%) ",
    "Completion rate, upper secondary education, both sexes (modelled data) (%) ",
    "School enrollment, secondary (% gross)",
    "School enrollment, tertiary (% gross)"
]

# Filter the dataset for the relevant indicators
filtered_data = dataset[dataset['indicator_name'].isin([independent_var] + dependent_vars)]

# Pivot the data to have years as rows and indicators as columns
pivoted_data = filtered_data.pivot(index='year', columns='indicator_name', values='value').reset_index()

# Drop rows with missing values to ensure clean regression
pivoted_data = pivoted_data.dropna()

# Define the independent variable (predictor) and dependent variables (outcomes)
X = pivoted_data[independent_var]
Y = pivoted_data[dependent_vars]

# Add a constant term to the predictor for the regression model
X_with_const = sm.add_constant(X)

# Fit the multiple regression model for each dependent variable and collect summaries
results = {}
for dependent_var in dependent_vars:
    model = sm.OLS(Y[dependent_var], X_with_const).fit()
    results[dependent_var] = model

    # Plotting
    plt.figure(figsize=(8, 6))

    # Scatter plot with regression line
    sns.regplot(x=X, y=Y[dependent_var], ci=None, line_kws={"color": "red"})
    plt.xlabel(independent_var)
    plt.ylabel(dependent_var)
    plt.grid(True)
    plt.show()



# Display regression summaries
for dependent_var, model in results.items():
    print(f"Regression Results for {dependent_var}:\n")
    print(model.summary())
    print("\n")
```

9.4.2 Create & Visualize Regression Statistic Summary

```python
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns

# Creating the dataframe from the given table data
data = {
    'Outcome': ['Completion Rate (Primary)', 'Completion Rate (Lower Secondary)', 'Completion Rate (Upper Secondary)', 'Enrollment Rate (Secondary)', 'Enrollment Rate (Tertiary)'],
    'R-squared': [0.256, 0.198, 0.245, 0.207, 0.030],
    'Coefficient': [0.0306, 0.0660, 0.7466, 9.0904, 2.2519],
    'P-value': [0.008, 0.023, 0.010, 0.020, 0.397]
}

df = pd.DataFrame(data)

# Setting up a seaborn heatmap to visualize the R-squared, Coefficients, and P-values
fig, ax = plt.subplots(figsize=(10, 6))

# Creating a heatmap to compare all metrics across outcomes
sns.heatmap(df[['R-squared', 'Coefficient', 'P-value']].T, annot=True, cmap='coolwarm', cbar=True, xticklabels=df['Outcome'], yticklabels=['R-squared', 'Coefficient', 'P-value'], ax=ax)

# Rotating the y-axis labels to horizontal
ax.set_yticklabels(ax.get_yticklabels(), rotation=0)

# Adding titles and labels
ax.set_title('Heatmap of R-squared, Coefficients, and P-values for Education Outcomes')
ax.set_xlabel('Outcomes')
ax.set_ylabel('Metrics')

plt.tight_layout()
plt.show()
```