Bangladesh Socio-Economic Analysis

MSDS 422 Final Project

Group 3: Jovana Petkovic, MJ Neev, Ryan Fallon and Yemi Adetutu

Executive Summary

- In Bangladesh, many challenges of regional disparities persist despite the nation's significant socio-economic transformation in following:
  - Declining poverty rate
  - Improving literacy
  - Rapid urbanization
- The goal is to create an automated ML pipeline to support the nation's policymaking with focus on
  - understanding of driver's of poverty and literacy
  - forecasting socio-economic trends
  - the impact of urbanization on infrastructure development

Problem Statement & Research Objectives

Problem: Regional disparities in poverty, literacy, and infrastructure hinder equitable growth.

Research Objectives:

1. Analysis of poverty and literacy trends in different regions.
2. Identifying socio-economic and demographic impacts on poverty and literacy.
   a. Education and healthcare spending by government.
   b. Impact of urbanization, GDP growth, and access to infrastructure.
3. Assess urban-rural disparities and implications for socio-economic status
4. Building a predictive model to forecast poverty and literacy under various policy scenarios.
5. Recommendations for balancing urban growth with equitable development.

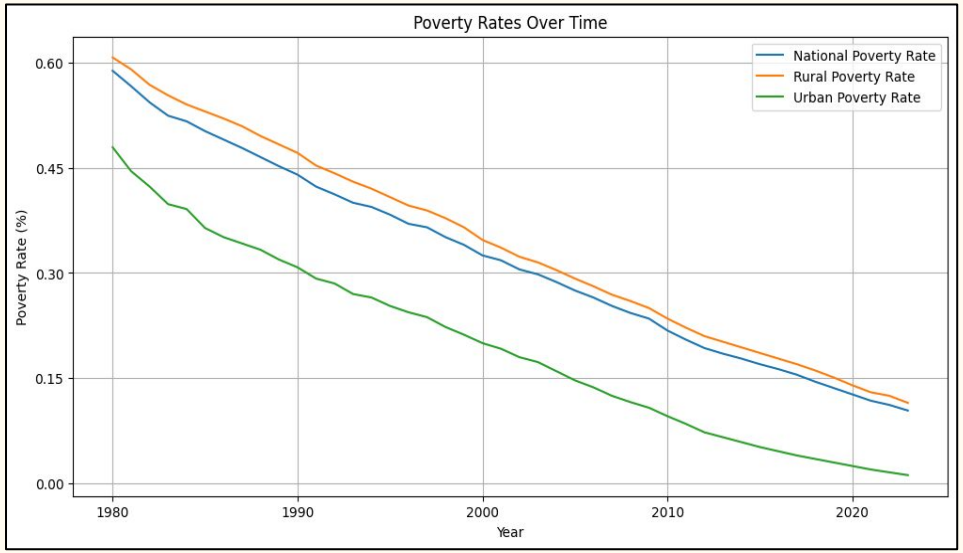
1. Exploratory Data Analysis

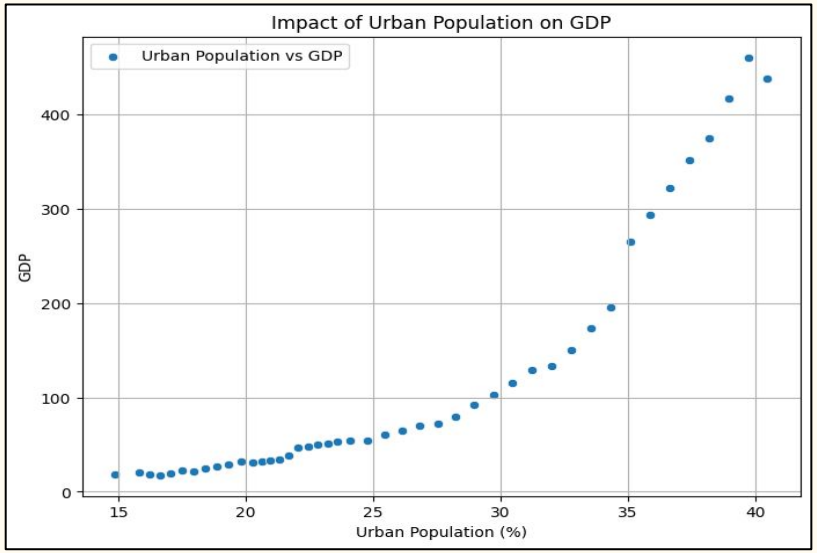
The dataset includes historical data on key indicators such as poverty rates, literacy, GDP, healthcare spending, education spending, urbanization metrics, and population trends. Key findings from the EDA include:

Key Highlights:

- Declining national poverty with pronounced urban-rural disparities.
- Urbanization positively impacts GDP growth but widens disparities.
- Strong correlation of literacy with urbanization and GDP; weak correlation with education spending.
- Strong positive relationship between unemployment and literacy.
- Strong negative relationship between life expectancy and literacy.

1. Exploratory Data Analysis, cont.





- Overall, the nation's poverty rate has been decreasing overtime, but in rural areas, poverty rate is the highest all period.

- Strong positive relationship with GDP growth and urbanization

1. Exploratory Data Analysis, cont.

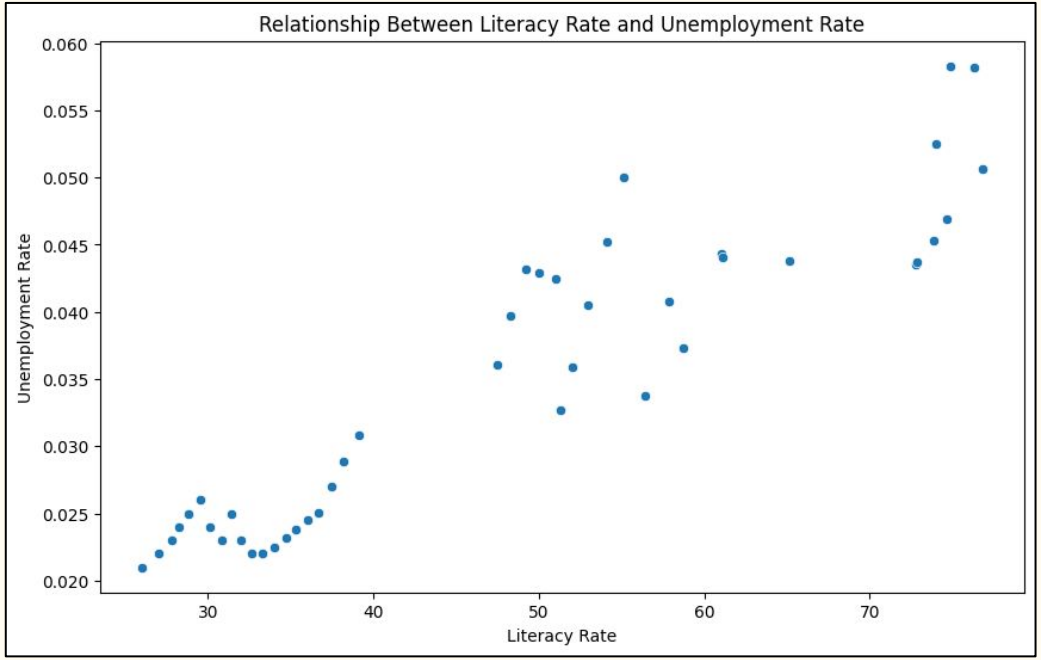
Correlation with Literacy Rate:	
Literacy Rate	1.000000
3-Year Literacy Rolling Avg	0.995405
Urban Population (%)	0.988159
Year	0.980226
Population	0.972824
Unemployment Rate	0.928044
GDP Per Capita	0.901279
GDP	0.897481
3-Year GDP Rolling Avg	0.888713
Electricity Access	0.711008
Export Growth	0.695975
Fossil Fuel Consumption	0.589086
Import Growth	0.585368
Urbanization × GDP Growth	0.535440
Healthcare Spending	0.416864
Clean Water Access	0.397836
Inflation Rate	0.294528
Education Spending	0.154902
Greenhouse Gas Emissions Change	0.086434
Annual Crime Rate Change	-0.083553
Suicide Rate	-0.305661
Maternal Mortality Rate	-0.370832
Net Migration Rate	-0.517146
Death Rate	-0.855011
Life Expectancy Growth Rate	-0.861326
Birth Rate	-0.883872
Infant Mortality Rate	-0.944369
Urban Poverty Rate	-0.955698
National Poverty Rate	-0.968302
Rural Poverty Rate	-0.970977

- Literacy rate has a strong correlation with urbanization and GDP
- Interestingly, unemployment rate shows a strong positive relationship with literacy rate
- Education spending has a weak correlation with literacy rate
- Life expectancy growth rate has a strong negative correlation with literacy rate

1. Exploratory Data Analysis, cont.

A strong positive relationship between Unemployment Rate and Literacy Rate

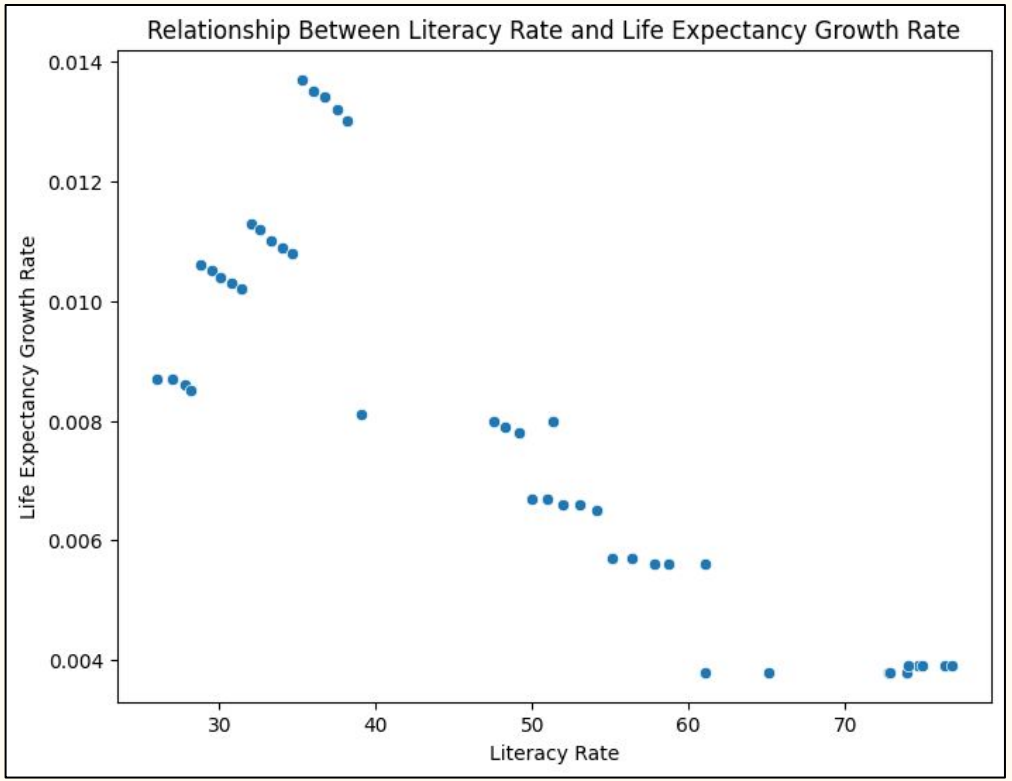
- Misalignment of educational system and job market
- Skill gap and overqualification
- Urban and rural disparity



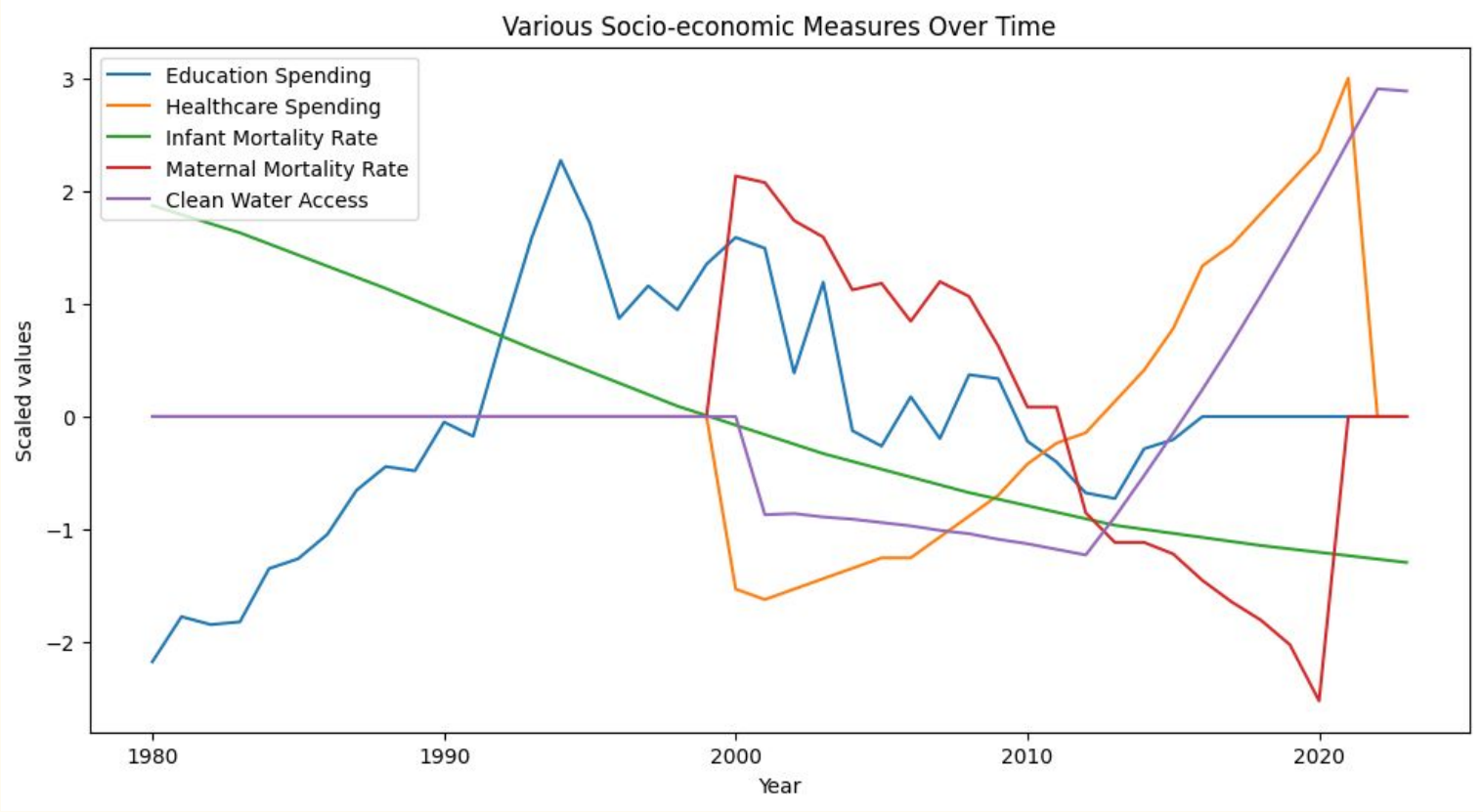
1. Exploratory Data Analysis, cont.

A strong negative relationship between Life Expectancy Growth Rate and Literacy Rate

- Early stage of economic growth
- Improvement in healthcare over education
- Disparities in literacy and healthcare access



1. Exploratory Data Analysis, cont.



2. Data Preparation & Feature Engineering

Highlights:

- Missing data imputation & outlier detection.
- Feature extraction (interaction terms, time-lagged variables).
- Data scaling (Min-Max, Standard Scaler) & transformations.
- Assumptions tested: multicollinearity, normality, stationarity.

Missing data imputation

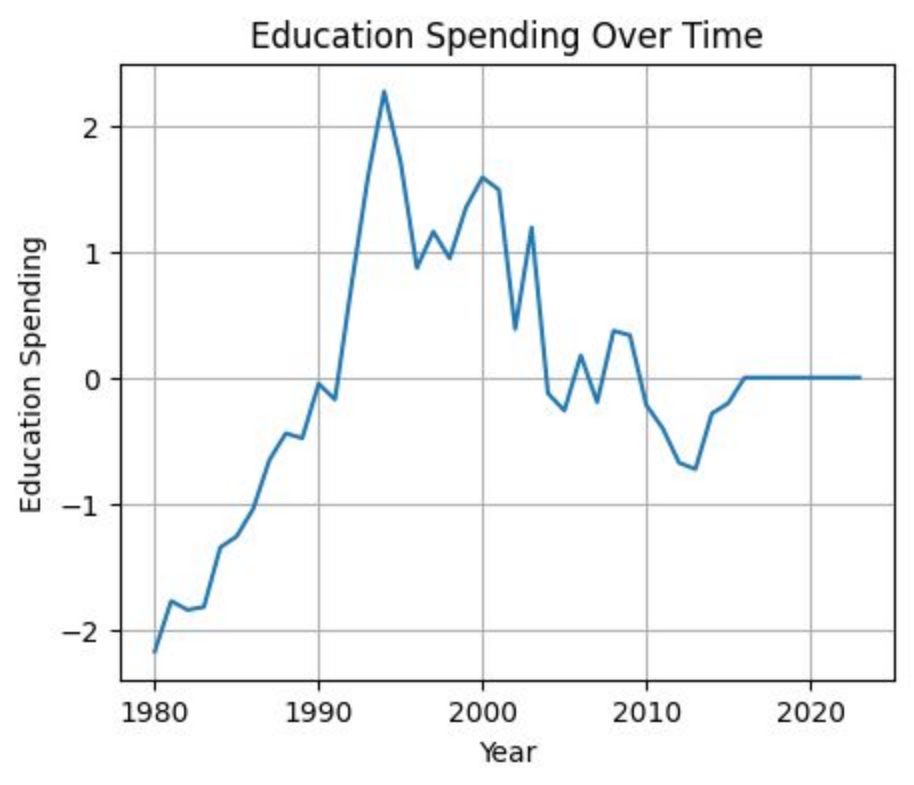
- Out of selected 27 features, several features have significant missing rates. These missing values are imputed with mean.

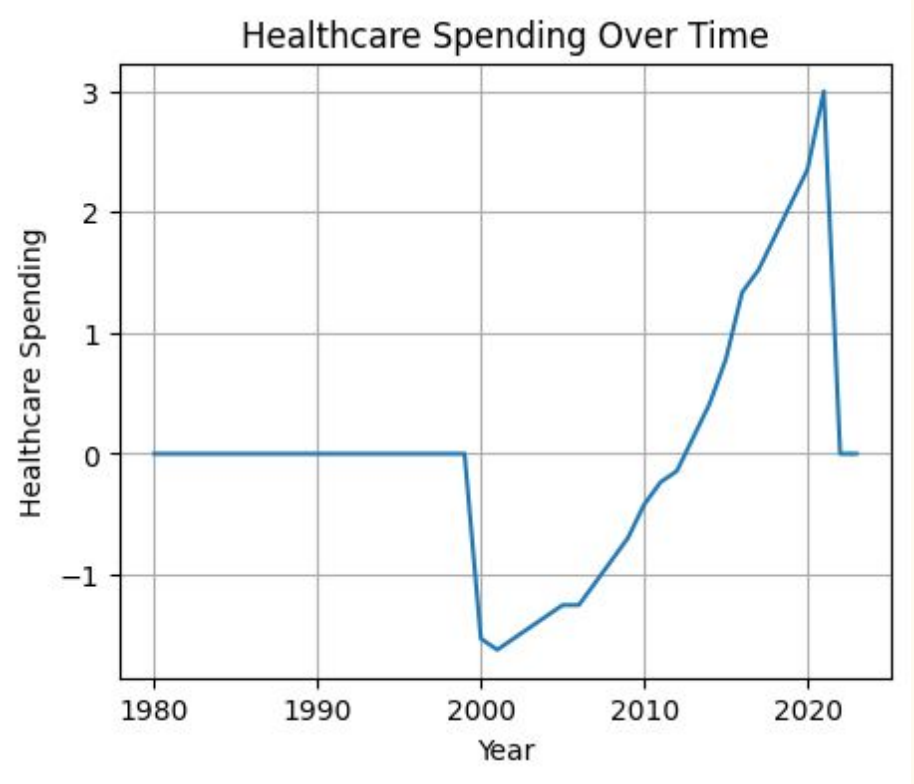
Missing value rate

Year	0.000000
Poverty Rate (National)	0.000000
Rural Poverty Rate	0.000000
Urban Poverty Rate	0.000000
Literacy Rate(%)	0.000000
GDP	0.000000
Education Spending (% of GDP)	18.181818
Healthcare Spending Per Capita (US$)	50.000000
Population	0.000000
Urban Population % of Total	0.000000
Unemployment Rate (%)	0.000000
Inflation Rate (%)	0.000000
Net Migration Rate	2.272727
GDP Per Capita	0.000000
Export Growth(%GDP)	0.000000
Import Growth(%GDP)	0.000000
Death Rate	0.000000
Birth Rate	0.000000
Infant Mortality Rate	0.000000
Life Expectancy Growth Rate (%)	0.000000
Annual % Crime Rate Change	56.818182
green house gas emissions Annual % Change	29.545455
Fossil Fuel consumption % of Total Energy Use	20.454545
Maternal Mortality Rate Per 100K Live Births	52.272727
Clean Water Access % of Population	47.727273
Electricity Access % of Population	27.272727
Suicide Rate	54.545455

dtype: float64

Outliers


Education Spending Over Time
Education Spending
Year
2
1
0
−1
−2
1980
1990
2000
2010
2020


Healthcare Spending Over Time
Healthcare Spending
Year
3
2
1
0
−1
1980
1990
2000
2010
2020

2. Data Preparation & Feature Engineering

Feature Selection & Engineering

Steps	Feature #	
Original dataset	118	Various socio-economic measures including GDP, C02 emission, fuel consumption, and migration rate
Columns of Interest	27	Based on the team's assessment, most relevant features were selected
Create new features (interaction terms)	3	"Urbanization × GDP Growth" = "Urban Population (%)" x "GDP" "3-Year GDP Rolling Avg", "3-Year Literacy Rolling Avg"
Drop multicollinearity	16	"GDP", "GDP Per Capita", "Urban Population (%)", "Urban Poverty Rate", "Rural Poverty Rate","Population", "Fossil Fuel Consumption", etc.
Drop weak correlation	2	"Greenhouse Gas Emissions Change", "Annual Crime Rate Change"
Final dataset	12	'Year', 'National Poverty Rate', 'Education Spending', 'Healthcare Spending', 'Inflation Rate', 'Net Migration Rate', 'Export Growth', 'Life Expectancy Growth Rate', 'Electricity Access', 'Suicide Rate', 'Urbanization × GDP Growth','3-Year Literacy Rolling Avg'

2. Data Preparation & Feature Engineering, cont.

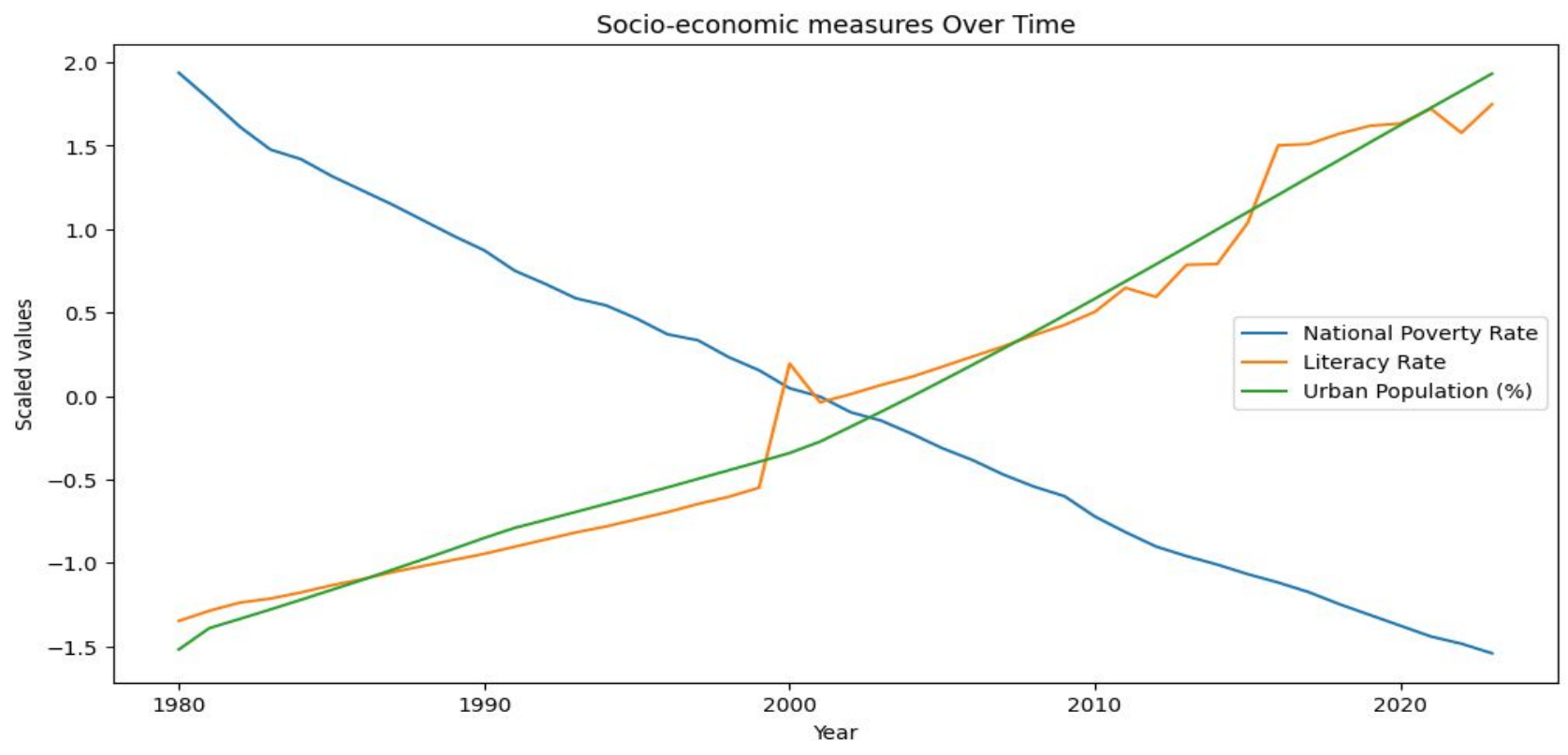
Data Scaling & Assumption Test

- Standardization was performed on each feature to prepare data for predictive modeling
- Following assumptions for time series data are tested
  - Stationarity
  - Multicollinearity

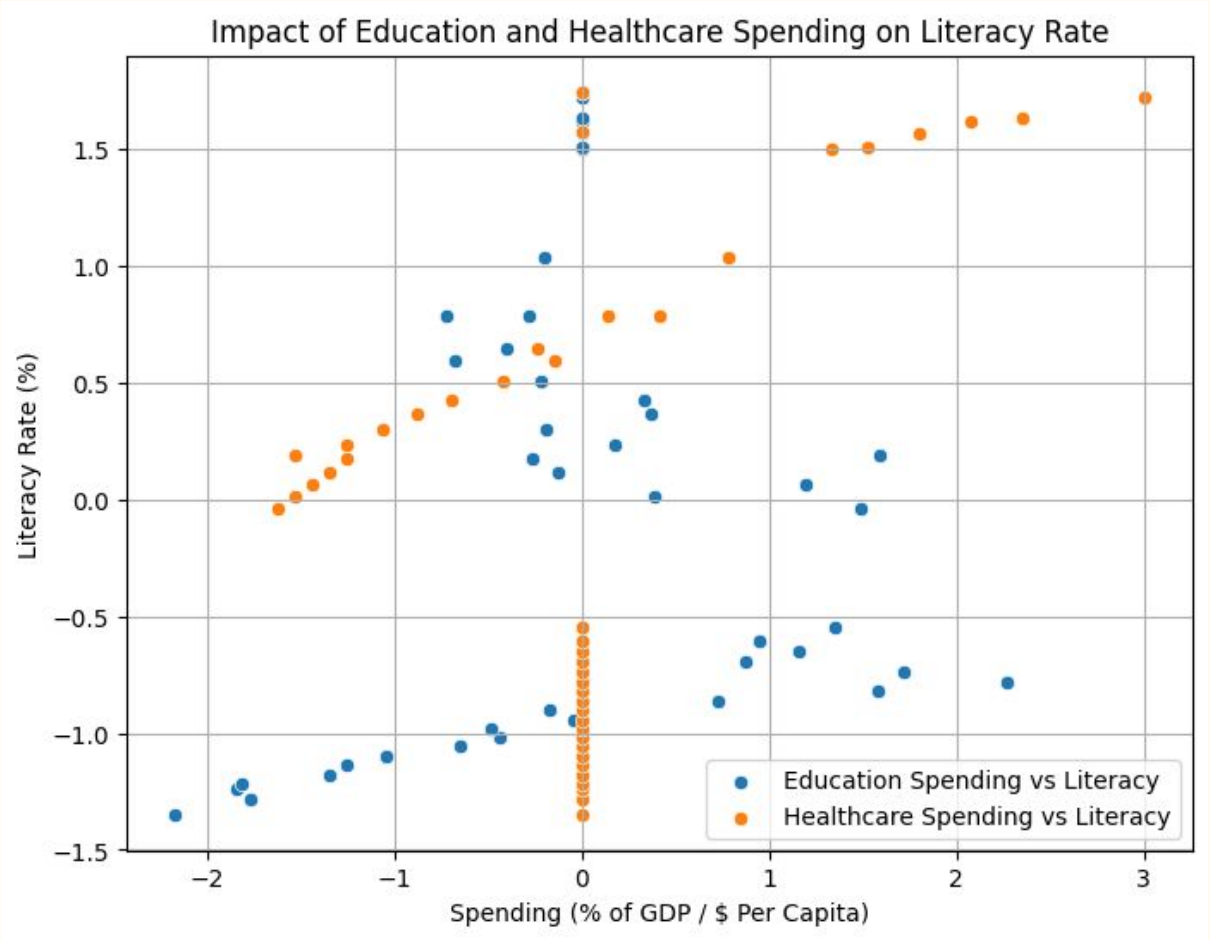
3. Visualizations

- **Trend Analysis**:
  - Line plots for poverty rates, literacy rates, and urbanization over time.
- **Feature Relationships:**
  - Scatterplots for literacy vs. poverty rates, and education/healthcare spending vs. literacy.
- **Correlations:**
  - Heatmap visualizing relationships among all variables, including engineered features.

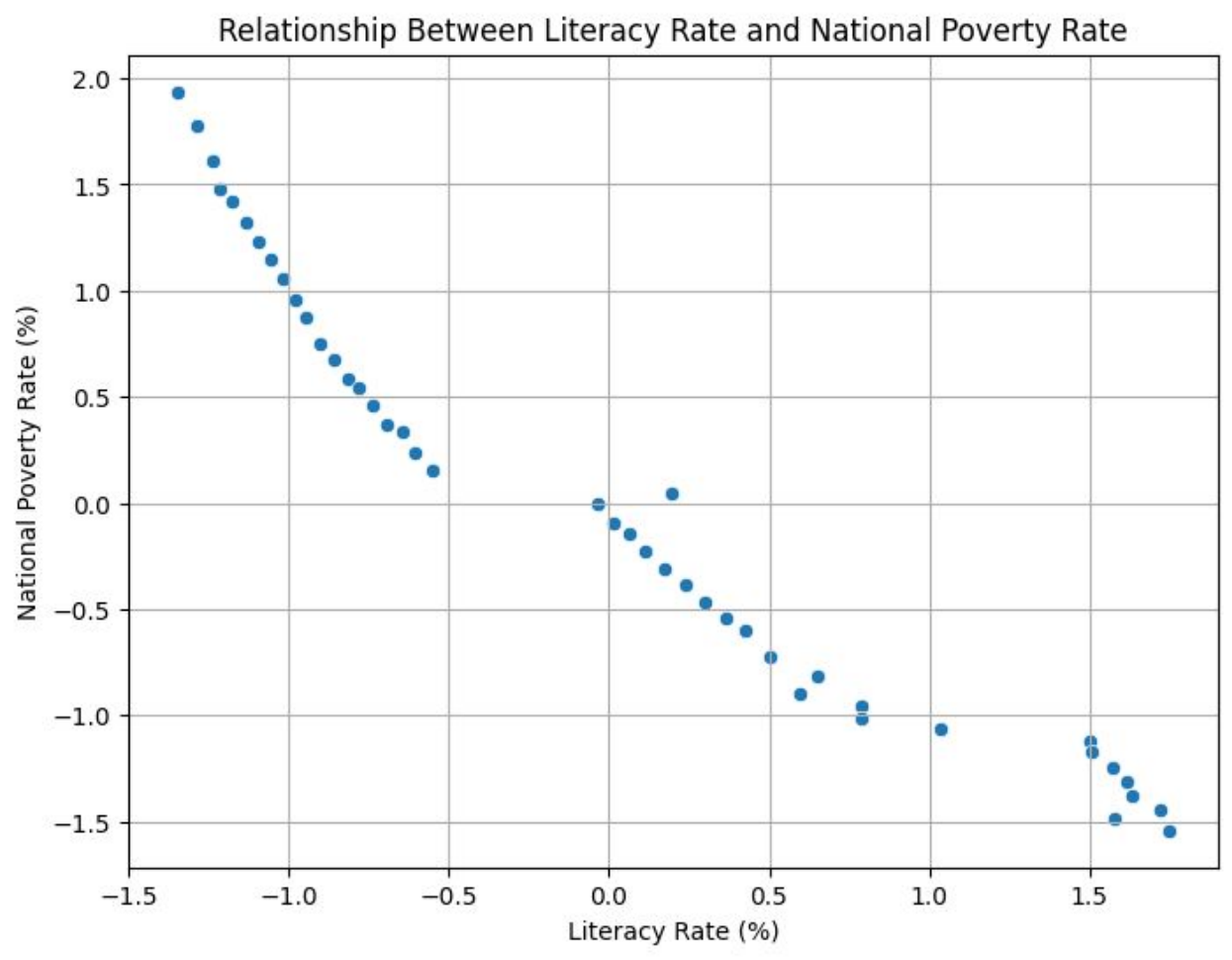
Trend Analysis



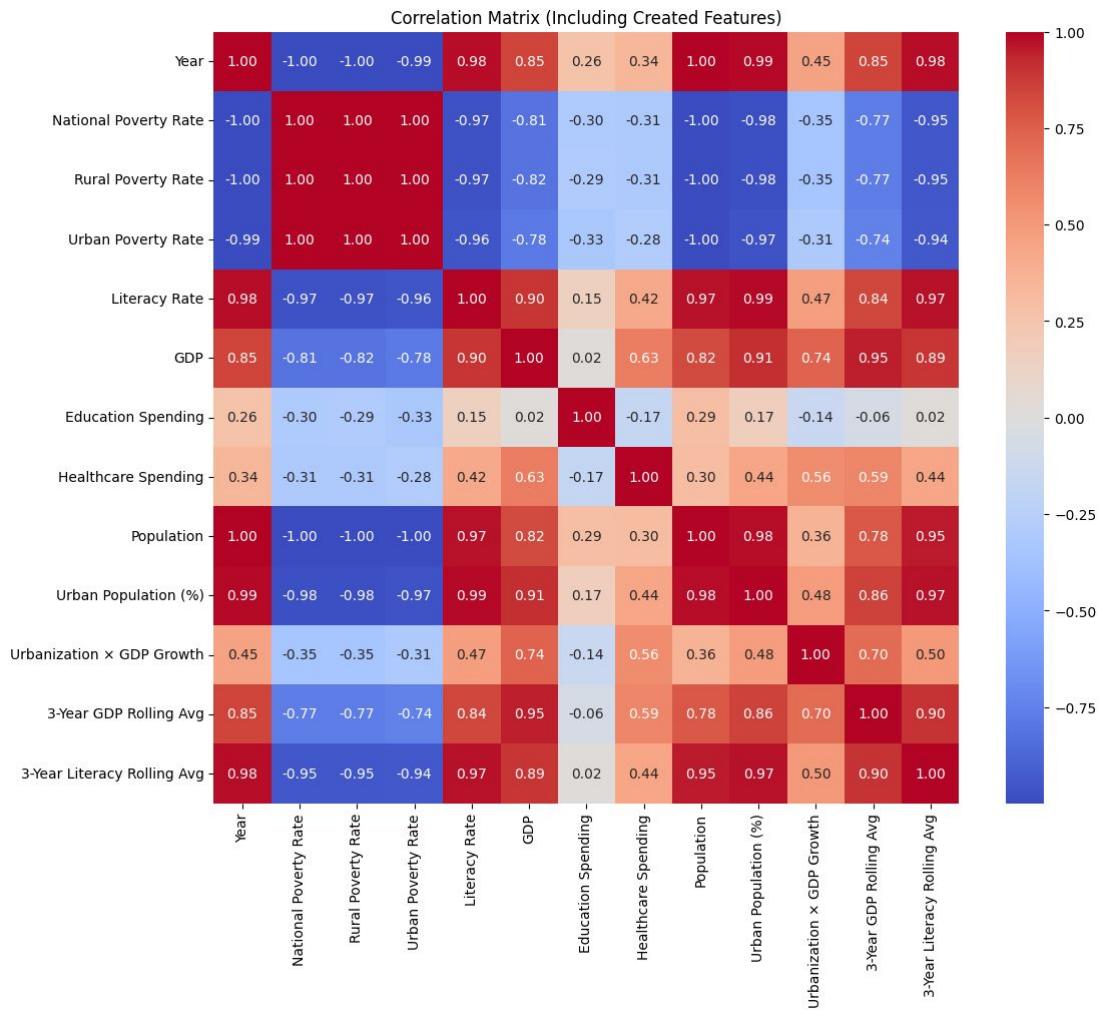
Feature relationships



Feature relationships



Correlation Analysis



4. Four predictive ML models created and metric output

a. Random Forest (RF)

Metric	Train	Test
RMSE	0.021702	0.269561
MAE	0.017697	0.245967
R^2	0.997422	-18.908087

b. Lasso Regression

Metric	Train	Test
MAE	0.0736	3.1801
MSE	0.0085	16.5092
R^2	0.9884	-830.0116

c. Vector Autoregression (VAR)

Metric	Train	Test
RMSE	0.013896	0.022142
MAE	0.011833	0.016789
R^2	0.998376	0.924304

d. Elastic Net Regression

Metric	Train	Test
R^2	0.959	0.938
MSE	0.040	0.060
RMSE	0.201	0.245
MAE	0.150	0.209

5. Findings and Conclusions

- Trends in poverty & literacy in Bangladesh: the analysis of 44 years of Bangladesh's socio-economic data highlights notable progress in poverty reduction and literacy improvement, influenced by economic growth, government policies, and social development programs.
- Challenges of a developing economy: economic fluctuations, income inequality, and rural-urban infrastructural differences continue to impact poverty and literacy rates, requiring sustained policy interventions.
- Predictive features: various ML models identify key independent variables that significantly influence the target variable.
  - Target variable: National poverty rate
  - Key features: Literacy rate, Life expectancy growth rate, Urbanization and GDP growth, Healthcare spending, Net migration rate

- ML Model Conclusion: VAR and Elastic Net models outperform the Lasso Regression and RF models based on the model metric tables detailed in section 4.
    - VAR and Elastic Net models are better suited to handle time series features, multicollinearity and complex data relationships present in the Bangladesh dataset. Lasso regression and RF models struggle with these factors as Lasso regression models are susceptible to multicollinearity between features and RF models are unable to capture temporal dependencies between features in time series datasets.
- Data limitations: the small dataset (44 years of data) restricts the ability to build a more complex and highly accurate predictive model, but it provides valuable insights into reading long-term trends.

- Future improvement: expanding the dataset or incorporating additional external factors could enhance model accuracy and generalization.
  - Incorporate additional years of historical data or external datasets such as global economic indicators or climate data.
  - Utilize real-time and high-frequency data where available (monthly or quarterly data instead of yearly)
  - Feature engineering to identify and include more relevant socio-economic indicators
  - Utilize dimension reduction such as PCA to eliminate redundant features and improve model efficiency.