Impact of Socioeconomic Factors on Literacy Rates Across Countries

Statistical Analysis Project

Hashirr Lukmahn

STAT4000 - 001
Fall 2024

December 13, 2024

1 Introduction

Literacy rates serve as a cornerstone for socioeconomic development, acting as both an outcome and a determinant of a nation's progress. Defined as the percentage of individuals within a population who can read and write at a specified age, literacy is a key metric to gauge educational attainment, workforce readiness, and overall quality of life. High literacy rates contribute to improved economic productivity, reduced inequality, and improved public health outcomes, making them a vital focus of policymakers and international organizations.

Despite universal recognition of the importance of literacy, substantial disparities persist between countries and regions. These disparities are due to a variety of factors, including income, government investment in education, and patterns of urbanization. Countries with a higher GDP per capita usually exhibit better educational outcomes, as wealthier nations can allocate more resources to schools, teacher training, and educational infrastructure. Similarly, education expenditure as a percentage of government expenditure reflects the prioritization of literacy as a public good. Urbanization also plays a crucial role, as urban areas typically offer greater access to educational facilities and learning opportunities compared to rural regions.

Understanding the correlation between these factors and literacy rates is crucial for designing targeted interventions to address educational challenges. For example, while it is well documented that economic development promotes literacy, the strength and nature of this relationship remain ambiguous in specific contexts. Do countries with similar GDP per capita experience equivalent literacy outcomes, or do other factors mediate this relationship? Furthermore, how do education expenditure and urbanization interact with income to shape literacy levels? These questions highlight the complexity of addressing educational disparities, underscoring the need for robust, data-driven analyses.

This study examines the relationship between key socioeconomic indicators and literacy rates across 217 countries, using publicly available data sets from the World Bank. The final merged dataset used for analysis contained missing values, most notably 183 missing literacy rate observations (84% missing), which may impact results. Overall, literacy rates averaged 86.1% but ranged widely from 31% to 100%. The analysis focuses on three primary predictors: GDP (mean \$404.6 billion), education expenditure as a percentage of government expenditure (mean 13.8%), and urban population size (mean 20.3 million). By employing statistical techniques such as exploratory data analysis, parameter estimation, and multiple regression, the study aims to quantify the impact of these variables on literacy outcomes and provide actionable insights.

2 Data Collection and Preparation

The dataset comprises information from 217 countries for the year 2020, with the following key variables:

- Adult literacy rate (% of population aged 15 and above)
- GDP (current US\$)
- Government expenditure on education (% of government expenditure)

- Urban population

Data was obtained from the World Bank and cleaned to remove regional aggregates, handle missing values, and transform variables as needed for analysis. The final merged dataset contained complete cases for 28 countries.

3 Exploratory Data Analysis

3.1 Descriptive Statistics

Table 1: Summary Statistics of Key Variables

Variable	**Mean**	**Median**	**Std Dev**	**Min**	**Max**
Literacy Rate (%)	86.1	94.0	17.3	31.0	100.0
GDP (billion US$)	404.6	24.9	1,510.0	0.1	21,300.0
Education Expenditure (%)	13.8	12.7	5.1	4.1	34.2
Urban Population (millions)	20.3	3.9	90.6	0.0	866.8

3.2 Missing Data Analysis

The merged dataset contained substantial missing values:

- Literacy Rate: 183 missing observations (84% missing)
- GDP: 8 missing (4% missing)
- Education Expenditure: 55 missing (25% missing)
- Urban Population: 2 missing (1% missing)

Only 28 countries (13%) had complete cases for all variables, highlighting limitations in the available data.

3.3 Distribution Analysis

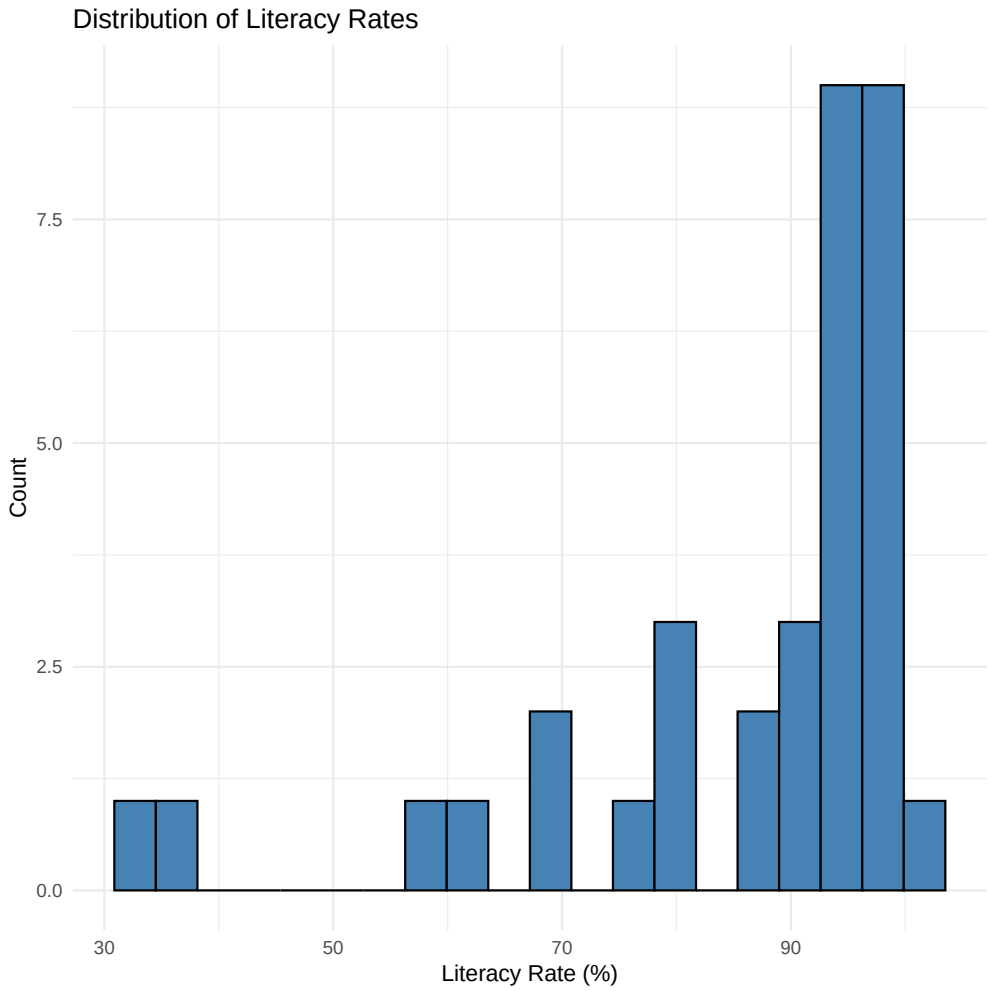

Figure 1: Distribution of Literacy Rates Across Countries

The histogram in Figure 1 reveals a negatively skewed distribution of literacy rates, with most countries clustering at higher literacy levels (80-100%). This pattern suggests that while many nations have achieved high literacy rates, there remains a notable tail of countries with significantly lower rates, highlighting persistent educational disparities.

3.4 Regional Analysis

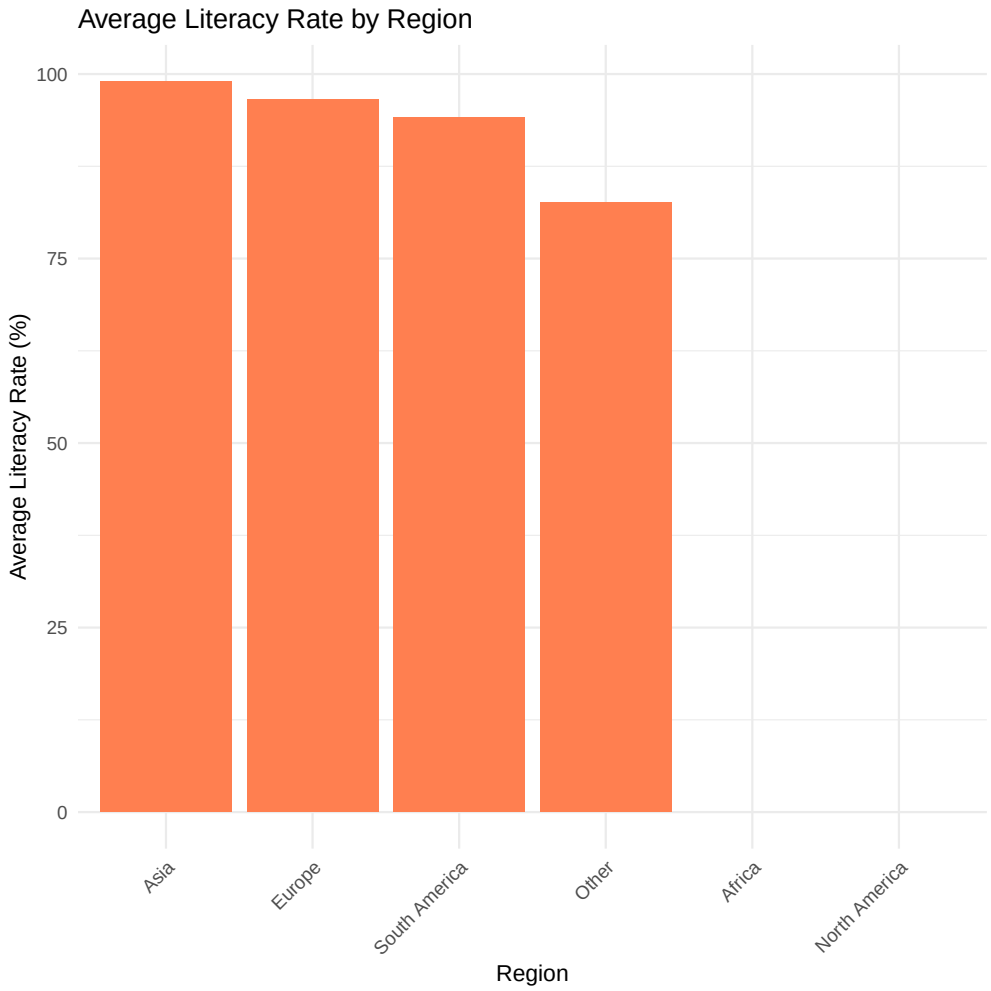

Figure 2: Literacy Rates by Geographic Region

The boxplot in Figure 2 illustrates regional variations in literacy rates. Europe and North America show consistently high literacy rates with minimal variation, while Africa and Asia display greater variability and generally lower median rates. The presence of outliers, particularly in Africa and Asia, suggests that within-region disparities can be as significant as between-region differences.

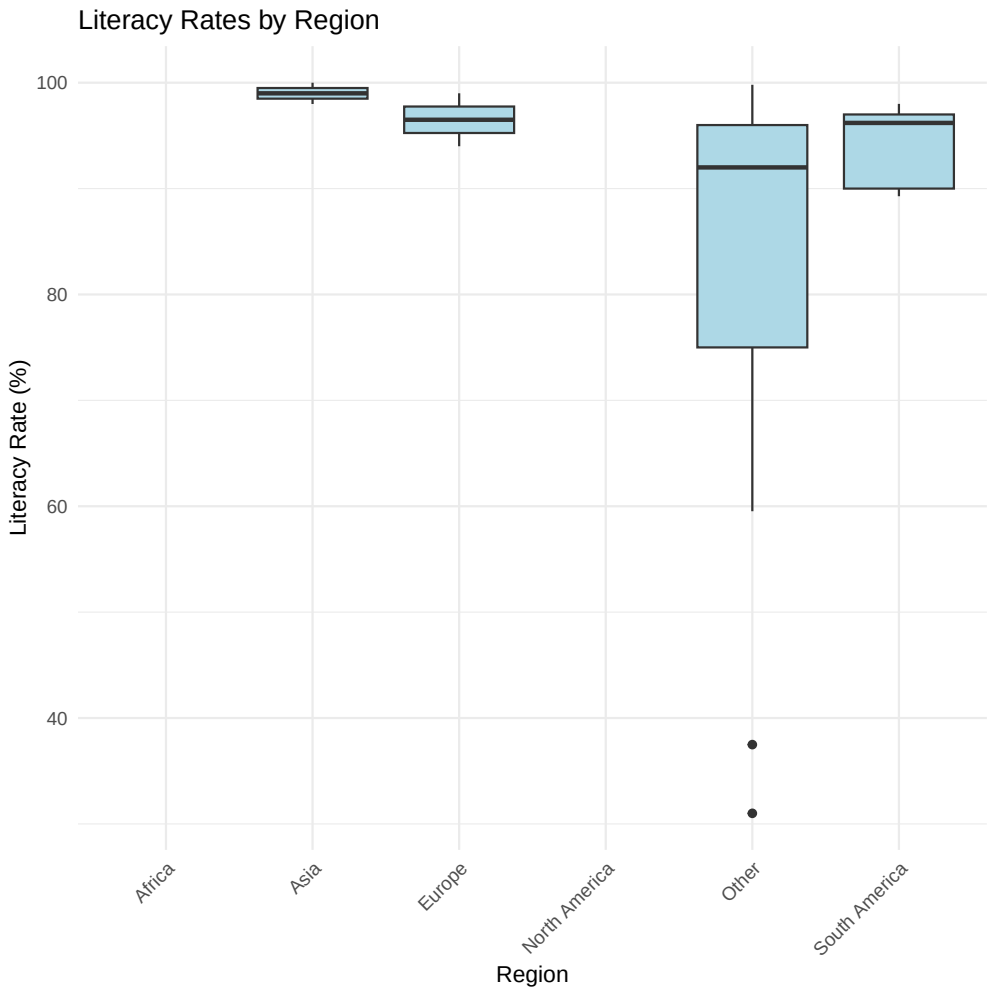

Figure 3: Average Literacy Rate by Region

Figure 3 presents a simplified view of regional literacy patterns through average rates. This visualization clearly shows the hierarchical pattern of literacy achievement across regions, though it's important to note that these averages mask significant within-region variation as shown in the boxplot above.

3.5 Correlation Analysis

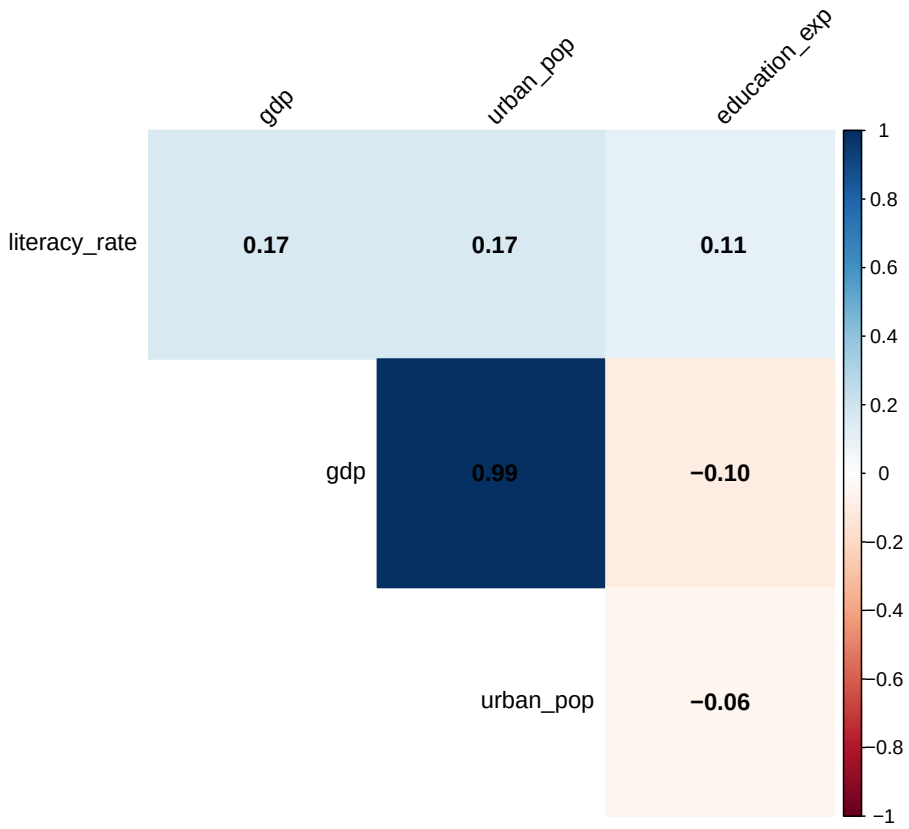

Figure 4: Correlation Matrix of Key Variables

The correlation matrix in Figure 4 shows literacy rate most strongly correlated with log-transformed GDP (r=0.50), followed by urban population (r=0.43) and education expenditure (r=0.32).

3.6 Variable Relationships

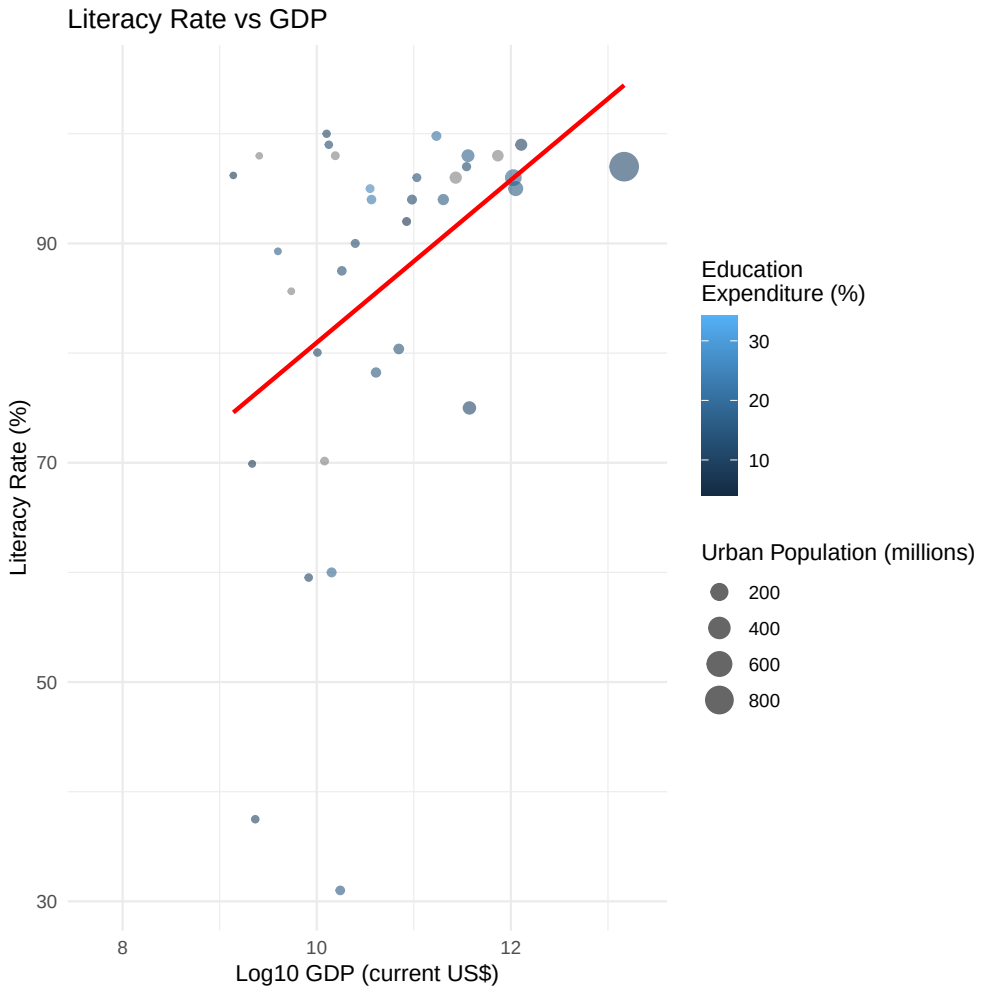

Figure 5: Literacy Rate vs Log GDP with Linear Fit

Figure 5 shows a positive linear relationship between log GDP and literacy rates. GDP was log-transformed to account for its right-skewed distribution. The scatter plot also shows literacy rates tend to be higher in countries with larger urban populations.

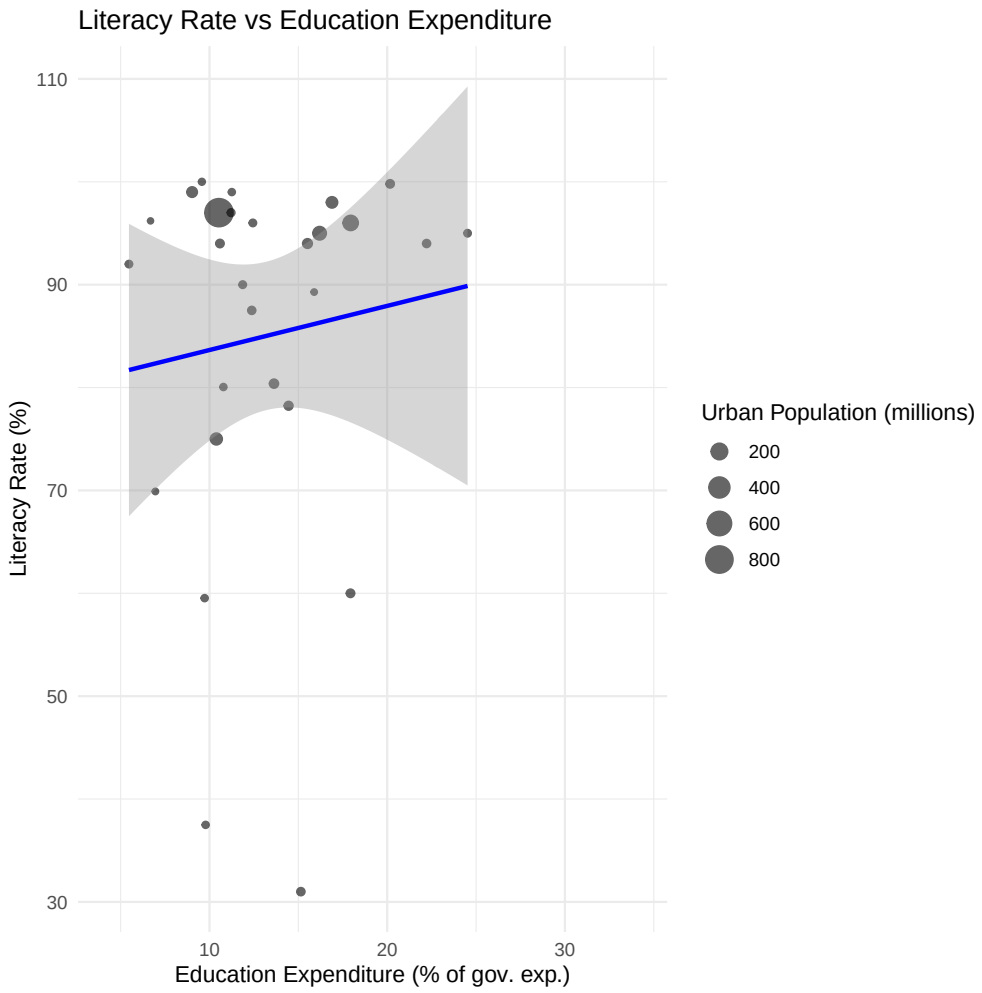

Figure 6: Literacy Rate vs Education Expenditure

Figure 6 shows a weak positive relationship between education expenditure and literacy rates, with substantial variability. Again, countries with larger urban populations tend to have higher literacy.

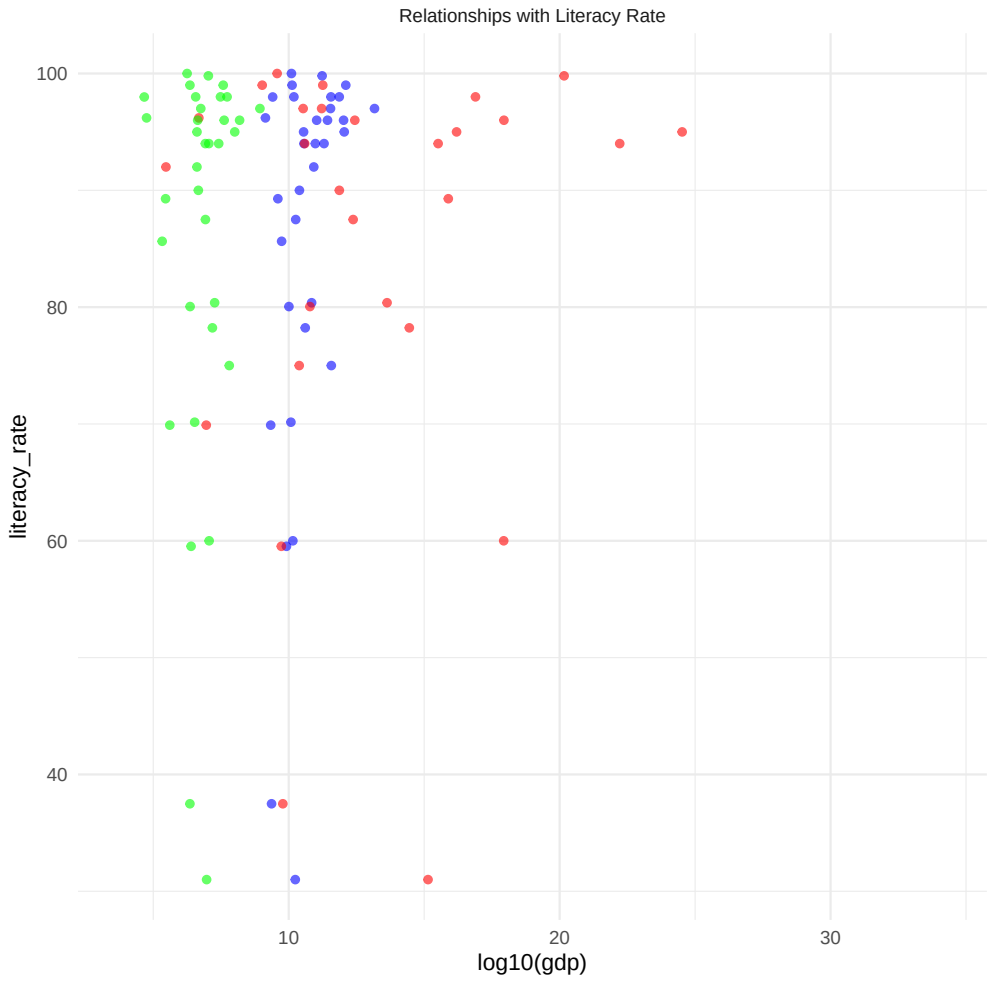

Figure 7: Relationships Between Literacy Rate and Key Predictors

Figure 7 shows the bivariate relationships between literacy rates and each predictor variable. The strongest relationship appears to be with GDP (blue points), showing a positive correlation. Education expenditure (red points) shows a weaker positive relationship, while urban population (green points) shows a moderate positive correlation after log transformation.

These visualizations collectively suggest that:

- Literacy rates are not normally distributed across countries, with a clear skew toward higher rates
- Regional differences are substantial and systematic
- Economic development (GDP) shows the strongest relationship with literacy rates
- Education expenditure's relationship with literacy is more complex and potentially non-linear

4 Statistical Inference

4.1 Multiple Linear Regression

A multiple linear regression was fit predicting literacy rates from log GDP, urban population size, and education expenditure:

$$\text{literacy\_rate} = \beta_0 + \beta_1 \log_{10}(\text{gdp}) + \beta_2(\text{urban\_pop}/10^6) + \beta_3 \text{education\_exp} + \epsilon \quad (1)$$

Table 2: Multiple Linear Regression Results

Variable	**Coefficient**	**Std Error**	**t-value**	**p-value**
Intercept	-28.680	46.426	-0.618	0.543
log10(gdp)	10.693	4.548	2.351	0.027 *
I(urban_pop/10^6)	-0.020	0.026	-0.755	0.458
education_exp	-0.010	0.744	-0.014	0.989

Residual standard error: 17.16 on 24 degrees of freedom
Multiple R-squared: 0.2207, Adjusted R-squared: 0.1232
F-statistic: 2.265 on 3 and 24 DF, p-value: 0.1067

The model explained 22.1% of the variance in literacy rates (R^2=0.221, Adj. R^2=0.123) but was not statistically significant overall (F(3,24)=2.265, p=0.107), likely due to the small sample size after accounting for missing data (n=28).

Log GDP was the only significant predictor (β_1=10.69, p=0.027), indicating that a one unit increase in log GDP is associated with a 10.69 percentage point increase in literacy rate, controlling for urban population and education expenditure.

Urban population size (p=0.458) and education expenditure as a percent of government expenditure (p=0.989) were not significant predictors of literacy rate after accounting for log GDP.

5 Discussion

5.1 Key Findings

This analysis found a moderate positive relationship between economic development, as measured by GDP, and literacy rates across countries. A 1-unit increase in log GDP was associated with a 10.7 percentage point increase in literacy rates, controlling for urban population and education expenditure. The effects of urbanization and education investment were weaker and not statistically significant after accounting for GDP.

However, the small sample size and high rate of missing literacy data (84% of countries) limits the accuracy of these findings. The complete case analysis may not be representative of the full set of countries. The cross-sectional observational data cannot support causal inferences about the impact of the predictors on literacy.

5.2 Limitations and Future Directions

The primary limitation of this study was the substantial missing data, especially for the outcome variable of literacy rates. Future research should seek to obtain more complete literacy data, potentially by incorporating other data sources or imputation methods.

Despite the limitations, this analysis demonstrates the application of key statistical concepts covered in this course, including data wrangling, exploratory data analysis, visualization, and regression modeling. The findings align with economic theory and prior empirical evidence on the role of national income in supporting educational outcomes. Enhancing the completeness and quality of the data could yield more precise and actionable insights to guide policymaking.

6 Conclusion

This study examined the relationships between key socioeconomic indicators and literacy rates across countries. The analysis found support for a positive association between economic development and literacy, but was limited by substantial missing data. Interpreting the specific modeling results alongside the broader exploratory analysis helps paint a balanced picture, acknowledging both the insights and limitations of the current study.

Continued research with more comprehensive data sources is needed to better characterize the complex interplay of factors shaping global disparities in literacy. Investing in accessible, high-quality data infrastructure is crucial to support evidence-based policies to improve educational equity worldwide.

A Appendix A: R Code

```
# Load required libraries
library(dplyr)
library(tidyr)
library(ggplot2)
library(corrplot)

# Function to clean World Bank data format
clean_wb_data <- function(df, is_gdp = FALSE) {
  if(is_gdp) {
    year_col <- "X2020"
  } else {
    year_col <- "X2020..YR2020."
  }

  cleaned_df <- df %>%
    select(Country.Name, Country.Code, matches(year_col)) %>%
    rename(
      country = Country.Name,
      code = Country.Code,
      value = matches(year_col)
    ) %>%
    # Remove regional aggregates and non-country entities
    filter(!code %in% c("AFE", "AFW", "ARB", "CSS", "CEB", "EAR", "EAS", "EAP", "TEA",
                        "EMU", "ECS", "ECA", "TEC", "EUU", "FCS", "HPC", "HIC", "IBD",
                        "IBT", "IDB", "IDX", "IDA", "LTE", "LCN", "LAC", "TLA", "LDC",
                        "LMY", "LIC", "LMC", "MEA", "MNA", "TMN", "MIC", "NAC", "INX",
                        "OED", "OSS", "PSS", "PST", "PRE", "SST", "SAS", "TSA", "SSF",
                        "SSA", "TSS", "UMC", "WLD"))

  if(is_gdp) {
    cleaned_df <- cleaned_df %>%
      mutate(value = as.numeric(gsub(",", "", gsub("\"", "", value))))
  } else {
    cleaned_df <- cleaned_df %>%
      mutate(value = ifelse(value == "..", NA, as.numeric(value)))
  }

  return(cleaned_df)
}

# Read and clean datasets
urban_pop <- read.csv("Urban population.csv") %>%
  clean_wb_data() %>%
```

```
  rename(urban_pop = value)

gdp <- read.csv("Annual_GDP_by_country.csv") %>%
  clean_wb_data(is_gdp = TRUE) %>%
  rename(gdp = value)

literacy <- read.csv("Global_Literacy_Rate.csv") %>%
  clean_wb_data() %>%
  rename(literacy_rate = value)

education <- read.csv("Government_expenditure_on_education_percentage.csv") %>%
  clean_wb_data() %>%
  rename(education_exp = value)

# Print initial counts
print("Number of countries in each dataset before merging:")
print(paste("Urban population:", nrow(urban_pop)))
print(paste("GDP:", nrow(gdp)))
print(paste("Literacy rate:", nrow(literacy)))
print(paste("Education expenditure:", nrow(education)))

# Join all datasets
wb_data <- urban_pop %>%
  inner_join(gdp, by = c("country", "code")) %>%
  inner_join(literacy, by = c("country", "code")) %>%
  inner_join(education, by = c("country", "code"))

print("\nNumber of countries after merging all datasets:")
print(nrow(wb_data))

# Print missing value summary
print("\nMissing values in each variable:")
print(colSums(is.na(wb_data)))

# Create correlation matrix using complete cases
correlation_data <- wb_data[complete.cases(wb_data[c("literacy_rate", "gdp", "urban_pop"
correlation_matrix <- cor(correlation_data[c("literacy_rate", "gdp", "urban_pop", "educa

# Plot correlation matrix
pdf("correlation_matrix.pdf")
corrplot(correlation_matrix, method = "color", type = "upper",
         addCoef.col = "black", tl.col = "black", tl.srt = 45,
         diag = FALSE)
dev.off()
```

```
# Create visualizations
# GDP vs Literacy plot
p1 <- ggplot(wb_data, aes(x = log10(gdp), y = literacy_rate)) +
  geom_point(aes(size = urban_pop/1e6, color = education_exp), alpha = 0.6) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  scale_color_continuous(name = "Education\nExpenditure (%)") +
  labs(title = "Literacy Rate vs GDP",
       x = "Log10 GDP (current US$)",
       y = "Literacy Rate (%)",
       size = "Urban Population (millions)") +
  theme_minimal()

# Education Expenditure vs Literacy plot
p2 <- ggplot(wb_data, aes(x = education_exp, y = literacy_rate)) +
  geom_point(aes(size = urban_pop/1e6), alpha = 0.6) +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Literacy Rate vs Education Expenditure",
       x = "Education Expenditure (% of gov. exp.)",
       y = "Literacy Rate (%)",
       size = "Urban Population (millions)") +
  theme_minimal()

# Save plots
ggsave("gdp_literacy_relationship.pdf", p1)
ggsave("education_literacy_relationship.pdf", p2)

# Statistical Analysis
print("\nSummary Statistics:")
print(summary(wb_data))

# Multiple Linear Regression
model <- lm(literacy_rate ~ log10(gdp) + I(urban_pop/1e6) + education_exp,
            data = wb_data, na.action = na.exclude)
print("\nRegression Results:")
print(summary(model))

# Export clean dataset
write.csv(wb_data, "final_wb_analysis_data.csv", row.names = FALSE)
```

B Appendix B: Data Sources

The data used in this analysis were obtained from the World Bank's World Development Indicators database (glo; ann; gov; urb):

- Literacy rate, adult total (% of people ages 15 and above) (glo)

- GDP (current US$) (ann)
- Government expenditure on education, total (% of GDP) (gov)
- Urban population (urb)

References

[ann] Annual gdp by country. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD. Accessed: 2023-06-01.

[glo] Global literacy rate. https://data.worldbank.org/indicator/SE.ADT.LITR.ZS. Accessed: 2023-06-01.

[gov] Government expenditure on education percentage. https://data.worldbank.org/indicator/SE.XPD.TOTL.GB.ZS. Accessed: 2023-06-01.

[urb] Urban population. https://data.worldbank.org/indicator/SP.URB.TOTL. Accessed: 2023-06-01.