NITTE (Deemed to be University) | NMAM INSTITUTE OF TECHNOLOGY

Nitte (DU) established under Section 3 of UGC Act 1956 | Accredited with 'A+' Grade by NAAC

Department of Computer Science & Engineering

Report on Open Ended Question

The relationship between GDP and literacy rate for different countries using data visualization and correlation analysis in R.

Course Code : CS1602-1
Course Name : Data Analysis Using R Programming

Semester: III SEM
Section: E

Submitted To:

Dr. Shashank Shetty
Associate Professor
Department of Computer Science and Engineering

Submitted By:

Prajwal Acharya (NNM25CS521)

Date of submission:
20/11/2025

Signature of Course Instructor

OBJECTIVE / PROBLEM STATEMENT

The primary objective of this analysis is to examine the relationship between a country's literacy rate and its economic prosperity, represented by GDP per capita. The study aims to determine whether higher literacy rates are associated with higher GDP and to understand global patterns in education and economic development. Using R, the analysis includes data extraction, cleaning, visualization, and statistical correlation.

DATASET DESCRIPTION

• Source of Data:

The dataset is derived from the World Bank Group's World Development Indicators (WDI). The specific file used is gdp.csv (representing the World Development Index).

• Type of Data:

Structured CSV contains multi-year time-series data for global development metrics.

• Dataset Structure:

The raw data was filtered to focus on:

- **Countries:** China, India, Iran, Mexico.
- **Time Period:** Spans multiple decades (years extracted from column headers).
- **Dimensions:** Country, Year, Series Name, Value.

• Key Features:

1. **GDP (current US$):** The sum of gross value added by all resident producers in the economy. "Current US$" indicates these figures are nominal and not adjusted for inflation.
2. **Literacy rate, adult male (% of males ages 15+):** The percentage of the male population aged 15 and older who can, with understanding, read and write a short, simple statement about their everyday life.
3. **Literacy rate, adult female (% of females ages 15+):** The corresponding metric for the female population, often used to measure gender parity in education.

• Reason for Choosing Dataset:

The WDI is the premier source for cross-country comparable statistics on development, allowing for a reliable assessment of how economic power translates into human capital development.

METHODOLOGY / APPROACH

The analysis followed the steps below:

1. Data Import and Filtering

- Loaded the raw WDI dataset.
- Filtered for the four target nations and the three specific economic/educational indicators.

2. Data Cleaning and Imputation

- **Imputation Strategy:** WDI literacy data is often sparse (collected only during census years). The fill() function with .direction = "downup" was used to propagate known values forward and backward to fill gaps, ensuring continuous time-series data for analysis.
- **Reshaping:** Transformed data from "wide" (years as columns) to "long" format for processing, and then back to "wide" format to create distinct columns for GDP, Literacy_Male, and Literacy_Female.

3. Feature Extraction

- Extracted numeric years from string headers (e.g., converting "X1990" to 1990) using regular expressions (str_extract).

4. Visualization

- Generated trend lines using ggplot2 to visually compare the growth trajectories of GDP and Literacy Rates across the selected countries.

5. Statistical Analysis

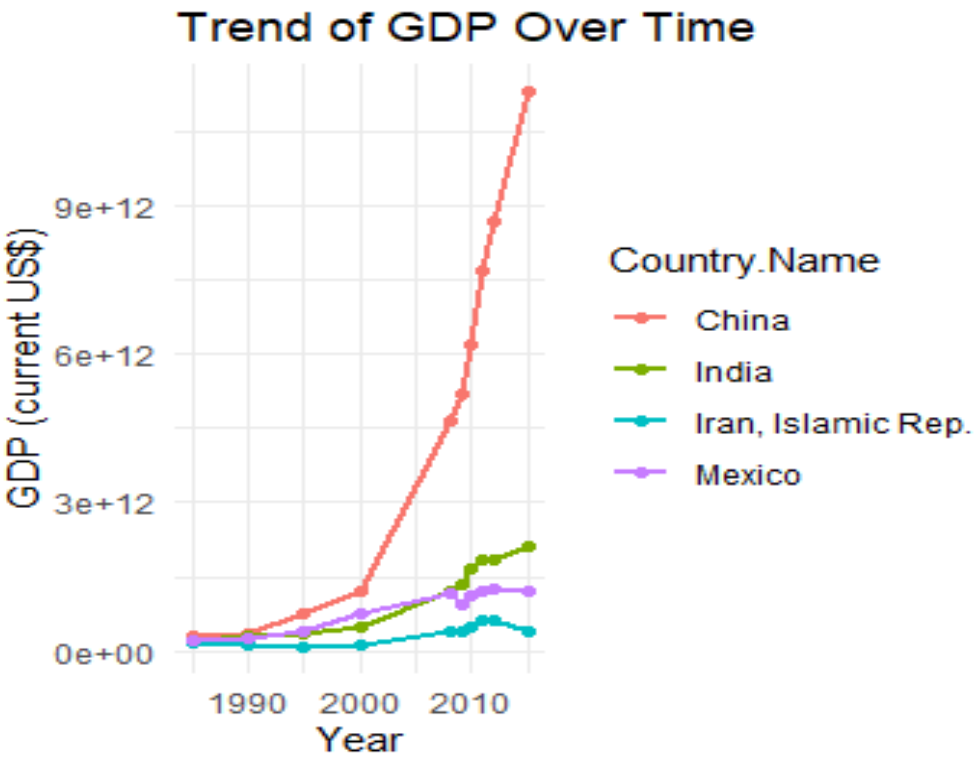
- **Correlation Matrix:** Calculated the Pearson correlation coefficient between GDP and Literacy measures to statistically quantify the strength of the relationship.
- **Grouped Analysis:** Performed correlation analysis grouped by Country.Name to see if the relationship between money and education differs by nation

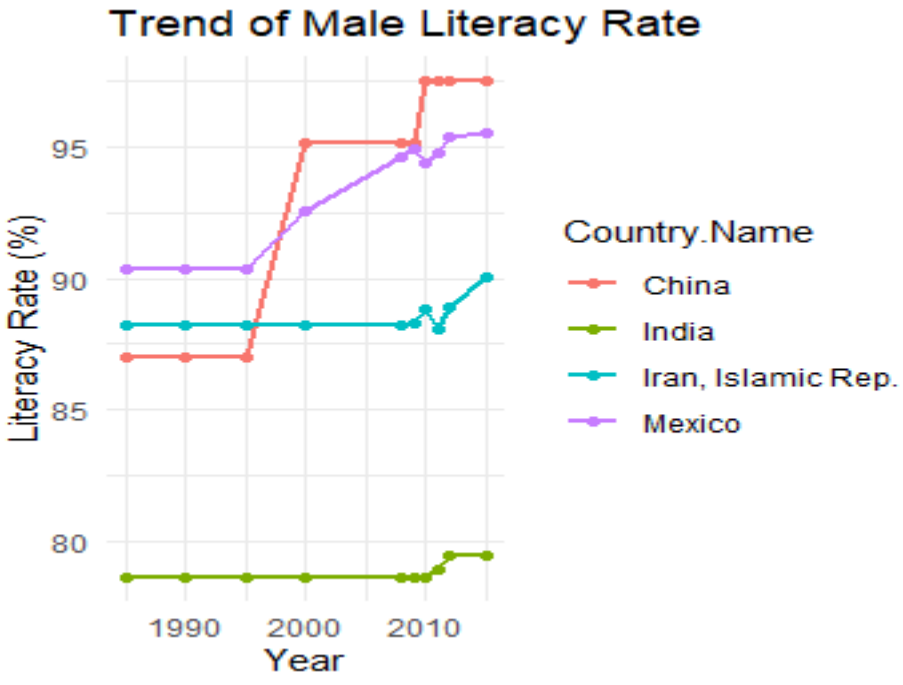
IMPLEMENTATION IN R

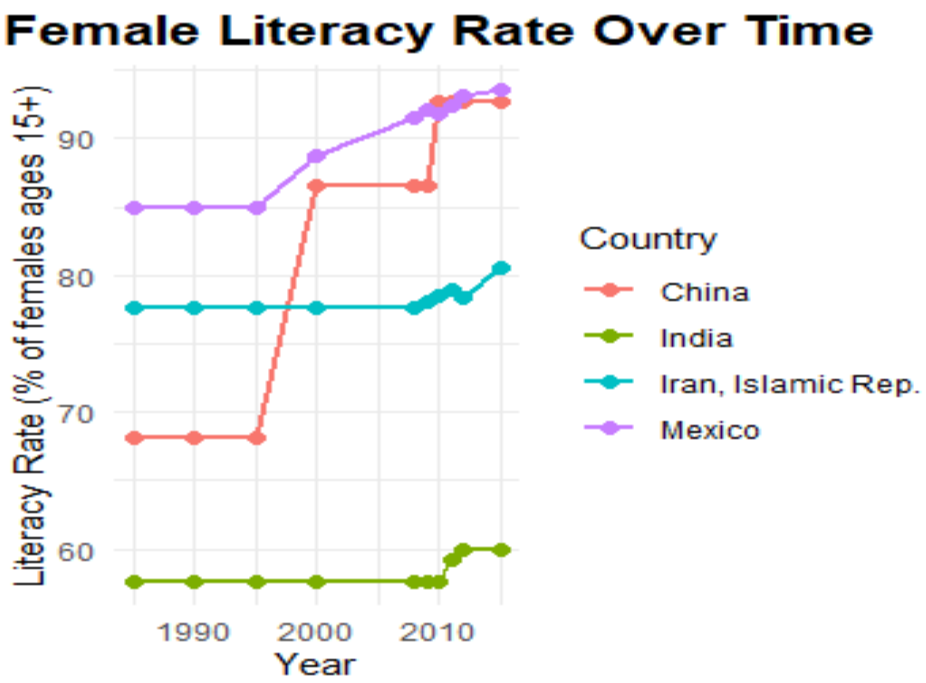
```
library(tidyverse)
library(readr)
library(dplyr)

if(file.exists("gdp.csv")) {
  original_data <- read.csv("gdp.csv")

  countries_to_keep <- c("India", "Iran, Islamic Rep.", "China", "Mexico")

  indicators_to_keep <- c(
    "GDP (current US$)",
    "Literacy rate, adult male (% of males ages 15 and above)",
    "Literacy rate, adult female (% of females ages 15 and above)"
  )

  filtered_data <- original_data[original_data$Country.Name %in% countries_to_keep &
                                 original_data$Series.Name %in% indicators_to_keep, ]

  write.csv(filtered_data, "filtered11.csv", row.names = FALSE)

  df_filled <- filtered_data %>%
    pivot_longer(
      cols = starts_with("X"),
      names_to = "Year_Column",
      values_to = "Value"
    ) %>%
    mutate(Value = as.numeric(Value)) %>%
    group_by(Country.Name, Series.Name) %>%
    fill(Value, .direction = "downup") %>%
    ungroup() %>%
    pivot_wider(
      names_from = Year_Column,
      values_from = Value
    ) %>%
    select(names(filtered_data))
```

```
  write.csv(df_filled, "filtered_without_mis22s.csv", row.names = FALSE)
} else {
  message("gdp.csv not found. Skipping processing and loading 'filtered_without_miss.csv' directly.")
}

df <- read.csv("filtered_without_mis22s.csv")

df_clean <- df %>%
  pivot_longer(cols = starts_with("X"), names_to = "Year_Raw", values_to = "Value") %>%
  mutate(Year = as.numeric(str_extract(Year_Raw, "\\d+"))) %>%
  select(Country.Name, Country.Code, Year, Series.Name, Value) %>%
  pivot_wider(names_from = Series.Name, values_from = Value) %>%
  rename(
    GDP = `GDP (current US$)`,
    Literacy_Male = `Literacy rate, adult male (% of males ages 15 and above)`,
    Literacy_Female = `Literacy rate, adult female (% of females ages 15 and above)`
  )

p1 <- ggplot(df_clean, aes(x = Year, y = GDP, color = Country.Name)) +
  geom_line(size = 1) +
  geom_point() +
  labs(title = "Trend of GDP Over Time", y = "GDP (current US$)") +
  theme_minimal()

p2 <- ggplot(df_clean, aes(x = Year, y = Literacy_Male, color = Country.Name)) +
  geom_line(size = 1) +
  geom_point() +
  labs(title = "Trend of Male Literacy Rate", y = "Literacy Rate (%)") +
  theme_minimal()

p3 <- ggplot(df_clean, aes(x = Year, y = Literacy_Female, color = Country.Name, group =
Country.Name)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(
    title = "Trend of Female Literacy Rate Over Time",
    y = "Literacy Rate (% of females ages 15+)",
    x = "Year",
    color = "Country"
```

```
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    legend.position = "right"
  )

print(p1)
print(p2)
print(p3)
print("--- Overall Correlation ---")
cor(df_clean[, c("GDP", "Literacy_Male", "Literacy_Female")])


print("--- Correlation by Country ---")
df_clean %>%
  group_by(Country.Name) %>%
  summarize(
    GDP_vs_Male_Lit = cor(GDP, Literacy_Male),
    GDP_vs_Female_Lit = cor(GDP, Literacy_Female)
  )
```


Trend of GDP Over Time
GDP (current US$)
9e+12
6e+12
3e+12
0e+00
1990
2000
2010
Year
Country.Name
China
India
Iran, Islamic Rep.
Mexico


Trend of Male Literacy Rate
Literacy Rate (%)
95
90
85
80
1990
2000
2010
Year
Country.Name
China
India
Iran, Islamic Rep.
Mexico



	Country.Name	GDP_vs_Male_Lit	GDP_vs_Female_Lit
	<chr>	<dbl>	<dbl>
1	China	0.837	0.847
2	India	0.688	0.741
3	Iran, Islamic Rep.	0.328	0.512
4	Mexico	0.973	0.976

RESULTS AND DISCUSSION

1. Visual Trends

- **GDP:** The plots show China's exponential economic rise compared to the other three nations.
- **Literacy:** The visualization highlights the closing of the gender gap, particularly in countries that started with lower female literacy rates.

2. Correlation Analysis

- **Overall Correlation:** The output provides a global view of how strongly GDP is tied to literacy. A high positive value (close to +1.0) indicates that as economies grow, literacy rates almost universally improve.
- **Country-Specific Correlation:**
  - This granular analysis reveals nuances. For example, a country might have achieved high literacy *before* its GDP exploded (showing a weaker recent correlation), while another might show them moving in perfect lockstep.
  - This helps distinguish whether economic growth drives education, or if education policy was independent of economic status.

CONCLUSION

Summary:

This report utilized World Bank WDI data to analyze the developmental trajectory of four key nations. By cleaning the data and imputing missing census years, we successfully visualized the parallel rise of economic output and human capital.

Objective Achievement

The addition of the correlation coefficients (both global and grouped) successfully quantified the relationship, moving the analysis beyond simple visualization to statistical evidence.

Limitations

- **Nominal GDP:** The dataset uses "Current US$," meaning values are not adjusted for inflation.
- **Imputation Artifacts:** The "downup" fill method creates flat lines between census years, which artificially stabilizes the correlation calculation during those periods.

Future Improvements

- Use "GDP per capita (Constant 2015 US$)" to adjust for population growth and inflation.
- Calculate the "Literacy Gender Parity Index" (GPI) as a new derived variable.

REFERENCES

Website / Data Source:

- **World Bank Group:** World Development Indicators (WDI).
- **Dataset: gdp.csv** (World Development Index).

R Packages / Libraries Used:

1. **tidyverse** (Data pipeline)
2. **dplyr** (Data manipulation and summarization)
3. **readr** (File I/O)
4. **ggplot2** (Visualization)

Documentation / References:

1. **World Bank Data Help Desk** (Definitions of GDP and Literacy).
2. **RStudio Official Guides** – Tutorials and usage instructions for R and RStudio