Exploratory of CO2 with Socio-economic Aspects

Matthew Anthony Tjahjadi, Harry Santosa, Wilbert Suwanto

2024-06-22

Introduction

Overview

This report presents the findings of an exploratory data analysis (EDA) focused on CO2 emissions and their relationship with various socio-economic factors. The primary aim is to uncover patterns and correlations that can inform efforts to mitigate climate change and promote sustainable development.

Context, Significance, Releavannce and Applications

CO2 emissions play a major role in climate change, impacting global tempertaures, weather patterns and ecosystems. Understanding the socio-economic factors that affect CO2 emissions is crucial for developing strategies to reduce greenhouse gas emissions, which will overall improve the quality of life within these countries. This analysis provides insights into how different countries contribute to CO2 emissions and the socio-economic variables associated with these emissions.

Objectives and Questions

The main objectives of this analysis are:

- To identify significant socio-economic factors that correlate with CO2 emissions.
- To explore how different countries and regions compare in terms of CO2 emissions and related socio-economic indicators.
- To provide insights that can help governments or policymakers on actions to take to reduce CO2 emissions

With those goals in mind, there are a few key questions we hope to answer by the end of this analysis:

- What socio-economic factors are most strongly associated with CO2 emissions?
- How do CO2 emissions vary across different regions and economic contexts?
- What trends can be observed in the relationship between economic development and CO2 emissions?

Analytical Techniques

The analysis employs various exploratory data analysis techniques, including:

- Descriptive statistics to summarize key variables related to CO2 emissions.
- Correlation analysis to identify relationships between CO2 emissions and socio-economic factors.

- Data visualization methods to graphically represent the trends and patterns in the data.

The insights derived from this analysis are crucial for addressing global climate challenges and formulating effective policies to reduce CO2 emissions. This analysis is personally meaningful as it contributes to a broader understanding of the factors driving climate change, which is essential for promoting sustainable development and international cooperation.

Data description

Data source

The dataset used for this exploratory data analysis was sourced from Kaggle. It provides a comprehensive overview of various socio-economic and demographic attributes of countries around the world as of 2023. The dataset can be accessed here.

Variables

The dataset includes a wide range of variables that provide insights into the socio-economic and demographic characteristics of different countries. However we will only take key variables relevant to our analysis of CO2 emissions which include:

1. Country: The name of the country.
2. Birth Rate: The number of births per 1,000 people.
3. CO2 Emissions: The total CO2 emissions produced by the country.
4. Population: The total population of the country.
5. Infant Mortality: The number of infant deaths per 1,000 live births.
6. Life Expectancy: The average number of years a person is expected to live
7. Fertility Rate: The average number of children born per woman.
8. Forested Area (%): Percentage of land area covered by forests.
9. Gasoline Price: Price of gasoline per liter in dollars.
10. Maternal Mortality Ratio: Number of maternal deaths per 100,000 live births

Data preprocessing

Selecting relevant columns

The dataset we used initially came with over 30 variables, however, we do this to focus on the columns relevant to the analysis of CO2 emissions and socio-economic factors.

```
#removing the columns that are not required in our observation
worlddata<-subset(worlddata,select = c("Country", "Birth.Rate", "Co2.Emissions", "Fertility.Rate","Gaso
```

Handling Missing Values

To ensure the dataset is complete and does not contain any missing or empty values, which can cause errors in analysis.

```r
# Remove rows with missing values (NA)
worlddata <- na.omit(worlddata)
# Remove rows with empty cells
worlddata <- worlddata[apply(worlddata, 1, function(row) !any(row == "")), ]
```

Converting Data Types

To convert columns to appropriate data types (numeric or factor) for accurate analysis, prevent errors and to facilitate numerical operations and visualizations.

```r
#convert forested area from char to numeric
worlddata$Forested.Area....<-gsub("%","",worlddata$Forested.Area....)
worlddata$Forested.Area....<-as.numeric(worlddata$Forested.Area....)

#convert co2 emissions from char to numeric
worlddata$Co2.Emissions<-(gsub(",","",worlddata$Co2.Emissions))
worlddata$Co2.Emissions<-as.numeric(worlddata$Co2.Emissions)

#convert population from char to numeric
worlddata$Population<-(gsub(",","",worlddata$Population))
worlddata$Population<-as.numeric(worlddata$Population)

#convert gasoline price from char to numeric
worlddata$Gasoline.Price <- as.numeric(gsub("\\$", "", worlddata$Gasoline.Price))
worlddata$Gasoline.Price <- as.numeric(worlddata$Gasoline.Price)

#convert countries to factor
worlddata$Country<-as.factor(worlddata$Country)
```

Removing Extreme Values (outliers)

To remove extreme values of CO2 emissions that could skew the analysis and lead to misleading conclusions.

```r
cleaned <- worlddata[worlddata$`Co2.Emissions` <= 5000000, ]
```

Data Exploration

Summary statistics

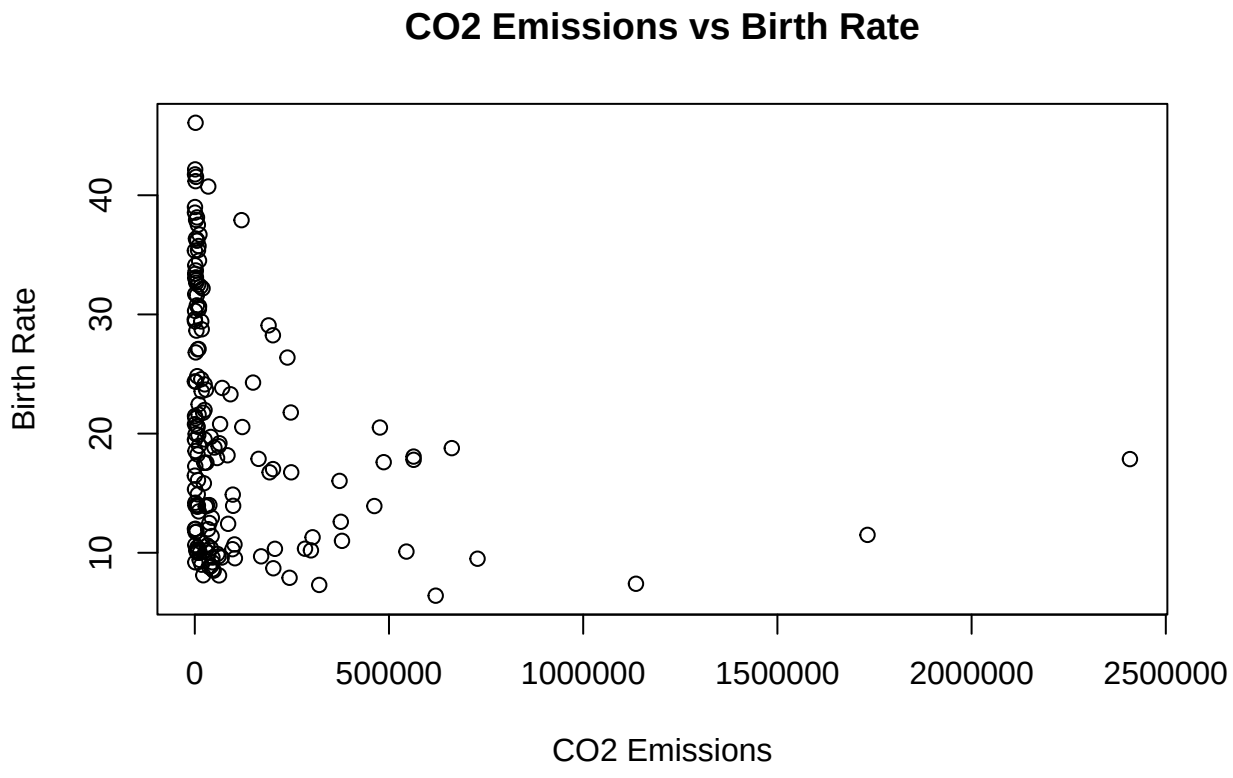
```r
summary(cleaned)
```

```
##                 Country      Birth.Rate    Co2.Emissions     Fertility.Rate
##  Afghanistan        :  1   Min.   : 6.40   Min.   :    147   Min.   :0.980
##  Albania            :  1   1st Qu.:11.22   1st Qu.:   4489   1st Qu.:1.688
##  Algeria            :  1   Median :18.01   Median :  17837   Median :2.230
##  Angola             :  1   Mean   :20.21   Mean   : 110216   Mean   :2.670
##  Antigua and Barbuda:  1   3rd Qu.:28.35   3rd Qu.:  74034   3rd Qu.:3.522
##  Argentina          :  1   Max.   :46.08   Max.   :2407672   Max.   :6.910
```

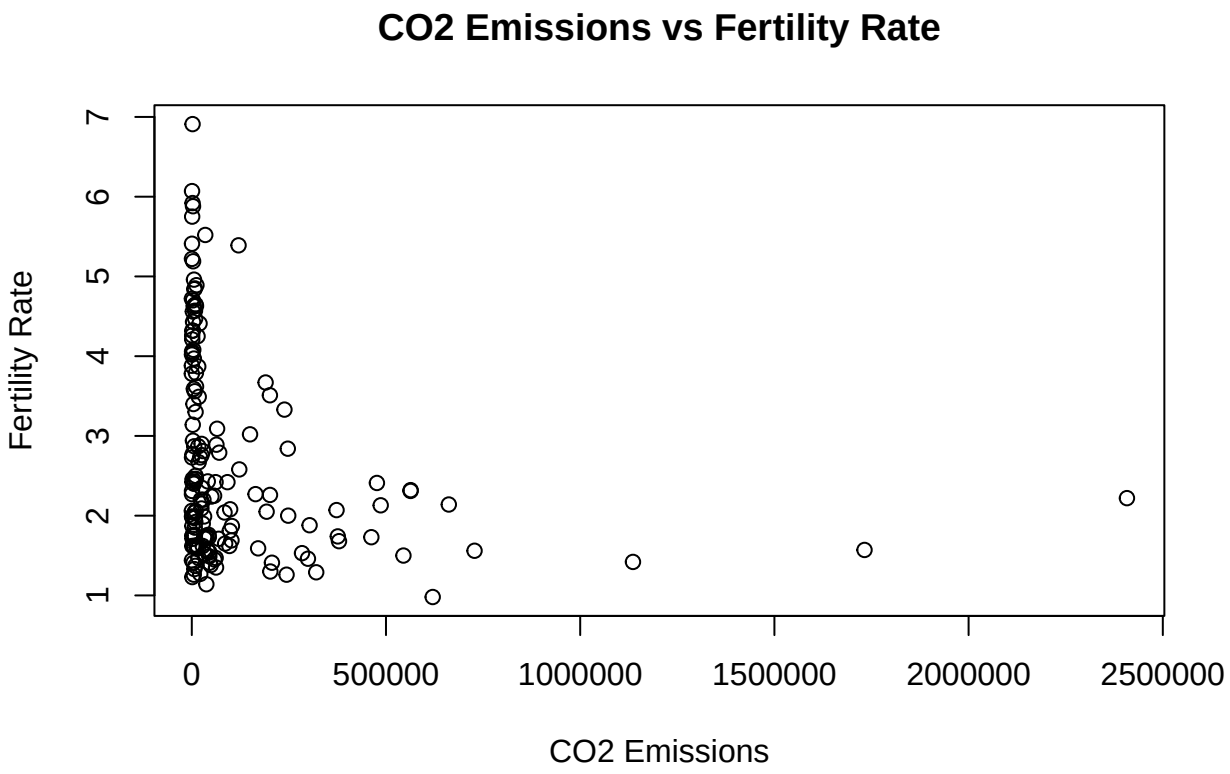
```
##  (Other)            :162
##  Gasoline.Price   Infant.mortality Life.expectancy Maternal.mortality.ratio
##  Min.   :0.0000   Min.   : 1.400   Min.   :52.80   Min.   :   2.00
##  1st Qu.:0.7575   1st Qu.: 5.875   1st Qu.:67.05   1st Qu.:  11.75
##  Median :0.9800   Median :13.700   Median :74.00   Median :  46.00
##  Mean   :0.9928   Mean   :20.977   Mean   :72.48   Mean   : 154.79
##  3rd Qu.:1.2125   3rd Qu.:32.025   3rd Qu.:77.65   3rd Qu.: 185.25
##  Max.   :2.0000   Max.   :84.500   Max.   :84.20   Max.   :1140.00
##
##    Population        Forested.Area....
##  Min.   :9.712e+04   Min.   : 0.00
##  1st Qu.:3.665e+06   1st Qu.: 9.95
##  Median :1.028e+07   Median :31.15
##  Mean   :3.508e+07   Mean   :29.99
##  3rd Qu.:3.198e+07   3rd Qu.:45.55
##  Max.   :1.366e+09   Max.   :98.30
##
```

Data visualization

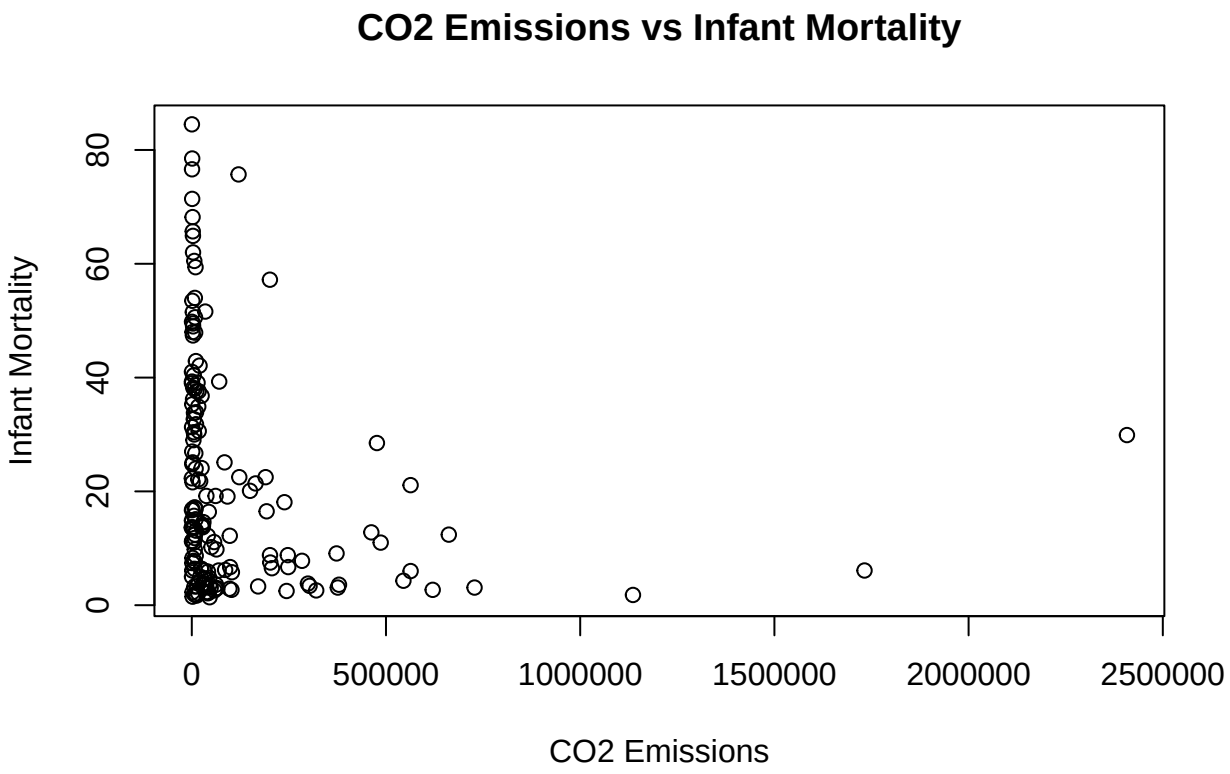
```
# Scatter plots for CO2 emissions vs key variables
plot(cleaned$Co2.Emissions, cleaned$Birth.Rate, main="CO2 Emissions vs Birth Rate", xlab="CO2 Emissions
```



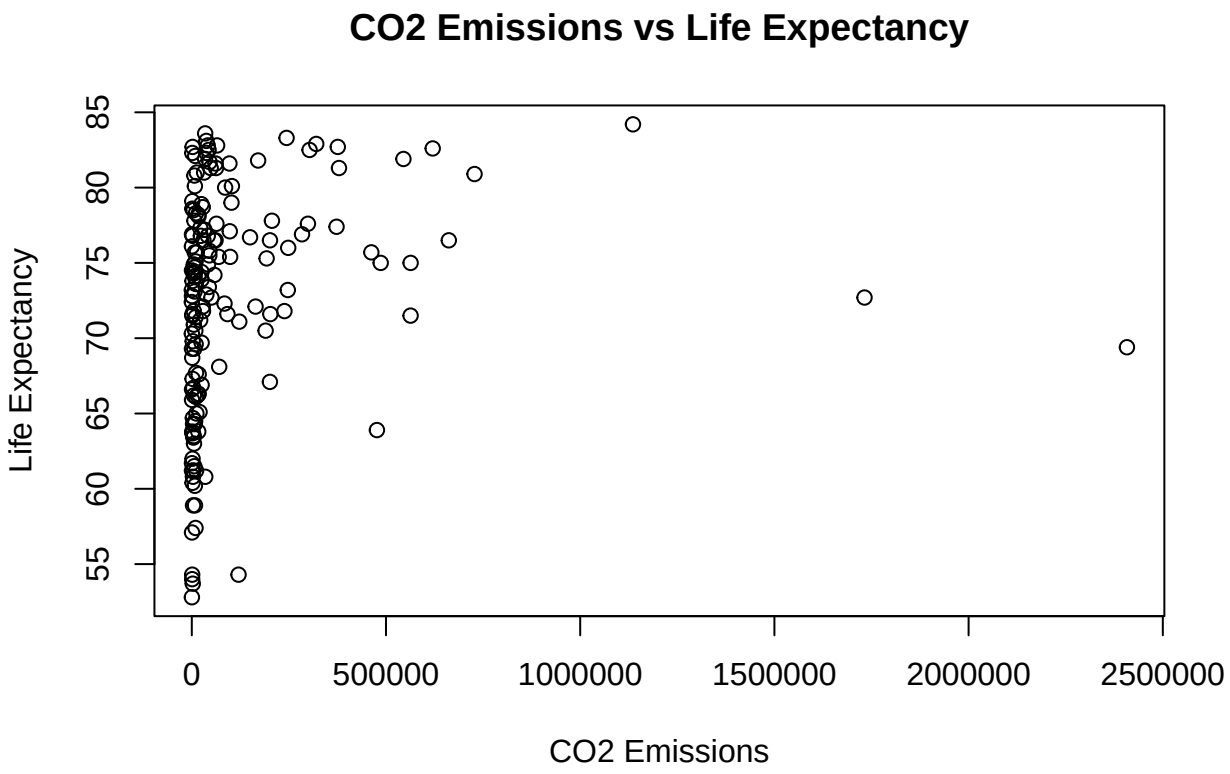
```r
plot(cleaned$Co2.Emissions, cleaned$Fertility.Rate, main="CO2 Emissions vs Fertility Rate", xlab="CO2 Em
```



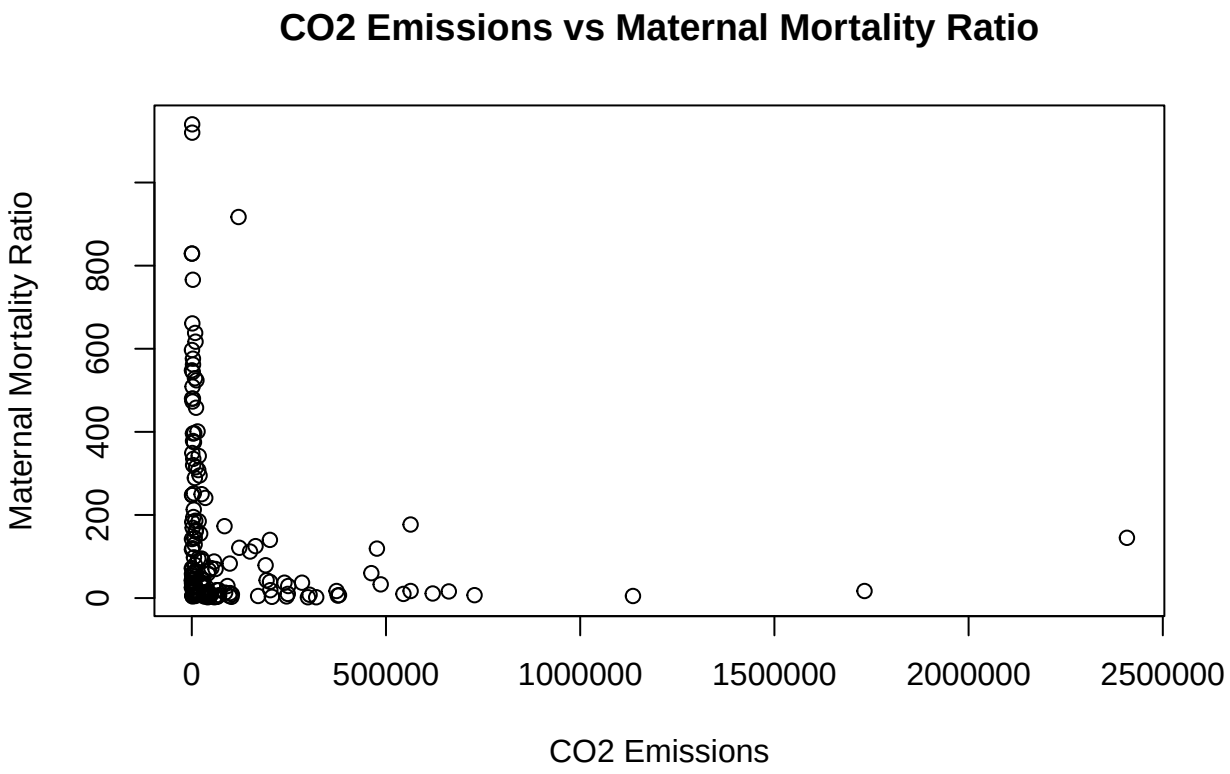

```r
plot(cleaned$Co2.Emissions, cleaned$Infant.mortality, main="CO2 Emissions vs Infant Mortality", xlab="C
```



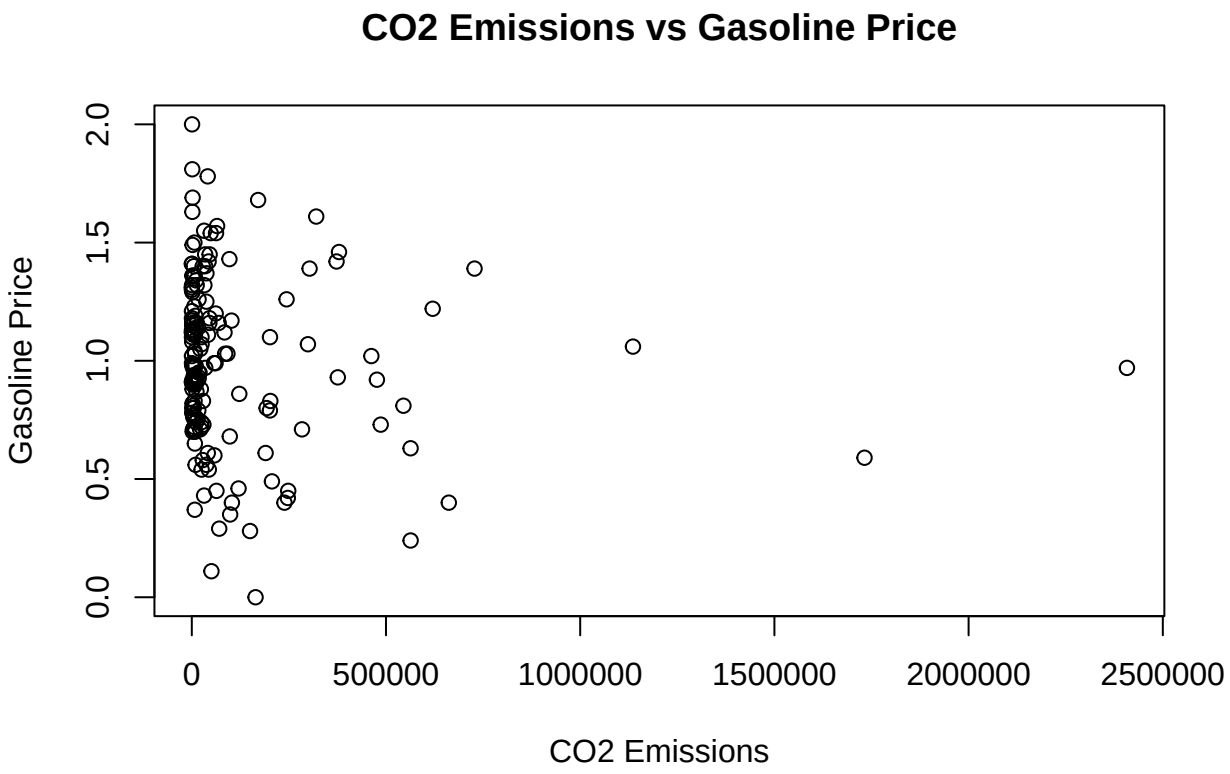

```
plot(cleaned$Co2.Emissions, cleaned$Life.expectancy, main="CO2 Emissions vs Life Expectancy", xlab="CO2
```



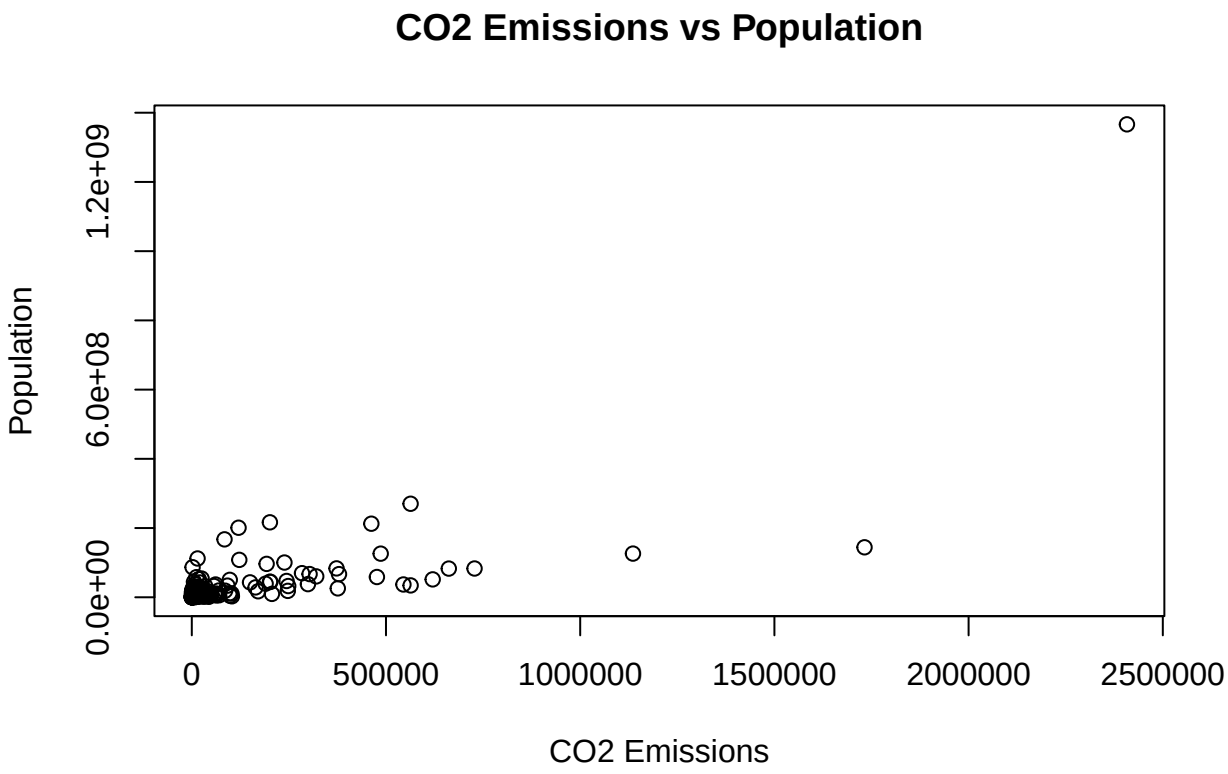

```
plot(cleaned$Co2.Emissions, cleaned$Maternal.mortality.ratio, main="CO2 Emissions vs Maternal Mortality
```



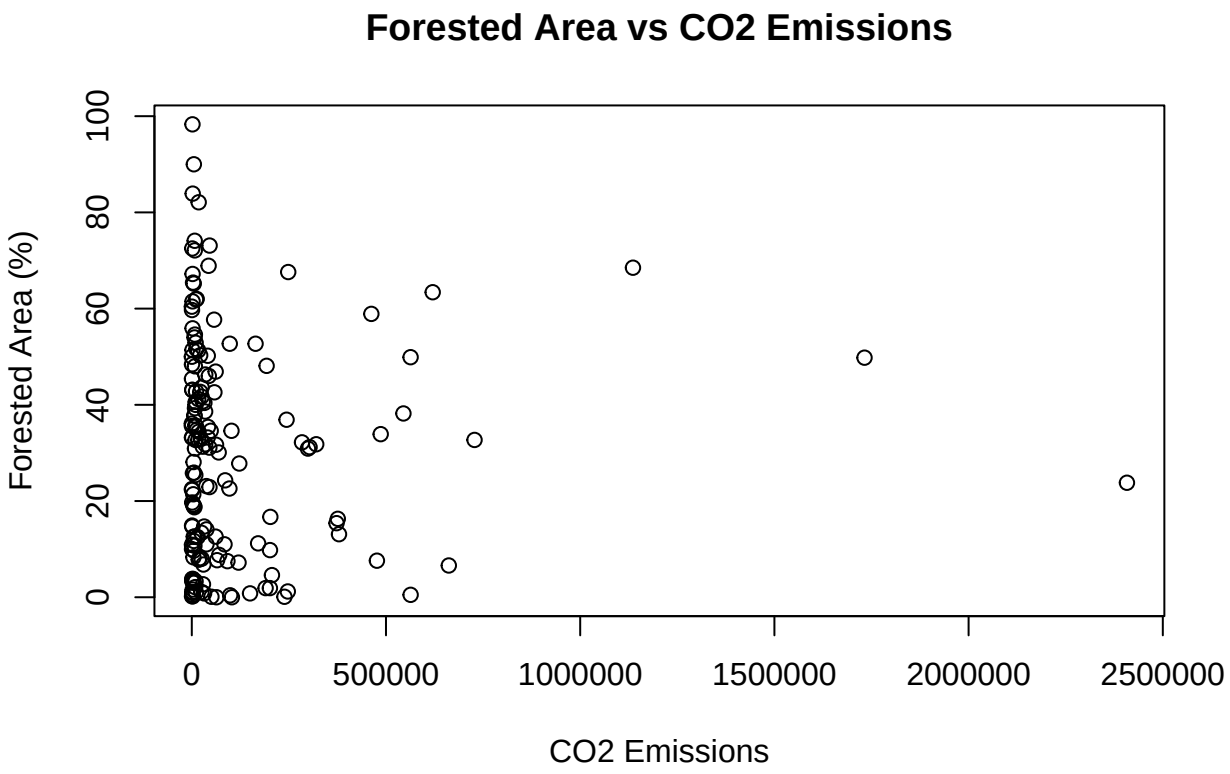

```
plot(cleaned$Co2.Emissions, cleaned$Gasoline.Price, main="CO2 Emissions vs Gasoline Price", xlab="CO2 Em
```



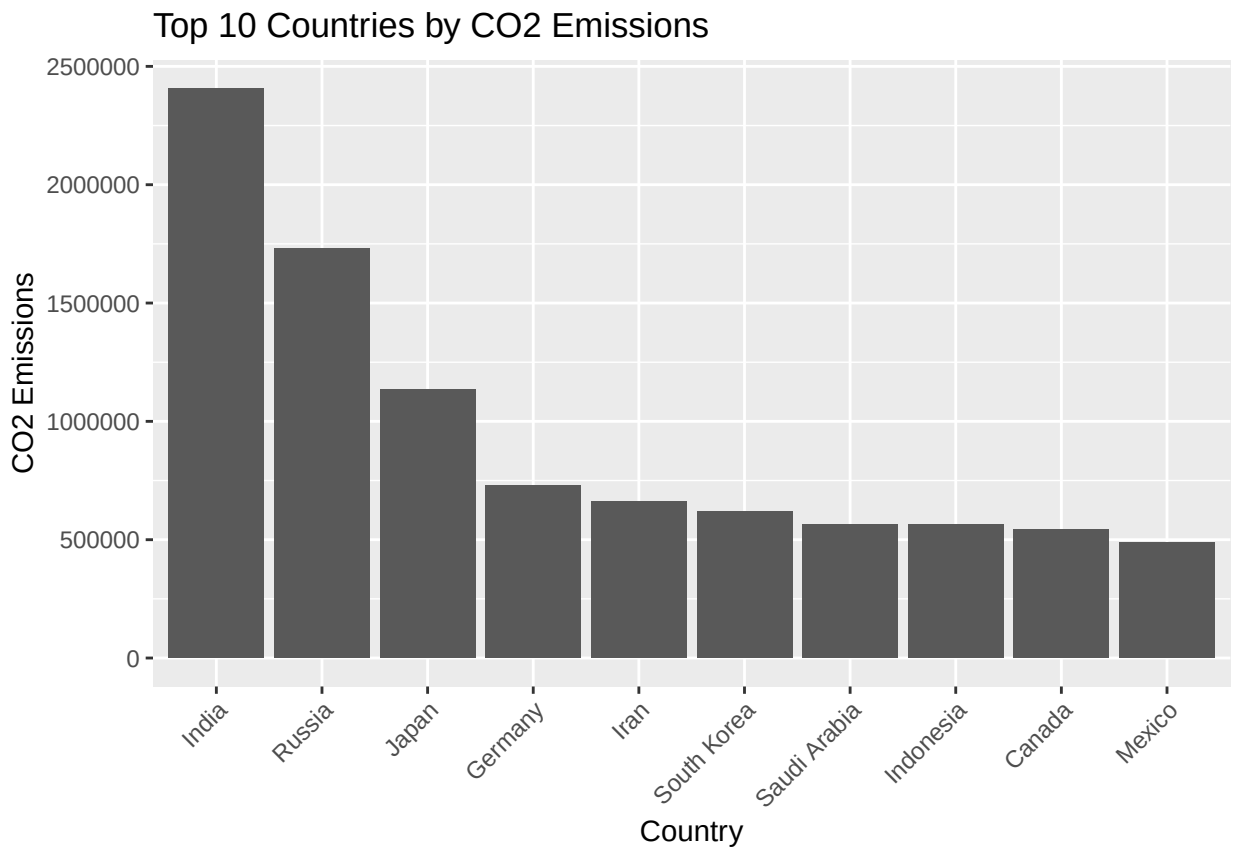

```
plot(cleaned$Co2.Emissions, cleaned$Population, main="CO2 Emissions vs Population", xlab="CO2 Emissions
```



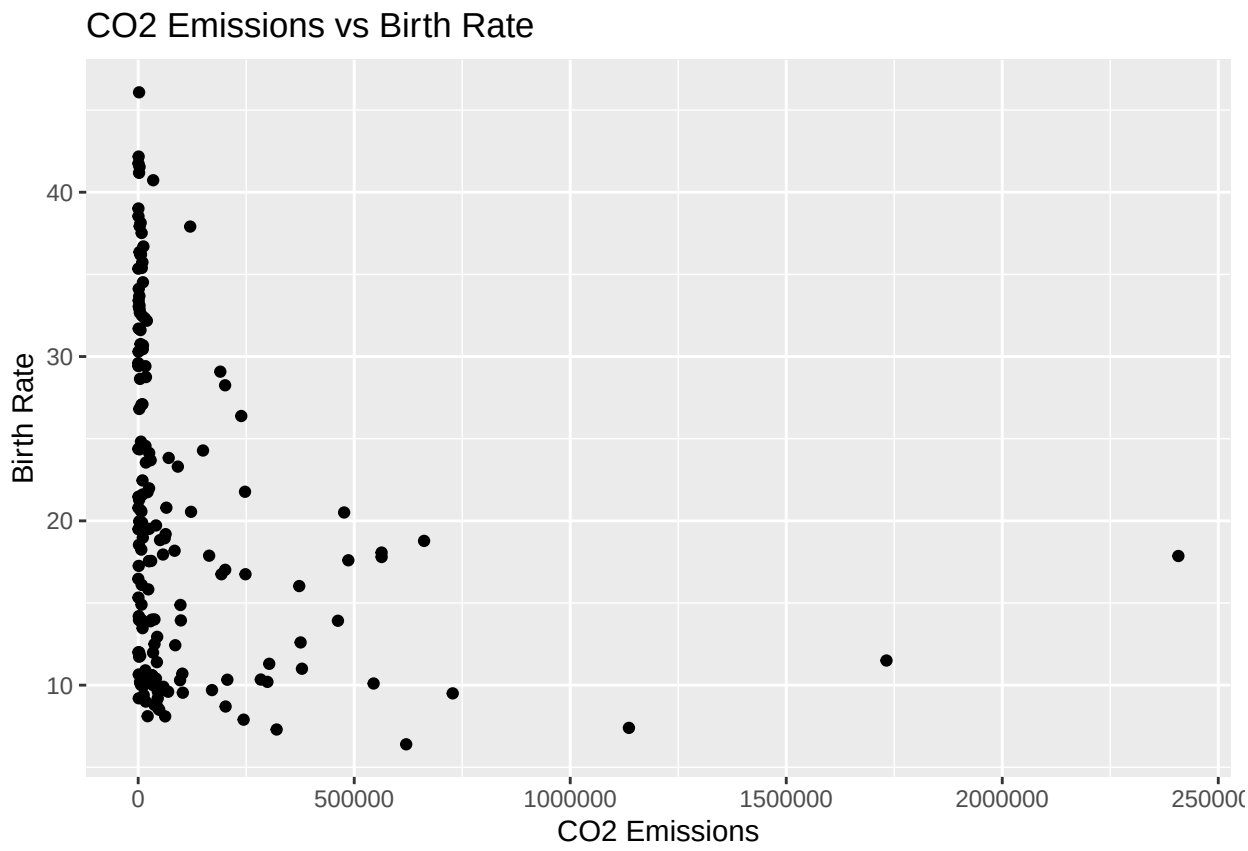

```
plot(cleaned$Co2.Emissions, cleaned$Forested.Area...., main="Forested Area vs CO2 Emissions", xlab="CO2
```



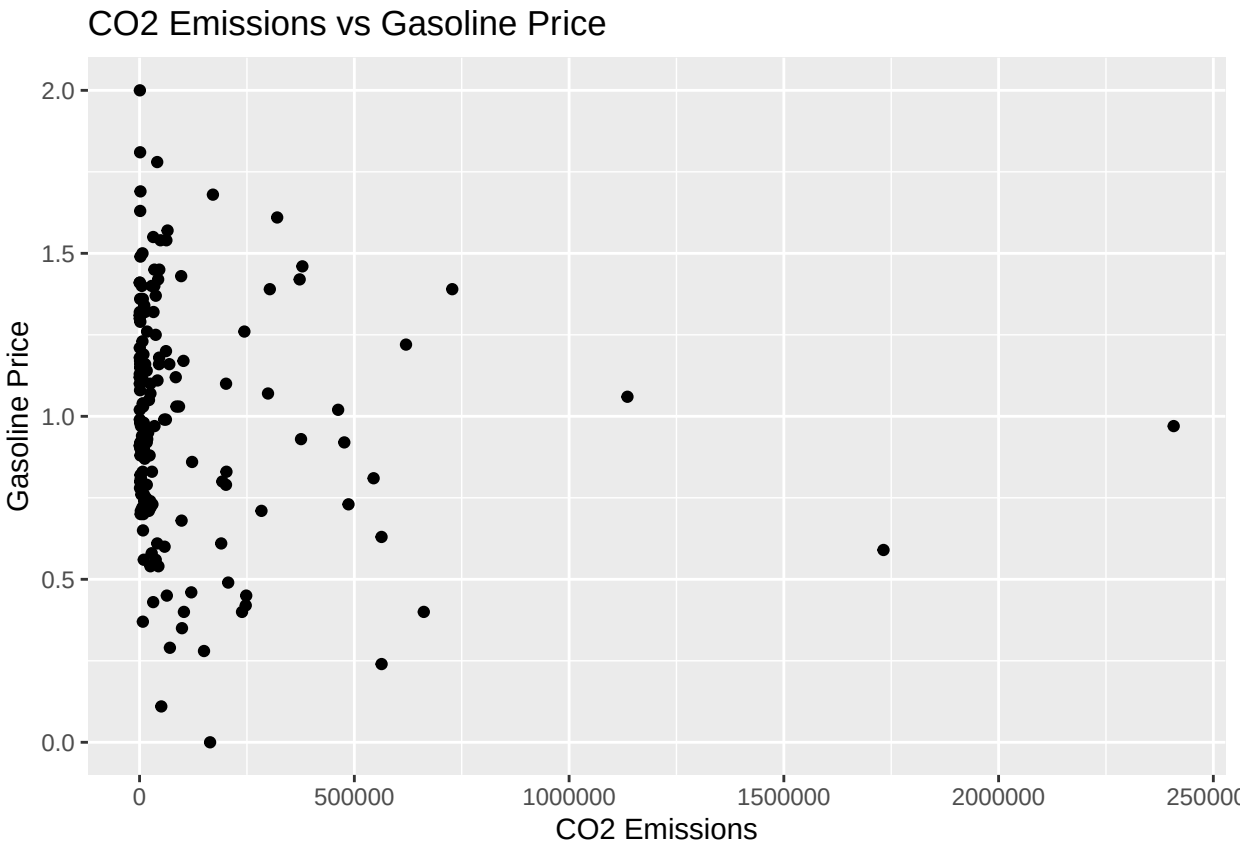

```r
# Top 10 countries by CO2 emissions
top_10_countries <- cleaned %>% arrange(desc(Co2.Emissions)) %>% head(10)

# Static bar chart for top 10 CO2 emitting countries
ggplot(top_10_countries, aes(x = reorder(Country, -Co2.Emissions), y = Co2.Emissions)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 10 Countries by CO2 Emissions", x = "Country", y = "CO2 Emissions") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Static scatter plot for CO2 emissions vs Birth Rate
ggplot(cleaned, aes(x = Co2.Emissions, y = Birth.Rate)) +
  geom_point() +
  labs(title = "CO2 Emissions vs Birth Rate", x = "CO2 Emissions", y = "Birth Rate")
```




```
# Static scatter plot for CO2 emissions vs Gasoline Price
ggplot(cleaned, aes(x = Co2.Emissions, y = Gasoline.Price)) +
  geom_point() +
  labs(title = "CO2 Emissions vs Gasoline Price", x = "CO2 Emissions", y = "Gasoline Price")
```

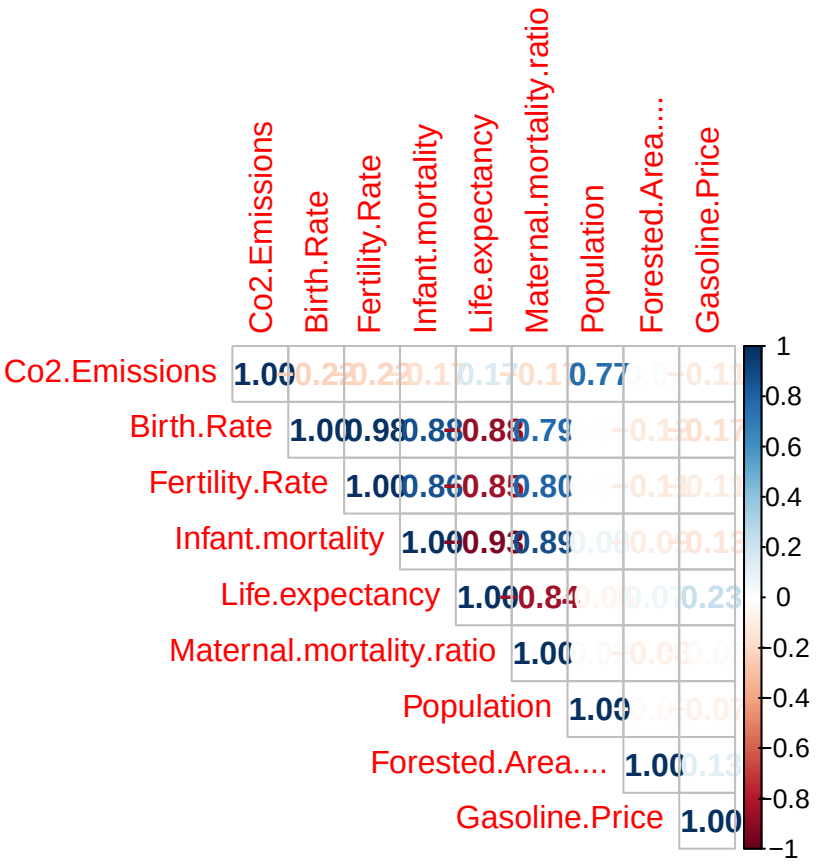


Statistical analysis

Correlation analysis

The correlation coefficients indicate the strength and direction of the relationships between CO2 emissions and other socio-economic factors.

```
# Correlation matrix
c <- cor(cleaned %>% select(Co2.Emissions, Birth.Rate, Fertility.Rate, Infant.mortality, Life.expectanc
corrplot(c, type="upper", method="number")
```



Correlation Analysis helps identify which socio-economic factors have a strong relationship with CO2 emissions. For example, if Population has a high positive correlation with CO2 emissions, it suggests that countries with larger populations tend to emit more CO2.

Discussion

Summary Statistics

The summary statistics provide an overview of the key variables related to CO2 emissions and their socio-economic determinants. A brief examination of these statistics reveals important patterns and trends:

- CO2 Emissions: The data exhibits a wide range of CO2 emissions across countries, indicating significant disparities in emissions levels.
- Population: There is also a notable variation in population sizes, which may influence the total CO2 emissions produced by a country.
- Birth Rate and Fertility Rate: These rates vary widely, reflecting different stages of demographic transition and development.
- Infant Mortality and Maternal Mortality Ratio: These indicators provide insights into the health and well-being of populations, which can be associated with economic development and environmental impacts.
- Life Expectancy: This variable can serve as a proxy for the overall quality of life and development status of a country.
- Forested Area: The percentage of forested land may be inversely related to CO2 emissions, as forests act as carbon sinks.

- Gasoline Price: This economic variable could influence the consumption patterns of fossil fuels and, consequently, CO2 emissions.

Data Visualization

The scatter plots and interactive visualizations offer a closer look at the relationships between CO2 emissions and various socio-economic factors:

- CO2 Emissions vs Birth Rate: No strong linear relationship is observed. However, countries with higher birth rates tend to have varying levels of CO2 emissions, suggesting other mediating factors at play.
- CO2 Emissions vs Fertility Rate: Similar to the birth rate, the fertility rate does not show a strong linear correlation with CO2 emissions.
- CO2 Emissions vs Infant Mortality: Higher infant mortality rates tend to correlate with lower CO2 emissions. This could be indicative of less industrialized nations with lower overall emissions.
- CO2 Emissions vs Life Expectancy: Countries with higher life expectancy generally have higher CO2 emissions, likely due to advanced industrialization and energy consumption.
- CO2 Emissions vs Maternal Mortality Ratio: Higher maternal mortality ratios are associated with lower CO2 emissions, reflecting the economic and development disparities.
- CO2 Emissions vs Gasoline Price: There appears to be an inverse relationship where countries with higher gasoline prices tend to have lower CO2 emissions. This can be attributed to more efficient energy use or policies promoting alternative energy sources.
- CO2 Emissions vs Population: A positive correlation is evident, indicating that more populous countries tend to emit more CO2, which is expected given higher energy demands and industrial activities.
- CO2 Emissions vs Forested Area: There is a slight inverse relationship, suggesting that countries with more forested areas tend to have lower CO2 emissions, possibly due to the carbon sequestration capabilities of forests.

Correlation Analysis

The correlation matrix reveals the strength and direction of relationships between CO2 emissions and socio-economic factors:

- Strong Positive Correlations: CO2 emissions have a strong positive correlation with population size, highlighting the impact of population density and industrial activities on emissions.
- Moderate Positive Correlations: Life expectancy shows a moderate positive correlation with CO2 emissions, suggesting that more developed countries with higher life expectancy tend to emit more CO2.
- Negative Correlations: Forested area and gasoline prices exhibit negative correlations with CO2 emissions, indicating the importance of environmental conservation and economic measures in controlling emissions.

Interpretation and Implications

The findings from this analysis provide several key insights:

- Economic Development and Emissions: More developed countries with higher life expectancy and lower maternal and infant mortality rates tend to have higher CO2 emissions. This underscores the challenge of balancing economic growth with environmental sustainability.
- Population Impact: Population size is a critical factor in CO2 emissions, emphasizing the need for population management and efficient resource use.
- Environmental and Economic Policies: Higher gasoline prices and greater forested areas are associated with lower CO2 emissions, suggesting the efficacy of policies promoting sustainable energy use and forest conservation.

Conclusion

The lack of clear correlations demonstrates the complexity and multifaceted nature of CO2 emissions and socio-economic determinants. This suggests that socio-economic aspects are affected by many factors beyond CO2 emissions, which may not be as simple as previously assumed. To effectively address climate change, we need to focus on variables beyond CO2 emissions alone. By diversifying our efforts to address other important variables, we will improve the climate more effectively.

For potential avenues for further research or analysis it could be beneficial from exploring additional datasets to identify other variables beyond CO2 that may correlate with socio-economic aspects. One particularly intriguing avenue is to investigate whether socio-economic conditions are influenced by the interplay of multiple variables. Specifically, examining the correlation between socio-economic factors and a combination of two or more variables could provide deeper insights and a more comprehensive understanding of the factors affecting socio-economic outcomes in the context of climate change.