MATH 485 - Explotary Data Analysis Term Project

CO2 Emission Worldwide

Berdan Yolcu

Contents

Introduction

- The primary focus of this project is to analyze global CO2 emissions with a particular
- emphasis on Europe and Asia. Using statistical methods, we explore the trends, correlations,
- and underlying relationships between population and CO2 emissions. The insights gained
- provide a deeper understanding of environmental impacts and help predict future emission trends.

Data Preprocessing

	Column <chr>	NA_C... <dbl>	NA_Percenta... <dbl>
country	country	0	0.00000
year	year	0	0.00000
population	population	10590	20.92968
gdp	gdp	36034	71.21625
co2	co2	19249	38.04301
co2_per_capita	co2_per_capita	23683	46.80620
co2_per_gdp	co2_per_gdp	34307	67.80308
coal_co2	coal_co2	25529	50.45456
oil_co2	oil_co2	25556	50.50793
gas_co2	gas_co2	25655	50.70359

1–10 of 14 rows Previous 1 2 Nex

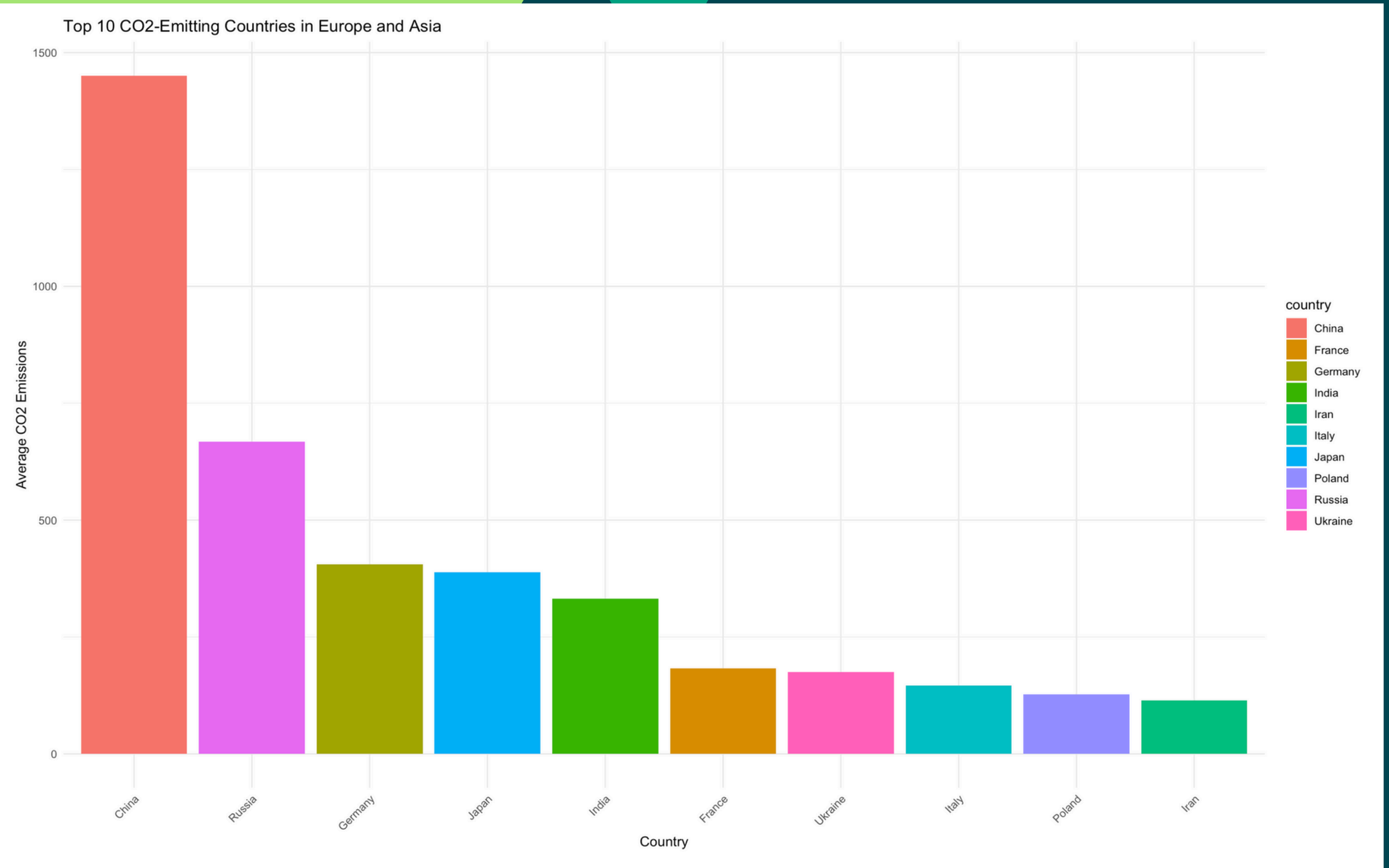
- Missing values for numerical columns (e.g., co2, population) were filled using their respective medians or means.
- Columns with more than 50% missing data (e.g., gdp, coal_co2) were removed
- To make the analysis more focused, the dataset was filtered to include only countries from Europe and Asia.

Descriptive Statistics and Visualizations

Let’s look at some graphs!

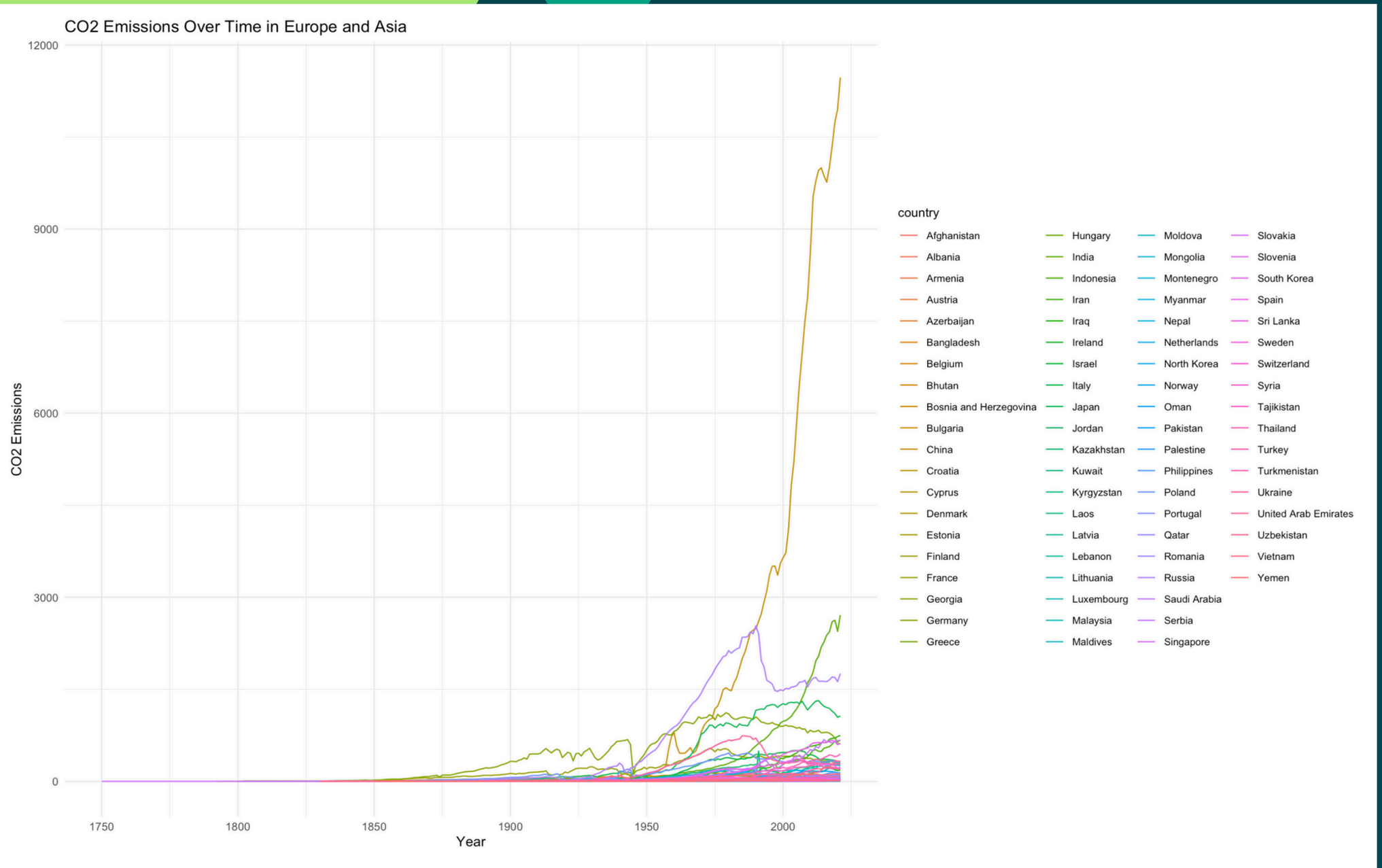
Top 10 CO2-Emitting Countries in Europe and Asia:

- A bar chart identified the top emitters, with industrialized nations dominatin the list.
- The most CO2-Emitting country is China by a wide margin.



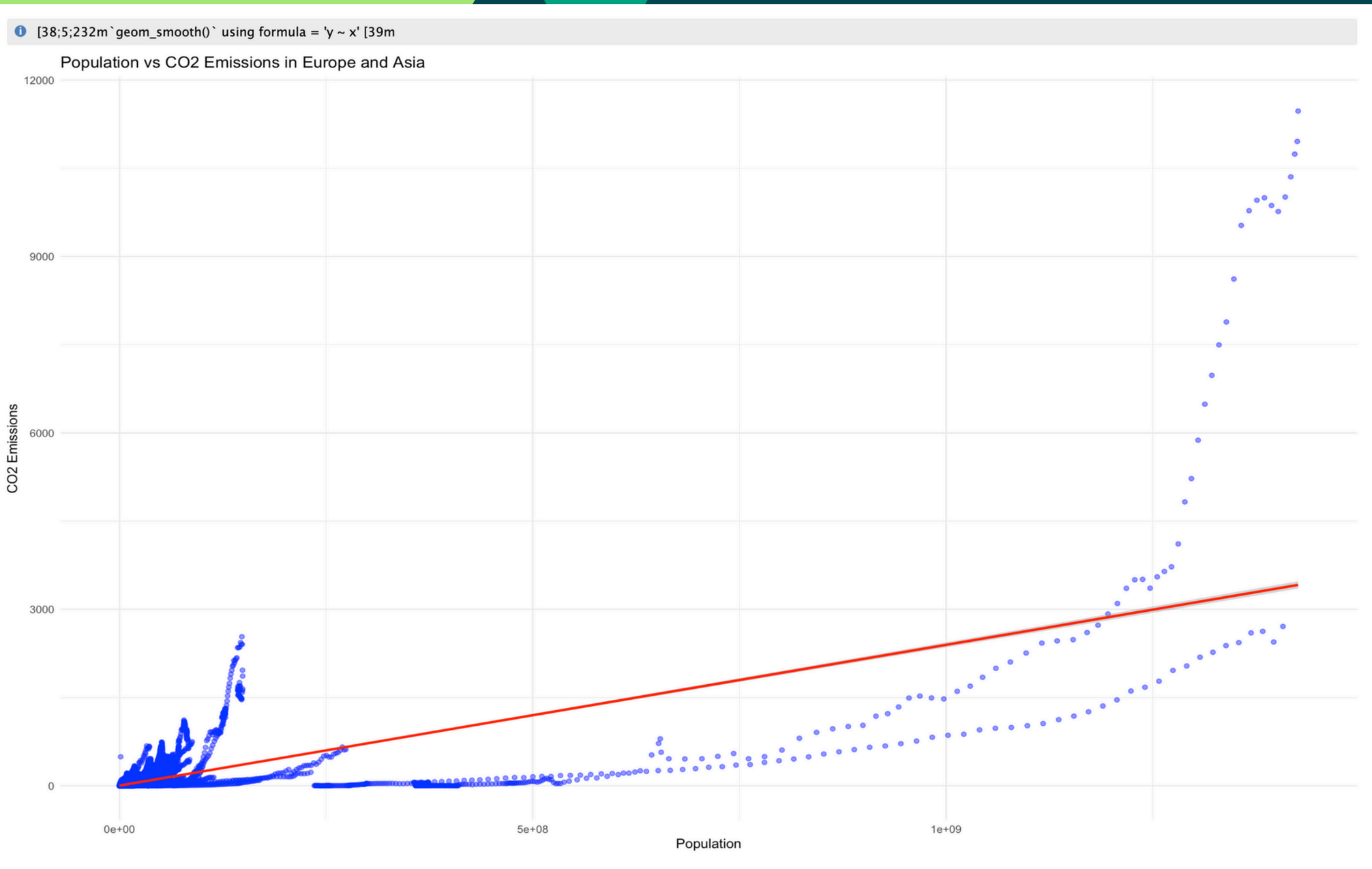
2. CO2 Emissions Over Time

- Line plots revealed trends in CO2 emissions across years, showcasing growth patterns in rapidly developing countries.



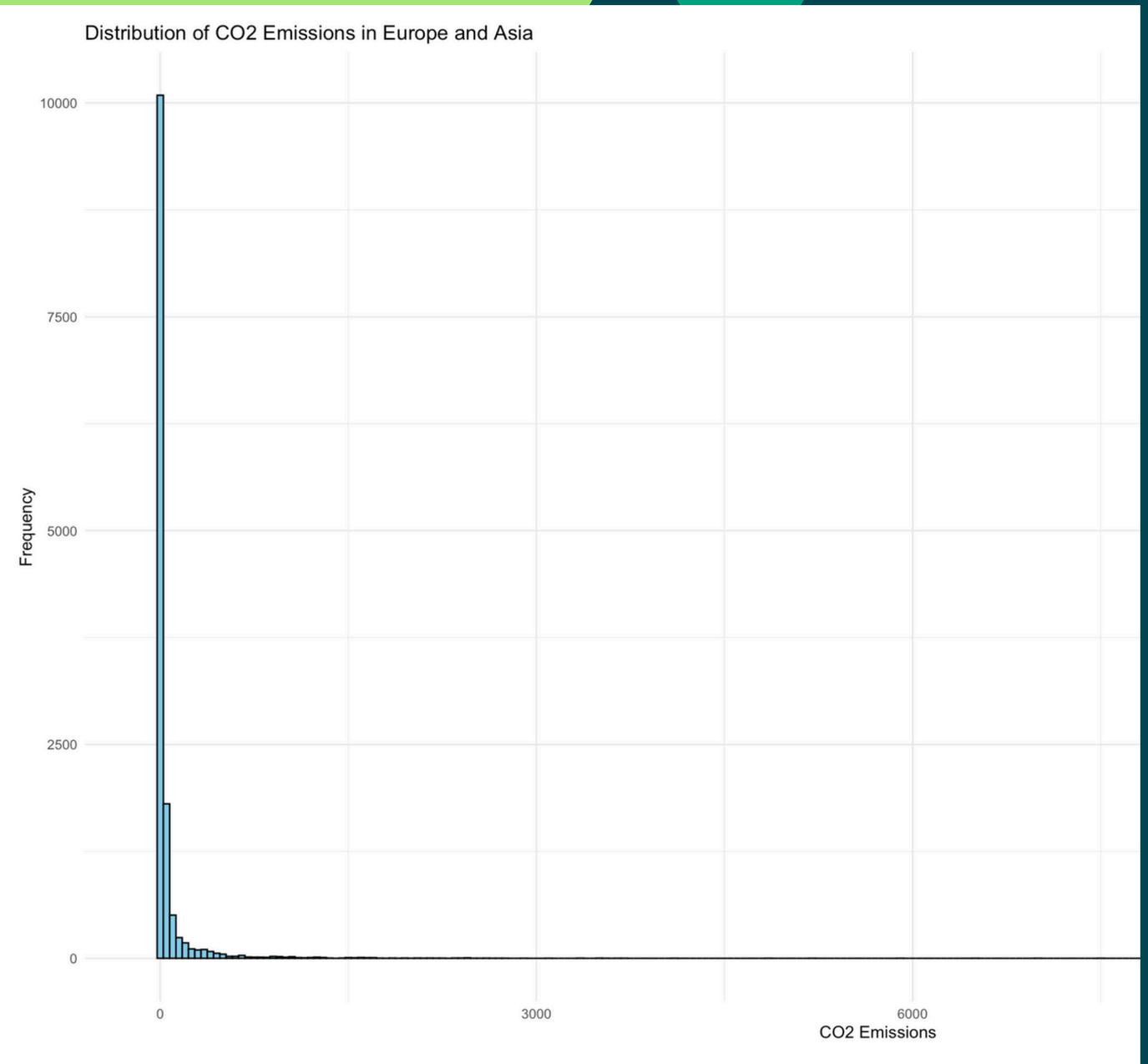
3. Population vs CO2 Emissions

- A scatter plot and regression line illustrated a direct correlation between population and CO2 emissions.



4. Distribution of CO2 Emissions

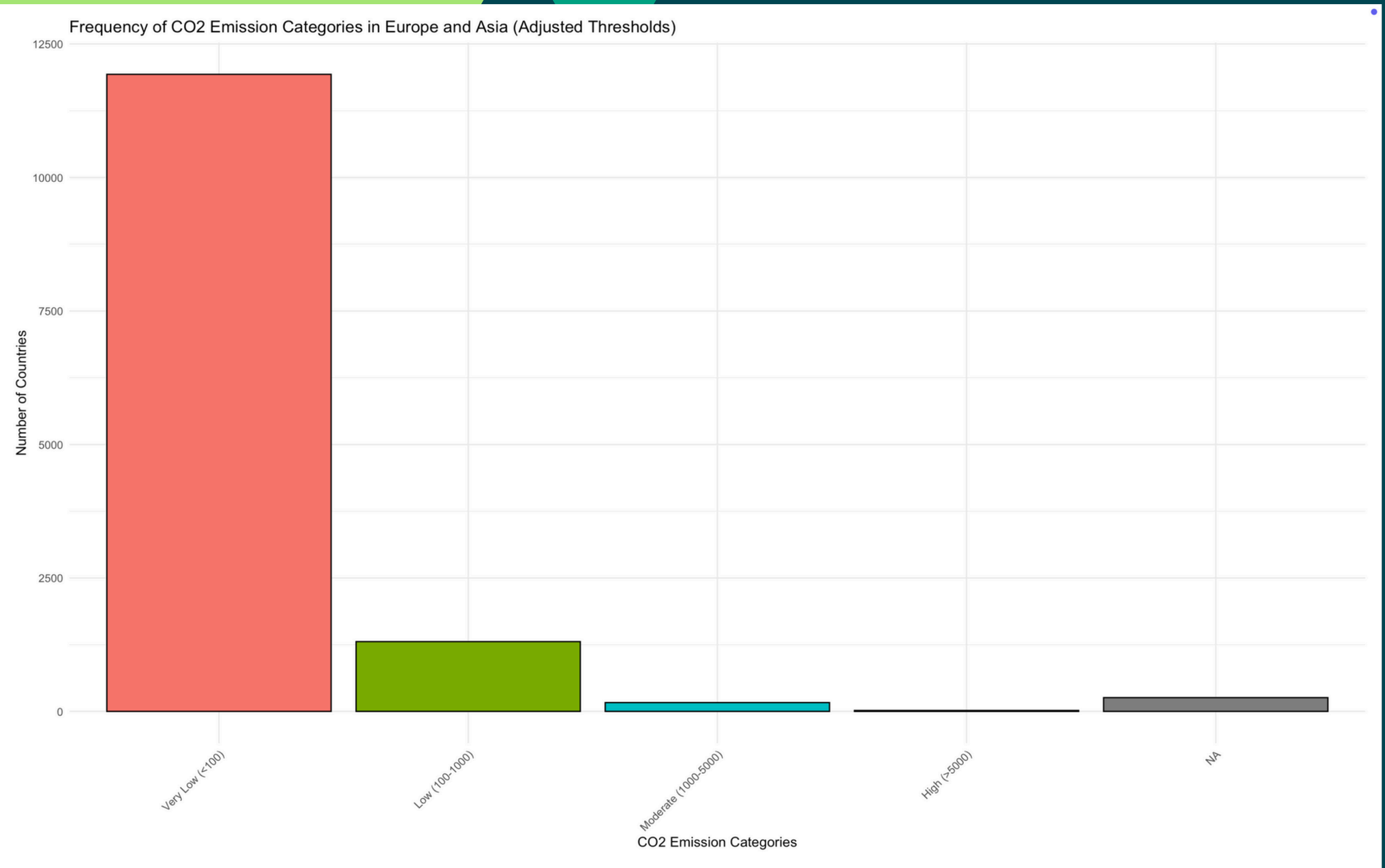
- This histogram illustrates the distribution of CO2 emissions in Europe and Asia, revealing a right-skewed pattern.



- The majority of countries emit relatively low levels of CO2, while a small number of countries contribute disproportionately high emissions. This disparity highlights the significant role of a few industrialized nations in global CO2 output.

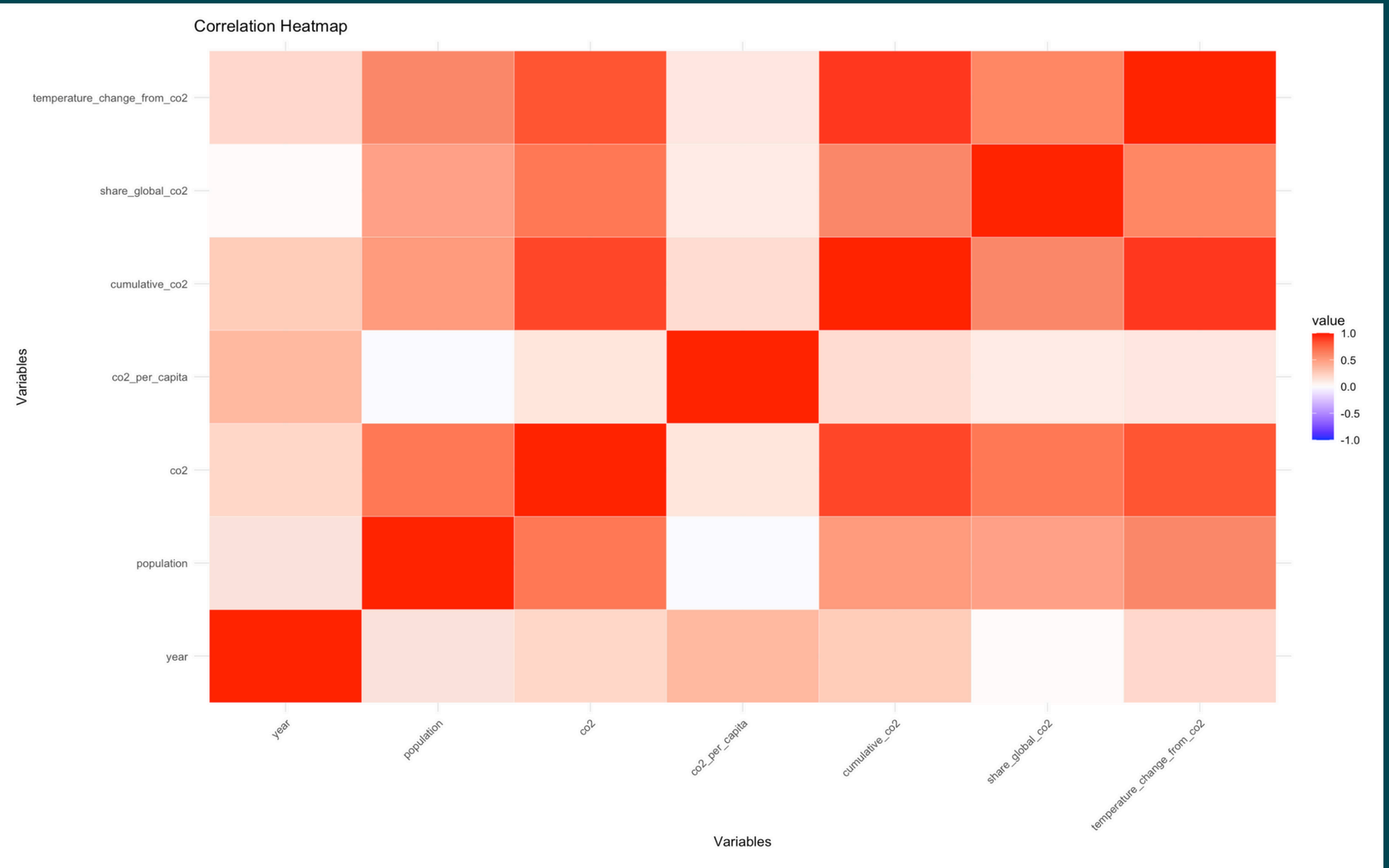
5- Frequency of CO2 Emission Categories:

- Countries were categorized into Very Low, Low, Moderate, and High emission groups, highlighting disparities in emission levels.



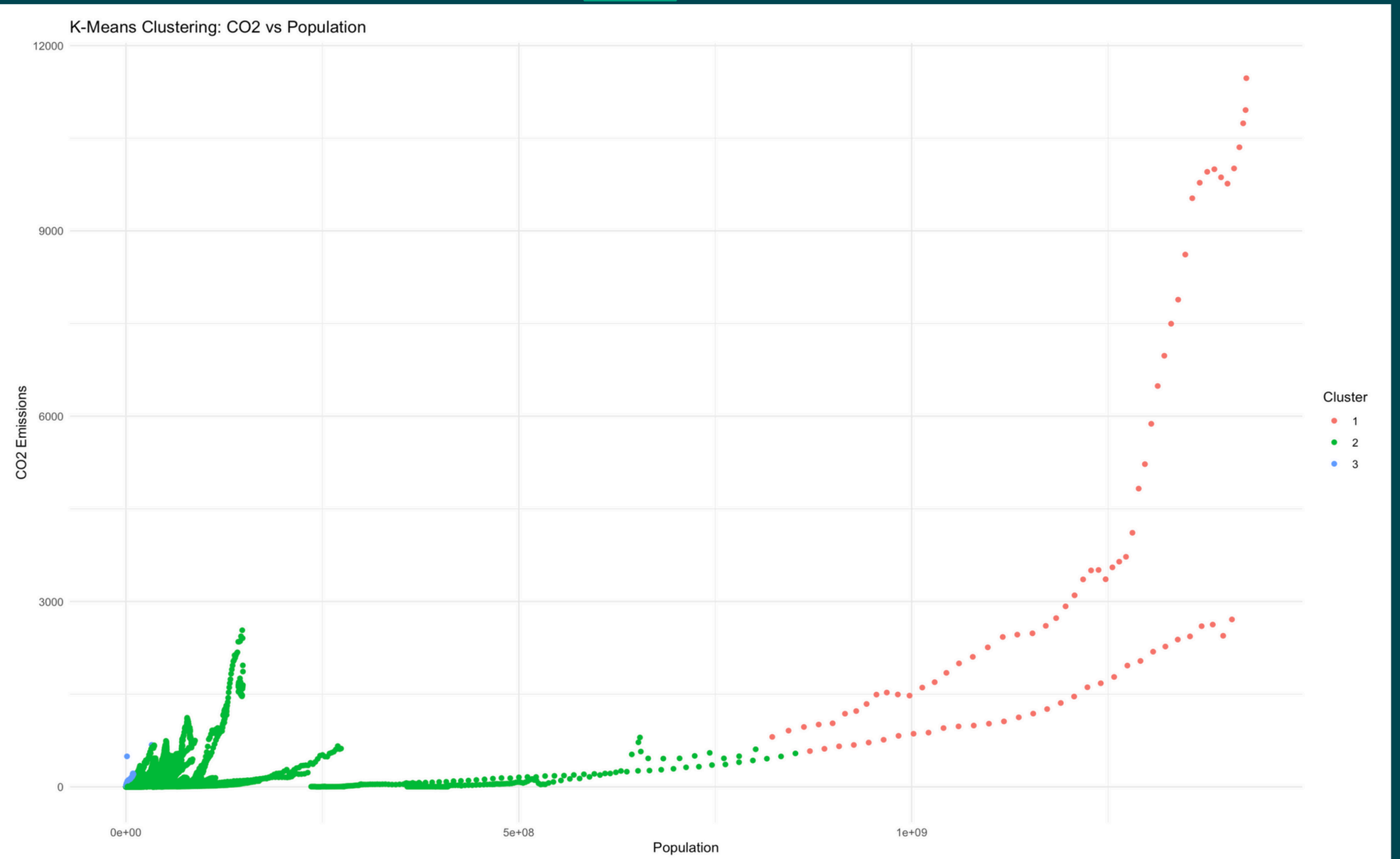
Correlation Inspection

- Strong positive correlation between population and co2.
- Expected high correlation between co2 and share_global_co2 (redundant for further analysis).



CLUSTER ANALYSIS

- K-means clustering grouped countries based on co2, population, and co2_per_capita into 3 clusters.



- Cluster 1 (Red): High population and high emissions (e.g., industrialized nations).
- Cluster 2 (Green): Moderate population and emissions (e.g., developing nations).
- Cluster 3 (Blue): Low population and emissions (e.g., smaller or less industrialized countries).

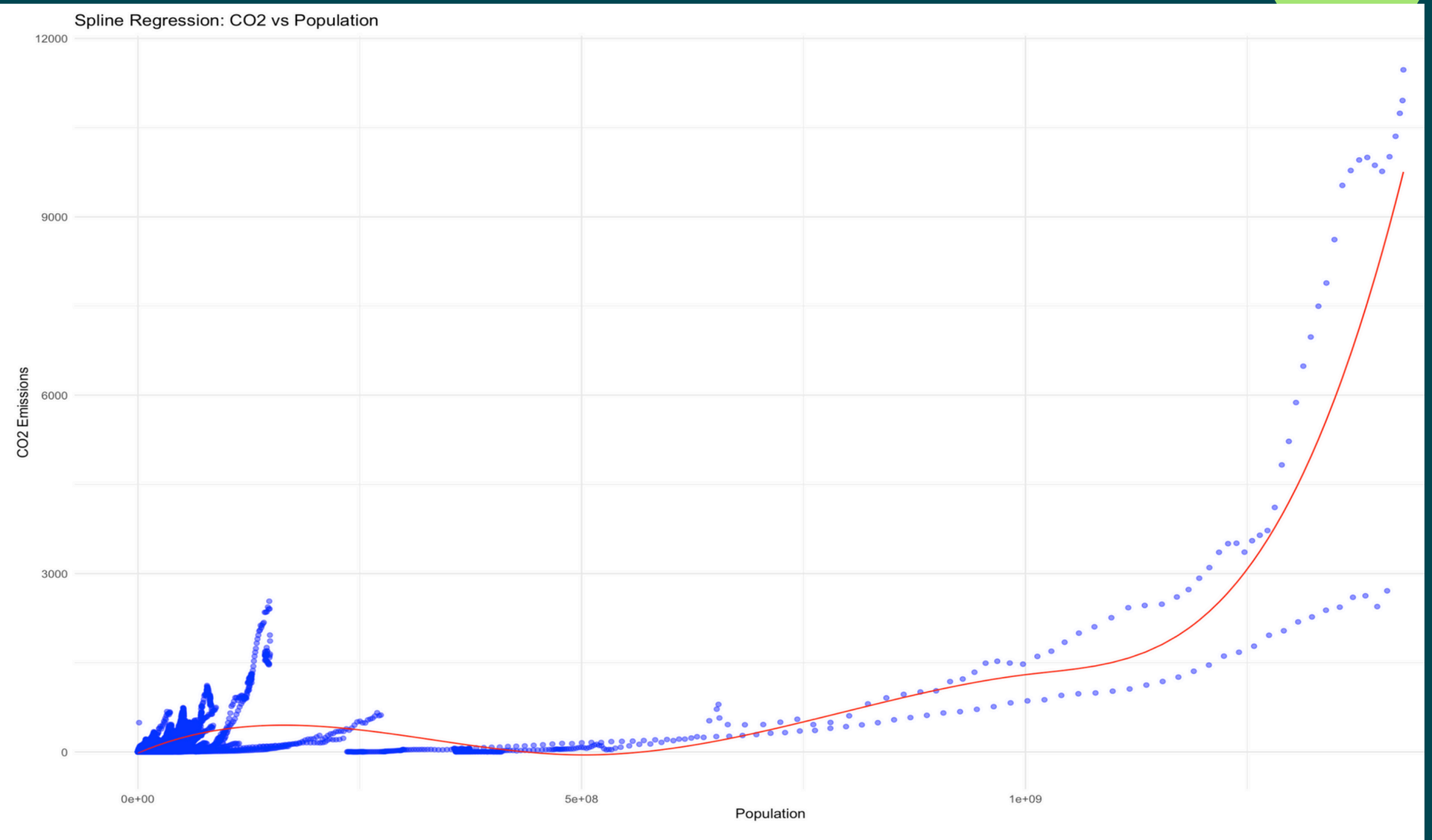
Regression Analysis

Target Variable	CO2
Independent Variable	Population
Method	Spline Regression

Spline Regression

- Spline regression is a flexible regression method that divides the data into segments using knots (breakpoints) and fits separate polynomial functions to each segment. These segments are then joined smoothly, ensuring continuity. This approach is useful when the relationship between the predictor and response variables is nonlinear. Also Spline Regression is helpfull for preventing overfitting.
- Knots: Points where the data is split for separate polynomial fits.
- Degree: The degree of the polynomial (e.g., cubic) used within each segment.

Regression Model:



Regression Model:

```
Call:
lm(formula = co2 ~ bs(population, degree = 3, knots = c(5e+08,
    1e+09)), data = df_europe_asia)

Residuals:
    Min      1Q  Median      3Q     Max
-5997.9   -17.4     9.0    20.3  3237.2

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                              -19.957      1.999  -9.985   <2e-16 ***
bs(population, degree = 3, knots = c(5e+08, 1e+09))1    1068.008     17.478  61.104   <2e-16 ***
bs(population, degree = 3, knots = c(5e+08, 1e+09))2   -1100.999     41.450 -26.562   <2e-16 ***
bs(population, degree = 3, knots = c(5e+08, 1e+09))3    1915.215     71.300  26.861   <2e-16 ***
bs(population, degree = 3, knots = c(5e+08, 1e+09))4    1292.642     72.847  17.745   <2e-16 ***
bs(population, degree = 3, knots = c(5e+08, 1e+09))5    9774.708     62.382 156.691   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 191.5 on 13675 degrees of freedom
Multiple R-squared:  0.7681,    Adjusted R-squared:  0.768
F-statistic:  9057 on 5 and 13675 DF,  p-value: < 2.2e-16
```

- The model has %76.8 Accuracy as can be seen.

Conclusion

1. Population as a Key Driver:

- A strong correlation between population size and CO2 emissions highlights the impact of population growth on environmental degradation.

2. Regional Insights:

- Europe and Asia demonstrate diverse emission patterns, with industrialized nations driving the bulk of emissions.

3. Regional Insights:

- The spline regression model provided robust predictions and underscored the importance of nonlinear relationships in environmental data.