MATH485 Exploratory Data Analysis Term Project

Worldwide CO2 Emission

Author: Berdan Yolcu

Student ID: 20200601062

Introduction

The primary focus of this project is to analyze global CO_2 emissions with a particular emphasis on Europe and Asia. Using statistical methods, we explore the trends, correlations, and underlying relationships between population and CO_2 emissions. The insights gained provide a deeper understanding of environmental impacts and help predict future emission trends.

Data Cleaning

Dataset Overview:

- The dataset was pre-processed to focus on key variables:
  - `country`, `year`, `population`, `co2`, `co2_per_capita`, `share_global_co2`, and `temperature_change_from_co2`.

Handling Missing Values:

- 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.

Data Filtering

- To make the analysis more focused, the dataset was filtered to include only countries from **Europe** and **Asia**.
- This resulted in a cleaner dataset representing significant contributors to global CO_2 emissions.

Descriptive Statistics and Visualizations

Key Statistics:

- Summary statistics showed a strong variability in `co2` emissions across countries.
- High variance was particularly evident in countries with large populations.

Visualizations:

1. **Top 10 CO_2-Emitting Countries in Europe and Asia**:
   - A bar chart identified the top emitters, with industrialized nations dominating the list.

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**:
    - A histogram revealed a right-skewed distribution, indicating most countries emit relatively low levels of CO2.
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

- A **correlation heatmap** was generated to explore relationships between variables.
- Insights:
    - Strong positive correlation between `population` and `co2`.
    - Expected high correlation between `co2` and `share_global_co2` (redundant for further analysis).

Clustering Analysis

Method:

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

Insights:

- **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` emissions were selected as the target variable, with `population` as the predictor.

Model:

- **Spline Regression** was employed to capture nonlinear relationships.
  - Knots at 500 million and 1 billion population split the data into meaningful segments.

Results:

- **R-squared = 0.7681**: The model explained 76.8% of the variability in `co2` emissions.
- **Interpretation**: Spline regression significantly outperformed linear models by accounting for the nonlinear relationship.

Conclusions

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

Future Work

1. Incorporate additional predictors like `GDP` and energy consumption to refine predictions.
2. Explore machine learning models (e.g., Random Forest) for improved accuracy.
3. Extend the analysis to other continents for a global perspective.

References

1. Dataset Source: CO2 Emissions Dataset, Global Carbon Budget (2024)

2. R Libraries: `dplyr`, `ggplot2`, `splines`, `reshape2`