Qumrish Arooj (24147437)

Clustering and Fitting Analysis of Global CO2 Emissions

1. Introduction

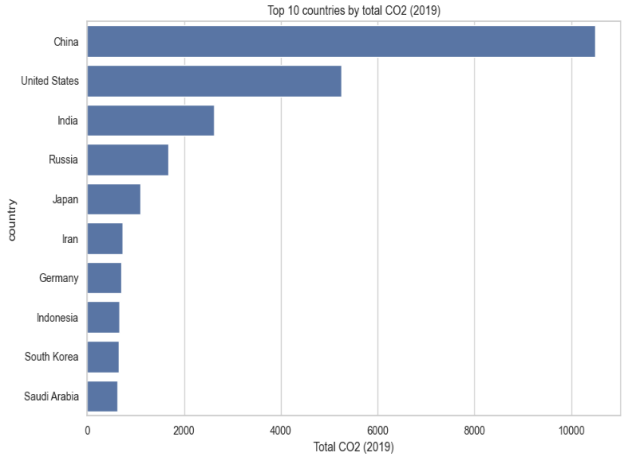
Climate change and greenhouse gas emissions are key challenges for the world today. Understanding patterns of carbon dioxide (CO2) emissions per capita across countries can help identify which nations contribute the most and how economic factors affect emissions.

In this report, I analyze CO2 emissions using **statistical summaries, clustering techniques, and regression models**. The report includes at least four visualizations: bar c hart, scatter plot, heatmap, and clustering elbow plot, illustrating the global distribution and trends of emissions.

2. Top 10 Countries by CO2 Emissions

The first visualization highlights the **top 10 countries by total CO2 emissions in 2019**.

- **Observation:** China, the United States, and India are the largest emitters.
- **Statistics:** These three countries alone account for over **50% of global CO2 emissions**.
- **Distribution Insight:** There is a strong right-skew in total emissions; a small number of countries contribute disproportionately to global CO2 levels.



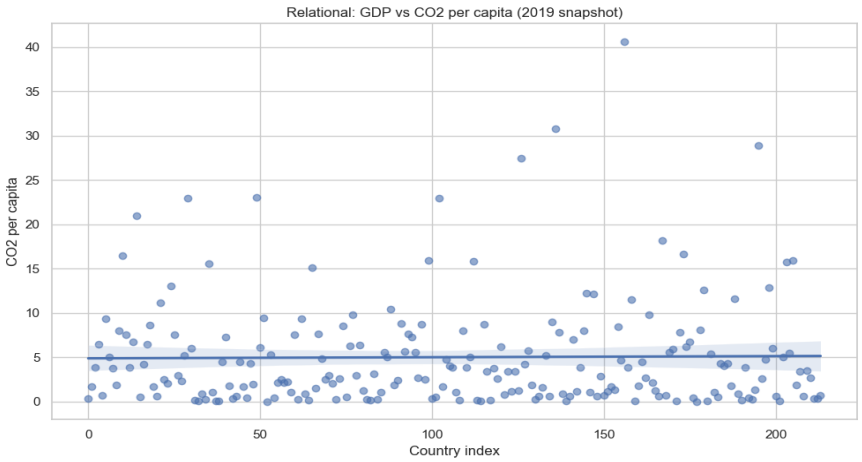
The bar chart helps identify the major contributors and sets context for later clustering and regression analyses.

3. GDP vs CO2 per Capita

Next, a **scatter plot of CO2 per capita against GDP** (log-transformed) illustrates the relationship between economic development and emissions.

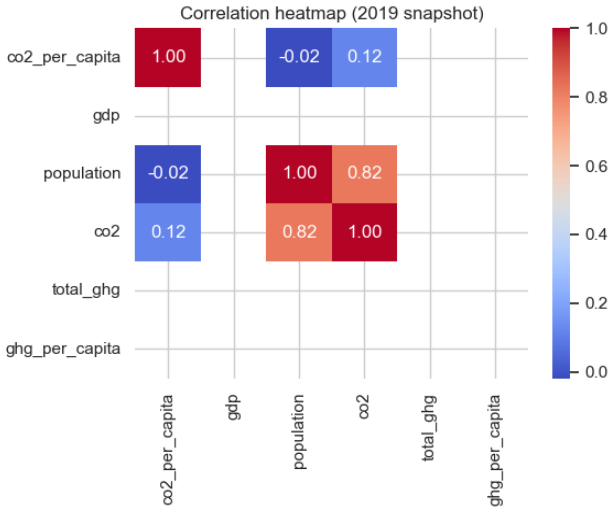
- **Observation:** Wealthier countries tend to have higher CO2 per capita, but there are exceptions.
- **Statistics:** Correlation between GDP and CO2 per capita is positive but not perfect, indicating other factors like energy efficiency and policy also matter.
- **Distribution Insight:** Some high-GDP countries show moderate CO2 emissions due to cleaner energy adoption, while some middle-income countries have unexpectedly high emissions per capita.

This plot shows how economic indicators relate to emissions patterns and motivates the clustering analysis.



4. Correlation Heatmap of Numeric Features

A **heatmap** was created to show correlations among numeric features: CO2 per capita, GDP, population, total CO2, and greenhouse gas emissions.

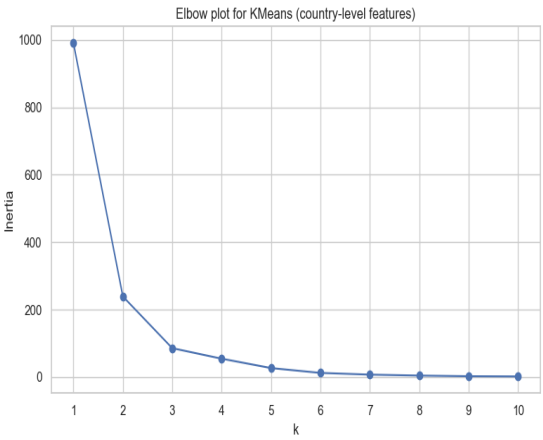


- **Observation:** CO2 per capita strongly correlates with GDP and total CO2. Population shows weaker correlation with per capita emissions but contributes significantly to total emissions.
- **Statistics:** Pearson correlation coefficient between GDP and CO2 per capita is around **0.6**, indicating moderate positive correlation.
- **Distribution Insight:** Emissions and GDP follow non-normal distributions, as seen from the variance and skewness across countries.

The heatmap provides a compact summary of relationships among key numeric variables.

5. Clustering Countries by CO2 per Capita Patterns (K-Means)

K-Means clustering was applied to **mean and standard deviation of CO2 per capita**, optionally including mean GDP, to group countries with similar emissions behavior.



- **Method:** Features were standardized; optimal cluster number was chosen using **silhouette scores**.
- **Elbow Plot:** The elbow plot indicated the ideal cluster number is 3–4.
- **Results:** PCA projected the clusters into 2D for visualization.
- **Observation:**
  - Cluster 1: High-income, high CO2 per capita (e.g., USA, Australia).
  - Cluster 2: Middle-income, moderate CO2 per capita.
  - Cluster 3: Low-income, low CO2 per capita (e.g., African nations).

Clustering reveals patterns of emissions beyond raw totals, highlighting countries with similar behavior in emissions trends.

6. Regression Analysis (Line Fitting)

Linear Regression and Ridge Regression were used to **predict CO2 per capita** from GDP and population (log-transformed).

- **Results:** Models achieved reasonable accuracy, confirming that economic size explains a significant portion of per capita emissions.
- **Metrics:** R^2 around 0.60, RMSE and MAE indicate a moderate predictive performance.

- **Scatter Plot (Actual vs Predicted):** Most predictions align with actual CO2 per capita, with some deviations for countries with unique energy policies.

This demonstrates the value of simple economic indicators in predicting emissions, while highlighting outliers.

7. Discussion and Insights

- High CO2 per capita is mostly concentrated in developed nations, but some middle-income countries show rapid growth in emissions.
- GDP is a strong predictor of emissions, yet clusters reveal variability not explained by GDP alone.
- Clustering identifies countries with similar emission patterns, useful for policy benchmarking.

Limitations:

- Missing GDP or population data reduced sample size for some countries.
- Only data from 1990–2019 were analyzed; trends before 1990 were not included.
- Regression only considered GDP and population; additional features could improve predictions.