מחט בערימת דאטה (67978) | תשפ"ד | פרויקט גמר

אחמד דלאשה 324059856 | סלאח מחאמיד 212001432 | סיוואר מנסור 204527680

30 בספטמבר 2024

Project Title

Analyzing CO2 Emissions Trends and Sector Contribution: A Clustering Approach to Understanding Global Emissions Patterns

Team Member Info

1. Ahmad Dallashe, ahmad.dallashe@mail.huji.ac.il, ahmad_dall7
2. Salah Mahmied, salah.mahmied@mail.huji.ac.il, salah.taher
3. Siwar Mansour, siwar.mansour@mail.huji.ac.il, siwarmansour99

Problem Description

The global rise in CO2 emissions is one of the most pressing environmental issues facing the world today, contributing significantly to climate change and global warming. Understanding the factors driving CO2 emissions at a national level, particularly how economic activity, population, and energy consumption contribute to emissions, is critical for formulating effective environmental policies. The goal of this project is to analyze global CO2 emissions trends and identify patterns in how different sectors (e.g., coal, oil, and gas) and economic activities (GDP, population) contribute to emissions. Specifically, this project aims to:

1. Identify key trends in CO2 emissions across a diverse range of countries over time.
2. Evaluate the impact of economic factors such as GDP and GDP per capita on CO2 emissions, both in total and on a per capita basis.
3. Cluster countries into groups based on their CO2 emissions and economic factors, to identify similarities and differences in their environmental impact.
4. Understand the contribution of different sectors (coal, oil, gas) to the total CO2 emissions of each country and how these sectors vary in importance between different economies.

The hypothesis behind this analysis is that countries with similar economic structures or energy reliance will exhibit similar patterns of CO2 emissions, regardless of their geographic or political context. The results of this project will provide insights into which countries are contributing most to global CO2 emissions, both in total and on a per-person basis, and which sectors are the most responsible for this contribution. These insights can inform both policy recommendations and future research aimed at mitigating climate change.

Data

We will include countries from different regions to get a comprehensive view of global emissions patterns. We are thinking of taking the following countries:

1. North America: United States
2. Europe: Germany, UK
3. Asia: China, India
4. Africa: South Africa, Nigeria
5. Latin America: Brazil, Mexico

DataSet

We used the "Our World in Data CO2 and Greenhouse Gas Emissions" dataset from https://github.com/owid/co2-data. We downloaded the CSV file (approximately 13 MB) containing CO2 emissions data for the last three decades (1990-2020). The dataset includes variables such as total CO2 emissions, CO2 emissions per capita, sector-specific emissions (coal, oil, gas), GDP, GDP per capita, and population, which were essential for our analysis.

Solution

o address the problem of analyzing CO2 emissions trends and the contribution of different sectors across various countries, we developed a systematic approach involving data preprocessing, feature engineering, and the application of two unsupervised clustering algorithms: K-Means Clustering and Hierarchical Agglomerative Clustering (HAC).

Data Preprocessing

- Data Source: We utilized the "Our World in Data CO2 and Greenhouse Gas Emissions" dataset, which provides comprehensive data on CO2 emissions, energy consumption, and economic indicators for countries worldwide.
- Data Cleaning:
  - Handling Missing Values: Removed non-critical columns with excessive missing data (e.g., trade_co2, cement_co2, trade_co2_share).
  - Imputation: Filled missing values in essential columns (population, co2, gdp, etc.) using forward-fill and interpolation methods to maintain data continuity.
- Data Filtering:
  - Country Selection: Focused on nine countries representing a mix of developed and developing economies with significant CO2 emissions: United States, China, India, Germany, Brazil, United Kingdom, Japan, Russia, and Canada.
  - Time Frame: Selected data from 1990 to 2020 to analyze recent trends.

Feature Engineering

- Calculating GDP per Capita: Computed GDP per capita by dividing GDP by population to normalize economic output across countries.
- Feature Selection:
  - Economic Indicators: gdp, gdp_per_capita
  - CO2 Emissions Metrics: co2, co2_per_capita
  - Sector-Specific Emissions: coal_co2, oil_co2, gas_co2
  - Energy Consumption: primary_energy_consumption
  - Global Contribution: share_global_co2
- Data Scaling: Standardized features using StandardScaler from Scikit-learn to ensure all features contributed equally to the clustering algorithms

Clustering Algorithms

K-Means Clustering

- Algorithm Overview: Partitioned the dataset into a predefined number of clusters (k) by minimizing the variance within each cluster.
- Implementation:
  - Number of Clusters (k): Set n_clusters=3 based on domain knowledge and validated using the Elbow Method.
  - Feature Sets: Applied clustering on different combinations of features to explore various relationships:
    - Feature Set 1: ['co2', 'gdp', 'oil_co2', 'gas_co2', 'gdp_per_capita']
    - Feature Set 2: ['co2_per_capita', 'gdp_per_capita', 'coal_co2', 'oil_co2', 'gas_co2']
    - Feature Set 3: ['co2', 'primary_energy_consumption', 'share_global_co2', 'coal_co2', 'oil_co2']
- Visualization: Created scatter plots to visualize clusters in relation to key variables like GDP per capita and CO2 emissions

Evaluation

Evaluation Criteria

To assess the performance and success of our clustering methods, we established the following evaluation criteria:

1. Cluster Cohesion and Separation: Measured how closely related the data points within the same cluster are (cohesion) and how well-separated different clusters are from each other (separation).
2. Silhouette Score: Utilized as a quantitative metric to evaluate the quality of clustering. The silhouette score ranges from -1 to 1, where a higher value indicates better clustering performance.
3. Interpretability of Clusters: Assessed based on how meaningfully the clusters group countries with similar CO2 emissions patterns and economic factors, aligning with real-world knowledge.
4. Consistency Between Methods: Compared results from both K-Means and Hierarchical Agglomerative Clustering (HAC) to ensure robustness and validate findings.

Definition of Success

Success in this context is defined by:

- Identifying Distinct Clusters: Successfully grouping countries into clusters that reveal meaningful patterns in CO2 emissions relative to economic and sectoral factors.
- Insights into CO2 Emissions Trends: Gaining a deeper understanding of how different countries contribute to global CO2 emissions and how economic factors influence these emissions.
- Policy-Relevant Findings: Deriving insights that could inform environmental policy decisions aimed at mitigating climate change.

Ensuring Findings Are Significant

To ensure that our findings are not due to chance:

- Statistical Validation: Used the silhouette score to validate the clustering results statistically.
- Cross-Validation of Methods: Applied two different clustering algorithms and compared the results to see if similar patterns emerged independently.
- Robust Data Preprocessing: Employed thorough data cleaning and preprocessing to minimize the impact of missing or erroneous data.

Setup

Experimental Design

Our experiments were set up as follows:

1. Data Selection: Chose nine countries with significant impacts on global CO2 emissions and diverse economic backgrounds to ensure a representative analysis.
2. Feature Selection: Selected relevant features based on domain knowledge, focusing on those most likely to influence CO2 emissions (e.g., GDP, energy consumption, sectoral emissions).
3. Clustering Algorithms: Implemented both K-Means and HAC to cluster the data.
4. Number of Clusters: Decided on three clusters based on the Elbow Method and domain knowledge, balancing simplicity and interpretability.

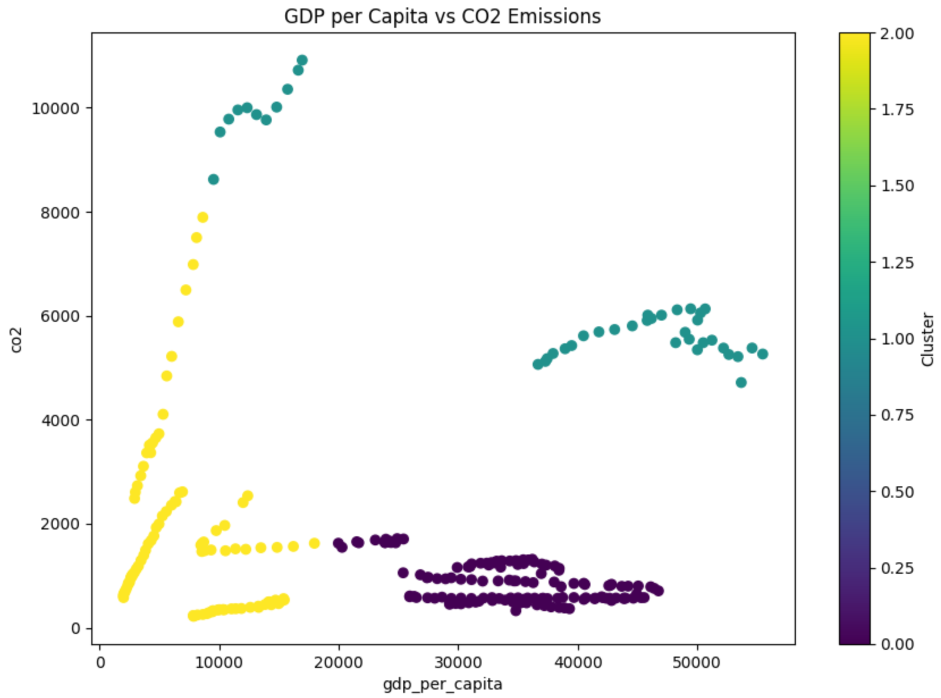
Avoiding Biases

- Data Bias: Mitigated by selecting countries from different regions and economic statuses.
- Feature Bias: Standardized all features to prevent variables with larger scales from dominating the clustering process.
- Algorithmic Bias: Used multiple clustering methods to reduce reliance on a single algorithm's assumptions.

Results

1. **K-Means Clustering**
   We applied K-Means clustering with n=3n=3 clusters across different feature sets to understand how countries group based on CO2 emissions and economic factors.
   (a) GDP per Capita vs. CO2 Emissions
   - Cluster Assignments:
     - Cluster 0: Brazil, Canada, Germany, India, Japan, Russia, United Kingdom
     - Cluster 1: China, United States
   - Observations:
     - Cluster 0 includes a mix of developed and developing countries. Developed nations like Germany, Japan, Canada, and the United Kingdom have high GDP per capita but manage to maintain relatively lower total CO2 emissions compared to Cluster 1. Developing countries like Brazil, India, and Russia have lower GDP per capita and total CO2 emissions.
     - Cluster 1 consists of China and the United States, the two largest CO2 emitters globally. The United States has a high GDP per capita and high CO2 emissions, while China has a lower GDP per capita but extremely high CO2 emissions due to its large industrial base and population.
   - Silhouette Score: 0.656, indicating well-defined clusters.
   - Inference from Plot:

Figure 1:

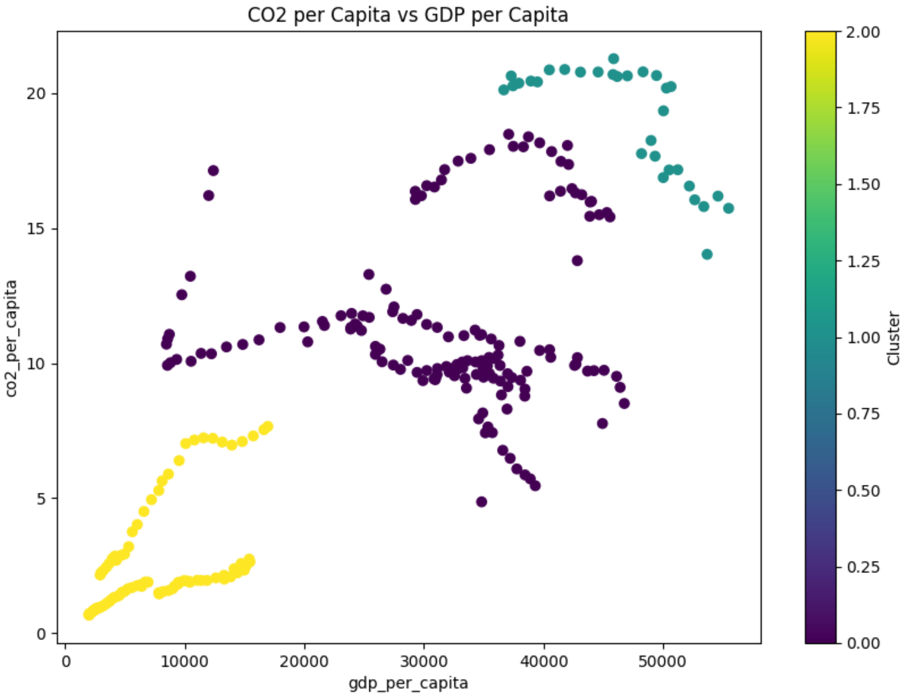


- The plot reveals a positive correlation between GDP per capita and total CO2 emissions. China and the United States stand out significantly, highlighting their substantial contribution to global emissions.
- This clustering suggests that economic scale and industrial activity are major drivers of CO2 emissions.

(b) CO2 per Capita vs. GDP per Capita

- Cluster Assignments:
  - Cluster 0: China, India, Russia
  - Cluster 1: Brazil, Canada, Germany, Japan, United Kingdom
  - Cluster 2: United States
- Observations:
  - Cluster 0 comprises countries with lower GDP per capita and lower CO2 emissions per capita. Despite China's and Russia's high total emissions, their per capita emissions are lower due to large populations.
  - Cluster 1 includes developed countries with high GDP per capita and moderate CO2 emissions per capita. These nations have advanced technologies and stricter environmental regulations that help control per capita emissions.
  - Cluster 2 is solely the United States, characterized by high GDP per capita and the highest CO2 emissions per capita among the countries analyzed.
- Inference from Plot:

Figure 2:

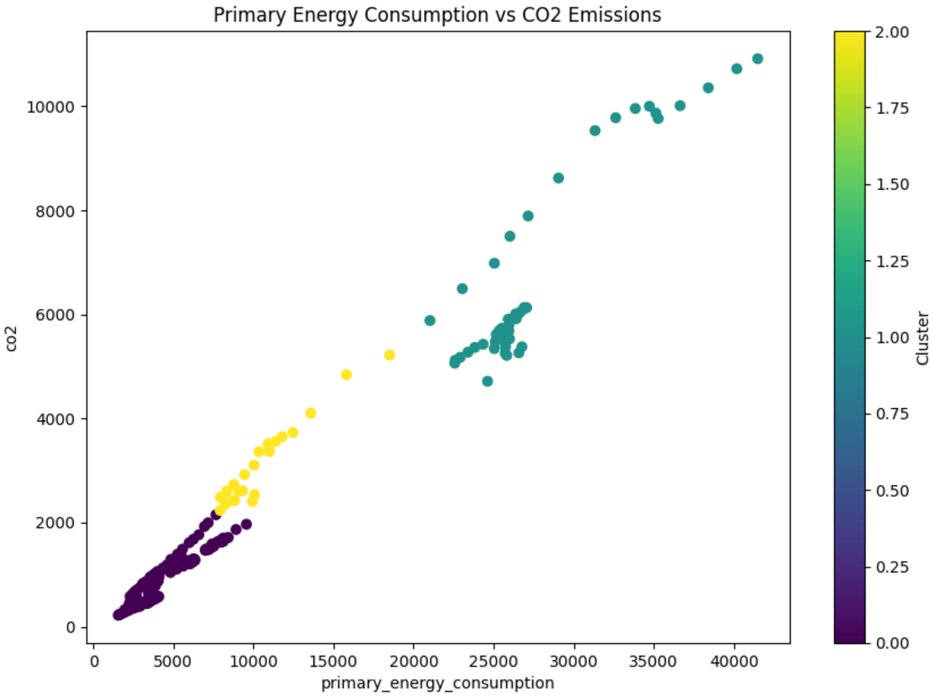


- The plot illustrates disparities in emissions efficiency. The United States appears as an outlier, indicating higher energy consumption and reliance on fossil fuels per person compared to other developed countries.
- This suggests potential for the U.S. to reduce emissions through energy efficiency and policy changes.

(c) Primary Energy Consumption vs. CO2 Emissions

- Cluster Assignments:
  - Cluster 0: Brazil, Canada, Germany, India, Japan, Russia, United Kingdom
  - Cluster 1: China, United States
- Observations:
  - Cluster 0 consists of countries with varying levels of energy consumption and CO2 emissions but generally lower than those in Cluster 1.
  - Cluster 1, containing China and the United States, shows extremely high primary energy consumption and CO2 emissions, underscoring their significant roles in global energy use and emissions.
- Silhouette Score: 0.656, indicating strong cluster cohesion.
- Inference from Plot:

Figure 3:



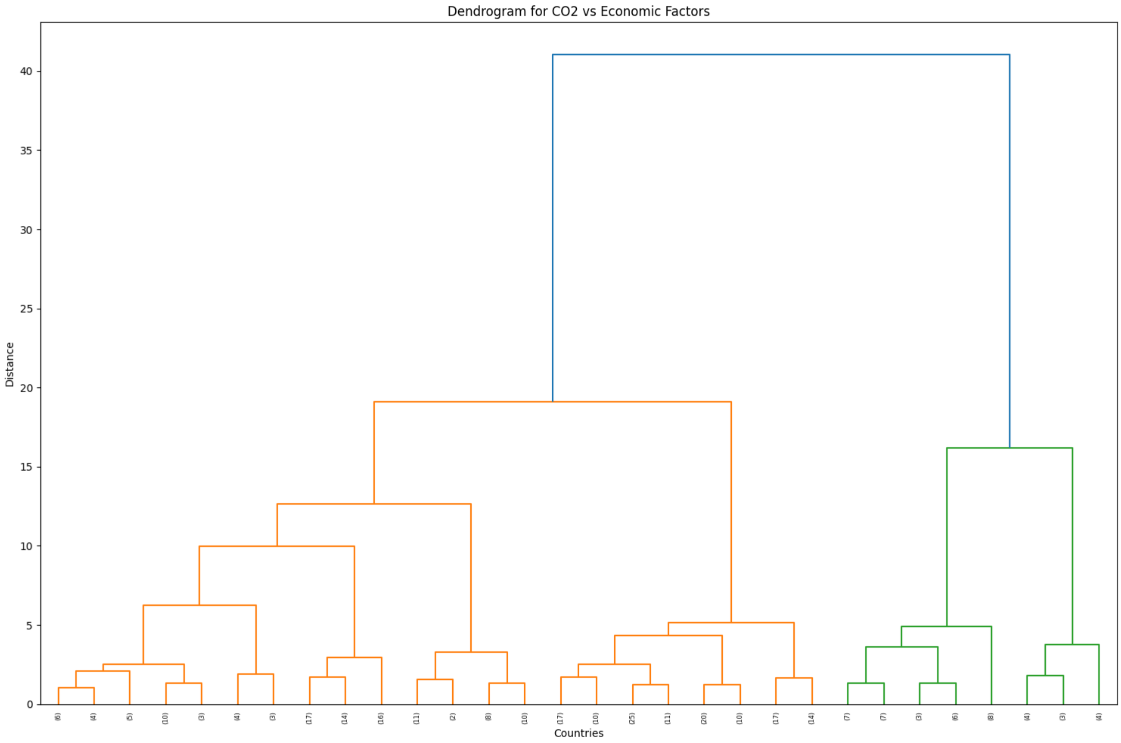
- The strong correlation between primary energy consumption and CO2 emissions is evident.
- This highlights the importance of energy consumption patterns in influencing total emissions and suggests that energy policy is crucial for emissions reduction.

2. **Hierarchical Agglomerative Clustering (HAC)**
   Using HAC with the Ward linkage method, we explored the hierarchical relationships between countries based on the same feature sets.

   (a) GDP per Capita vs. CO2 Emissions

   - Cluster Assignments:
     - Cluster 0: China, United States
     - Cluster 1: Brazil, China, India, Russia
     - Cluster 2: Canada, Germany, Japan, United Kingdom
   - Observations:
     - Cluster 0 again groups China and the United States due to their high CO2 emissions..
     - Cluster 1 includes developing countries with lower GDP per capita and varying CO2 emissions.
     - Cluster 2 comprises developed nations with high GDP per capita and relatively lower CO2 emissions, indicating efficient energy use and effective environmental policies.
   - Inference from Plot:

Figure 4:

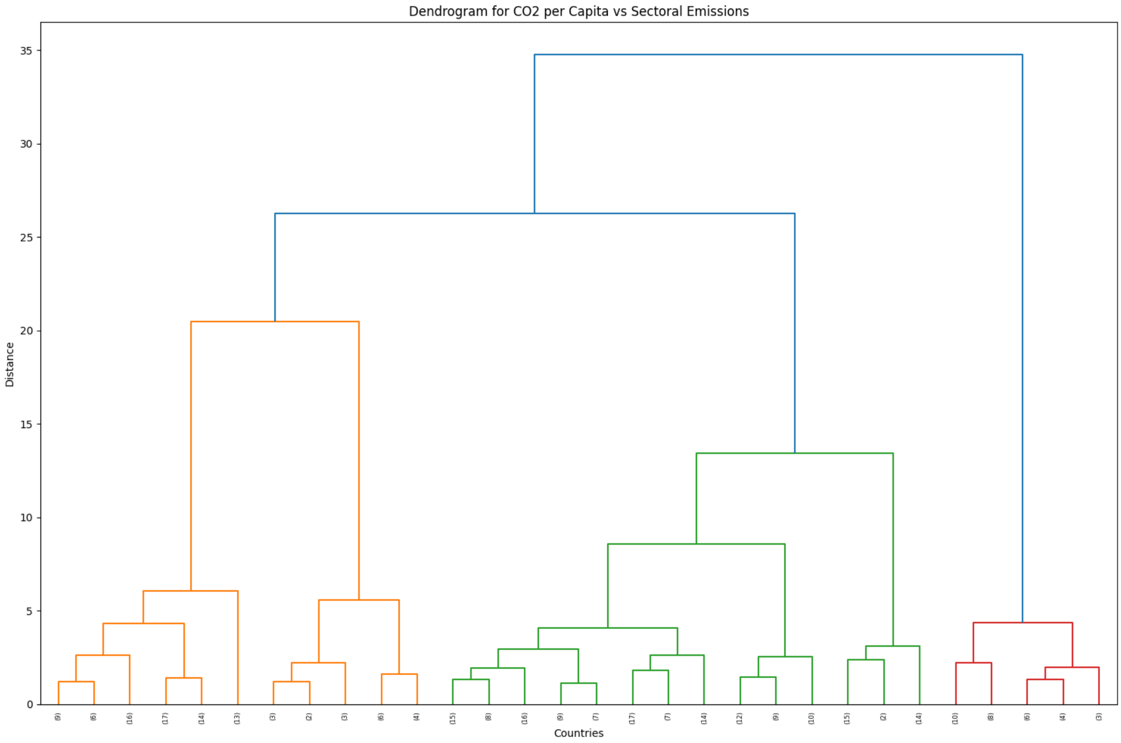


- The dendrogram illustrates the hierarchical separation, with China and the United States branching off at higher distances due to their significant emissions.
- This suggests a clear distinction between the highest emitters and other countries.

(b) CO2 per Capita vs. GDP per Capita

- Cluster Assignments:
  - Cluster 0: Brazil, China, India
  - Cluster 1: United States
  - Cluster 2: Canada, Germany, Japan, Russia, United Kingdom
- Observations:
  - Cluster 0 consists of countries with lower GDP per capita and lower CO2 emissions per capita.
  - Cluster 1 is the United States, isolated due to its high emissions per capita.
  - Cluster 2 includes developed countries with high GDP per capita and moderate CO2 emissions per capita.
- Inference from Plot:

Figure 5:

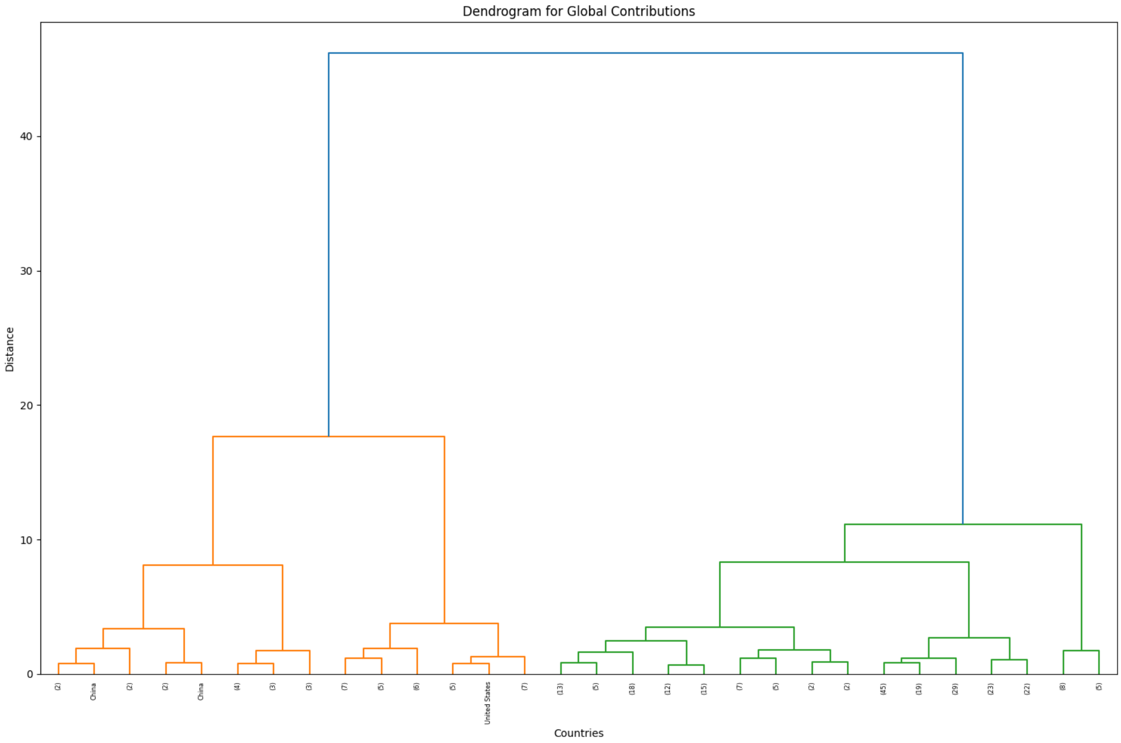


- The dendrogram highlights the United States as an outlier among developed nations regarding emissions per capita.
- This emphasizes the potential for the U.S. to adopt practices from its peers to reduce per capita emissions.

(c) Primary Energy Consumption vs. CO2 Emissions

- Cluster Assignments:
  - Cluster 0: Brazil, Canada, China, Germany, India, Japan, Russia, United Kingdom
  - Cluster 1: China
  - Cluster 2: United States
- Observations:
  - Cluster 0 groups most countries together, indicating similar patterns in energy consumption and CO2 emissions.
  - Cluster 1 isolates China, reflecting its massive increase in energy consumption and emissions over recent decades.
  - Cluster 2 is the United States, separated due to its consistently high energy consumption and emissions.
- Inference from Plot:

Figure 6:



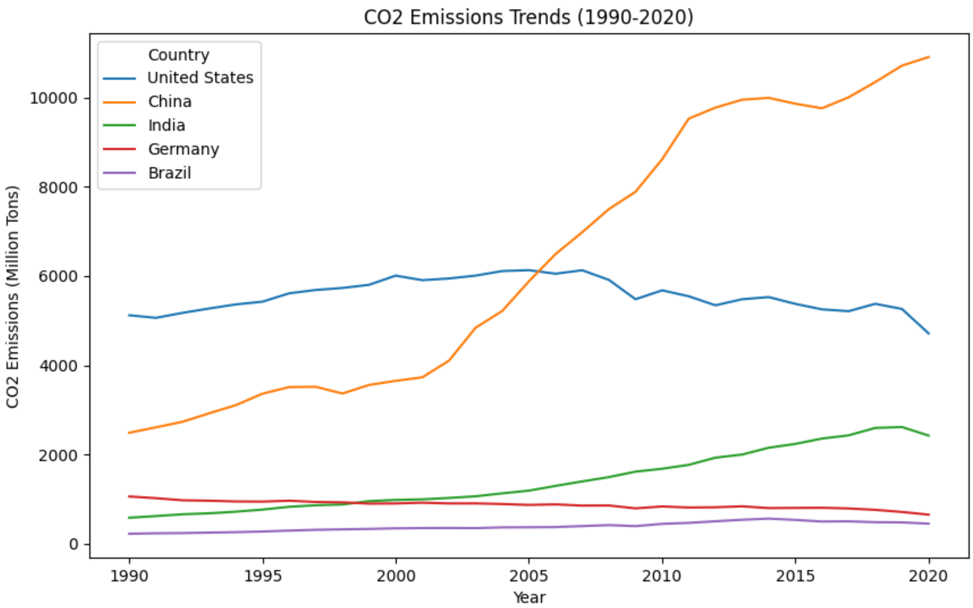
    - The dendrogram shows China and the United States as significant outliers, reinforcing their pivotal roles in global emissions.
    - This suggests that global emissions reduction efforts should prioritize these two nations.

3. **Comparison Between K-Means and HAC**
    - Consistency in Findings:
        - Both clustering methods consistently identified China and the United States as the primary contributors to CO2 emissions.
        - Developed countries with advanced economies and environmental policies tend to cluster together.
    - Differences:
        - K-Means provides distinct cluster boundaries, which are useful for straightforward classification.
        - HAC reveals hierarchical relationships and sub-cluster formations, offering deeper insights into the degrees of similarity among countries.
    - Inference:
        - The similarity in cluster assignments across both methods validates the robustness of our findings.
        - The hierarchical nature of HAC complements K-Means by uncovering the nested structure of the data.

4. CO2 Emissions Trends Over Time
    - Observations:
        - China: Exhibits a sharp increase in CO2 emissions, particularly after 2000, surpassing the United States around 2007 due to rapid industrialization.
        - United States: Shows a gradual decline in emissions since 2010, attributed to shifts toward renewable energy and efficiency improvements.
        - India: Demonstrates a steady rise in emissions, reflecting economic growth and increased energy demand.
        - Germany: Displays a slight decrease in emissions, highlighting successful implementation of environmental policies and renewable energy adoption.
        - Brazil: Maintains relatively stable emissions with minor fluctuations, possibly due to balanced energy sources and deforestation policies.
    - Inference:
        - Global Trend: CO2 emissions are generally rising, driven largely by developing economies expanding their industrial capacities.
        - Policy Impact: Developed countries with proactive environmental policies show stabilization or reduction in emissions, indicating policy effectiveness.

  - Future Implications: The data underscores the need for sustainable development practices in rapidly growing economies.

- Figure 7:



Visualization

Choice of Visualization

- Scatter Plots:
  - Purpose: To display relationships between two quantitative variables and visualize cluster groupings.
  - Application: Used in both K-Means and HAC to plot GDP per capita against CO2 emissions, and CO2 emissions per capita against GDP per capita.
- Dendrograms:
  - Purpose: To represent the hierarchical structure of data, showing how clusters are formed at various levels.
  - Application: Used in HAC to visualize how countries merge into clusters based on similarity in emissions and economic factors.
- Line Charts:
  - Purpose: To illustrate trends over time for individual countries.
  - Application: Plotted CO2 emissions from 1990 to 2020 for key countries to observe changes and patterns.
- Explanation
  - Effectiveness:
  - The chosen visualizations effectively communicate complex relationships and patterns within the data.
  - They allow for quick identification of outliers, trends, and cluster compositions.
  - Interpretation:
  - Scatter Plots: Facilitate understanding of how countries compare on key metrics and how they are grouped by the clustering algorithms.
  - Dendrograms: Provide insights into the hierarchical similarities and differences between countries, revealing deeper structural relationships.
  - Line Charts: Highlight temporal changes, indicating whether emissions are rising, falling, or stable.

Impediments

Challenges Encountered

1. Data Quality Issues:
   - Missing Data: Some countries lacked complete data for all variables and years, which could affect cluster accuracy.
   - Resolution: Employed interpolation and forward-filling to estimate missing values, ensuring a more complete dataset.
2. Optimal Number of Clusters:
   - Determining k: Choosing the appropriate number of clusters was critical and non-trivial.
   - Resolution: Used the Elbow Method and evaluated silhouette scores to select k=3k=3 as a balance between simplicity and explanatory power.
3. Feature Correlation:
   - Multicollinearity: High correlation between features like GDP and GDP per capita could bias clustering.
   - Resolution: Standardized features and considered using Principal Component Analysis (PCA) but ultimately retained features for interpretability.
4. Interpretation of Clusters:
   - Complex Relationships: Economic and environmental factors are interrelated, making it challenging to attribute causality.
   - Resolution: Focused on observable patterns and supported interpretations with domain knowledge.

Dealing with Issues

- Data Validation: Cross-checked data with alternative sources where possible to ensure reliability.
- Methodological Rigor: Applied consistent preprocessing and scaling techniques to maintain data integrity across analyses.
- Collaborative Review: Discussed findings with peers to validate interpretations and ensure that conclusions were well-founded.

Future Work

Future work could involve expanding the analysis to include more countries and a longer historical timeframe to gain a comprehensive global perspective on CO2 emissions trends. Incorporating additional variables such as renewable energy adoption and environmental policies could provide deeper insights into factors influencing emissions. Applying advanced clustering techniques or time-series analysis may also uncover new patterns and enhance the robustness of the findings.

Conclusion from Results

The analysis confirms that global CO2 emissions have been generally rising over the past three decades, primarily due to rapid industrialization and economic growth in developing nations like China and India. Our clustering results consistently identified China and the United States as the most significant contributors to CO2 emissions, both in total output and energy consumption. These two countries often formed distinct clusters, underscoring their outsized impact on global emissions.

Developed countries such as Germany, Japan, Canada, and the United Kingdom clustered together, characterized by high GDP per capita but relatively lower CO2 emissions per capita. This suggests that economic prosperity does not necessarily require proportional increases in emissions. Effective environmental policies, technological advancements, and a shift toward renewable energy sources in these nations have likely contributed to more sustainable emission levels.

The consistency between the K-Means and Hierarchical Agglomerative Clustering results strengthens the validity of our findings. Both methods revealed similar patterns and groupings, highlighting the strong relationship between economic factors and CO2 emissions. Hierarchical clustering provided additional insights into the degrees of similarity among countries, enhancing our understanding of the hierarchical nature of global emissions patterns.

These findings emphasize the crucial role of energy consumption patterns and economic structures in influencing CO2 emissions. The disparities in emissions per capita among developed nations indicate opportunities for high-emitting countries to adopt best practices from their peers. Policy interventions focusing on energy efficiency, adoption of cleaner technologies, and sustainable industrial practices could significantly impact emissions trends.

In conclusion, our project demonstrates that clustering techniques are effective tools for analyzing complex environmental data, providing valuable insights into global CO2 emissions trends. The rising emissions in developing economies highlight the urgent need for international collaboration and targeted policies to address climate change. Our results support the hypothesis that countries with similar economic structures or energy reliance exhibit comparable emissions patterns, offering a foundation for future research and policy development aimed at mitigating global CO2 emissions.