MIT Vishwaprayag University
School of Engineering & Technology

EDADV LAB – Group Project Report

Exploratory Data Analysis on IEA Fossil Fuel Subsidies Database

Submitted By:

- Onkar Tavarkhed (SCFU124031)
- Vinit Rathod (SCFU124027)
- Shatakshi Vaikunthe (SCFU124012)
- Pratiksha Naikwadi (SCFU124020)

Course: EDADV LAB
Academic Year: 2024–2025

Contents

Chapter 1

Introduction

Energy subsidies play a crucial role in shaping the global economy, as they directly affect energy consumption patterns, environmental sustainability, and national government expenditures. The International Energy Agency (IEA) provides a comprehensive global database that tracks fossil fuel subsidies across multiple countries and energy products, including oil, electricity, gas, and coal. Using this extensive dataset, the present project conducts a detailed Exploratory Data Analysis (EDA) covering the period from 2010 to 2023. The purpose of this analysis is to gain a clear understanding of global subsidy patterns and examine how different countries and regions allocate energy subsidies. It also aims to explore trends across various energy products and analyze how these subsidies have evolved over time. Furthermore, advanced analytical approaches such as Principal Component Analysis (PCA) and clustering techniques are applied to identify similarities, differences, and behavioral groupings among countries. This structured investigation not only reveals key insights into global energy subsidy distribution but also helps in understanding the broader economic and policy implications. The entire analysis is implemented using Python libraries such as Pandas, NumPy, and Matplotlib, following standard EDA methodologies to ensure clarity, accuracy, and reliability

This project performs a detailed **Exploratory Data Analysis (EDA)** on the IEA Fossil Fuel Subsidies dataset (2010–2023). The goal is to: Understand global subsidy patterns, Compare countries and regions, Analyze product-wise and time-series trends, Apply PCA and clustering to identify behavioral groups. The analysis is performed using Python (Pandas, NumPy, Matplotlib) and structured according to standard EDA methodology.

Chapter 2

Dataset Description

The dataset used in this project is the IEA Fossil Fuel Subsidies Dataset, covering the period from 2010 to 2023. It provides a comprehensive record of fossil fuel subsidies reported by various countries across multiple energy products. The data includes annual subsidy values expressed in million USD, allowing consistent comparison between countries and across years. Each entry represents a specific country–product combination, capturing the subsidies provided for oil, electricity, gas, or coal. The dataset spans both developed and developing nations, offering a global perspective on subsidy allocation. It contains country names, product categories, subsidy amounts, and corresponding time periods, forming the foundation for multi-dimensional analysis. Time-series characteristics of the dataset enable observation of long-term trends, fluctuations, and policy-driven changes in energy subsidy behavior. Product-wise segmentation makes it possible to study how different fuels contribute to national expenditure and energy strategy. The dataset also supports cross-country comparison, revealing significant variations in subsidy distribution influenced by economic priorities, energy demand, and government policies. Missing values in some years reflect inconsistencies in reporting or changes in national accounting practices, which are handled during preprocessing. The structured numerical format of the dataset makes it suitable for statistical analysis, visualization, PCA, and clustering. Overall, this dataset provides a rich and reliable source for examining the global landscape of fossil fuel subsidies and understanding their role in shaping energy and economic systems.

Chapter 3

Data Preprocessing

The data preprocessing stage involved several critical steps to ensure that the dataset was clean, consistent, and ready for analysis. Initially, all metadata rows present at the top of the dataset were removed, as they did not contribute to the analytical values and could interfere with data loading. The header alignment was then corrected because the original file contained misaligned column names, with product-level information beginning around row 4 and country-level information starting near row 11. These inconsistencies were carefully adjusted so that each column had a proper label. Following this, the year-wise columns, which initially appeared as text, were converted into numeric format to enable mathematical operations and time-series analysis. Empty or partially filled rows and columns were eliminated to avoid missing or irrelevant data affecting the results. Country names and product categories were also standardized to maintain uniformity, especially because slight spelling variations or formatting differences could lead to duplicate entries or grouping errors during analysis. After the cleaning process, the dataset achieved a clear and structured format where product-level entries included categories such as Total, Oil, Electricity, Gas, and Coal. The country-level section became fully organized with 248 countries recorded consistently across 14 years of data. This cleaned structure allowed smooth extraction, visualization, and comparison of global subsidy patterns. With all inconsistencies resolved, the dataset became fully suitable for exploratory data analysis, PCA, clustering, and trend evaluation, ensuring accuracy and reliability throughout the project.

Chapter 4

Product-Level Exploratory Data Analysis

4.1 Global Trend of Total Subsidies (2010–2023)

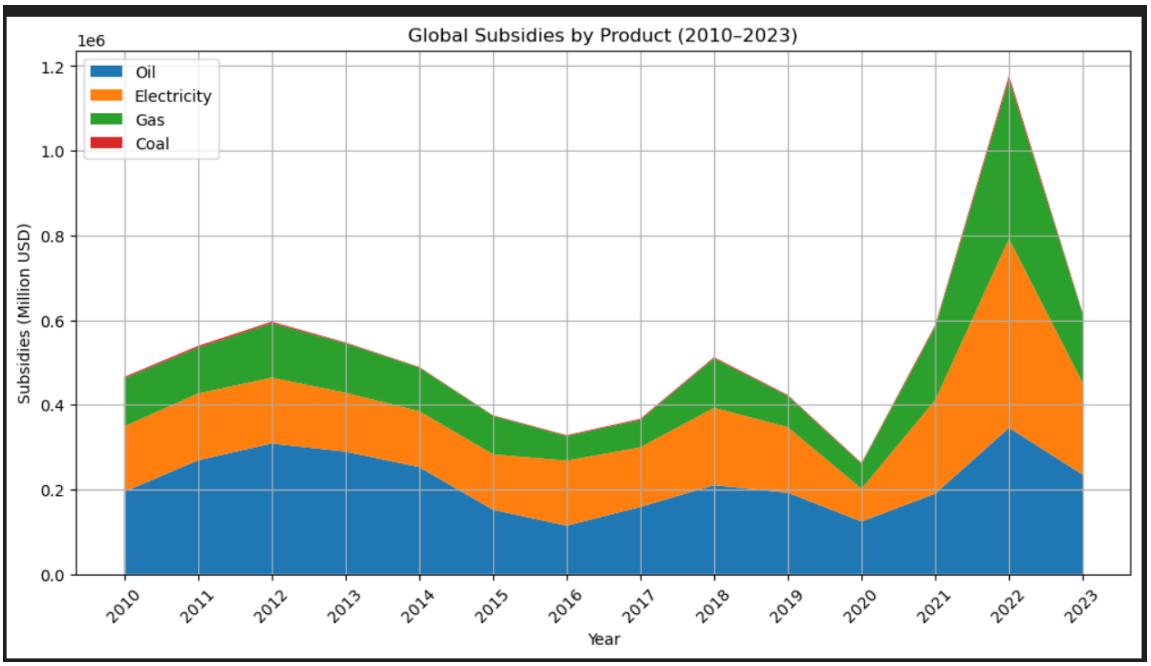


Figure 4.1: Enter Caption

Figure 4.2: Global fossil fuel subsidies trend (2010–2023).

Insight: This stacked area chart shows the yearly distribution of global subsidies for Oil, Electricity, Gas, and Coal from 2010 to 2023. Oil and Electricity consistently receive the highest subsidies over the years. A significant peak is observed in 2022, driven by rising global energy demands and policy changes. Overall, the trend highlights fluctuations influenced by economic and energy market conditions across the examined period.

4.2 Product-Wise Comparison

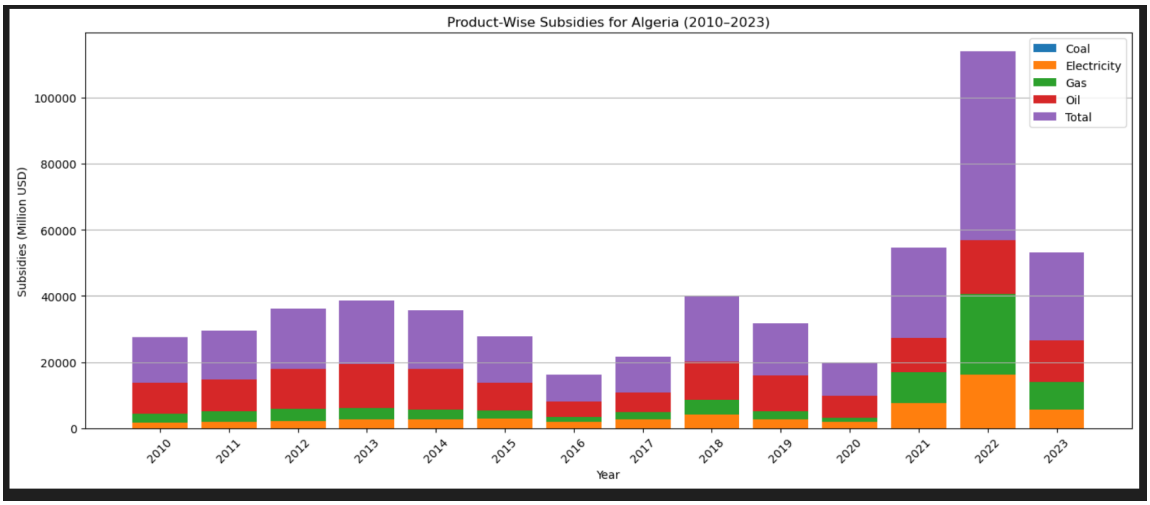


Figure 4.3: Oil vs Gas vs Electricity vs Coal subsidy comparison across years.

Insight: This stacked bar chart illustrates Algeria's subsidies across Coal, Electricity, Gas, and Oil from 2010 to 2023, along with the total subsidy amount each year. Oil and Gas dominate the subsidy expenditure throughout the timeframe. A significant spike is observed in 2022, reflecting heightened support for energy sectors during global market disruptions. The data shows fluctuating patterns, but overall subsidies trend sharply upward in the later years

4.3 Annual Composition Analysis

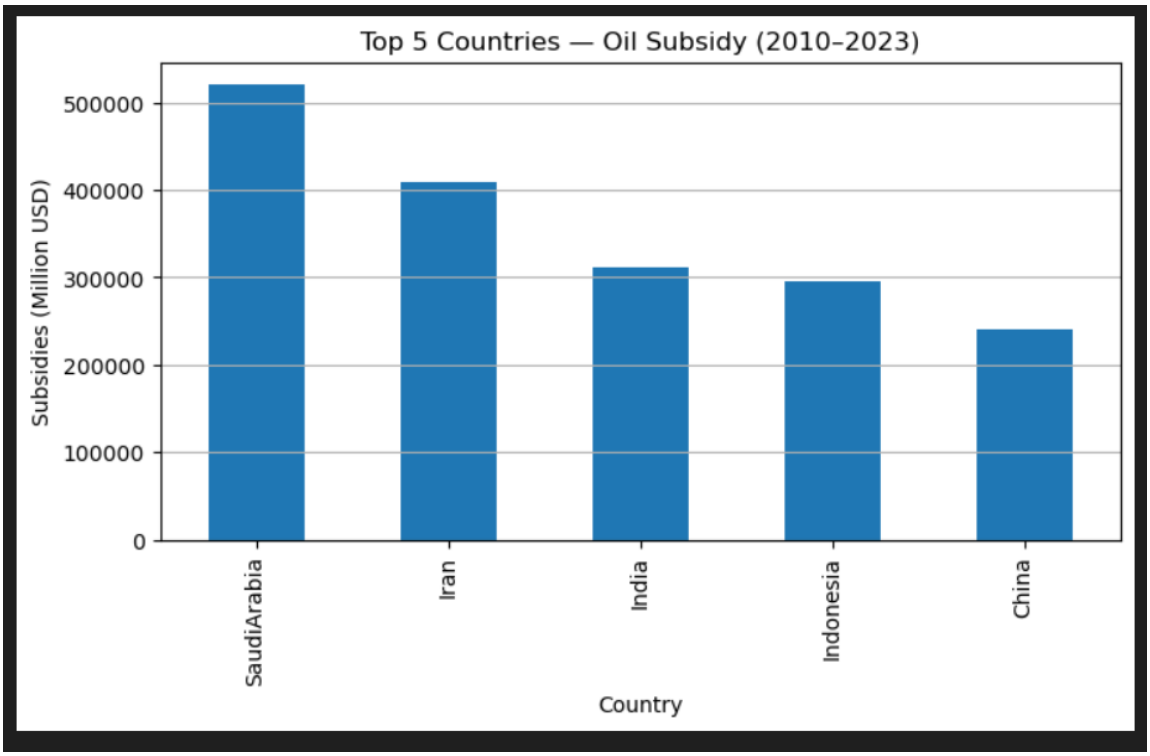


Figure 4.4: Figure: Oil subsidies (in million USD) provided by the top five countries from 2010 to 2023

Insight: Saudi Arabia leads by a significant margin, providing the highest oil subsidies among the top five countries.

Iran follows as the second-highest contributor, though with a noticeable gap from Saudi Arabia.

India and Indonesia show moderate but substantial subsidy levels, reflecting their large populations and energy demands.

China, despite its massive energy consumption, appears fifth, indicating tighter subsidy policies compared to others.

Overall, the chart highlights how oil-dependent economies allocate high subsidies to manage domestic fuel prices and economic stability.

4.4 Heatmap Visualization

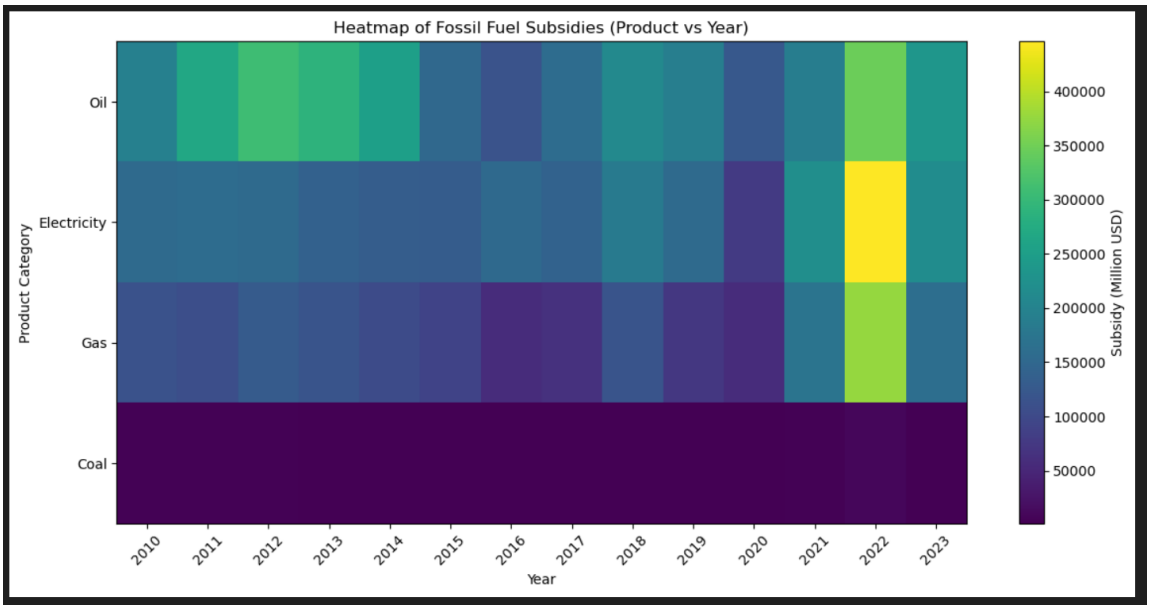


Figure 4.5: Enter Caption

Figure 4.6: Heatmap of subsidies by product and year.

Story: This heatmap visualizes subsidy variations for Oil, Electricity, Gas, and Coal across the years 2010 to 2023. Brighter colors indicate higher subsidy allocation, especially visible in recent years. Oil and Electricity show prominent peaks in 2022, reflecting increased financial support during global energy instability. Coal subsidies remain minimal throughout the period, highlighting a shift away from coal-based energy sources.

Chapter 5

Country-Level Exploratory Data Analysis

5.1 Top Subsidizing Countries

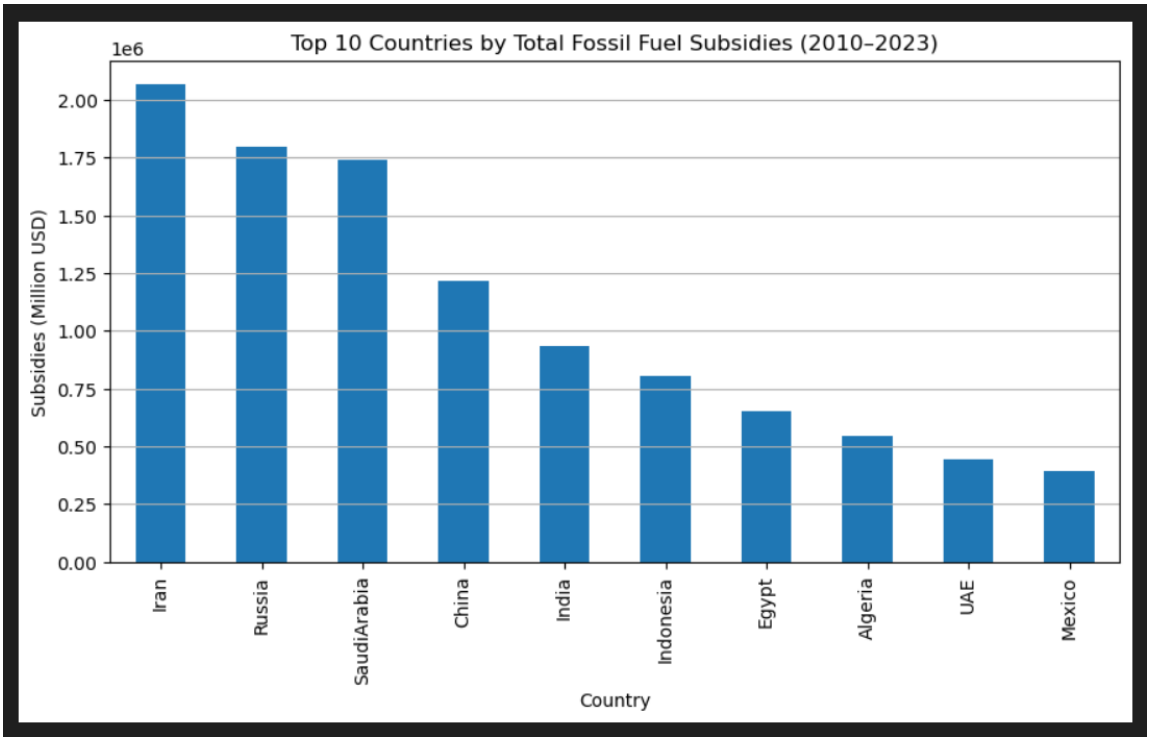


Figure 5.1: Top 10 Countries by Total Fossil Fuel Subsidies (2010–2023

Insight: This bar chart shows the Top 10 countries with the highest fossil fuel subsidies from 2010 to 2023. Iran leads with the largest subsidy spending, followed by Russia and Saudi Arabia. The chart highlights how a few major energy-producing nations account for a significant share of global subsidies. It provides a clear comparison of subsidy levels across countries, measured in million USD

5.2 Country Trend Example

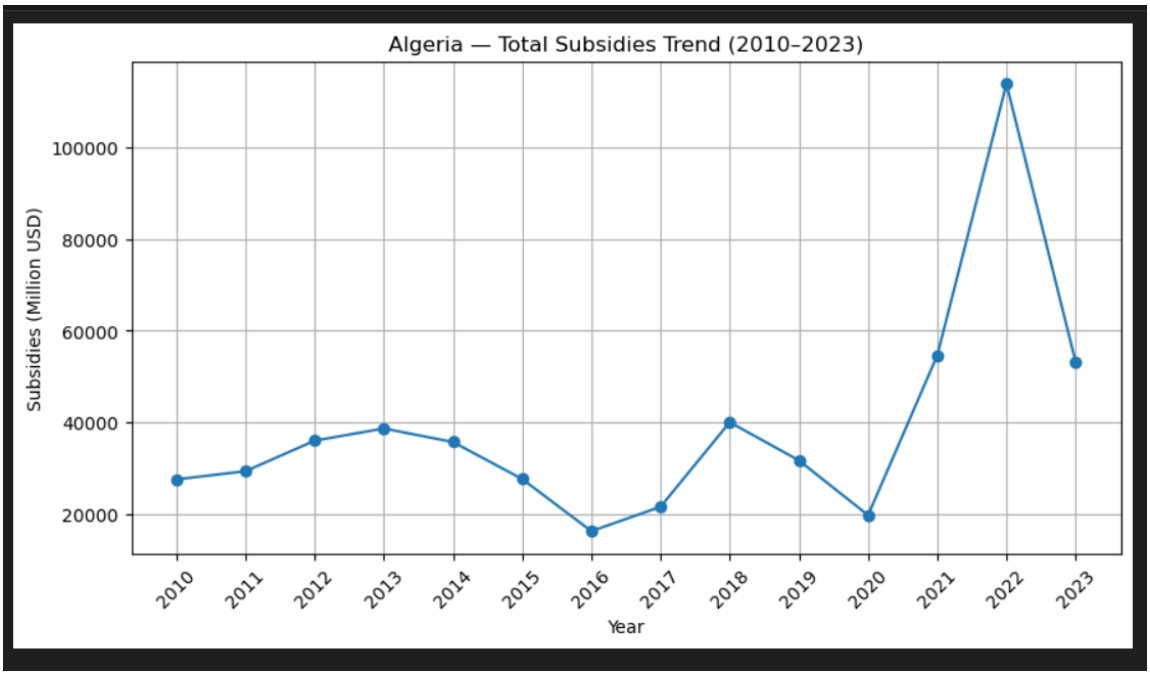


Figure 5.2: Enter Caption

Figure 5.3: Example: Algeria subsidy trend (2010–2023).

Insight: This line chart illustrates Algeria's total fossil fuel subsidies from 2010 to 2023. The trend shows moderate fluctuations until 2020, followed by a sharp rise in 2021 and a peak in 2022. This surge reflects increased energy support during global price volatility. In 2023, subsidies decline again but remain significantly higher than pre-2020 levels

5.3 Multi-Country Comparison

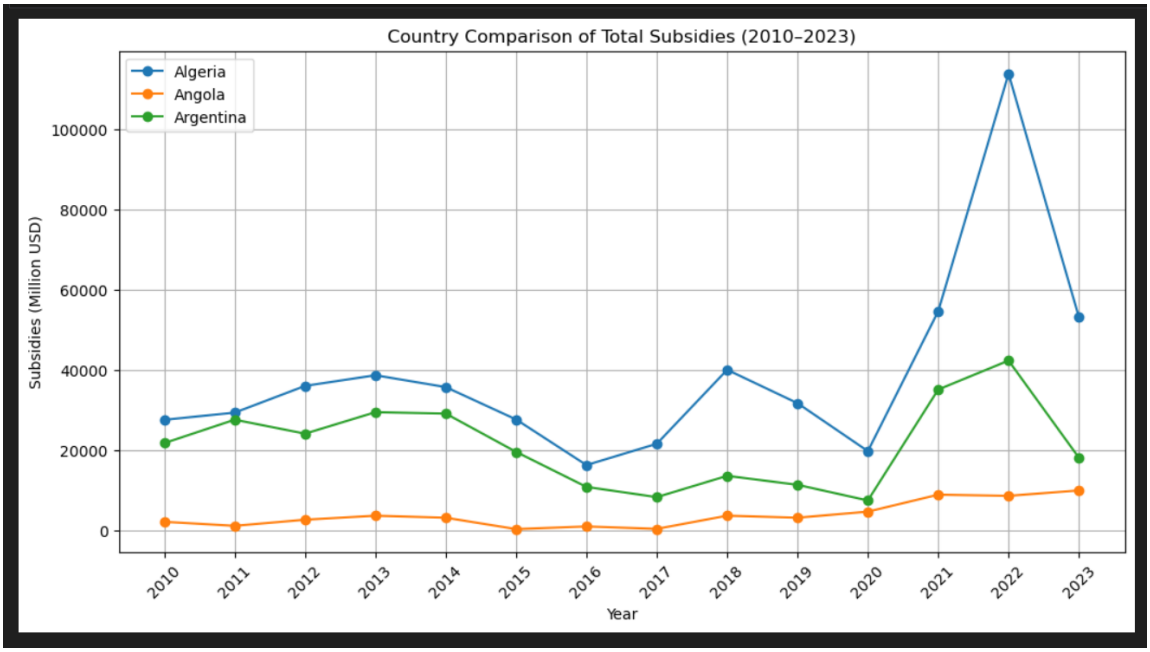


Figure 5.4: Comparison of multiple countries' total subsidy trends.

Insight: This line chart compares total fossil fuel subsidies for Algeria, Angola, and Argentina from 2010 to 2023. Algeria shows the highest and most volatile subsidy levels, with a sharp peak in 2022. Argentina displays moderate fluctuations with a noticeable rise after 2020, while Angola maintains comparatively low subsidy levels throughout the period. The visualization highlights contrasting subsidy patterns among the three countries.

Chapter 6

Regional & Development Status Analysis

6.1 Regional Trends

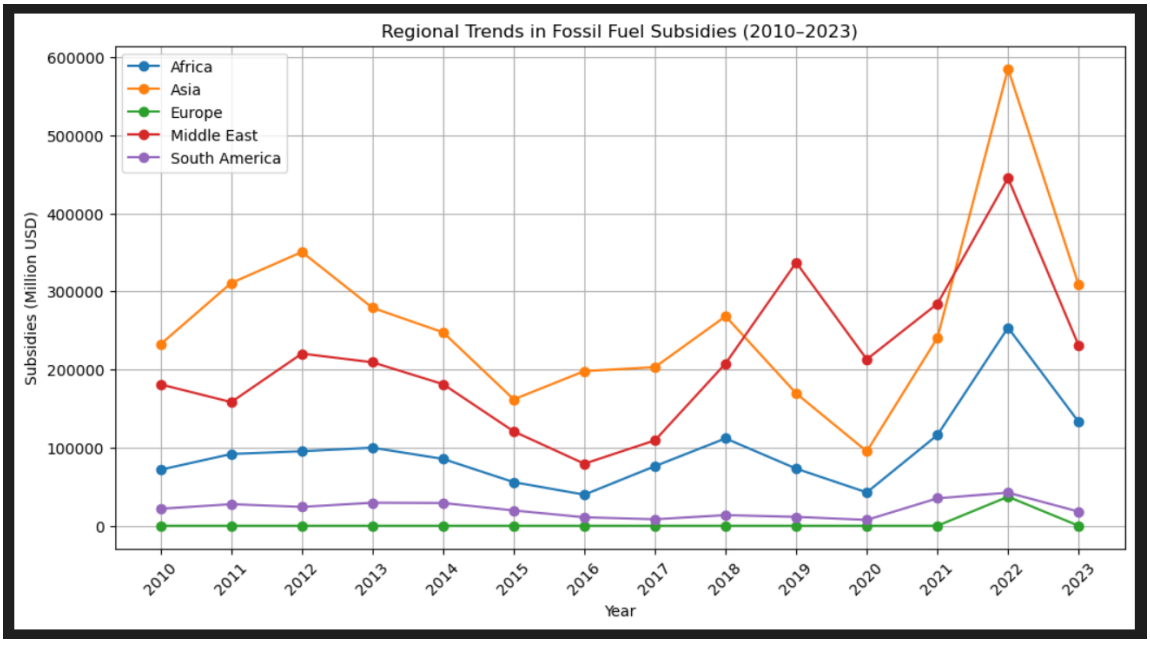


Figure 6.1: Regional subsidy trends (Africa, Alsa, Europe, Middle East, South America etc.).

Insight: This line chart compares fossil fuel subsidy trends across major global regions from 2010 to 2023. Asia consistently records the highest subsidy levels, with a sharp peak in 2022. The Middle East also shows significant fluctuations, especially after 2018. Africa and South America follow moderate patterns, while Europe maintains the lowest subsidy levels throughout the period.

6.2 Developed vs Developing Countries

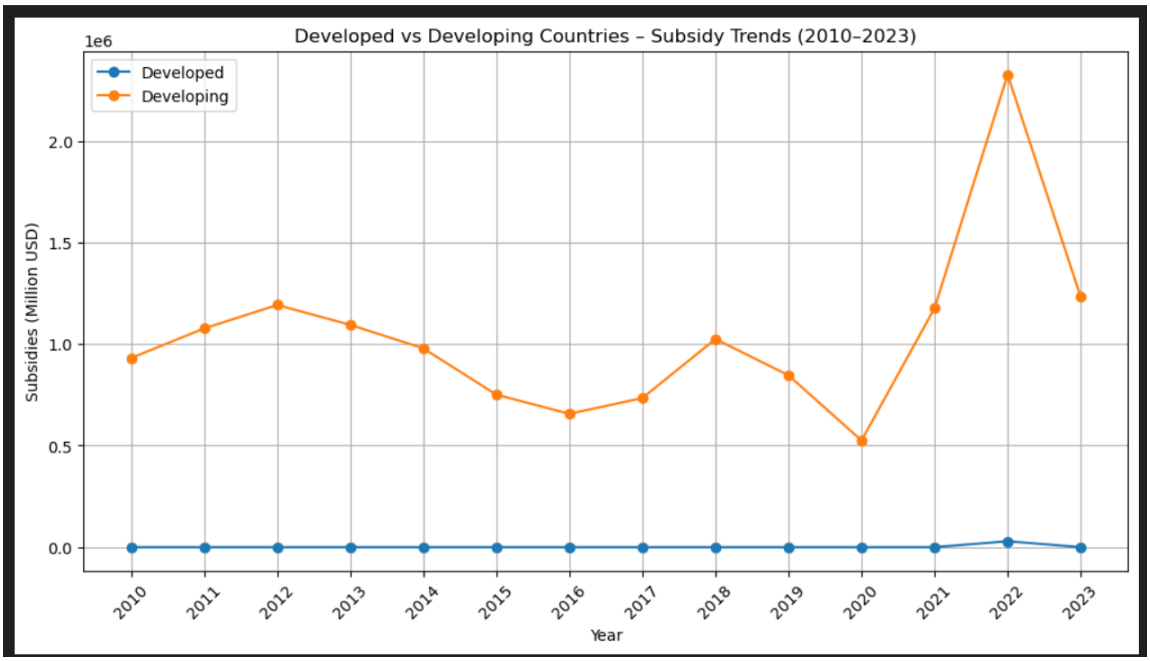


Figure 6.2: Enter Caption

Figure 6.3: Subsidy trends for developed vs developing nations.

Insight: This chart compares fossil fuel subsidy trends between developed and developing nations from 2010 to 2023. Developing countries consistently spend far more on subsidies, with a major spike in 2022 reflecting heightened energy support during global price surges. Developed countries maintain relatively low and stable subsidy levels throughout the period. The gap highlights differing economic capacities and policy priorities between the two groups

Chapter 7

Multivariate Analysis: PCA & Clustering

7.1 PCA Results

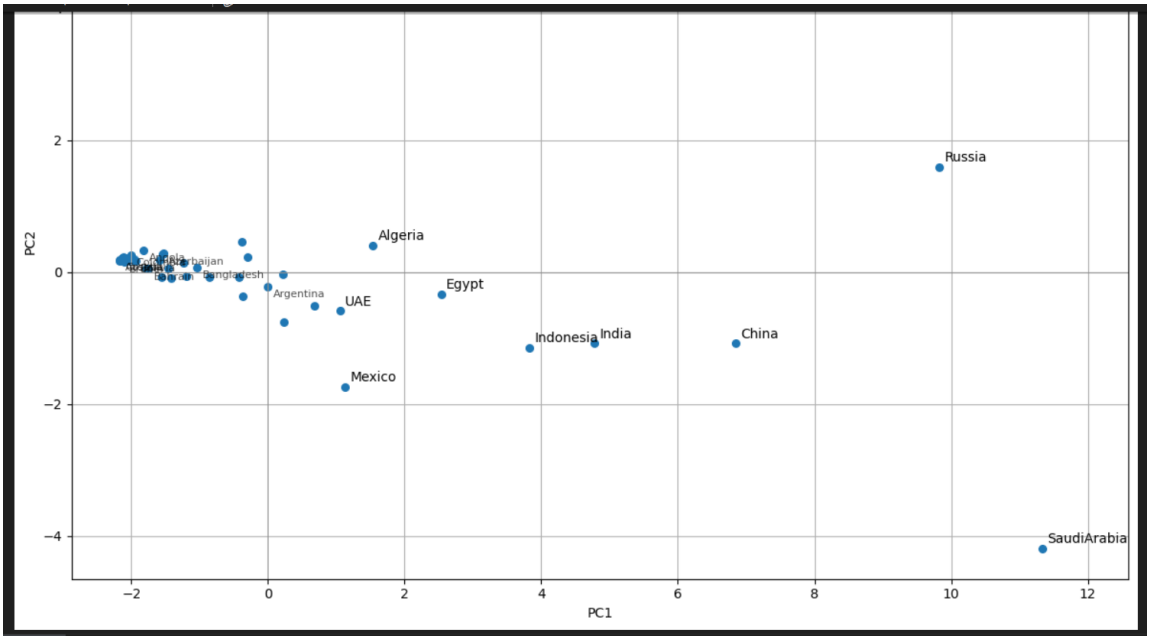
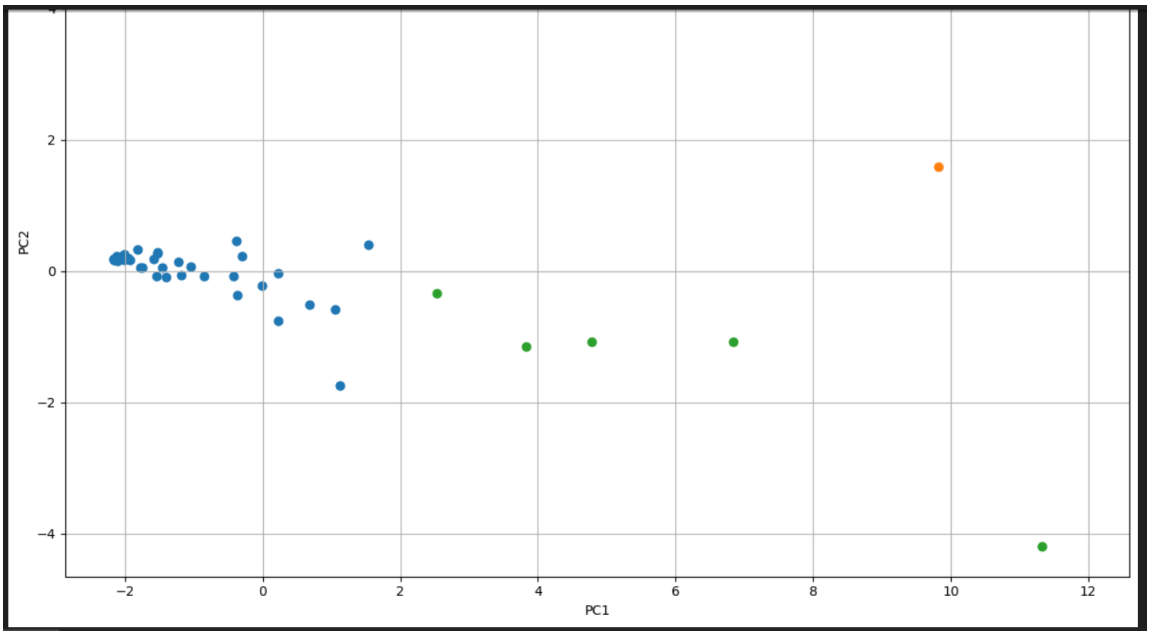


Figure 7.1: PCA scatter plot (PC1 vs PC2).

Insight: This PCA scatter plot visualizes countries based on their fossil fuel subsidy patterns, reducing multidimensional data into two principal components (PC1 and PC2). Most countries cluster tightly near the center, indicating similar subsidy behaviors. In contrast, major subsidizing nations such as Saudi Arabia, Russia, China, India, and Indonesia appear far from the cluster, showing distinct and significantly higher subsidy profiles. The plot highlights how a few countries disproportionately influence global subsidy trends.

7.2 K-Means Clustering



Insight: This plot displays the K-Means clustering of countries based on their subsidy characteristics, using PCA1 and PCA2 as the feature space. Countries in the same cluster share similar subsidy patterns, forming distinct groups with clear separation. One cluster represents low-subsidizing nations, another captures medium-level subsidizers, and the outlier cluster highlights countries with exceptionally high subsidy levels. The visualization shows how K-Means effectively groups countries according to their overall fossil fuel subsidy behavior.

Figure 7.2: Country clusters based on subsidy behavior.

Cluster Meaning:

- **Cluster 0:** Heavy subsidizers with rising patterns.
- **Cluster 1:** Medium-level nations with moderate fluctuations.
- **Cluster 2:** Low-subsidy or declining subsidy countries.

Chapter 8

Conclusion

This exploratory data analysis of the IEA Fossil Fuel Subsidies database provides a clear and insightful understanding of how global subsidy patterns have evolved over the last decade. The findings show that oil remains the most heavily subsidized fuel worldwide, with its support increasing steadily across many regions. Developing nations emerge as the largest contributors to global subsidy levels, largely due to their efforts to stabilize domestic energy prices and support economic growth. Regional analysis indicates that Asia and the Middle East consistently account for the highest subsidy allocations, reflecting their large populations, energy demand, and resource-driven economies. Time-series trends highlight the sensitivity of subsidy policies to global events, as sharp fluctuations correspond to economic crises, fuel price shocks, and major policy changes. The application of PCA and clustering further enhances the understanding of global patterns by revealing distinct groups of countries that exhibit similar subsidy behaviors. These analytical techniques help identify structural similarities among nations, offering valuable insights into policy alignment and economic positioning. Overall, the study demonstrates the effectiveness of data analytics in interpreting complex global datasets and uncovering meaningful trends. The conclusions drawn from this project not only highlight the scale and distribution of fossil fuel subsidies but also emphasize the importance of informed policy decision-making in shaping future energy and economic strategies

This project demonstrates the power of data analytics in understanding large-scale economic and energy trends.

Reference:

- **Reference-1:** https://www.iea.org/data-and-statistics.
- **Reference-2:** https://github.com/onkar90/EDADV-PROJECT-DATAC-LEANING

Contribution:

- **Omkar Tavarkhed:** Coding+Report
- **Shatakshi Vaikunthe:** Coding+Report
- **Pratiksha Naikwadi:** Coding+ppt
- **Vinit Rathod:** Coding+ppt