Analysis of Energy Consumption and Emission Data Using Orange Data Mining

Introduction

This report presents an analysis of energy usage patterns and their relationship with carbon emissions in the years 2019 and 2020. It focuses on the consumption of various fuel types by countries, as detailed in the BP Statistical Review of World Energy dataset. Utilizing Orange Data Mining software, the study seeks to unravel the intricate relationships between different types of energy consumption across nations and over time, offering insights that could potentially guide policy and decision-making in the energy sector. The BP dataset, known for its comprehensive and reliable energy metrics, provides the opportunity to explore patterns in climate change and its environmental implications, yielding valuable insights that can support the global shift towards sustainable energy practices.

Data Selection and Preprocessing

The study draws from BP's comprehensive dataset detailing fuel consumption by country and CO2 emissions, focusing on 2019 and 2020 to track recent energy trends. During preprocessing, data underwent rigorous standardization to rectify inconsistencies in formatting, labels, and numeric values, ensuring data integrity. Merging two key datasets—one for energy consumption by fuel type and another for emissions—allowed a unified analysis. Non-standard country names were harmonized, with certain countries grouped under a generic "other" due to historical or political changes. Features were engineered for clarity, combining various energy sources into consolidated renewable and non-renewable metrics, along with calculating the year-over-year emissions change. This refinement enabled a more straightforward comparative analysis. Stringent normalization and standardization ensured a high-quality dataset ready for detailed exploration.

Structure & Design

The overarching strategy was to uncover latent patterns and trends within the energy consumption and carbon emissions data. The exploratory analysis focus involved examining the dynamics between renewable and non-renewable energy sources, aiming to discern the trends for countries with a higher renewable to non-renewable ratio to stabilize or reduce their carbon emissions between 2019 and 2020.

Using Orange Data Mining software, the data journey began by loading datasets on emissions and fuel consumption through File Widgets. A Merge Data Widget brought these datasets together under the 'Country' category for a cohesive analysis. The Feature Constructor was then used to craft new attributes vital for trend spotting.

The initial exploration included using Distribution and Box Plot Widgets for spotting data anomalies and understanding data characteristics. Select Columns Widget refined the dataset to essential variables, while Select Rows Widget isolated data according to hypothesized conditions like changes in emissions or energy ratios.

The workflow involved identifying data patterns and deepening variable relationship analysis with Correlation and Heat Map Widgets. Scatter Plot Widgets visualized the data for further interpretative analysis and Save Data Widgets ensured all findings were documented for subsequent use.

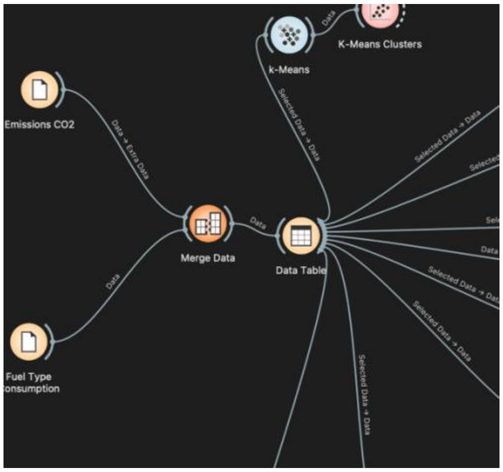


Figure 1 Initial Workflow & Design

In summary, the data was meticulously prepared, analyzed, and visualized to reveal insights on energy consumption's impact on emissions, providing a solid foundation for data-driven decision-making in energy policy.

Analysis

A central hypothesis for this analysis was on the interconnection between renewable and non-renewable energy consumption and their respective CO2 emissions in various global regions throughout the target years. Whilst some trends were simpler to observe such as which regions consume the most and least coal or nuclear energy, other measurements required feature construction.

For example, through feature engineering, while a global trend towards reduced emissions was evident, the analysis identified exceptions. Notably, 'Other CIS', 'Other Africa', 'Iran', and 'China' recorded emission increases, potentially due to their distinct economic, regulatory, and industrial energy profiles. These increases may reflect the complex interplay of economic expansion, dependence on fossil fuels, and varying environmental policies. Contrastingly, the same regions that saw emissions rise also demonstrated a marked increase in renewable energy uptake, particularly solar and wind. This juxtaposition highlights a critical phase in the energy transition where the expansion in renewable use and the persistence of high emissions coexist.

Continuing our analysis, we delve deeper into specific scatter plots which shed light on relationships between energy source and emissions.

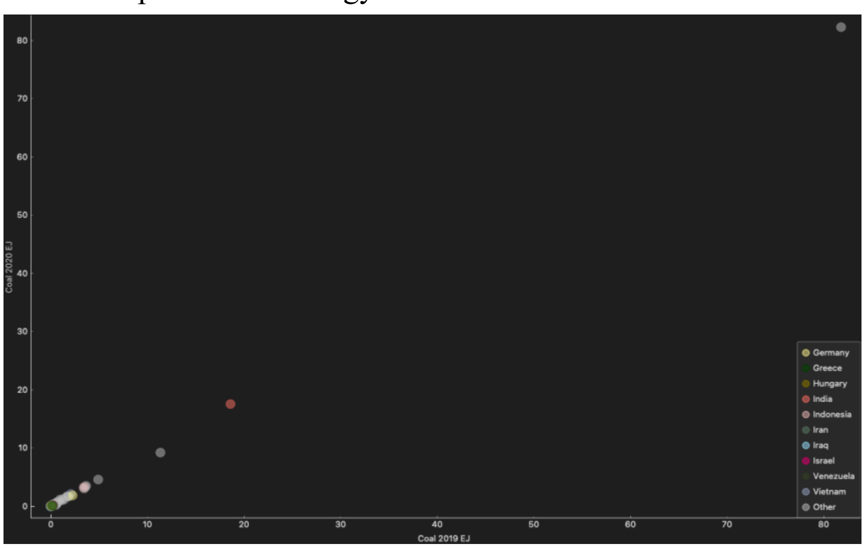


Figure 2 Coal 2019 vs Coal 2020

Figure 2 (Coal 2019 vs Coal 2020) illustrates a positive correlation between coal consumption and emissions. This correlation suggests that countries with high coal consumption tend to exhibit higher CO2 emissions. Such findings underscore the substantial contribution of coal to emission profiles and emphasize the importance of transitioning away from this high-emission energy source.

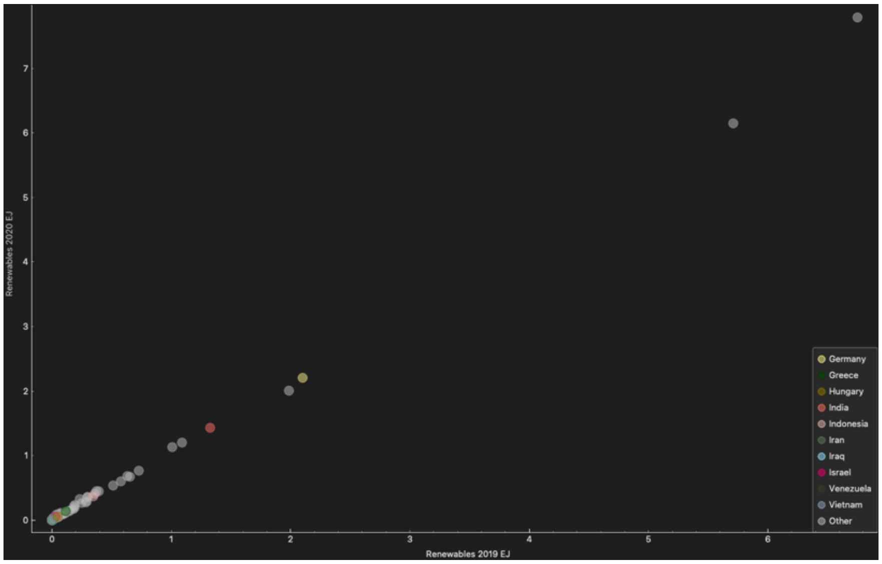


Figure 3 Renewable Energy 2019 vs Renewable Energy 2020

Figure 3 shows the linear relationship between renewable energy consumption in 2019 vs 2020. Even more insightful, comparing emissions between two years reveals trends that reflect the effectiveness of policies or the influence of shifts towards renewable energy. This analysis allows us to gauge the impact of various factors, including energy source transitions, regulatory changes, and economic developments, on CO2 output. Furthermore, insights into the adoption of different energy types, such as Hydroelectric vs. Nuclear Energy provides valuable information on the potential effects of these energy sources on emissions. If countries with higher hydroelectric or nuclear energy usage exhibit lower emissions, it highlights the positive impact of these sources in reducing CO2 output.

Finally, Figure 4 (Renewable Energy 2019 vs Carbon Dioxide Emissions 2019) presents scatter plots contrasting renewable energy consumption against CO2 emissions. These visualizations are critical in reinforcing the role of renewables in emissions reduction strategies. If countries with higher renewable energy usage tend to have lower emissions, it underscores the significance of renewable sources in mitigating environmental impact.

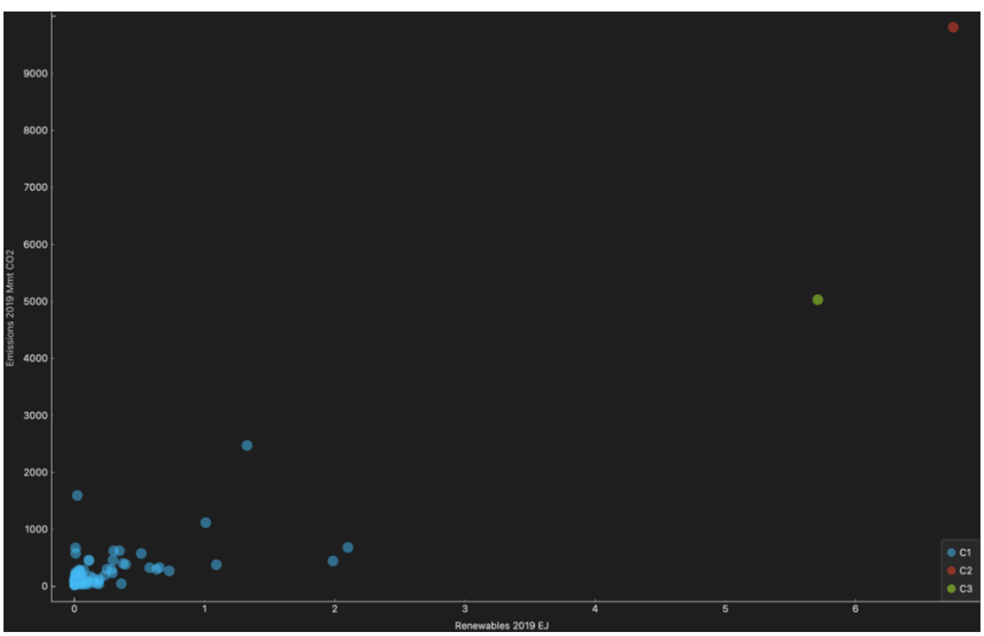
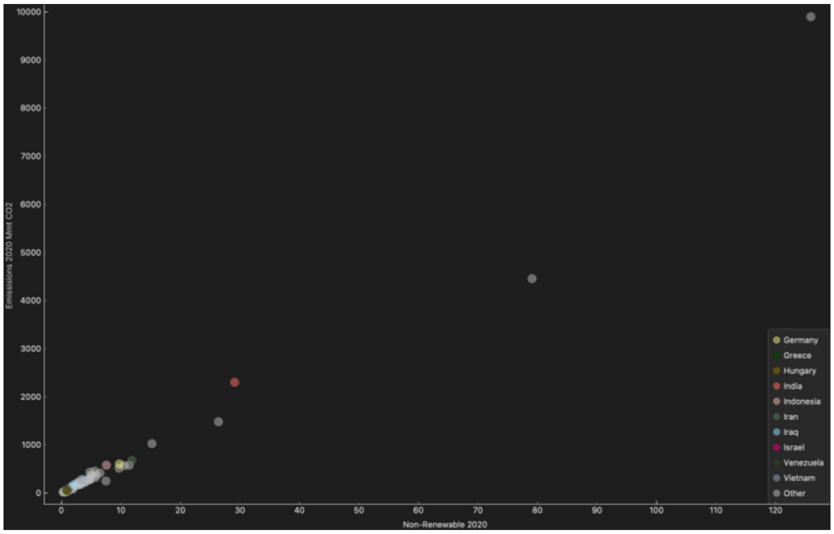


Figure 4 Renewable Energy 2019 vs Carbon Dioxide Emissions 2019

By examining these correlations and trends, we gain a comprehensive understanding of the complex relationship between energy consumption patterns and their consequences on carbon emissions, further supporting our analysis of the global energy landscape.



As illustrated in Figure 5 (Non-Renewables 2020 vs Emissions 2020), countries with higher non-renewable energy consumption tend to exhibit higher emissions. This observation reinforces the notion that non-renewable energy sources are closely associated with elevated CO2 emissions. However, it's worth noting that the relationship appears to be more variable, as evidenced by the spread of data points in the scatter plot.

Figure 5 non-renewables 2020 vs Emissions 2020

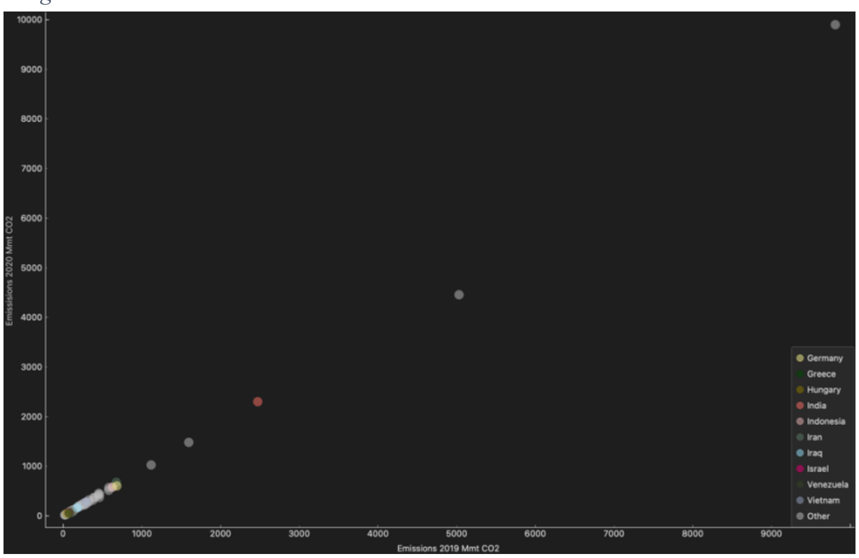


Figure 6 (Emissions 2019 vs Emissions 2020) showcases the changes in emissions from 2019 to 2020. A distinct group of countries with reduced emissions stands out, potentially indicating the effectiveness of policies or shifts towards greener energy sources. This analysis allows us to identify countries that are making significant strides in reducing their carbon footprint within a single year.

Figure 6 Emissions 2019 vs Emissions 2020

The K-Means clustering analysis, as previously discussed, further enriches our understanding. From these clusters and patterns, we can make several observations. For example, Cluster 1 (C1): This cluster likely represents countries heavily reliant on non-renewable energy sources with high emissions. Cluster 2 (C2): Countries in this cluster may have a more balanced energy mix, resulting in moderate emissions. Cluster 3 (C3): This group appears to be at the forefront of renewables adoption, effectively managing lower emissions. This analysis supports the hypothesis that renewable energy consumption is inversely related to carbon emissions, and the clustering helps to identify which countries are transitioning effectively towards renewable energy. The patterns suggest that investing in renewables could be a viable strategy for emissions

reduction. However, the relationship is complex, and other factors not captured in the plots may also influence emissions.
When examining correlation coefficients, they provide additional valuable insights.
A perfect positive correlation (+1.000) between Renewable 2019 and Renewable 2020 indicates consistent renewable energy usage from 2019 to 2020. Additionally, strong positive correlations (above +0.9) among Non-Renewable 2019, Non-Renewable 2020, and Emissions for both years suggest a consistent relationship across these variables from year to year. This implies that countries with high non-renewable energy usage in one year tend to have high usage in the following year, reflecting continuity in energy patterns. A high correlation between Energy Consumption and Emissions reinforces the direct link between higher non-renewable energy consumption and elevated carbon emissions. This correlation holds true for both 2019 and 2020, highlighting the persistent impact of non-renewable sources on emissions.
Lastly, the intriguing dynamic between Renewable vs. Non-Renewable Energy shows high correlation coefficients between these variables for both years. This suggests that countries with high renewable energy consumption often also have high non-renewable energy consumption, possibly due to larger economies requiring a mix of energy sources. These strong positive correlations provide valuable insights for forecasting and policymaking aimed at reducing emissions, highlighting the interconnected nature of energy choices on a global scale.

Conclusion
Looking ahead, future analysis endeavors could dive deeper into the assessment by incorporating more sophisticated predictive modeling techniques. Both supervised and unsupervised learning approaches could be explored to extract more precise insights from the data. The employment of larger datasets, encompassing a broader spectrum of countries and years, could enhance the robustness of the analysis. Time series analyses can be conducted to explore the trends in emissions in countries that have implemented carbon taxes or green energy incentives. Moreover, rigorous testing and validation procedures would be implemented to ensure the reliability of predictive models. In terms of future data analysis projects using Orange, one potential direction is to explore economic development hypotheses, investigating whether developed countries experience lower emission increases relative to energy consumption due to more efficient technologies. We can filter countries by GDP or development status and analyze emissions per unit of energy consumed. These future directions hold the potential to provide a more comprehensive understanding of the intricate relationship between energy consumption, economic development, and environmental impact.

Comparison to Conventional Programming
Orange (anaconda) Vs Python (Jupyter Notebook) - Orange adapts more seamlessly to changes in data, tables, and variables. In Python, adjusting often requires manual code modifications which is more error prone. From initial use Orange's visual interface, including the speed of dragging and dropping widgets compared to writing code in Python. Furthermore, any bugs are more interpretable and faster detected. Therefore, using Orange for data analysis offers a visual and interactive approach, making it quicker for those less familiar with coding. On the other hand, Python provides greater flexibility, especially for complex tasks, and offers transparency and replicability. In conclusion, the ODM workflow provided a robust framework for navigating

the complex data landscape, from data access and cleaning to in-depth analysis and mining. Though its performance on more time/computationally expensive tasks requires more evaluation.

Dataset Reference: BP Statistical Review of World Energy Dataset:
https://data.subak.org/dataset/bp-statistical-review-of-world-energy/resource/3a0753f8-7ebe-450c-95d9-4625a62b7fcd?inner_span=True

Software Used:
Orange Data Mining Software
Microsoft Excel
ChatGPT for Conversational Dialogue around Planning and Approach