Data Visualization - Assignment 3

Harsh Kumar
IMT2021016

Subhajeet Lahiri
IMT2021022

Rohit Shah
IMT2021027

***Abstract*—This report introduces a practical visual analytics based method to examine how a country's energy use, production, and wealth are connected. Using tools like statistics and machine learning, we explore patterns in a diverse energy dataset. Our goals include understanding energy use trends, examining production, looking at wealth connections, finding outliers, and studying renewable energy's impact.**

The workflow, including a loop for feedback, follows the diamond workflow proposed by Kiem et al. We expect to gain a good understanding of how energy and wealth are linked, identify trends, and offer useful insights. The different components in our workflow are all implemented in Python - enabling greater reproducibility.

This report provides a simple foundation for better decision-making, offering important insights for sustainable energy practices and economic growth.

Link to the attached figures and the demos

Link to the GitHub repository

Link to our A1 report

I. Introduction

This report utilizes Kiem et al.'s Diamond Workflow to analyze the World Energy dataset (Fig. 1). We initiated our study by defining clear objectives, establishing a foundation for a systematic exploration of global energy dynamics. Our approach involved some preprocessing to get the raw data in the desired shapes, paving the way for exploratory data analysis (EDA).

During EDA and subsequent phases, we applied both supervised (regression) and unsupervised (k-means clustering) learning techniques to extract insights. Our focus on data augmentation kickstarted our analytical journey, branching into specific data transformation approaches aligned with our observations and objectives.

Our exploration delves into discerning trends within non-renewable and renewable energy sources, presented through inferences and visualizations. This report provides a comprehensive understanding of global energy data, showcasing the diverse pathways we navigated to uncover valuable insights.

II. Problem and Objectives Definition

Overall Objective:

Explore and analyze the correlation between energy production, consumption, and wealth metrics across different countries over time.

Specific Objectives:

1) **Understand Energy Consumption Patterns:** Analyze the trends and patterns in energy consumption (across different types) for various countries.
2) **Examine Energy Production Metrics:**
   - Investigate energy production metrics and their distribution globally.
   - Identify countries leading in renewable energy production and traditional energy sources.
3) **Assess Wealth Metrics:**
   - Explore wealth-related metrics (such as GDP) for countries in the data set.
   - Investigate the correlation between wealth metrics and energy consumption/production.
4) **Identify Outliers and Anomalies:**
   - Detect outliers or anomalies in the dataset that may signify unique energy consumption or production behavior.
5) **Understand the Impact of Renewable Energy:**
   - Assess the contribution of renewable energy sources to the overall energy production.
   - Identify countries with notable achievements in renewable energy utilization.

Expected Outcomes:

- A comprehensive understanding of the relationship between energy metrics and wealth indicators.
- Identification of key trends, patterns, and outliers in energy consumption and production.
- Insights into the impact of renewable energy sources on a country's energy portfolio.

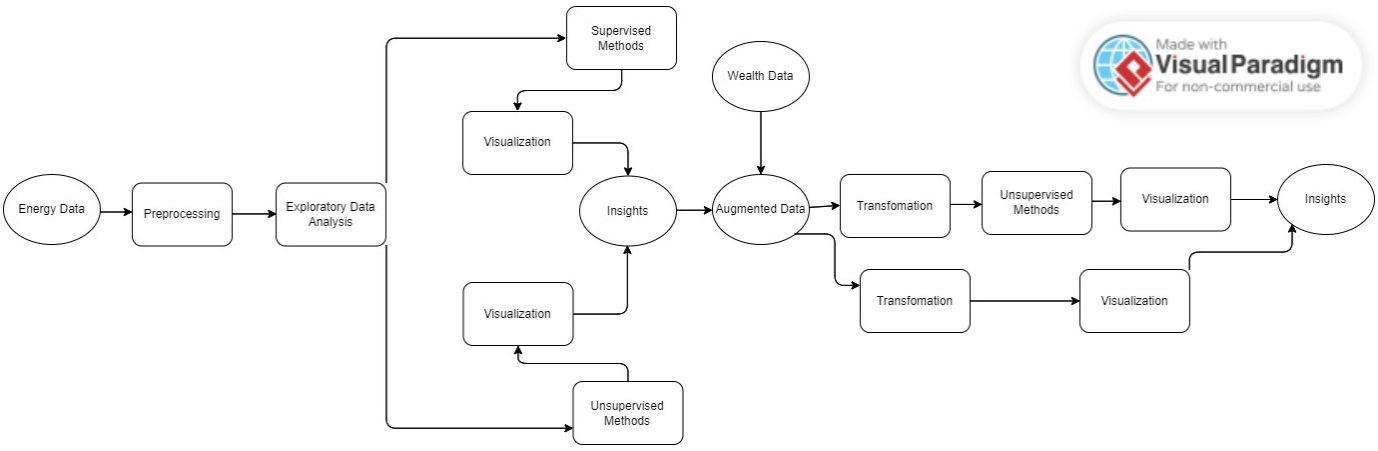

Fig. 1. The workflow that was adhered to

III. Data Collection and Preprocessing

A. *Data Inconsistency:*

In the context of our project, we encountered a disparity in the structure of our World Wealth dataset compared to the World Energy dataset obtained during Assignment 1. The initial World Energy dataset, represented in CSV format, featured a tabular structure with individual columns for various features, such as *Feature1*, *Feature2*, and so forth, organized by *Country* and *Year*.

TABLE I
Original World Energy Dataset

Country	Year	Feature1	Feature2	...
Country1	1980	..	..	...
Country1	1981	..	..	...
...	...	...	...	...
Country1	2022	..	..	...
Country2	1980	..	..	...

In contrast, the new dataset assumed a different format, employing a wide format where features were represented as columns and each observation denoted by the combination of *CountryName*, *FeatureName*, and respective *years*.

TABLE II
New World Wealth Dataset

CountryName	FeatureName	1980	1981	...	2022
Country1	Feature1	..	..	...	..
Country1	Feature2	..	..	...	..
...	...	...	...	...	...
Country2	Feature1	..	..	...	..

Although visualization of the new data was feasible, our objective was to create a consolidated dataset for enhanced visualizations. This necessitated the generation of a new data file by combining selected columns from the Assignment 1 data and relevant columns from the newer dataset. The key criterion for this amalgamation was the ability to perform joins based on [*Country*, *Year*].

B. *Missing Country, Year pairs from A1 dataset:*

It was crucial to make the datasets consistent to facilitate subsequent join operations on the (*Country*, *Year*) pairs. Ensuring uniformity in the (*Country*, *Year*) pairs across both datasets was imperative, requiring an equivalent count for each such pair.

We executed a set difference operation on countries. This operation retained only those countries that were present in both datasets, ensuring a synchronized set of countries for your subsequent analyses.

Concerning the temporal dimension, while the wealth dataset spanned all years from 1980 to 2022, the A1 dataset exhibited missing years for certain countries within this period. To rectify this, countries lacking sufficient year entries in the A1 datasets were systematically excluded from both datasets.

IV. Exploratory Data Analysis - The General Scheme of Things

In data mining, Exploratory Data Analysis (EDA) is a crucial initial step. It involves examining and understanding the dataset you're working with by summarizing its main characteristics and identifying patterns or insights. EDA uses various statistical and visualization techniques such as histograms, box plots, scatter plots, and more to explore the data's structure, relationships between variables, and any potential anomalies or trends.

This section lists the Exploratory Data Analysis done on the data :

A. *Getting the data*

: Our data set consists of time series data, so we developed an API (FIG 2)that returns a view of data for a particular year, where user can study the pattern in various things like energy consumption and production and how it relates to other features.

B. *Removing Duplicate Columns*

Our data set consisted of various columns that had the same data presented in different formats. For example, oil

```
def get_year(self,year):
    grouped_data = self.X.groupby(['Year'])
    return grouped_data.get_group(year)
```

Fig. 2. Users can set the year value to get the data view of different years.

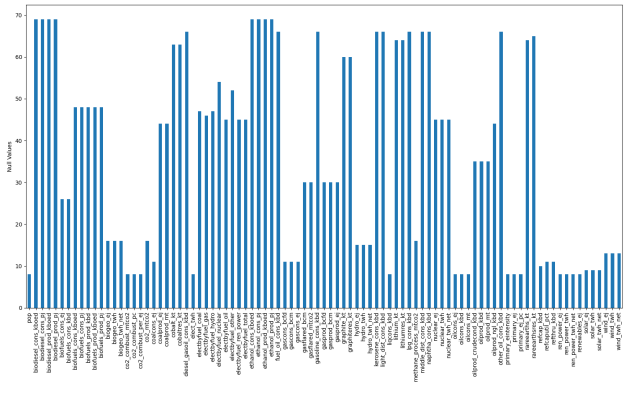

Fig. 3. Null values in each columns (70 is the total rows in this particular year, which is 2022).

production data was given in both barrels and in tonnes, Now for analysis purpose, we are scaling the data and using ML methods like clustering to find the patterns in the data and hence, having multiple columns representing the same thing in different part creates redundancy. All such columns were identified manually and removed from the data by looking at the data description.

C. Handling Missing Data

Our data set is a time series data set, where data is taken from year 1980 to 2022. Now for various reasons, data in different years and in different columns are missing and thus represented by 0. This can create bias and thus needs to be dealt with. Upon further analysis (FIG 3), we found that few columns for the year 2022, had more than 98% missing values and thus these columns for a particular time were removed. Now for other columns with relatively lower null values, these missing data were imputed using the mean strategy.

D. Detecting and Handling Outliers

An outlier in a data set is a data point or observation that significantly differs from other data points. It's a value that appears to deviate or lie far outside the typical range or pattern of the majority of the data. Outliers can occur due to various reasons, such as measurement errors, natural variations, or rare events. Outliers can have a substantial impact on statistical analyses, modeling, and machine learning algorithms. They can distort the results of data analysis, affecting measures like the mean and standard deviation. Thus identifying and handling outliers is very important in data analysis.

We first identify the outliers using box plots. For this again we have built an api (FIG 4) that can easily show the box

```
def show_outliers(self,column):
    try:
        sns.boxplot(x=self.X[column])
        plt.show()
    except:
        print("There was some error while accessing the given column")
```

Fig. 4. Users can set the column name to get the box plot for a particular data view of a particular year.

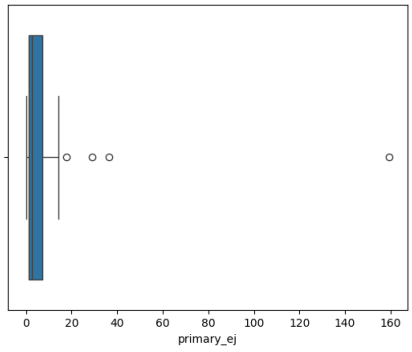

Fig. 5. Box Plot of primary energy consumption in the year 2022. We can clearly see there are 4 outliers which can affect the resulting analysis and thus needs to be dealt with.

plot for identifying outlier, where user just needs to pass the feature/column name that the want to analyse for outliers.

For example, we plot the box plot for primary energy consumption (FIG 5), renewable energy consumption (FIG 6), total electricity generation (FIG 7.

We provide two APIs , one for removing outliers from a specific column (FIG 8) and one from entire data set (FIG 9). The reason for this was suppose we want to find pattern between total electricity generated and the total primary energy consumed. Now, total primary energy consumed and total renewable energy consumed are two separate thing and thus removing outlier from entire data set might cause some data that were outlier in renewable energy consumed to be removed, while it might not have been outlier in primary

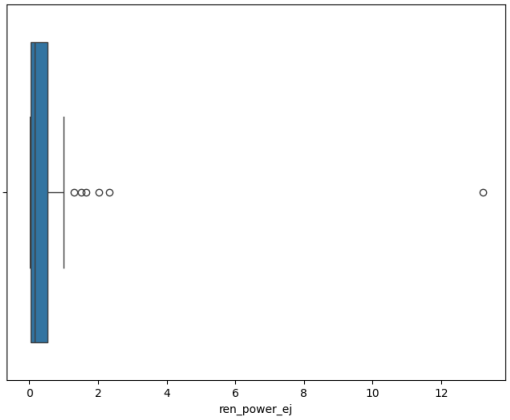

Fig. 6. Box Plot of renewable primary energy consumption in the year 2022.

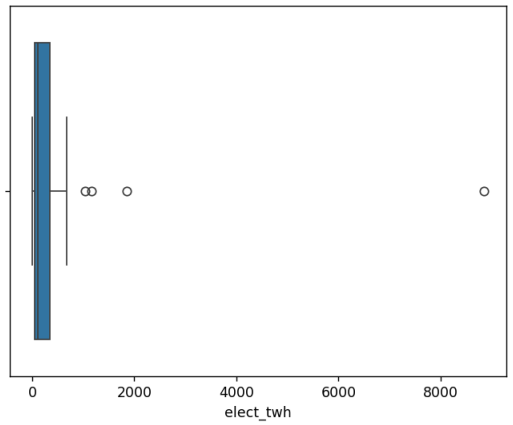

Fig. 7. Box Plot of total electricity generation in the year 2022.

```
def remove_outliers(self,col):

    sorts = self.X[col].sort_values()
    Q1 = sorts.quantile(0.25)
    Q3 = sorts.quantile(0.75)

    IQR = Q3-Q1

    prev = len(self.X.index)

    self.X = sorts[~((sorts < (Q1 - 1.5 * IQR)) |(sorts > (Q3 + 1.5 * IQR)))]

    print(f"{prev-len(self.X.index)} outliers were removed")
```

Fig. 8. Users can select the column to remove outlier from.

energy consumed. So, if a user is analysing the pattern in specific subset of data set, then we can choose to remove outliers from that subset only.

Method of Removal : : We are using IQR method to remove outliers from the dataset. The data is sorted in increasing order of the value and split into 4 equal parts and Q1, Q2 and Q3 are the values that separate the 4 equal parts.

- Q1 : Represents the 25th percentile of the data.
- Q2 : Represents the 50th percentile of the data.
- Q3 : Represents the 75th percentile of the data.

IQR is the range between the first and the third quantiles namely Q1 and Q3: $IQR = Q3 - Q1$. The data points which fall below $Q1 - 1.5IQR$ or above $Q3 + 1.5IQR$ are outliers. These data points are identified and subsequently removed from the data set.

In 2022, there were 29 data points that were considered outlier out of 70 data points when entire data set was used for outlier detection. This is why, we build another API to remove outliers from specific columns that we were trying to analyze.

```
def remove_all_outliers(self):
    Q1,Q3,IQR = self.IQR()
    prev = len(self.X.index)
    self.X = self.X[~((self.X < (Q1 - 1.5 * IQR)) |(self.X > (Q3 + 1.5 * IQR))).any(axis=1)]
    print(f"{prev-len(self.X.index)} outliers were removed")
```

Fig. 9. Users can choose to remove outlier from the entire dataset.

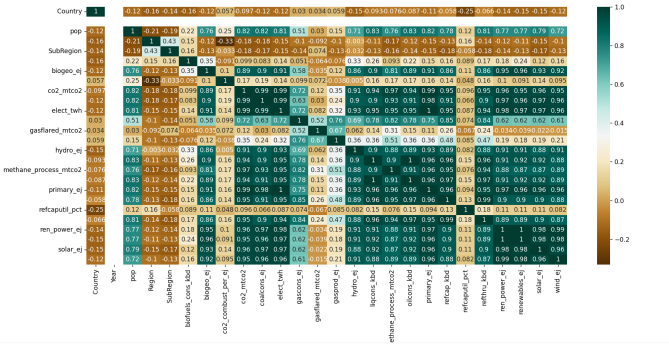

Fig. 10. Correlation Heatmap.

Correlation Heatmap

Correlation heatmaps are a visual representation used in data analysis to display the correlation coefficients between variables in a dataset. These heatmaps provide a quick and intuitive way to identify relationships, patterns, or dependencies between different variables.

We tried to understand the dependence of columns with one another using the correlation heatmap (FIG 10).

We can see that primary energy consumption is highly correlated (positively) to the production of oil, electricity, carbon dioxide emission. This was expected as with increasing consumption of primary energy, we would expect the electricity generation, carbon dioxide emission would also increase. This correlation heatmap is very useful for selecting features to study for pattern. For example, if we want to study the pattern of primary energy consumption then we would want to select the features that highly correlates with primary energy consumption.

V. EDA - Non-renewable energy sources

This section - and the subsequent sections dealing with non-renewable energy sources - will be concerned with the following columns in the World Energy data set :

1) **'coalprod_ej'**: Total coal production in exajoules for a specific country in a specific year.
2) **'coalcons_ej'**: Total coal consumption in exajoules for a specific country in a specific year.
3) **'oilprod_kbd'**: Total oil production in kilobarrels per day for a specific country in a specific year.
4) **'oilcons_kbd'**: Total oil consumption in kilobarrels per day for a specific country in a specific year.

A. Outlier detection and removal

Removing outliers in the given energy data comes with its own pitfalls. If we remove all outliers using the IQR based scheme detailed in the previous section, we end up removing a lot of data (Fig. 11).

If we generate attribute-wise box plots, we get Fig. 12. We can thus make a case for deferring outlier removal to

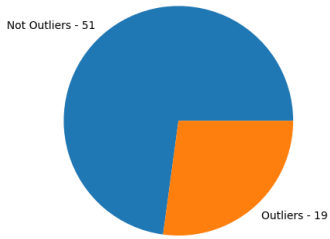

Fig. 11. Outlier counts with respect to the columns named above (2021)

be conducted on a more **case-by-case basis** by giving the following reasons :

1) **Dominance of Some Countries:**
   In energy data, a small number of countries often dominate production and consumption. Removing outliers could disproportionately affect these dominant countries, distorting the representation of global energy patterns.
2) **Inherent Variability in Energy Data:**
   Energy data exhibits inherent variability due to geopolitical, economic, and environmental factors. Extreme values in energy production and consumption may reflect real-world events, and removing them risks oversimplifying and distorting the representation of the energy landscape.

B. Pairwise Correlation Analysis

Refer to Fig. 13 and Fig. 14

1) Methodology:

- Scatter plots were created to visually represent the correlation between coal and oil consumption and production. Coal consumption vs. production was depicted in blue, while oil consumption vs. production was depicted in red.
- Linear regression lines were fitted to the data to capture trends and quantify the relationships.
- A second plot was created using similar methods but with outliers removed.

2) Inferences:

- **Positive Correlation:**
  - Both coal and oil consumption exhibited a positive correlation with production, indicating a general trend of increasing consumption with increasing production.
- **Differential Consumption Patterns:**
  - **Coal:** The regression line's steeper slope for coal production vs. consumption suggests a more pronounced increase in coal consumption compared to oil. Coal production appears to have a stronger influence on consumption.
  - **Oil:** The shallower slope for oil production vs. consumption implies that oil consumption does not rise as rapidly as oil production. The relationship between oil production and consumption is less direct.
- **Effect of Outlier Removal**
  - Outlier removal accentuated the disparities in the slopes between production and consumption, emphasizing the impact of extreme values on the observed correlations.
  - The second plot, with outlier removal, provided a clearer representation of the relationships between production and consumption for both coal and oil.

VI. EDA - Renewable Energy Sources

We will focus on the following important columns in the World Energy data for this section and all the subsequent sections that deal with renewable sources of energy:

1) **electbyfuel_hydro**: Total electricity generated by hydro power for a given country in a given year (in exajoules).
2) **hydro_ej**: Total electricity consumption from hydro power for a given country in a given year (in exajoules).
3) **biofuels_prod_pj**: Total electricity generated by biofuels for a given country in a given year (in petajoules).
4) **biofuels_cons_pj**: Total electricity consumption from biofuels for a given country in a given year (in petajoules).

A. Outlier detection and removal

Refer to (Fig. 15 & Fig. 16), which clearly shows that removing these outliers using IQR strategy (detailed in previous section), will result in losing important data. Following are some reasons, why we defer from removing outliers in addition to reasons mentioned in Subsection VI-A:

1) **Developing Technologies**: Rapid advancements in technology can lead to significant shifts in energy practices. These changes may initially appear as outliers but could represent the emergence of new trends that need to be studied rather than excluded.
2) **Interconnected Energy Systems**: The world energy ecosystem is interconnected, with changes in one region affecting others. Outliers in one country's data may be linked to broader global trends or interconnected energy systems, making isolated removal of outliers less meaningful.
3) **Long-Term Trends due to temporal data**: Our World Energy dataset spans multiple years (1960 to 2022), and long-term trends in energy practices may not be adequately captured by a simplistic outlier detection approach. Ignoring these shifts could lead to misinterpretation of the data.

B. Pairwise Correlation Analysis

Refer to Fig. 17 & Fig. 18

1) Methodology:

- **Scatter Plot Creation for Correlation Analysis**: To understand the relationships between electricity consumption and productions from hydro and biofuels sources.
- **Distinguishing Hydro & Biofuels plots**: Achieved by portraying hydro consumption versus production in a

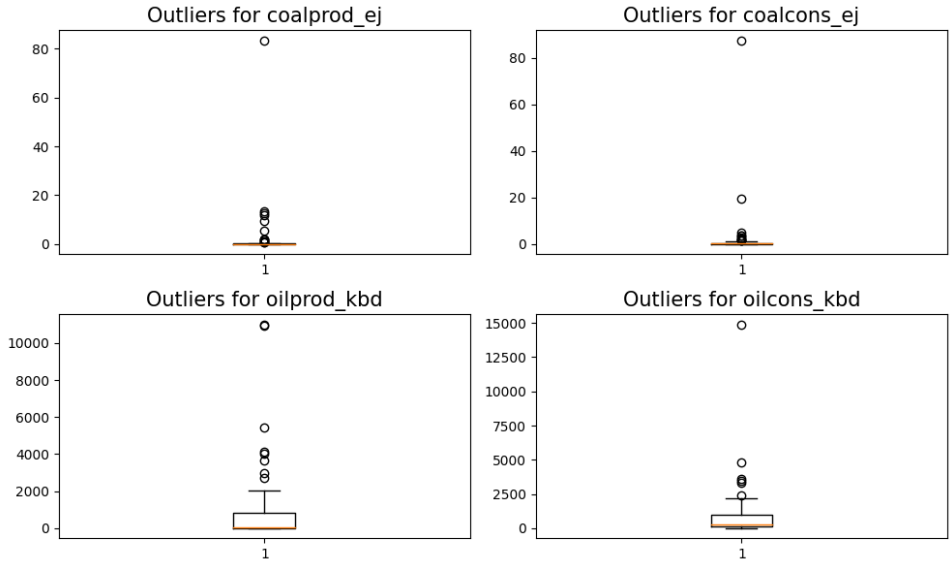

Fig. 12. Attribute-wise box plots related to non-renewable energy

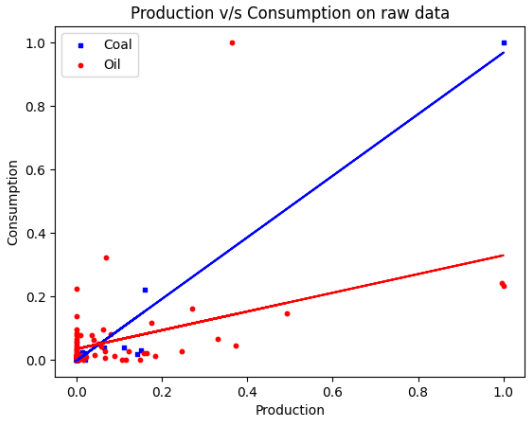

Fig. 13. With outliers

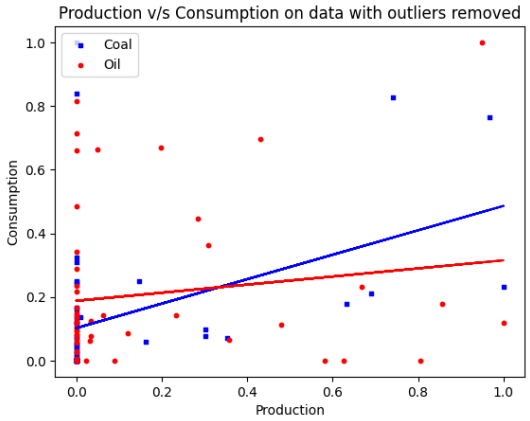

Fig. 14. Without outliers

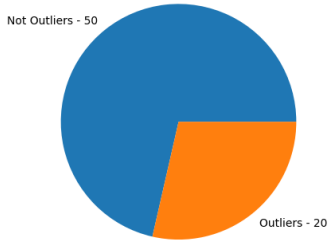

Fig. 15. Outlier counts with respect to the columns named in EDA of renewable energy sources (2021)

distinct blue hue, while biofuels consumption versus production was depicted in a contrasting red shade.

- A subsequent set of visualizations was generated using similar techniques; however, this time, outliers were intentionally excluded to offer an alternative perspective on the underlying data trends.

2) Inferences:

- **Overall trend**:
  - Both hydro and biofuels show a positive correlation between production and consumption, with hydro having a stronger correlation.
  - The linear regression lines for both energy sources show an upward trend, indicating that production and consumption tend to increase together.
- **Further Scrutiny**:
  - Biofuels's steeper slope suggests a more significant increase in production compared to consumption.
  - Hydro's shallower slope indicates a slower rate of production growth relative to consumption.

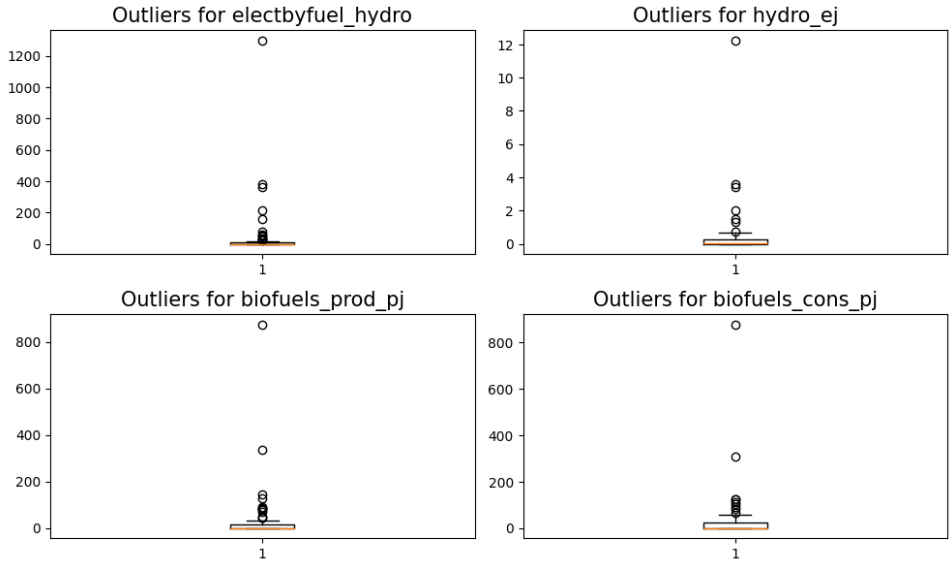

Fig. 16. Attribute-wise box plots related to renewable energy

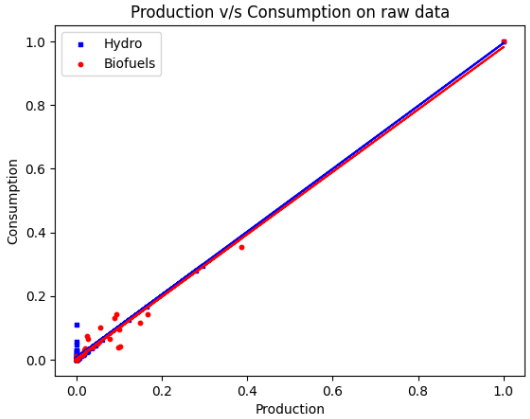

Fig. 17. Scatterplot for renewable sources With outliers

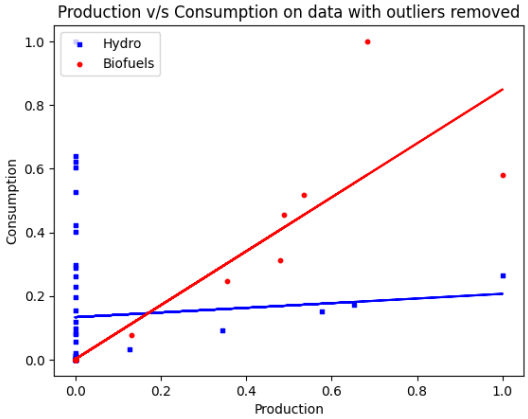

Fig. 18. Scatterplot for renewable sources without outliers

- **Effect of outlier removal**:
  - The impact of excluding outliers became evident in the examination of production and consumption slopes, highlighting the influence of extreme values on the observed correlations.
  - The subsequent plot, created after the deliberate removal of outliers, presented a more lucid depiction of the relationships between production and consumption for both biofuel and hydro sources.

VII. Unsupervised Learning

A. *Primary Energy Consumption*

1) Methods:

- **Attribute Selection:** Using the correlation heatmap built in Section IV, we identified the attributes that were highly correlated with primary energy consumption and selected a subset of those features. More specifically, we selected following features for clustering :
  - Country
  - CO_2 Emission
  - Electricity Generation
  - Oil Consumption
  - Coal Consumption
  - Primary Energy Consumption

  The selection has also been done based on domain knowledge, as we would expect electricity generation to be highly related to primary energy consumption, and oil and coal consumption to be highly related with primary energy consumption and more energy consumption would lead to more and more CO_2 Emission.
- **Outlier Handling:** Using the API built in Section IV, we removed outliers from the attributes selected and

found that 8 out of 70 data point were outliers. The remaining data except country were normalized to be used for clustering algorithm.

- **K-Means Clustering:**
  - The unsupervised learning algorithm, KMeans clustering, was applied to identify inherent patterns and groupings within the data set.
  - To choose the number of clusters, we first plotted the cost vs clusters plot (FIG 19) and using the elbow method, we selected the number of clusters to be 6.
- **Visualization:**
  - We then plot the distribution of each attribute in the cluster in a juxtaposed visualization (FIG 20 - 24)
  - We then convert the multi-dimensional data to 2 dimension using Principal Component Analysis so that we can visualize the country segmentation into clusters. (FIG 25)

2) Inference: With the help of juxtaposed visualizations (FIG 20 - 24) and the cluster feature importance winscss method ([1]), we found that CO_2 emission, oil consumption and electricity generation were the main features used in formation of clusters.

Cluster Interpretations :

- Cluster 0 contains the countries that produces CO_2 in high amounts while consuming other resources like oil, primary energy in small to moderate amounts.
- Cluster 1 contains the countries that consume coal in moderate to high amounts while consuming all the other resources moderately.
- Cluster 2 contains the outlier that could not be removed during the data handling.
- Cluster 3 contains the countries that consumes primary energy in high amounts.
- Cluster 4 contains the countries that consumes resources in less amount and produces carbon dioxide in less amount.
- Cluster 5 contains the countries that consumes resources in less amount and produces carbon dioxide in moderate amount.

We then plotted the cluster visualization without removing the outliers (FIG 26) and found that China and India, which were outliers before, formed their own cluster having only single country (themselves). This is consistent with our expectation since these two are the countries with one of the highest population in the world and naturally these two countries would consume primary energy and emit carbon dioxide at a scale which is much different from rest of the world.

B. Non-renewable Energy

1) Methods:

- **Attribute Selection and Plot Creation:**
  - The two sets of attributes led to the creation of two separate plots:
    1) Points represented by (coal production, coal consumption) (Fig. 27).

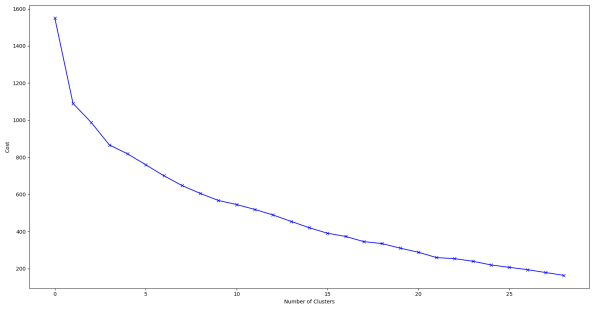

Fig. 19. Cost vs Clusters to apply elbow method for number of cluster selection

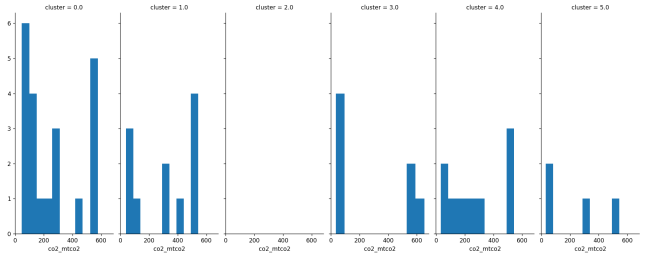

Fig. 20. Distribution of CO_2 in each cluster

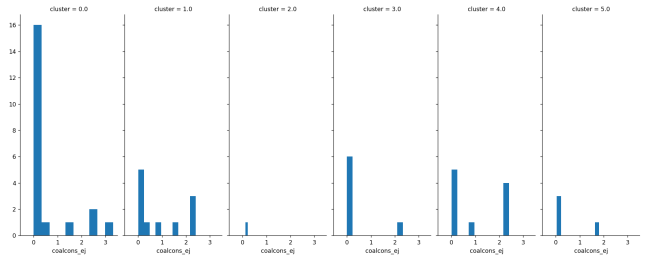

Fig. 21. Distribution of coal in each cluster

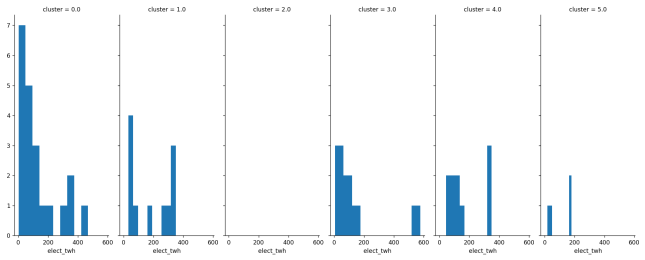

Fig. 22. Distribution of electricity generated in each cluster

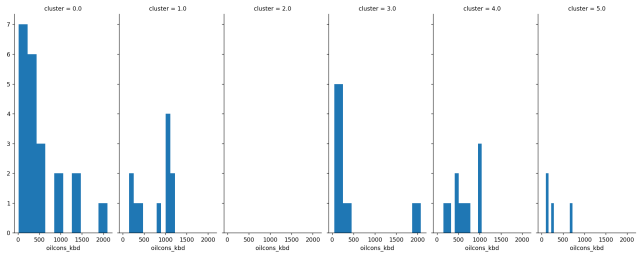

Fig. 23. Distribution of oil in each cluster

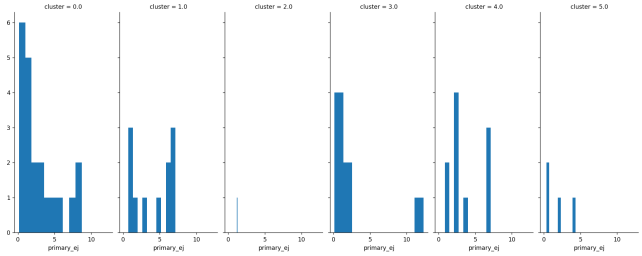

Fig. 24. Distribution of energy consumption in each cluster

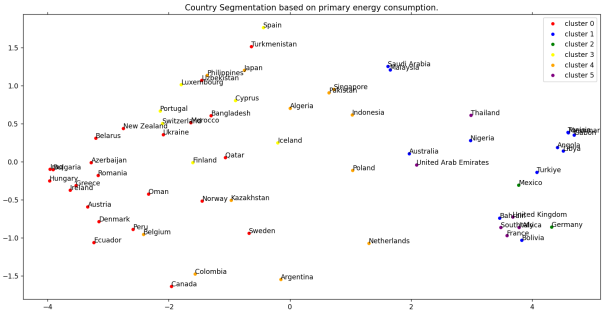

Fig. 25. Cluster Visualization (Data converted to 2-dimensional for visualization using principal component analysis.)

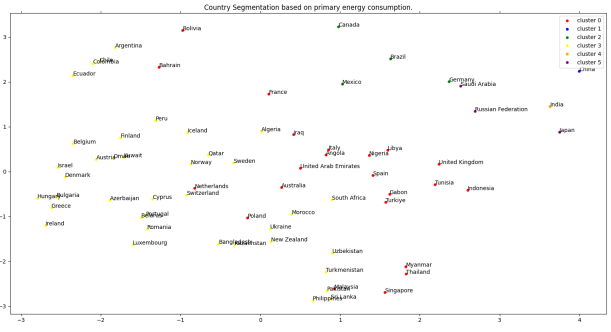

Fig. 26. Cluster Visualization without removing outliers (We can see that China and India, being outliers, form their own clusters.)

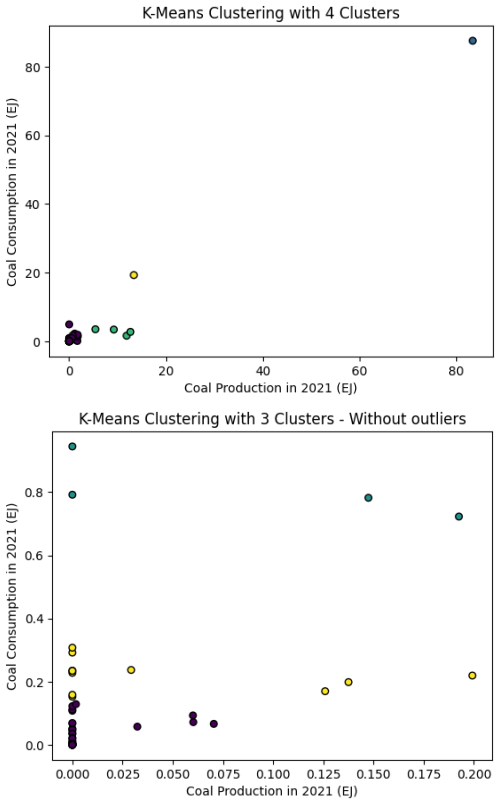

Fig. 27. Clusters in coal production and consumption

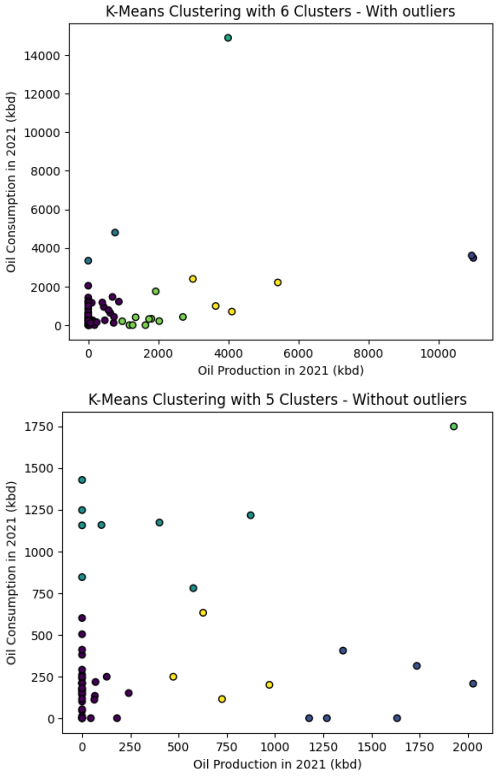

Fig. 28. Clusters in oil production and consumption

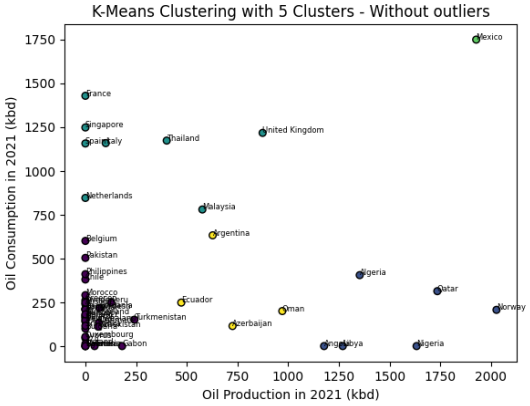

Fig. 29. Scatter plot with labels

2) Points represented by (oil production, oil consumption) (Fig. 28).

– This approach allowed for distinct analyses of the relationships between production and consumption for both coal and oil.

- **Outlier Handling:**
  – Two separate analyses were conducted, one with outlier-free data and another with data containing outliers.
  – Outliers were identified and removed using the criterion defined earlier.
- **K-Means Clustering:**
  – The unsupervised learning algorithm, KMeans clustering, was applied to identify inherent patterns and groupings within the data set.
  – The number of clusters was varied from 2 to 5 to explore different potential groupings.
- **Visualization:**
  – Scatter plots were generated to visually represent the clusters in the 2D space, with production on the x-axis and consumption on the y-axis.

2) Inferences: The scatter plots were generated without labels to avoid clutter. The inferences are drawn using labelled plots. One such plot is attached for reference (Fig. 29).

1. Coal Production vs. Consumption Plot with Outliers: - The plot reveals significant outliers, prominently represented by China (blue) and India (yellow), indicating that these countries have exceptional coal production and consumption levels. - To discern patterns in the remaining data, we focus on the other two clusters in the same plot, specifically after removing outliers.

2. Coal Production vs. Consumption Plot without Outliers: - Three distinct strata become evident in the absence of outliers. These strata are aligned along the family of lines $x + ay = b$, where both a and b are positive. - The orientation of these strata implies that as countries scale coal production, they also scale consumption. Notably, both India and China align on the $x = y$ line. - Clusters tend to gravitate toward

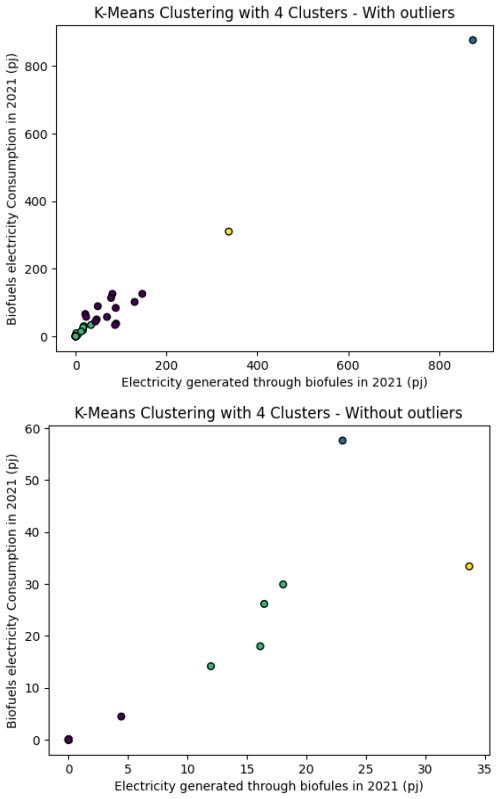

Fig. 30. Clusters in electricity production and consumption through biofuels power

the half below the $x = y$ line, suggesting that countries with substantial coal production also engage in exports.

3. Oil Production vs. Consumption Plot: - The distribution of clusters differs from the coal plots. Each $x + ay = b$ stratum observed in coal is now divided into two groups: one in the top half (consume more than they produce) and the other in the bottom half (produce more than they consume). - The variation within a stratum is attributed to sheer production + consumption capacity, influenced by some unknown set of parameters. This division is exemplified by China (cyan) in the top half and Saudi Arabia, Russia (blue) in the bottom half.

4. Oil Production vs. Consumption Plot without Outliers: - Similar patterns persist even after removing outliers. Notably, the blue cluster is dominated by Gulf nations, while the cyan cluster consists of West European nations. - The purple cluster suggests a hypothesis that it may include poorer/smaller countries, though further verification is required to substantiate this assumption. - For oil, as opposed to coal, countries gravitate towards the upper half of the identity line.

These observations provide insights into the relationships between production and consumption for both coal and oil, shedding light on outliers, strata orientations, and variations within clusters. Further analysis and verification can help refine these initial hypotheses.

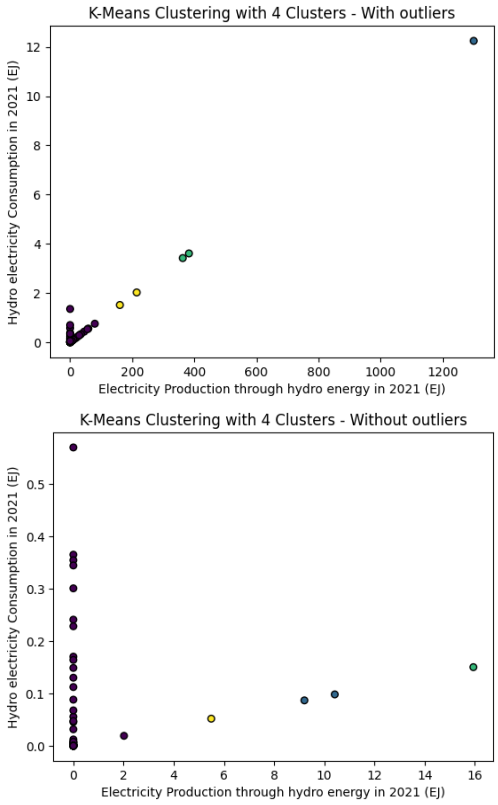

Fig. 31. Clusters in electricity production and consumption through hydro power

C. *Renewable Energy*

1) Methods:

- **Attribute Selection and Plot Creation:**
  - The two sets of attributes led to the creation of two separate plots:
    - Points represented by (biofuel production, biofuel consumption) (Fig. 30).
    - Points represented by (hydro production, hydro consumption) (Fig. 31).
  - This approach allowed for distinct analyses of the relationships between production and consumption for both hydro and oil.
- Refer to Subsection VII-B1 for details on Outlier handling, K-Means Clustering, Visualization. These methods are the same for renewable energy sources.

2) Inferences: The creation of scatter plots intentionally omitted labels to maintain clarity and avoid visual clutter. Conclusions and interpretations are derived from labeled plots, and a specific example is provided for reference. (Fig. 29)

- **Biofuels**:
  Cluster Analysis: The scatterplot diagram provide insights into distinct clusters of countries based on biofuels consumption and production. The following clusters were identified:
  - **Cluster 1 (Yellow, Blue)**: This cluster comprises countries with high biofuels consumption and production of electricity through biofuels. It includes countries like the United States, Brazil, and China. The prominence of these countries suggests a leading role in both biofuels production and consumption on the global stage.
  - **Cluster 2 (Green)**: Countries within this cluster exhibit intermediate levels of consumption and production of electricity through biofuels. This may indicate a phase of development or a transition from conventional fossil fuels to renewable energy sources.
  - **Cluster 3 (Violet)**: Countries in this cluster demonstrate low consumption and production of electricity through biofuels, potenti ally due to limited access to biofuels resources or a lack of robust policy frameworks supporting biofuels development.

  Possible reasons & explanations:
  - **Resource and Policy Dynamics**: Clustering can be attributed to a synergy between resource availability and supportive government policies. Countries in Cluster 1, exhibiting high consumption and production, likely benefit from abundant agricultural resources and favorable policy frameworks.
  - **Economic, Import, and Other Influences**: While economic factors, as indicated by GDP per capita, play a role in consumption, the mix in Cluster 1 suggests a multi-faceted influence beyond economics alone. Import reliance (Cluster 4) and other factors, such as technological advancements & consumer preferences, further contribute to the intricate clustering patterns observed in the scatterplot.

- **Hydro**:
  Cluster Analysis:
  - **Cluster 2 (blue)**: This cluster contains countries with high hydro electricity consumption and production. These countries are likely to have abundant hydro resources and well-developed hydro power infrastructure.
  - **Cluster 3 (yellow)**: This cluster contains countries with medium hydro electricity consumption and production. These countries may be developing their hydro power sectors or transitioning from fossil fuels to renewable energy.
  - **Cluster 4 (purple)**: This cluster contains countries with high hydro electricity consumption but low production. These countries are likely to be net importers of hydro electricity. Examples of countries in this cluster include Japan, South Korea, and Italy.

  Possible reasons & explanations:

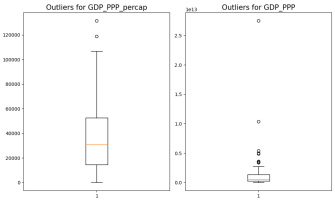

Fig. 32. Boxplots for the GDP PPP and the GDP PPP per capital metrics for 2021

– **Regional Dynamics**: The geographical distribution of countries may play a role in hydroelectricity consumption and production clustering. Regional variations in hydro resources and energy demand could contribute to distinct patterns observed in the scatterplot.

– **Infrastructure Investment**: The level of investment in hydroelectric infrastructure can influence a country's capacity for production. Countries with significant investment in modern hydroelectric technologies may exhibit higher efficiency and greater production capabilities.

VIII. Data Transformation and Further Visualization

A. *Non renewable energy sources*

*1) **Data augmentation with wealth data**:* In the initial phase of data transformation, two significant variables, Gross Domestic Product (GDP) Purchasing Power Parity (PPP) per capita, and overall GDP PPP, were incorporated into the dataset. These augmentations provide a broader context for understanding the economic dimensions associated with energy production and consumption.

Visualization Method

- **Boxplots Analysis** To gain insights into the distribution and variability of the augmented data, boxplots were constructed for the newly added columns - GDP PPP per capita and GDP PPP (Fig. 32). These boxplots offer a visual representation of the central tendency, spread, and potential outliers within the economic metrics.
- **Consumption and Production Scatterplots:** To further explore the interplay between economic metrics and energy dynamics, scatter plots were generated. In these scatter plots:
  – **X-axis:** Represents energy consumption.
  – **Y-axis:** Represents energy production.
  – **Color Channel:** Reflects the GDP PPP metric.

  Four distinct scatter plots were created, encompassing both coal and oil data, each with two variations: one including outliers and another with outliers removed. These visualizations aim to unveil potential patterns and correlations between energy consumption, production, and the economic landscape.

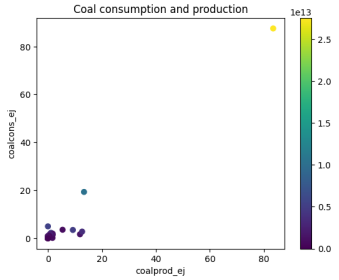

Fig. 33. Coal consumption and production (2021)

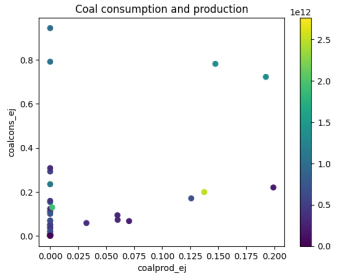

Fig. 34. Coal consumption and production without outliers(2021)

Inferences

- **Boxplots for GDP PPP Metrics:**
  – The box plot for GDP PPP per capita exhibits a wide box and whiskers with very few outliers, suggesting a relatively stable and uniform distribution of wealth across countries.
  – In contrast, the box plot for overall GDP PPP reveals numerous outliers, and the whiskers are not placed far apart. This observation implies significant variability in the overall economic landscape, potentially contributing to the prevalence of outliers in energy data.
  – The presence of outliers in GDP PPP aligns with the variability seen in energy data - which makes sense because both are aggregates. This emphasizes the importance of exploring normalized data to gain a more nuanced understanding of the relationships between economic metrics and energy variables.
- **Color-mapped scatter plots:**
  – In the **Coal Consumption and Production Plot:**

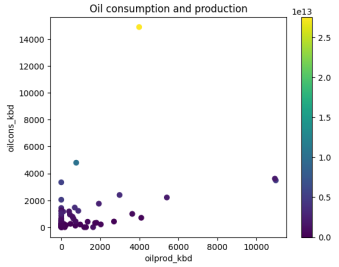

Fig. 35. Oil consumption and production (2021)

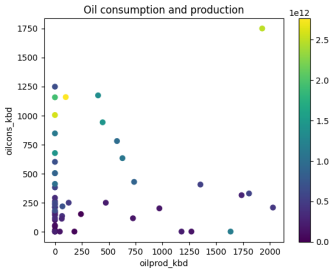

Fig. 36. Oil consumption and production without outliers(2021)

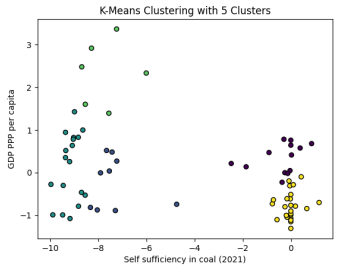

Fig. 37. Clusters in (self sufficiency in coal, GDP PPP per capita) - 2021

  - ∗ China (yellow) and India (blue) prominently stand out, indicating their significant roles in coal consumption and production.
  - ∗ After removing outliers, the color map patterns become more diffuse, but the trend persists. Countries with higher GDP PPP tend to both produce and consume more coal.
- In the **Oil Consumption and Production Plot:**
  - ∗ The color map in the oil plot exhibits a more apparent and uniform gradation compared to coal.
  - ∗ Even after removing outliers, the correlation between higher GDP PPP and increased oil production and consumption remains evident.
  - ∗ The patterns observed in the color map emphasize the strong association between economic wealth and oil-related activities across various countries.

*2) **Log Transforms**:*

- Applied log transforms to the data to address skewed distributions and enhance interpretability.
- Log transforms were particularly useful when dealing with the ratio of two variables.

Refer to Fig. 37 and Fig. 38.

Visualization Method

- **Introduction of Self-Sufficiency Metrics:**
  - Introduced two new metrics to incorporate wealth data into the clusters.
  - **Self Sufficiency in Coal:**

$$\log\left(\frac{\text{coal consumption} + \epsilon}{\text{coal production} + \epsilon}\right)$$

where ϵ is a very small number.

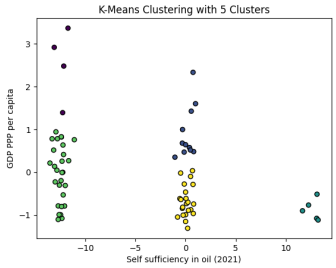

Fig. 38. Clusters in (self sufficiency in oil, GDP PPP per capita) - 2021

  - **Self Sufficiency in Oil:**

$$\log\left(\frac{\text{oil consumption} + \epsilon}{\text{oil production} + \epsilon}\right)$$

similarly defined.

- **Correlation Study with Wealth:**
  - Created two scatter plots to study the correlation between self-sufficiency metrics and wealth (GDP PPP per capita).
  - **Data Normalization:** The GDP values were scaled using a Standard Scaler.
  - **Visualization:** Plotted self-sufficiency on the x-axis and GDP PPP per capita on the y-axis. The points were color-coded according to the assigned clusters.
- **KMeans Clustering:**
  - Employed KMeans clustering on the normalized data points to identify patterns and groupings.
  - Cluster assignments were visualized using color-coding on the scatterplots.
  - This approach allowed for the exploration of how self-sufficiency metrics relate to wealth and the emergence of distinct clusters in the data.

Inferences

- **Coal**
  - The **Purple Cluster:** Consists of developed countries such as the UK, Spain, Japan, and Oceania nations like Australia and New Zealand. These countries, despite being developed, have not phased out coal yet, possibly due to existing infrastructure or economic considerations.
  - The **Yellow Cluster:** Encompasses poorer countries with substantial coal reserves, including India, China, Russia, Thailand, Brazil, Myanmar, Indonesia, and Greece. These nations, with significant coal resources, may rely on coal for economic development.
  - The **Green Cluster:** Dominated by affluent European nations such as Switzerland, Ireland, Luxembourg, and Norway, who haven't phased out coal and depend on imports to meet their coal needs. This reliance on imports might be driven by economic factors or the availability of alternative energy sources.
  - The upper half of the **Blue and Cyan Clusters:** Resembles the constitution of the Green Cluster. The

Cyan Cluster includes more European nations like France, Italy, and the Netherlands, while the Blue Cluster consists of Gulf nations like Saudi Arabia and Oman. These countries, despite their economic wealth, have not phased out coal and rely on similar import strategies.
  - The lower half of the **Blue and Cyan Clusters:** Comprises poorer countries that are still dependent on coal imports, such as Bangladesh, Sri Lanka, Nigeria, and others. These nations might face economic challenges in transitioning away from coal or have limitations in domestic coal production.
- **Oil**
  - The **Cyan Cluster:** Comprises African nations like Libya, Tunisia, and Nigeria, where oil production capabilities significantly outpace oil consumption. This surplus in oil production may indicate an emphasis on oil exports, contributing to economic dynamics.
  - The **Yellow Cluster:** Consists of nations blessed with oil wealth, including Russia and Gulf nations. Despite their oil wealth, these countries are not as economically rich as their counterparts in the Blue Cluster. Economic factors or resource management strategies may contribute to this distinction.
  - The **Blue Cluster:** Encompasses nations blessed with oil wealth that have become economically prosperous. Includes Norway, Qatar, UAE, Denmark, Australia, and Canada. These countries have effectively leveraged their resource wealth for economic development.
  - The **Purple Cluster:** Comprises rich European nations such as Switzerland, Ireland, Luxembourg, the Netherlands, and Singapore. These countries are not well-endowed with oil wealth but rely on oil imports. Economic factors and strategic considerations may drive their dependence on oil imports.
  - The **Green Cluster:** Represents poorer nations not possessing significant oil wealth. Spans a spectrum in GDP PPP per capita, including European countries like the Netherlands and France, as well as Asian countries like Indonesia and Bangladesh. These nations exhibit diverse economic profiles despite their shared lack of significant oil resources.

*3) **Aggregates over the years**:* **Visualization Method**

- Grouped the coal production, coal consumption, oil production, oil consumption, and GDP PPP metrics by year to obtain time series data.
- Created a line chart to visualize the temporal trends, with the year on the x-axis and normalized values (using Min-Max scaling) on the y-axis.
- Each metric was represented by a colored line, allowing for a comprehensive analysis of the temporal evolution of energy-related variables and economic indicators (Fig. 39).

Inferences

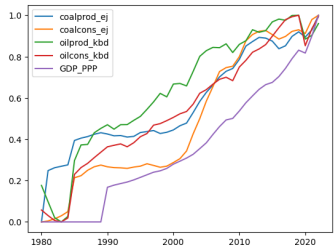

Fig. 39. Total values summed across all countries over the years

- The time series data reveals a consistent increasing trend in coal production, consumption, oil production, oil consumption, and GDP PPP from 1980 to 2021. Despite environmental efforts, this upward trajectory suggests that the consumption of fossil fuels, including coal and oil, is not slowing down. Economic and geopolitical factors may contribute to the persistence of fossil fuel consumption, posing challenges for sustainable energy transition.
- There is a noticeable dip in oil production and consumption in the early 1980s. This dip can be attributed to geopolitical factors, particularly the events surrounding the Iran-Iraq War (1980-1988) and the subsequent oil market disruptions. The conflict disrupted oil production and export capabilities in the region, leading to a temporary decline in global oil production and consumption.
- A slight dip is observed in 2020, primarily attributed to the global COVID-19 pandemic. The pandemic led to widespread lockdowns, travel restrictions, and economic slowdowns, resulting in decreased demand for oil. The energy sector, closely tied to economic activities, experienced a temporary contraction during this period, reflected in the dip in oil production and consumption.
- The slope of the increase in consumption and production of fossil fuels (coal and oil) is not as steep since the 1990s compared to the growth in GDP PPP. This decoupling suggests a relative improvement in energy efficiency. While economic growth continues, the slowdown in the rate of fossil fuel consumption growth indicates a potential silver lining, showcasing advancements in energy efficiency and diversification of energy sources.

B. *Renewable energy sources*

*1) **Data augmentation with wealth data**:* In a parallel approach to the non-renewable energy sources analysis, the augmentation of data for hydro and biofuels incorporated two vital variables: Gross Domestic Product (GDP) Purchasing Power Parity (PPP) per capita and overall GDP PPP.

Visualization Method

Refer to subsection VIII-A1 for more details on Boxplots analysis, Consumption and Production Scatterplots. Four scatter plots were crafted for biofuels and hydro data, each

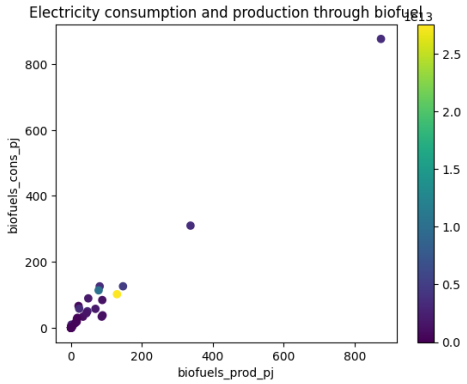

Fig. 40. Electricity consumption and production through biofuels (with outliers) (2021)

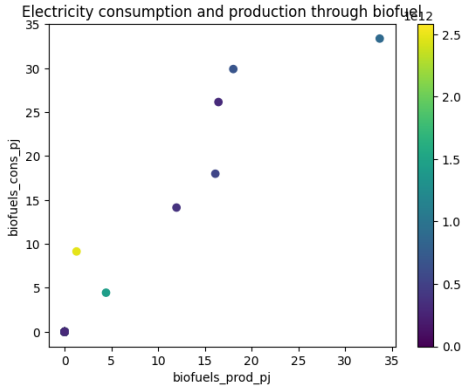

Fig. 41. Electricity consumption and production through biofuels (without outliers) (2021)

with two variations: one with outliers included and another with outliers removed.

Inferences

- **Boxplots for GDP PPP Metrics:** Refer to VIII-A1 for more details as to why exploring normalized data is crucial for gaining a more nuanced understanding of

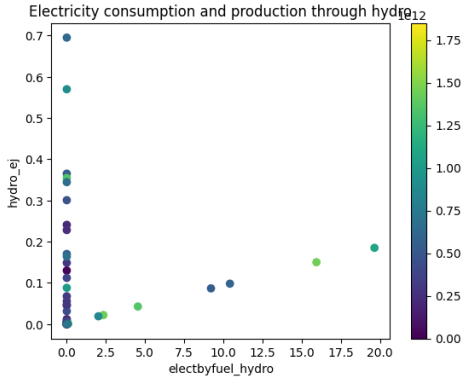

Fig. 42. Electricity consumption and production through hydro (without outliers) (2021)

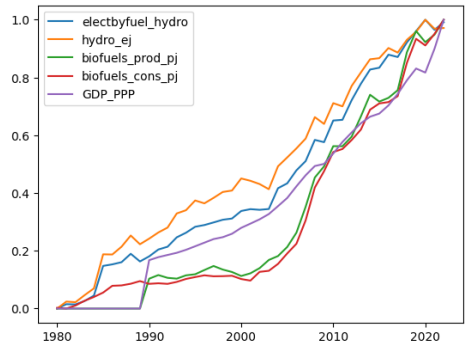

Fig. 43. Total values summed across all countries over the years

the relationships between economic metrics and energy variables.

- **Color-mapped scatter plots**

  Biofuel Consumption and Production Plot: Refer to Fig 40 & Fig 41
  - The USA and Brazil emerge as major producers and consumers, clearly identifiable as individual points. However, in the bottom left, there is a dense cloud of violet points, making it challenging to extract meaningful information.
  - Following outlier removal, the previously dense cluster becomes sparser, providing clearer visibility of individual countries while still adhering to the observed trend.

  Hydro Consumption and Production Plot: Refer to Fig 42
  - **Positive Correlation in plot:** Higher hydro electricity production corresponds to higher consumption in countries, even when outliers are included.
  - **Categorization by GDP Per Capita:** We have 3 main types of countries:
    * High GDP per capita countries (yellow, light green) tend to exhibit elevated levels of both hydroelectricity consumption and production.
    * Middle GDP per capita countries (dark green) show a mixed pattern in hydroelectricity consumption and production levels.
    * Low GDP per capita countries (dark violet) display lower levels of hydroelectricity consumption and production.

*2) **Aggregates over the years**:* Grouped biofuels, hydro, and GDP PPP metrics by year for time series data. Created a line chart with Min-Max scaling for normalization, showcasing temporal trends. Colored lines facilitated comprehensive analysis of energy and economic evolution (Fig 43). Refer to VIII-A3 for more details.

Inferences: We will be analyzing the trends for these 3 years: (1980, 2000, 2020). Each of these years serves as a significant milestone, capturing distinct phases in the development,

adoption, and transformation of energy sources.

- **Year 1980:**
  - **Low Biofuel Production and Consumption:** In the infancy of biofuel technology, high costs, and limited government support contribute to low production and consumption.
  - **Low GDP per Capita:** Reflects an early stage in global economic development.
- **Year 2000:**
  - **Plateauing Hydro Production:** Limited investment in new infrastructure, environmental concerns, and competition from renewables shape the trend.
  - **Stable Hydro Consumption:** Hydro maintains significance, filling the gap in renewable energy development.
  - **Initial Biofuel Growth:** Government support and technological advancements yield results.
  - **Growing Biofuel Consumption:** Increasing availability and awareness contribute to rising demand.
  - **Significant GDP Growth:** Technological advancements and globalization drive economic prosperity.
- **Year 2020:**
  - **Renewed Hydro Growth:** Rekindled interest due to cleanliness, reliability, and advancements in efficiency and environmental impact reduction.
  - **Slight Decline in Hydro Consumption:** Other renewables gain a growing share in the energy mix.
  - **Substantial Biofuel Growth:** Continued government support, technological advancements, and cost reduction contribute to increased production.
  - **Significant Biofuel Consumption:** Rising environmental awareness and cost-competitiveness drive demand.
  - **Slower GDP Growth:** A global economic landscape experiences slower growth compared to previous decades.

REFERENCES

[1] https://github.com/YousefGh/kmeans-feature-importance/tree/kmeans-feature-importance-v01