STAT 112 - Introduction to Data Processing and Visualization Project (Group 8)

Energy Consumption and The Main Features That Affect It

Tunahan Turgut  
Statistics Department  
*Middle East Technical University*Ankara, Türkiye  
tunahan.turgut@metu.edu.tr

Ahmet Furkan Köşedaşı  
Statistics Department  
*Middle East Technical University*Ankara, Türkiye  
furkan.kosedasi@metu.edu.trMehmet Toprak Özen  
Statistics Department  
*Middle East Technical University*Ankara, Türkiye  
e266673@metu.edu.tr

Bilginalp Büyüktaş  
Statistics Department  
*Middle East Technical University*Ankara, Türkiye  
e269444@metu.edu.trDavut Varlık  
Statistics Department  
*Middle East Technical University*Ankara, Türkiye  
e266681@metu.edu.tr

Arda Ergüven  
Statistics Department  
*Middle East Technical University*Ankara, Türkiye  
arda.erguven@metu.edu.tr

Sertaç Kandemir  
Statistics Department  
*Middle East Technical University*Ankara, Türkiye  
e266661@metu.edu.tr

*Abstract:* The aim of this research is to analyze and visualize energy consumption, the factors influencing it, and the factors that, in turn, affect these influences. The dataset given required precise tidying and cleaning, which is significant for our visualizations. This is because we cannot visualize data if it is not clean, as it would result in irrelevant information.

# Introduction

Energy consumption may be influenced by many factors. Our report’s main objective is to examine these factors to get relevant and rational conclusions from the visualizations. If we concentrate on our dataset, we would observe that the dirty dataset “Dirty\_Energy\_Consumption” contains 736 rows of information and 14 variables.

# DATA TIDYING AND CLEANING STEPS

In the data cleaning step we tried each delimiter using pd.read\_csv(). When we found some data into a single column, we used .split() to carefully separate the values.

*String formatting*

Next for the column names with random spaces and inconsistent capitalization. We used a chain of string methods to clean them up: first .str.strip() to get rid of extra spaces, then .str.lower() to make everything lowercase, and finally .str.replace(' ', '\_') to connect words with underscores.

For numeric columns like population and GDP, we had to do some serious cleaning. Using .str.replace(r'[^\d.-]', '') with regex, we stripped away any characters that weren't numbers or decimal points. Then we used pd.to\_numeric() to convert these cleaned strings into proper numbers.

The categorical columns needed their own special attention. We used .str.lower() and .str.strip() to make everything consistent. For income groups, we created a mapping dictionary converting various versions of "high income" or "upper middle" into our standardized categories using .map(income\_group\_mapping).

*Uniformity*

The data cleaning process enforced uniformity using .clip() for numeric ranges, .map() for categorical standardization, and .str methods for text formatting, ensuring consistent data representation throughout the energy consumption dataset..

*Outliers*

We used the interquartile range method (calculating Q1 and Q3 with .quantile()) to identify them, and replacing them with median values using .loc

*Missing Values*

We used .isnull().mean() to count how many we had. If a column had too many missing values (more than 65%), we used .drop() for the all collumn. For the remaining gaps, we filled numeric columns with their median values and categorical columns with their most common values using .fillna().

Finally, we gave everything its proper type. Numeric columns became floats with .astype(float), categorical columns became categories with .astype('category'), and years became integers with .astype(int).

Throughout this whole process, we kept a watchful eye for any errors, wrapping everything in a try-except block. If anything went wrong, we had debugging tools ready to inspect the data using .head() and check the shape of our DataFrame.

*Duplicates*

We checked the duplicates using pd.DataFrame.duplicated() both before and after the cleaning process. It did not contain any duplicated rows.

# EXPLORATORY DATA ANALYSIS (EDA)

After the data cleaning process, we started working on analyzing the dataset, which contains energy consumption and environmental data for 700 observations across different countries and regions. There are 14 columns in total, 7 of which are continuous variables and the remaining 7 are categorical variables. The variable information is as follows:

*(variable name, description, scale)*

Year: The year of observation (2000–2023), Interval

Country: Name of the country. Matches its corresponding region, Nominal

Population: Population of the country (continuous variable, in range 1,000,000–300,000,000), Ratio, Continuous

GDP\_per\_Capita: Gross Domestic Product per capita in USD (continuous variable, in range 1,000–60,000 USD), Ratio, Continuous

Energy\_Consumption\_per\_Capita: Annual energy consumption per capita in MWh (continuous variable, in range 1–10 MWh), Ratio, Continuous

Carbon\_Emission: Carbon dioxide emissions in tons per capita (continuous variable, derived from energy consumption), Ratio, Continuous

Renewable\_Energy\_Share: Percentage of energy derived from renewable sources (continuous variable, in range 10%–90%), Ratio, Continuous

Electricity\_Price: Electricity price per kWh in USD (continuous variable, in range $0.05–$0.30), Ratio, Continuous

Region: Geographical region of the country. Categories: North America, Europe, Asia, South America, Africa, Oceania, Nominal

Energy\_Policy: Stringency of energy policies ('Strict', 'Moderate', 'Lenient'), Ordinal

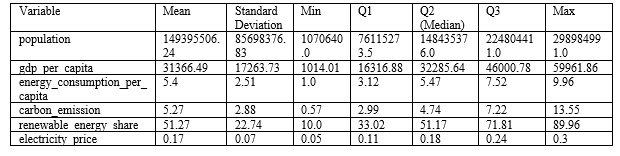
Climate\_Agreement: Climate agreement the country adheres to ('Paris', 'Kyoto', 'None'), Nominal

Urbanization: Level of urbanization ('High', 'Medium', 'Low'), Ordinal

Energy\_Source: Dominant energy source ('Coal', 'Natural Gas', 'Renewables', 'Nuclear', 'Oil').

Income\_Group: Income classification of the country ('High', 'Upper-Middle', 'Lower-Middle', 'Low'), Ordinal

We created the following table that contains the descriptive statistics for numerical columns, so that we can investigate and understand the data properly. In this way, we were able to observe the mean, the standard deviation, the quartiles, the minimum, and the maximum values of each continuous variable.



GDP\_per\_Capita (Gross Domestic Product per Capita)

The Total Gross Domestic Product divided by the population. Usually used to account for an estimate of economic welfare and standards of living, where a high value will be interpreted with high levels of economic welfare and income. Normally, countries with high GDP\_per\_Capita show modern infrastructures, more advanced technologies, and higher energy use. The variable is of extreme relevance to explaining the relationship between energy demand, economic well-being, industrial development, and environmental impacts.

Renewable\_Energy\_Share (Percentage of Energy from Renewable Sources)

Renewable\_Energy\_Share is that part of a country's total energy consumption derived from renewable sources: solar, wind, hydropower, and geothermal. Indirectly, it provides evidence of a country's sustainable energy policy and commitment to moving away from dependency on fossil fuels. Generally speaking, a country with a larger share of renewable energy focuses on the reduction of greenhouse gas emissions, combating climate change, and reaching long-term energy security. Understanding the Renewable\_Energy\_Share will enable assessment of the environmental performance of countries in light of efforts toward global sustainability and international agreements on climate change.

We tried to analyze the given dataset by asking 7 research questions.

1. How does per capita energy consumption progress over time in regions? (Tunahan Turgut)

2. Electricity prices and what affects it? (Mehmet Toprak Özen)

3. What potential trends or correlations can be identified between energy consumption per capita, GDP per capita, and renewable energy share for different climate agreement statuses? (Davut Varlık)

4. How do countries' participation in climate agreements and their energy policies correlate with changes in carbon emissions? (Ahmet Furkan Köşedaşı)

5. How does the distribution of renewable energy share differ across levels of urbanization? (Bilginalp Büyüktaş)

6. How does the distribution of carbon emissions vary across different energy sources for various regions? (Arda Ergüven)

7. How does energy consumption per capita differ among low, higher-middle, lower-middle, and high-income groups? (Sertaç Kandemir).

*Question 1: How does per capita energy consumption progress over time in regions?*

Energy consumption is a type of continuous variable that we can analyze across different regions and dates. In this experimental data analysis phase, we tried to reveal how it changes according to regions and time intervals, with the help of a table. Since there were multiple data rows containing the same dates and regions in our cleaned dataset, it was appropriate to use the average value of per capita energy consumption in these rows to represent the per capita energy value in the table. Then, representing the continuous variable energy\_consumption\_per\_capita using color intensity levels in a heat map turned out to be a suitable visualization choice. Furthermore, dividing the 2000–2023 year range given to us in the dataset into 4-year parts is useful both to improve visualization quality and to reduce information loss in the dataset to a negligible level

This heat-map illustrates the energy consumption per capita with measured in MWh (Megawatt hour), across various regions over 6 time intervals 4 years each. Each cell represents energy consumption per capita for a specific region and time period, with color ranging from light yellow ,which represents lower consumption, to dark blue that represents higher consumption. Moreover the exact mean value of the specific time interval and region can be reached by looking; hence, numerical comparisons can be made while interpreting the table.

*Key findings:*

• The North American and European regions follow a relatively close energy consumption trend, and the southern hemisphere regions of Oceania and South America also follow a relatively close consumption trend among themselves.

• There was a remarkable decrease in per capita energy consumption in Europe and North America between 2012 and 2015. In fact, these two regions saw by far the lowest energy consumption values ,with 4.2 MWh and 3.7 MWh, in this time period; unlike them and other regions, there was an increase in per capita energy consumption values in Africa and Asia. In fact, Africa saw its peak value with 6.5 MWh per capita between these years.

• The highest energy consumption per capita values, with 6.7 MWh per capita, belonged to the Oceania region between 2004 and 2007.

• Looking at recent years (2020–2023), South America is clearly at the top in energy consumption per capita, with 6.6 MWh per capita.

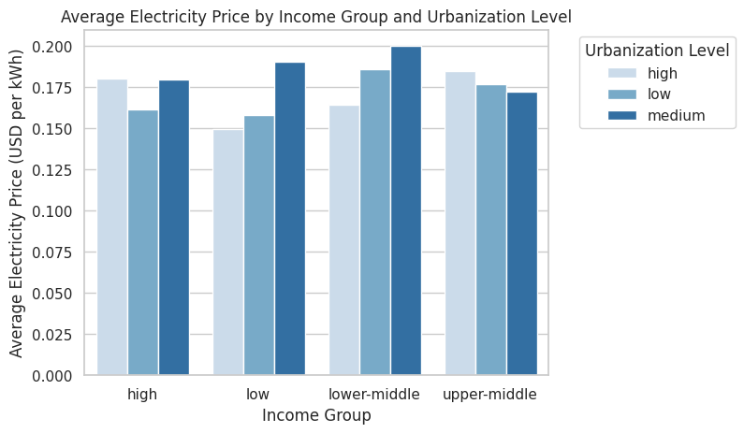
• Energy consumption in the Asian region remains on trend more stable than in other regions. In addition, while the energy consumption trend changes faster in the Europe region, the energy consumption trend decreases and increases over longer periods in the Oceania region.

• No region is far ahead of the other in terms of total energy consumption values per capita.

*Question 2: Electricity Prices and What Affects it*

Electricity price is a qualitative type of data. It is affected by some variables such as Energy Policy, Income Group, Urbanization Level, Energy Source, and Region. We created three plots to visualize and answer our question.

*a) Average Electricity Price by Income Group and Urbanization Level*



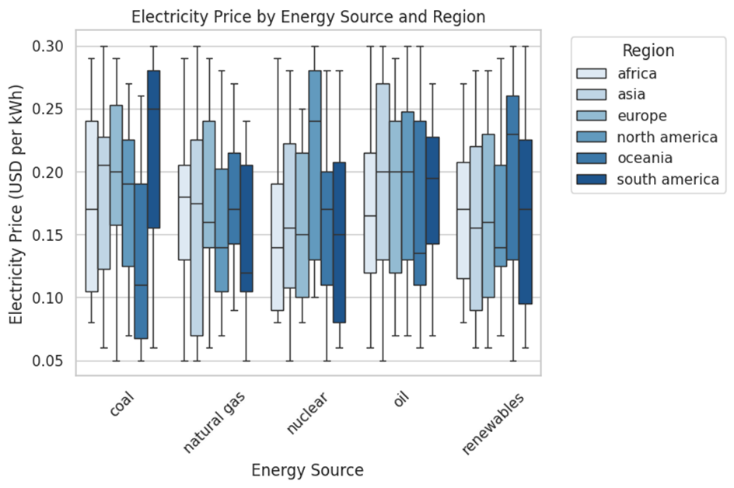
This bar chart compares the average price of electricity in USD per kWh across income groups classified into three different levels of urbanization: high, medium, and low. There are four classes of income: high, upper-middle, lower-middle, and low. Bars shaded differently for each category of urbanization show the fluctuations in prices across these classes.

The graph shows that, on average, low-income groups pay much more for electricity compared to the lower-middle and upper-middle income. Particularly large, it has reached above an average of 0.18 USD per kWh in regions with a low and middle level of urbanization. That low-income area would bear these high costs signals underdeveloped infrastructures, very expensive imported energy, and inefficiencies within energy distribution channels.

Electricity prices are more stable in the lower-middle-income and upper-middle-income groups, with relatively small differences between urbanization levels. The price of electricity for the lower-middle-income group, for example, is nearly constant across all classes of urbanization, standing at about 0.17 USD per kWh. The upper-middle-income group maintains moderate prices, hovering between 0.14 and 0.16 USD per kWh regardless of the level of urbanization.

A very interesting trend in high-income regions is that the price of electricity for highly urbanized areas is only marginally higher than for medium- or low-urbanized areas. This might be a function of advanced infrastructure and maybe accompanying policies considering the environment or other sources of energy.

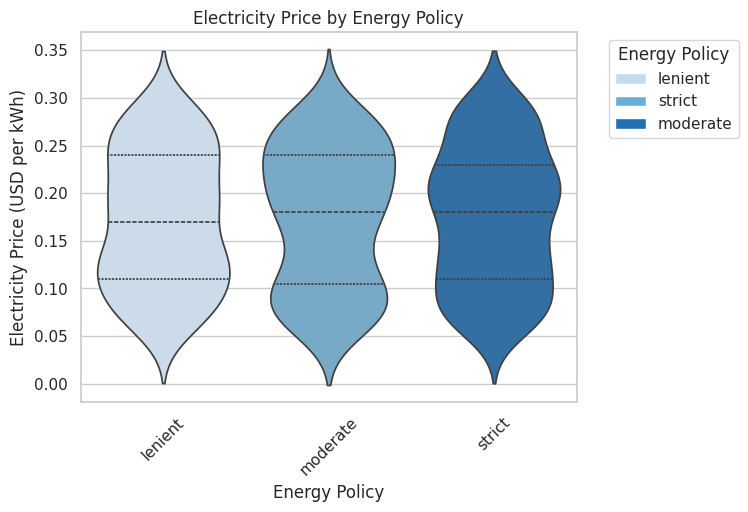
*b) Electricity Price by Energy Source and Region*



This box-and-whisker plot compares the spreads in electricity prices, based on energy sources (coal, natural gas, nuclear, oil, renewables) for six regions worldwide: Africa, Asia, Europe, North America, Oceania, and South America. The chart illustrates how the trends in prices and distributions for each source vary across regions.

The largest variability in electricity prices is found in regions that are highly dependent on renewable sources of energy. This variability reflects differences in the adoption of renewable technologies, governmental subsidies, and regional capacities for producing renewable energy. For example, Europe and North America have higher shares of renewables and thus tend to have higher electricity prices. In contrast, regions like Africa have lower renewable-related prices, probably because of smaller-scale deployments and therefore lower production costs.

Regional Disparities: Europe and North America generally have higher electricity prices for all energy sources compared to Africa and South America. This trend reflects regional disparities in energy policies, access to resources, and economic structures. While developed regions often invest in cleaner but costlier technologies, developing regions prioritize affordability and resource availability.



This violin plot shows the distribution of electricity prices by different energy policies, categorized into lenient, moderate, and strict. This visualization shows how policy can affect electricity pricing.

Strict Policies: The distribution of the price of electricity is widest across regions with strict energy policies. The higher average prices of electricity in such regions are indicative of the greater costs associated with their transition towards sustainable energy systems, meaning investments in renewable energy infrastructure and phasing out fossil fuels.

Lenient Policies: In regions with lenient policies, electricity prices are relatively low. The regions tend to use inexpensive and available sources of fossil fuel to provide more predictable and stable pricing structures.

Moderate Policies: The price distribution of the moderate policies falls between lenient and strict policies.

*Question 3: What potential trends or correlations can be identified between energy consumption per capita, GDP per capita, and renewable energy share for different climate agreement statuses?*

ekran görüntüsü, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

People in countries part of the Paris Agreement use less energy per person (5.26). Surprisingly, countries with no agreements consume more energy than them (5.65). The amount of energy consumed by countries with the Kyoto Agreement falls in between (5.58), which raises questions about the actual impact of these accords on energy consumption practices.

Richest nations without climate pacts hold the highest GDP per person ($32,251), perhaps indicating that high-income nations opt out of climate accords due to business or political interests. Kyoto protocol nations show a GDP ($31,325) that overshadows the GDP of Paris signatories ($31,195), exhibiting minor economic disparities.

Nations under the Kyoto Agreement take the lead in renewable resource use (52.97%) showcasing a longer commitment compared to Paris Agreement nations. Coincidentally, countries without formal climate agreements maintain almost similar figures to Kyoto (51.98%), indicating that push towards renewable sources is not confined to set agreements. Whereas, Paris-based countries have a smaller portion (50.20%), showing that recent commitments haven't yet lead significant renewable surge.

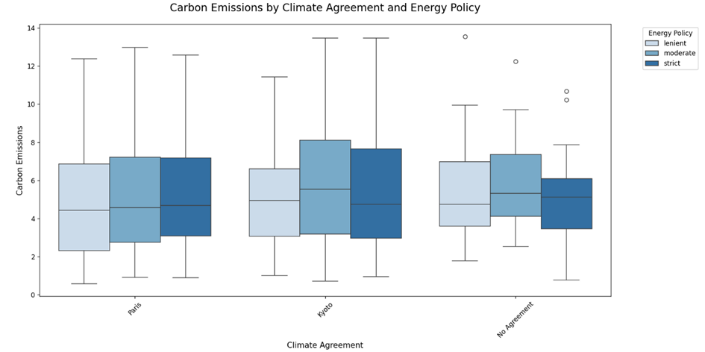
In both Kyoto and non-agreement nations, increased renewable usage correlates with reduced energy consumption ‒ suggesting improvements in efficiency or maybe higher renewable incorporation where consumption is already moderate.

Paris acts as an anomaly: it has lower shares of renewables yet lesser overall energy consumption ‒ could be due to gradual transitions towards greener resources.

Affluent nations without climate pacts may still pursue renewable energy goals independently, mirroring their economic capacity or national preferences. Long-standing commitments like Kyoto tend to assist countries in achieving a greater percentage of renewables as compared to Paris Agreement nations.

Nations respecting varying climate agreements display differences in energy consumption, GDP and renewable usage. Adoption of renewable resources doesn't seem strictly tied to formal agreements, implying an inverse connection between consumption levels and increased green solutions, particularly for Kyoto and non-signatory nations. Paris stands as an example of nations working towards including greener energy options

*Question 4: How do countries' participation in climate agreements and their energy policies correlate with changes in carbon emissions?*



The box plot is a great way to show how carbon emissions vary under different combinations of climate agreements and energy policies. This box plot shows how carbon emissions are influenced by different combinations of climate agreements and the strictness of energy policies. The goal is to explore trends and patterns which reveal how these factors shape carbon emissions levels.

*a) Paris Climate Agreement*

Lenient Policy:

Under Paris, countries with lenient policies have the lowest emissions among all categories, with minimal variability. This would suggest that even lenient policies, when supported by Paris, might keep emissions in check.

Moderate Policy:

With more variation between countries, emissions increase slightly compared to lenient policies. However, the framework of Paris keeps emissions relatively stable.

Strict Policy:

Strict policies result in slightly higher emissions and greater variability than moderate ones. This suggests that challenges in implementing stricter measures effectively, even under Paris.

*b) Kyoto Protocol:*

Lenient Policy:

Compared to Paris, emissions are more variable and higher. This suggests that lenient policies make Kyoto's framework less effective.

Moderate Policy:

Kyoto's moderate policies resulted in the highest emissions, indicating that they are not sufficiently enforced to encourage reductions.

Strict Policy:

Strict policies lead to lower emissions than lenient or moderate ones, but they are still higher, and more variable compared to Paris.

*c) No Agreement*

Lenient Policy:

Lenient policies without agreements show the highest emissions and the most variability, demonstrating their efficiency in reducing emissions.

Moderate Policy:

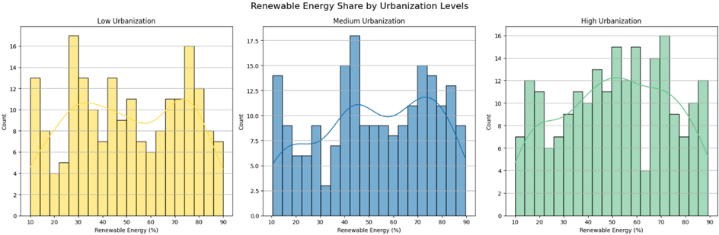
Emissions remain high, though slightly less variable than lenient policies, showing limited impact without international coordination.

Strict Policy:

Strict policies, even without an agreement, lead to high emissions and significant variability. This demonstrates the importance of international cooperation in improving the effectiveness of domestic policies.

*Question 5: How does the distribution of renewable energy share differ across levels of urbanization?*

As part of the analysis, we considered the distribution of renewable energy adoption across areas with different levels of urbanization—low, medium, and high. The visualization is to understand how the share of renewable energy differs and what patterns might emerge in these distinct settings. Below, we summarize our findings based on the histogram and data.

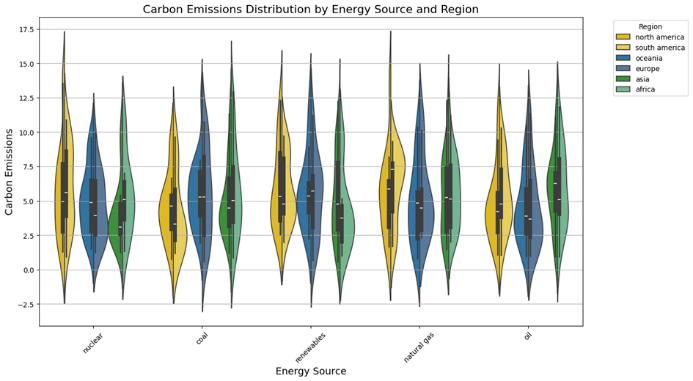


In the areas with low urbanization, we find much variation in renewable energy adoption: Two big clusters are noticed, one at the lower range of 10–20% and another at the higher range of 70–80%. Adoption in the middle range between 30–60% is far less. The trend looks very irregular, which may suggest poor access to renewable energy resources or supportive policies.

For medium urbanization, in these regions, the renewable energy adoption distribution was more concentrated, with peaks at 40–50% and 70–80%, reflecting moderate to high dependence on renewables. It was smoother than that of the lowly urbanized areas, showing a more even pattern. However, adoptions below 30% were still not common to suggest that the use of non-renewable energy was still low.

In highly urbanized, the shares of renewable energy are most evenly distributed, ranging between 40% to 80%. The density curve for this is smooth; this could be systematic, even adoption due to entrenched policies and structures. Counts below 30% remained modest, showing limited reliance on low-renewable energy shares.

*Question 6: How does the distribution of carbon emissions vary across different energy sources for various regions?*



The carbon emissions are correlated to the energy sources from different regions in order to be able to show the ways through which energy choices or sources of energy produce environmental outcomes. A violin plot has been created to display carbon emissions across various energy sources, colored by region, for easy comparison.

The outcome has shown that the carbon emission by different sources of energy is very far from each other, with some of them giving very high emissions, such as those emitted by coal and natural gas, while others are giving much lower values, like solar and wind. This finding underlines the importance of energy sources in defining carbon footprints.

Great variability can be observed in regional analysis. Some regions, such as Asia and Oceania, have higher ranges of dispersion for the same energy vectors, denoting different energy practices and dependences on high-emission fuels. In Europe and North America, it is smaller and lower; this might be explained by more restrictive environmental policies and greater diffusion of renewable energy.

These results mean that not only is the choice of energy source significant but also the regional energy policy for carbon emission variability. Promotion of renewable energy sources and creating their effective polices would, therefore, be in a position to reduce the same especially for regions with high variability or a high dependence on non-renewable sources.

*Question 7: How does energy consumption per capita differ among low, higher-middle, lower-middle, and high-income groups?*

**metin, ekran görüntüsü, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu***Average Energy Consumption per Capita by Income Group*

This bar plot compares the average energy consumption per capita among the three income groups: lower-middle, upper-middle, and high income. High-income countries continue to have the highest energy consumption per capita, driven by industrialization, advanced infrastructure, and access to diverse energy resources. Upper-middle-income countries show moderate energy consumption, likely reflecting their transition to higher development stages with increasing energy demand. Lower-middle-income countries have the lowest energy consumption per capita, showing limited energy access and infrastructure challenges.

A diagram of a graph

Description automatically generated*Distribution of Energy Consumption per Capita by Income Group*

The box plot highlights the variability and spread of energy consumption per capita within each income group, showing the interquartile range, median, and outliers .High-income countries have lower variability, showing more uniform energy consumption levels across this group. Upper-middle-income countries show a wider spread, with some countries consuming as much as high-income nations, potentially due to large industrial hubs. Lower-middle-income countries display significant variability, with notable outliers consuming disproportionately higher energy resource-rich but economically developing nations.

**çizgi, metin, ekran görüntüsü, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu***Temporal Trends of Energy Consumption per Capita by Income Group*

The energy consumption trends from 2014-2022 show that rich countries kept using high amounts of energy with a slight increase, while upper-middle income countries used less energy over time. Both lower-middle and low income countries increased their energy use, likely because they are developing and building more infrastructure. This pattern suggests that developed nations have stable energy needs, while developing countries are increasing their energy use to support economic growth.

IV CONCLUSION

Study indicates that selecting appropriate energy sources and implementing specific energy policies for different regions have a great impact on carbon emission variability. Strict policies tend to be more effective, and their successful implementation is more likely to be achieved when supported by international agreements. Urban areas have broader access to renewable energy sources compared to low urbanization areas. Emphasizing the need for equitable energy transitions across different areas. It is crucial for attaining both ecological and financial sustainability in times ahead. Transitioning from nonrenewable to renewable energy sources raises challenges. Namely, higher electricity prices and imbalances between emission reductions and energy affordability. Countries that are more developed have consistent energy requirements, while nations still in progression are increasing their energy consumption to support economic growth. It is inferred that the global energy system still may not be ready for a full transition to the use of renewable energy sources