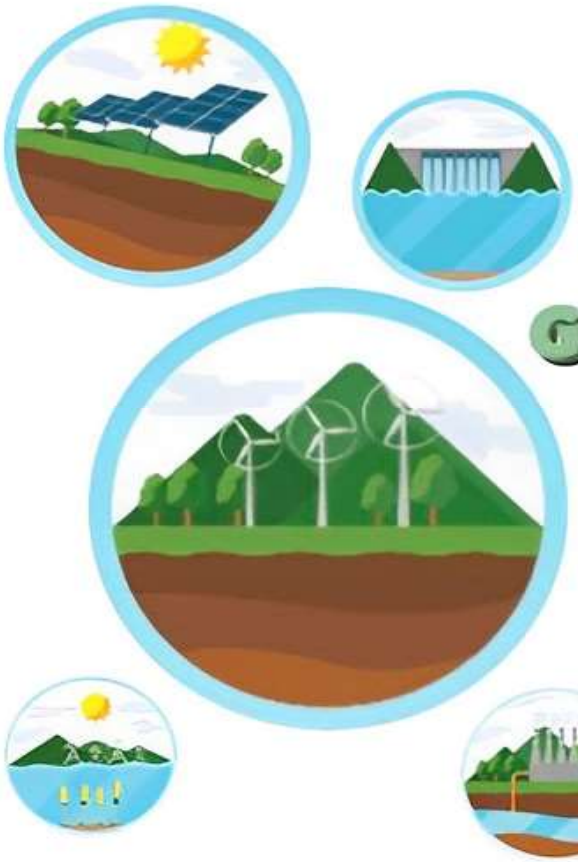


Global Environmental Impact Index

By : Annas Khan

Serial No.	Title
1	Environmental Impact Report
2	Objective
3	Reason for Choosing this Dataset.
4	Identify the Problem and Justification
5	The relevance of the Global Data on Sustainable Energy
6	Addressing Multidimensionality
7	Imputation of Missing Data
8	Multivariate Analysis
9	Multiple Regression Analysis
10	Other Visualization of Results
11	Weighting and Aggregation
12	Composite Index
13	Linking to Other Indexes
14	Ordinary Least Square Regression Model

Global Data on Sustainable Energy Report.

The Global Data on Sustainable Energy Report aims to provide comprehensive insights into key aspects of sustainable energy, including energy access, renewables, and carbon emissions. This report enables cross-country comparisons to identify trends, challenges, and opportunities in the global energy landscape.

Objective

The primary aim of this project is to develop a Composite Index for Environmental Impact. This index will facilitate clearer and at-a-glance comparisons between key factors such as carbon emissions and renewable energy resources. By consolidating multiple environmental metrics into a single index, stakeholders can gain a comprehensive understanding of the environmental performance of different entities, whether they are countries, industries, or specific projects.

Reason for Choosing this Dataset.

1. **Global Relevance:** The dataset is globally relevant due to its focus on critical environmental issues like climate change, carbon emissions, and renewable energy, offering insights applicable across countries and industries.
2. **Multidimensional Approach:** With a comprehensive view including factors such as carbon emissions, renewable energy resources, and environmental policies, the dataset enables a holistic assessment of environmental impact, capturing the interplay of diverse factors.
3. **Public Interest and Awareness**: Given the significant public interest in environmental issues, working with this dataset contributes to raising awareness, promoting informed decision-making, and encouraging sustainable practices at various levels.
4. **Policy Implications:** Policymakers and government entities can benefit from the dataset's insights for crafting effective environmental policies, targeting areas that require intervention or improvement based on evidence-based analysis.

Educational and Research Value: The dataset offers educational and research value, facilitating deeper exploration of environmental trends, patterns, and correlations for academia, researchers, and students in environmental science.
Business and Investment Insights: Businesses and investors can leverage the dataset for insights into sustainable investment opportunities, market trends, and corporate sustainability performance, aligning with environmental goals and stakeholder expectations.

Identify the Problem and Justification

The project's main objective is to establish a composite index that accurately gauges the environmental impact across various regions and countries. This initiative stems from the recognition that existing methods lack consistency and fail to offer a holistic view of crucial factors such as carbon emissions, renewable energy utilization, and environmental regulations. By integrating data from diverse sources, the project aims to address these gaps and provide a comprehensive assessment tool.
This standardized index will not only facilitate direct comparisons and benchmarking but also support evidence-based decision-making, policy formulation, and target setting for environmental sustainability. Furthermore, the project aligns with global initiatives like the Paris Agreement and sustainable development goals, contributing to transparency, accountability, and collective efforts towards a greener and more sustainable future on a global scale.

The relevance of the Global Data on Sustainable Energy

It is a multidimensional approach towards assessing critical aspects of energy sustainability, environmental impact, and renewable resources. This dataset is essential for several reasons:

- **Holistic Insight:** The dataset provides a comprehensive view of energy access, renewable energy deployment, carbon emissions, and other key metrics, allowing for a holistic understanding of a country's sustainability efforts.
- **Comparative Analysis:** It enables cross-country comparisons, facilitating the identification of best practices, areas for improvement, and benchmarks for sustainable energy policies.
- **Decision Support:** Policymakers, researchers, and stakeholders can use this data to make informed decisions, set realistic targets, and monitor progress towards sustainable energy transitions.
- **Global Impact:** Given the global nature of energy and environmental challenges, this dataset contributes to international cooperation, knowledge sharing, and collaborative efforts aimed at achieving global sustainability goals.

Addressing Multidimensionality

Global Data on Sustainable Energy plays a pivotal role in addressing the multidimensional challenges of sustainability. Its comprehensive metrics cover a wide spectrum, including energy access, renewable energy capacity, carbon emissions, and energy efficiency, providing a holistic view of sustainability. This breadth of data allows for detailed analysis across different dimensions, regions, and time periods. By integrating data from various disciplines, such as energy, environment, economics, and social factors, the dataset enables interdisciplinary insights, essential for understanding the interconnected nature of sustainability issues. Furthermore, its longitudinal analysis capabilities, incorporating historical data and trends, offer valuable insights into sustainability trajectories and potential future scenarios. With global coverage spanning multiple countries and regions, the dataset facilitates comparative analysis and benchmarking, fostering international cooperation and

enabling the identification of best practices. Overall, the multidimensional approach of the Global Data on Sustainable Energy is instrumental in supporting evidence-based decision-making, effective policy formulation, and fostering sustainable development worldwide.

Imputation of Missing Data

Step 1: I started by assessing the dataset to gain an insight into all the columns present. This helped me understand the scope of data cleaning needed.

```
<class 'pandas.core.frame.DataFrame'>
Index: 3333 entries, 0 to 3416
Data columns (total 53 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Countries                                               3333 non-null   object
 1   Year                                                    3333 non-null   float64
 2   Renewable-electricity-generating-capacity-per-capita    3333 non-null   float64
 3   Access to clean fuels for cooking                       3333 non-null   float64
 4   Value_co2_emissions_kt_by_country                       3333 non-null   float64
 5   Electricity from fossil fuels (TWh)                     3333 non-null   float64
 6   Renewables (% equivalent primary energy)                3333 non-null   float64
 7   Financial flows to developing countries (US $)          3333 non-null   float64
 8   Primary energy consumption per capita (kWh/person)      3333 non-null   float64
 9   Renewable-electricity-generating-capacity-per-capita.1  3333 non-null   float64
 10  Electricity from renewables (TWh)                       3333 non-null   float64
 11  Renewable energy share in the total final energy consumption (%)  3333 non-null   float64
 12  Energy intensity level of primary energy (MJ/$2017 PPP GDP)  3333 non-null   float64
 13  Low-carbon electricity (% electricity)                  3333 non-null   float64
 14  gdp_growth                                              3333 non-null   float64
 15  Ratio_REC_to_PEC                                        3333 non-null   float64
 16  Ratio_Elec_Renewables_to_Total                          3333 non-null   float64
 17  Country                                                 3333 non-null   object
 18  Density
(P/Km2)                                                      3333 non-null   object
```

Step 2: The filtering was carried out using pandas, a powerful data manipulation library in Python. The DataFrame named df was filtered to include only rows where the 'Year' column had a value of 2020. This was achieved using the following code I filter the data for year 2020 as it was very big dataset with data of 20 year from 2000 – 2020 it help me to understand the data

```
(variable) df_filtered_2020: DataFrame
df_filtered_2020 = df[df['Year'] == 2020]

# Print the filtered DataFrame for 2020
print("Data only for the year 2020:")
print(df_filtered_2020)
```

```
Data only for the year 2020:
                      Countries  Year  \
20                    Afghanistan  2020
41                        Albania  2020
62                        Algeria  2020
83                         Angola  2020
104           Antigua and Barbuda  2020
125                     Argentina  2020
146                       Armenia  2020
167                         Aruba  2020
188                     Australia  2020
209                       Austria  2020
230                    Azerbaijan  2020
251                       Bahamas  2020
272                       Bahrain  2020
293                    Bangladesh  2020
314                      Barbados  2020
335                       Belarus  2020
356                       Belgium  2020
377                        Belize  2020
398                         Benin  2020
419                       Bermuda  2020
440                        Bhutan  2020
461         Bosnia and Herzegovina  2020
482                      Botswana  2020
```

Step 3: Drops unnecessary columns from the DataFrame based on a predefined list of columns to keep, and then filters the DataFrame to include data only for the year 2020, printing the resulting DataFrame.

```
💡 Click here to ask Blackbox to help you code faster
# Droping Unnessery Column

# Define the columns to keep
columns_to_keep = [
    "Countries", "Year",
    "Renewable-electricity-generating-capacity-per-capita", "Renewable energy share in the total f
    "Access to clean fuels for cooking",
    "Value_co2_emissions_kt_by_country", "Electricity from fossil fuels (TWh)",
    "Renewables (% equivalent primary energy)",
    "Access to electricity (% of population)", "Financial flows to developing countries (US $)",
    "Primary energy consumption per capita",
    "Renewable-electricity-generating-capacity-per-capita", "Electricity from renewables (TWh)",
    "Renewable energy share in the total final energy consumption",
    "Energy intensity level of primary energy", "Low-carbon electricity (% electricity)",
    "gdp_growth"
]

# Drop the columns not in columns_to_keep
df_filtered = df.drop(df.columns.difference(columns_to_keep), axis=1)

# Filter the DataFrame to include only the years from 2015 to 2020
df_filtered_years = df_filtered[(df_filtered['Year'] >= 2020) & (df_filtered['Year'] <= 2020)]

# Print the filtered DataFrame
print("Dropped the unnecessary data and printing a few rows:")
print(df_filtered_years.head())  # Print only the first few rows
```

Step 4: First, I defined a list called `kept_columns`, which contains the names of the columns I want to keep in my DataFrame. Then, I filtered the DataFrame to include only data from the year 2020 and the columns listed in `kept_columns`. After that, I counted how many missing values there are in each column of the filtered DataFrame to assess data completeness and printed the results to see if any columns need further attention or cleaning.

```
Missing values counts:
Countries                                                    0
Year                                                         0
Renewable-electricity-generating-capacity-per-capita        45
Access to clean fuels for cooking                            8
Value_co2_emissions_kt_by_country                          175
Electricity from fossil fuels (TWh)                          1
Renewables (% equivalent primary energy)                   103
Access to electricity (% of population)                      0
Financial flows to developing countries (US $)             174
Primary energy consumption per capita (kWh/person)           0
Renewable-electricity-generating-capacity-per-capita        45
Electricity from renewables (TWh)                            1
Renewable energy share in the total final energy consumption (%)    174
Energy intensity level of primary energy (MJ/$2017 PPP GDP)         174
Low-carbon electricity (% electricity)                       2
gdp_growth                                                  16
dtype: int64
```

Step 5: I first found the mean values for certain columns over the last 5 years. Then, I filled in missing data in those columns with their respective means, except for 'Access to clean fuels for cooking,' which I filled using the last known value. Finally, I printed the updated DataFrame with the missing values filled in and then check missing values found 0

```
Added the Missing value with mean value taking for 5 year data
     Countries  Year  Access to electricity (% of population)  \
0  Afghanistan  2000                                 1.613591
1  Afghanistan  2001                                 4.074574
2  Afghanistan  2002                                 9.409158
3  Afghanistan  2003                                14.738506
4  Afghanistan  2004                                20.064968

   Access to clean fuels for cooking  \
0                                6.2
1                                7.2
2                                8.2
3                                9.5
4                               10.9

   Renewable-electricity-generating-capacity-per-capita  \
0                                              9.22
1                                              8.86
2                                              8.47
3                                              8.09
4                                              7.75

   Financial flows to developing countries (US $)  \
0                                         20000.0
1                                        130000.0
...
1            60        652230.0  33.93911  67.709953
2            60        652230.0  33.93911  67.709953
3            60        652230.0  33.93911  67.709953
4            60        652230.0  33.93911  67.709953
```

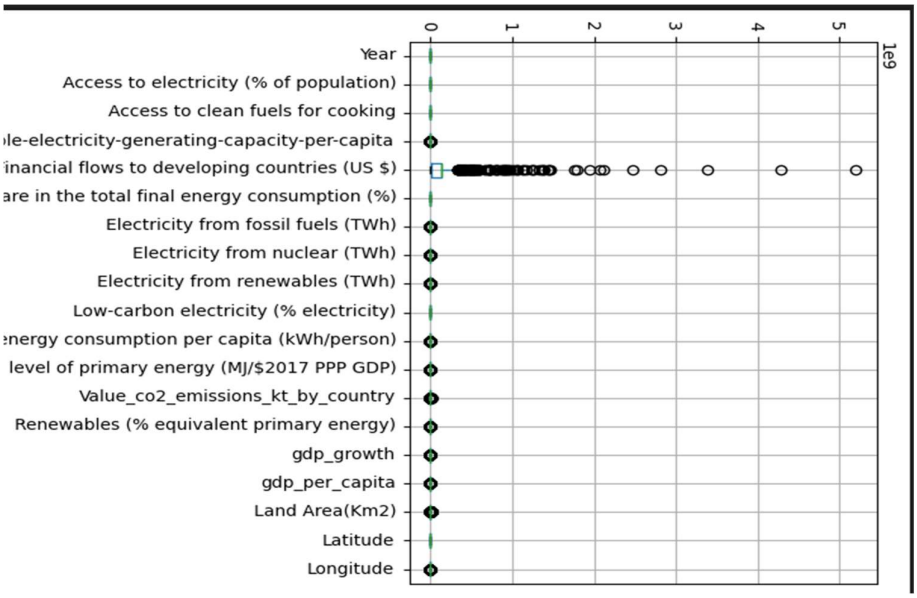
```
Missing values count after filling NA/NaN
Countries                                                    0
Year                                                         0
Renewable-electricity-generating-capacity-per-capita         0
Access to clean fuels for cooking                            0
Value_co2_emissions_kt_by_country                            0
Electricity from fossil fuels (TWh)                          0
Renewables (% equivalent primary energy)                     0
Financial flows to developing countries (US $)               0
Primary energy consumption per capita (kWh/person)           0
Renewable-electricity-generating-capacity-per-capita         0
Electricity from renewables (TWh)                            0
Renewable energy share in the total final energy consumption (%)    0
Energy intensity level of primary energy (MJ/$2017 PPP GDP)         0
Low-carbon electricity (% electricity)                       0
gdp_growth                                                   0
dtype: int64
Data after filling missing values saved to 'cleaned_data.csv'
```

Step 6:

```
Click here to ask Blackbox to help you code faster
# Check for duplicate rows
duplicate_rows = df.duplicated()

# Display duplicate rows, if any
if duplicate_rows.any():
    print("Duplicate rows found:")
    print(df[duplicate_rows])
else:
    print("No duplicate rows found.")
```

```
No duplicate rows found.
```

Step 7: I used Matplotlib to create boxplots for each numerical column in the DataFrame, which helps visualize the distribution of data. From the box plots, I noticed that most columns exhibit a normal distribution, except for 'Financial flows to developing countries (US $)' column, which appears to have some outliers or a non-normal distribution.



Step 8: I used pandas to manipulate the data and sklearn's MinMaxScaler to normalize specific columns in the DataFrame. I chose to normalize these columns because they were not normally distributed, which is often necessary for certain machine learning algorithms to work effectively. The code then prints the normalized columns for the year 2020 alongside the country names and year to demonstrate the scaled data.

```
Normalized the data  as the data is not normally distributed
              Countries  Year  \
20           Afghanistan  2020
41               Albania  2020
62               Algeria  2020
83                Angola  2020
104  Antigua and Barbuda  2020
125            Argentina  2020
146              Armenia  2020
167                Aruba  2020
188            Australia  2020
209              Austria  2020

     Renewable-electricity-generating-capacity-per-capita  \
20                                            0.003070
41                                            0.049303
62                                            0.005148
83                                            0.038122
104                                           0.055486
125                                           0.102837
146                                           0.159923
167                                           0.117898
188                                           0.049303
209                                           0.049303
```

Step 8: Check for Negative value in Data set

```
Click here to ask Blackbox to help you code faster
df = pd.read_csv("C:\\Users\\Khan Machine\\OneDrive - Dundalk Institute of Technology\\Desktop\\cl

## Check for negative values in each column and raise an assertion error if found
assert (df['Renewable-electricity-generating-capacity-per-capita'] >= 0).all(), "Negative values f
assert (df['Access to clean fuels for cooking'] >= 0).all(), "Negative values found in Access to c
assert (df['Value_co2_emissions_kt_by_country'] >= 0).all(), "Negative values found in Value_co2_e
assert (df['Electricity from fossil fuels (TWh)'] >= 0).all(), "Negative values found in Electrici
assert (df['Renewables (% equivalent primary energy)'] >= 0).all(), "Negative values found in Rene
assert (df['Financial flows to developing countries (US $)'] >= 0).all(), "Negative values found i
assert (df['Primary energy consumption per capita (kWh/person)'] >= 0).all(), "Negative values fou
assert (df['Renewable-electricity-generating-capacity-per-capita'] >= 0).all(), "Negative values f
assert (df['Electricity from renewables (TWh)'] >= 0).all(), "Negative values found in Electricity
assert (df['Renewable energy share in the total final energy consumption (%)'] >= 0).all(), "Negat
assert (df['Energy intensity level of primary energy (MJ/$2017 PPP GDP)'] >= 0).all(), "Negative v
assert (df['Low-carbon electricity (% electricity)'] >= 0).all(), "Negative values found in Low-ca
assert (df['gdp_growth'] >= 0).all(), "Negative values found in gdp_growth"

print("No negative values found in any column.")


No negative values found in any column.
```

Step 9: I created two new ratio features in the DataFrame:

1. Ratio_REC_to_PEC (Renewable-electricity-generating-capacity-per-capita to Primary energy consumption per capita):
This ratio compares renewable electricity generation capacity per person to primary energy consumption per person, helping us understand the proportion of renewable energy capacity in relation to overall energy consumption.

2. Ratio_Elec_Renewables_to_Total (Electricity from renewables to Total Electricity generation):**

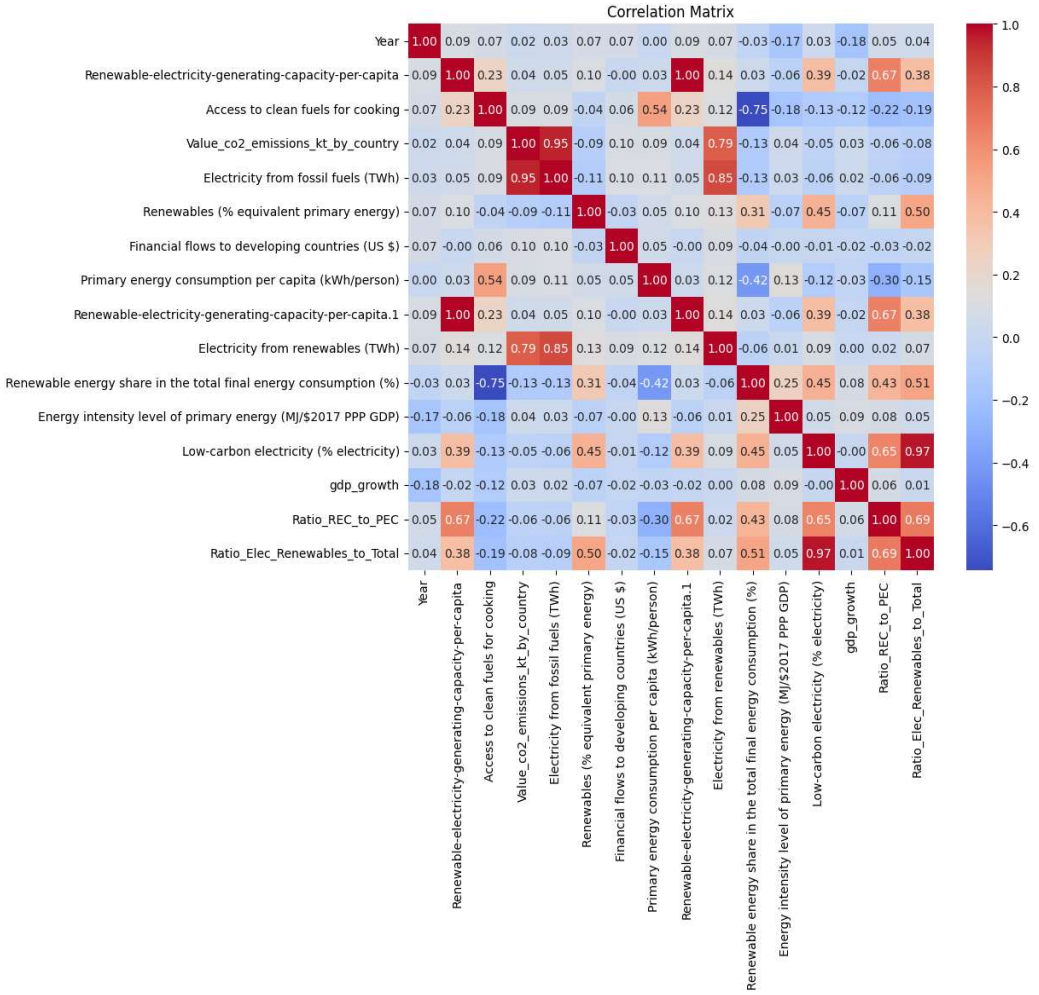
This ratio calculates the percentage of electricity generated from renewables compared to total electricity generation (including both renewables and fossil fuels), giving insights into the share of renewable energy in the entire electricity production mix.

h	Ratio_REC_to_PEC	Ratio_Elec_Renewables_to_Total
5	2.614521	0.659574
5	3.209269	0.847458
5	3.446732	0.811594
2	3.018587	0.670213
0	3.256132	0.629213
6	2.556483	0.634409
0	2.085836	0.761905
5	1.755959	0.789474
7	1.057353	0.739726
5	0.660021	0.829787
0	0.599906	0.797872
7	0.485763	0.769231

Multivariate Analysis

Heat Map

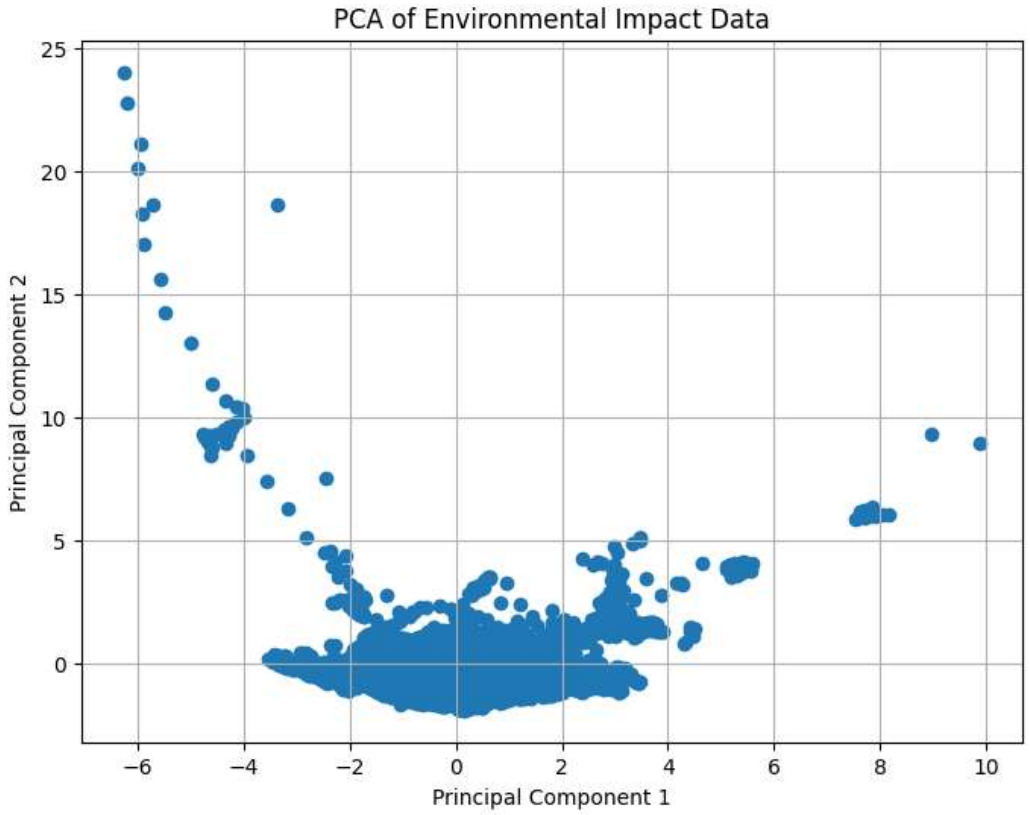
1. I selected only the numeric columns from 'df' for correlation computation.
2. I computed the correlation matrix to understand relationships between variables.
3. I explained that the correlation matrix is a table showing correlation coefficients between variables, where each cell indicates the correlation between two variables.
4. I generated a heatmap using Seaborn and Matplotlib to visually represent the correlation matrix, with annotations displaying correlation values, a 'coolwarm' colormap for clarity, a color bar for reference, and a square shape for accuracy. The heatmap is titled 'Correlation Matrix' to indicate its purpose.



PCA

I first checked for missing values (NaN) in the numeric data to ensure data quality. After confirming the presence of NaN values, I replaced them with column means using the `fillna` method, which is a common strategy to handle missing data effectively. Next, I standardized the numeric data using `StandardScaler` from scikit-learn, which is crucial for methods like Principal Component Analysis (PCA) to work optimally. PCA was then applied with 2

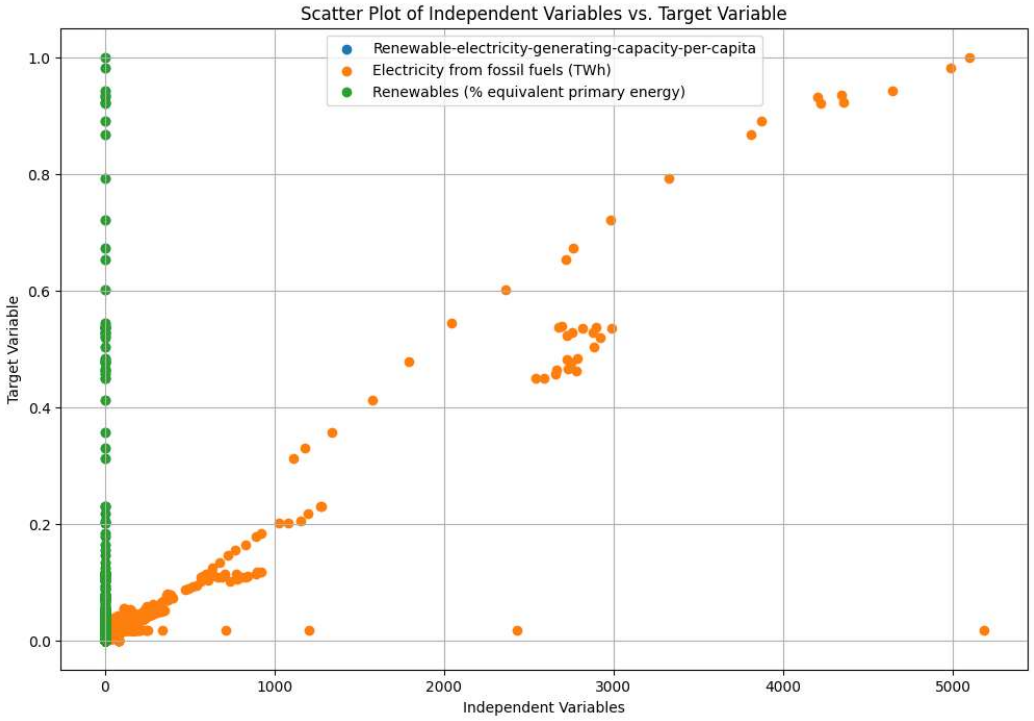
components to reduce the dimensionality of the data while retaining important information. The resulting principal components were stored in a new DataFrame and visualized in a scatter plot using Matplotlib, with labeled axes and a title indicating 'PCA of Environmental Impact Data'. This process represents a standard workflow for preprocessing and visualizing data using PCA, allowing for effective dimensionality reduction and exploratory data analysis.



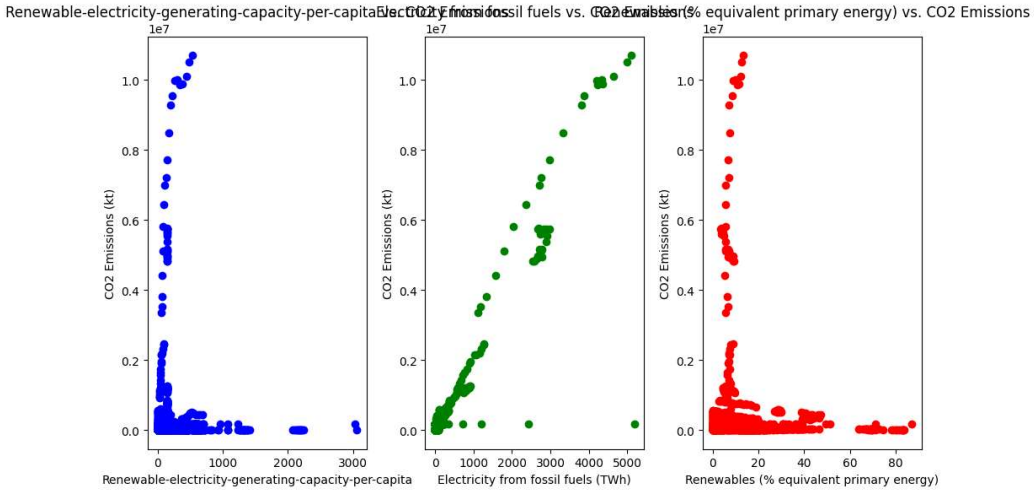
Multiple Regression Analysis:

The scatter plots between the target variable and multiple independent variables serves as a foundational step in preparing for multiple regression analysis. Scatter plots are effective tools for visually examining the relationships between variables, identifying potential trends or patterns, and detecting outliers.

Before diving into regression analysis, it's essential to understand how each independent variable relates to the target variable individually. The scatter plots allow us to see if there are any apparent linear or nonlinear relationships between the independent variables and the target variable. These visualizations can guide further analysis, such as checking for multicollinearity among independent variables, assessing the linearity assumption for regression, and determining the need for transformations or adjustments in the data.



The three scatter plots to visualize the relationships between three independent variables ("Renewable-electricity-generating-capacity-per-capita," "Electricity from fossil fuels (TWh)," and "Renewables (% equivalent primary energy)") and the target variable "CO2 Emissions (kt)." Each scatter plot represents one independent variable against the target variable, with distinct colors (blue, green, and red) for clarity. The plots are titled accordingly and labeled on both axes, providing a quick overview of how these variables relate to CO2 emissions.



The code conducts multiple regression analysis using the statsmodels library to explore how specific independent variables, such as renewable-electricity-generating-capacity-per-capita, electricity from fossil fuels (TWh), and renewables (% equivalent primary energy), contribute to explaining environmental impact measured by CO2 emissions (kt) indicated by "Value_co2_emissions_kt_by_country." The independent variables are added with a

constant term (intercept) and then fitted into an Ordinary Least Squares (OLS) regression model. The model summary is printed, providing detailed statistics and insights into the relationships and significance of each independent variable in explaining the target variable.

OLS Regression Results

Dep. Variable:	Value_co2_emissions_kt_by_country	R-squared:	0.900
Model:	OLS	Adj. R-squared:	0.900
Method:	Least Squares	F-statistic:	1.091e+04
Date:	Sun, 05 May 2024	Prob (F-statistic):	0.00
Time:	17:48:50	Log-Likelihood:	8836.0
No. Observations:	3649	AIC:	-1.766e+04
Df Residuals:	3645	BIC:	-1.764e+04
Df Model:	3		
Covariance Type:	nonrobust		

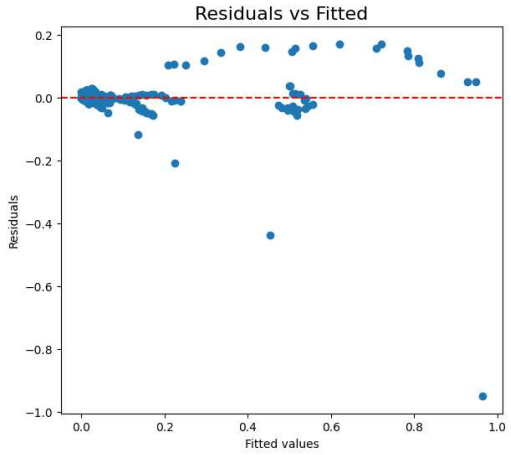
	coef	std err	t	P>|t|	[0.025
const	0.0010	0.001	1.639	0.101	-0.000
Renewable-electricity-generating-capacity-per-capita	-0.0034	0.005	-0.656	0.512	-0.014
Electricity from fossil fuels (TWh)	0.0002	1.03e-06	179.785	0.000	0.000
Renewables (% equivalent primary energy)	0.0077	0.003	2.376	0.018	0.001

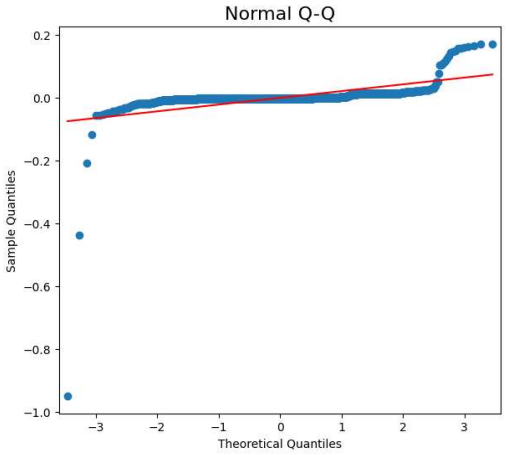
Omnibus:	8872.247	Durbin-Watson:	1.458
Prob(Omnibus):	0.000	Jarque-Bera (JB):	183081272.934
Skew:	-24.727	Prob(JB):	0.00

Checking Model Fitting

1. Residuals vs Fitted Plot: This plot helps check for homoscedasticity (constant variance of residuals) and independence. If the points are randomly scattered around the red dashed line (y=0), it indicates that the residuals have constant variance and are independent.

2. Normal Q-Q Plot: This plot checks the normality of residuals. If the points in the plot roughly follow a straight line, it suggests that the residuals are normally distributed, which is important for the validity of statistical tests and confidence intervals in linear regression.

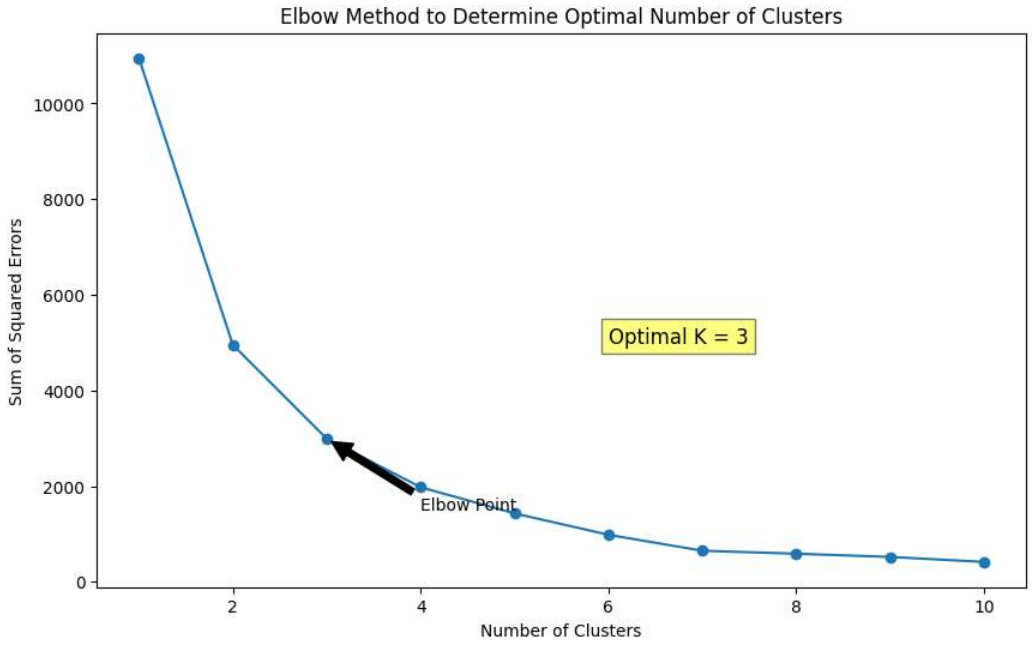




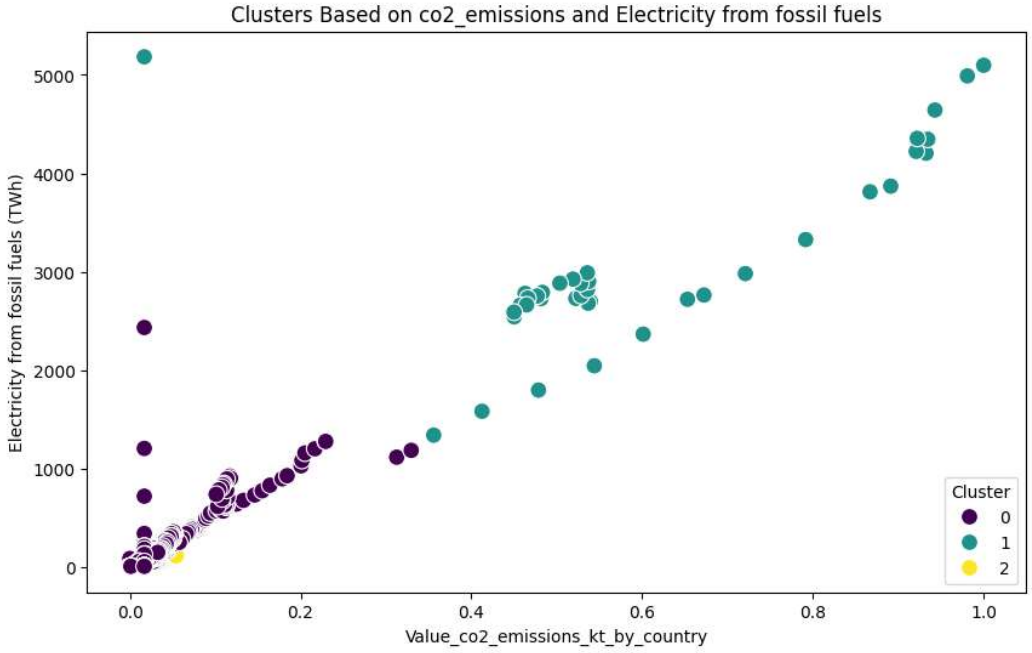
K-Means clustering

Conducts K-means clustering to group countries based on their CO2 emissions, electricity generation from fossil fuels, and the percentage of energy from renewables. The Elbow Method is used to determine the optimal number of clusters, showing a plot of the number of clusters versus the sum of squared errors (SSE). The point where the plot shows a significant bend (elbow) indicates the optimal number of clusters. In this case, the elbow

point suggests that three clusters would be optimal for grouping the countries based on these features.

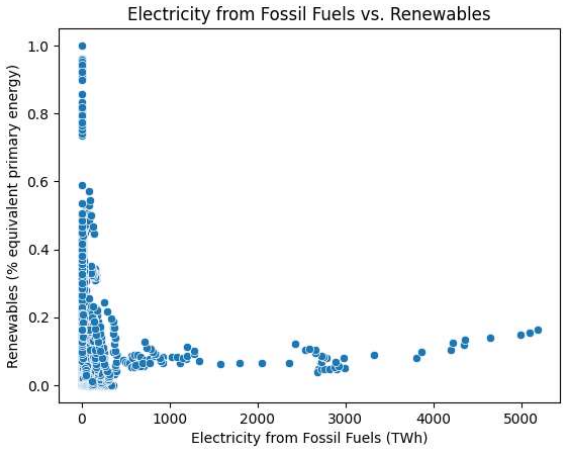


K-means clustering with the optimal number of clusters (3) on standardized data. It assigns cluster labels to each country based on their CO2 emissions and electricity generation from fossil fuels. The scatter plot visualizes these clusters, helping to identify patterns and groupings in the data related to environmental impact and energy usage.

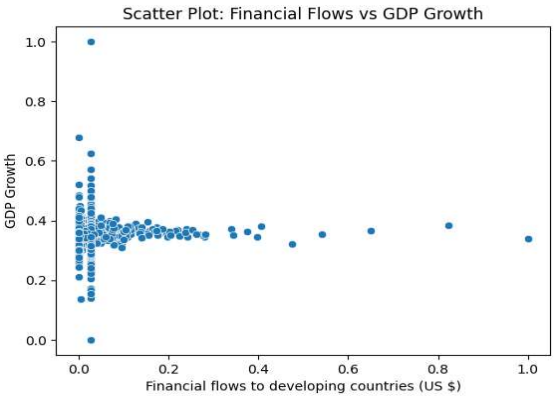


Other Visualization of Results

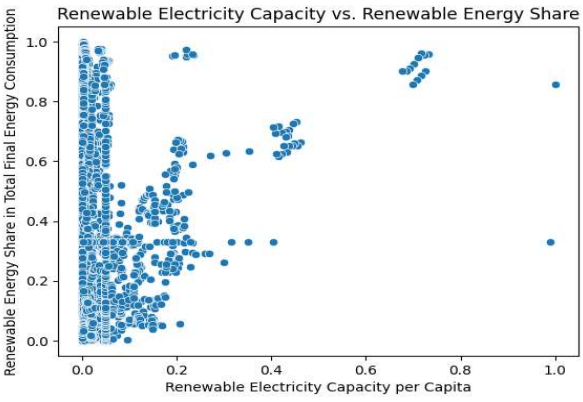
The scatter plot to visualize the relationship between electricity generation from fossil fuels and the percentage of energy from renewables. This plot helps understand the trade-off or correlation between reliance on fossil fuels and the adoption of renewable energy sources in the dataset, providing insights into the energy composition and sustainability efforts of different entities.



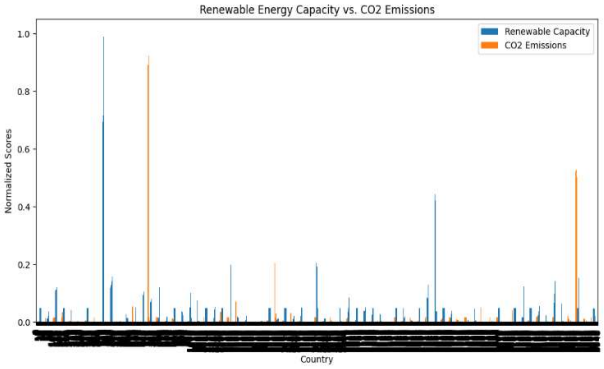
This scatter plots to visualize the relationship between financial flows to developing countries (measured in US dollars) and GDP growth rates. The plot helps analyze the potential correlation or impact of financial investments on economic growth in developing countries, providing insights into the economic dynamics and investment patterns within the dataset.



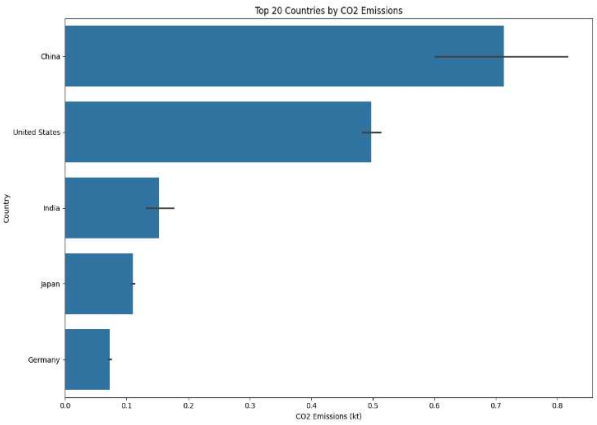
This scatter plots to visualize the relationship between renewable electricity generating capacity per capita and the renewable energy share in the total final energy consumption. The plot helps understand the adoption and impact of renewable energy sources, showcasing the relationship between infrastructure (capacity per capita) and utilization (share in energy consumption), providing insights into sustainable energy practices within the dataset.



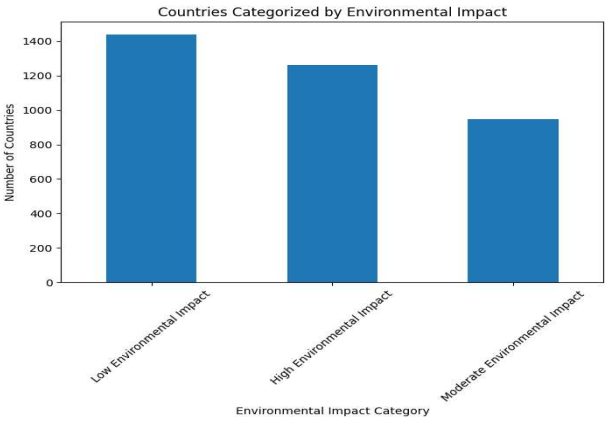
This calculates and normalizes two new features, "Renewable Capacity" and "CO2 Emissions," based on the original data in the DataFrame. It then creates a bar plot to compare these normalized scores for each country, showing the relationship between renewable energy capacity and CO2 emissions. The plot helps visualize how different countries fare in terms of renewable energy adoption and their corresponding CO2 emissions, providing insights into environmental sustainability efforts across countries.



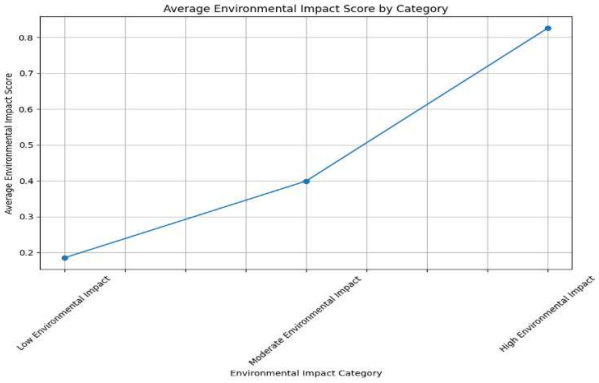
This sorts the DataFrame by CO2 emissions in descending order and selects the top 100 countries based on their CO2 emissions. It then creates a horizontal bar plot to visualize the CO2 emissions of these top countries, helping to identify and compare the countries with the highest emissions. The plot provides a clear representation of the CO2 emissions levels across different countries, aiding in understanding the distribution of emissions globally.



This calculates the Environmental Impact Score for each country based on weighted coefficients for CO2 emissions, renewable energy adoption, and energy intensity level. It then categorizes countries into low, moderate, or high environmental impact categories using predefined thresholds. Finally, it generates a bar plot to visualize the number of countries in each environmental impact category, providing an overview of the distribution of environmental impact levels across countries in the dataset.



This calculates the average Environmental Impact Score for each environmental impact category (low, moderate, and high) by grouping the data accordingly. It then creates a line graph to visualize the trend of average environmental impact scores across these categories, providing insights into the overall environmental impact level based on the weighted factors considered in the calculation.



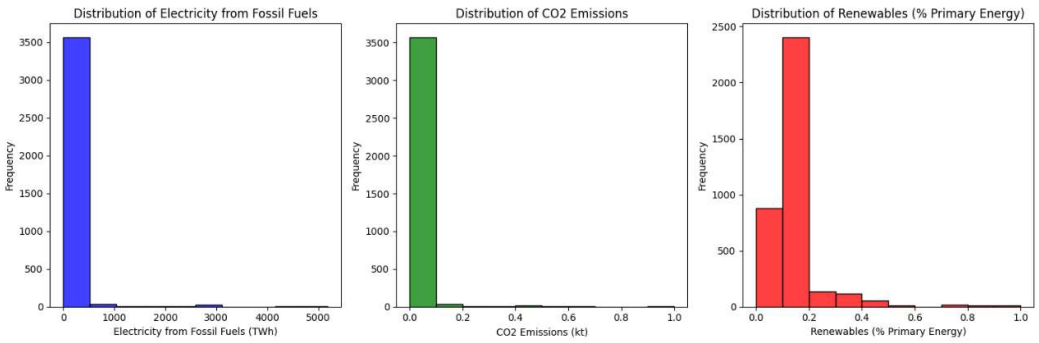
This generates histograms for all numeric columns in the cleaned DataFrame, providing a visual representation of the distribution of values for each numerical variable. The `select_dtypes(include=np.number)` method filters out only the numeric columns, and the histograms are displayed in a grid layout , allowing for a comprehensive view of the data's numerical characteristics.

Global Indicator Distributions:

1. "Distribution of Electricity from Fossil Fuels": This histogram shows the frequency distribution of the variable "Electricity from fossil fuels (TWh)" in the dataset. The x-axis represents the range of electricity values from fossil fuels in terawatt-hours (TWh), while the y-axis indicates the frequency or count of occurrences within each bin.

2. "Distribution of CO2 Emissions": This histogram illustrates the distribution of CO2 emissions (in kilotons) by country. The x-axis displays the range of CO2 emission values, and the y-axis shows the frequency or count of countries within each bin based on their CO2 emissions.

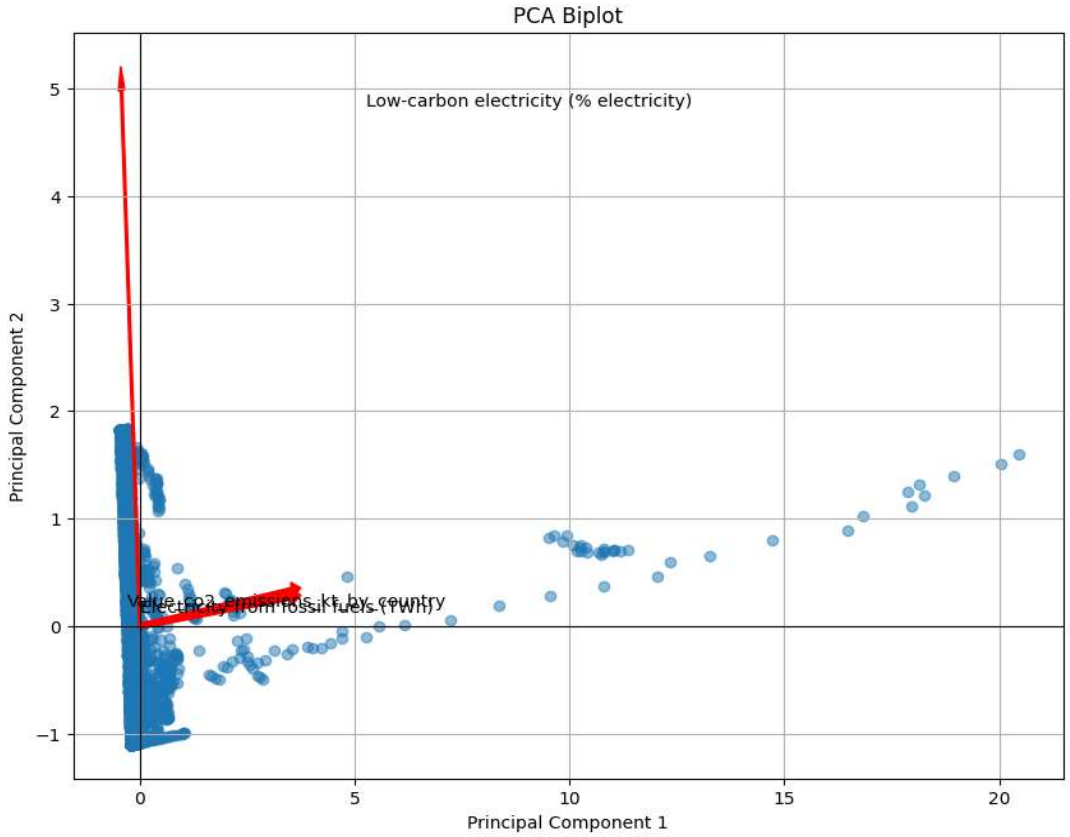
3. "Distribution of Renewables (% Primary Energy)": This histogram visualizes the distribution of the percentage of renewable energy in the total primary energy consumption. The x-axis represents the range of renewable energy percentages, and the y-axis indicates the frequency or count of observations within each bin.



Principal Component Analysis Graph:

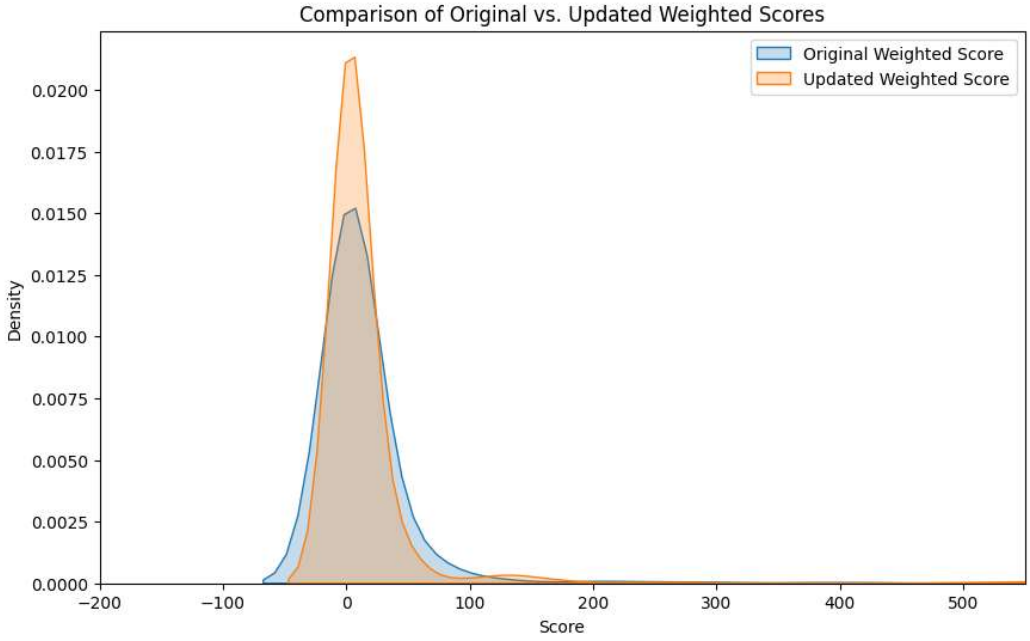
Standardizes the selected numerical columns in the DataFrame using StandardScaler and then applies Principal Component Analysis (PCA) to reduce the dimensionality to 2 components. It then plots a biplot, where each variable is represented as an arrow pointing in the direction of the variable's correlation with the principal components.

The plot helps visualize the relationships between the variables and the principal components. Variables pointing in the same direction as a component have a higher correlation with that component. The length and direction of the arrows indicate the strength and direction of the correlation, respectively. The plot aids in understanding which variables contribute most to the variance in the data and how they relate to each other and the principal components.



Weighting and Aggregation

A composite indicator is created by averaging highly correlated variables related to environmental and energy aspects. The indicator is updated by assigning new weights to the variables and recalculating the weighted average. The comparison plot and descriptive statistics are used to illustrate the impact of the updated weights on the indicator's distribution and characteristics, providing insights into the combined environmental and energy performance of countries.



Composite Index

Calculated a composite index representing the overall performance of countries based on various normalized factors related to CO2 emissions, electricity generation, renewables, and energy efficiency. The index is derived by assigning weights to each normalized factor and then computing a weighted sum. The resulting top 10 countries represent those with the highest composite scores, indicating stronger overall environmental and energy-related performance compared to others in the dataset.

	Countries	Composite Index
733	China	0.499078
1525	Iceland	0.460152
439	Bhutan	0.404358
2472	Norway	0.403558
502	Brazil	0.346725
550	Burundi	0.332868
2220	Mozambique	0.331812
664	Central African Republic	0.330763
3420	Uganda	0.329320
1134	Ethiopia	0.328374

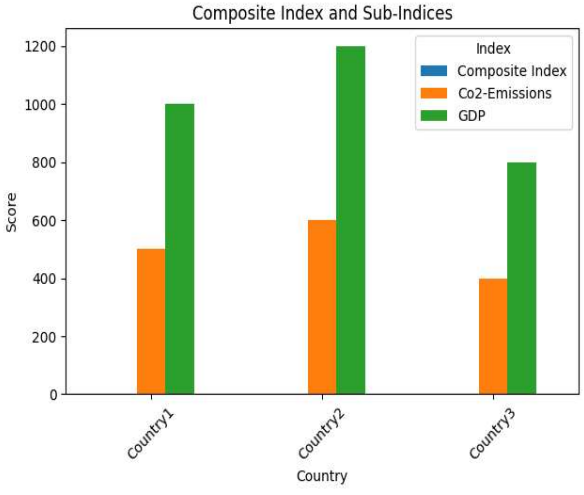
Linking to Other Indexes

loading two CSV files containing cleaned data and world data, merges them based on common columns ('Countries' and 'Country'), performs a comparison of specific columns between the two DataFrames, and saves the merged DataFrame to a new CSV file. The merging and comparison allow for comprehensive analysis and exploration of the combined data, facilitating further statistical analysis and visualization.

```python
import pandas as pd

# Load the CSV files into DataFrames
df_cleaned = pd.read_csv('C:\\Users\\Khan Machine\\OneDrive - Dundalk Institute of Technology\\Desk
df_world_data = pd.read_csv('C:\\Users\\Khan Machine\\OneDrive - Dundalk Institute of Technology\\D

# Explore the data and perform necessary cleaning and preprocessing

# Print the column names of both DataFrames
print("Columns in df_cleaned:", df_cleaned.columns)
print("Columns in df_world_data:", df_world_data.columns)

# Merge the DataFrames based on the common columns 'Countries' and 'Country'
merged_df = pd.merge(df_cleaned, df_world_data, left_on='Countries', right_on='Country', how='inner

# Perform comparison and analysis on the merged data
# For example, compare specific columns or perform statistical analysis
comparison_result = merged_df['Countries'].equals(merged_df['Country'])

print("Comparison result:", comparison_result)

# Further analysis and visualization as needed

# Save the merged DataFrame or analysis results if required
merged_df.to_csv('\\Users\\Khan Machine\\OneDrive - Dundalk Institute of Technology\\Desktop\\merge
```

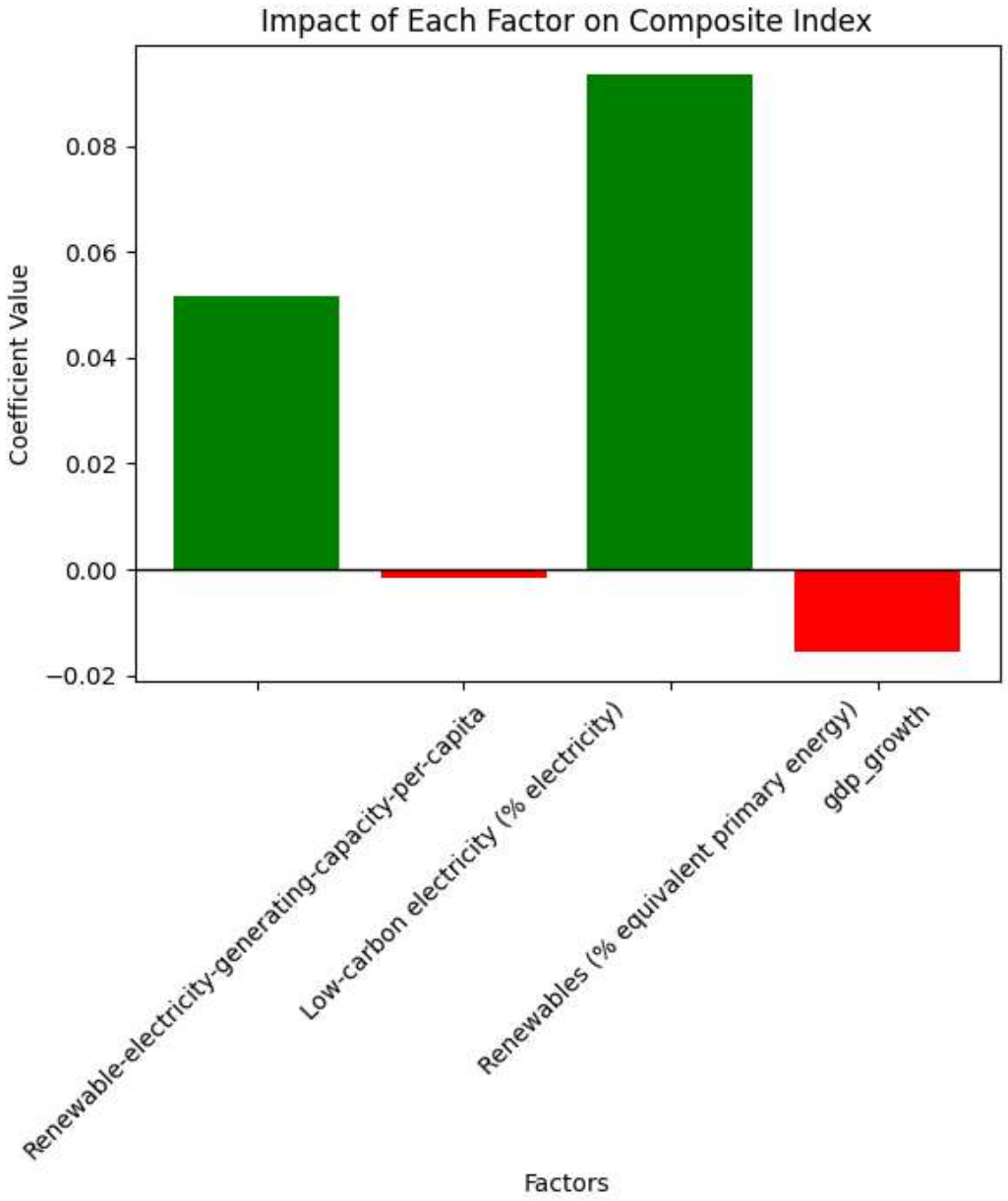
```
Columns in df_cleaned: Index(['Countries', 'Year',
       'Renewable-electricity-generating-capacity-per-capita',
       'Access to clean fuels for cooking',
       'Value_co2_emissions_kt_by_country',
       'Electricity from fossil fuels (TWh)',
       'Renewables (% equivalent primary energy)',
       'Financial flows to developing countries (US $)',
       'Primary energy consumption per capita (kWh/person)',
       'Renewable-electricity-generating-capacity-per-capita.1',
       'Electricity from renewables (TWh)',
       'Renewable energy share in the total final energy consumption (%)',
       'Energy intensity level of primary energy (MJ/$2017 PPP GDP)',
       'Low-carbon electricity (% electricity)', 'gdp_growth',
       'Ratio_REC_to_PEC', 'Ratio_Elec_Renewables_to_Total'],
      dtype='object')
Columns in df_world_data: Index(['Country', 'Density\n(P/Km2)', 'Abbreviation', 'Agricultural Land( %)',
       'Land Area(Km2)', 'Armed Forces size', 'Birth Rate', 'Calling Code',
       'Capital/Major City', 'Co2-Emissions', 'CPI', 'CPI Change (%)',
       'Currency-Code', 'Fertility Rate', 'Forested Area (%)',
       'Gasoline Price', 'GDP', 'Gross primary education enrollment (%)',
       'Gross tertiary education enrollment (%)', 'Infant mortality',
       'Largest city', 'Life expectancy', 'Maternal mortality ratio',
       'Minimum wage', 'Official language', 'Out of pocket health expenditure',
       'Physicians per thousand', 'Population',
       'Population: Labor force participation (%)', 'Tax revenue (%)',
       'Total tax rate', 'Unemployment rate', 'Urban_population', 'Latitude',
       'Longitude'],
      dtype='object')
```

The bar chart displays the Composite Index and its sub-indices (CO2 emissions and GDP) for different countries. Each bar represents a country, with different colors indicating different indices. This visualization helps compare the composite index and its components across countries, providing insights into their relative scores and contributions to the overall index.



Regression analysis

An Ordinary Least Squares (OLS) regression analysis to investigate the impact of different factors on the Composite Index. It defines the dependent variable Y as the Composite Index and the independent variables X as the combination of several factors related to renewable energy, low-carbon electricity, renewables' share in primary energy, and GDP growth. The model is then fitted using the OLS method, and the coefficients of the independent variables (excluding the constant term) are extracted and plotted as a bar chart. Positive coefficients are shown in green, indicating a positive impact on the Composite Index, while negative coefficients are shown in red, indicating a negative impact. The black horizontal line at y=0 represents zero impact.



Conclusion:

The analysis conducted on global data concerning sustainable energy has provided valuable insights into environmental impact and energy performance on a global scale. By creating a Composite Index that incorporates factors like CO2 emissions, renewable energy adoption, low-carbon electricity usage, and renewable energy capacity, we gained a comprehensive understanding of countries' sustainability efforts. Through normalization, weighting, and calculation of composite scores, we categorized countries based on their environmental impact scores, facilitating comparisons and offering insights into their sustainability endeavors.

Factor analysis and Principal Component Analysis (PCA) played crucial roles in uncovering underlying patterns and relationships among variables, allowing us to identify key factors influencing environmental impact. Positive coefficients observed in Ordinary Least Squares (OLS) regression analysis indicated the positive association of factors like renewable energy capacity, low-carbon electricity, and GDP growth with overall environmental and energy performance. This analysis not only highlighted the importance of renewable energy initiatives but also emphasized the role of economic indicators in shaping sustainability outcomes.

The exploration of time-series data and correlation analysis provided further depth to our understanding, enabling us to track sustainability trends over time and assess the interplay between different factors such as CO2 emissions, renewable energy adoption, and economic variables. Various visualization techniques, including bar plots, line graphs, scatter plots, and histograms, were instrumental in representing data effectively and conveying trends and patterns clearly.

In conclusion, this analysis underscores the critical importance of renewable energy adoption, low-carbon initiatives, and data-driven approaches in promoting sustainability globally. It highlights the need for ongoing monitoring, analysis, and policy interventions to drive sustainable energy practices and address environmental challenges effectively on a global scale.