CSE 587
Spring 2024

DIC Project Phase 1

Group Members:

Andrew Balotin

Deep Shahane

Bhushan Mahajan

1. Problem Statement

a. Background:
Gross Domestic Product (GDP) is an economic index that shows an economy's health by measuring its total output of goods and services. It indicates an important part in measuring the growth rate and overall economic performance. The Kaggle dataset, offering extensive data on global energy consumption, can be instrumental in predicting GDP fluctuations and energy use trends. Understanding these interconnections is key to optimizing resource allocation, crafting effective energy policies, and developing informed economic strategies.

b. Contribution Potential:
This study aims to predict GDP using machine learning techniques, which uses a dataset that contains different energy production, consumption, and population. The goal is to enhance economic forecasting by creating a robust model that is capable of simulating the impact of energy consumption changes on GDP. This is especially relevant for strategic planning and predicting economic shifts, given the increasing importance of energy management amidst the transition to sustainable energy and global economic shifts. The insights gained through the dataset could help us in optimizing economic performance based on energy production, consumption, and population patterns, and inform the development of energy policies.

2. Data Sources
The main data source for the study on estimating GDP using different metrics is an extensive Kaggle dataset named "World Energy Consumption." The wide range of characteristics in this dataset about energy production, consumption, and other socioeconomic elements that could affect a nation's GDP made it a desirable choice. With more than 20000 records in the dataset, there is enough information for a thorough study. Additional data sources, such as the World Bank for economic indicators and the International Energy Agency for comprehensive energy statistics, might be considered to enhance the dataset and increase the model's accuracy. In the final report, every source will be correctly cited and linked following the research standards.

Dataset link:
https://www.kaggle.com/datasets/pralabhpoudel/world-energy-consumption

3. Data Cleaning/Processing

The dataset from Kaggle will undergo a thorough cleaning and processing procedure to ensure its suitability for predictive modeling. Initially, there were 22012 rows and 129 columns in the dataset. We performed the following steps to clean the data:

1. Finding Null Values: We discovered that our dataset contained null values. Numerous null values were discovered, with some counts over 16,000.

2. Filling missing values: We used interpolation to fill null values in 22 columns and interpolated the population data based on country. Furthermore, as all cases with missing continent data were recognized as African countries, we imputed the null values in the "continent" column for African countries.

3. Deleting Rows and Columns: We eliminated the area without an ISO code and eliminated rows in the GDP column that had null values. Rows about Antarctica that had irreversible null values were also removed. Additionally, the national codes found in the ISO column were eliminated. Moreover, columns that displayed over 10,000 null values were eliminated.

4. Outlier Detection: The Z-score approach was utilized to identify any outliers. Only values with scores within three standard deviations of the mean were included in the new dataset.

5. Data Type Conversion: To enable precise analysis, we made sure that every column has the appropriate data type (numeric, categorical, etc.) allocated to it.

6. Float Precision: Everything else was rounded to three decimal places, except for the columns labeled "year" and "country".

7. Removing special characters: To remove special characters from the string type "country" column, we used regular expressions (regex).

8. Added column: Continent column which contains every continent like Asia, North America, South America, etc

9. Data selection: We narrowed down our dataset to focus on the continents of North and South America, which contains 2,170 rows and 28 countries.

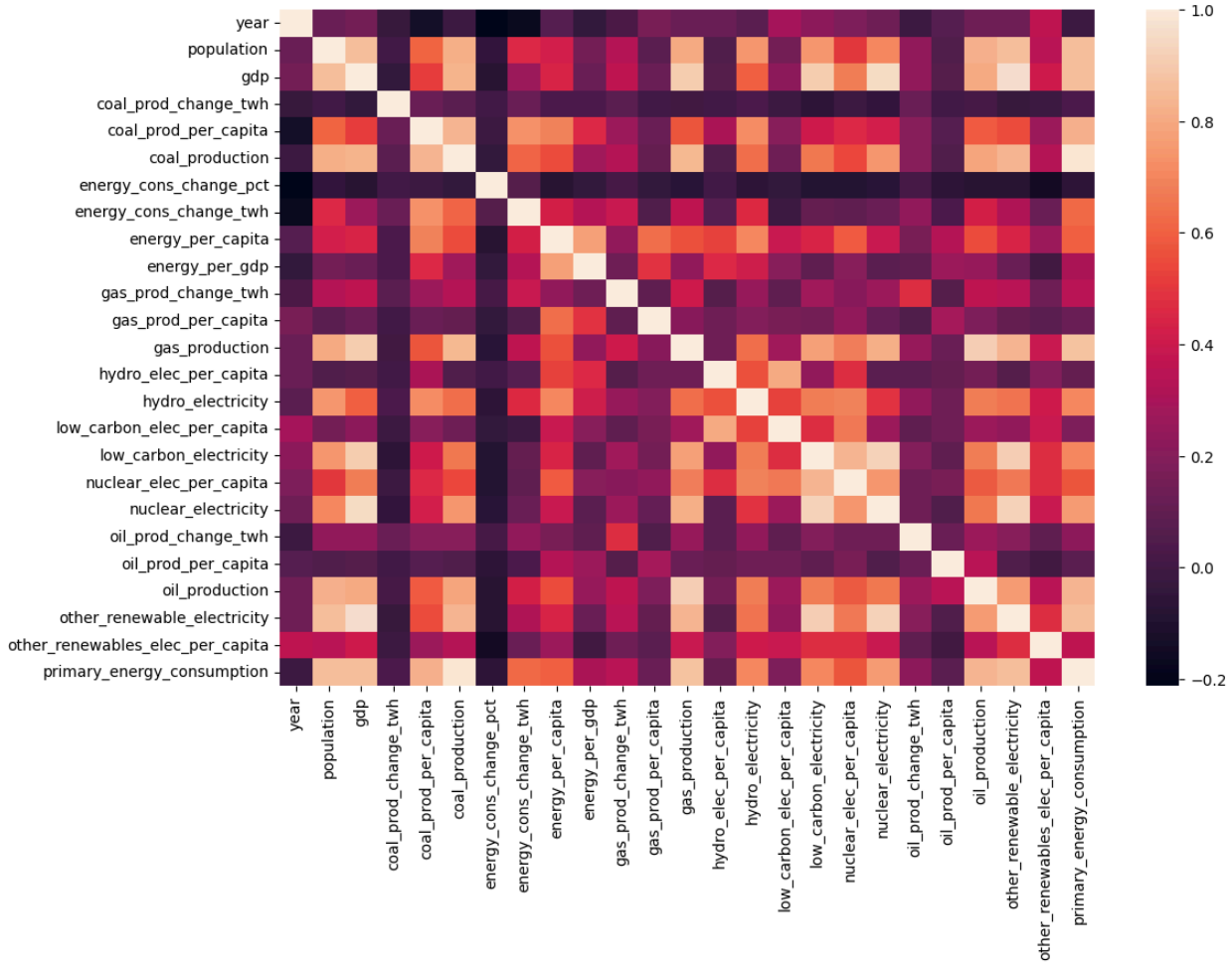
10. Standardization: For the standardization part, we used the Standard Scaler approach, which scales the features to guarantee that the sum of the standard deviations equals one and this approach rescales the data to a mean of zero.

11. Encoding Categorical Variables:
    Using encoding techniques, we converted categorical information into numerical values, such as 1s and 0s.
12. Data Splitting:
    We separated the dataset into subsets for testing and training while keeping the 80:20 ratio.

4. Exploratory Data Analysis (EDA)

An important first part in comprehending the relationships and underlying structure of the dataset is exploratory data analysis or EDA. The following EDA activities will be carried out for this project:

1. Heatmap:



The image shows a heatmap visualization of the correlation matrix for various energy-related variables, such as production and consumption metrics across different energy sources. Lighter colors represent stronger correlations between the variables. We can find a correlation between GDP with other columns.

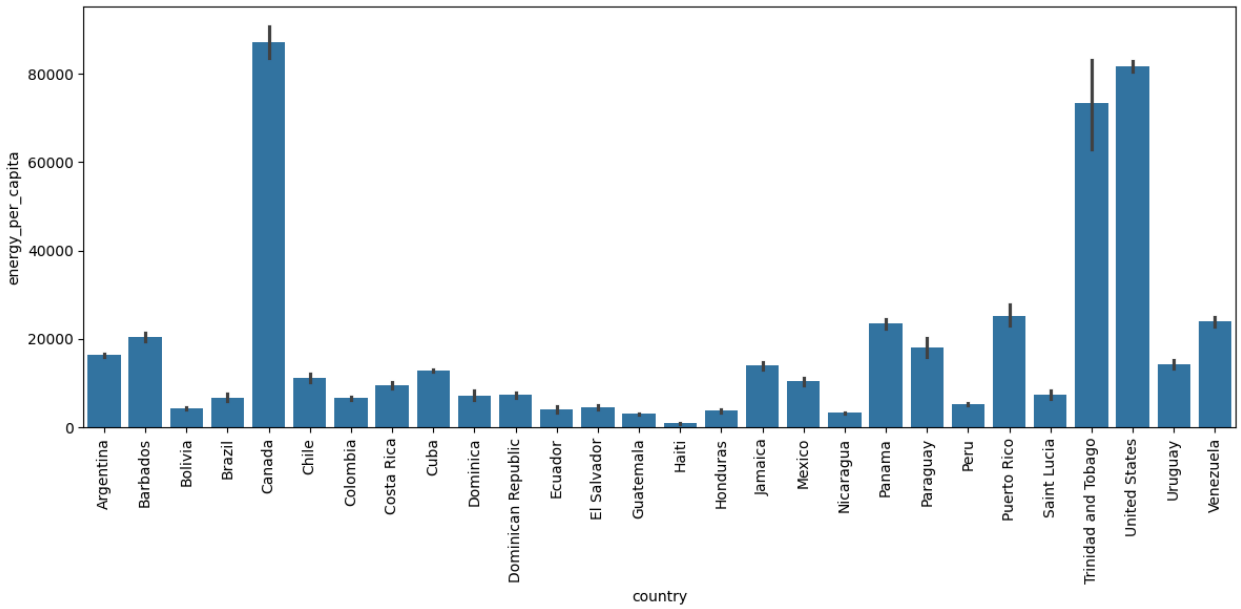
2. Summary Statistics:

```
cleaned_csv_specific_continents.describe()
```

	year	population	gdp	coal_prod_change_twh	coal_prod_per_capita	coal_production	energy_cons_change_pct	energy_cons_change_twh	energy_per_capita	energy_pe
count	2170.000000	2.170000e+03	2.170000e+03	2170.000000	2170.000000	2170.000000	2170.000000	2170.00000	2170.000000	2170.0
mean	1970.445622	2.533447e+07	4.550474e+11	1.707614	2345.474274	259.336830	4.576671	41.04332	20449.563983	1.6
std	33.957368	5.052151e+07	1.773466e+12	89.173953	6556.804464	1027.135147	8.263905	165.49628	27636.466988	1.0
min	1900.000000	6.817100e+04	1.608543e+08	-1126.270000	-13.003000	-0.509000	-36.288000	-857.41400	382.382000	0.3
25%	1943.000000	3.417788e+06	1.315517e+10	0.000000	0.000000	0.000000	0.948000	0.33250	4666.116000	0.9
50%	1980.000000	7.362767e+06	3.737957e+10	0.000000	0.000000	0.000000	4.292500	3.22100	9855.295000	1.2
75%	1999.000000	1.947408e+07	1.397927e+11	0.033000	335.521000	8.808750	7.245000	12.26100	19747.168000	1.8
max	2018.000000	3.321400e+08	1.819372e+13	1170.106000	45648.847000	6704.618000	75.146000	1059.16400	166508.463000	6.3

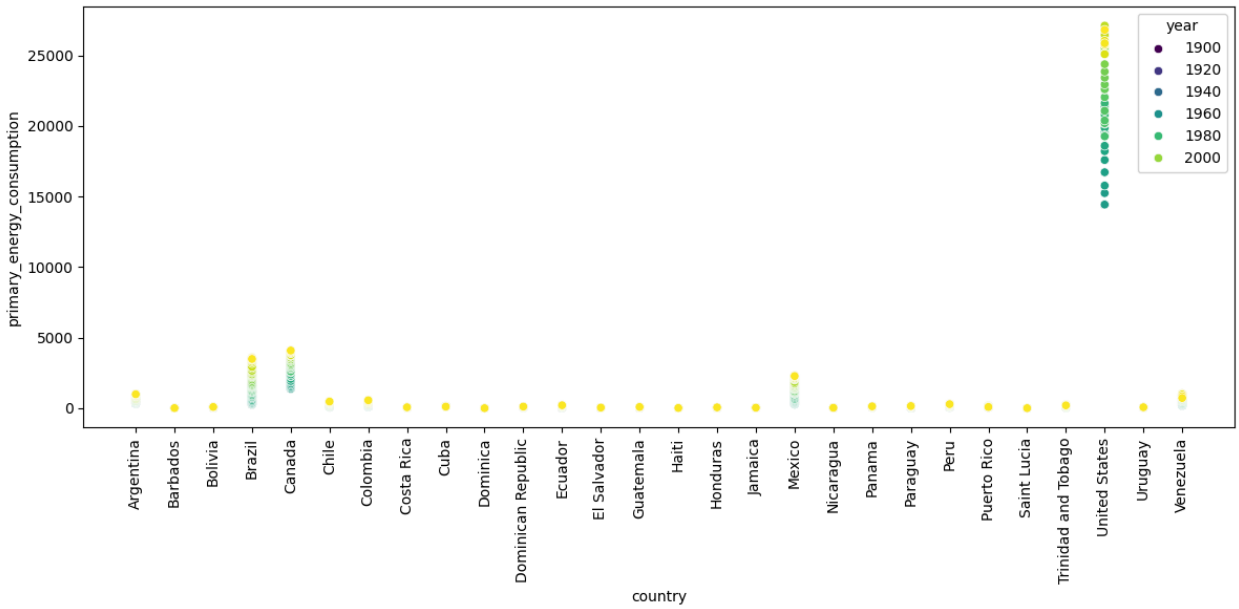
The numerical variables' mean, median, mode, standard deviation, (SD), quartiles, and range are calculated. This helps a rapid interpretation of the data's distributions and primary patterns.

3. Bar Graph:



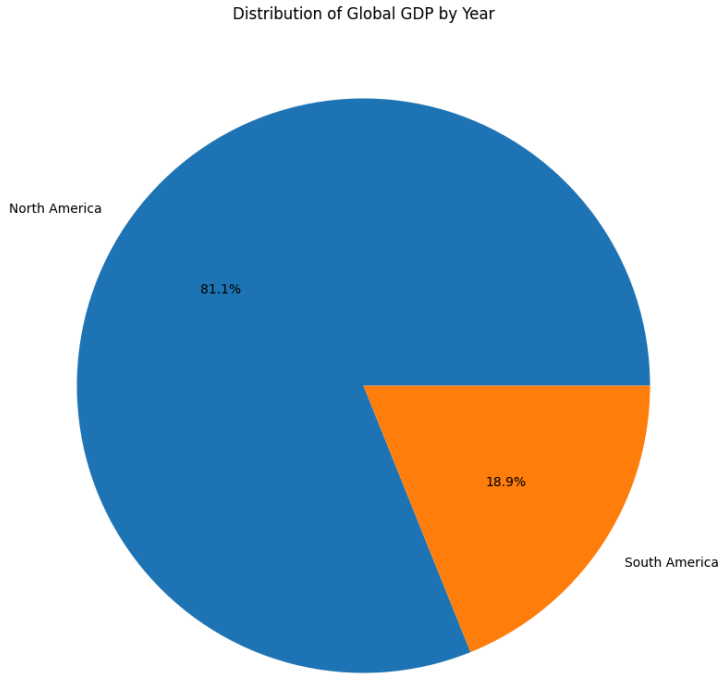
The bar graph displays notable differences between countries when comparing the amount of energy used per capita in each. In comparison to other countries, Canada, Trinidad and Tobago, and the United States exhibit a much higher per capita energy consumption, suggesting a larger per capita energy demand in these regions.

4. Scatter Plot:



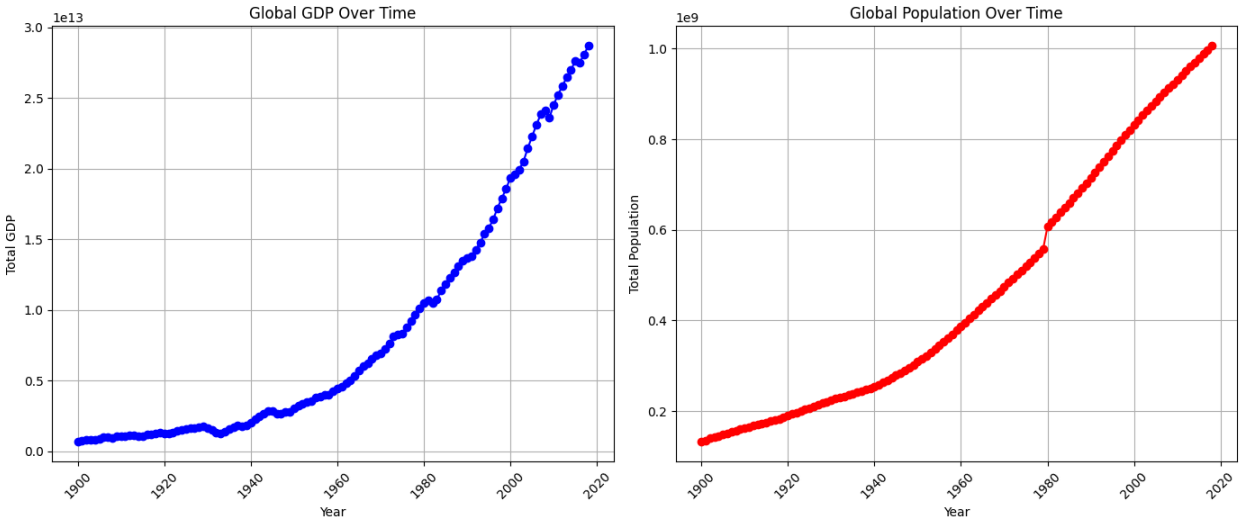
The main energy consumption of different countries over a range of years is displayed as a scatter plot, where each point represents the consumption for that particular year. While other nations display lower and more dispersed patterns of energy consumption, the United States, Canada, and Trinidad and Tobago exhibit a dense concentration of points at higher consumption levels, indicating substantial energy use over time. Dark blue denotes the most recent year, whereas yellow represents the most recent year.

5. Pie Chart:



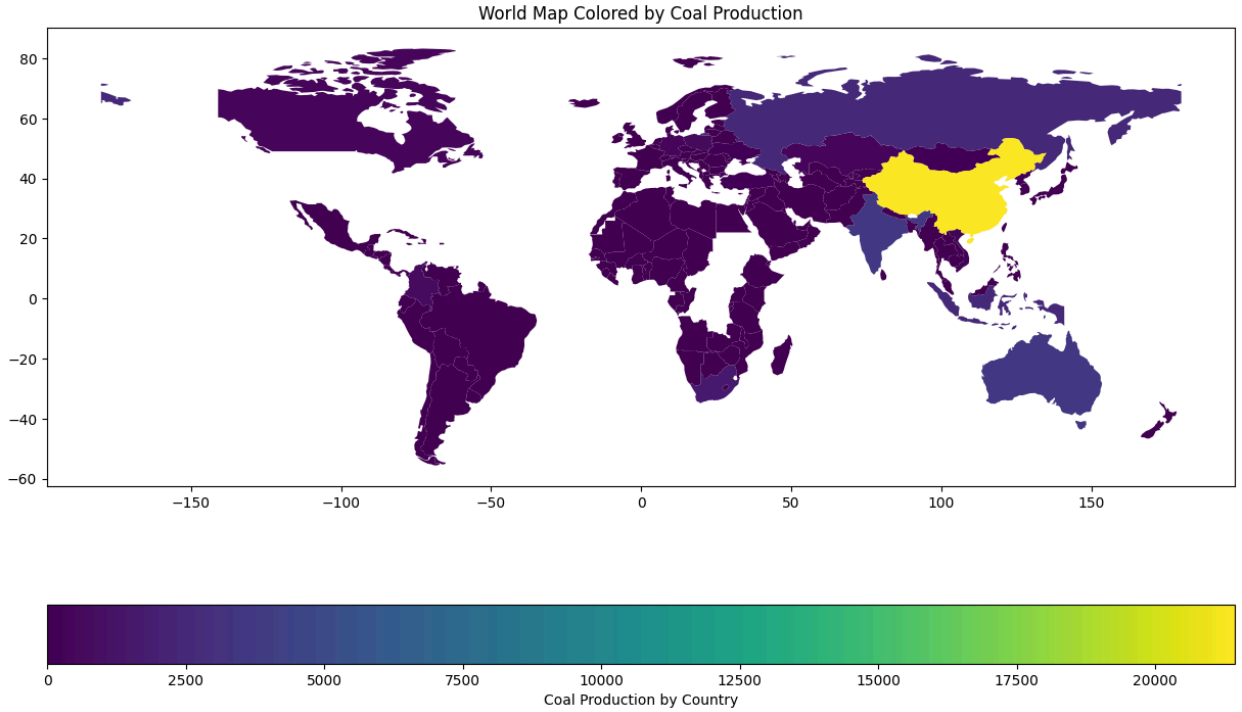
The pie chart depicts the distribution of Global GDP between North America and South America, showing that North America contributes the larger share with 81.1%, while South America accounts for 18.9%.

6. Line Graph:



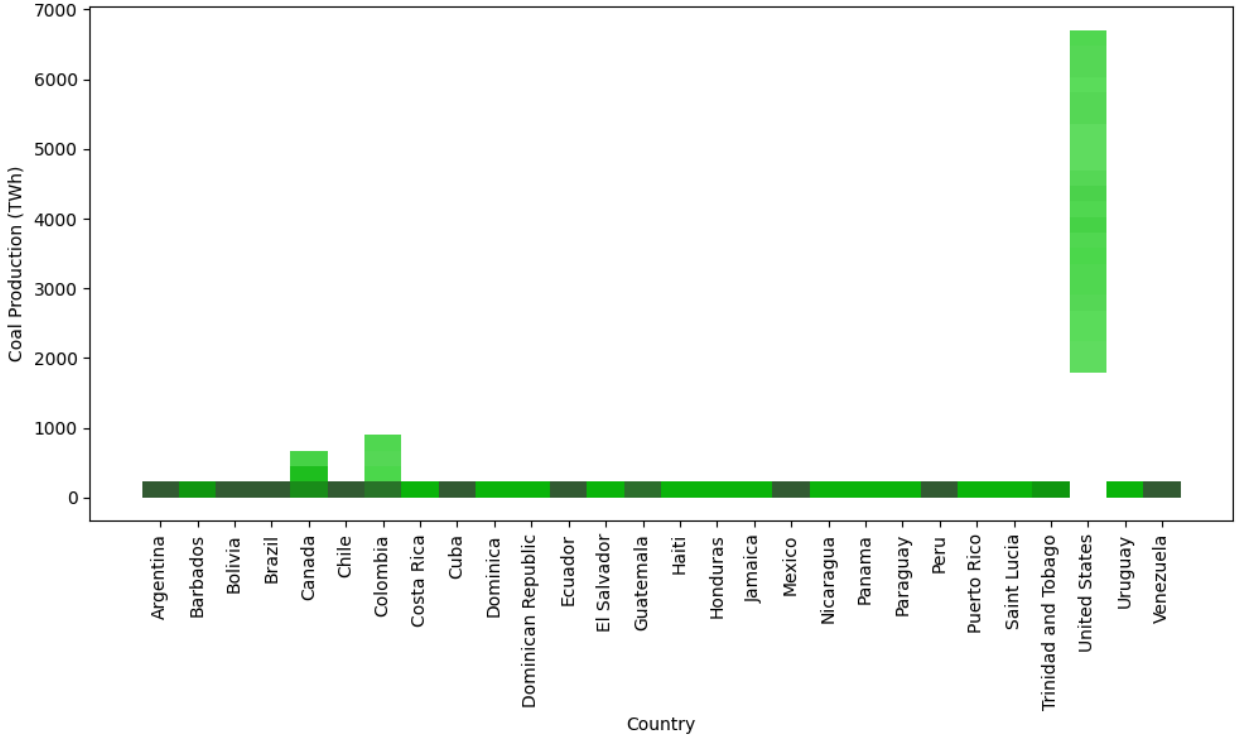
The line graphs display the exponential growth of Global GDP and population over time. The left graph shows a steep increase in Global GDP, after the 1950s. The right graph shows a similar exponential rise in the global population.

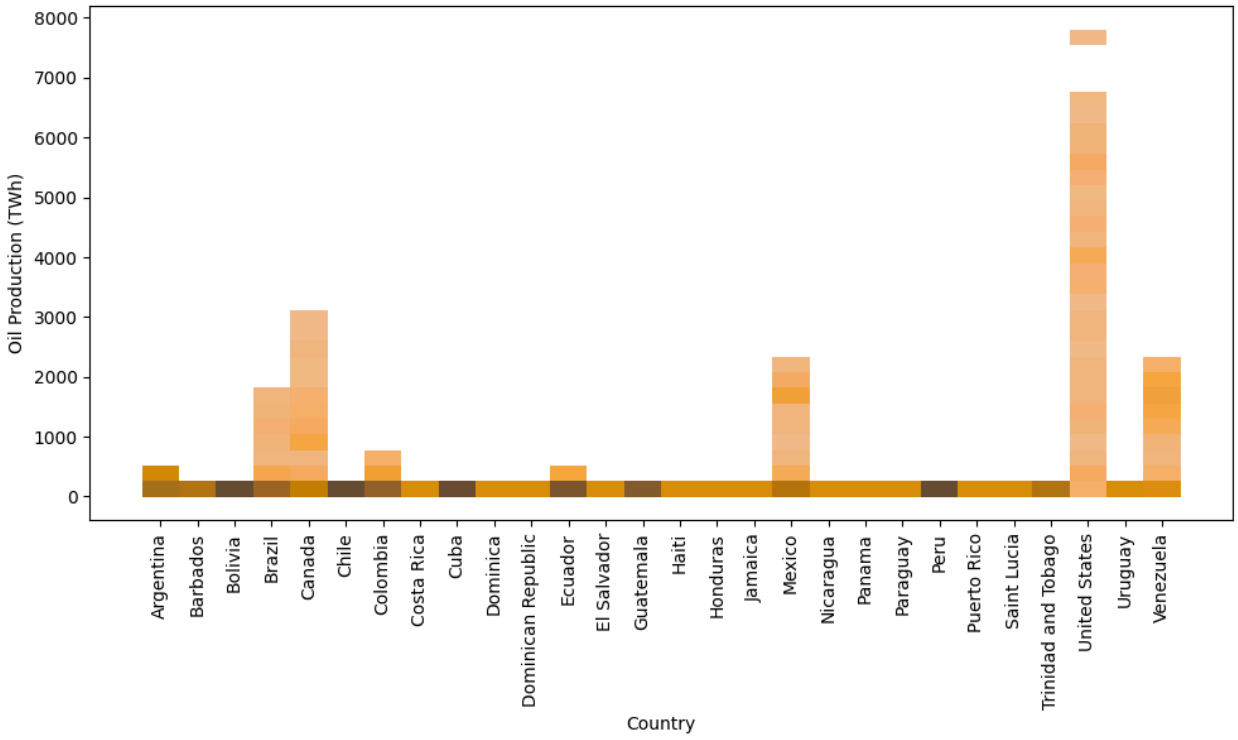
7. GeoSpatial Analysis:



The world map is color-coded to represent coal production by country. The country highlighted in yellow signifies the highest coal producer on the map, while other countries are shown in varying shades of purple to indicate lower levels of production.

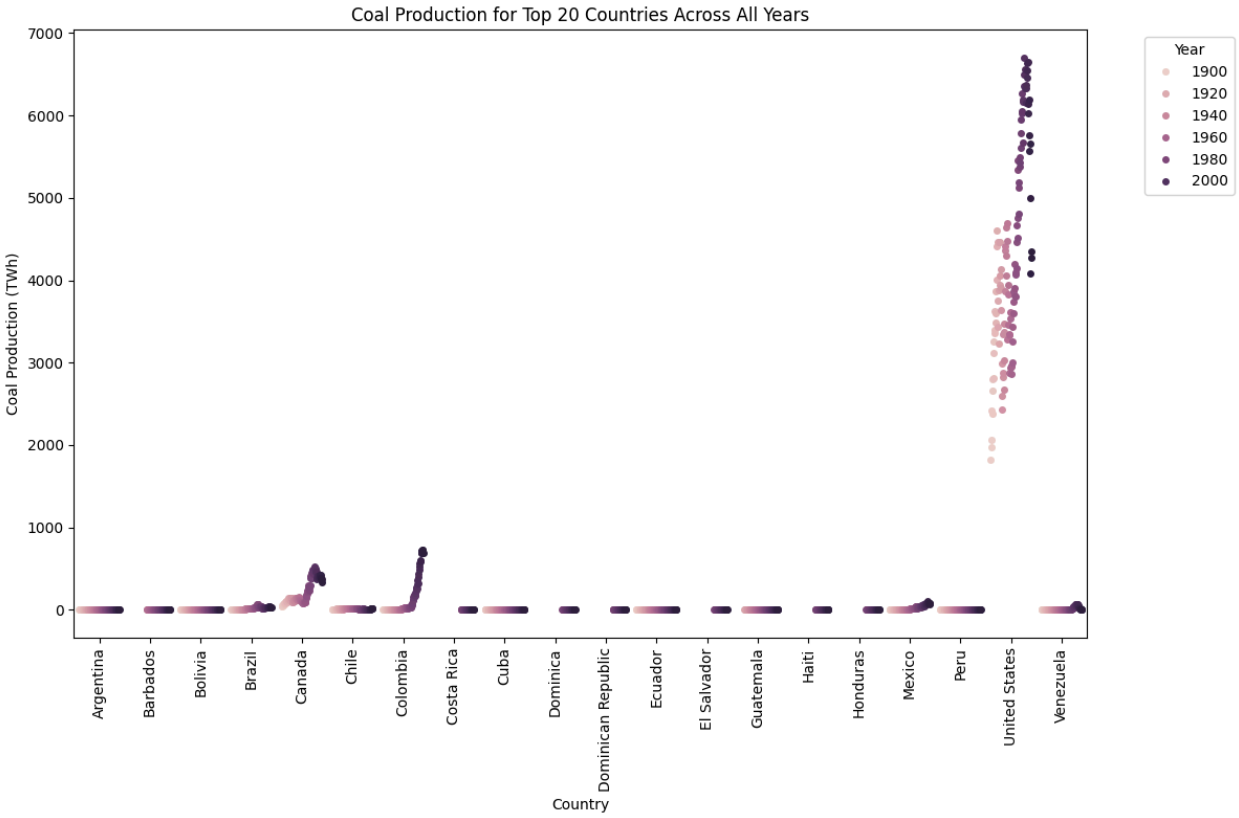
8. Histogram:





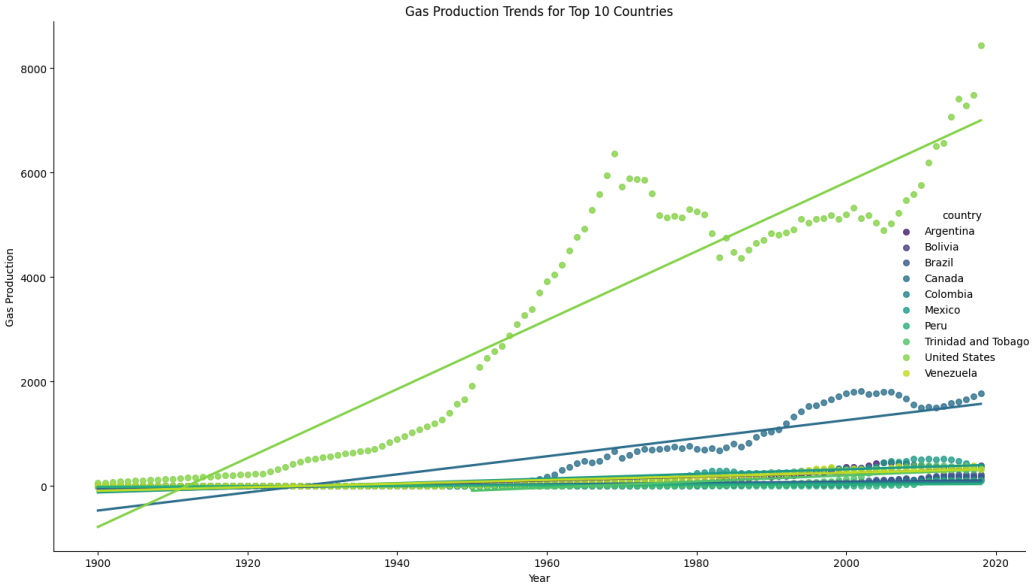
The histogram compares coal and oil production across various countries. The top chart, in green, shows coal production, with one country far exceeding the others. The bottom chart, in orange, represents oil production, where a few countries show moderately high production levels, and one stands out with exceptionally high oil production.

9. Swarm Plot:



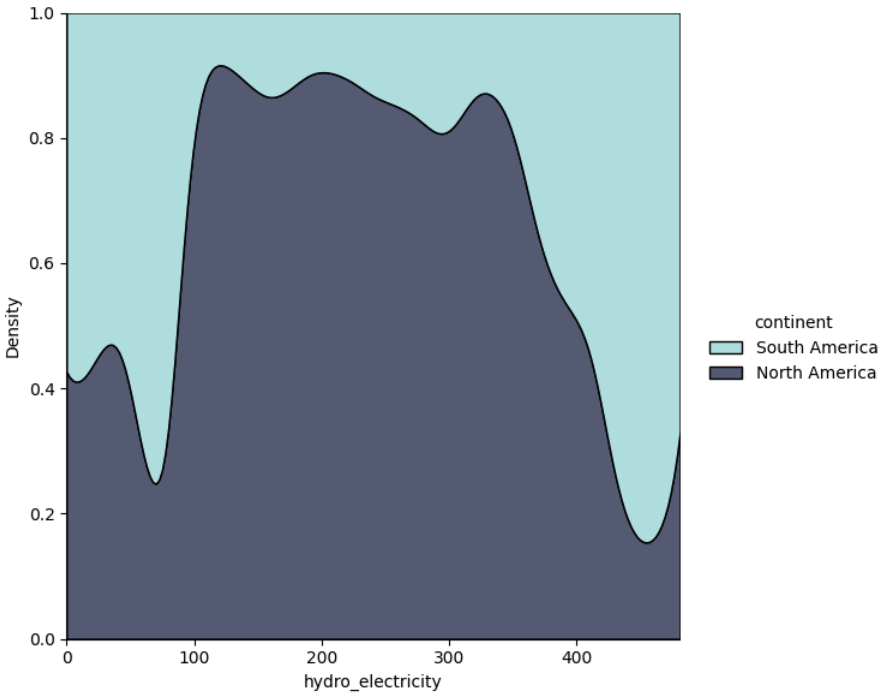
The swarm plot visualizes coal production for the top 20 countries across various years, with each dot representing a specific year's production. The United States stands out with a dense cluster of points at higher levels of production, showing its dominant role in coal output over time.

10. LM Plot :



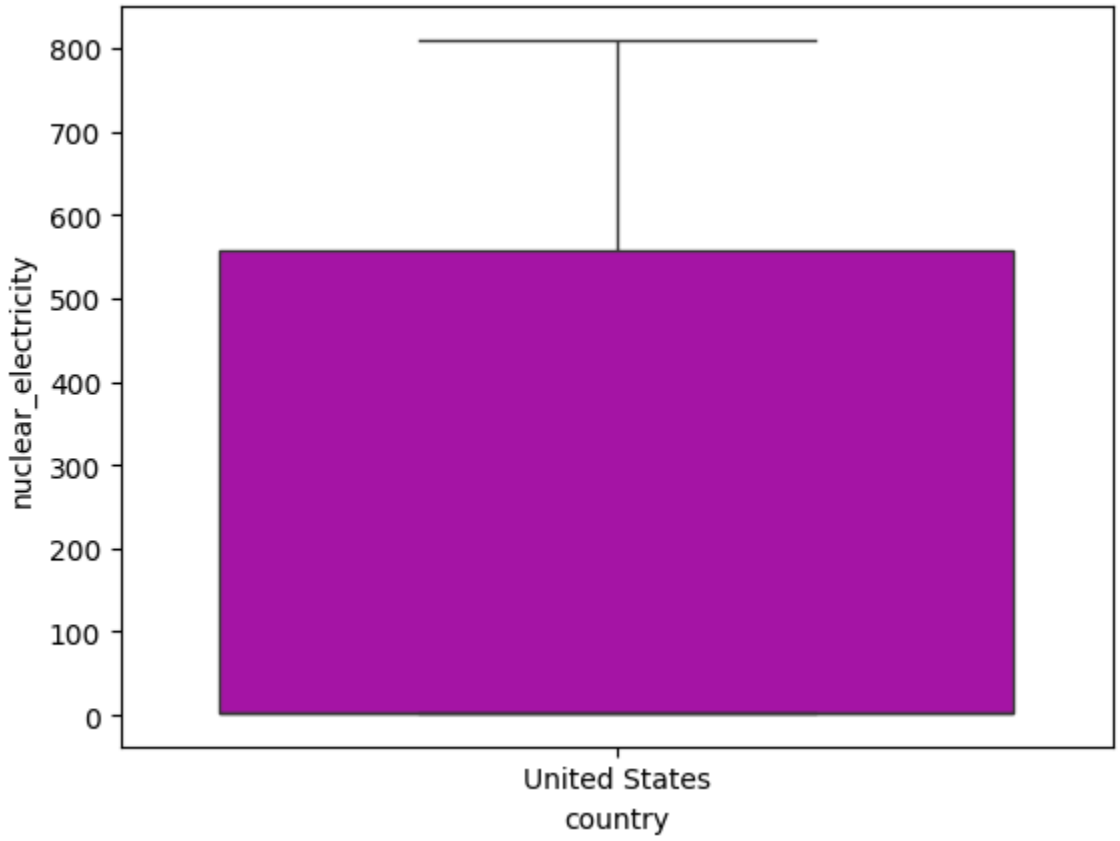
The LM Plot showcases the historical gas production trends for the top 10 countries. The graph indicates that several countries have steadily increased their gas production. Line represents the regression line which indicates the trend.

11. Density Plot:



The density plot illustrates the distribution of hydroelectricity production between South America and North America. The overlapping areas show the common ranges of hydroelectric power generation, with the peaks indicating the most common values of hydroelectricity production in each continent.

12. Box Plot:



The box plot represents the distribution of nuclear electricity production in the United States. showing where the middle 50% of data points lie, with the line inside the box denoting the median

Reference:

https://geopandas.org/en/stable/docs/reference/api/geopandas.GeoDataFrame.plot.html

https://seaborn.pydata.org/

https://pycountry-convert.readthedocs.io/en/latest/

https://github.com/owid/energy-data/blob/master/owid-energy-codebook.csv/