# Project Report

## GitHub URL

https://github.com/Viduthalaiselvi2106/UCDPA_ViduthalaiSelviGovindasamy

## Abstract

Global energy statistics can give us an insight into the relation between production and consumption of energy resources. Based on their relation, redistribution of the energy resources can be a solution to handle energy requirements across the globe. This project analyzes the global energy statistics data and proposes a way to predict the future production and consumption data.

## Introduction

We require different types of energy to facilitate our daily routine work. There are different sources of energy such as coal, gas, petroleum, hydroelectric power etc. This energy basically drives economies globally. Every country has different available natural sources of energy and thus, the production of energy using those available sources differs according to their availability.

Generation of energy from natural sources has a greater impact on global warming. Global energy production requires natural resources that emit carbon dioxide and various greenhouse gases. These emitted greenhouse gases will directly harm our ecosystem and it will be a threat to human life. To tackle this, every country has policies to handle production of energy such that it has minimal impact on global warming. But, in some countries, the consumption is nearly more than the energy being produced. So, there comes the need to redistribute natural resources across different countries.

This project aims to analyze and predict the relation between the consumption and production of energy in each continent. This relation will be helpful in redistributing the resources without having an impact on the global economy. The data analysis in this project has been restricted to continent level as the distribution of resources is feasible and cost effective at continent level.

## Dataset

Source of Dataset: https://www.kaggle.com/datasets/akhiljethwa/world-energy-statistics

1. **Production Total [1980-2021]**
    - This dataset has details about total production in different countries grouped into continents.
    - The production total has been recorded from 1980 to 2021.
    - This dataset is more trustable as it has been collected from **U.S. Energy Information Administration [2]**
2. **Consumption Total [1980-2021]**
    - This dataset has details about total consumption in different countries grouped into continents.
    - The total consumption in different countries has been recorded from 1980 to 2021.
    - This dataset has been collected from **U.S. Energy Information Administration [2]**

# Implementation Process

## 1.Importing

Importing required packages for the data analysis. The major packages used in this project are numpy, scipy, pandas, matplotlib, sqlite3 and seaborn. Pandas package was used for reading csv files containing Production and Consumption data.

### 1.1. Importing a CSV file into a table using sqlite3 tool

- To store in an SQL database, a new database 'GlobalEnergy.db' was created initially.
- Using 'CREATE TABLE' query, a table to store production data was created and populated using the data from csv file containing corresponding data.
- The stored values were retrieved using SELECT query.
- Later, these stored values were uploaded in a csv file with the required data.

## 2. Data Exploration

Initially, the data frame is used to explore the columns in the csv file. Using describe method in pandas, information such as mean, standard deviation, number of rows, minimum and maximum are found. All the information for the production table is as shown in table-1 below.

**Table-1: Information about production data**

| | Total_production |
|---|---|
| count | 229.000000 |
| mean | 79.280368 |
| std | 308.217753 |
| min | 0.000000 |
| 25% | 0.039002 |
| 50% | 2.272900 |
| 75% | 27.855034 |
| max | 3149.020415 |

An aggregate function is applied over each continent to find the minimum and maximum production values. Similarly, an aggregate function is applied on consumption data. The resultant table after applying aggregate function on production and consumption data are as shown in table-2 and table-3 below,

**Table 2 – Minimum and Maximum values on production data**

| Continent | min | max |
|---|---|---|
| Africa | 0.000000 | 238.118858 |
| Asia & Oceania | 0.000000 | 2620.639503 |
| Central & South America | 0.000000 | 294.823104 |
| Eurasia | 0.104255 | 1549.828759 |
| Europe | 0.000000 | 355.330260 |
| Middle East | 0.005926 | 905.009573 |
| North America | 0.000000 | 3149.020415 |

**Table 3 – Minimum and Maximum values on consumption data**

| Continent | min | max |
|---|---|---|
| Africa | 0.004458 | 192.697680 |
| Asia & Oceania | 0.000000 | 2953.320240 |
| Central & South America | 0.018534 | 351.277664 |
| Eurasia | 4.030892 | 894.408884 |
| Europe | 0.398382 | 437.349547 |
| Middle East | 1.122861 | 258.846097 |
| North America | 0.000000 | 3830.095107 |

## 2.1. Data Manipulation

The total consumption of resources was used to calculate the percentage across countries and stored in a new column as 'Total_Consumption_percent'. Indexing and slicing were also used in the data frame to retrieve relevant information for analysis.

## 3. Data Cleansing

Initially, the dataset is checked for any missing values using isnull and sum functions in pandas. But there seems to be some values in the dataset which are string values such as '--', 'in' etc. These strings values were of no use as we are interested in numerical values. The replace function is used to replace all such strings by 0. After replacing there is possibility of having duplicates in the dataset and these duplicates were removed using drop_duplicates function. In addition to cleaning the dataset, there is also a need to update missing values so that it won't affect the analysis. The missing values are replaced using fillna method.

## 4. Merging Dataset

This dataset has multiple files showing production and consumption of individual resources such as coal, petroleum etc. So, these different datasets are being merged based on continent and country names. Merge function is used to perform inner join over the dataset containing production of gas, coal, and petroleum. To remove duplicates after merging the datasets, the drop_duplicates method is being used.

## 5. Data Visualization

### 5.1 Analysing and Customizing data

The distribution of production and consumption data with reference to country name and continent name is visualized using sunburst plot. Similarly, a box plot and violin plot are used to visualize the data with respect to continent name. The dataset is also analyzed categorically using mean estimator and visualized using point plot. To visualize the relation between parameters in the dataset, a relational plot is used. The percentage of production across the world is being analyzed using scatter plot over percentage column created earlier.

Since the data analysis is between production and consumption of resources, the relation between them is being analyzed and visualized using line plot. But this analysis is at the continent level.

## 6.Machine Learning

Based on data analysis, it has been found that the production and consumption data have a high relation with its preceding year data. To predict the current year's production and consumption data, a linear regression model is being chosen. The linear regression model is applied on previous year data and then it is used to predict the present year data. Using the linear regression model, I used the production/consumption value in the year 2020 and predicted the values in the year 2021. This model is analyzed using Pearson Correlation Coefficient and it is found to be almost 1. Pearson Correlation Coefficient estimates the linear correlation between dependent and independent variables.

## Results

### Results of Data Visualization

Sunburst plot is used to show distribution of production and consumption data at country level as well as continent level. The sunburst plot is as shown in figure-1 and figure-2. Similarly, the distribution is also visualized using box plot, violin plot and point plot as shown in Figure-3, 4 and 5.

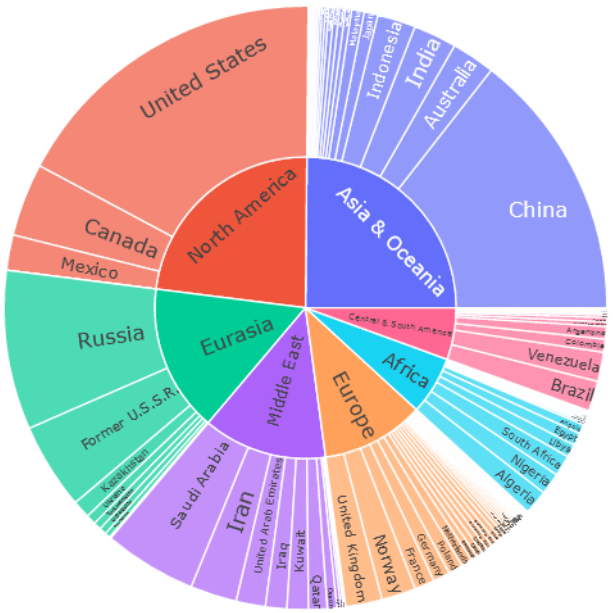


**Figure 1 – Distribution of Production Data using Sunburst**

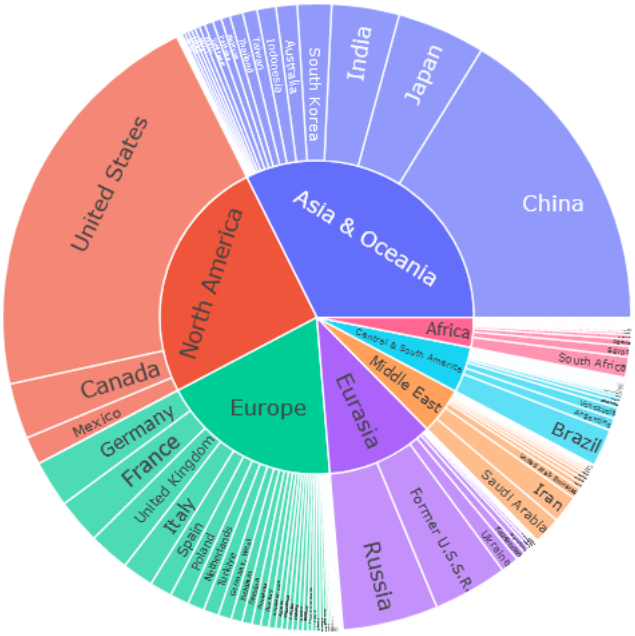


**Figure 2 – Distribution of Consumption Data using Sunburst**

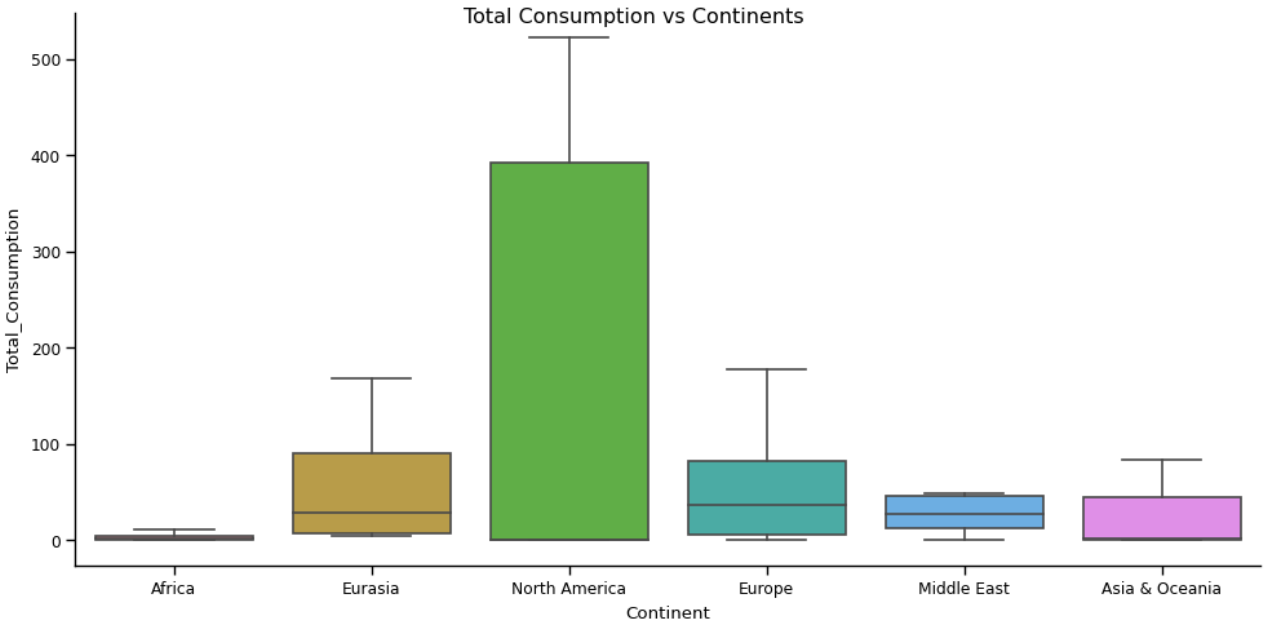


**Figure 3 - Distribution of Total Consumption using Box plot**

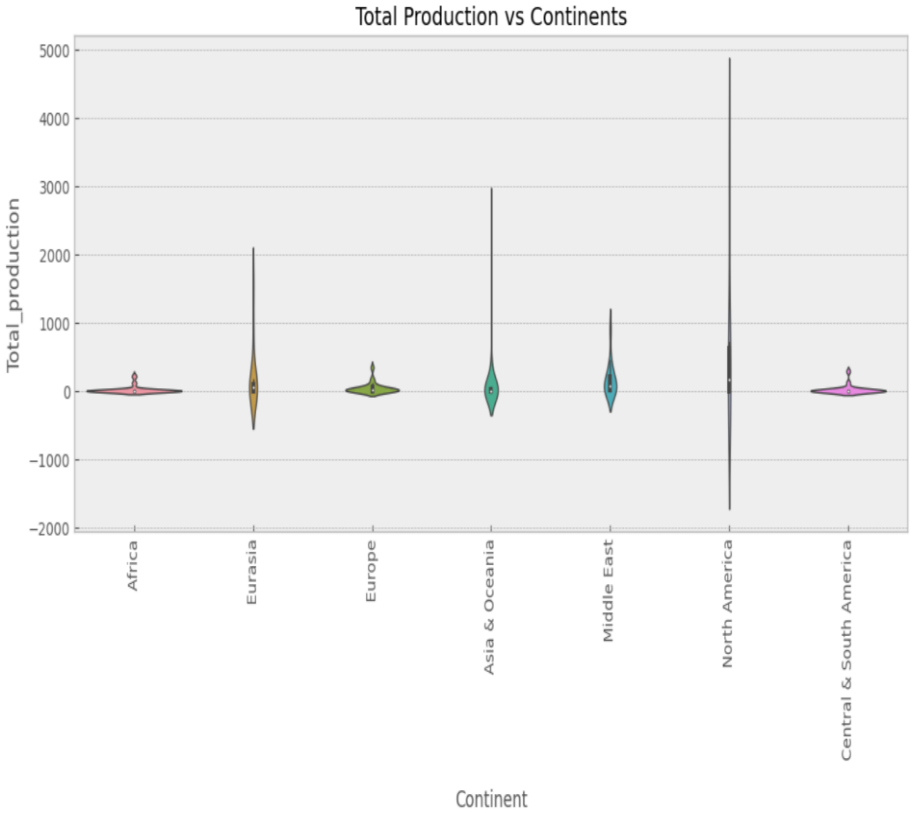


**Figure 4 – Violin plot showing distribution of total production over different continents**

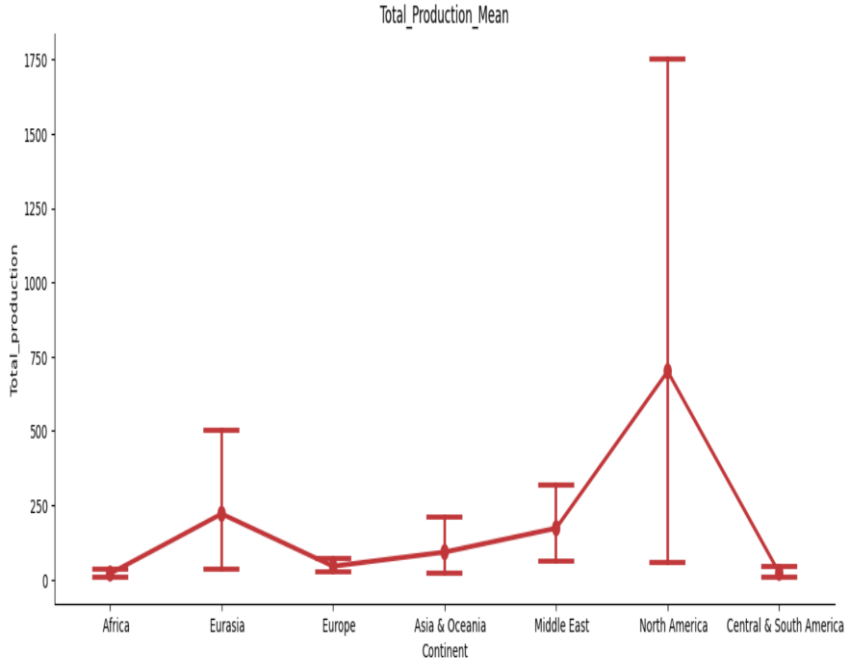


**Figure 5 – Categorical plot (point plot) showing mean total production across continents**

Relational plot is used to identify relation across years at individual resource level. This plot is as shown in figure-6, 7 and 8.

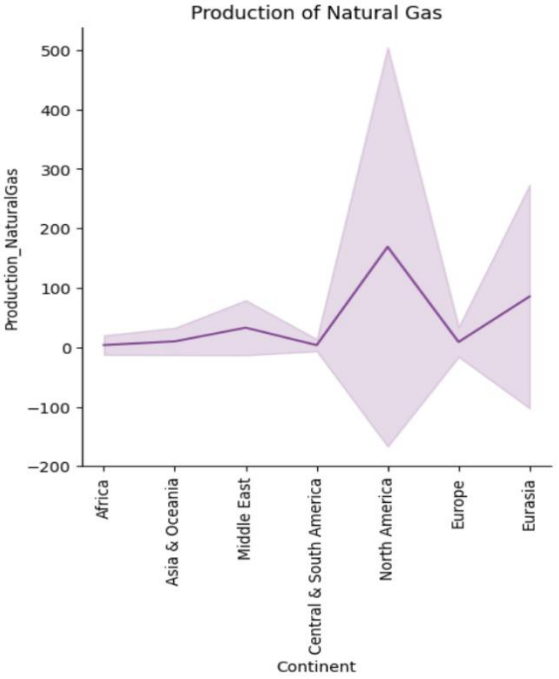


**Figure 6 – Relational plot showing production of natural gas and its relation over years**

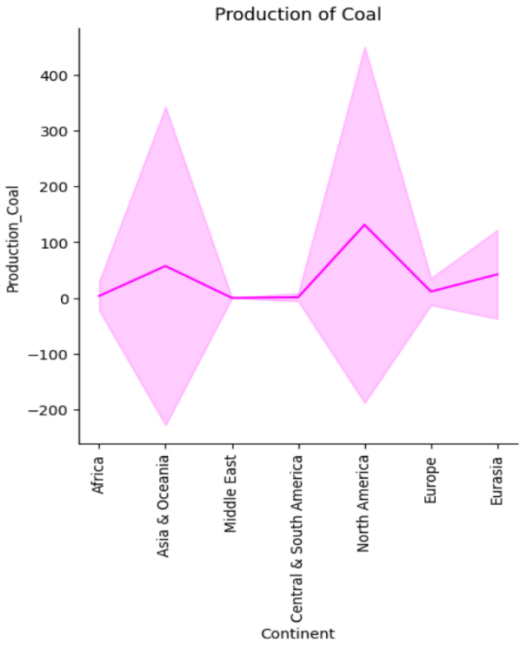


**Figure 7 – Relational plot showing production of Coal and its relation over years**

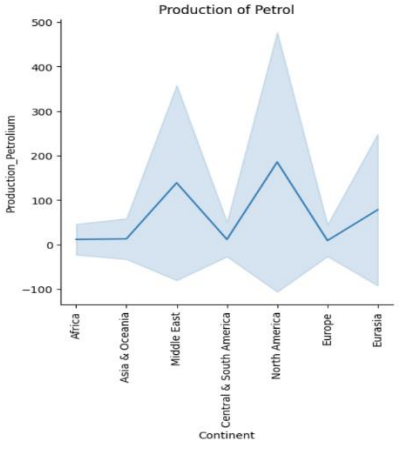


**Figure 8 – Relational plot showing production of petrol and its relation over years**

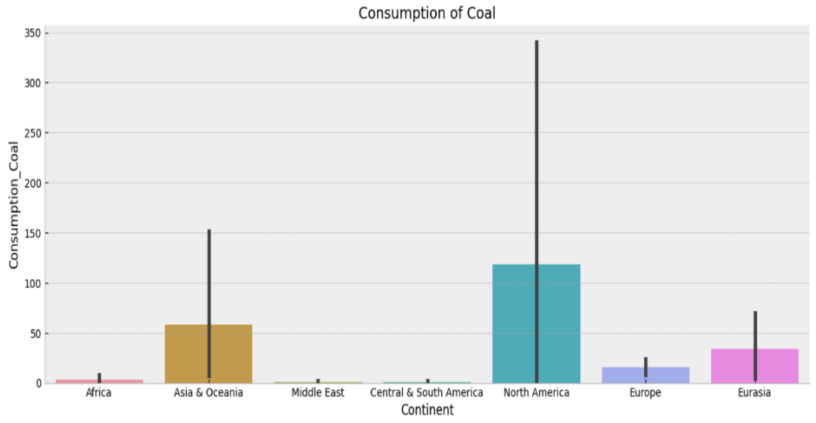


**Figure 9 – Bar plot showing Consumption of coal in different continents**

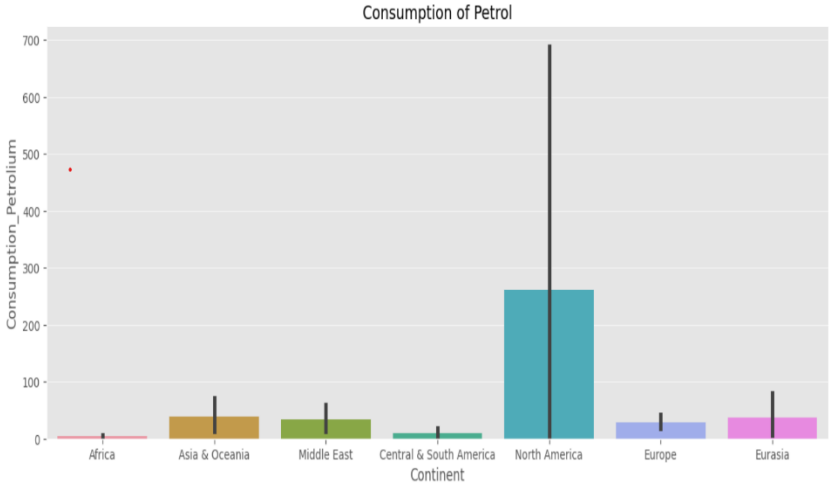


**Figure 10– Bar plot showing Consumption of petrol in different continents**

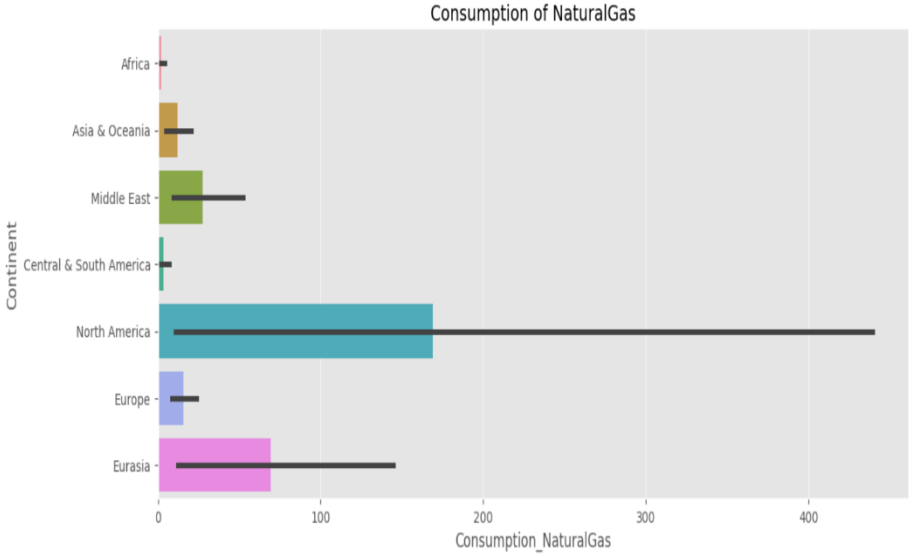


**Figure 11 – Bar plot showing Consumption of natural gas in different continents**

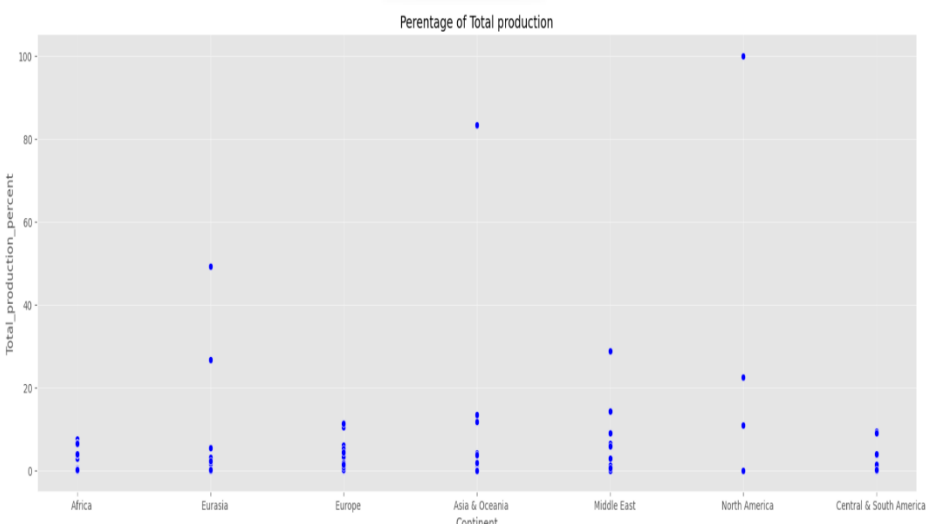


**Figure 12– Percentage of total production in each country across different continents**

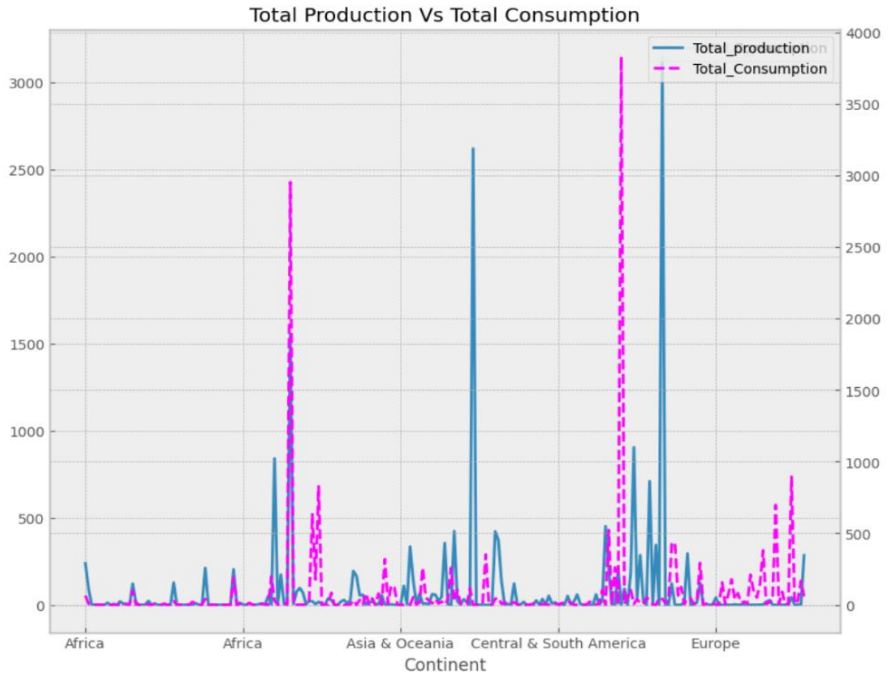


**Figure 13 – Relation between total production and total consumption over different continents**

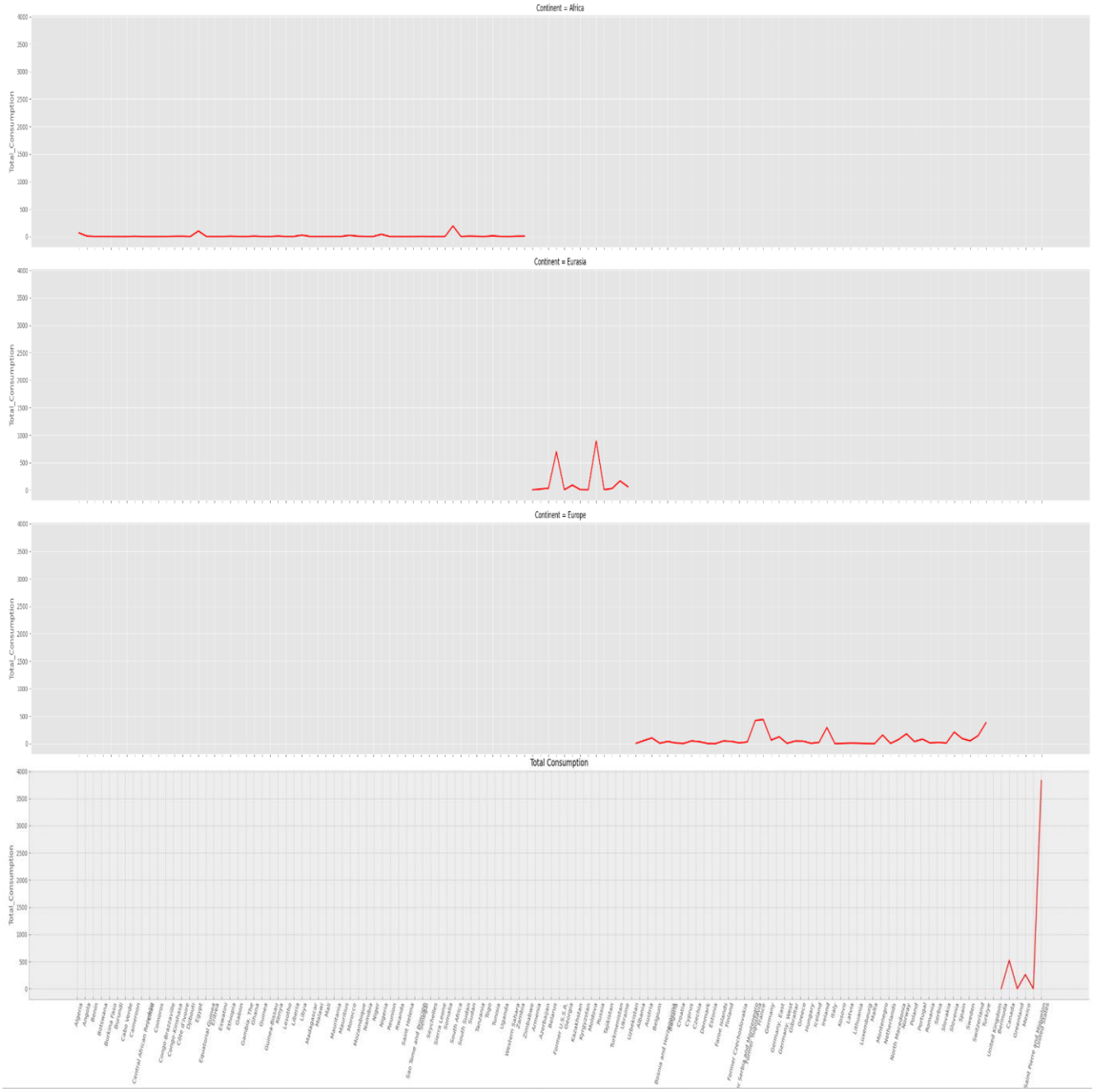


**Figure 14 – Total consumption over different countries**

**Results of Data modelling**

The Linear Regression model has been applied on production data in the year 2020 (**x – dependent variable**) and production data in 2021 (**y – independent variable**). In figure 15, a scatter plot is used to visualize the developed linear regression model and its Pearson Correlation Coefficient is also measured. Similar approach has been followed in consumption data and its corresponding scatter plot is as shown in figure 16.

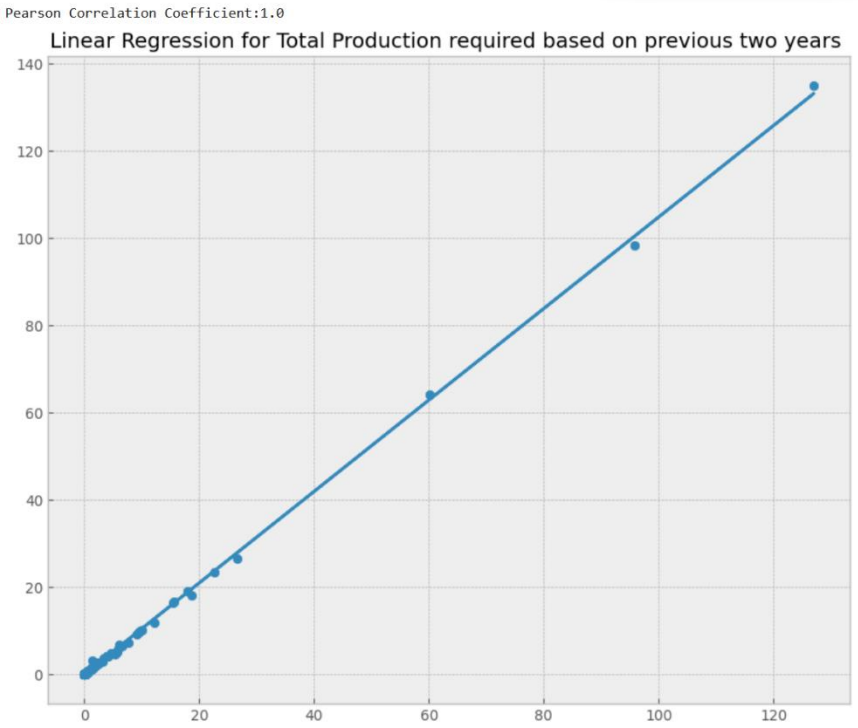


**Figure 15 – Linear regression on total production data**

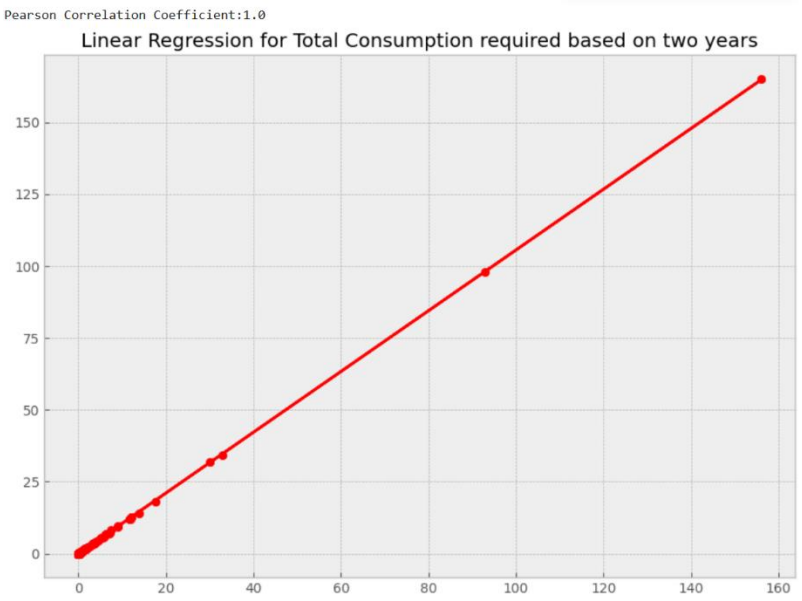


**Figure 16 – Linear regression on total consumption data**

## Insights

- The production in some countries seems to be more than the required consumption as in the case of United Kingdom.
- There is also a case where consumption is more than production in some countries such as in Germany.
- Since there is more consumption in certain countries, redistributing the produced resources can be a feasible solution.
- Based on data analysis on the production data and consumption data, it is evident that both the production and consumption has dependency on its preceding year data.
- Linear regression will be a more relevant model for this dataset as the future data has dependency on preceding year data (dependent variable).

# References


1. https://www.kaggle.com/datasets/akhiljethwa/world-energy-statistics - Global Energy Statistics, Apr 2023.
2. https://www.eia.gov/ - U.S. Energy Information Administration, Apr 2023
3. S. Bilgen, Structure and environmental impact of global energy consumption, Renewable and Sustainable Energy Reviews, Volume 38, 2014, PP 890-902, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2014.07.004.
4. Kumari, Khushbu & Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33. 10.4103/jpcs.jpcs_8_18.
5. Mahesh B. Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet]. 2020 Jan;9:381-6.