Uncovering Insights in Nuclear Dataset Using Exploratory Data Analysis

Example Dataset :“Nuclear Share of Electricity Generation" from Kaggle.

Objective:

- Analyze nuclear electricity production data to uncover trends and patterns.
- Identify key factors influencing nuclear electricity generation.
- Provide actionable insights and recommendations for policy and decision-makers.

Introduction

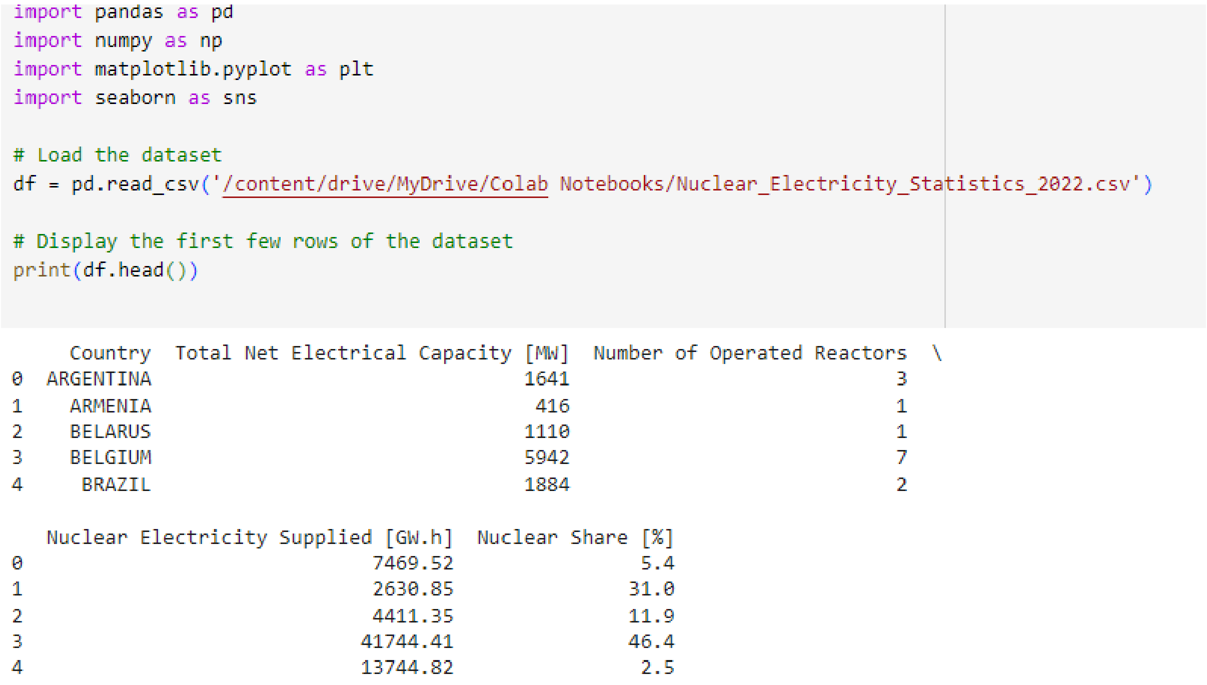
Nuclear energy is a significant source of electricity generation worldwide, offering a low-carbon alternative to fossil fuels. Understanding the trends and patterns in nuclear electricity production is crucial for energy planning and policy formulation. This project aims to perform an exploratory data analysis (EDA) on a dataset containing nuclear electricity statistics for various countries. By analyzing this data, we aim to uncover insights into the global nuclear energy landscape and identify factors that influence nuclear electricity generation.

Dataset Description

The dataset contains information on nuclear electricity production for different countries over several years. Key attributes include:

- Country: The name of the country.
- Year: The year of the record.
- Electricity production (TWh): The total nuclear electricity production in terawatt-hours.

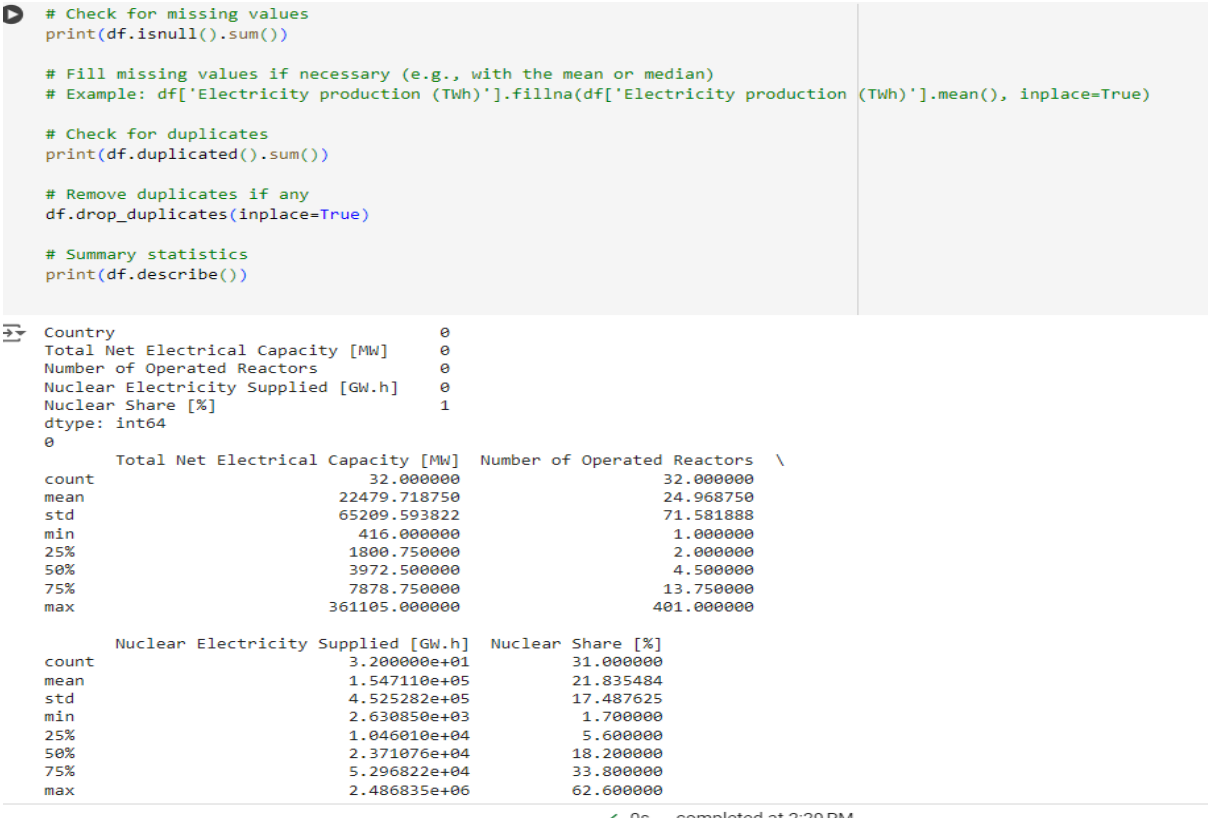
Exploratory Data Analysis: Load the data



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/Nuclear_Electricity_Statistics_2022.csv')

# Display the first few rows of the dataset
print(df.head())
```

```
     Country  Total Net Electrical Capacity [MW]  Number of Operated Reactors  \
0  ARGENTINA                                1641                            3   
1    ARMENIA                                 416                            1   
2    BELARUS                                1110                            1   
3    BELGIUM                                5942                            7   
4     BRAZIL                                1884                            2   

   Nuclear Electricity Supplied [GW.h]  Nuclear Share [%]  
0                              7469.52                5.4  
1                              2630.85               31.0  
2                              4411.35               11.9  
3                             41744.41               46.4  
4                             13744.82                2.5  
```

Data Cleaning :Handle missing values, outliers, and duplicates.



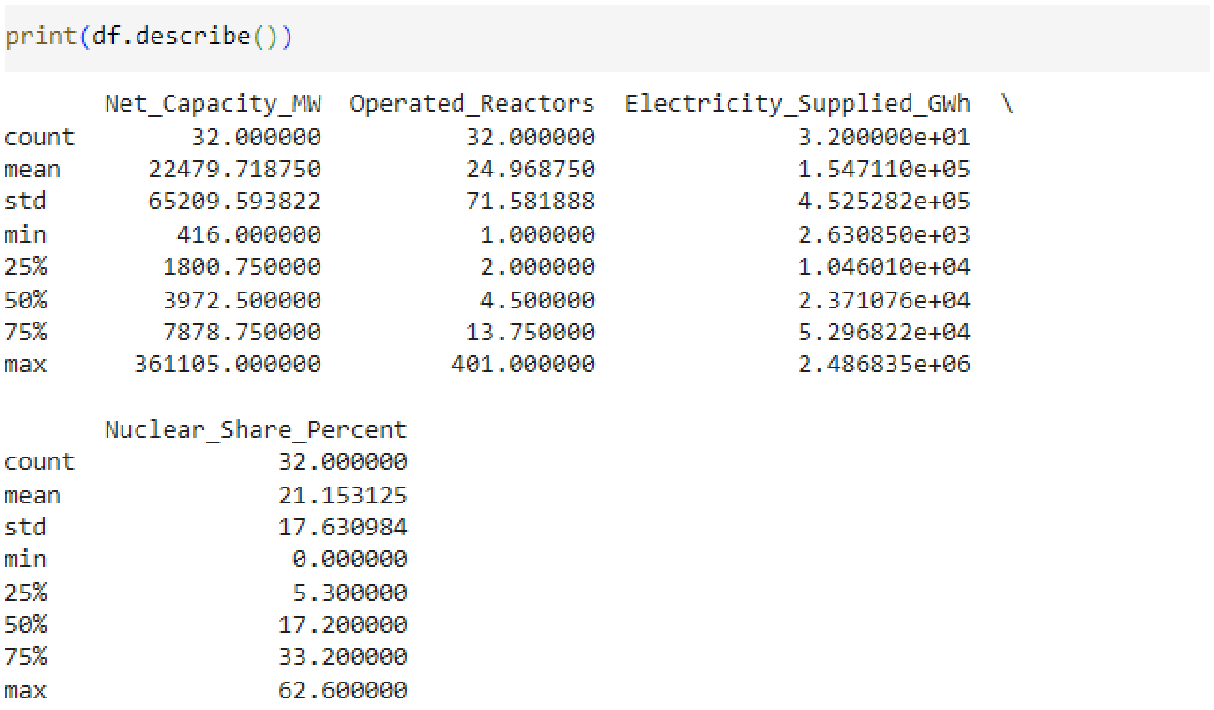
```
# Check for missing values
print(df.isnull().sum())

# Fill missing values if necessary (e.g., with the mean or median)
# Example: df['Electricity production (TWh)'].fillna(df['Electricity production (TWh)'].mean(), inplace=True)

# Check for duplicates
print(df.duplicated().sum())

# Remove duplicates if any
df.drop_duplicates(inplace=True)

# Summary statistics
print(df.describe())
```

```
Country                                0
Total Net Electrical Capacity [MW]     0
Number of Operated Reactors            0
Nuclear Electricity Supplied [GW.h]    0
Nuclear Share [%]                      1
dtype: int64
0
       Total Net Electrical Capacity [MW]  Number of Operated Reactors  \
count                           32.000000                    32.000000   
mean                         22479.718750                    24.968750   
std                          65209.593822                    71.581888   
min                            416.000000                     1.000000   
25%                           1800.750000                     2.000000   
50%                           3972.500000                     4.500000   
75%                           7878.750000                    13.750000   
max                         361105.000000                   401.000000   

       Nuclear Electricity Supplied [GW.h]  Nuclear Share [%]  
count                         3.200000e+01          31.000000  
mean                          1.547110e+05          21.835484  
std                           4.525282e+05          17.487625  
min                           2.630850e+03           1.700000  
25%                           1.046010e+04           5.600000  
50%                           2.371076e+04          18.200000  
75%                           5.296822e+04          33.800000  
max                           2.486835e+06          62.600000  
```

✓ 0s completed at 2:30 PM

Summary Statistics

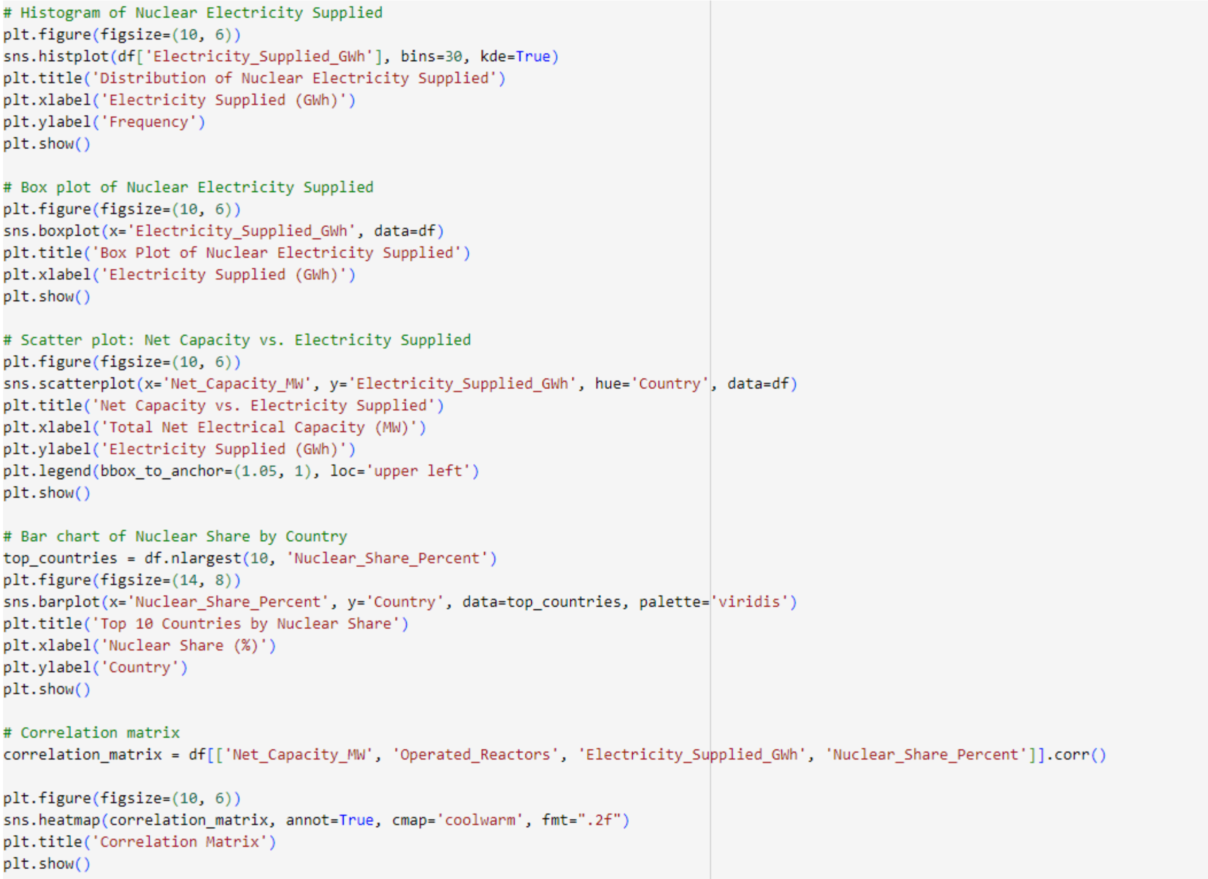
Summarize the dataset to understand its distribution



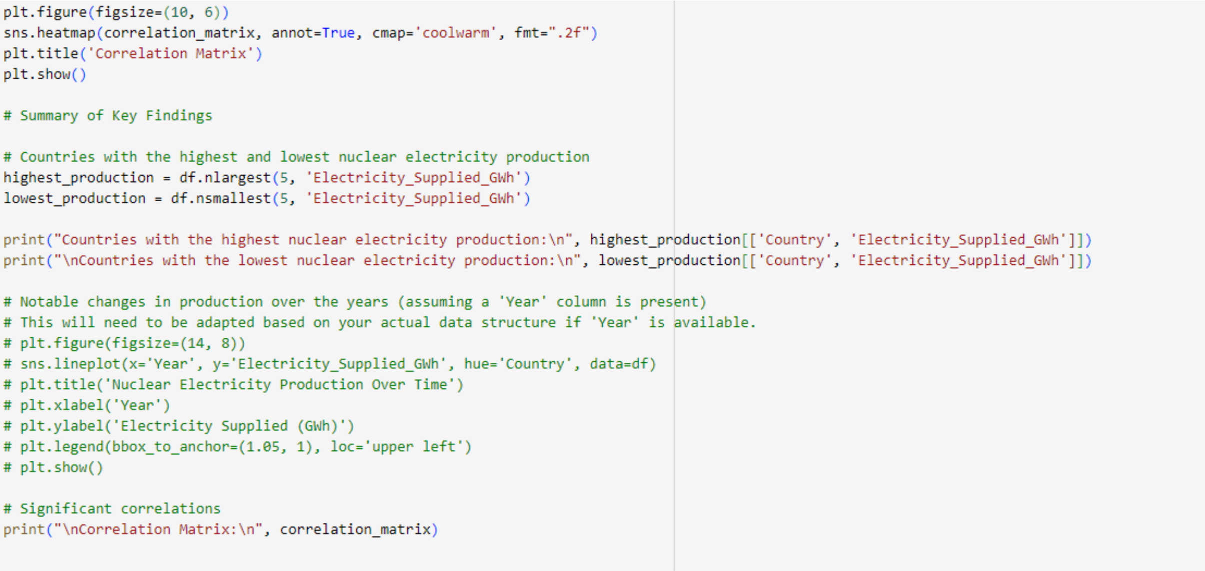
```python
print(df.describe())
```

```
       Net_Capacity_MW  Operated_Reactors  Electricity_Supplied_GWh  \
count        32.000000          32.000000              3.200000e+01   
mean      22479.718750          24.968750              1.547110e+05   
std       65209.593822          71.581888              4.525282e+05   
min         416.000000           1.000000              2.630850e+03   
25%        1800.750000           2.000000              1.046010e+04   
50%        3972.500000           4.500000              2.371076e+04   
75%        7878.750000          13.750000              5.296822e+04   
max      361105.000000         401.000000              2.486835e+06   

       Nuclear_Share_Percent  
count              32.000000  
mean               21.153125  
std                17.630984  
min                 0.000000  
25%                 5.300000  
50%                17.200000  
75%                33.200000  
max                62.600000  
```

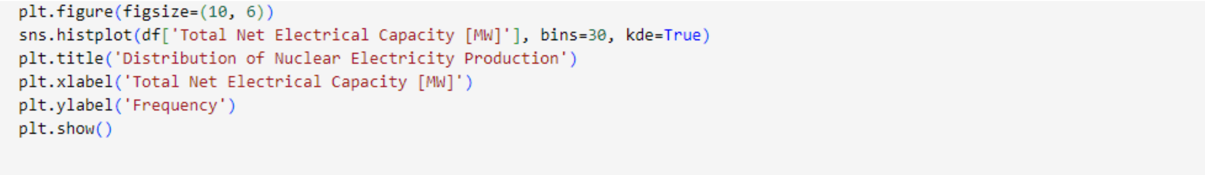
Data Visualization and Discussion



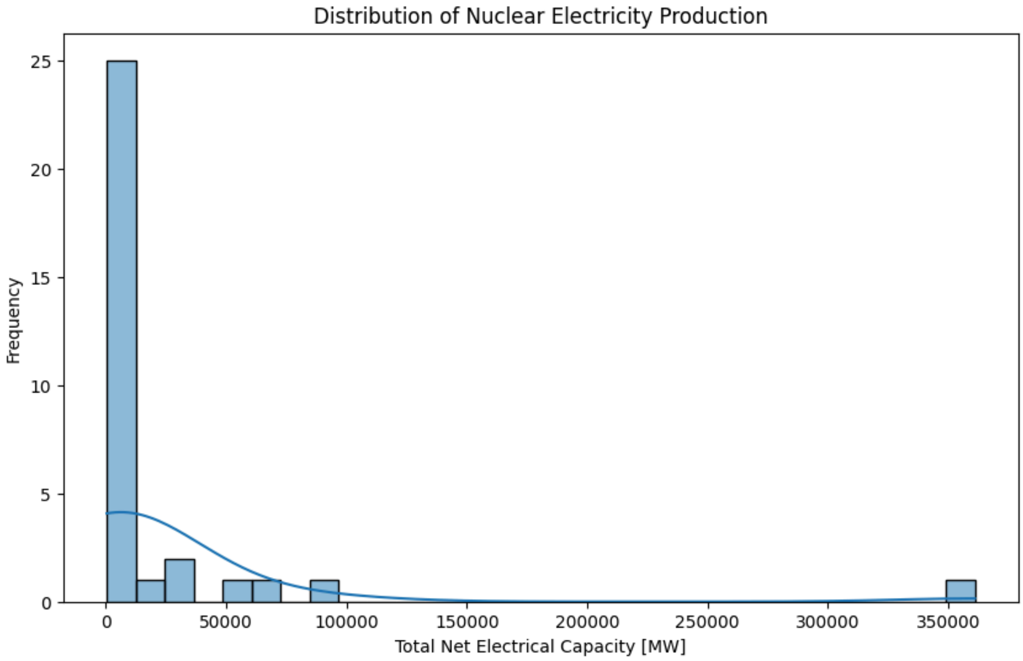
```python
# Histogram of Nuclear Electricity Supplied
plt.figure(figsize=(10, 6))
sns.histplot(df['Electricity_Supplied_GWh'], bins=30, kde=True)
plt.title('Distribution of Nuclear Electricity Supplied')
plt.xlabel('Electricity Supplied (GWh)')
plt.ylabel('Frequency')
plt.show()

# Box plot of Nuclear Electricity Supplied
plt.figure(figsize=(10, 6))
sns.boxplot(x='Electricity_Supplied_GWh', data=df)
plt.title('Box Plot of Nuclear Electricity Supplied')
plt.xlabel('Electricity Supplied (GWh)')
plt.show()

# Scatter plot: Net Capacity vs. Electricity Supplied
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Net_Capacity_MW', y='Electricity_Supplied_GWh', hue='Country', data=df)
plt.title('Net Capacity vs. Electricity Supplied')
plt.xlabel('Total Net Electrical Capacity (MW)')
plt.ylabel('Electricity Supplied (GWh)')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()

# Bar chart of Nuclear Share by Country
top_countries = df.nlargest(10, 'Nuclear_Share_Percent')
plt.figure(figsize=(14, 8))
sns.barplot(x='Nuclear_Share_Percent', y='Country', data=top_countries, palette='viridis')
plt.title('Top 10 Countries by Nuclear Share')
plt.xlabel('Nuclear Share (%)')
plt.ylabel('Country')
plt.show()

# Correlation matrix
correlation_matrix = df[['Net_Capacity_MW', 'Operated_Reactors', 'Electricity_Supplied_GWh', 'Nuclear_Share_Percent']].corr()

plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```

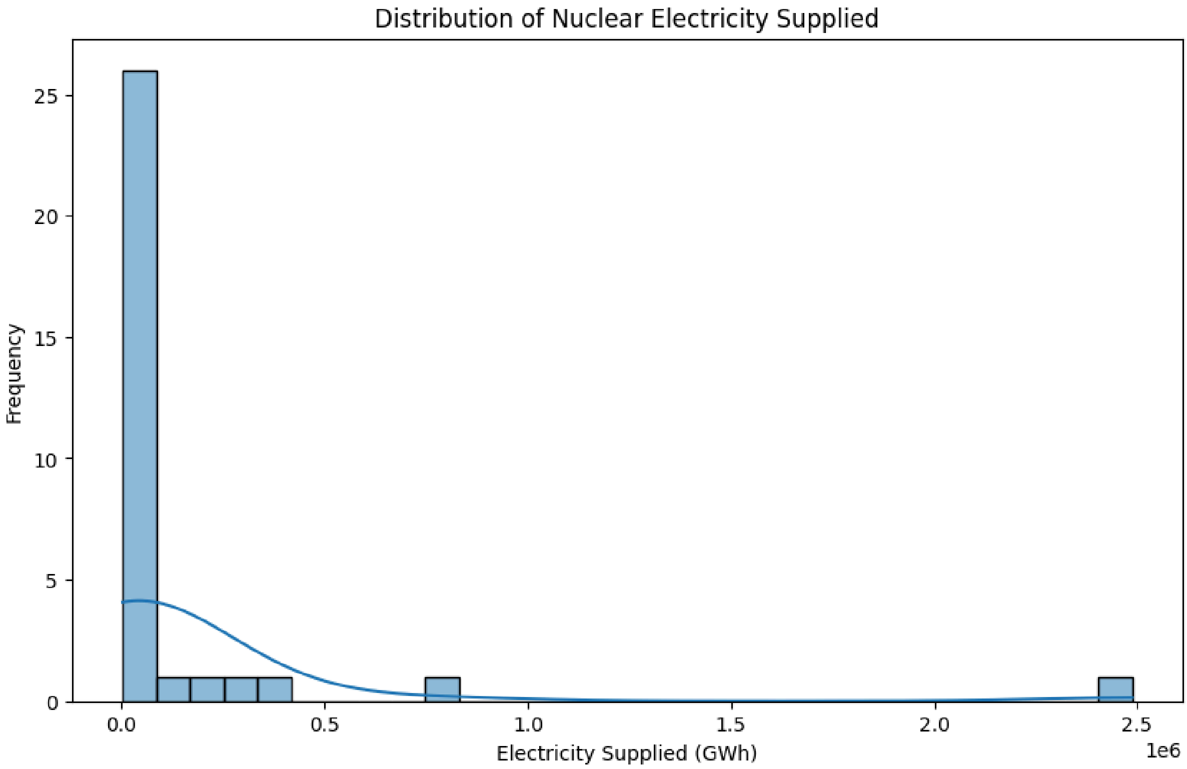


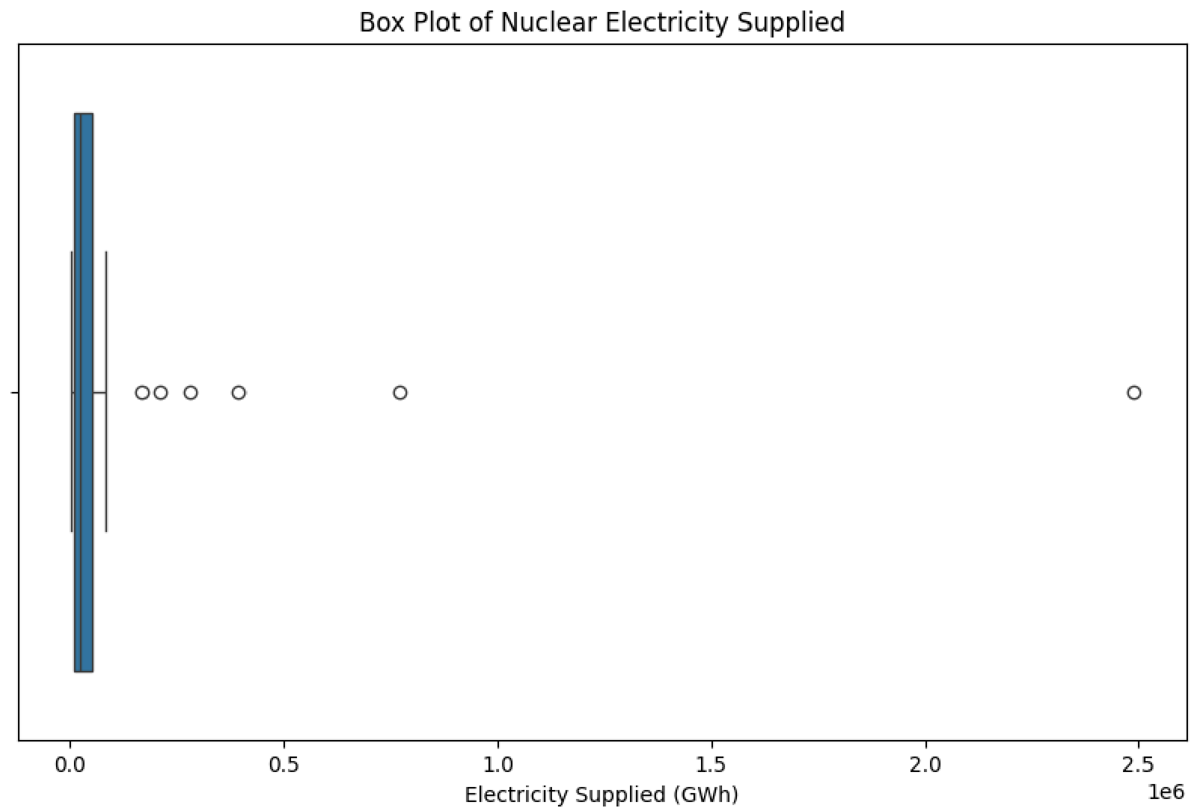
```python
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()

# Summary of Key Findings

# Countries with the highest and lowest nuclear electricity production
highest_production = df.nlargest(5, 'Electricity_Supplied_GWh')
lowest_production = df.nsmallest(5, 'Electricity_Supplied_GWh')

print("Countries with the highest nuclear electricity production:\n", highest_production[['Country', 'Electricity_Supplied_GWh']])
print("\nCountries with the lowest nuclear electricity production:\n", lowest_production[['Country', 'Electricity_Supplied_GWh']])

# Notable changes in production over the years (assuming a 'Year' column is present)
# This will need to be adapted based on your actual data structure if 'Year' is available.
# plt.figure(figsize=(14, 8))
# sns.lineplot(x='Year', y='Electricity_Supplied_GWh', hue='Country', data=df)
# plt.title('Nuclear Electricity Production Over Time')
# plt.xlabel('Year')
# plt.ylabel('Electricity Supplied (GWh)')
# plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
# plt.show()

# Significant correlations
print("\nCorrelation Matrix:\n", correlation_matrix)
```

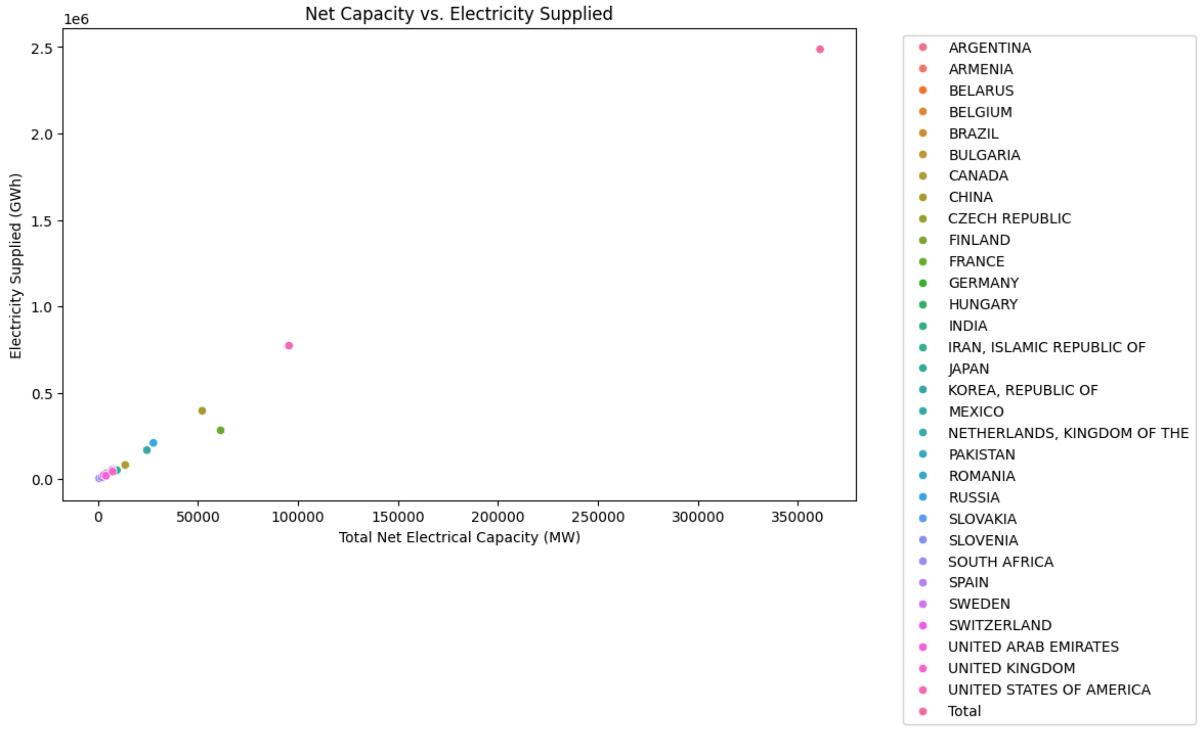


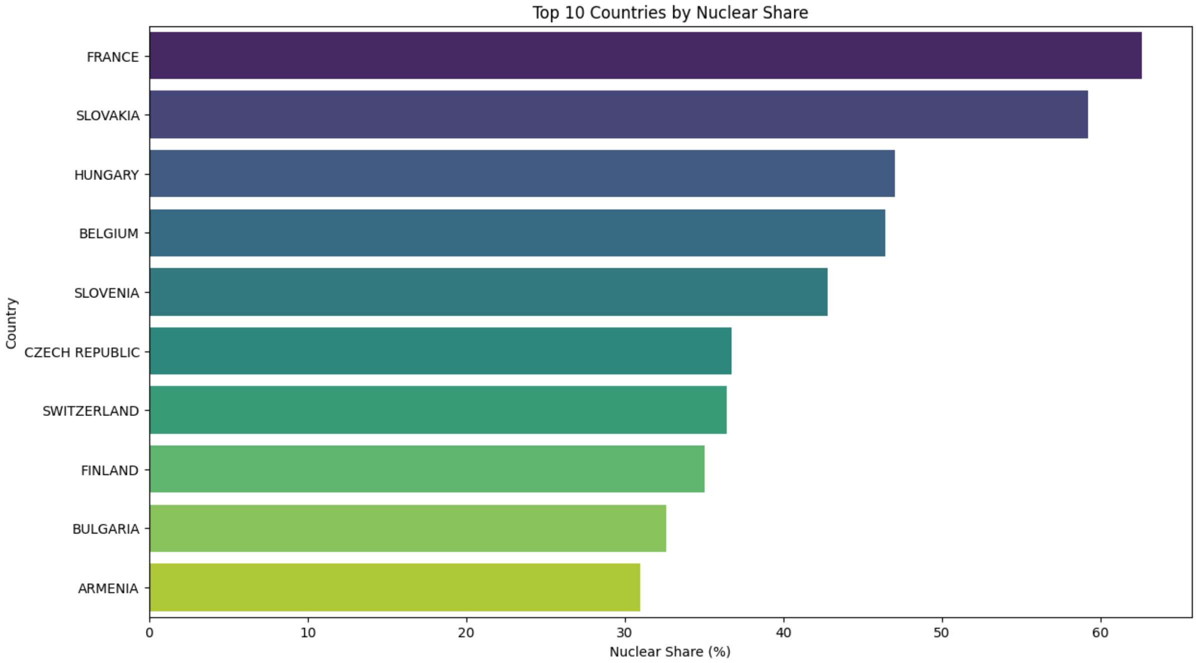
```
plt.figure(figsize=(10, 6))
sns.histplot(df['Total Net Electrical Capacity [MW]'], bins=30, kde=True)
plt.title('Distribution of Nuclear Electricity Production')
plt.xlabel('Total Net Electrical Capacity [MW]')
plt.ylabel('Frequency')
plt.show()
```

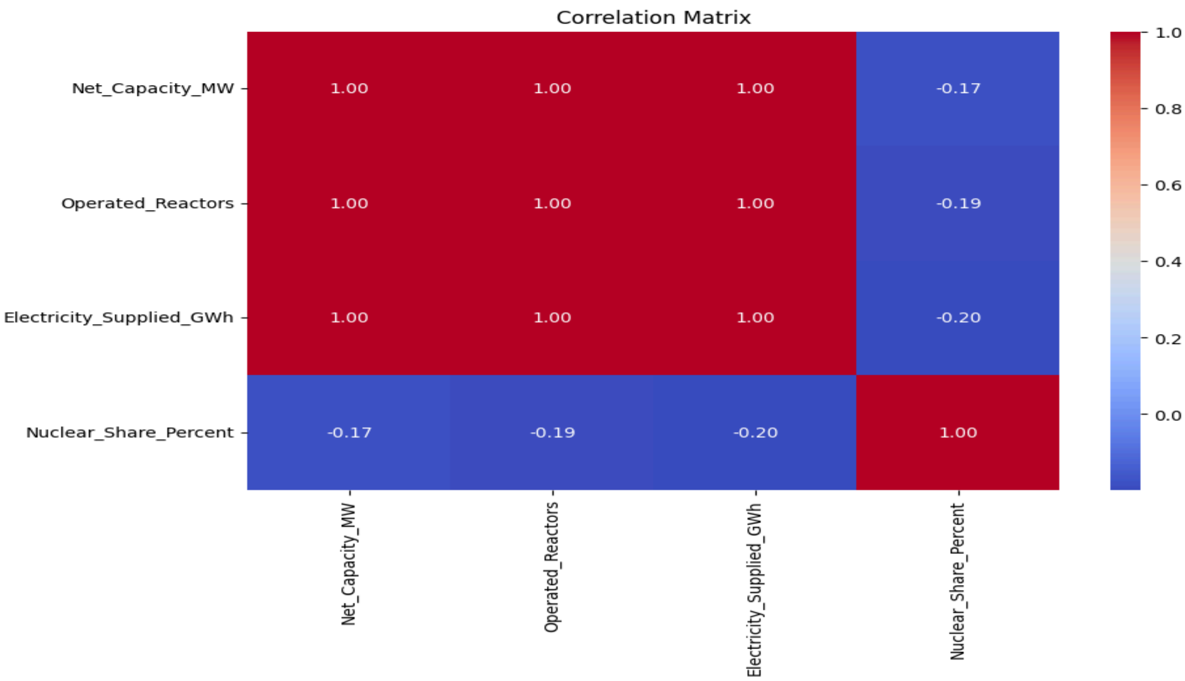


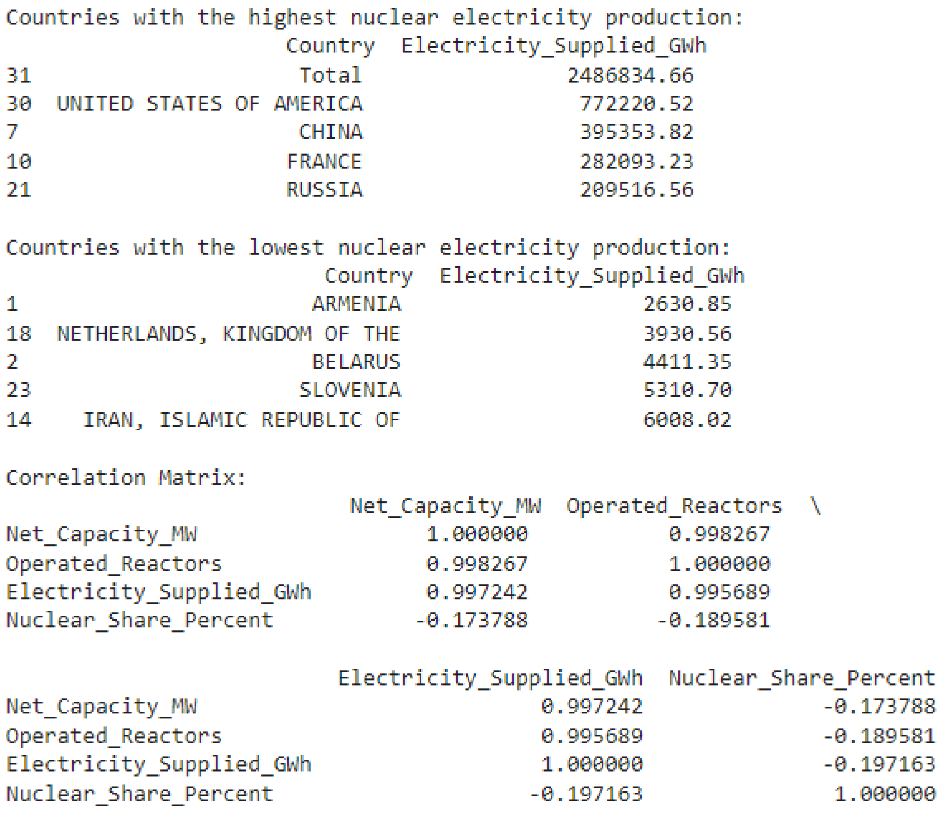



Box Plot of Nuclear Electricity Supplied
0.0
0.5
1.0
1.5
2.0
2.5
1e6
Electricity Supplied (GWh)


Net Capacity vs. Electricity Supplied
1e6
2.5
2.0
1.5
1.0
0.5
0.0
Electricity Supplied (GWh)
0
50000
100000
150000
200000
250000
300000
350000
Total Net Electrical Capacity (MW)
ARGENTINA
ARMENIA
BELARUS
BELGIUM
BRAZIL
BULGARIA
CANADA
CHINA
CZECH REPUBLIC
FINLAND
FRANCE
GERMANY
HUNGARY
INDIA
IRAN, ISLAMIC REPUBLIC OF
JAPAN
KOREA, REPUBLIC OF
MEXICO
NETHERLANDS, KINGDOM OF THE
PAKISTAN
ROMANIA
RUSSIA
SLOVAKIA
SLOVENIA
SOUTH AFRICA
SPAIN
SWEDEN
SWITZERLAND
UNITED ARAB EMIRATES
UNITED KINGDOM
UNITED STATES OF AMERICA
Total


Top 10 Countries by Nuclear Share
Country
FRANCE
SLOVAKIA
HUNGARY
BELGIUM
SLOVENIA
CZECH REPUBLIC
SWITZERLAND
FINLAND
BULGARIA
ARMENIA
0
10
20
30
40
50
60
Nuclear Share (%)




```
Countries with the highest nuclear electricity production:
                      Country  Electricity_Supplied_GWh
31                      Total                2486834.66
30  UNITED STATES OF AMERICA                 772220.52
7                       CHINA                 395353.82
10                     FRANCE                 282093.23
21                     RUSSIA                 209516.56

Countries with the lowest nuclear electricity production:
                         Country  Electricity_Supplied_GWh
1                        ARMENIA                   2630.85
18  NETHERLANDS, KINGDOM OF THE                   3930.56
2                        BELARUS                   4411.35
23                      SLOVENIA                   5310.70
14    IRAN, ISLAMIC REPUBLIC OF                   6008.02

Correlation Matrix:
                           Net_Capacity_MW  Operated_Reactors  \
Net_Capacity_MW                   1.000000           0.998267   
Operated_Reactors                 0.998267           1.000000   
Electricity_Supplied_GWh          0.997242           0.995689   
Nuclear_Share_Percent            -0.173788          -0.189581   

                          Electricity_Supplied_GWh  Nuclear_Share_Percent  
Net_Capacity_MW                           0.997242              -0.173788  
Operated_Reactors                         0.995689              -0.189581  
Electricity_Supplied_GWh                  1.000000              -0.197163  
Nuclear_Share_Percent                    -0.197163               1.000000  
```

Summarize key findings:

- Most of the countries have Total net electric Capacity is less than 50000 Mw.
- United States of America has the most Nuclear Electric Production.
- Armenia has the lowest Nuclear Electric Production.
- There is a strong correlation between Net_Capacity_MW and Electricity_Supplied_GWh, indicating that countries with higher net capacity tend to produce more nuclear electricity.
- The correlation matrix provides insights into how various factors like the number of operated reactors and the nuclear share percentage are related.

Conclusion

- The exploratory data analysis revealed significant insights into nuclear electricity production across different countries. Countries like the USA and France are leading in nuclear electricity production, while countries like Armenia and Netherlands have much lower production. The strong correlation between net electrical capacity and electricity supplied underscores the importance of infrastructure in nuclear energy generation.

References:https://www.kaggle.com/datasets/kanchana1990/nuclear-share-of-electricity-generation