Milestone 1 Report

Data Collection, Preprocessing, and EDA   Feb 5 – Feb 23

Student: Tzu-Chieh Chao

1. Project Objective

This project aims to elucidate the complex relationships between carbon dioxide emissions, sustainable energy practices, and key economic indicators in the G20 nations from 2000-2020. Leveraging comprehensive datasets sourced from Kaggle, the analysis will examine how variables such as renewable energy share, low-carbon electricity percentage, and CO2 emissions correlate with economic measures like GDP per capita, GDP growth rates, and exchange rates.

2. Tools

- Programming Language: Python
- Libraries:
    - Data Manipulation: Pandas, NumPy, re
    - Visualization: Matplotlib, Seaborn, Plotly, mplcursors
    - Statistical Analysis: SciPy
- Version Control: Git & GitHub

3. Data Sources

To ensure a comprehensive analysis, I’ve selected the following datasets from Kaggle:

1. Global Data on Sustainable Energy(2000-2020)

    - Description: This dataset presents an analysis of sustainable energy indicators and various relevant factors across all countries from the year 2000 to 2020. It encompasses critical dimensions such as access to electricity, the utilization of renewable energy sources, carbon emissions, energy intensity, financial flows, and economic growth. Researchers can engage in comparative analyses among nations, monitor advancements towards Sustainable Development Goal 7, and derive significant insights into the trends of global energy consumption over the specified period.
    - Source: global-data-on-sustainable-energy

2. CO2 Emission by countries Year wise (1750-2022)

    - Description: This dataset will facilitate predictions regarding global warming.
    - Source: co2-emission-by-countries-

3. G20 Countries' Currency Exchange Rates against USD

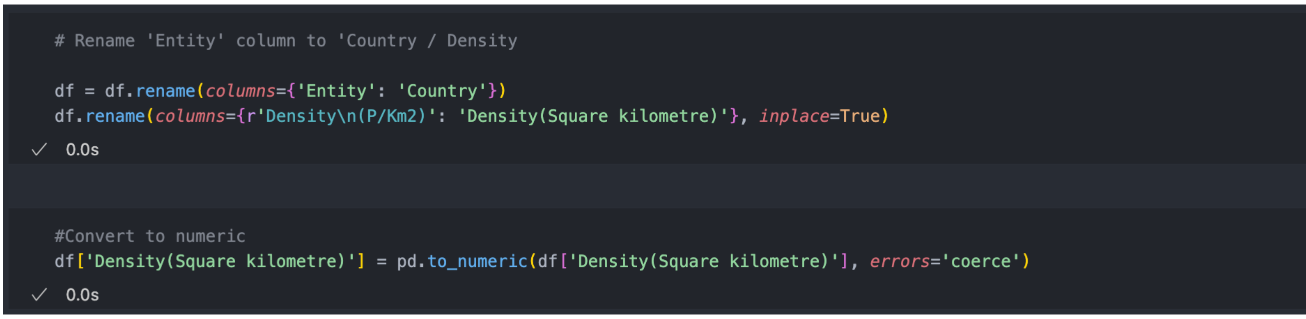
    - Description: This dataset includes the currency exchange rates of G20 countries in relation to the US dollar, thereby facilitating a comprehensive understanding of the macroeconomic landscape.

- Source: g20-countries-currency-exchange-rates-against-usd
-

4. Data Preprocessing

Initially, I perform data cleaning and analysis on three separate datasets. I remove any countries that are not part of the G20 from all three datasets. Furthermore, after relocating the columns with a significant amount of missing data, all missing values have been resolved.

In the Global Data on Sustainable Energy dataset . I changed the column name "Entity" to "Country" and renamed "Density. (P/Km2)" to "Density (Square kilometre)." Additionally, I converted the data type of "Density (Square kilometre)" from object to float.



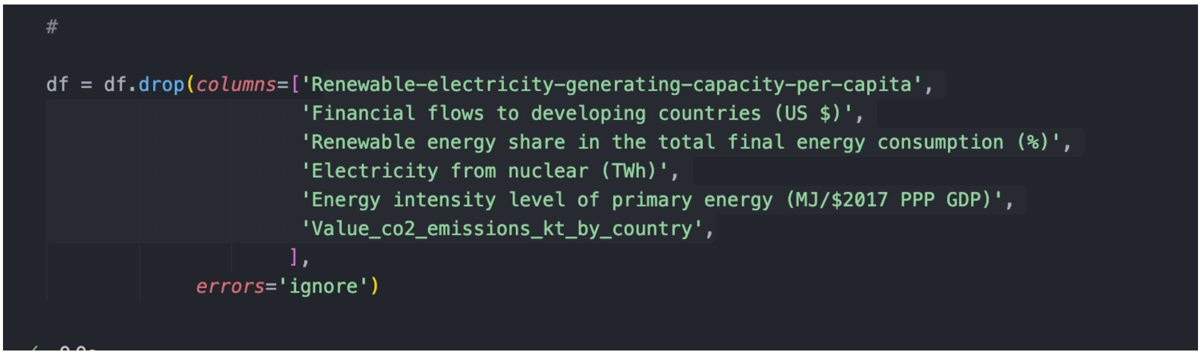
```
# Rename 'Entity' column to 'Country / Density

df = df.rename(columns={'Entity': 'Country'})
df.rename(columns={r'Density\n(P/Km2)': 'Density(Square kilometre)'}, inplace=True)
```

✓ 0.0s

```
#Convert to numeric
df['Density(Square kilometre)'] = pd.to_numeric(df['Density(Square kilometre)'], errors='coerce')
```

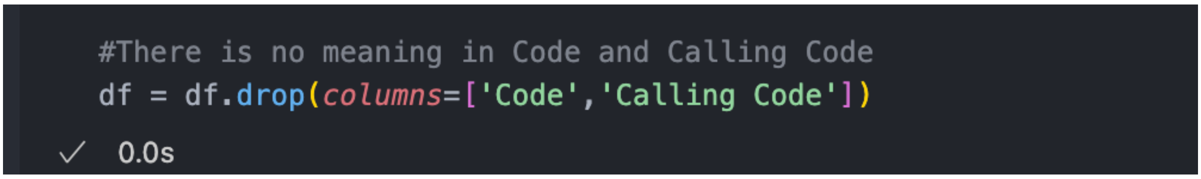
✓ 0.0s

In addition, I will remove the columns ['renewable-electricity-generating-capacity-per-capita', 'Financial flows to developing countries (US $)', 'Renewable energy share in total final energy consumption (%)', 'Electricity from nuclear (TWh)', 'Energy intensity level of primary energy (MJ/$2017 PPP GDP)', 'Value_co2_emissions_kt_by_country'] due to their excessive missing values.



```
#

df = df.drop(columns=['Renewable-electricity-generating-capacity-per-capita',
                      'Financial flows to developing countries (US $)',
                      'Renewable energy share in the total final energy consumption (%)',
                      'Electricity from nuclear (TWh)',
                      'Energy intensity level of primary energy (MJ/$2017 PPP GDP)',
                      'Value_co2_emissions_kt_by_country',
                     ],
             errors='ignore')
```

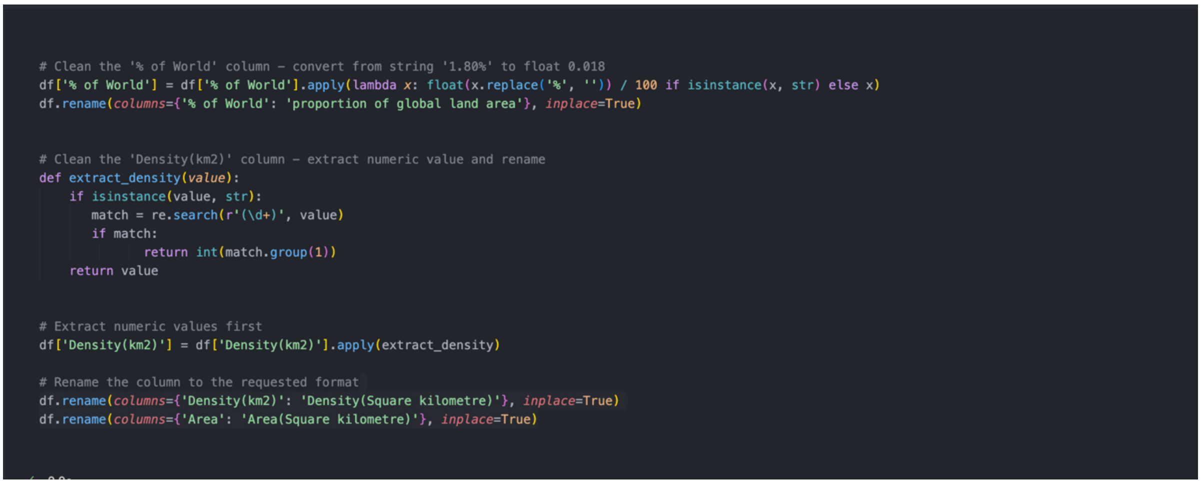
In the CO2 emission by countries(1750-2022) datasets. I eliminated the "Code" (and "Calling Code" columns as they are not meaningful.



```
#There is no meaning in Code and Calling Code
df = df.drop(columns=['Code','Calling Code'])
```

✓ 0.0s

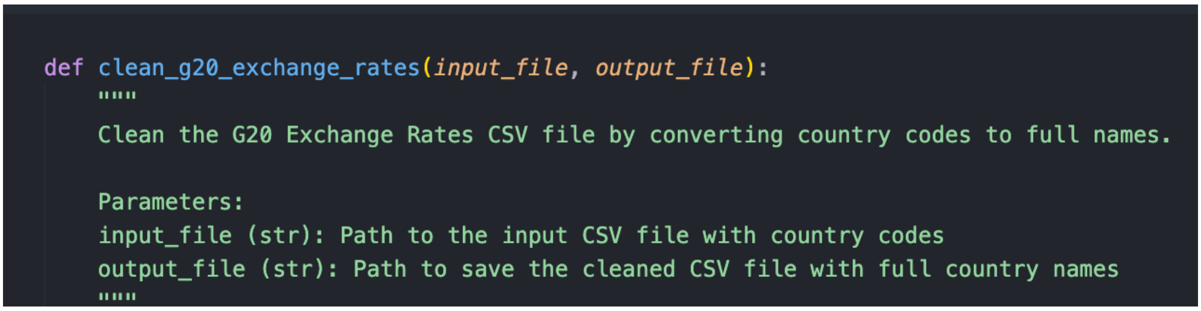
In addition, I updated the column name "% of World" to "proportion of global land area," changed "Density(Km2)" to "Density (Square kilometre)," and renamed "Area" to "Area (Square kilometre)." I also converted the data types of both "Area (Square kilometre)" and "Density (Square kilometre)" from object to numeric.



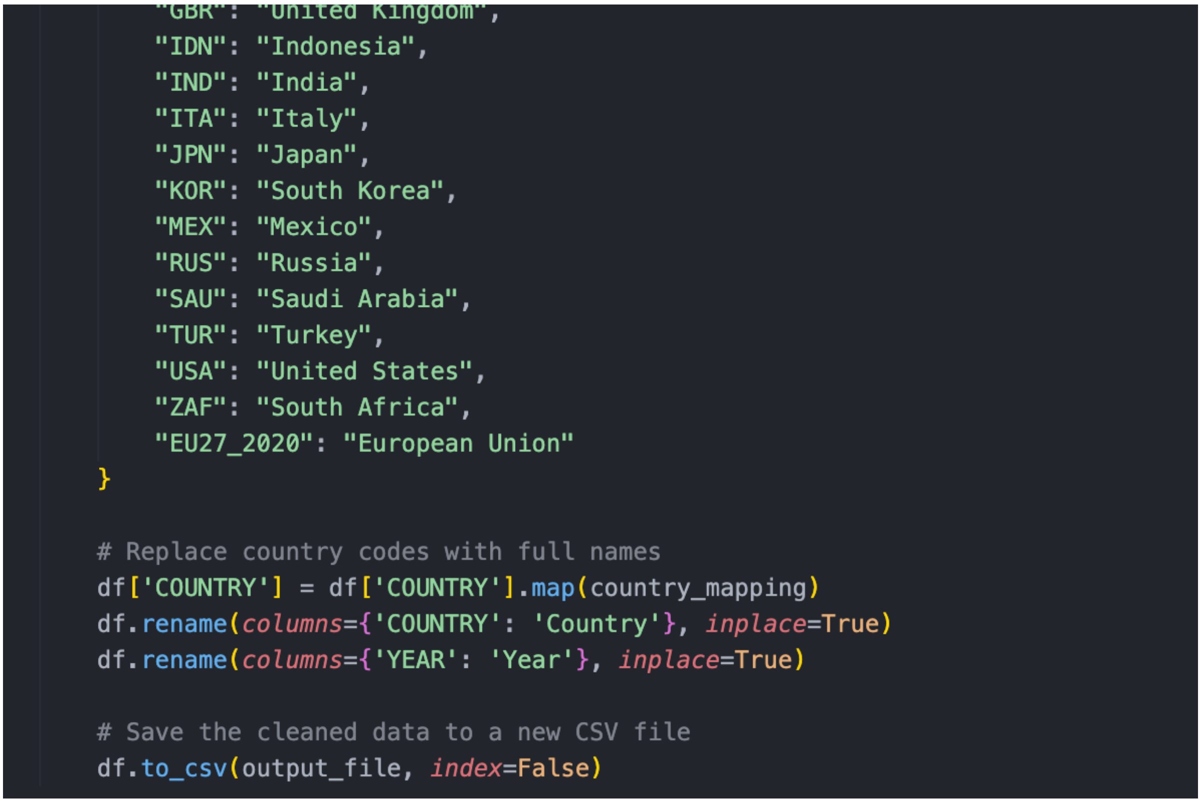
```python
# Clean the '% of World' column - convert from string '1.80%' to float 0.018
df['% of World'] = df['% of World'].apply(lambda x: float(x.replace('%', '')) / 100 if isinstance(x, str) else x)
df.rename(columns={'% of World': 'proportion of global land area'}, inplace=True)


# Clean the 'Density(km2)' column - extract numeric value and rename
def extract_density(value):
    if isinstance(value, str):
        match = re.search(r'(\d+)', value)
        if match:
            return int(match.group(1))
    return value


# Extract numeric values first
df['Density(km2)'] = df['Density(km2)'].apply(extract_density)

# Rename the column to the requested format
df.rename(columns={'Density(km2)': 'Density(Square kilometre)'}, inplace=True)
df.rename(columns={'Area': 'Area(Square kilometre)'}, inplace=True)
```

In the G20 Countries’ Currency Exchange rates against USD dataset.

I modified the column name "COUNTRY" to "Country" and changed "YEAR" to "Year." I I also replace the country code with its full name.

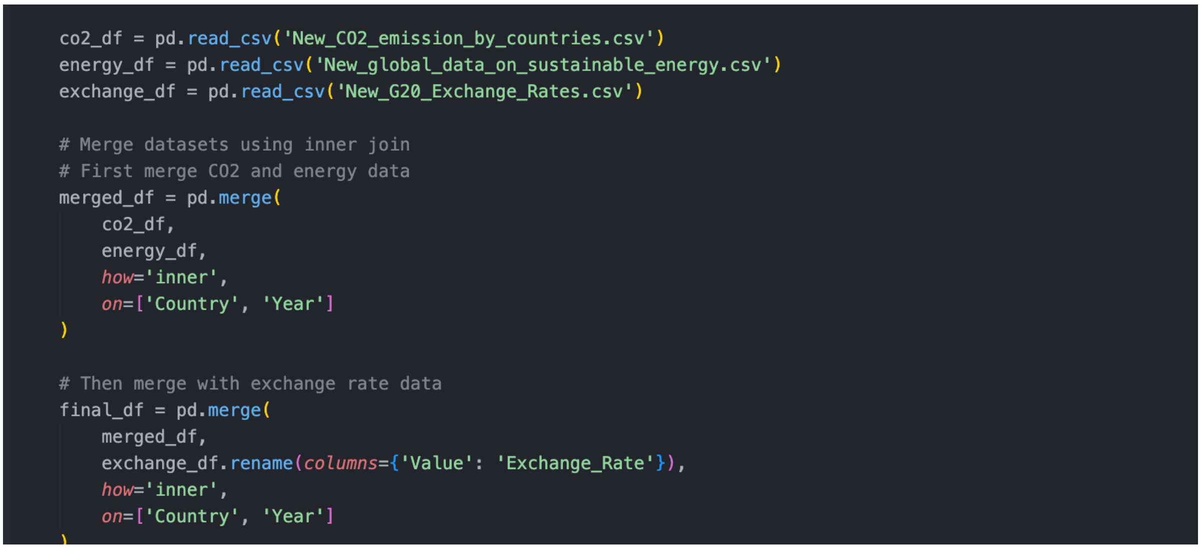


```python
def clean_g20_exchange_rates(input_file, output_file):
    """
    Clean the G20 Exchange Rates CSV file by converting country codes to full names.

    Parameters:
    input_file (str): Path to the input CSV file with country codes
    output_file (str): Path to save the cleaned CSV file with full country names
    """
```



```
        "GBR": "United Kingdom",
        "IDN": "Indonesia",
        "IND": "India",
        "ITA": "Italy",
        "JPN": "Japan",
        "KOR": "South Korea",
        "MEX": "Mexico",
        "RUS": "Russia",
        "SAU": "Saudi Arabia",
        "TUR": "Turkey",
        "USA": "United States",
        "ZAF": "South Africa",
        "EU27_2020": "European Union"
    }

    # Replace country codes with full names
    df['COUNTRY'] = df['COUNTRY'].map(country_mapping)
    df.rename(columns={'COUNTRY': 'Country'}, inplace=True)
    df.rename(columns={'YEAR': 'Year'}, inplace=True)

    # Save the cleaned data to a new CSV file
    df.to_csv(output_file, index=False)
```

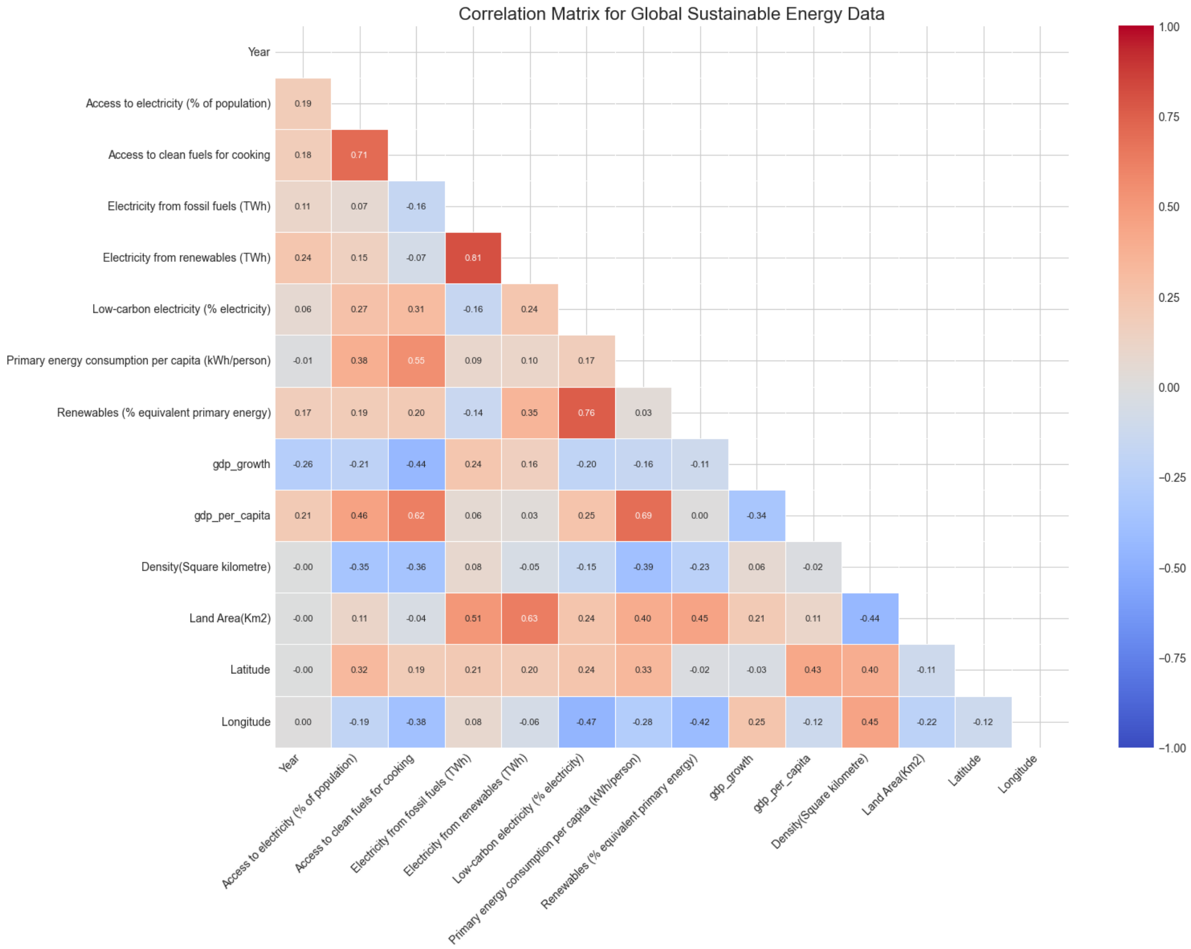
Finally, merging three distinct datasets.



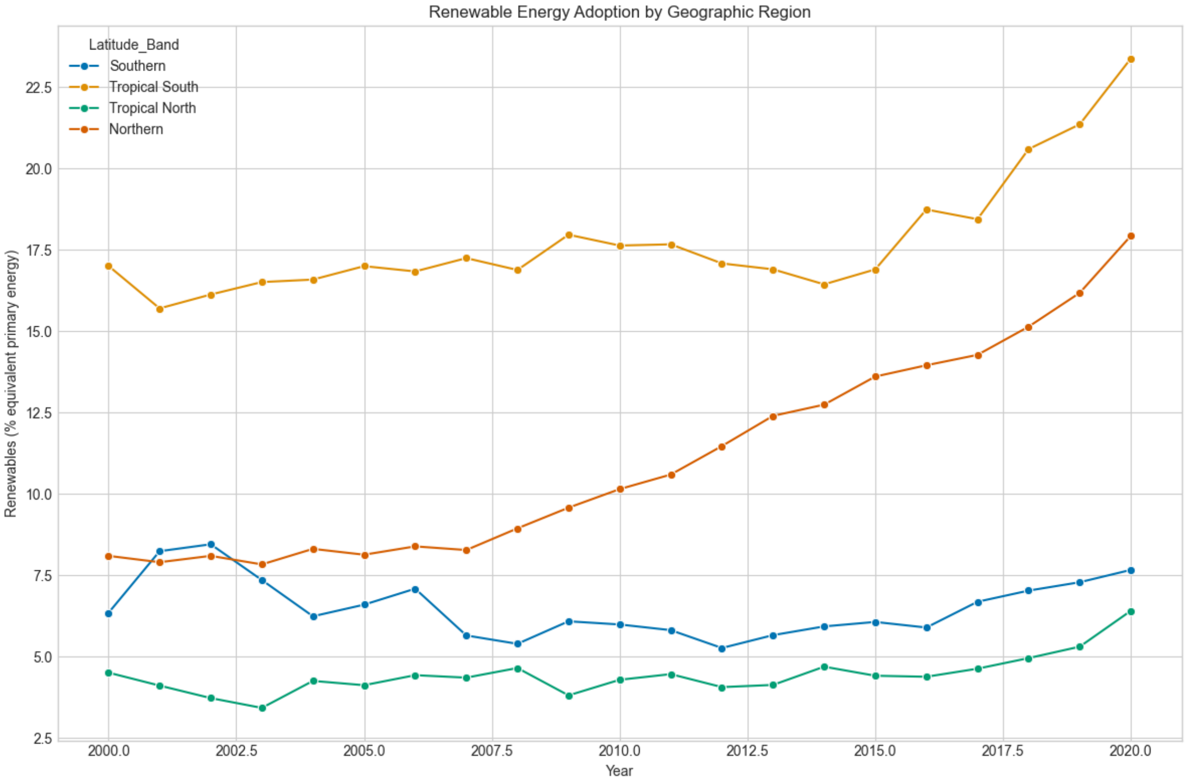
```
co2_df = pd.read_csv('New_CO2_emission_by_countries.csv')
energy_df = pd.read_csv('New_global_data_on_sustainable_energy.csv')
exchange_df = pd.read_csv('New_G20_Exchange_Rates.csv')

# Merge datasets using inner join
# First merge CO2 and energy data
merged_df = pd.merge(
    co2_df,
    energy_df,
    how='inner',
    on=['Country', 'Year']
)

# Then merge with exchange rate data
final_df = pd.merge(
    merged_df,
    exchange_df.rename(columns={'Value': 'Exchange_Rate'}),
    how='inner',
    on=['Country', 'Year']
)
```

5. EDA

Correlation Matrix on Sustainable energy data



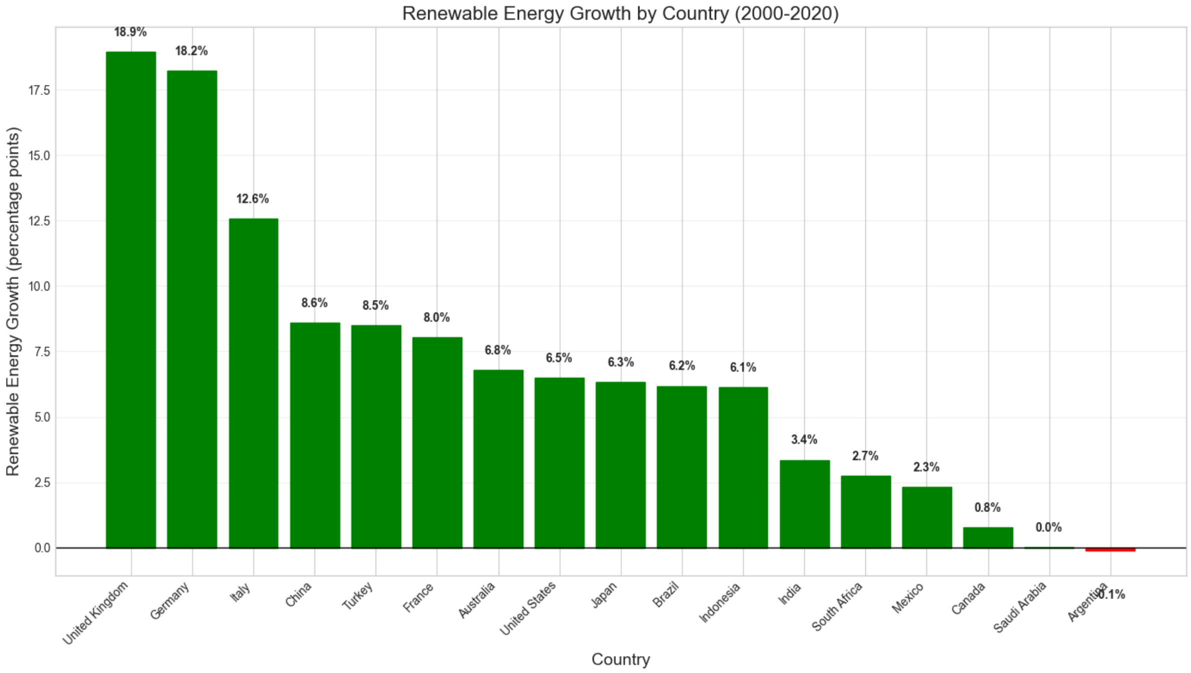
Renewable Energy Regional Analysis(Latitude)



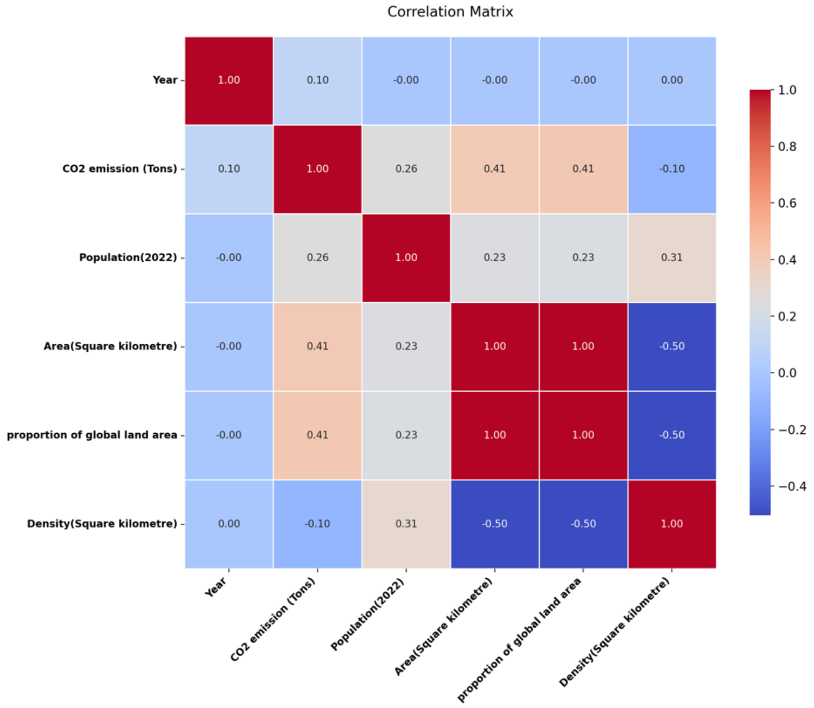
Renewables (% equivalent primary energy): Equivalent primary energy that is derived from renewable sources. Higher values may indicate:

- Higher levels of industrialization Libraries:
- Generally, climatic conditions that demand more energy (such as very cold or hot areas)
- Higher standard of living, using more energy-intensive products and services
- Energy efficiency may be low

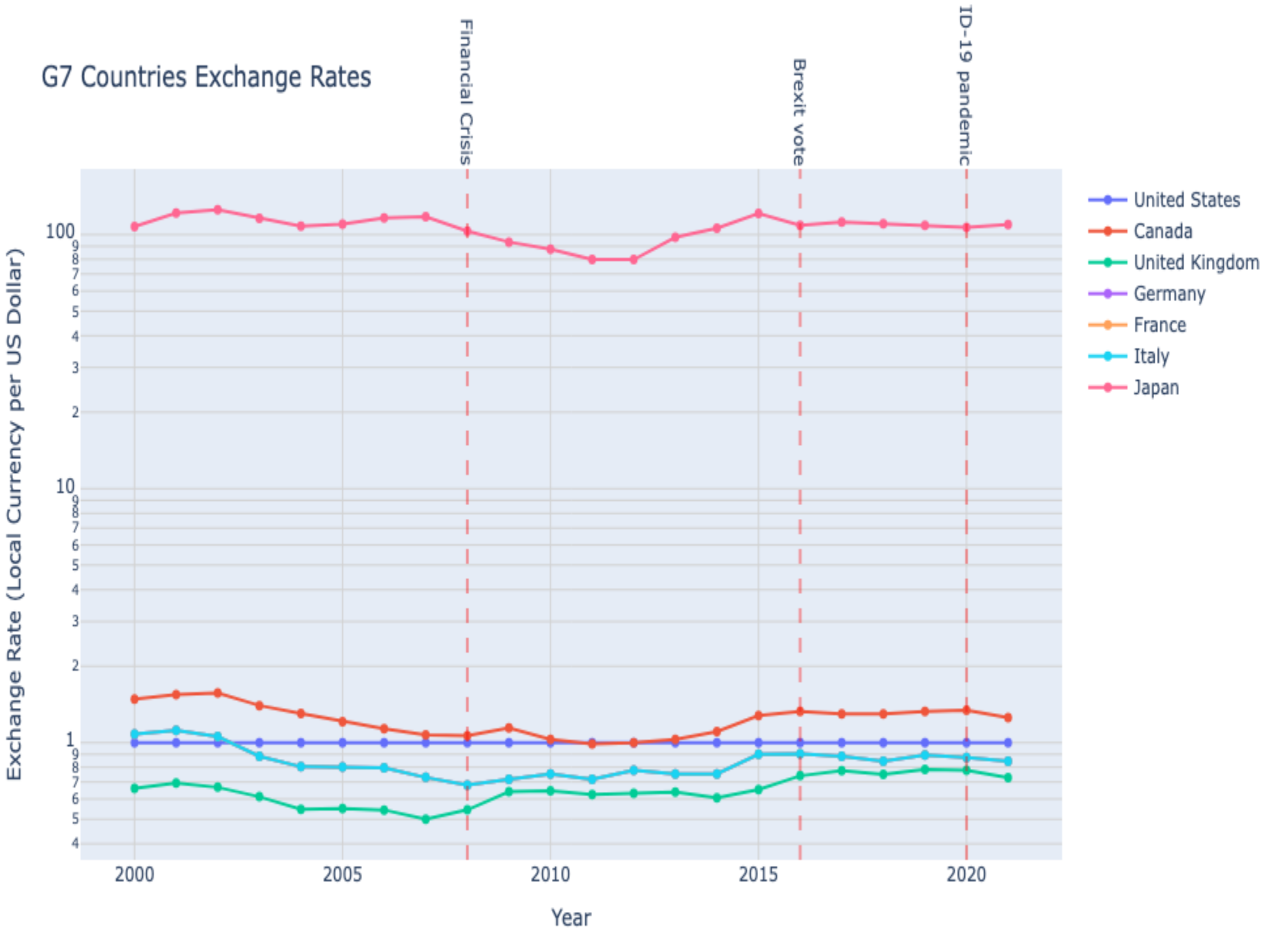
Renewable Energy Growth(2000-2020)



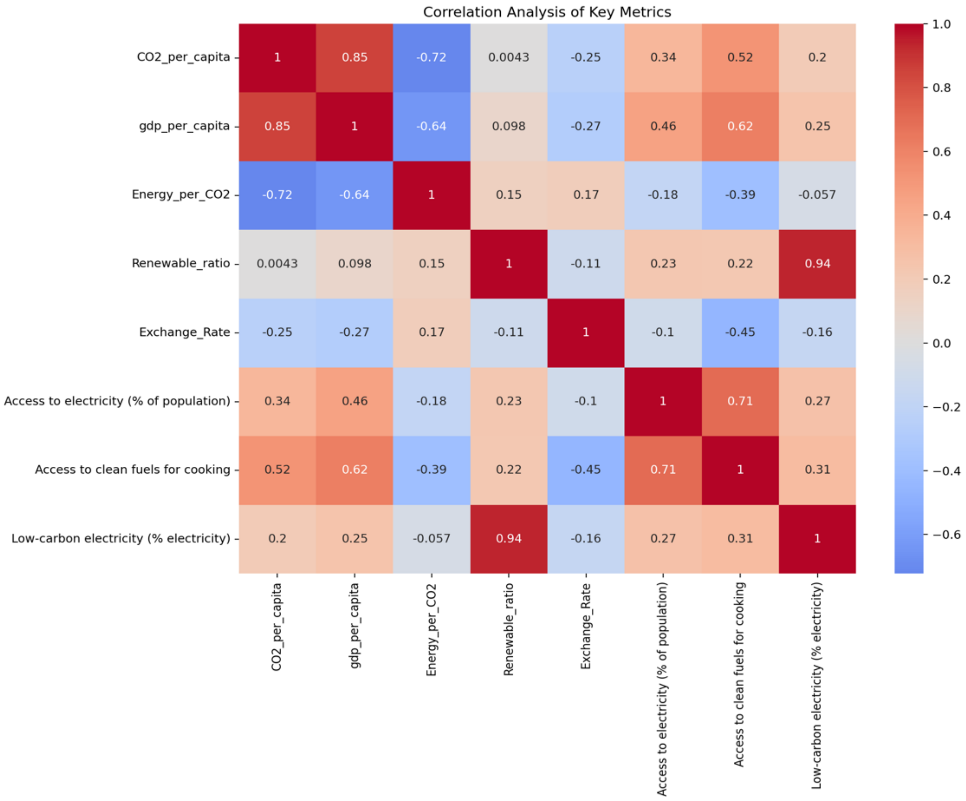
Correlation Matrix on CO2 data



Trend of Exchange Rates

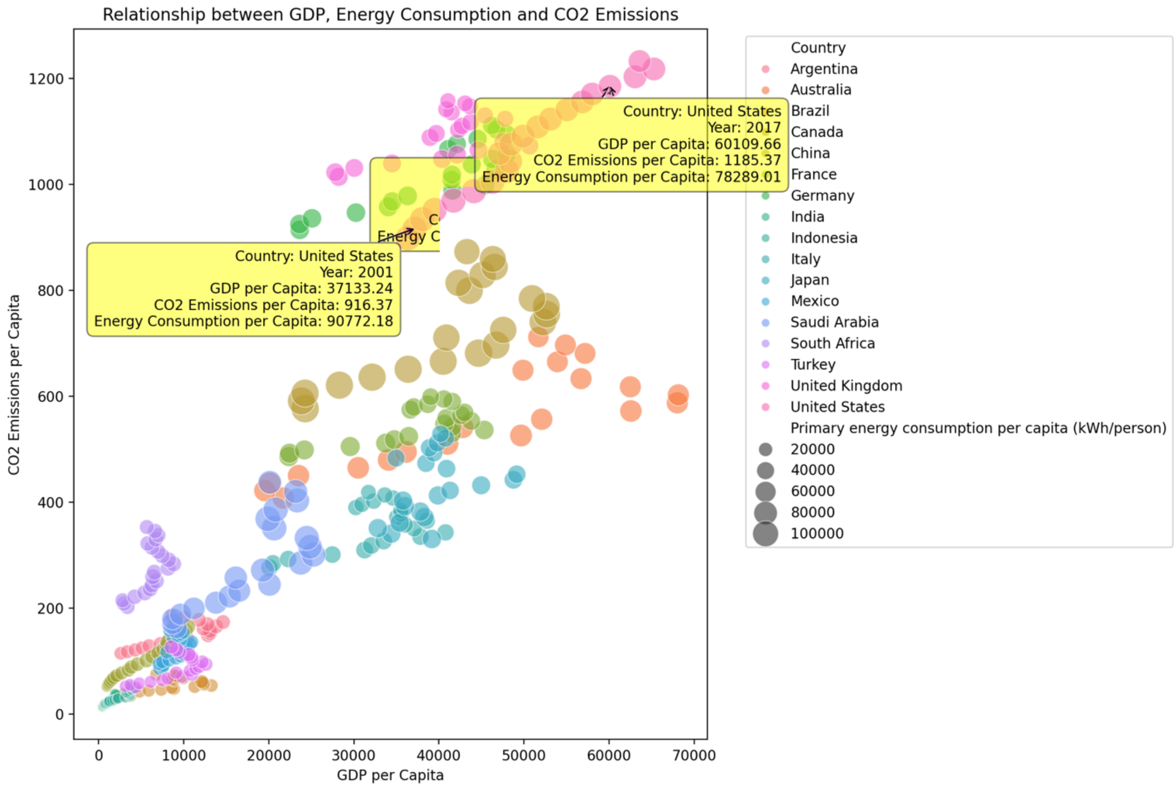


Total Correlation Matrix



Comparison(GDP/CO2/Energy)

Primary energy consumption per capita (kWh/person): Energy consumption per person in kilowatt-hours.


Relationship between GDP, Energy Consumption and CO2 Emissions
CO2 Emissions per Capita
GDP per Capita
Country: United States
Year: 2017
GDP per Capita: 60109.66
CO2 Emissions per Capita: 1185.37
Energy Consumption per Capita: 78289.01
Country: United States
Year: 2001
GDP per Capita: 37133.24
CO2 Emissions per Capita: 916.37
Energy Consumption per Capita: 90772.18
Country
Argentina
Australia
Brazil
Canada
China
France
Germany
India
Indonesia
Italy
Japan
Mexico
Saudi Arabia
South Africa
Turkey
United Kingdom
United States
Primary energy consumption per capita (kWh/person)