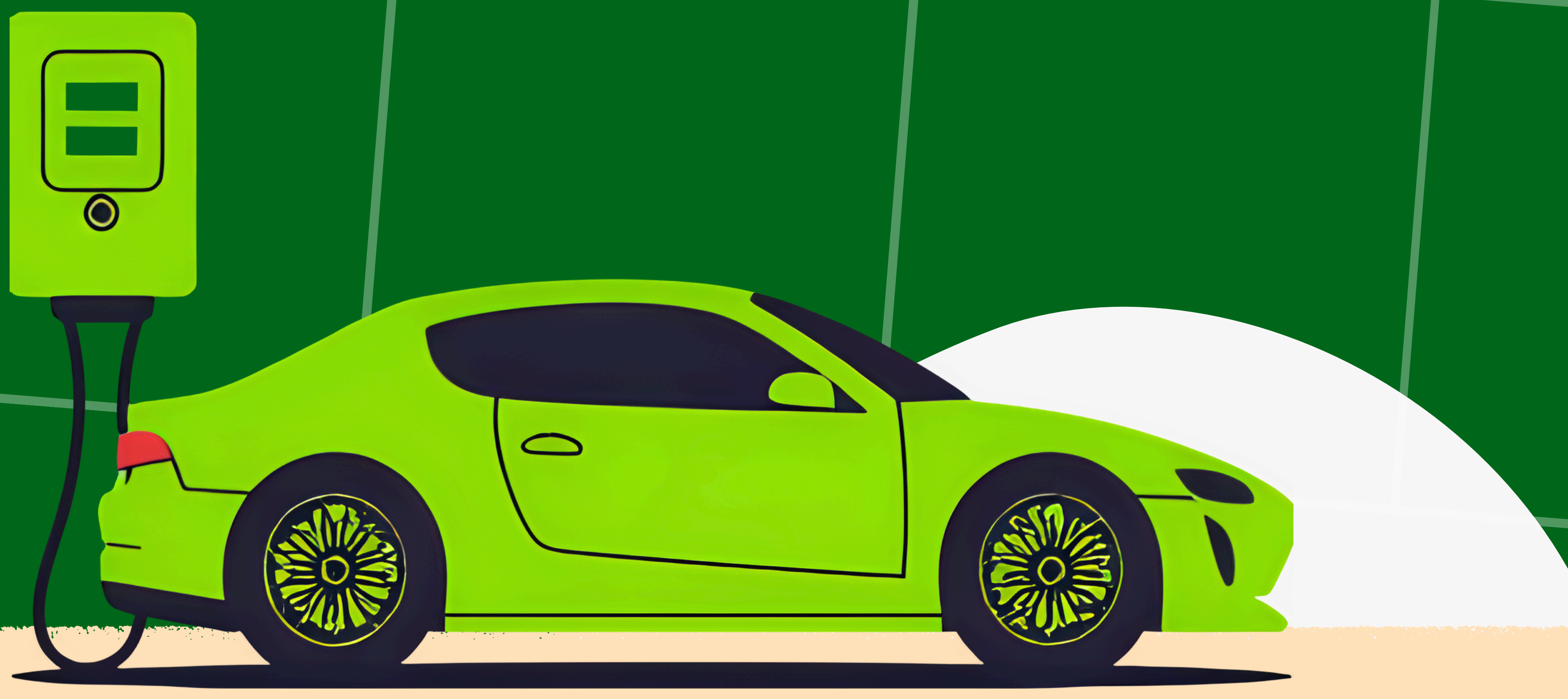
ELECTRIC VEHICLE CAPSTONE PROJECT

Exploratory Data Analysis



INTRODUCTION

The United States has witnessed a transformative shift toward sustainable transportation and clean energy.As concerns about climate change, air pollution, and fossil fuel dependency intensify, the adoption of electric vehicles (EVs) and the expansion of renewable energy sources have emerged as critical components of the national strategy for environmental sustainability.

This project explores the dynamic relationship between electric car registrations and renewable energy production across the United States. By analyzing state-level data from 1990 to 2019, we aim to uncover patterns, trends, and correlations that reflect how the growth in electric vehicle adoption aligns with the development of renewable energy infrastructure, including solar, wind, hydro, and geothermal power.

Through geospatial analysis, time-series exploration, and predictive modeling, this study provides insights into which states are leading the clean energy transition and how renewable energy availability might influence EV adoption rates. The findings can support policy-making, investment decisions, and further research aimed at accelerating the U.S. shift to a low-carbon economy.

OBJECTIVE

This project combines data from multiple sources to analyze trends from 1990 to 2019 , with a focus on the following goals:

1. Data Integration and Preprocessing

To import multiple datasets related to EV registrations, renewable energy generation (e.g., solar, wind, hydro, geothermal), and U.S. state codes. It perform data cleaning and preprocessing, including handling missing values, ensuring consistent data formats, and preparing the data for analysis through merging and transformation.

2. Spatial Distribution Analysis of EV Registrations

To map the distribution of electric vehicle registrations across U.S. states using geospatial visualization techniques (e.g., choropleth maps). It highlight regional patterns in EV adoption and identify states that are leading or lagging in the transition to electric mobility.

3. Comparison with Renewable Energy Projects

To compare the spatial distribution of EV registrations with the locations and outputs of major renewable energy projects. It determine whether states with higher renewable energy production tend to have higher levels of EV adoption.

4. Time Series Analysis of Growth Trends

To analyze the evolution of electric vehicle registrations and renewable energy production over time (1990–2019). It identify periods of accelerated growth and determine if trends in both areas are interconnected or influenced by common external factors such as policy changes or technological advancements.

5. Correlation and Relationship Analysis

To evaluate the statistical relationship between the growth in electric vehicle adoption and the increase in renewable energy generation using correlation coefficients. To assess whether a positive or negative correlation exists and how strong that relationship is across states and over time.

6. Identification of Regional Leaders and Laggards

To identify which states are leading in the adoption of electric vehicles and renewable energy, and which are falling behind. It explore possible socioeconomic, infrastructural, or policy-driven reasons behind these disparities.

7. Deriving Actionable Insights for Policy and Planning

To generate meaningful insights that can help inform policymakers, researchers, and stakeholders on how to support and accelerate the clean energy and transportation transition. To recommend areas where coordinated policy efforts or increased investment could drive greater alignment between EV adoption and renewable energy availability.

By achieving these objectives, the project not only provides a comprehensive picture of the current state of EV and renewable energy progress in the U.S., but also lays the groundwork for future planning, forecasting, and policy development aimed at fostering a more sustainable future.

IMPORTED LIBRARIES

1. NumPy (import numpy as np):

NumPy is a foundational library for numerical computing in Python. It enables the creation and manipulation of large, multi-dimensional arrays and provides a wide range of mathematical functions that support efficient numerical calculations throughout the project.

2. Pandas (import pandas as pd)

Pandas is essential for structured data handling. It offers flexible data structures like DataFrames and Series, allowing for seamless data cleaning, transformation, filtering, and analysis—all critical for working with tabular datasets such as EV registrations and energy production records.

3. Seaborn (import seaborn as sns)

Built on top of Matplotlib, Seaborn simplifies the process of creating visually appealing and informative statistical graphics. It is particularly useful for generating heatmaps, distribution plots, and correlation matrices to uncover relationships in the data

4. Matplotlib (import matplotlib.pyplot as plt)

Matplotlib is a powerful plotting library used to create static, customizable 2D visualizations such as bar charts, line plots, and scatter plots. It provides detailed control over plot formatting, which is useful for precise data representation.

5. Plotly Express (import plotly.express as px)

Plotly Express enables the creation of interactive and web-friendly visualizations with minimal code. It is ideal for building dynamic charts, maps, and time series visualizations that allow users to explore data through zooming and hovering features.

DATA OVERVIEW

1. Loading Data from Google Drive

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
sc = '/content/drive/MyDrive/Colab Notebooks/Copy of state_codes.xlsx'
state_code=pd.read_excel(sc)
r = '/content/drive/MyDrive/Colab Notebooks/Copy of States_Electric_Vehicle_Registrations_2018.xlsx'
registration_EV =pd.read_excel(r)
e='/content/drive/MyDrive/Colab Notebooks/Copy of States_Annual_Energy_Generation_Sources_1990_2019.xlsx'
energy_generation =pd.read_excel(e)
re = '/content/drive/MyDrive/Colab Notebooks/Copy of States_All_Vehicle_Registrations_2018.xlsx'
all_vehicle_registration =pd.read_excel(re)
```

I've mounted my Google Drive using drive.mount ('/content/drive') to access the datasets stored in my Drive. Using the Pandas library and the pd.read_excel() function, I've loaded four Excel files into separate variables: state_code for the state codes, registration_EV for electric vehicle registration in 2018, energy_generation for annual energy production from 1990 to 2019, and all_vehicle_registration for total vehicle registrations in 2018.

2 .Data Cleaning and Preprocessing

```
registration_EV.drop(["Unnamed: 0","Unnamed: 3","Unnamed: 4"],axis=1,inplace=True)
```

```
registration_EV.drop([0,1,53],axis=0,inplace=True)
```

```
registration_EV.rename(columns={'Unnamed: 1':'State','Unnamed: 2':'Registration Count'},inplace=True)
```

```
registration_EV.sort_values(by='Registration Count').reset_index().drop('index',axis=1)
```

I cleaned and organized all four datasets to prepare them for analysis. For the EV registration data, I removed unnecessary columns and rows, renamed columns for clarity, and sorted the data by registration count. Similar cleaning steps were applied to the other datasets to ensure consistency and accuracy.

3. Data Snapshot

```
energy_generation.head()
```

	YEAR	STATE	TYPE OF PRODUCER	ENERGY SOURCE	Generation(MWh)
1	1990	AK	Total Electric Power Industry	Total	5599506
2	1990	AK	Total Electric Power Industry	Coal	510573
3	1990	AK	Total Electric Power Industry	Hydroelectric Conventional	974521
4	1990	AK	Total Electric Power Industry	Natural Gas	3466261
5	1990	AK	Total Electric Power Industry	Petroleum	497116

I used energy_generation.head() to view the first five rows of the dataset, allowing me to quickly inspect its structure, columns, and initial values for better understanding before analysis.

Explanation of Columns:

1. **state** – The name of the U.S. state where the energy was generated.
2. **year** – The year of energy generation (ranging from 1990 to 2019).
3. **energy_source** – The type of energy source (e.g., Solar, Wind, Hydro, Coal, etc.).
4. **generation_(mwh)** – The total amount of electricity generated in megawatt-hours (MWh) for that year, state, and energy source.

4.Dataset Dimensions

```
energy_generation.shape
(53756, 5)
```

I used energy_generation.shape to view the number of records and features in the energy generation dataset helping understand its size and structure

5. Data Info

```
energy_generation.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53756 entries, 1 to 53756
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   YEAR              53756 non-null  object
 1   STATE             53756 non-null  object
 2   TYPE OF PRODUCER  53756 non-null  object
 3   ENERGY SOURCE     53756 non-null  object
 4   Generation(MWh)   53756 non-null  object
dtypes: object(5)
```

I used energy_generation.info() to get a concise summary of the dataset, including the number of entries, column names, non-null counts, and data types. This helps identify missing values and understand the structure of the data.

6.Descriptive Statistics

```
energy_generation.describe()
```

	Generation(MWh)	GENERATION (GWh)
count	5.375600e+04	5.375600e+04
mean	1.693131e+07	1.693085e+04
std	1.309890e+08	1.309890e+05
min	-8.823445e+06	-8.823000e+03
25%	2.684825e+04	2.600000e+01
50%	3.279665e+05	3.270000e+02
75%	3.405300e+06	3.405000e+03
max	4.178277e+09	4.178277e+06

I used energy_generation.describe() to view summary statistics like mean, min, max, and standard deviation, helping understand the distribution of energy generation values.

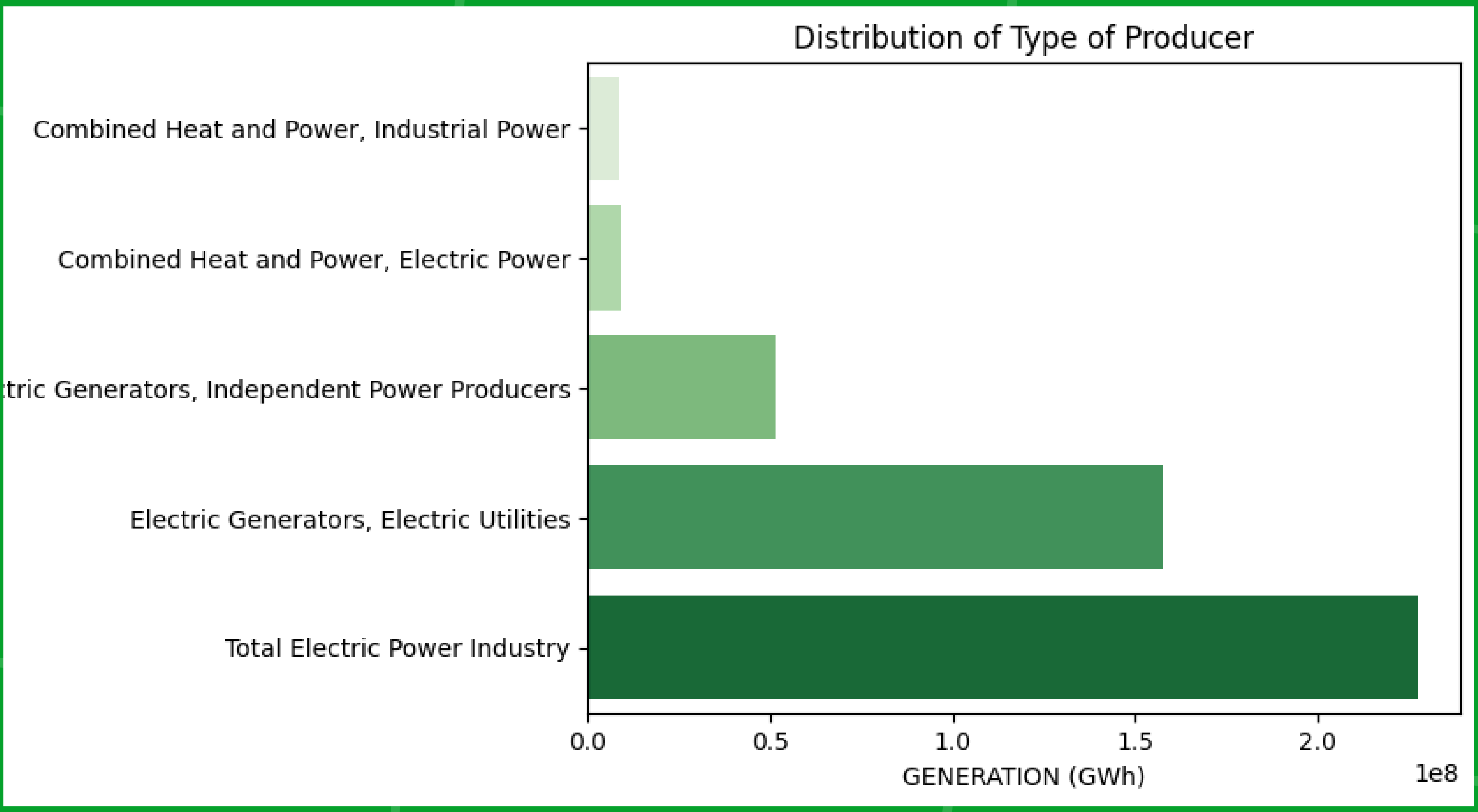
7. Data Type Conversion

```
energy_generation['Generation(MWh)']= energy_generation['Generation(MWh)'].astype('int')
energy_generation['GENERATION (GWh)']= energy_generation['GENERATION (GWh)'].astype('int')
```

I converted the 'Generation(MWh)' and 'GENERATION (GWh)' columns to integers using .astype('int') to ensure consistent and accurate data types for numerical analysis.

DATA VISUALIZATION

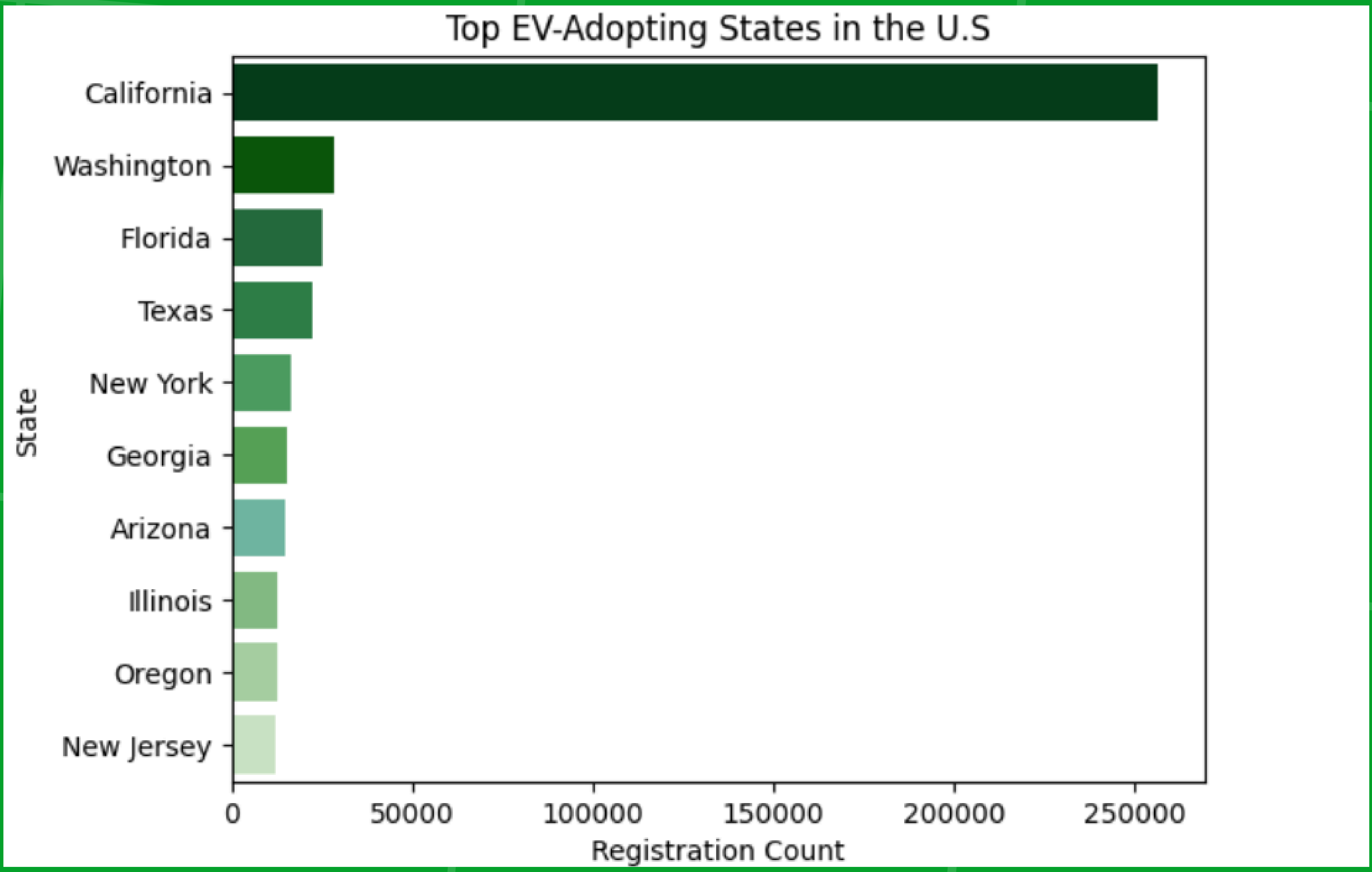
1.Producer-wise Total Energy Generation



I removed rows labeled 'Total' from the 'ENERGY SOURCE' column to focus on actual energy sources. Then, I grouped the data by 'TYPE OF PRODUCER' and calculated the total energy generated in megawatt-hours (MWh) using .sum(). I sorted the results from lowest to highest and created a horizontal bar chart with sns.barplot() to show which producer types generate the most energy.

2.Top EV-Adopting States in the U.S

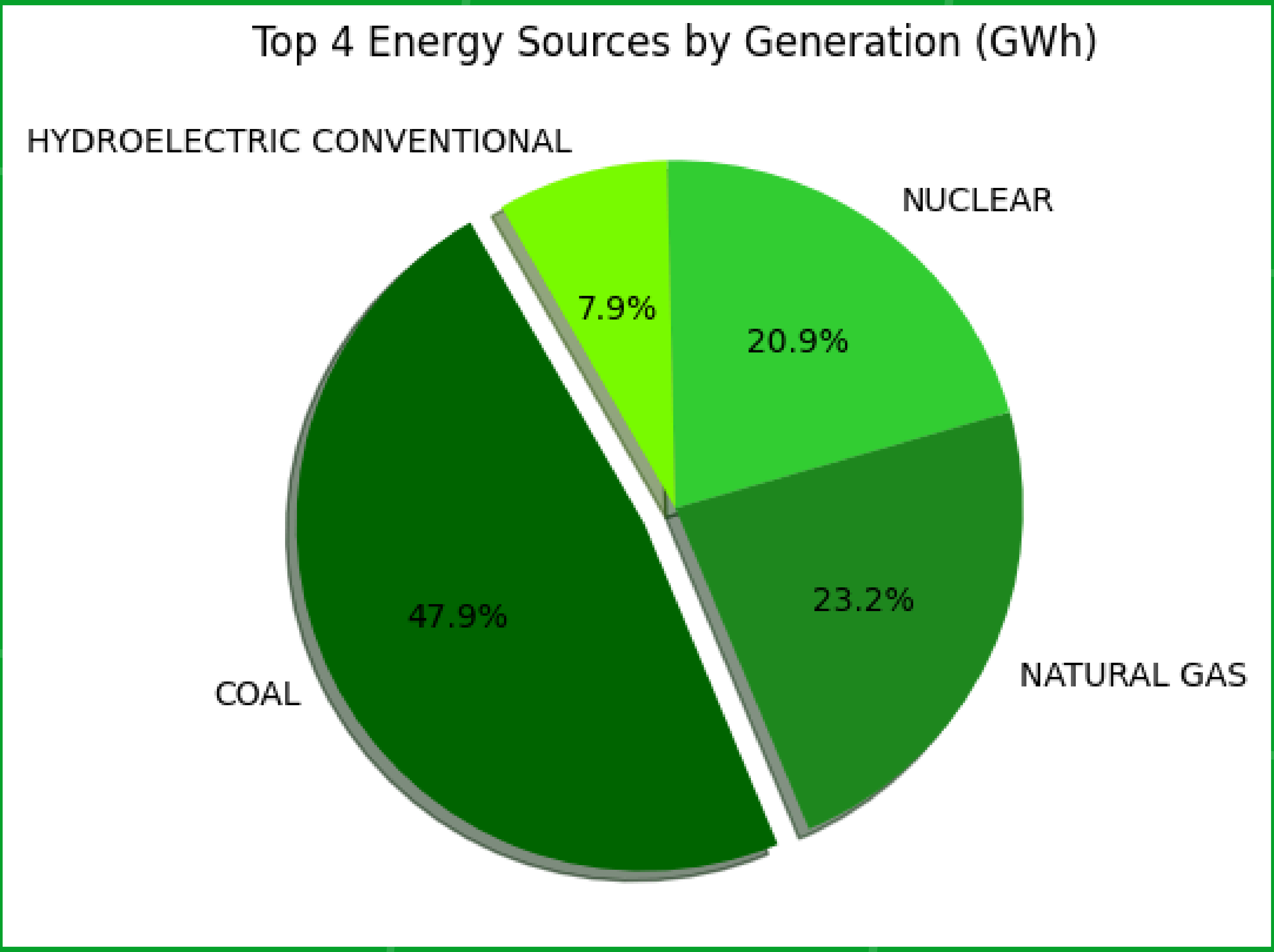
```
green_colors = ['#00441b','#006400','#1b7837','#238b45','#41ab5d','#4daf4a','#66c2a4','#78c679','#a1d99b','#c7e9c0' ]
sns.barplot(data=registred_EV.head(10), x='Registration Count', y='State',palette=green_colors,orient='h')
plt.title('Top EV-Adopting States in the U.S')
plt.show()
```



This chart shows the top 10 U.S. states with the highest EV registrations in 2018. California leads by a wide margin, followed by states like New York and Florida, highlighting where EV adoption is most concentrated.

3.Top 4 Energy Sources by Generation (GWh)

```python
labels=(['COAL','NATURAL GAS','NUCLEAR','HYDROELECTRIC CONVENTIONAL'])
myexplode=([0.1,0,0,0])
green_colors = ['#006400', '#228B22', '#32CD32', '#7CFC00']
plt.pie(Energy_Source_Generation['GENERATION (GWh)'].head(4),colors=green_colors,labels=labels,explode=myexplode,shadow=True,
        startangle=120,autopct='%1.1f%%')
plt.title("Top 4 Energy Sources by Generation (GWh)")
plt.show()
```

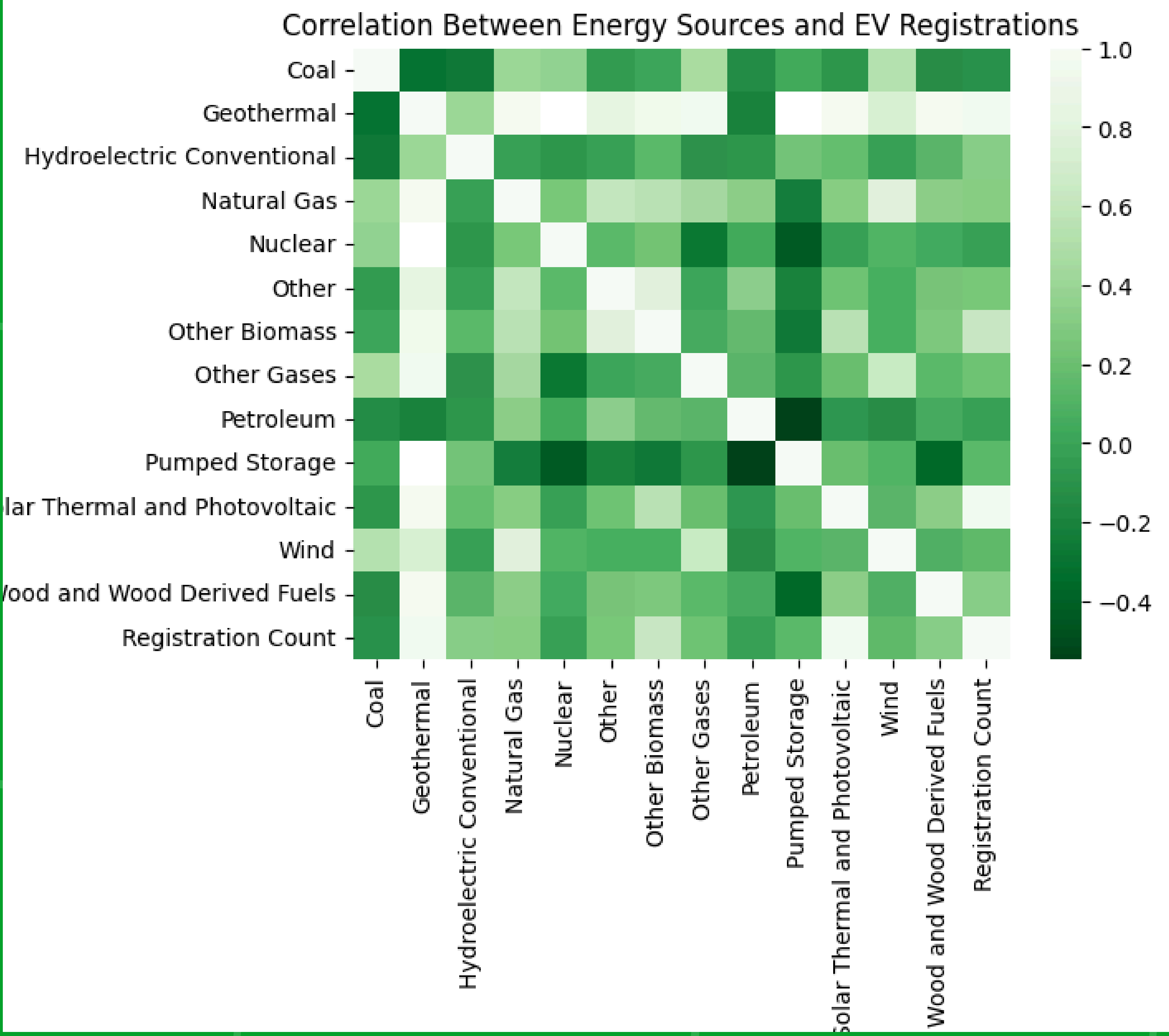


I calculated the total energy generation (in GWh) for each source and selected the top 4. Using plt.pie(), I created a pie chart to show their generation share, highlighting COAL with an explode effect, adding shadow for depth, and setting a start angle of 120° for better visual balance.

4.Correlation Between Energy Sources and EV Registrations

```
[45] correlation=energy_production_by_registration.corr(numeric_only=True)
     correlation.fillna(0)
```

```
sns.heatmap(correlation,cmap='Greens_r')
```

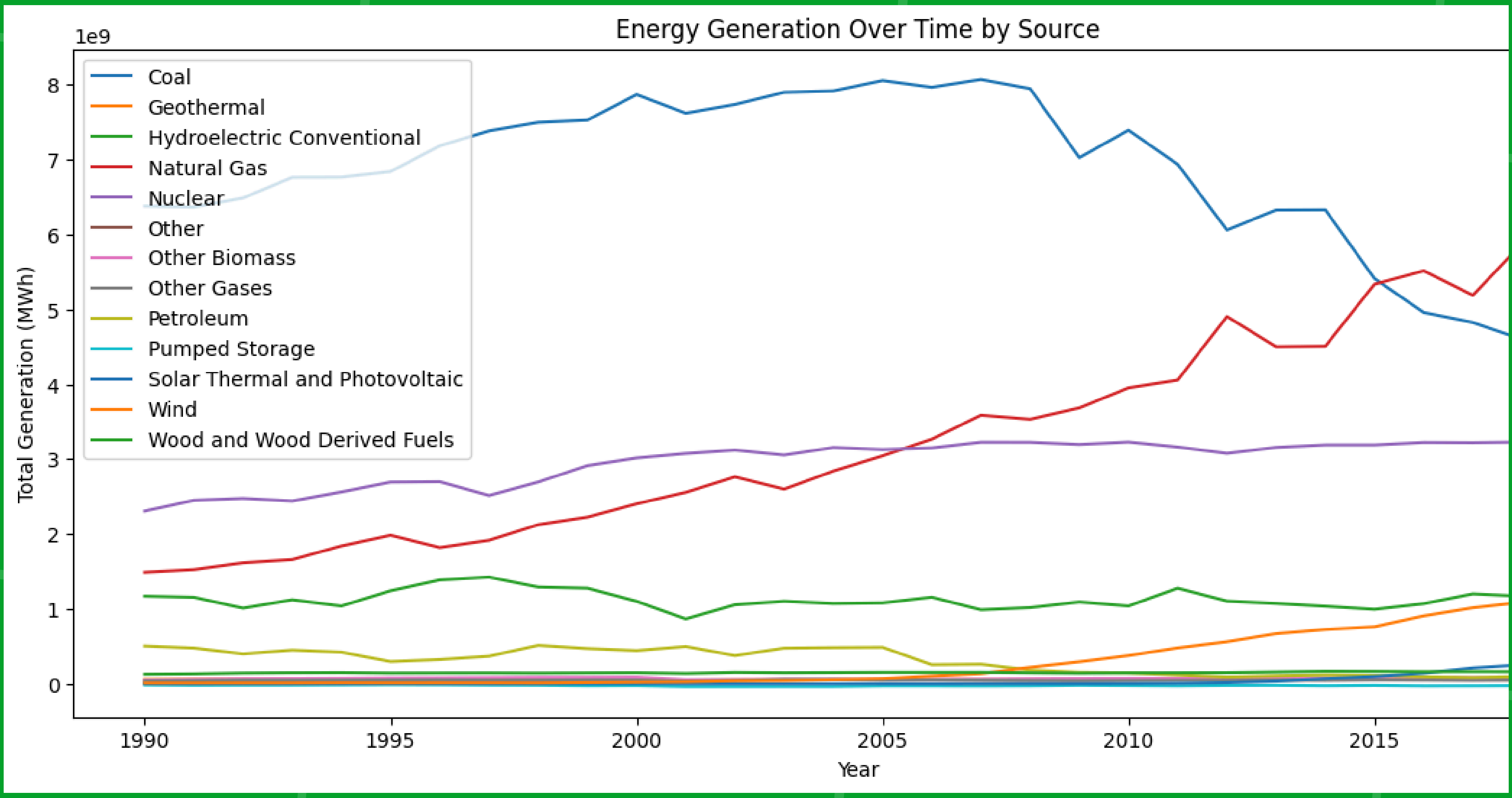


I filtered the data for 2018, calculated total energy generation by state and source, and reshaped it into a pivot table. Then, I merged it with EV registration data using state codes. Finally, I computed and visualized a correlation matrix to explore relationships between energy sources and EV registrations.

5. Trend of Energy Generation by Source (1990–2019)

```
[47] energy_ts = energy_generation[energy_generation['ENERGY SOURCE'] != 'Total']
     ts_energy = energy_ts.groupby(['YEAR', 'ENERGY SOURCE'])['Generation(MWh)'].sum().reset_index()
     ts_pivot_table = ts_energy.pivot(index='YEAR', columns='ENERGY SOURCE', values='Generation(MWh)')
```

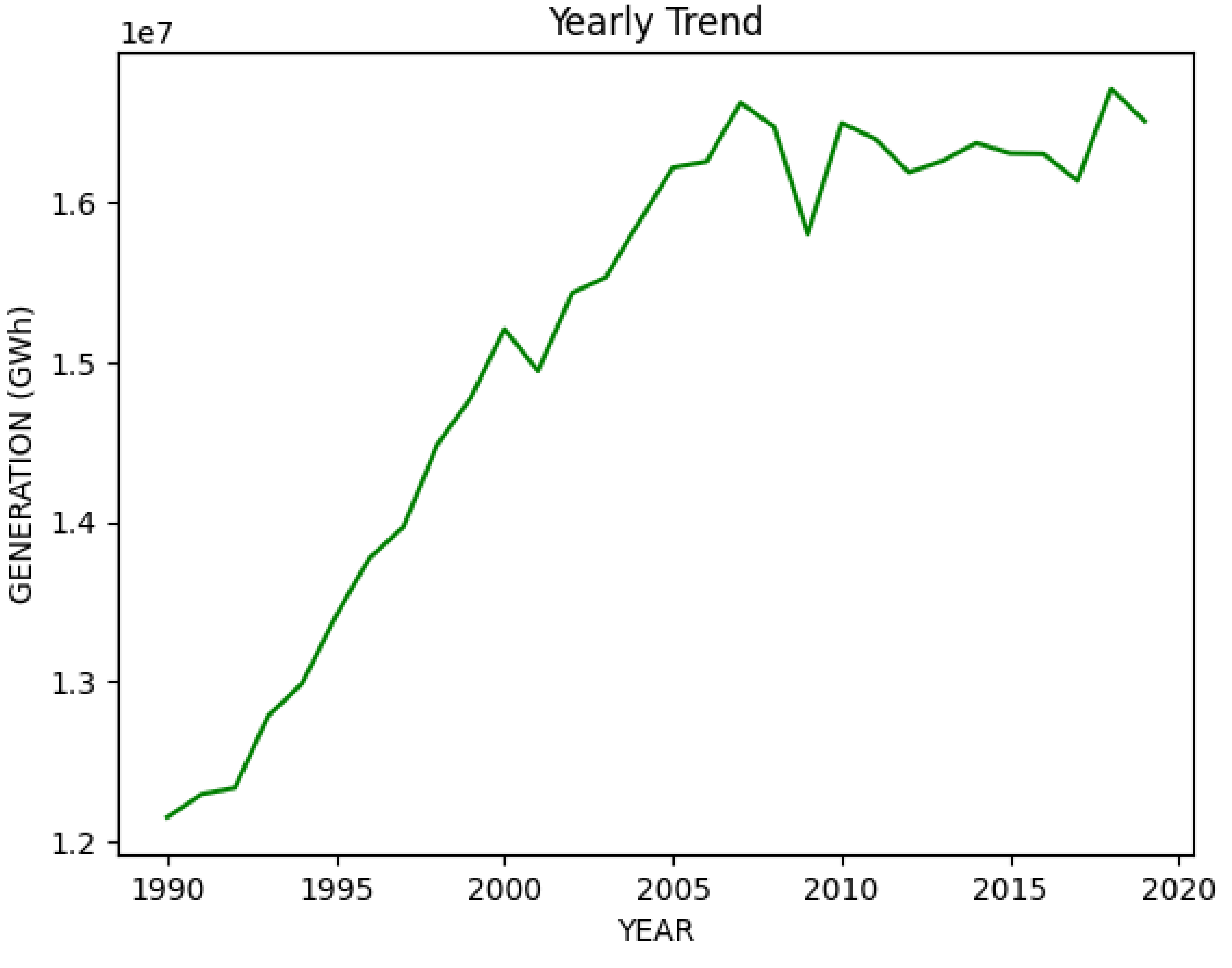
```
ts_pivot_table.plot(figsize=(14, 6))
plt.title('Energy Generation Over Time by Source')
plt.xlabel('Year')
plt.ylabel('Total Generation (MWh)')
plt.legend()
plt.show()
```



I filtered out the 'Total' entries to focus on specific energy sources, then grouped the data by year and energy source to calculate total generation in MWh. I reshaped it into a pivot table and plotted a multi-line time series chart to show how energy generation from each source has changed over time.

6.Yearly Trend

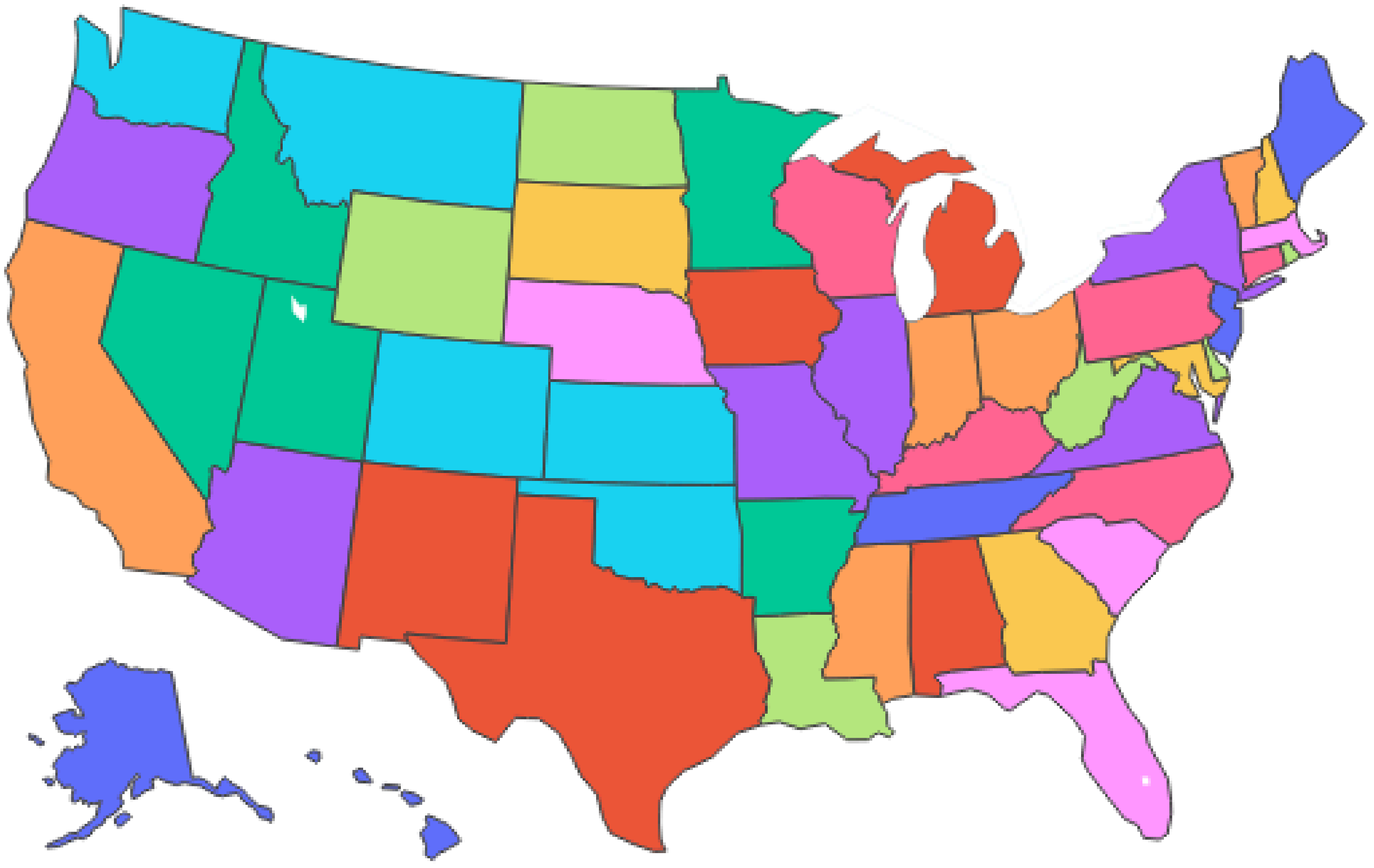
```
yearly_trend = Energy_Source.groupby('YEAR')['GENERATION (GWh)'].sum().reset_index()
sns.lineplot(data=yearly_trend,x='YEAR',y='GENERATION (GWh)',color='green')
plt.title('Yearly Trend')
plt.show()
```



I grouped the energy dataset by year and calculated the total energy generation in GWh for each year. Then, I used sns.lineplot() to plot a green line chart showing the overall trend in energy production over time.

7.EV Registrations Across U.S. States (2018)

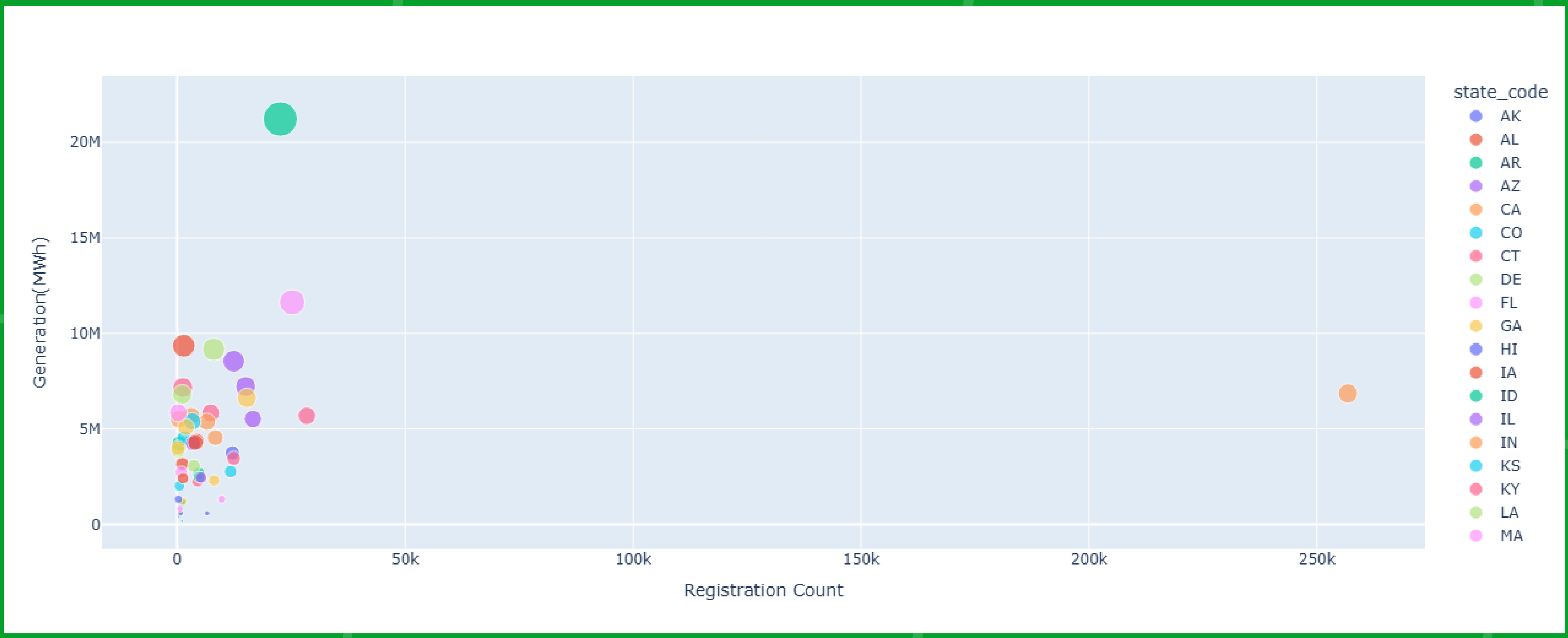
```
GEOSPATIAL_VISUALIZATION = px.choropleth(
    merge_of_states,
    locations='state_code',
    locationmode='USA-states',
    color='Registration Count',
    scope='usa',
    title='EV Registrations by State (2018)')
GEOSPATIAL_VISUALIZATION.show()
```



I used px.choropleth() to create a U.S. map showing the number of electric vehicle registrations in each state for 2018. States are colored based on their registration count, helping to visually compare EV adoption across the country.

8.DISTRIBUTION WITH THE LOCATIONS OF RENEWABLE ENERGY

```
[52] fig = px.scatter(t,x='Registration Count',y='Generation(MWh)',color='state_code',size='Generation(MWh)')
     title='DISTRIBUTIONS WITH THE LOCATIONS OF RENEWABLE ENERGY '
     fig.show()
```



I merged EV registration data with 2018 energy generation data by state, cleaned unnecessary columns, and created a pivot table summarizing total energy generation and EV registrations per state. Then, I used px.scatter() to plot a bubble chart showing the relationship between EV registrations and energy generation for each state. Each bubble's size and color represent the amount of energy generated and the state code, respectively.

ANALYTICAL FINDINGS

1. Natural Gas, Nuclear, and Hydroelectric are Major Energy Contributors : These energy sources lead in total generation, highlighting their role in powering U.S. electricity demand.

2. Correlation Analysis Reveals Moderate Links : A moderate positive correlation was observed between renewable energy sources (like wind and solar) and EV registrations, but it's not uniform across all states.

3. States Dependent on Coal Show Lower EV Adoption: High fossil-fuel states tend to have lower EV penetration, possibly due to policy gaps, infrastructure limits, or lower environmental prioritization.

4.Western and Northeastern States Lead in EV Adoption: Geospatial analysis highlights regional leadership in EV use, likely due to state incentives, infrastructure, and environmental awareness.

5.Geospatial Maps Show Uneven EV Adoption Across States: EV registration density is concentrated in a few coastal and urban states, while central and rural states lag behind.

RECOMMENDATIONS

1.Target EV Awareness Campaigns in Low-Adoption States: Focus educational efforts in regions where both renewable energy and EV use are low to accelerate clean transitions.

2.Incentivize Industrial and Private Producers to Adopt Renewables: Expand grants or tax credits for producers (not just consumers) who shift toward cleaner energy portfolios.

3. Incentivize Solar and Wind Expansion in High-Registration States: States with growing EV use should prioritize green electricity sources to make EV adoption more sustainable.

4.Integrate Clean Energy Into Transportation Planning: States should align transportation electrification with clean energy production goals to maximize carbon reduction impact.

5. Urban Centers Should Lead EV-Grid Integration : Urban areas with dense EV use should explore smart grid and vehicle-to-grid (V2G) programs to stabilize electricity demand.

CONCLUSION

This capstone project analyzed the relationship between electric vehicle (EV) registrations and energy production across the U.S., focusing on 2018 data within a 1990–2019 range. Through exploratory data analysis (EDA), we identified key patterns linking clean energy growth with EV adoption. While natural gas, coal, and nuclear remain dominant energy sources, renewable sources like wind and solar have shown notable growth. States with higher renewable energy generation also tend to have more EV registrations, suggesting a positive link between clean energy development and the transition to electric mobility.

EV adoption is highest in coastal states with strong policies and infrastructure, while central, fossil-fuel-dependent states lag behind. A moderate positive correlation exists between renewable energy generation and EV use, though factors like population and policy also influence this relationship.
In summary, the EDA process provided valuable insights into the state-wise distribution of energy production and EV registrations. It highlighted both the progress and disparities in the transition to sustainable energy and mobility in the U.S., setting a strong foundation for future predictive modeling and policy impact evaluation.

THANK YOU

FOR VIEWING OR CODES

https://colab.research.google.com/drive/1pVhC4AyRnF7-FYHuuI0pukCGxqSLxKrt#scrollTo=5nsZFGbXTlUC