Deliverable 2

Group-10

Brijesh Patel Sai Reddy Busi Reddy Chandra Manohar Reddy Kusam
Sanjana Gedela Koushik Vijay Kulkarni

Dataset:

The dataset includes energy generation records for all states in the United States of America, collected monthly by various types of energy producers from different non-renewable and renewable energy sources between 2001 and 2022. It contains columns like year and month of production, state in which the source was generated, type of energy producer, energy source, and energy generated in megawatt hours (MWh).

Source: [US Energy Generation 2001-2022 | Kaggle] (organised_Gen.csv) adapted from [Electricity - U.S. Energy Information Administration (EIA)]

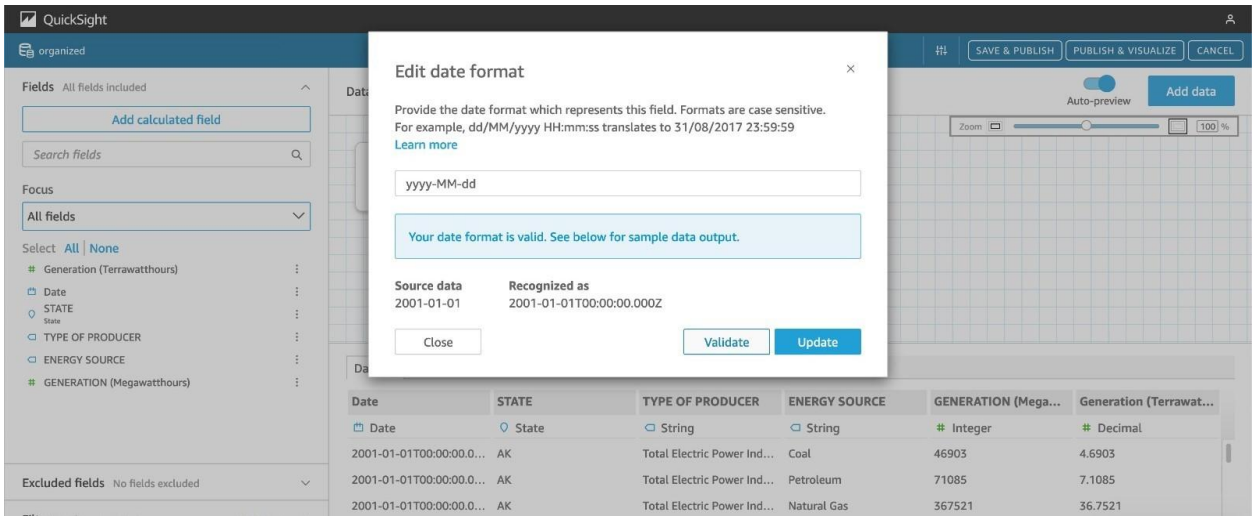
Tools used for Data Analysis:

Amazon S3, Amazon QuickSight, and Amazon Sagemaker

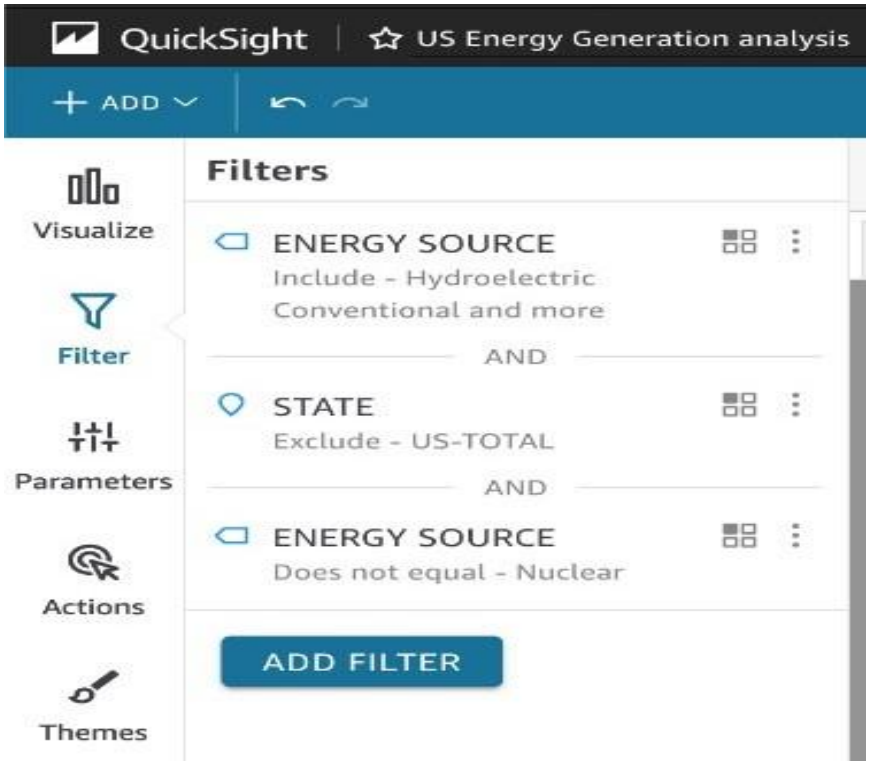
Data Preparation:

- organised_Gen.csv contains a few negative values in the Generation (MWh) column that add no value in analysis, hence those records are cleaned.
- As part of the analysis, calculated fields are generated for the Generation column, which may contain high values, so to improve readability, a new column is created with energy measured in 1,000,000 MWh, which is Generation in TerraWatt hours (TWh).

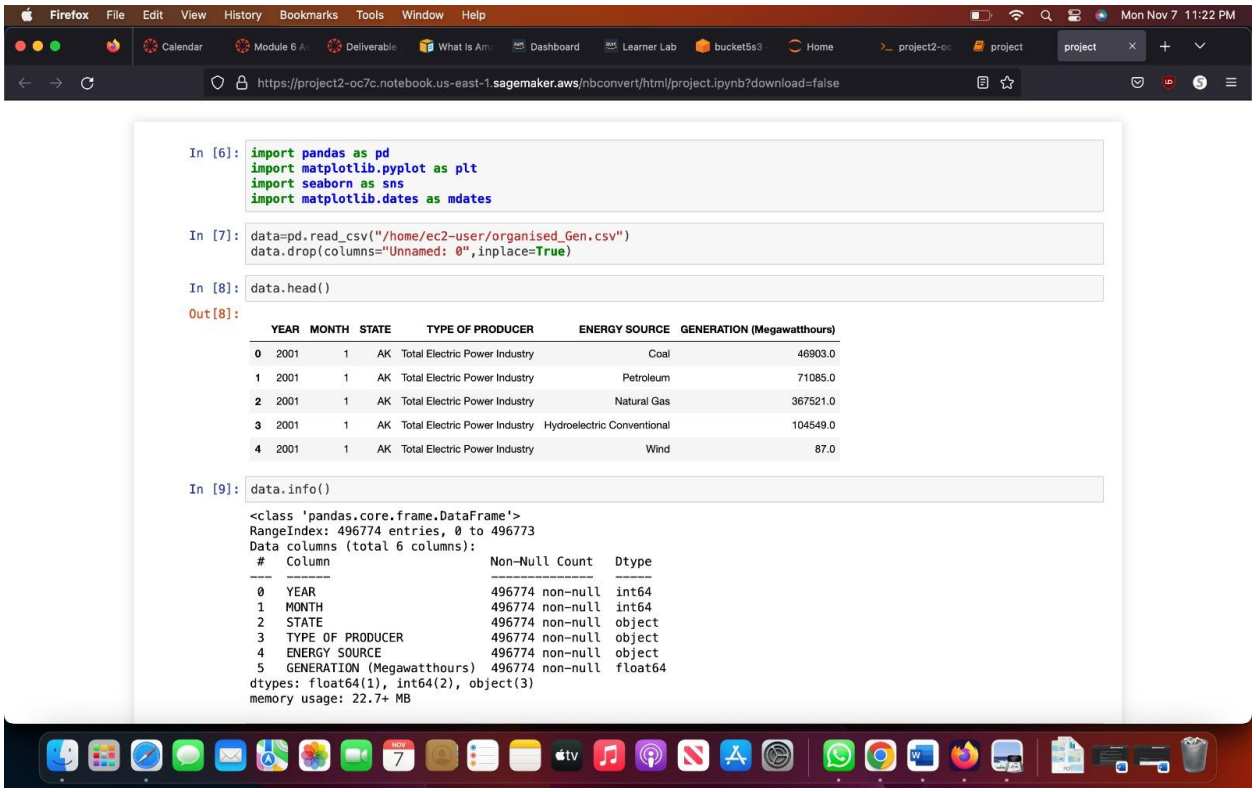
- Month and year columns are grouped as one Date field for filtering data.



- Filters are added for State, Type of producer and Generation fields to exclude records with “total” generation and for Energy source field is filtered to categorize as Renewable and Non-renewable which include/exclude nuclear.



Data Exploration:



In [6]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.dates as mdates
```

In [7]:
```
data=pd.read_csv("/home/ec2-user/organised_Gen.csv")
data.drop(columns="Unnamed: 0",inplace=True)
```

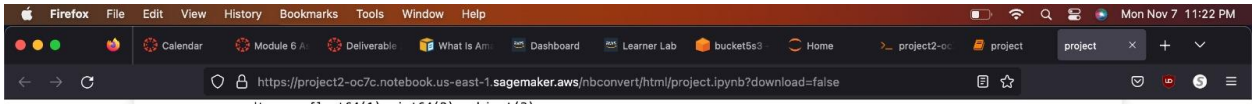
In [8]:
```
data.head()
```

Out[8]:

	YEAR	MONTH	STATE	TYPE OF PRODUCER	ENERGY SOURCE	GENERATION (Megawatthours)
0	2001	1	AK	Total Electric Power Industry	Coal	46903.0
1	2001	1	AK	Total Electric Power Industry	Petroleum	71085.0
2	2001	1	AK	Total Electric Power Industry	Natural Gas	367521.0
3	2001	1	AK	Total Electric Power Industry	Hydroelectric Conventional	104549.0
4	2001	1	AK	Total Electric Power Industry	Wind	87.0

In [9]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 496774 entries, 0 to 496773
Data columns (total 6 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   YEAR                          496774 non-null  int64
 1   MONTH                         496774 non-null  int64
 2   STATE                         496774 non-null  object
 3   TYPE OF PRODUCER              496774 non-null  object
 4   ENERGY SOURCE                 496774 non-null  object
 5   GENERATION (Megawatthours)    496774 non-null  float64
dtypes: float64(1), int64(2), object(3)
memory usage: 22.7+ MB
```

```
dtypes: float64(1), int64(2), object(3)
memory usage: 22.7+ MB
```

In [10]:
```
data.describe()
```

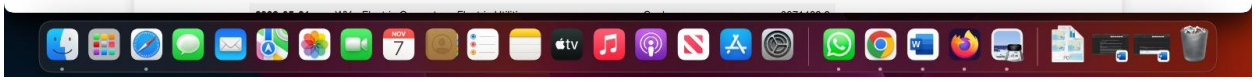
Out[10]:

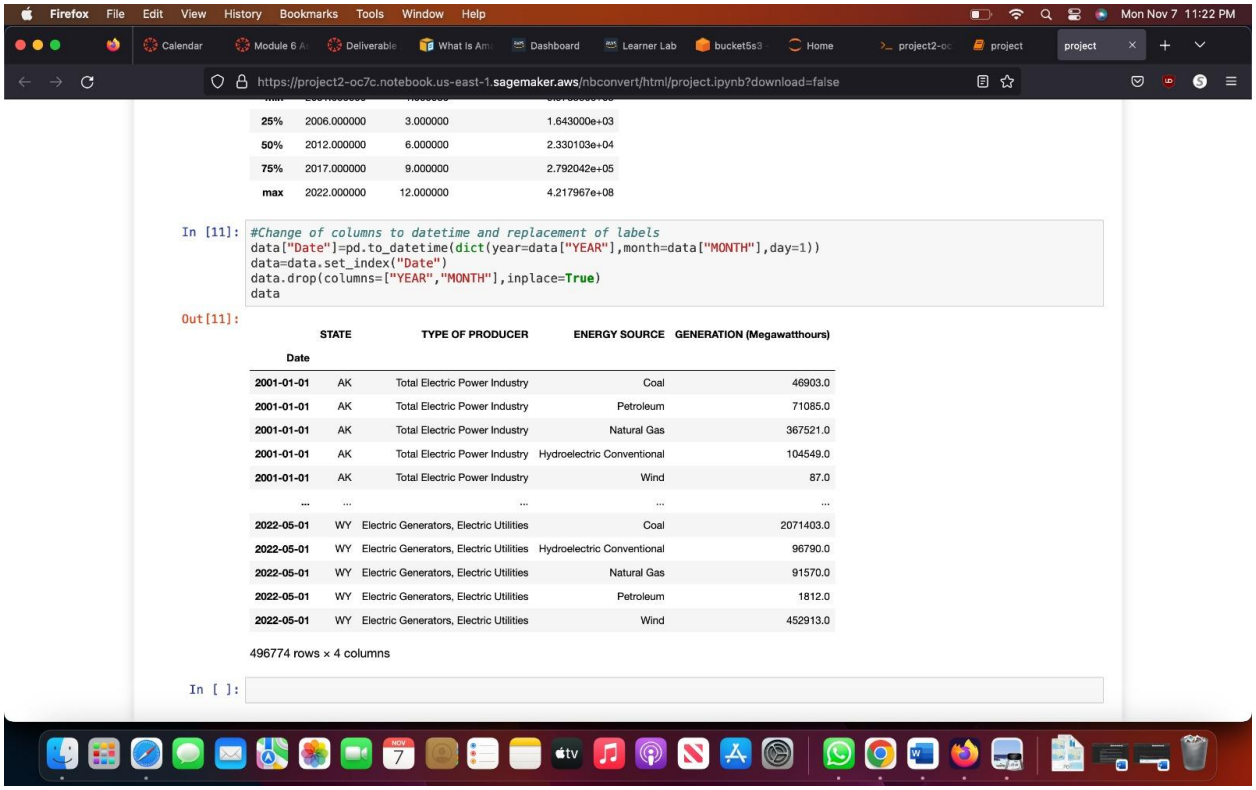
	YEAR	MONTH	GENERATION (Megawatthours)
count	496774.000000	496774.000000	4.967740e+05
mean	2011.649768	6.431452	1.393384e+06
std	6.168041	3.458944	1.078685e+07
min	2001.000000	1.000000	-9.978550e+05
25%	2006.000000	3.000000	1.643000e+03
50%	2012.000000	6.000000	2.330103e+04
75%	2017.000000	9.000000	2.792042e+05
max	2022.000000	12.000000	4.217967e+08

In [11]:
```
#Change of columns to datetime and replacement of labels
data["Date"]=pd.to_datetime(dict(year=data["YEAR"],month=data["MONTH"],day=1))
data=data.set_index("Date")
data.drop(columns=["YEAR","MONTH"],inplace=True)
data
```

Out[11]:

Date	STATE	TYPE OF PRODUCER	ENERGY SOURCE	GENERATION (Megawatthours)
2001-01-01	AK	Total Electric Power Industry	Coal	46903.0
2001-01-01	AK	Total Electric Power Industry	Petroleum	71085.0
2001-01-01	AK	Total Electric Power Industry	Natural Gas	367521.0
2001-01-01	AK	Total Electric Power Industry	Hydroelectric Conventional	104549.0
2001-01-01	AK	Total Electric Power Industry	Wind	87.0
...	...	...	...	...





25%	2006.000000	3.000000	1.643000e+03
50%	2012.000000	6.000000	2.330103e+04
75%	2017.000000	9.000000	2.792042e+05
max	2022.000000	12.000000	4.217967e+08

In [11]:
```
#Change of columns to datetime and replacement of labels
data["Date"]=pd.to_datetime(dict(year=data["YEAR"],month=data["MONTH"],day=1))
data=data.set_index("Date")
data.drop(columns=["YEAR","MONTH"],inplace=True)
data
```

Out[11]:

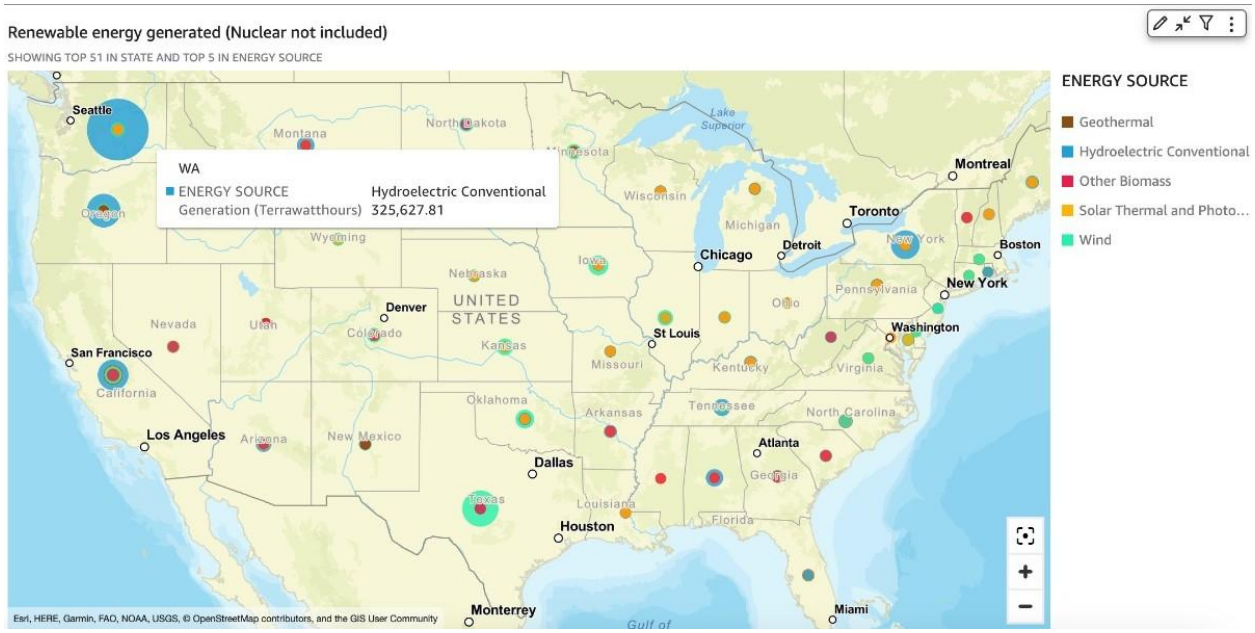
Date	STATE	TYPE OF PRODUCER	ENERGY SOURCE	GENERATION (Megawatthours)
2001-01-01	AK	Total Electric Power Industry	Coal	46903.0
2001-01-01	AK	Total Electric Power Industry	Petroleum	71085.0
2001-01-01	AK	Total Electric Power Industry	Natural Gas	367521.0
2001-01-01	AK	Total Electric Power Industry	Hydroelectric Conventional	104549.0
2001-01-01	AK	Total Electric Power Industry	Wind	87.0
...	...	...	...	...
2022-05-01	WY	Electric Generators, Electric Utilities	Coal	2071403.0
2022-05-01	WY	Electric Generators, Electric Utilities	Hydroelectric Conventional	96790.0
2022-05-01	WY	Electric Generators, Electric Utilities	Natural Gas	91570.0
2022-05-01	WY	Electric Generators, Electric Utilities	Petroleum	1812.0
2022-05-01	WY	Electric Generators, Electric Utilities	Wind	452913.0

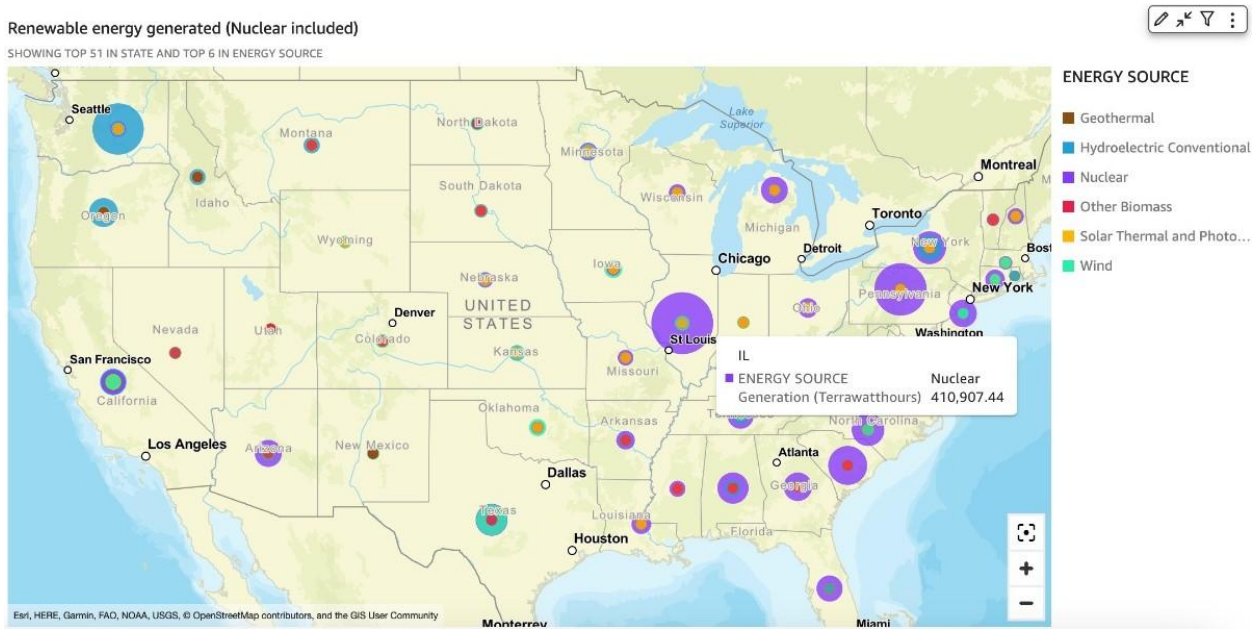
496774 rows × 4 columns

In [ ]:

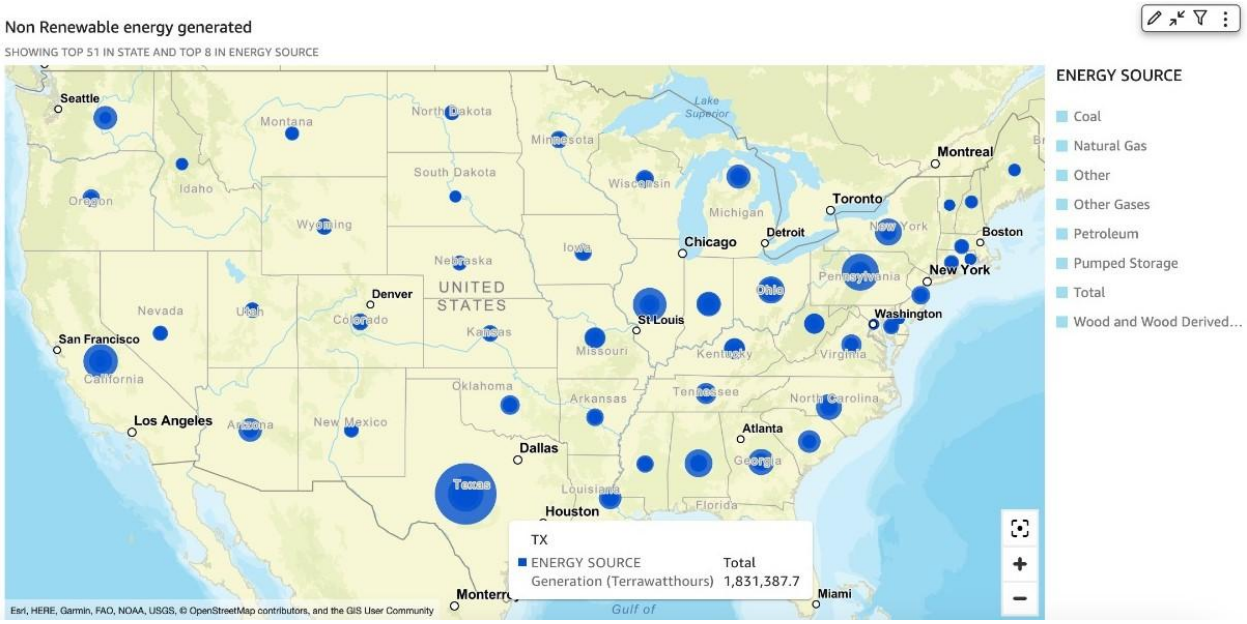
Data Visualization:

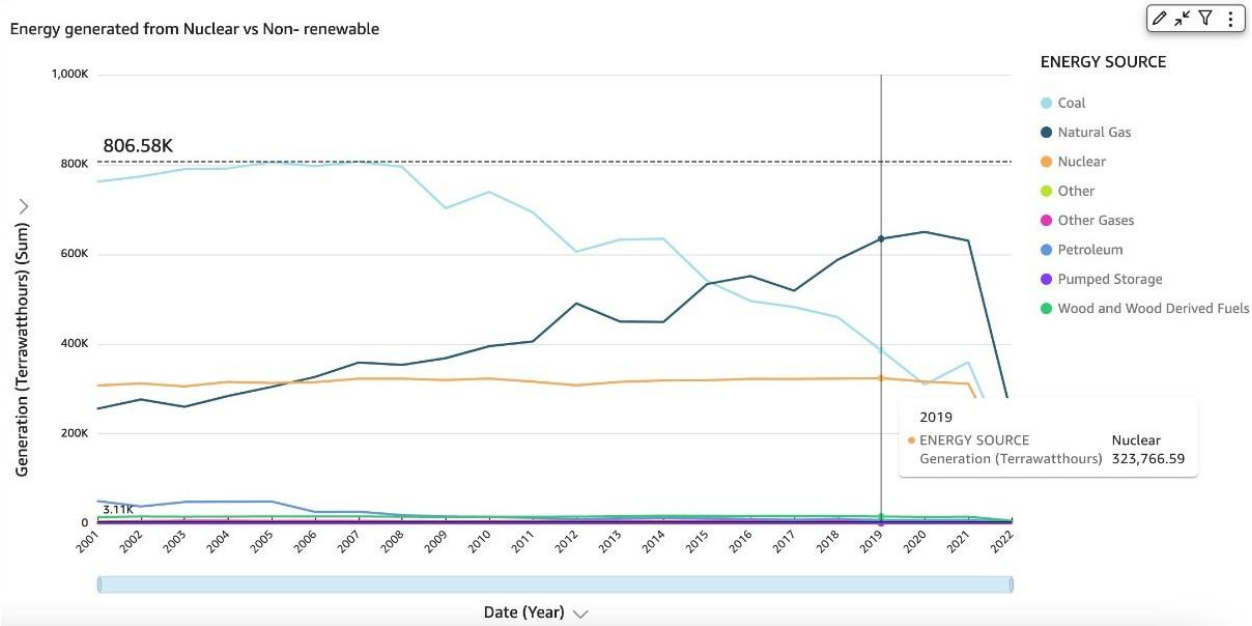
- Below, 2 maps show the renewable energy generated per state with or without nuclear sources included. Without a nuclear source included, Washington generates the most, and Illinois leads the energy generation with a nuclear source.



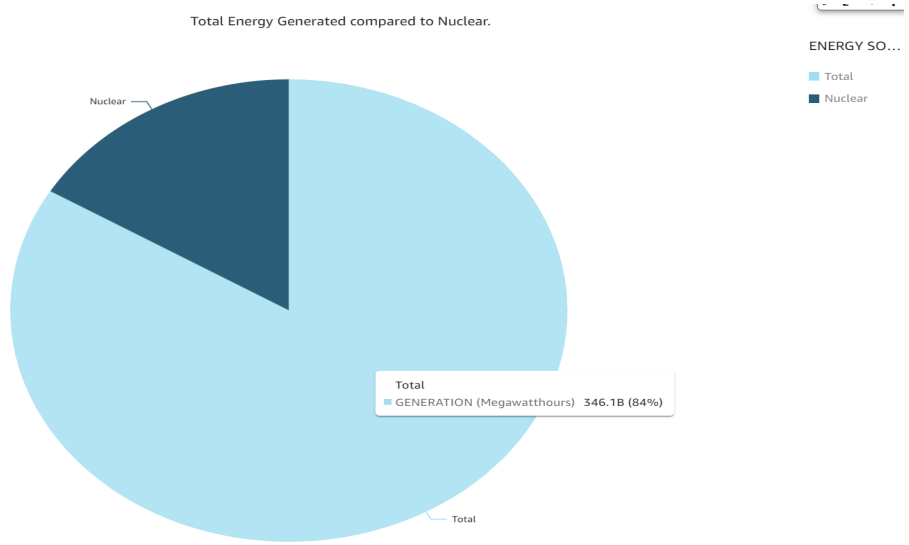


- Below, a filled and line graph show non-renewable energy generated per state and also comparisons between the energy generated throughout the past two decades from non-renewable and nuclear sources. Texas generates the most energy from non-renewable sources.

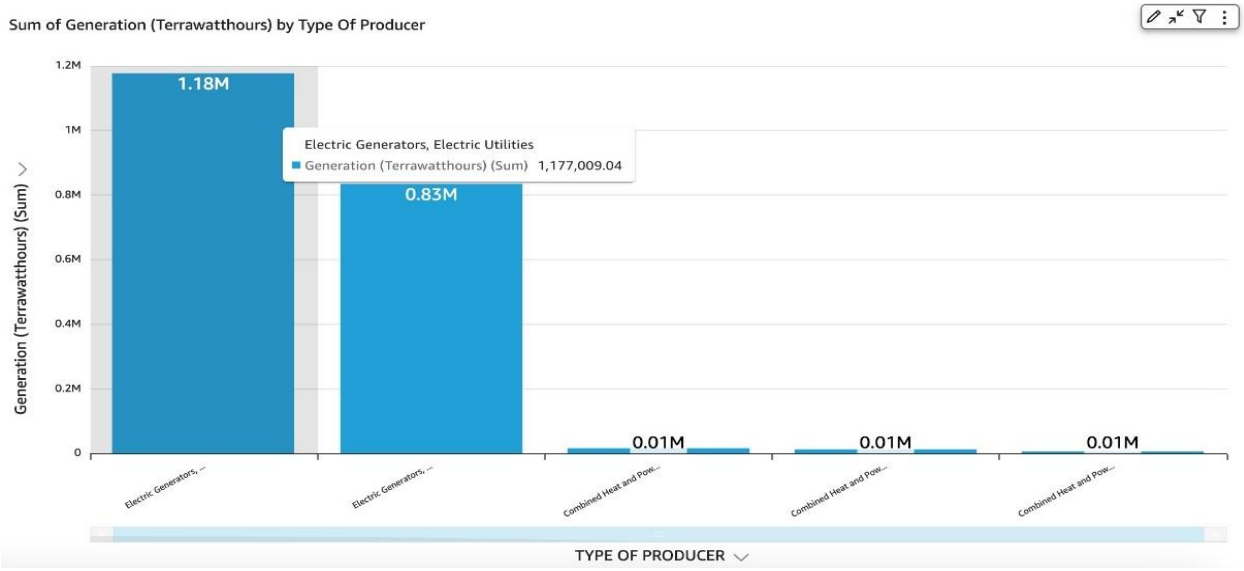


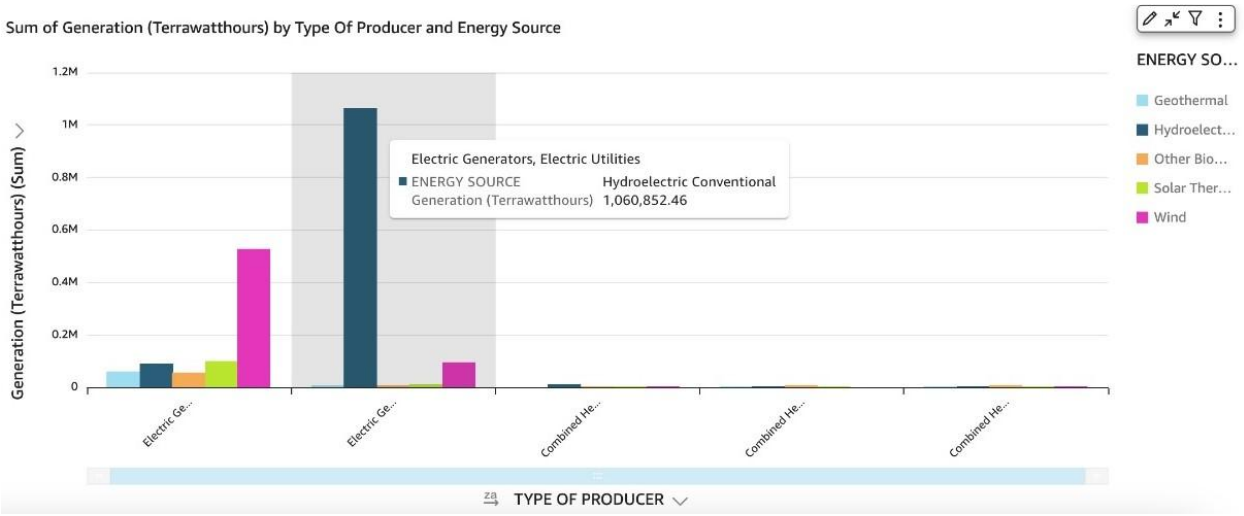


- The below pie chart draws a comparison between the energy generated from nuclear sources and the total energy from all sources.

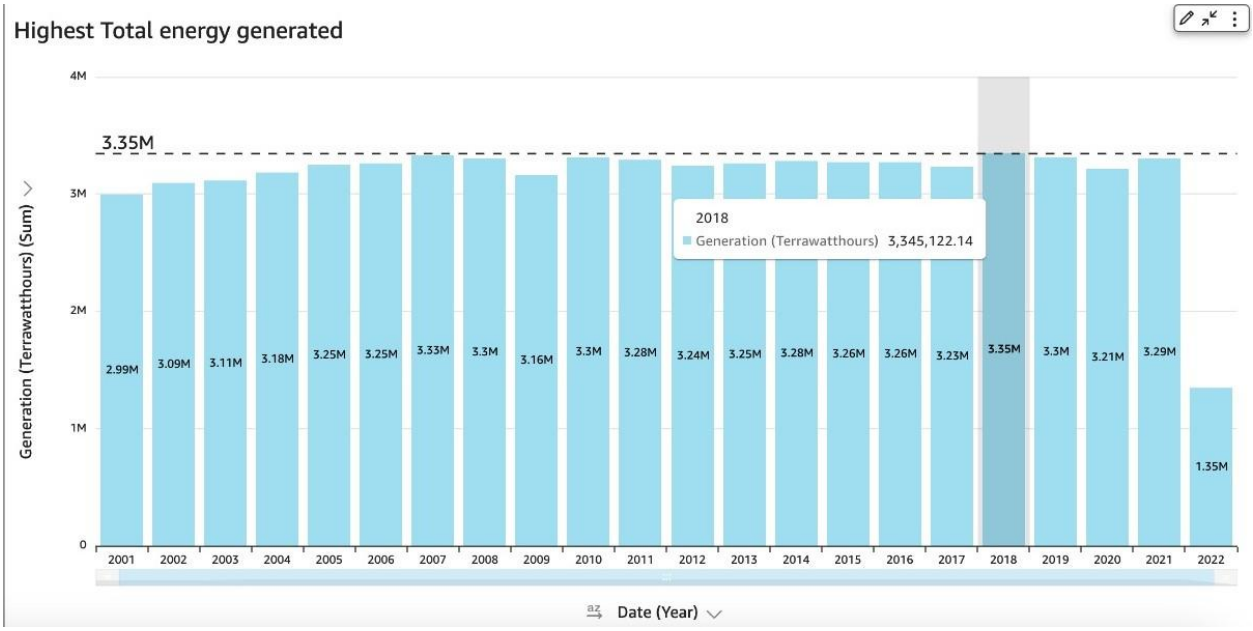


- Below, bar graph and clustered bar graphs describe energy generated from different types of producers, with energy generators and utilities being the highest with 1.18 million TWh of generation.

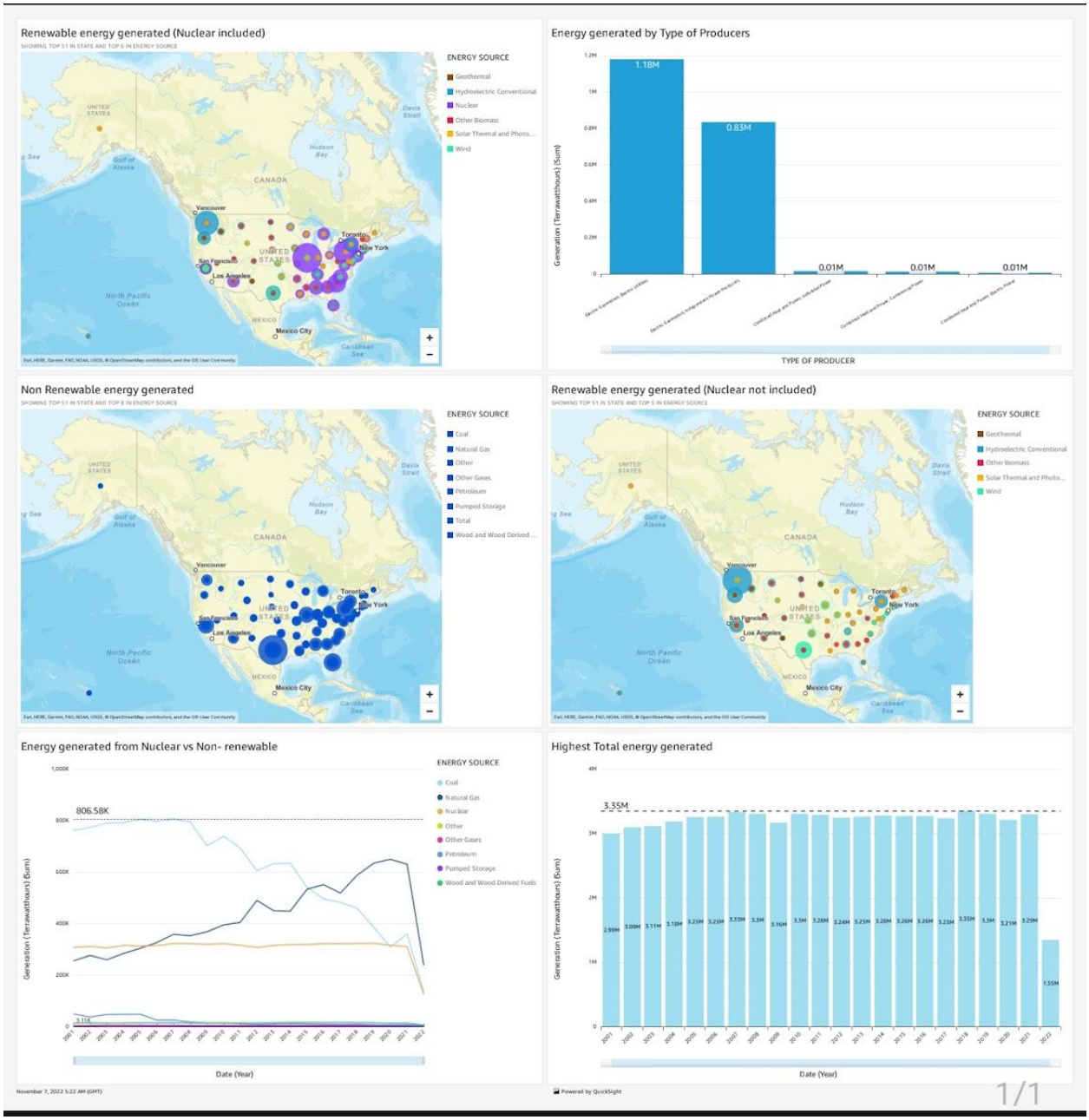




- Below Bar graph with maximum energy generated as a reference line describes total energy generated across 2 decades (2001–2022). In 2018, 3.35 million TWh of total energy generated was the highest.



Dashboard:



- The dashboard above was created in Amazon Quicksight and it shows the graphs that were generated based on the questions we have inb terms of the research objective. We used those visualizations to describe the dataset. It is showing some of the graphs from above.