

ECE 9014 Final Project

Deliverable 2

**US Electricity Information database**

Yuanpeng Wang – 251343797

Hongjin Zi - 251389742

Table of Contents

[Data Warehousing 1](#_Toc152247762)

[Data Cube Design 1](#_Toc152247763)

[Diagram 1](#_Toc152247764)

[SQL scripts 4](#_Toc152247765)

[Data Mining and Visualization 13](#_Toc152247766)

[Analysis and Charts 13](#_Toc152247767)

# Data Warehousing

### Data Cube Design

What is the “measure” you’re trying to analyze in your data cube?

**Production amount**. **Usage / Unit price / Revenue**

What are your dimensions?

**User - User Type**

**Area – Region – State**

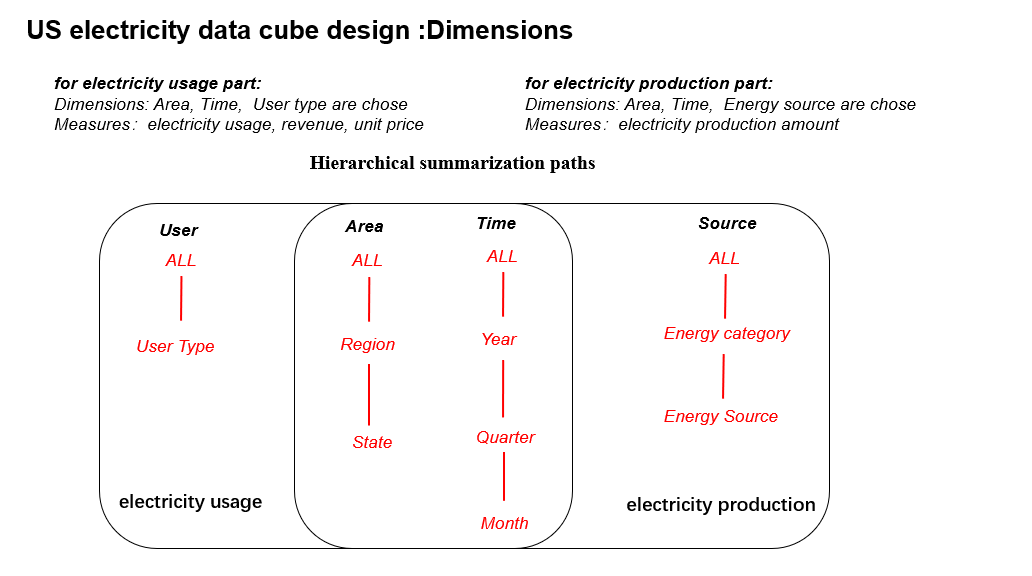
**Time – Year - Quarter - Month**

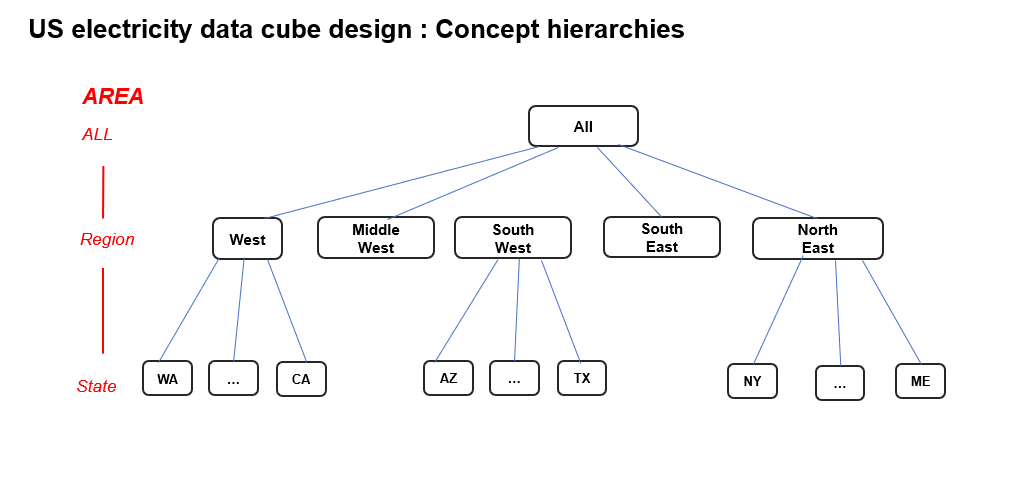
**Source - Energy category - Energy Source**

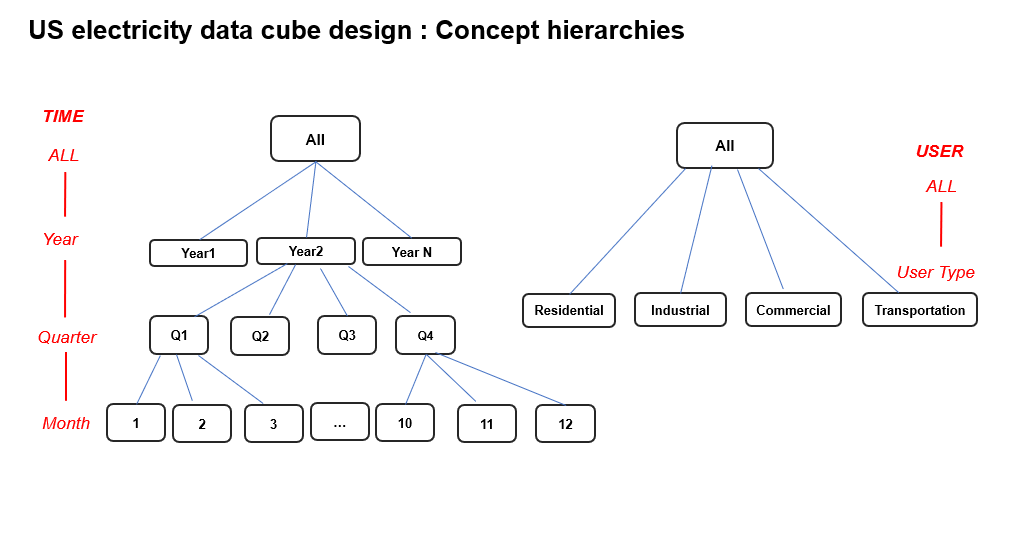
Why did you choose them?

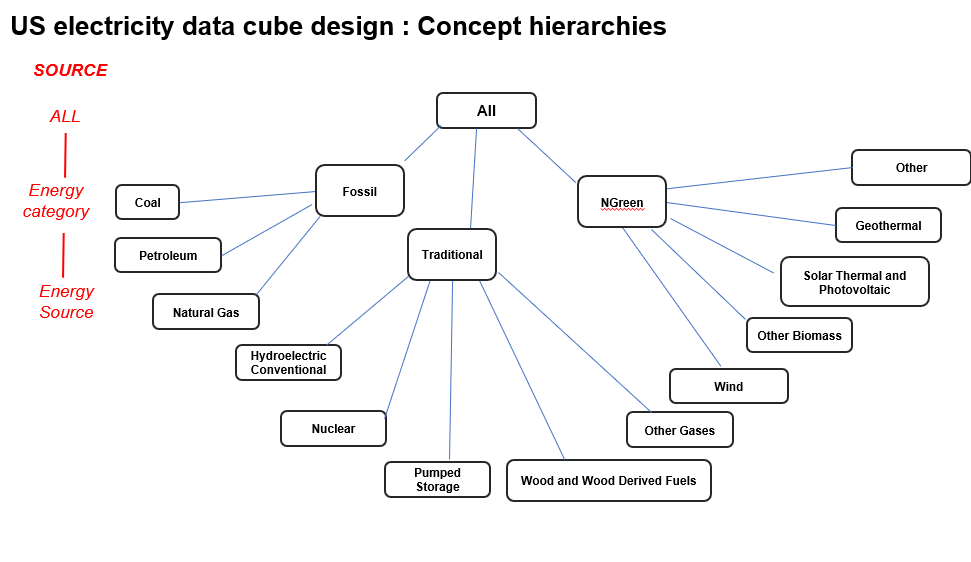
Target is to build data cube to support analysis the whole us electricity market from production and usage perspective, like market volume, user behaviors, usage trend prediction, region gap, energy source, and green source utilization percentage, … to support plant planning, expansion, regulating, policymaking.

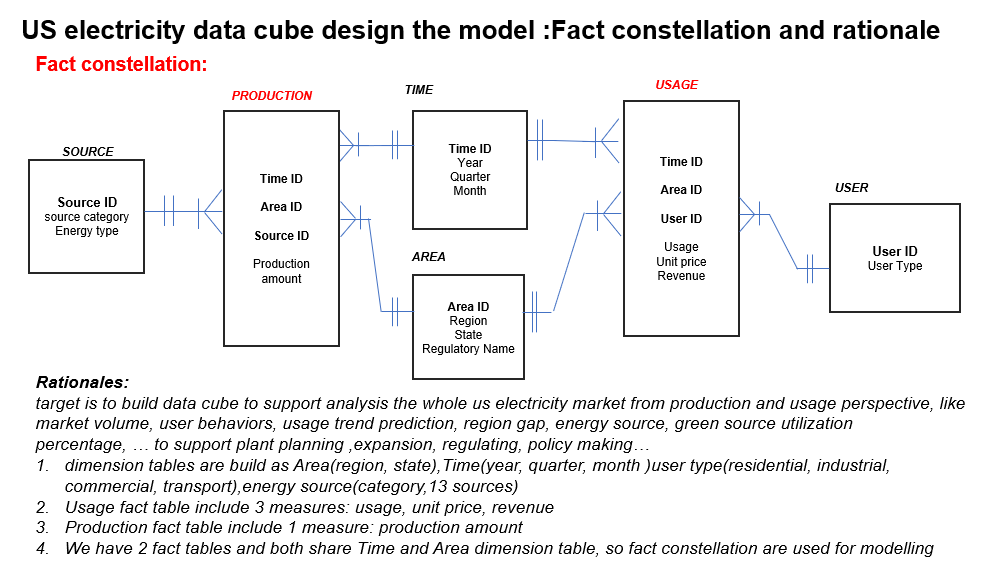
### Diagram

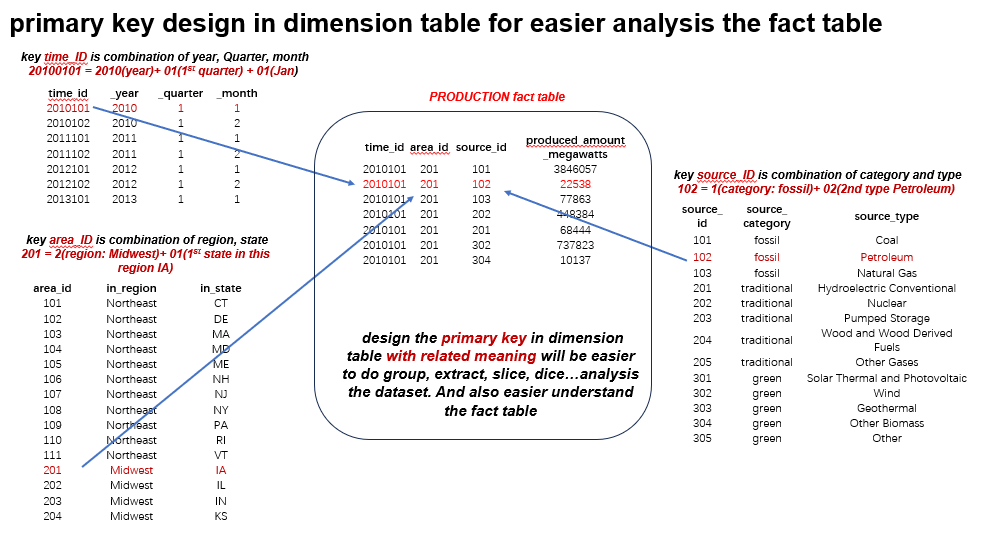


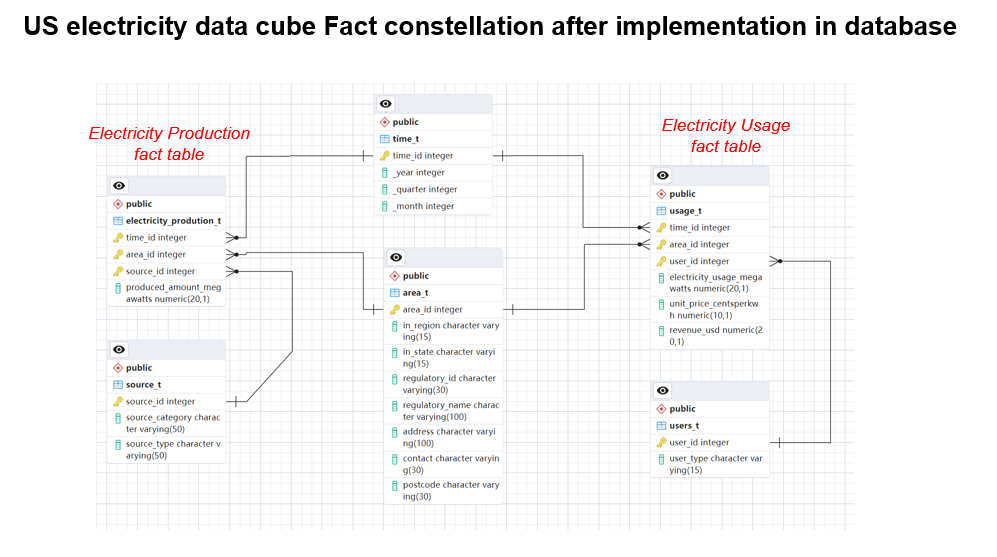












### SQL scripts

create 4 dimension tables and 2 fact tables, and also implement its data.

**Create dimension table 1 time\_t and import rows**

CREATE TABLE time\_t (

time\_ID INT,

\_year INT,

\_quarter INT,

\_month INT,

PRIMARY KEY (time\_ID));

COPY time\_t(time\_ID,\_year,\_quarter,\_month)

FROM 'D:\databaseSQL\time\_t.csv'

DELIMITER ','

CSV HEADER;

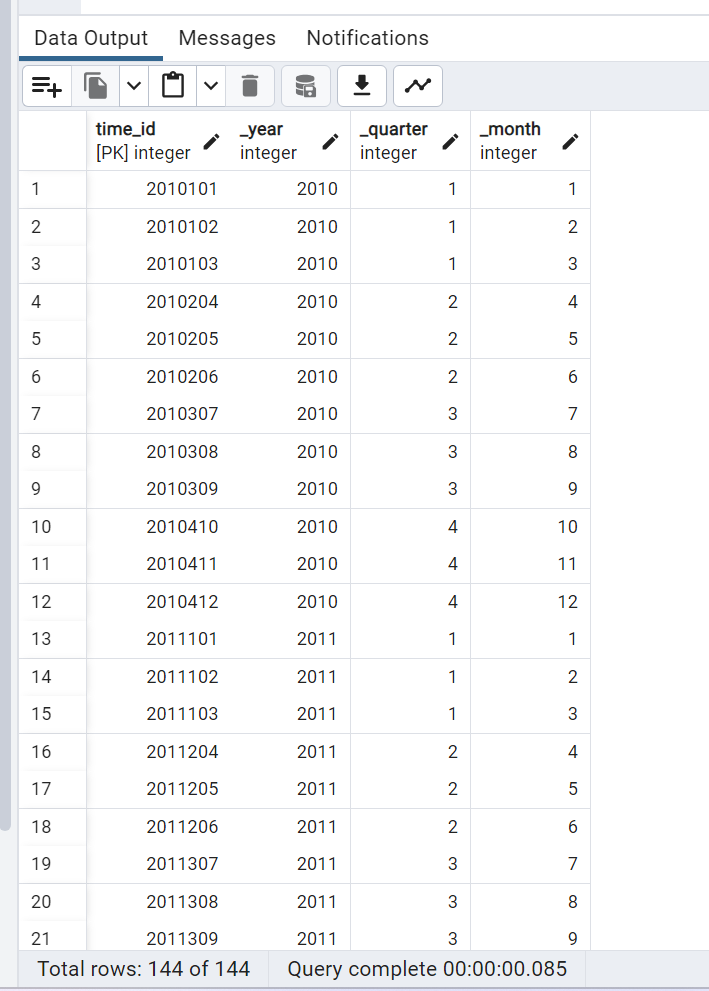


Fig.1 Data Table

**Create dimension table 2 area\_t and import rows**

CREATE TABLE area\_t (

area\_ID INT,

in\_region VARCHAR(15),

in\_state VARCHAR(15),

regulatory\_ID VARCHAR(30),

regulatory\_name VARCHAR(100),

address VARCHAR(100),

contact VARCHAR(30),

postcode VARCHAR(30),

PRIMARY KEY (area\_ID)

);

COPY area\_t(area\_ID,in\_region,in\_state,regulatory\_ID,regulatory\_name,address,contact,

postcode)

FROM 'D:\databaseSQL\area\_t.csv'

DELIMITER ','

CSV HEADER;

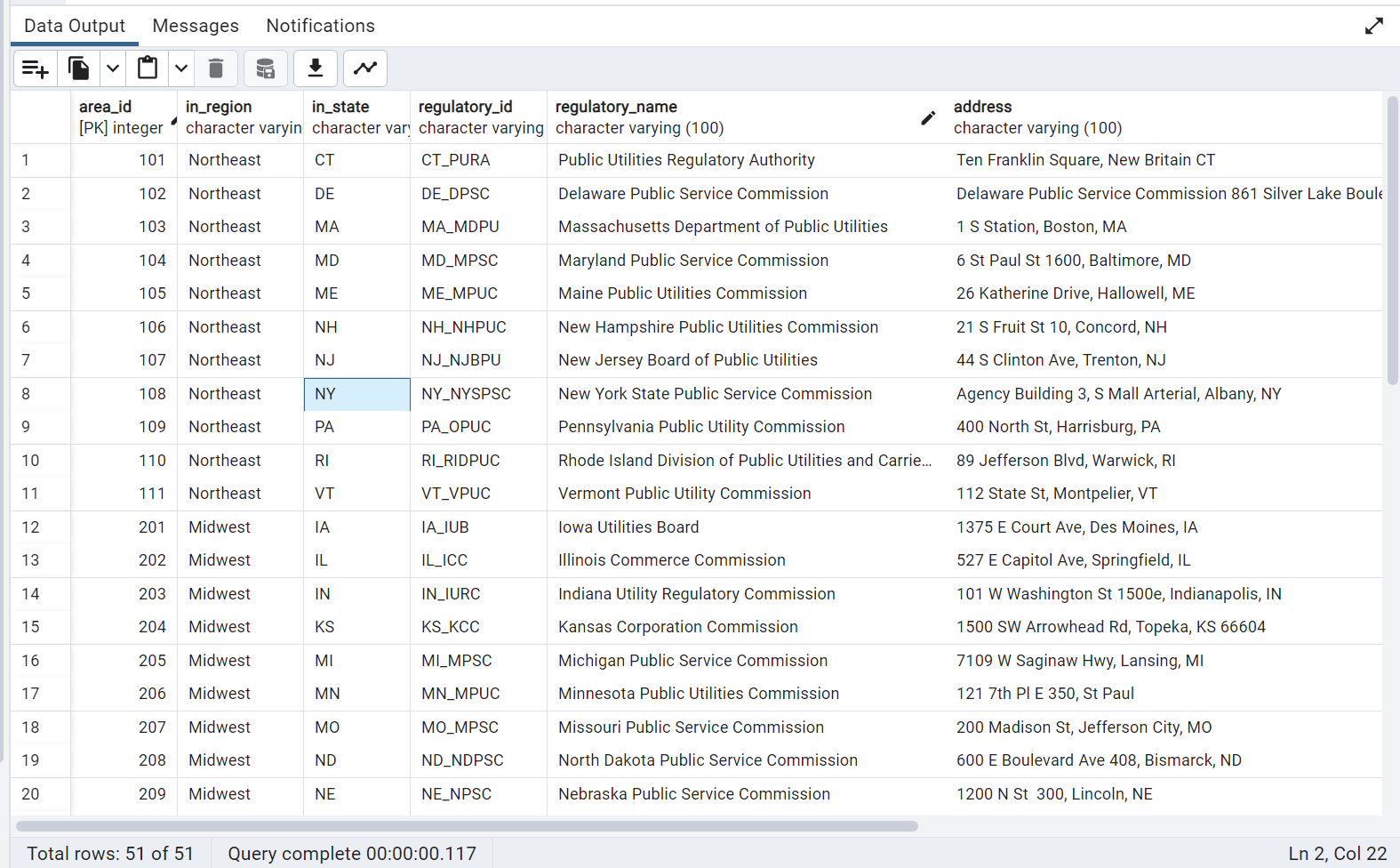


Fig.2 Data Table

**Create dimension table 3 users\_t and import rows**

CREATE TABLE users\_t (

user\_ID INT,

user\_type VARCHAR(15),

PRIMARY KEY (user\_ID)

);

COPY users\_t(user\_ID,user\_type)

FROM 'D:\databaseSQL\users\_t.csv'

DELIMITER ','

CSV HEADER;

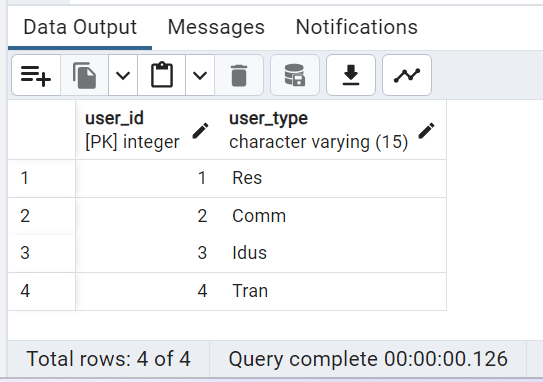


Fig.3 Data Table

**Create dimension table 4 energy\_source\_t and import rows**

CREATE TABLE energy\_source\_t (

source\_ID INT,

source\_category VARCHAR(50),

source\_type VARCHAR(50),

PRIMARY KEY (source\_ID)

);

COPY energy\_source\_t(source\_ID,source\_category,source\_type)

FROM 'D:\databaseSQL\energy\_source\_t.csv'

DELIMITER ','

CSV HEADER;

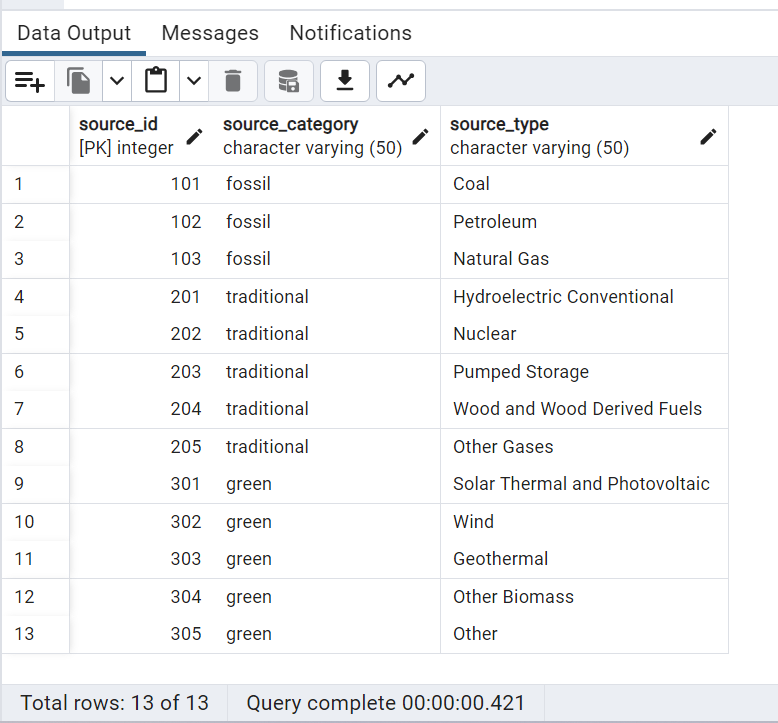


Fig.4 Data Table

**Create fact table 1 usage\_t and import rows**

CREATE TABLE usage\_t

(time\_ID INT,

area\_ID INT,

user\_ID INT,

Electricity\_Usage\_Megawatts DECIMAL(20, 1),

Unit\_Price\_CentsPerkWh DECIMAL(10, 1),

revenue\_usd DECIMAL(20, 1),

PRIMARY KEY (time\_ID,area\_ID,user\_ID),

FOREIGN KEY (time\_ID) REFERENCES time\_t (time\_ID),

FOREIGN KEY (area\_ID) REFERENCES area\_t (area\_ID),

FOREIGN KEY (user\_ID) REFERENCES users\_t (user\_ID));

COPY usage\_t(time\_ID,area\_ID,user\_ID,electricity\_usage\_megawatts,unit\_price\_centsperkwh,revenue\_usd)

FROM 'D:\databaseSQL\usage\_t.csv'

DELIMITER ','

CSV HEADER;

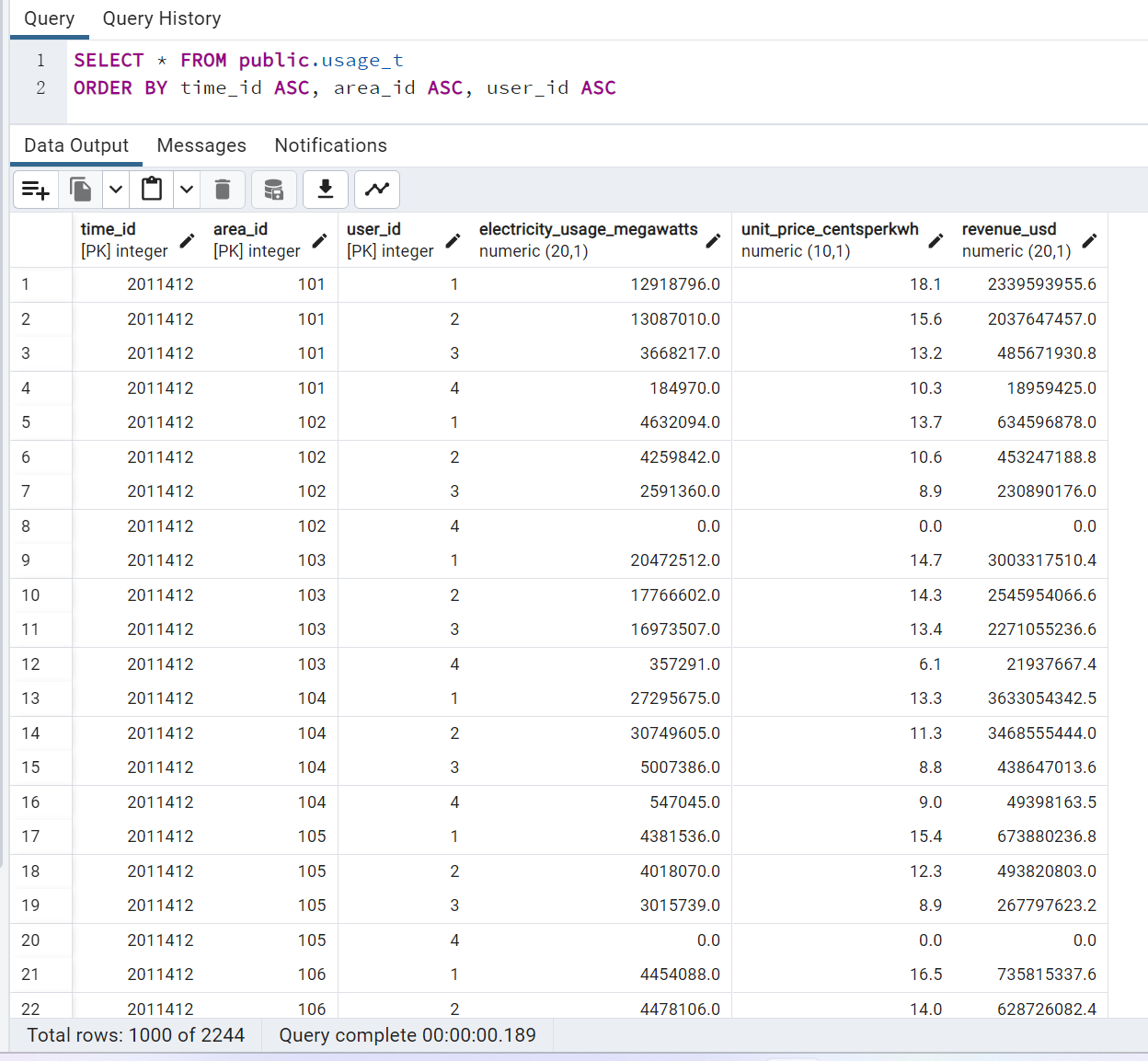


Fig.5 Data Table

**Create fact table 2 electricity\_prodution\_t and import rows**

CREATE TABLE electricity\_prodution\_t

(time\_ID INT,

area\_ID INT,

source\_ID INT,

produced\_amount\_megawatts DECIMAL(20, 1),

PRIMARY KEY (time\_ID,area\_ID,source\_ID),

FOREIGN KEY (time\_ID) REFERENCES time\_t (time\_ID),

FOREIGN KEY (area\_ID) REFERENCES area\_t (area\_ID),

FOREIGN KEY (source\_ID) REFERENCES source\_t (source\_ID));

COPY electricity\_prodution\_t(time\_ID,area\_ID,source\_ID,produced\_amount\_megawatts)

FROM 'D:\databaseSQL\electricity\_prodution\_t.csv'

DELIMITER ','

CSV HEADER;

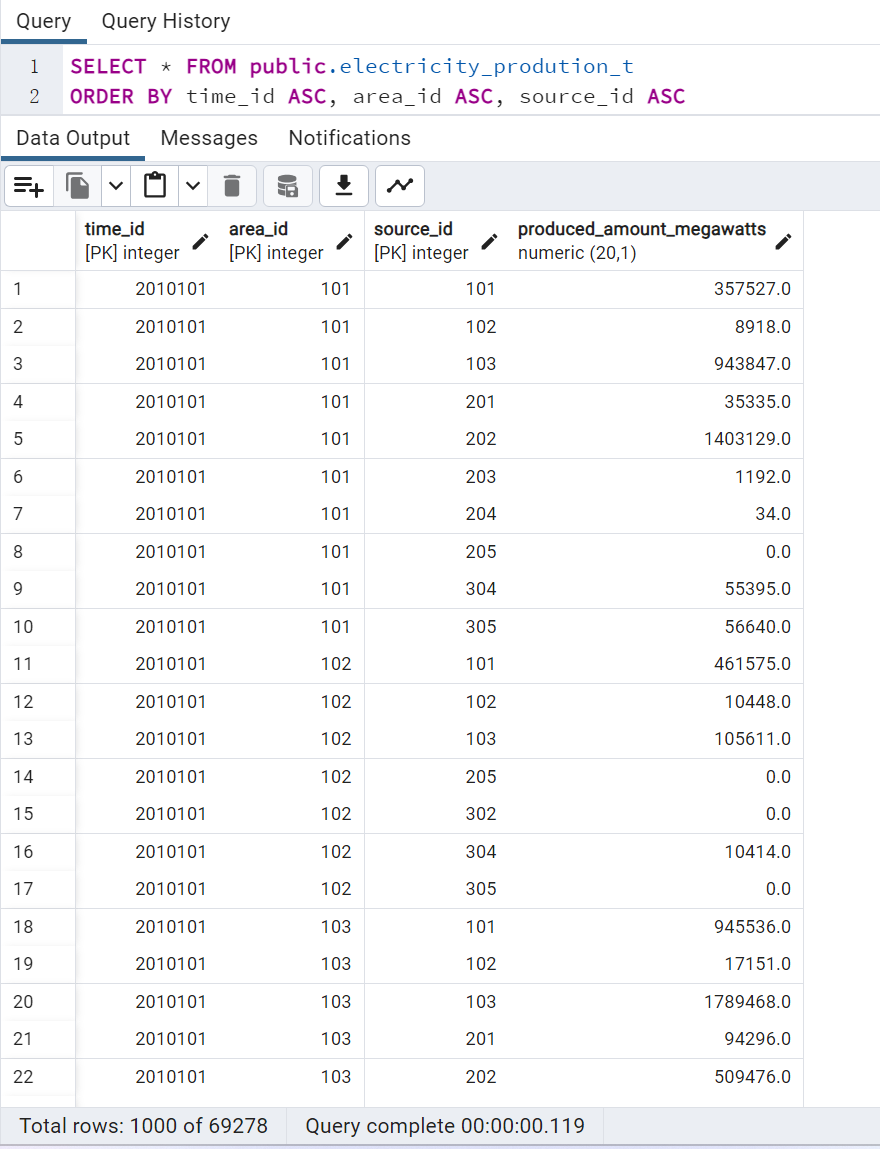


Fig.1 Data Table

**Create a full view for data mining, with 3 dimensions and measures,**

**Production view:**

SELECT

area\_t.in\_region, area\_t.in\_state,

time\_t.\_year, time\_t.\_quarter, time\_t.\_month,

source\_t.source\_category, source\_t.source\_type,

electricity\_production\_t.produced\_amount\_megawatts

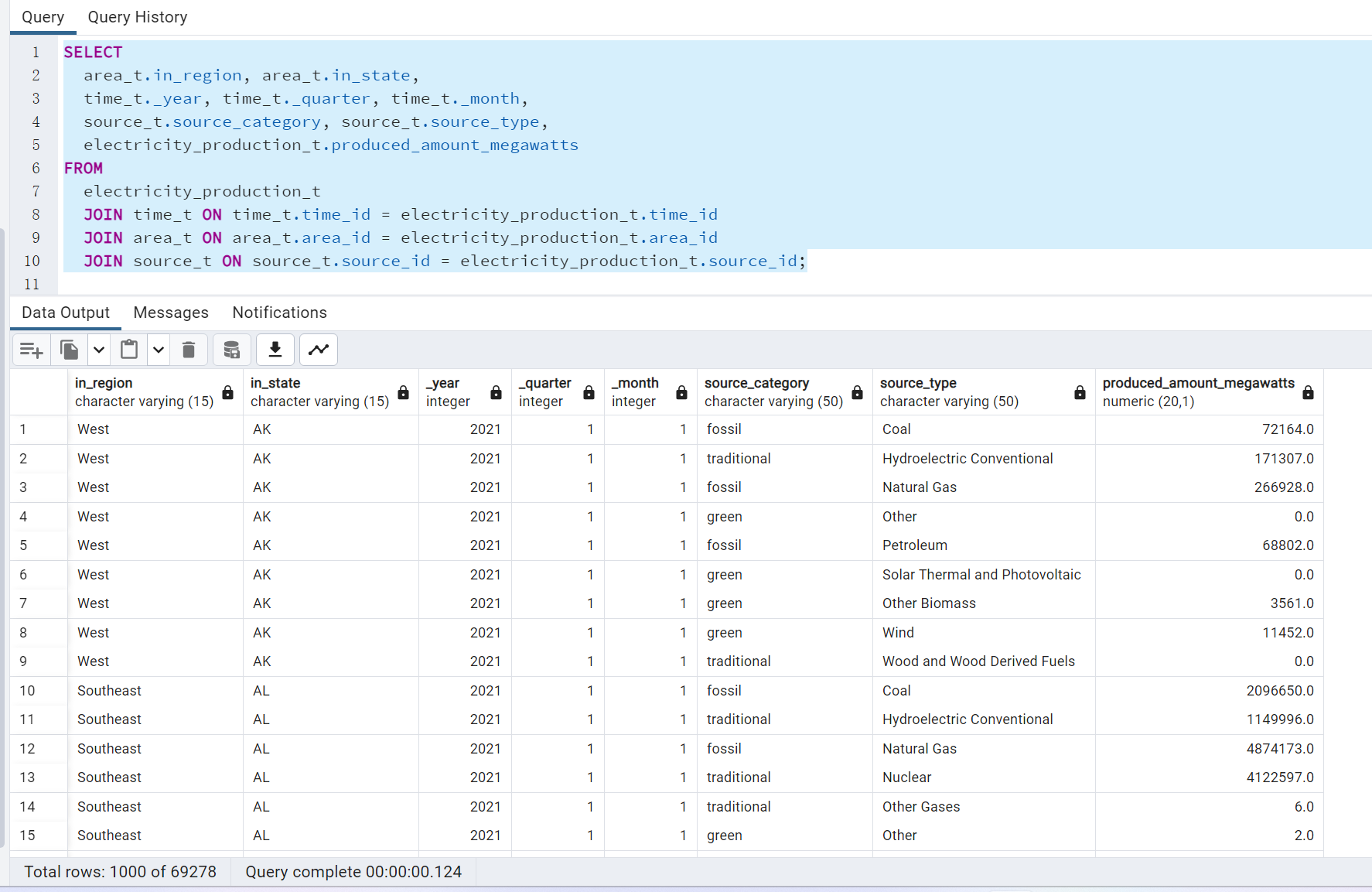
FROM

electricity\_production\_t

JOIN time\_t ON time\_t.time\_id = electricity\_production\_t.time\_id

JOIN area\_t ON area\_t.area\_id = electricity\_production\_t.area\_id

JOIN source\_t ON source\_t.source\_id = electricity\_production\_t.source\_id;



**Usage view:**

SELECT

time\_t.time\_id,

area\_t.in\_region, area\_t.in\_state,

users\_t.user\_type,

usage\_t.electricity\_usage\_megawatts,usage\_t.unit\_price\_centsperkwh,usage\_t.revenue\_usd

FROM

usage\_t

JOIN time\_t ON time\_t.time\_id = usage\_t.time\_id

JOIN area\_t ON area\_t.area\_id = usage\_t.area\_id

JOIN users\_t ON users\_t.user\_id = usage\_t.user\_id;

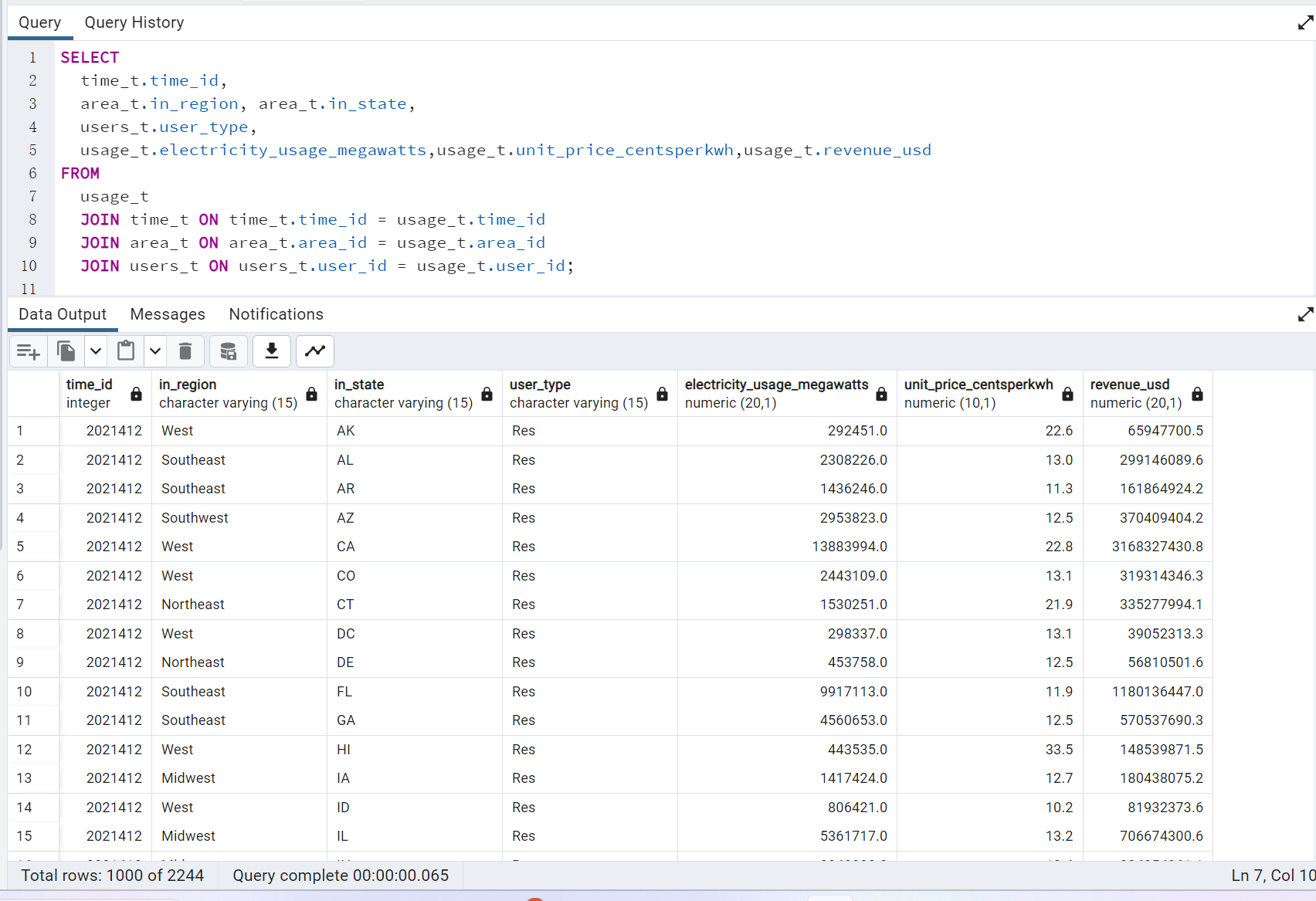
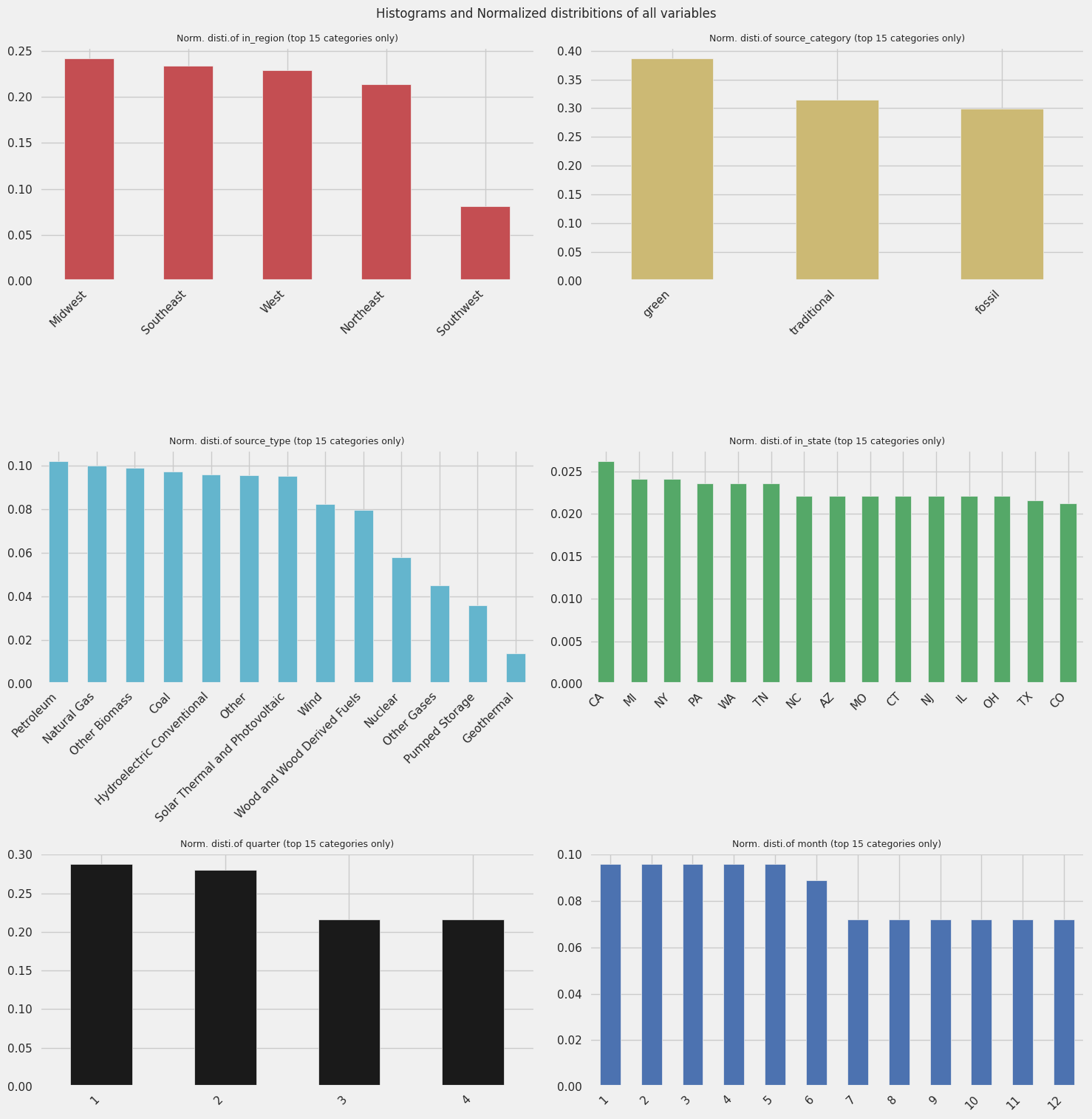


Fig.1 Data Table

# Data Mining and Visualization

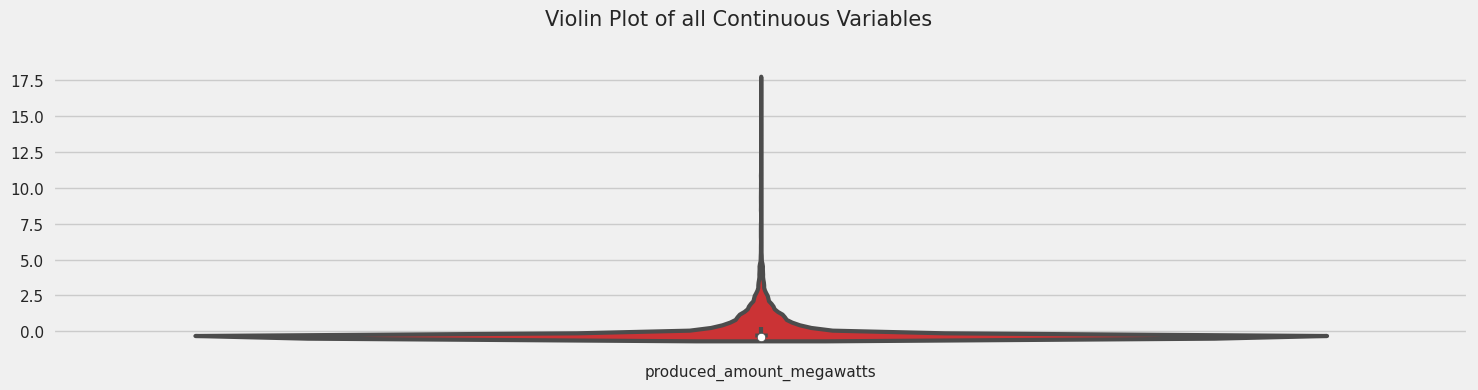
### Analysis and Charts

**Electricity Production:**



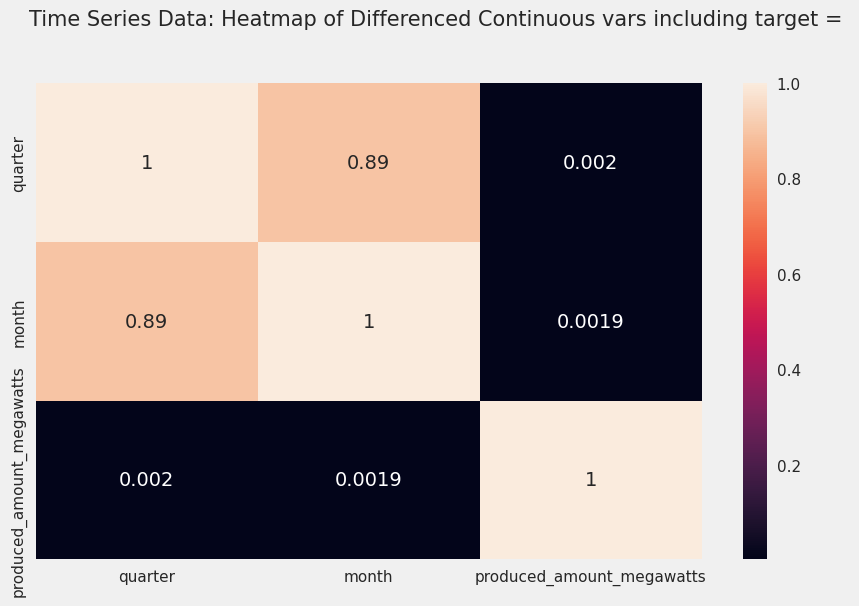
1. From the histograms, we can observe that the Midwest produces the greatest amount of electricity in normal distribution in the US.
2. The green energy category is the first source of electricity production in normal distribution in the US.
3. Petroleum is the first source of electricity production in the normal distribution in the US.
4. California produces the greatest amount of electricity in normal distribution in the US.
5. Electricity is mainly produced in the first and the second quarter of one year in normal distribution in the US.
6. Electricity is mainly produced from Jan. to Jun. in normal distribution in one year in the US.

Violin Plot of electricity production:



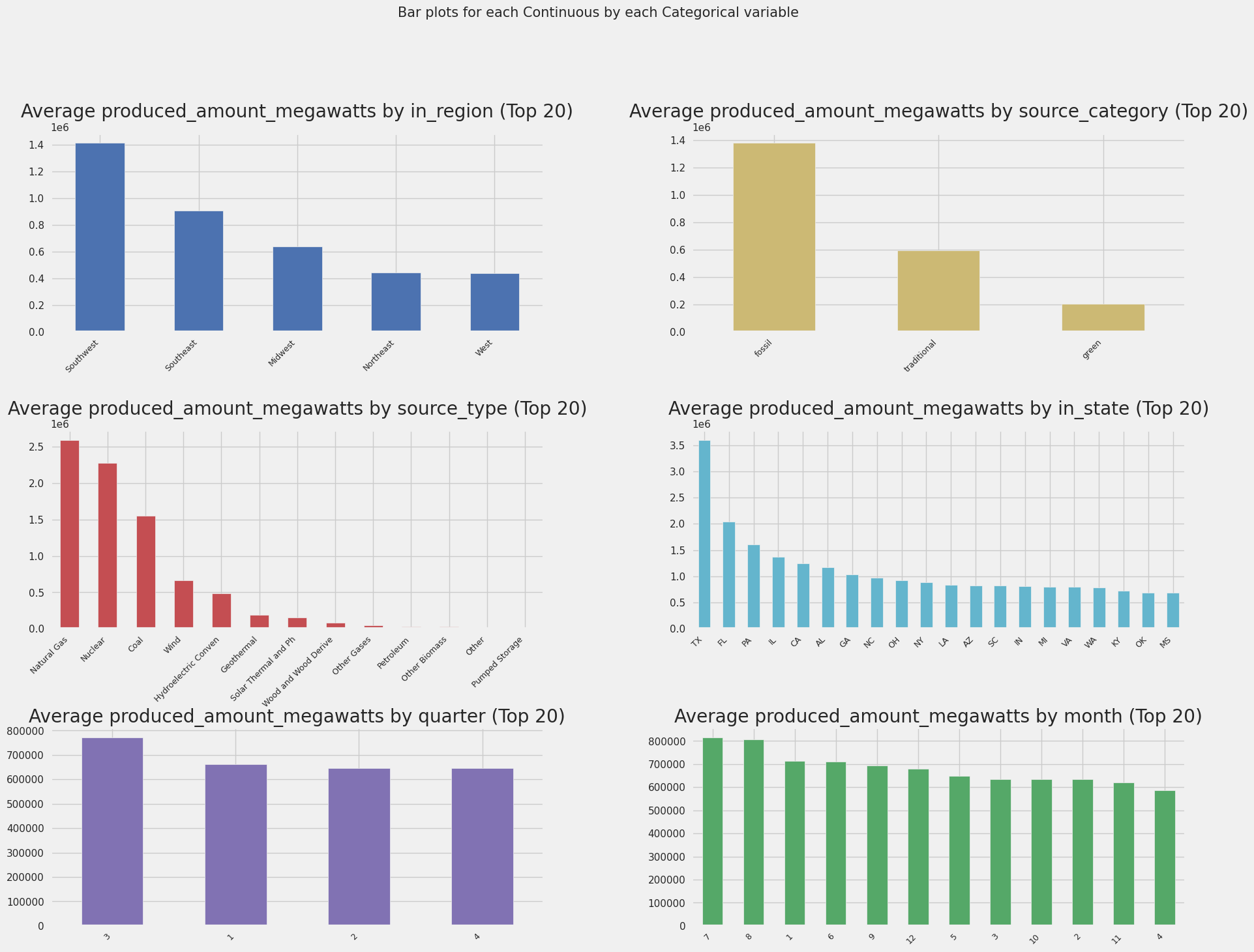
The amount of electricity production is mainly distributed in the [0, 2.5] range.

Heatmap:



The difference in the amount of electricity production in each month is 0.0019. The difference in the amount of electricity production in each quarter is 0.002.

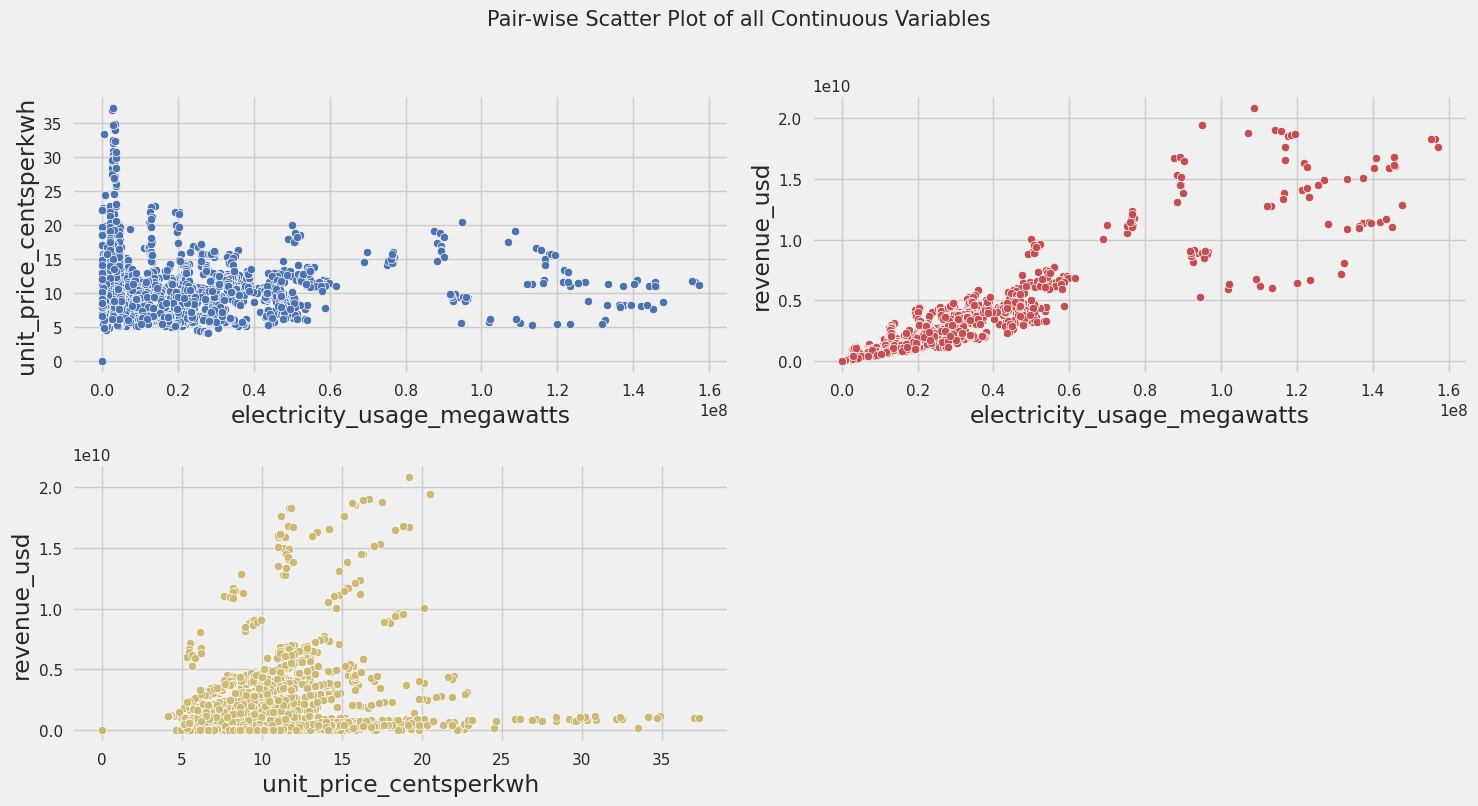
Average of the amount of Electricity Production:



1. Southwest produces the greatest amount of electricity on average in the US.
2. Fossil fuel is the first source of electricity on average in the US.
3. The Natural Gas category is the first source of electricity on average in the US.
4. TX produces the greatest amount of electricity on average in the US.
5. The US produces the greatest amount of electricity in the third quarter of one average in the US.
6. The US produces the greatest amount of electricity in July and August on average in the US.

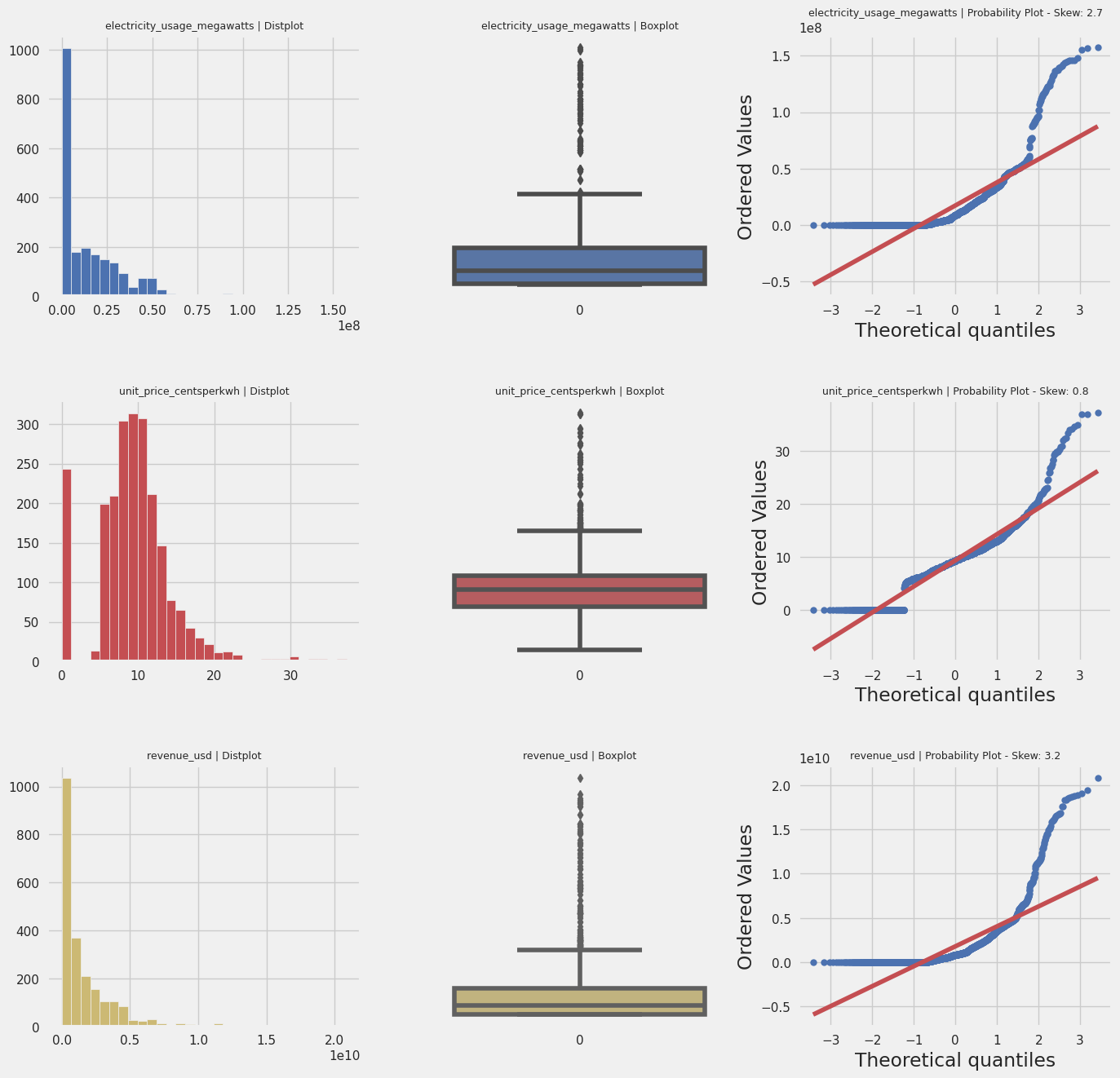
**Electricity Usage:**

Pair-wise Scatter Plots:



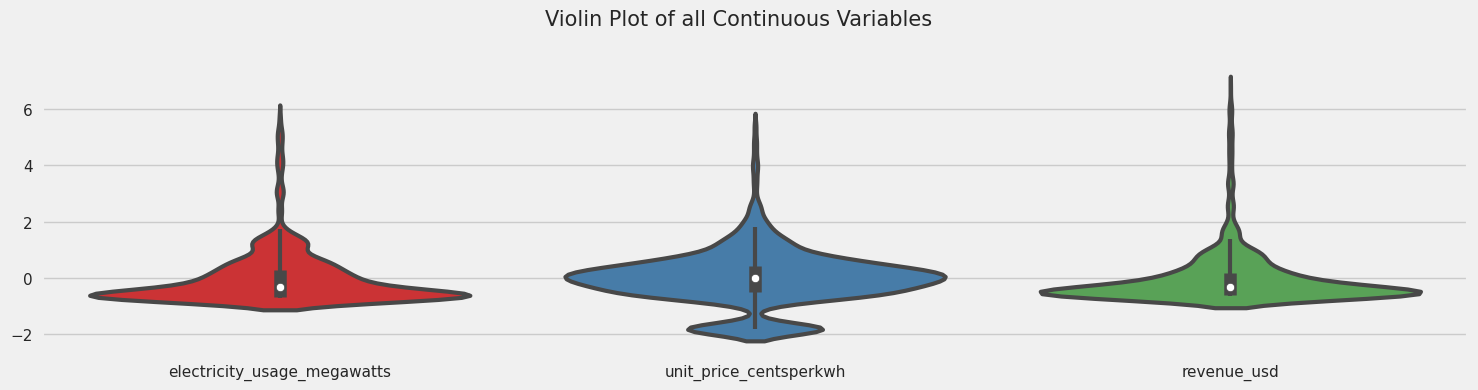
1. Electricity usage and Unit price are mainly located in the [0 - 0.6] and [5, 18], respectively. The higher the unit price the lower the electricity usage. When the unit price is mainly located between the 10 and 15 range, electricity usage can achieve the greatest usage.
2. Electricity usage and Revenue are mainly located in the [0 - 0.6] and [0 - 1], respectively. It indicates that higher electricity does not necessarily bring more revenue to this industry.
3. this industry. Unit price and revenue are mainly located between [5 - 15] and [0 – 0.5]. It indicates that a higher unit price does not necessarily bring more revenue to this industry.

Distribution plots and probability plots:

****

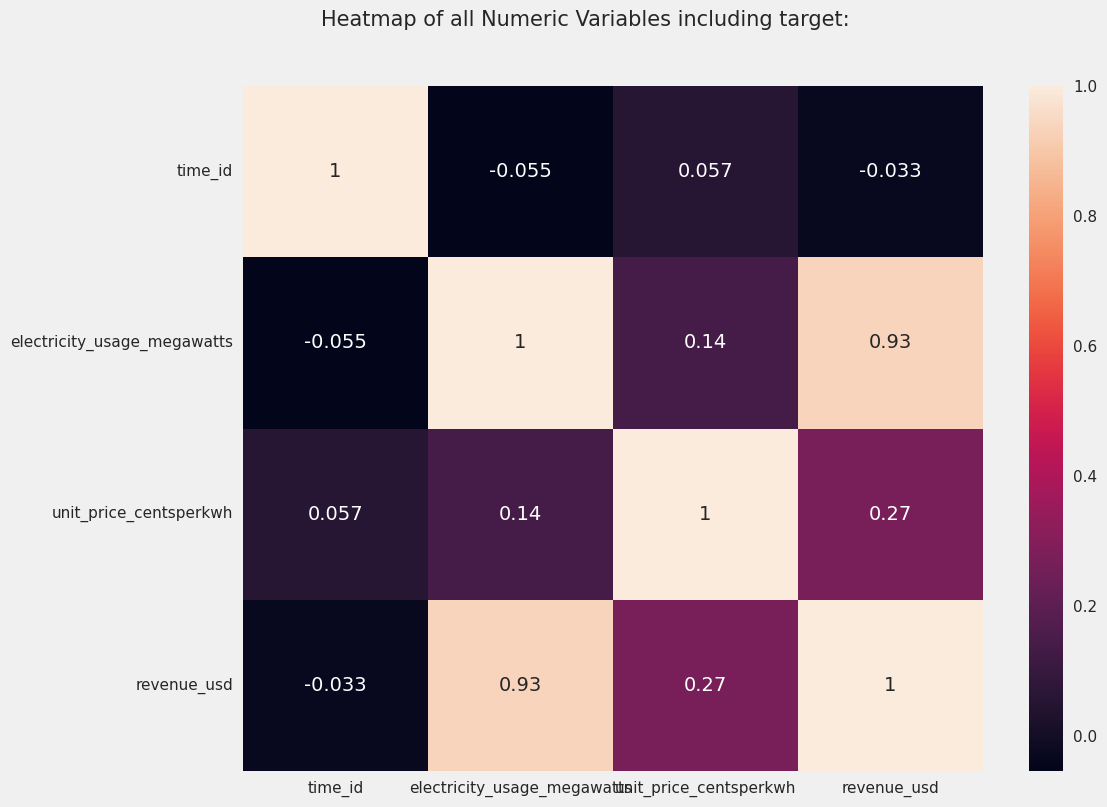
1. Electricity usage is mainly distributed in the range between 0 to 0.25.
2. In the usage box plot, the lower whisker is located around the middle of the quartile group overlapping the middle quartile group and the upper whisker is above the quartile group.
3. The probability plot of electricity usage is shown above.
4. In the unit price box plot, the Electricity unit price is mainly distributed in the range between 6 to 18.
5. The lower whisker is located below the low quartile and the upper whisker is above the upper quartile.
6. The probability plot of the unit price is shown above.
7. Electricity revenue is mainly distributed in the range between 0 to 0.5.
8. In the revenue box plot, The lower whisker is located around the middle of the quartile group overlapping the middle quartile group and the upper whisker is above the quartile group.
9. The probability plot of the revenue is shown above.

Violin Plots:

****

1. According to the electricity usage violin plot we can observe that the electricity is mainly distributed below 2 units.
2. We can observe the unit price is mainly distributed in the -2 to 2 range.
3. We can observe the revenue is mainly distributed below 2 units.

Heatmap:

****

According to the heatmap, we can observe the relationships between unit price, revenue, electricity usage, and time.

****

1. The Southwest region uses the greatest amount of electricity on average the US.
2. The Northeast region has the most expensive unit price of electricity on average in the US.
3. The Southwest region creates the greatest amount of electricity revenue on average in the US.
4. Commercial and residential are the main users of electricity on average in the US.
5. The unit price of electricity for residential is the most expensive on average in the US.
6. Residential user group is the first source of electricity revenue on average in the US.
7. Texas has the highest electricity usage on average in the US.
8. Hawaii has the most expensive unit price on average in the US.
9. California brings the highest revenue on average in the US.
10. The electricity usage maintains content across ten years on average in the US. (date as of June 15, 2021)
11. The electricity price maintains content across ten years on average in the US. (date as of June 15, 2021)
12. The electricity revenue maintains content across ten years on average in the US. (date as of June 15, 2021)

Comparison and Summary:

1. Texas produces the greatest amount of electricity and at the same time consumes the greatest amount of electricity. Texas has a relatively low unit price since Texas produces a lot electricity.
2. The U.S. government should consider increasing the investment in electricity facilities in California since California created the greatest amount of electricity Revenue in the US, while its electricity production is the highest.
3. Hawaii state has the highest unit price. Since the amount of electricity production in Hawaii is almost equal to zero. Due to the geometric location and shipping cost, the unit price of Hawaii is the highest in the US. Therefore, we can learn that the living expense is very high for people living in Hawaii.

Data Visualization:

Now you can visit these links: <https://elepro.yuanpengwang.me/>

And <https://eleusa.yuanpengwang.me/> to check my data visualization.