U.S. Electricity Market

A CRISP-DM Methodology

Project by:
Namrata Adhikari

U.S. Electricity Market

Business Understanding

The United States ranks among the world's largest electricity producers and consumers, with a significant and growing demand. In 2022, the country consumed approximately 4 trillion kilowatt-hours ranking as the second leading electricity consumer worldwide (Statista, 2024). This consumption level represents a 14-fold increase since 1950 (U.S. Energy Information Administration, 2024), and the trend is expected to continue, with consumption projected to rise by 27% by 2050 (statista, 2023). To address this growing demand, it is essential to analyze electricity pricing and consumption patterns to inform energy policy and business decisions.

Hence, this project aims to leverage data analytics to conduct an in-depth analysis of the "U.S. Electricity Market" to uncover trends and patterns in electricity pricing, consumption, and revenue patterns across different states, sectors, and years from 2001 to 2024. This will facilitate stakeholders to make data-driven decisions regarding pricing strategies, demand forecasting, resource planning and allocation, operational investments, and policy formulation. By providing actionable insights, this project will support the development of a sustainable and efficient electricity market, meeting the growing demands of the future.

Key Business Questions:

- How have electricity prices, consumption, and revenue changed over time across different US regions, sectors, and seasons and what underlying factors could drive these changes?
- What is the relationship between electricity price, sales volume, and revenue, and how do changes in one impact the others?

Data Understanding

DATA Collection :

The original/primary U.S. electricity data source is the U.S. Energy Information Administration (EIA). However, this project opted for a secondary dataset from Kaggle, which has been meticulously cleaned and modified from the original data. The initial data collection methodology employed by the Kaggle author has undergone a rigorous process, involving:

- API queries to retrieve data from EIA
- Data transformation from JSON to yearly CSV files
- Data cleaning to remove redundancies and irrelevant information
- Standardization of data formats
- Integration into a single, comprehensive CSV file

The major objectives and benefits of utilizing this secondary dataset from Kaggle are substantial time-saving on data collection and modification, focus on in-depth analysis, and the use of the subject expertise, and resources of the Kaggle author. While the data collection process is systematic, the author has not mentioned or detailed the exact data modification processes causing concerns regarding data quality and consistency.

U.S. Electricity Market

DATA Description :

The dataset U.S. Electricity prices (2001 – 2024) dataset sourced from Kaggle is a .csv file last updated on April 7, 2024. This dataset is refreshed and updated annually, providing a comprehensive overview of the U.S. electricity market. The dataset's structure and description, as outlined in the data dictionary consists of the following:

Columns/Variables: **8**	Rows/observations: **85,870**
Year	year of observation(2001:2024)
Month	month of observation(1:12)
stateDescription	the name of the state
sectorName	the sector of the electricity market
customers	the number of customers
price	average price of electricity per kilowatt-hour (kWh) in cents
revenue	the total revenue generated from electricity sales (in millions $)
sales	total electricity sales volume (in millions of kilowatt-hours).

DATA Exploration :

The data exploration phase aims to thoroughly examine the dataset to identify errors, missing values, patterns, or inconsistencies and prepare it accordingly in the next step for modeling. Leveraging descriptive statistics and visualizations like histograms and boxplots from the base-R, and tidyverse, packages, the following observations are made:

1. Data type identification reveals that the variables "stateDescription" and 'sectorName" are identified as character variables, while the remaining including "Month" and "Year" are considered numeric. Since the year and month in the dataset are labeled as numbers, for instance, 2001:2024 and 1:12 instead of a proper date format and month names, it could have been considered a numeric variable.
2. The "customers" column contains 30.3% of missing values, that is 26,040 observations out of 85,870. This needs to be addressed to overcome potential data quality issues.
3. The "year" column for the years 2001-2023 consists of 3720 observations each, except for 2024 which consists of only 310 observations indicating incomplete information for the given year.
4. The distribution of "price", "revenue", and "sales" data is heavily skewed to the right, with outliers present far to the right, suggesting a possible upward trend over time. However, the pace and magnitude of this increase are questionable. Hence, further exploration of the dataset revealed that observations for the "stateDescription" and "sectorName" variables are duplicated, with each state further being grouped into regions under the same variable and the sector variable including an additional "all sectors" category that aggregates data from the four actual sectors (commercial, industrial, other, and residential). This duplication has resulted in inflated sales, revenue, and price figures.

5. The data dictionary specifies 5 distinct "sectorName" values; commercial, industrial, other, residential, and all sectors, a combination of all 4 of them. However, an unexpected sixth value called transportation is present in some of the observations indicating data inconsistency.
6. The numeric variable contains some unit and scale inconsistencies like the price variable representing the average price of electricity per kilowatt-hour (kWh) in "cents" and the revenue in millions of dollars($).

Data Preparation:

The data preparation phase has applied various data cleaning and transformation processes, all of which are discussed in detail as follows:

1. Data Cleaning

1.1. Handling missing values:

As identified in the earlier stage, a substantial portion of the data is missing from the customers column. To address this issue, methods like Imputation could have been used, however, the nature of the missing data is "missing not at random (MNAR)" meaning, the data for the customers is missing for the first 26,040 observations which include years 2001-2007. Such methodical cases of missingness, are the most difficult to handle and the only way is to find the actual missing data, or else, imputation may introduce bias in the analysis (Harrison & Pius, 2020). Therefore, the "customers" column is removed from the dataset to maintain data integrity. Moreover, since 2024 is an incomplete year, with data only available up to April, the observations for 2024 are also discarded. This ensures a more reliable analysis.

2. Data Transformation:

The data transformation stage employs several processes such as renaming column names, changing variable data types, creating new variables, and standardizing the scale/units of numerical variables to facilitate comparison and modeling. The transformations are detailed below.

2.1. Renaming/Recoding Variables and Observations:

To simplify the analysis, the variables 'stateDescription' and 'sectorName' are renamed to 'state' and 'sector', respectively. Additionally, the outlier value 'transportation' in the sector column is reclassified as 'other' to align with the data dictionary, which defines five distinct sectors: commercial, industrial, other, residential, and all sectors.

2.2. Date Type Conversion:

The data type variables for state, and sector, are changed to a factor variable from a character variable to represent them as distinct categories or groups within a variable. Similarly, to perform a time-series analysis with the year column (2001:2023), it is left as a numeric variable for now.

2.3. Resolving inconsistencies & Adding New Variables :

To facilitate seasonal analysis of the U.S. electricity market, a new 'season' variable is created based on the 'Month' variable, categorizing the year into four seasons: Spring (March to May), Summer (June to August), Fall (September to November), and Winter (December to February). The 'season' variable is designated as a factor variable. To address inconsistencies in units and scales, a new column is created with 'Price' converted to dollars per kilowatt-hour ($/kWh). Additionally, all three numeric variables ('Revenue', 'Price', and 'Sales') are rounded to two decimal places, and three new columns are created to represent the rounded values. This ensures data consistency and facilitates further analysis.

3. Variable Selection and Final Dataset :

In the final step, only the transformed and necessary variables are selected and saved as a new CSV file named "U.S_Electricity" for use in modeling. To prevent data duplication and aggregation, only the regional observations are filtered and included within the 'state' variable. Notably, since this project relies on a single dataset, data integration is not required, and the new structure of the prepared file is explained as follows:

Columns/Variables: 7	Rows/observations: 15,180
year	year (2001:2023)
season	four seasons of the year (spring, summer, fall, winter)
state	the name of the state grouped by regions
sector	the sector of the electricty market
PRICE	average price of electricity per kilowatt-hour (kWh) in dollars ($)
SALES	total electricity sales volume (in millions of kilowatt-hours).
REVENUE	the total revenue generated from electricity sales (in millions $)

Modeling

During the data preparation stage, careful attention was given to address issues uncovered during data exploration and to ensure the data's suitability for modeling. However, following the iterative nature of the CRISP-DM process, data preparation and modeling phases are often intertwined and changes to the data may be necessary as required by the modeling process and its results. Therefore, the modeling phase begins with a renewed data exploration phase to highlight and report key summary and to see if further data cleaning is required.

Firstly, a 5-number summary statistic is executed for the three numeric variables price, sales, and revenue, providing an initial understanding of their distribution. This summary includes Minimum, First Quartile (Q1), Median, Third Quartile (Q3), and Maximum. Subsequently, a boxplot is created to visualize the summary statistics although it is important to note that the boxplot does not explicitly display the mean. The boxplot displays the interquartile range (IQR), the middle line (median), whiskers representing

the range of values within 1.5 times the IQR, and potential outliers. After confirming that the data distribution is ready for further analysis, now the focus has shifted towards answering the key questions of the project.

Usually, as per the law of demand (fundamental principles in economics), the price of a product and the quantity demanded by consumers have an inverse relationship meaning that when the price of a product increases, consumers tend to buy less of it, and when the price decreases, consumers tend to buy more of it. Hence, to gain further insights into the U.S. Electricity market a correlation and regression analysis is performed. Regression analysis can provide insights into how changes in the independent variable (PRICE) are associated with changes in the dependent variable (SALES). Multiple regression with Price and sales as independent variables and Revenue as a dependent variable is also performed. Correlation measures the strength and direction of the linear relationship between two continuous variables. Establishing this model beforehand will not only provide a foundational understanding of the relationships between variables but also facilitate and ease the interpretations of subsequent visualizations and analyses.

To analyze the trends in consumption, revenue, and price from 2001 to 2023, a line graph is created for each variable, displaying the time-series data. Additionally, a combination of bar and column graphs is used to reveal regional, seasonal, and sectoral patterns in electricity consumption, revenue, and price. To gain insights into revenue patterns, a combination chart is created, showcasing revenue and consumption percentages across U.S. regions. This visualization aims to provide valuable insights into the financial performance and market position of different states or regions in the electricity sector. Likewise, to facilitate easier interpretation and visualization of the large dataset, a percentage of total sales is calculated to understand electricity consumption patterns. This approach enables a more comprehensive understanding of the relative contribution of each category or group to the whole, making it easier to communicate and understand the data insights.

Evaluation:

From the modeling phase, several key findings emerged that can guide decision-making in the electricity sector. These findings are elaborated below:

Correlation and Regression Model (price/sales):

The correlation analysis between price and the sales variable returns a negative correlation coefficient of -0.4712. This indicates an inverse linear relationship meaning that as the cost of electricity increases, the sales volume tends to decrease and vice-versa. The magnitude of -0.4712 further shows a moderate relationship, which is not extremely strong, but significant enough to suggest a noticeable pattern. Interestingly, even in the essential electricity market, which is required for daily activities, especially for residential and commercial customers, the law of demand still holds.

Likewise, the linear regression model revealed a significant negative relationship between PRICE and SALES, with a moderate explanatory power. The regression coefficient for the price is -195028.9 indicating that for every 1 dollar/kWh increase in PRICE, SALES are estimated to decrease by 195028.9 million kWh. The intercept value of 50338.9 further suggests that at a point where the independent variable price is zero, there are still approximately 50338.9 million kWh of sales of electricity. Similarly, the Multiple R-squared values of 0.222 explain that 22.2% of the proportion of variance in SALES is explained by PRICE in the model.

According to the multiple regression model, both price and sales are significant predictors of revenue in the U.S. electricity sector with higher prices and higher sales volumes associated with higher revenue and vice-versa. Although, the price has a more substantial impact on revenue than sales volume. The coefficients indicate that a dollar/kWh change in PRICE is associated with an approximate 9065.001 (million $) change in REVENUE and a one-unit change in SALES in (million kWh) is associated with a 0.1011 (million $) change in REVENUE. Moreover, the proportion of variance in REVENUE explained by PRICE and SALES, is 0.901 (or 90.1%).

In conclusion, a price change has a noticeable inverse impact on electricity sales. This relation is further visualized through a scatter plot. However, it's essential to remember that correlation does not imply causation, and other influential factors not included in the model such as weather patterns, economic conditions, population growth, and more may have a more profound effect on electricity sales than price alone. By acknowledging these limitations, a deeper understanding of the complex dynamics in the electricity market can be gained, leading to more informed decision-making.

Electricity Sales/Consumption Trend:

The overall electricity consumption trend shows noticeable fluctuations over the years 2001-2023. Initially, there was a steady increase in consumption from 2001 to 2007, followed by a sharp decline between 2008 – 2009. The consumption then rebounded in 2010, marking the beginning of a period of gradual fluctuations that lasted until 2017. In 2018, there was a sudden surge, followed by a second significant drop in 2020. The consumption then began to rise again, only to experience another decline in 2023. Contrary to the expected inverse relationship between sales and price as depicted by the correlation and regression model, the electricity price consistently increased over the given period, while sales heavily fluctuated. A closer examination reveals that the decline in sales during 2008-2009 and 2020 was attributable to external factors - the economic crisis and the COVID-19 pandemic, respectively (EIA, 2023). This highlights a crucial statistical principle that correlation does not imply causation. In this case, the economic conditions, rather than the price increase, drove the sales decline, serving as a reminder to consider multiple factors when interpreting data.

Moreover, between 2001 and 2023, the “South Atlantic” region consistently had the highest electricity consumption, accounting for 21.601% of the total. The “West South Central” and “East North Central” regions followed closely, with each contributing around 15% of the total. In contrast, the District of

Columbia had the lowest electricity consumption, with a mere 0.299%, and “Pacific Noncontiguous” had the second lowest consumption with 0.425%. Similarly, Electricity consumption is highest during the summer season with 28.19% consumption out of the total, followed by 24.75% in the winter, 24.2% in the Fall, and 22.86% in the Spring. This result is logical given the fact that hot summers and cold winters often lead to increased use of air conditioning and heating systems, respectively, driving up electricity consumption. Further exploration into sector-specific consumption by season revealed that In the fall and spring seasons, the commercial sector is the predominant consumer of electricity with approximately 18% consumption, while during the summer and winter seasons, the residential sector demonstrates higher electricity utilization of approximately 19% compared to other sectors.

Electricity Revenue Trend:

The electricity revenue pattern over the years shows a consistent and gradual increase with few noticeable declines in the years 2009 and 2020 likely attributed to economic conditions. Seasonally, summer generates the highest average revenue at $3265.73 million, followed by fall, winter, and spring, which exhibit similar revenue ranges. Across all seasons, the residential sector consistently dominates as the largest revenue source, followed by commercial, industrial, and other sectors, respectively.

Likewise, the “Pacific Noncontiguous” region has the highest revenue($/kWh) among all the regions in the dataset. The value of 0.22 indicates that, on average, each kilowatt-hour of electricity sold in the Pacific Noncontiguous region generates $0.22 in revenue. Interestingly, in terms of market share, however, it is the second lowest with only 0.425% of total electricity consumption. This disparity suggests that either electricity prices are higher in this region, revenue generation is more efficient, or a combination of both factors is at play. On the other hand, the “East South Central” region falls on the bottom as it only generates $0.082 in revenue for each kilowatt-hour of electricity sold.

Electricity Price Trend:

Firstly, a boxplot is used to see a visual representation of the variability and distribution of prices across regions and to investigate the disparity revealed earlier. The Electricity price is significantly higher in the “Pacific Noncontiguous” region. This is due to Hawaii, a state in the Pacific Noncontiguous region reliance on imported petroleum fuels for electricity generation, resulting in higher prices and revenue (EIA, 2023). Despite having the lowest consumption and sales, the Pacific Noncontiguous region generates the highest revenue, a disparity explained by its unique energy generation circumstances. The boxplot also highlights extreme price values in regions like New England, Pacific Contiguous, East North Central, South Atlantic, and West South Central. Notably, the District of Columbia shows an unusual data point with a price of $0.0, potentially indicating a data anomaly or inaccuracy. Alternatively, this could be related to government initiatives or subsidies, suggesting further investigation in a future project scope.

Electricity prices fluctuate seasonally, reaching a peak in summer at $0.119/kWh, followed by fall, spring, and winter. Interestingly, electricity consumption is also highest in the summer. As noted by EIA (2023), the surge in demand drives prices upward due to more expensive generation sources, hence the price is

highest in the summer. Notably, while winter ranks second in terms of consumption, its prices are the lowest among the four seasons, potentially due to the commercial sector's increased consumption during fall and spring. Moreover, residential sector prices remain consistently higher across all seasons, followed by commercial sector prices, reflecting the higher costs of electricity distribution to these groups compared to the more efficient and cost-effective high-voltage supplies for industrial consumers (EIA, 2023).

Deployment:

By acknowledging the insights gained from the analysis of the electricity market, decision-makers can develop more informed strategies, taking into account regional, seasonal, and sector-specific patterns, as well as the inverse relationship between price and sales, to optimize revenue generation and meet electricity demand efficiently. However, while the insights are valuable, deployment may not be feasible at this stage. Hence, to further explore the implications of these findings, the following next steps are recommended.

1. Conduct additional analysis to investigate the relationship between price, sales, and revenue by including external factors such as weather patterns, economic conditions, and population growth. This will provide a more comprehensive understanding of the electricity market dynamics.
2. Investigate the data anomalies identified in the analysis, such as the District of Columbia‘s $0.0 price, to uncover potential opportunities for improvement or innovation.
3. While this project provides a comprehensive overview of the electricity market trends by region, sector, and season, further analysis can be conducted to drill down into specific regions, states, sectors, and seasons for more detailed and accurate insights. This will enable the development of targeted strategies and effective decision-making.

By taking these next steps, a more detailed understanding of the electricity market can be gained, enabling stakeholders to refine their understanding of the market and develop effective solutions to meet the increasing demand.

Ethical Considerations:

The project demonstrates a commitment to responsible and ethical data management to promote trust and credibility in the results and conclusions by addressing the following business and data ethical considerations.

- **Data Transparency and Objectivity:** The data collection, processing, modeling methods, and even limitations are thoroughly documented clearly and understandably for all stakeholders. The original dataset and all additional sources used for information and analysis are properly cited and referenced, maintaining integrity and accountability. No subjective judgments or biases are introduced during the analysis, ensuring balanced insights.

- **Data ownership and permissions:** The data owners, both the U.S Energy Information Administration (EIA) and Kaggle, have given permission and rights to use the data. Indeed, the electricity data is made available to the public through open data tools for transparency, promoting accountability and public benefit.
- **Data Quality:** Kaggle is a reputable dataset source site, and the most recently updated version directly sourced from EIA is used for the analysis, ensuring the accuracy and reliability of the data. Moreover, only the necessary information is used. This minimizes the risk of errors or inconsistencies that could impact the project's conclusions.
- **Data Anonymization:** The dataset does not include any individual's private or sensitive information, reducing the risk of privacy violations or harm to individuals.

In this way, by embracing these key ethical principles, the project has contributed to a more responsible and ethical data science community.

REFERENCES:

EIA. (2023). *Electricity Explained - Factors affecting electricity prices*. Retrieved from U.S. Energy Information Administration: https://www.eia.gov/energyexplained/electricity/prices-and-factors-affecting-prices.php

EIA. (2023). *Use of Energy Explained*. Retrieved from U.S. Energy Information Administration: https://www.eia.gov/energyexplained/use-of-energy/#:~:text=Before%202020%2C%20the%20largest%20recorded,4.9%25%20during%20the%20economic%20recession.

EIA. (2024). *Electricity consumption in the United States was about 4 trillion kilowatthours (kWh) in 2022*. Retrieved from U.S. Energy Information Administration: https://www.eia.gov/energyexplained/electricity/use-of-electricity.php

Harrison, E., & Pius, R. (2020). *R for Health Data Science.* Chapman & Hill.

kaggle. (2024). *U.S. Electricity Prices.* Retrieved from kaggle: https://www.kaggle.com/datasets/alistairking/electricity-prices/data

statista. (2023). *Projected electricity use in the United States from 2022 to 2050*. Retrieved from statista: https://www.statista.com/statistics/192872/total-electricity-use-in-the-us-since-2009/

statista. (2024). *Electricity consumption worldwide in 2022, by leading country.* Retrieved from statista: https://www.statista.com/statistics/267081/electricity-consumption-in-selected-countries-worldwide/