USC University of Southern California

DSCI-510 : Principles of Programming for Data Science

FALL SEMESTER 2025-26

PROJECT REPORT

Name	USC ID	E-Mail
Abhey Sabesan Mageswaran Aryaan	7162-7286-71	sabesanm@usc.edu
Rylan Nathan Lewis	8358-1308-73	rylannat@usc.edu

https://github.com/Rylannat/DSCI_510_Final_Project

How Do Heating Fuel Prices Compare Across California, and What Is Their Relationship to Energy Efficiency?

This project analyzes California's monthly electric power operational data from the U.S. Energy Information Administration (EIA) to understand how heating fuel prices relate to energy production, efficiency, and environmental impact. We collected fuel-specific records, cleaned and standardized the dataset, and performed exploratory analysis to identify trends, compare fuel performance, and evaluate how cost, energy value, and emissions indicators evolve over time.

Submitted to

Prof. Itay Hen

Q.2. What data did you collect? How did you collect it? How many data samples did you collect?

a) **Specify exact data sources and your approach.**
b) **Describe what has been changed from your original plan, what challenges you encountered or resolved.**

We collected California's monthly electric power operational data from the U.S. Energy Information Administration (EIA). The dataset includes fuel costs, cost per BTU, electricity generation, fuel consumption, heat content, Sulfur content, and other technical performance indicators. These fields allow us to compare fuel prices, energy output, and efficiency across fuel types and across time.

We used the EIA v2 API to pull the data programmatically. API calls were set up to request all available monthly data for California across multiple fuel categories. Each response was first stored as a raw CSV file containing the original data. The data was then loaded into a pandas DataFrame for cleaning and preprocessing. After handling missing values, removing duplicates, converting dates, and standardizing column names, we saved a processed CSV file ready for analysis.

In total, after combining all API pulls and completing the cleaning process, we obtained **23,819 monthly records** across all fuel types. This represents multiple years of monthly observations for each major fuel category.

a. Exact data sources and approach

The data was collected from the **U.S. Energy Information Administration (EIA) Open Data v2 API**. We used the following endpoint:

- **electric-power-operational-data:**
  https://api.eia.gov/v2/electricity/electric-power-operational-data/data/

This endpoint provides monthly California records on fuel costs, fuel consumption, heat content, generation, and related operational metrics. The API returned data in JSON format, which we converted into pandas DataFrames for cleaning and analysis.

Our approach involved:

1) Generating API requests using the EIA key
2) Downloading the raw JSON responses
3) Converting them into DataFrames
4) Standardizing column names and data types
5) Cleaning missing values and removing duplicates
6) Merging all returned tables into a single analysis-ready dataset
7) Creating monthly, yearly, and fuel-type summaries for further analysis

b. Changes from original plan and challenges

Originally, the goal was to analyze heating fuel prices alone. However, the operational dataset included several related fields, such as heat content, consumption, and generation, so the scope expanded to include these variables to properly evaluate energy efficiency.

During data collection and cleaning, several challenges came up:

Missing and inconsistent numeric values were common across the raw EIA API pulls. To address this, we implemented a multi-step hierarchical imputation strategy for the key cost fields (cost and cost_per_btu).

Missing values were filled using:
(i) monthly fuel-type means, then
(ii) yearly fuel-type means, and finally
(iii) global fuel-type means.

This ensured that critical economic variables were never left missing.

- Time fields appeared in multiple string formats, so all period values were converted to a standardized datetime type, enabling reliable monthly and yearly grouping across the project.
- Column naming inconsistencies were resolved by normalizing all column names into a uniform lowercase snake-case format. Unit descriptor fields (e.g., *_units) were removed after confirming they were redundant.
- Duplicate observations were checked for if produced when combining several API responses. These were detected and removed using exact-match deduplication to ensure each fuel-type record per period was unique.
- The raw dataset contained many features unnecessary for the analysis, so we filtered out irrelevant columns such as location, state description, stocks, and various receipt-related fields to keep the final dataset lightweight and analysis-ready.

Q.3. What kind of analysis and visualizations did you do?

a) **What analysis techniques did you use, and what are your findings?**
b) **Describe the figures you made. Explain its setup, meaning of each element.**
c) **Describe your observations and conclusion.**
d) **Describe the impact of your findings.**

The analysis focused on understanding how heating fuel prices connect to overall efficiency in California's electric power system. Once the monthly dataset was cleaned and organized, we explored it using several techniques.

We looked at time-series trends to see how fuel costs, generation, heat content, and consumption changed over the years. We compared fuels using **cost-per-BTU**, which shows the real energy value of each fuel. To understand performance, we calculated **efficiency metrics**, such as how much electricity was generated for each unit of fuel consumed. We also used **sulfur content** as a simple indicator of emissions, and finally, we examined **economic efficiency** by measuring how much energy each fuel produced per dollar spent.

Key findings:

- ✓ Fuel prices showed clear long-term patterns, and some fuels became noticeably more expensive over time.
- ✓ Higher fuel prices did not always mean better efficiency.
- ✓ Fuels with lower sulfur content were generally cleaner, but they were not always the most cost-effective.
- ✓ Efficiency differed a lot across fuel types, showing that California's energy mix includes both fuels that are expensive but inefficient and fuels that are cheaper but perform better.

b) Across the project, we created several visualizations to help interpret data more clearly.

- ✓ **Fuel price trends:**
  Line plots showing how the cost of each fuel changed month by month. The x-axis represents time, the y-axis shows the price, and each line corresponds to a different fuel type. This helped reveal long-term patterns and sudden spikes.
- ✓ **Cost-per-BTU comparison:**
  A bar chart comparing the cost per BTU across fuels. The x-axis lists the fuel types and the y-axis shows the cost. This made it easier for us to see which fuels provide more energy value for the money.
- ✓ **Efficiency over time:**
  A line chart displaying how efficiently each fuel produced electricity (generation divided by consumption). This showed us which fuels stayed consistent and what all fluctuated over the years.
- ✓ **Sulfur content trends:**
  A time-series plot showing sulfur levels, used as a rough indicator of emissions. Lower lines suggest cleaner fuels.
- ✓ **Energy-per-dollar chart:**
  A bar chart comparing how much electricity each fuel produced per dollar spent. This offered a straightforward look at cost effectiveness.

c) Observations and conclusions

From the visualizations and calculations, several patterns stood out:

- Different fuels behave very differently in terms of price, efficiency, and emissions.
- Some fuels are stable and predictable, while others show sharp price jumps.
- Fuels with higher BTU values weren't always the most efficient once operational factors were considered.
- Cleaner fuels (those with lower sulfur content) didn't always provide the best economic value.
- Overall efficiency changed noticeably over time, suggesting shifts in technology, fuel quality, or operational practices.

Overall conclusion:
Fuel price alone doesn't tell the full story. True performance depends on a combination of cost, heat content, efficiency, and environmental impact. Looking at multiple metrics gives a much more reliable understanding of how each fuel contributes to California's energy system.

Kindly note that detailed inferences are added in the **github repo>results>inferences.txt**

d. Impact of the findings

The results of the analysis have several meaningful implications:

- **Policy and planning:**
  The insights can help identify which fuels support California's long-term clean energy goals and which ones may be less sustainable.
- **Cost decisions:**
  Understanding which fuels deliver more energy per dollar can guide budgeting and operational choices.
- **Environmental considerations:**
  Tracking sulfur content helps highlight which fuels are cleaner and more aligned with environmental targets.
- **Energy transition awareness:**
  The trends show which fuels are becoming less viable and which ones might take on a bigger role in the future.

Together, the findings offer a data-backed view of how cost, efficiency, and environmental factors interact in California's energy landscape.

4. Given more time, what direction would you take to improve your project?

Given more time, **we** would expand the project in several directions. First, **we** would bring in additional EIA datasets, such as detailed emissions records or generator-level efficiency data, to build a more complete view of how different technologies affect performance. We would also develop forecasting models to predict future fuel prices and efficiency trends, which would add a forward-looking dimension to the analysis.

Another improvement would be to examine regional differences within California by combining the dataset with weather variables, such as heating-degree and cooling-degree days. This would help explain seasonal patterns in consumption and efficiency. Finally, **we** would create an interactive dashboard using tools like Plotly or Tableau to make the results easier to explore and more accessible to a wider audience.

Overall, with additional time, **we** could make the project more comprehensive, predictive, and user-friendly.