Amy Zhang (Instance 332.0323)
Perplexity (Instance ID: Perplexity-20250323-1100)

DATA PROFILE

‘Cooling System Water Usage Dataset (2023) - merging datasets from the US Energy Information Administration (EIA)

(filtered_water_df_WMETRICS_cleaned.csv)

1. SUMMARY

Data Sourcing: “The U.S. Energy Information Administration (EIA) collects, analyzes, and disseminates independent and impartial energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment.”
This is an external data source. It is government data, and as such a trustworthy data source.

Data Collection: ‘Cooling System Water Usage Dataset (2023)’ is my merge of Schedule 8 (Environmental Data) from the survey Form EIA-923 for 2023 (https://www.eia.gov/electricity/data/eia923/) with ‘Thermoelectric Cooling Water Data’ from 2023 (also EIA: https://www.eia.gov/electricity/data/water/). According to the website, “The survey Form EIA-923 collects detailed electric power data -- monthly and annually -- on electricity generation, fuel consumption, fossil fuel stocks, and receipts at the power plant and prime mover level.” Observations are all at the month level: “Monthly data (M) -approximately 3,034 plants from the monthly survey.” Further details as to methodology are as follows: “Automatic, computerized verification of keyed input, review by subject matter specialists, and follow-up with nonrespondents assure quality statistics. To ensure the quality standards established by the EIA, formulas that use the past history of data values in the database have been designed and implemented to check data input for errors automatically. Data values that fall outside the ranges prescribed in the formulas are verified by telephoning respondents to resolve any discrepancies. All survey nonrespondents are identified and contacted.” (https://www.eia.gov/electricity/monthly/pdf/technotes.pdf)

Data Contents: The data contains 15,677 observations at the Year-Month-Plant Id-Cooling System Id and Boiler plus Generator Id level. It contains Water Metrics– most importantly, Total Withdrawal, Discharge, and Consumption Volume; there are fuel metrics as well, but these are more complete on the Plant Id aggregated level, for which we will draw upon ‘US Power Plant Fuel and Generator Dataset’.

2. LIMITATIONS

Sampling and Non-Sampling Errors: As acknowledged by the EIA, the data is subject to both sampling and non-sampling errors. While annual census data isn't subject to sampling error, monthly sample survey data is. Non-sampling errors can arise from various sources, including nonresponse, response errors, definitional ambiguities, differences in question interpretation, recording mistakes, and coverage or estimation errors for missing data.

- Incomplete Water Metrics: While total withdrawal, discharge, and consumption volumes are provided, the absence of water rate metrics (e.g., withdrawal rate or flow rate) limits the ability to assess short-term impacts on water bodies.
- Incomplete Fuel Consumption Information: Although fuel consumption data is included, its completeness is limited compared to plant-level aggregated data (as available in the 'US Power Plant Fuel and Generator Dataset'). This may hinder comprehensive water-energy nexus analyses.
- Potential Biases:
    - Reporting Bias: There's a potential for reporting bias, as plant operators may have incentives to underreport water usage if it reflects negatively on their environmental performance.
    - Selection Bias: The dataset may not be representative of all power plants in the US, as participation in the EIA surveys is mandatory for some but voluntary for others. This could lead to a selection bias if plants with certain characteristics are more or less likely to participate.
- Definitional Issues: The dataset may contain ambiguities in the definition of certain terms (e.g., "cooling system," "water source"), which could lead to inconsistencies in reporting across different plants.

Ethical Considerations:

- Environmental Justice: The use of water for cooling power plants can have disproportionate impacts on local communities, particularly those that are already vulnerable to water scarcity or pollution. It's essential to consider environmental justice implications when analyzing this data.
- Transparency and Accountability: The EIA data is publicly available, promoting transparency and accountability in the energy sector. However, it's essential to ensure that the data is interpreted and presented in a responsible and unbiased manner.

- Data Privacy: Although the dataset doesn't contain personal information, it's essential to protect the identity of individual plants and operators when presenting or sharing the data. Anonymization techniques may be necessary.
- Informed Decision-Making: The data should be used to inform decision-making related to energy policy and water management in a way that promotes sustainability and considers the needs of all stakeholders.

3. DATA RELEVANCE

One of the main objectives of this project is to understand the impact the current electricity grid has on US water resources; the ecological impact of our current infrastructure. Therefore, water usage related to cooling systems is highly relevant for understanding the operations of our current technologies.

'US Power Plant Fuel and Generator Dataset (2023)' + 'US Energy Storage Plant Data (2023)' - Data from the US Energy Information Administration (EIA)

(df_1_USPR_filtered.csv; df_3_transformed.csv)

1. SUMMARY

Data Sourcing:
"The U.S. Energy Information Administration (EIA) collects, analyzes, and disseminates independent and impartial energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment." This is an external data source. As government data, it is generally considered reliable.

Data Collection:

- 'US Power Plant Fuel and Generator Dataset (2023)' (df_1_USPR_filtered.csv): A filtered version of Schedule 2, 3, and 4 from the EIA-923 survey form for 2023 (https://www.eia.gov/electricity/data/eia923/). This filtering was performed to include only the plants also present in the water usage dataset (filtered_water_df). The data encompasses power plants in the Continental US and Puerto Rico.

- 'US Energy Storage Plant Data (2023)' (df_3_transformed.csv): Data from the "Energy Storage" sheet within the same EIA-923 survey form. It's important to note that this dataset was not filtered to match the water usage dataset. Instead, it is used to provide broader context around energy storage operations.

According to the EIA, the EIA-923 survey collects detailed monthly and annual electric power data on electricity generation, fuel consumption, fossil fuel stocks, and receipts at the power plant and prime mover level.
The EIA employs quality control measures such as automatic computerized verification of keyed input, subject matter specialist review, and follow-up with nonrespondents. Formulas using the past history of data values are implemented to automatically check data input for errors, and discrepancies are resolved by contacting respondents.
(https://www.eia.gov/electricity/monthly/pdf/technotes.pdf)

Data Contents:

- Power Plant Data (df_1_USPR_filtered.csv): This dataset focuses on the unique Plant IDs also found in the Water Usage dataset, representing power plants primarily located in the Continental US and Puerto Rico. It includes key metrics such as fuel consumption and electricity generation at the monthly level.
- Energy Storage Plant Data (df_3_transformed.csv): This dataset includes energy storage plant data and will be used to provide context for water consumption analyses, specifically in relation to fuel usage and electricity generation within the energy storage sector. Key metrics include plant characteristics, monthly fuel consumption, and electricity generation.

2. LIMITATIONS

- Sampling and Non-Sampling Errors: As acknowledged by the EIA, the data is subject to both sampling and non-sampling errors. Annual census data isn't subject to sampling error, but monthly sample survey data is. Non-sampling errors arise from various sources, including nonresponse, response errors, definitional ambiguities, differences in question interpretation, recording mistakes, and coverage or estimation errors for missing data.
- Potential Biases:

  - Reporting Bias: Plant operators may have incentives to underreport fuel consumption or other metrics if it reflects negatively on their environmental performance, leading to a potential reporting bias.
  - Selection Bias: The dataset may not be fully representative of all power plants in the US, as participation in EIA surveys can be mandatory for some and voluntary for others.
- Energy Storage Data Limitations: The Energy Storage dataset was not filtered based on the water usage dataset. This limits the ability to directly correlate water usage data with energy storage plant operations.

Ethical Considerations:

- Transparency and Accountability: The EIA data is publicly available, which promotes transparency and accountability in the energy sector. Ensure that the data is interpreted and presented responsibly and without bias.
- Data Privacy: Protect the identity of individual plants and operators when presenting or sharing the data. Consider using anonymization techniques if needed.
- Informed Decision-Making: Use the data to inform decisions related to energy policy and water management in a manner that promotes sustainability and considers the needs of all stakeholders.
- Environmental Justice Considerations:
  The selection and operation of energy facilities can disproportionately impact some communities. These factors are of utmost consideration in our project.

3. DATA RELEVANCE

These datasets provide essential plant metadata—from Census and NERC region details to primary mover types, EIA sector classifications, and electricity generation statistics. This metadata is crucial for contextualizing and analyzing water usage patterns across different types of power generation facilities and in relation to energy storage operations. The datasets allow for a nuanced understanding of the interplay between water, energy, and the environment, which is central to the project's objectives.

QUESTIONS

1) What characterizes plants that are more water efficient than others?
2) Which Water Types are the most drawn upon for cooling systems and what are the associated technologies?
3) Can we predict which plants might benefit most from Edge AI monitoring?

Perplexity (Instance ID: Perplexity-20250323-1100) Assessment:

- Question 1: What characterizes plants that are more water efficient than others?
  - Strong Question: This is a great starting point for understanding how different plant characteristics influence water usage.
  - Feasibility: Answering this question is feasible because `df_1_USPR_filtered.csv` contains a range of plant characteristics, like `Reported Prime Mover`, `Reported Fuel Type Code`, `NERC Region`, and capacity-related metrics. When combined with your water usage data, you can calculate water efficiency metrics.
  - Potential Analysis: Look for correlations between water efficiency and factors like plant type, cooling technology, location, and operational characteristics.
- Question 2: Which Water Types are the most drawn upon for cooling systems and what are the associated technologies?
  - Strong Question: Important for understanding the sources of cooling water and potential environmental impacts.
  - Feasibility: Directly addressable using the `filtered_water_df` (assuming it contains information on water source/type and cooling system technology).
  - Potential Analysis: Cross-tabulate water types with cooling system types to identify the most common combinations. Consider looking at regional variations in water source and technology choices.
- Question 3: Can we predict which plants might benefit most from Edge AI monitoring?
  - Ambitious and Relevant Question: This question looks towards potential solutions for optimizing water usage.
  - Feasibility (More Challenging): This is the most complex question and requires a few assumptions and careful framing. The data likely doesn't directly tell you which plants *would* benefit, but you can identify factors that suggest a plant *could* benefit.
    - Define "Benefit": You'll need to define what "benefit" means in quantifiable terms (e.g., potential for water savings, reduction in operational costs, improved energy efficiency).
    - Proxy Variables: Look for proxy variables in your dataset that might indicate potential for improvement. For example:
      - Plants with high variability in monthly water consumption might benefit from more precise monitoring.

        - Older plants or plants with less efficient cooling systems could be prime candidates for optimization.
        - Plants in water-stressed regions might have a greater incentive to improve water efficiency.
    - Modeling (Optional): You could potentially build a predictive model to estimate the potential water savings from Edge AI monitoring based on plant characteristics and operational data.
- Potential Analysis: Focus on identifying factors that are associated with high water usage or inefficiency. Create a "benefit score" based on these factors. This won't be a perfect prediction, but it can help prioritize plants for further investigation.

Suggestions for Refining the Questions:

- Add Specificity: For each question, consider adding a time dimension (e.g., "How has water efficiency changed over time?") or a regional focus (e.g., "Are there regional differences in the types of cooling technologies used?").
- Consider Causation vs. Correlation: Be mindful of the distinction between causation and correlation. Your analysis can identify factors that are associated with water efficiency, but it can't necessarily prove that those factors *cause* the efficiency differences.
- Ethical Considerations: Be mindful of the ethical implications of your findings. For example, if you identify certain types of plants as being less efficient, avoid language that stigmatizes those plants or operators. Instead, focus on identifying opportunities for improvement and promoting sustainable practices.