Power Consumption Analysis Under Varying Weather Conditions

Abstract
This report presents a comprehensive exploratory analysis of household electricity consumption in relation to external weather conditions. By integrating minute-level power consumption data from a French household with hourly historical weather data (temperature, humidity, and precipitation), the study applies advanced data processing, statistical correlation, time series smoothing, and unsupervised clustering techniques. The analysis demonstrates that temperature is the most influential weather parameter affecting power demand, especially under cold and humid conditions. Daily energy behaviors are segmented using clustering techniques into interpretable groups that improve understanding of environment-driven consumption patterns.

Keywords
Power consumption, temperature, humidity, precipitation, exploratory data analysis, clustering, KMeans, correlation, rolling average, lag analysis, weather impact

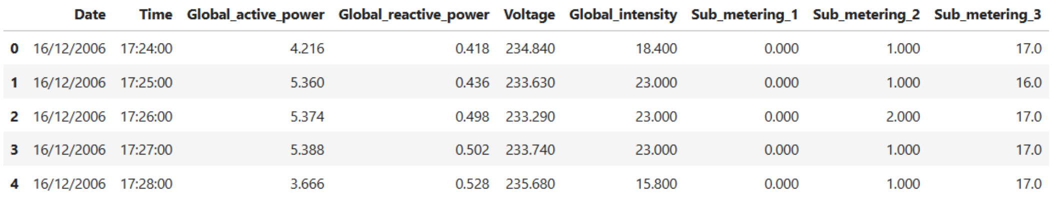
1.Introduction

Residential electricity consumption is inherently dynamic, shaped by both human behavior and external conditions. To capture and understand this variation, this study integrates time-series electricity usage data with environmental variables and applies data science methods to extract insight. The methodology includes merging datasets, performing correlation analysis, visualizing rolling trends, investigating time-lagged dependencies, and segmenting daily patterns using clustering. The findings aim to support better planning for demand forecasting and smart energy systems.

2. Dataset Description

Two distinct datasets were combined:

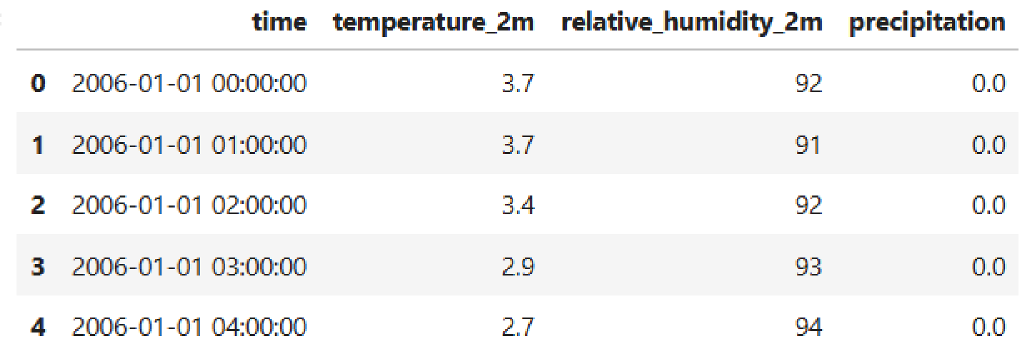
- Power consumption data was acquired from the UCI Machine Learning Repository. It provides minute-level measurements of Global Active Power, Reactive Power, Voltage, Current, and Sub-metering from 2006 to 2010 for a French household.



	Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
0	16/12/2006	17:24:00	4.216	0.418	234.840	18.400	0.000	1.000	17.0
1	16/12/2006	17:25:00	5.360	0.436	233.630	23.000	0.000	1.000	16.0
2	16/12/2006	17:26:00	5.374	0.498	233.290	23.000	0.000	2.000	17.0
3	16/12/2006	17:27:00	5.388	0.502	233.740	23.000	0.000	1.000	17.0
4	16/12/2006	17:28:00	3.666	0.528	235.680	15.800	0.000	1.000	17.0

- Weather data was sourced using Open-Meteo API(*url = "https://archive-api.open-meteo.com/v1/archive?latitude=48.776&longitude=2.2906&start_date=2006-01-01&end_date=2010-12-*

31&hourly=temperature_2m,relative_humidity_2m,precipitation"), containing hourly historical records for the same geographic location. Parameters included temperature at 2 meters, relative humidity, and precipitation levels. Both datasets were cleaned, validated, and synchronized temporally for analysis.



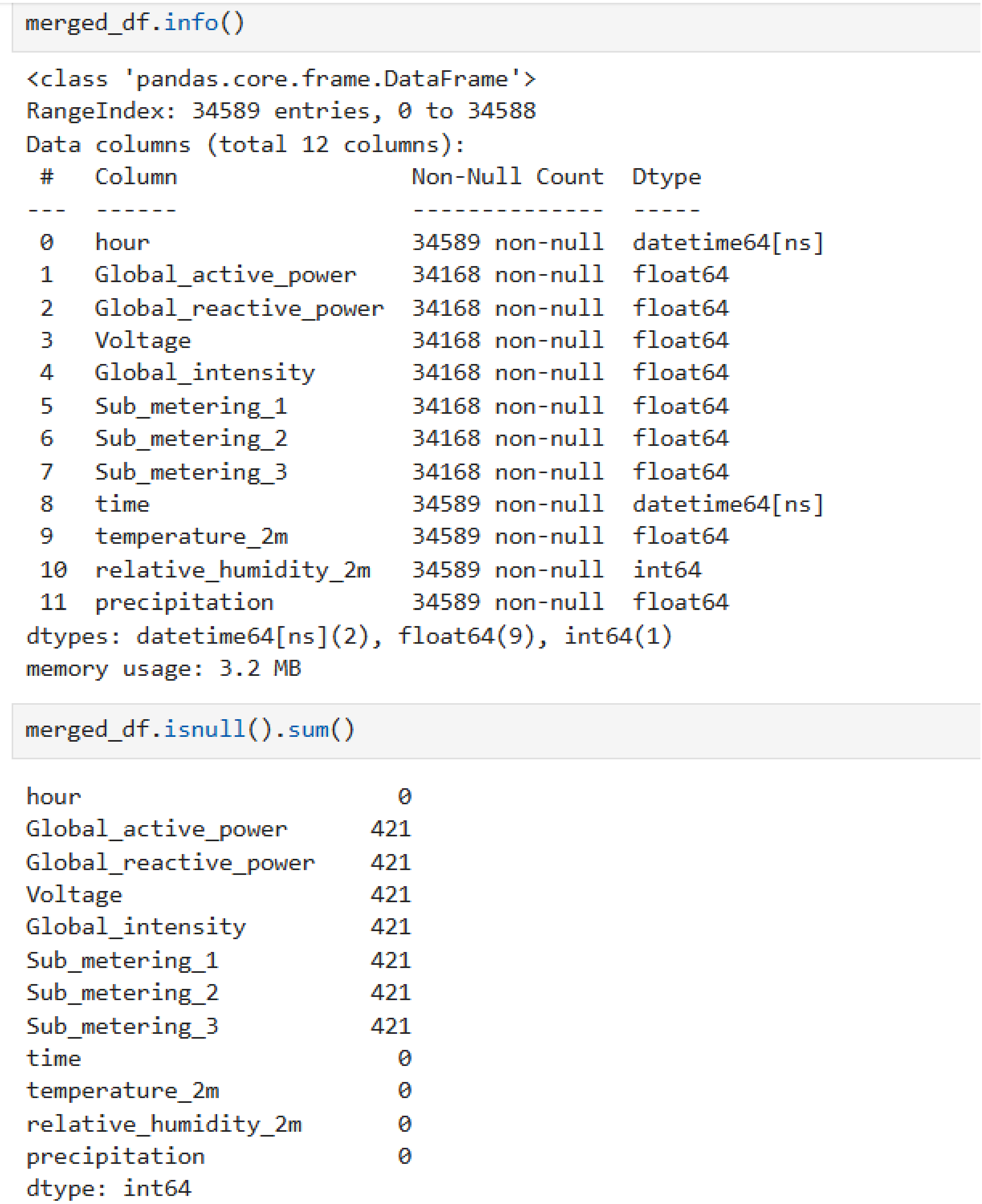
	time	temperature_2m	relative_humidity_2m	precipitation
0	2006-01-01 00:00:00	3.7	92	0.0
1	2006-01-01 01:00:00	3.7	91	0.0
2	2006-01-01 02:00:00	3.4	92	0.0
3	2006-01-01 03:00:00	2.9	93	0.0
4	2006-01-01 04:00:00	2.7	94	0.0

3.DataProcessing

The power dataset contained string-based placeholders (e.g., '?') for missing values. These were first identified and replaced with NaNs. Date and Time columns were merged into a single datetime index. All relevant fields such as power and voltage were explicitly converted to numeric types. The data was resampled from minute-level to hourly and daily frequencies using the resample() method for analysis at broader time scales.

To handle missing values systematically, a conditional approach was applied. A list of core power columns was defined: ['Global_active_power', 'Global_reactive_power', 'Voltage', 'Global_intensity', 'Sub_metering_1', 'Sub_metering_2', 'Sub_metering_3']. The dataset was filtered using dropna() with the argument how='all', ensuring rows were dropped only if **all** power values were missing simultaneously. This allowed partial observations to be retained while eliminating rows with no usable power data.

Following this, null value checks were performed using isnull().sum() to confirm that all remaining power and weather fields were fully populated. The dataset was then successfully merged with weather data using floored hourly timestamps to preserve temporal alignment. This ensured every entry had complete information for all power and weather variables, thereby making the dataset robust for downstream statistical analysis and modeling.



```
merged_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34589 entries, 0 to 34588
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype         
---  ------                 --------------  -----         
 0   hour                   34589 non-null  datetime64[ns]
 1   Global_active_power    34168 non-null  float64       
 2   Global_reactive_power  34168 non-null  float64       
 3   Voltage                34168 non-null  float64       
 4   Global_intensity       34168 non-null  float64       
 5   Sub_metering_1         34168 non-null  float64       
 6   Sub_metering_2         34168 non-null  float64       
 7   Sub_metering_3         34168 non-null  float64       
 8   time                   34589 non-null  datetime64[ns]
 9   temperature_2m         34589 non-null  float64       
 10  relative_humidity_2m   34589 non-null  int64         
 11  precipitation          34589 non-null  float64       
dtypes: datetime64[ns](2), float64(9), int64(1)
memory usage: 3.2 MB
```

```
merged_df.isnull().sum()
```

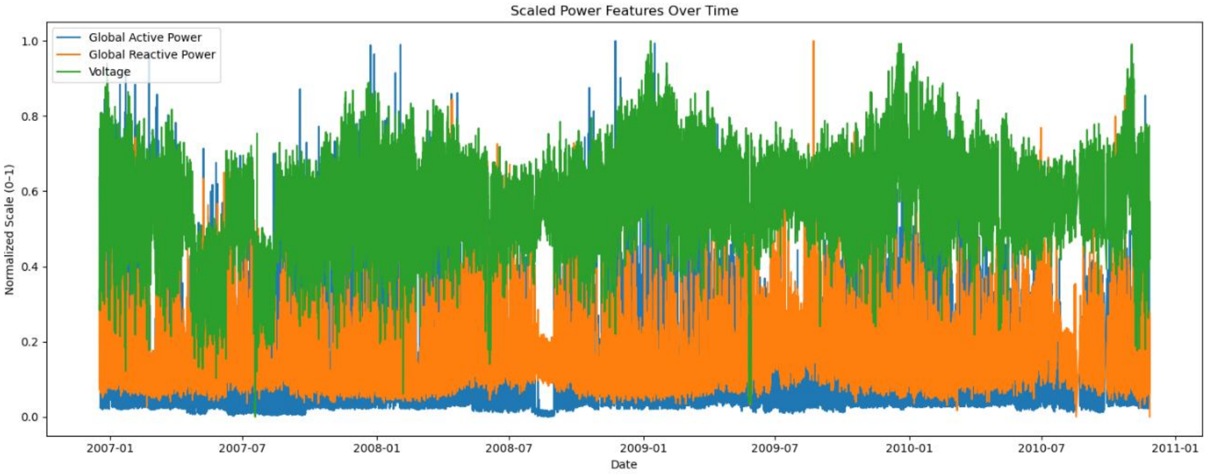
```
hour                       0
Global_active_power      421
Global_reactive_power    421
Voltage                  421
Global_intensity         421
Sub_metering_1           421
Sub_metering_2           421
Sub_metering_3           421
time                       0
temperature_2m             0
relative_humidity_2m       0
precipitation              0
dtype: int64
```

4. Exploratory Data Analysis

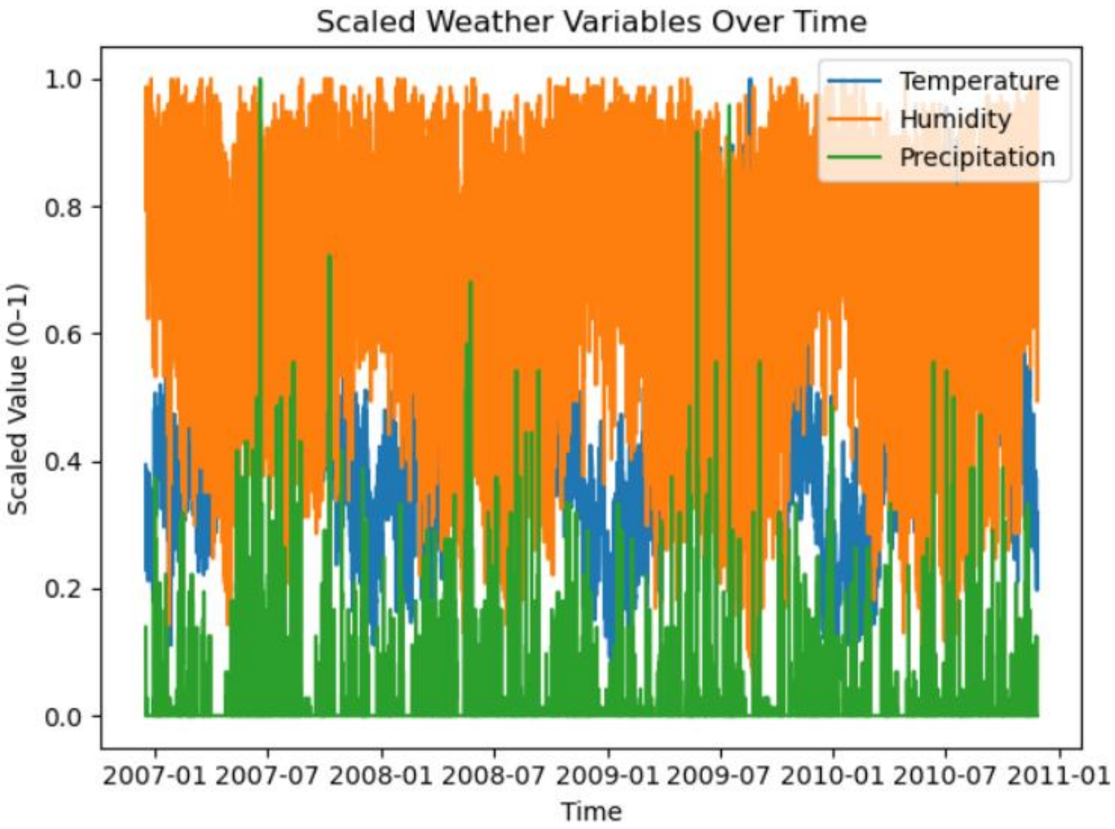
Quantitative relationships were explored using Pearson correlation on cleaned, numeric fields. A correlation matrix was generated for power and weather variables. The strongest correlation observed was between Global Active Power and Global Intensity (r = 1.00), which is expected due to their formulaic relationship. Global Active Power showed a weak negative correlation with temperature (r ≈ -0.19), indicating higher consumption during colder conditions. Correlations with humidity and precipitation were negligible.

To understand long-term dependencies among internal power variables, three key metrics—Global Active Power, Global Reactive Power, and Voltage—were plotted together over time in both unscaled and normalized formats.

In the unscaled plot, raw magnitudes were preserved and plotted on separate y-axes to represent their original units. The figure clearly shows different scales: active power (kW), reactive power (kW), and voltage (V). This visualization allows assessment of high-level fluctuation patterns across the full duration of the dataset. It revealed that periods of high active power often coincide with slightly elevated reactive power and changes in voltage stability.



In the normalized (scaled) version, all three variables were transformed to a [0,1] scale using MinMaxScaler. This enabled direct visual comparison of their shape and periodicity over time, without magnitude differences affecting interpretation. Despite the difference in raw units, the scaled plot highlighted synchronized peaks and valleys, suggesting a periodic dependency pattern and joint influence on consumption profiles.



Correlation Matrix Overview

A correlation heatmap was constructed to quantify and visualize the linear relationships between various power-related and weather-related variables. The matrix displays pairwise Pearson correlation coefficients, where values range from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no linear relationship.

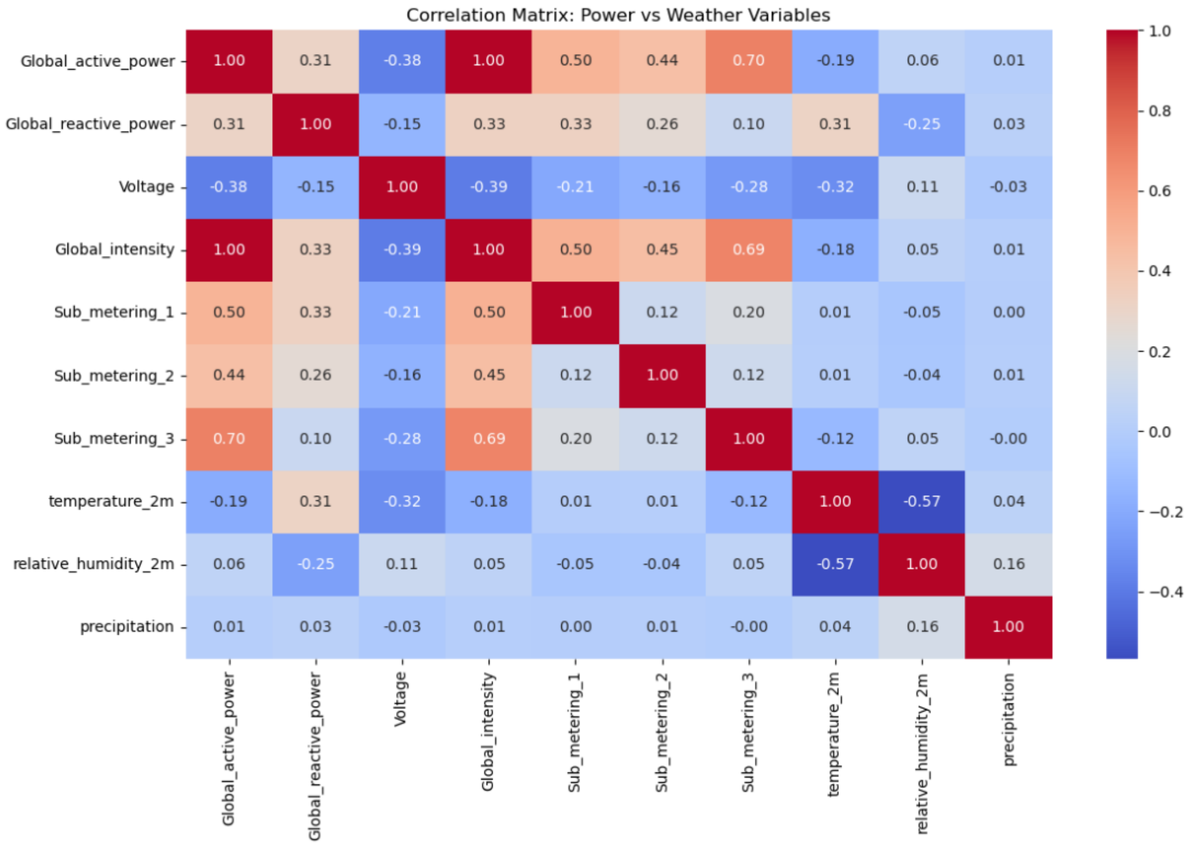
Key findings from the matrix include:

- **Strong intra-power correlations**:
    - Global_active_power shows perfect correlation with Global_intensity (r = 1.00), affirming their direct mathematical relationship.
    - It also has a high correlation with Sub_metering_3 (r = 0.70), suggesting a consistent energy draw from the devices in that channel.
- **Voltage has inverse relationships**:
    - It is weakly negatively correlated with both Global_active_power (r = -0.38) and Global_intensity (r = -0.39), indicating that high consumption periods may be accompanied by voltage drops.
- **Weak external dependencies**:

- Temperature showed a mild negative correlation with Global_active_power (r = -0.19), suggesting that colder weather is loosely associated with higher energy usage (likely due to heating).
- Relative_humidity and precipitation displayed very weak correlations, indicating limited or no direct impact on power usage in this dataset.

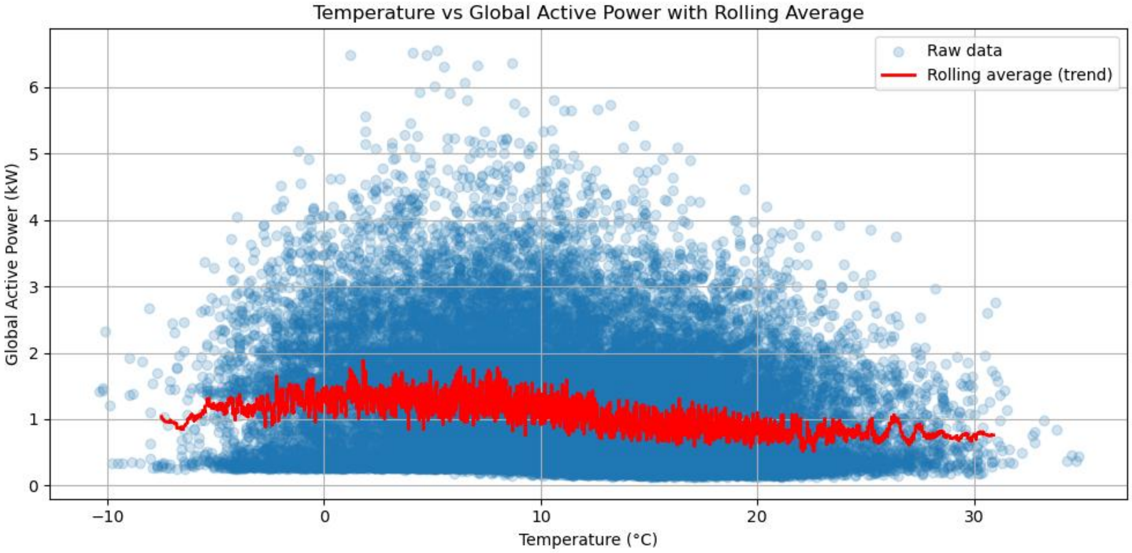
- **Inter-weather correlations**:
  - Temperature and Relative_humidity have a strong inverse correlation (r ≈ -0.57), aligning with seasonal atmospheric behavior.

This matrix was essential for selecting features for further modeling and provided early insight into which environmental variables were worth exploring in more detail through trend plots and clustering.

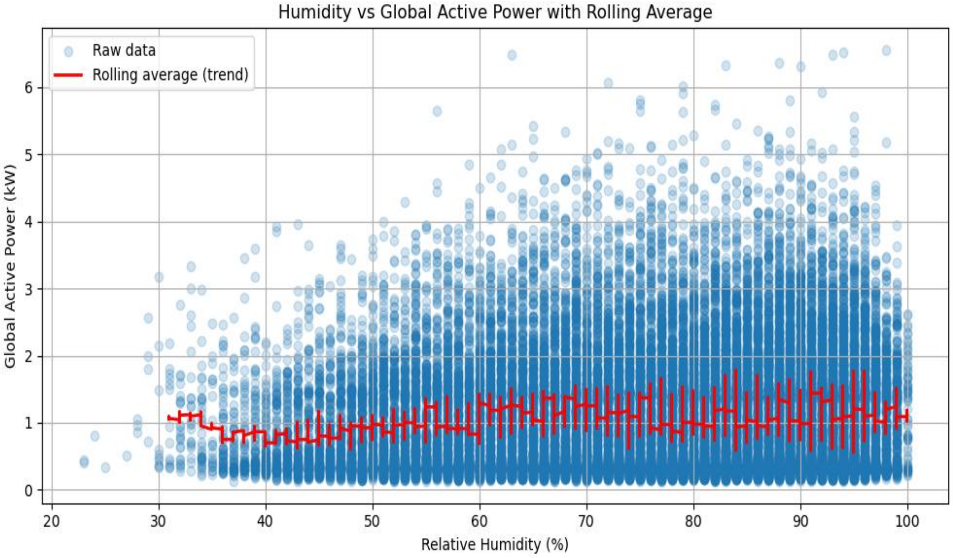


5. Rolling Average Trend Analysis

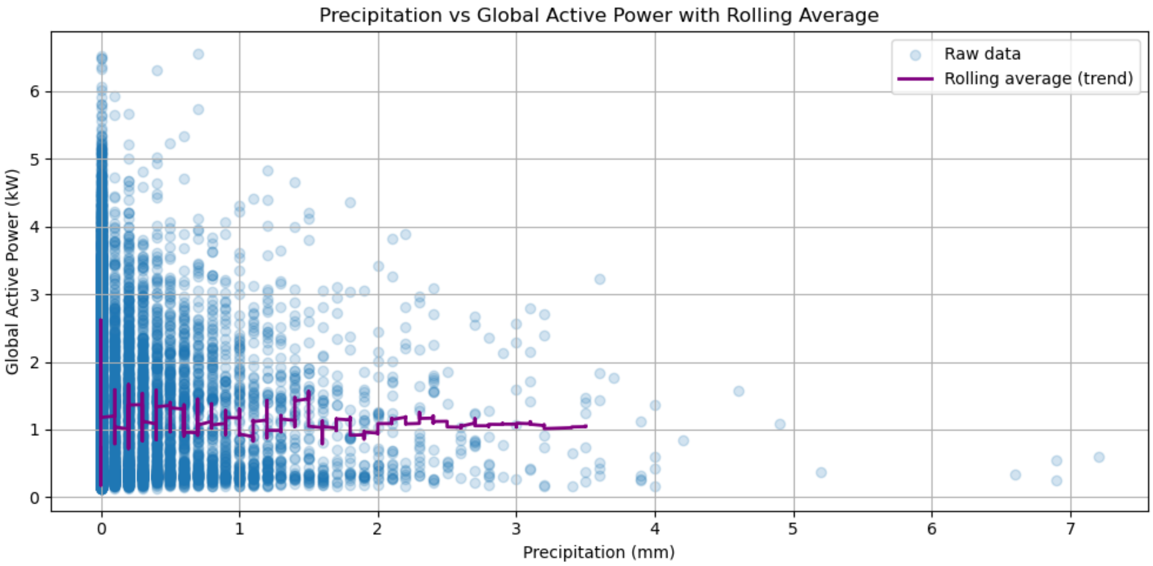
To explore temporal trends while suppressing noise, rolling averages were applied. The dataset was sorted by weather variable and power usage, then smoothed using a centered rolling window of 50 observations. These rolling averages were plotted over the scatter distributions:



- For temperature, a clear inverse trend with power usage was observed.



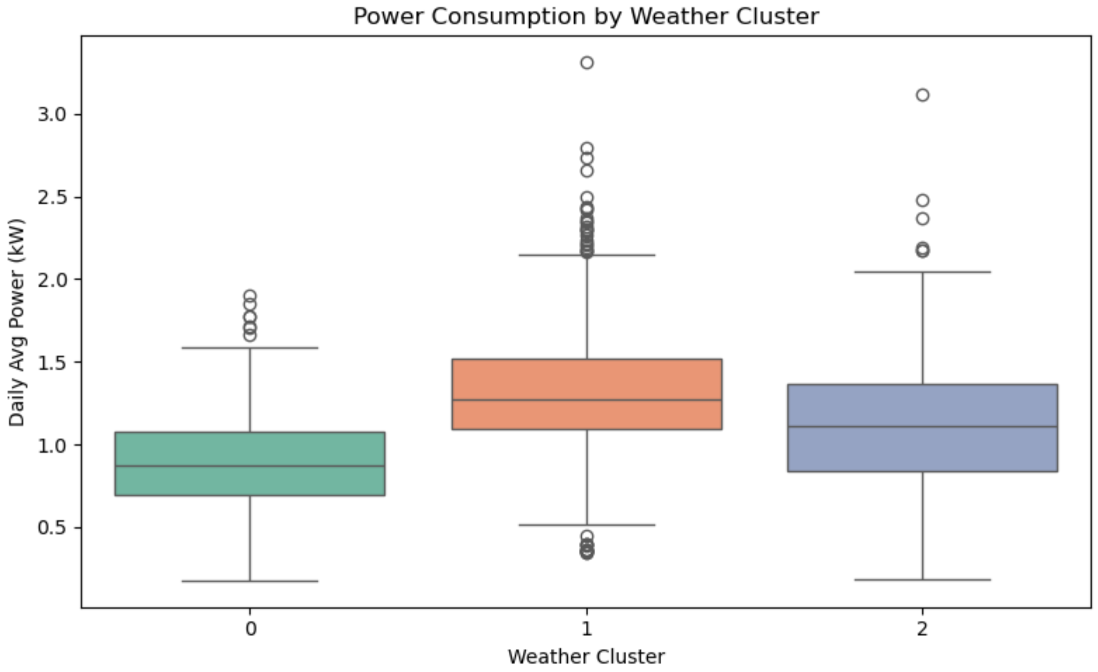
- For humidity, a mild increasing trend emerged, indicating a possible secondary effect.



- For precipitation, the trend was flat, suggesting no measurable influence on power usage. This method allowed for more interpretable visual patterns that raw scatter plots could not reveal.

6. Lagged Weather Effect Analysis
To detect potential delayed responses, lagged versions of temperature, humidity, and precipitation were introduced by shifting their values by one day. These lag features were then correlated with same-day power consumption. Results showed that the lagged temperature variable had a slightly stronger correlation than the current-day value, indicating possible thermal inertia effects—where cold weather conditions impact energy demand the following day (e.g., heating systems).

7. Clustering Daily Weather Patterns
To classify days into weather-driven consumption profiles, KMeans clustering was applied. Prior to clustering, features were scaled using StandardScaler to prevent magnitude bias. The input features used for clustering were: daily average temperature, relative humidity, and precipitation. The number of clusters (k=3) was selected to balance interpretability and intra-cluster variance. KMeans was then applied, and the resulting cluster labels were merged back into the daily dataset.



8. Cluster Profile and Interpretation

Each cluster was profiled by computing its mean weather and power statistics:

- **Cluster 0 (Warm-Dry)**: Avg Temp ≈ 15.9°C, Humidity ≈ 69%, Precip ≈ 0.02 mm, Power ≈ 0.88 kW
- **Cluster 1 (Cold-Humid)**: Avg Temp ≈ 5.7°C, Humidity ≈ 83%, Precip ≈ 0.03 mm, Power ≈ 1.32 kW
- **Cluster 2 (Mild-Rainy)**: Avg Temp ≈ 12.3°C, Humidity ≈ 84%, Precip ≈ 0.37 mm, Power ≈ 1.13 kW
  Boxplots confirmed that Cluster 1 days consistently consumed the most energy. Cluster 0 had the lowest consumption and most stable weather conditions. The segmentation clearly demonstrated that environmental clustering captures meaningful behavioral patterns in energy usage.

9. Figures and Visual Interpretations

To complement numerical analysis, various visualizations were generated:

- A heatmap visualized the correlation matrix for power and weather variables.
- Rolling average plots were used to show trends for temperature, humidity, and precipitation against power.
- Cluster-wise boxplots were constructed to highlight variance and median differences.
- Daily scatter plots with rolling smoothed lines demonstrated non-linear behavior not visible in raw data.

These visuals were essential for confirming statistical trends and explaining model outputs.

10. Conclusion

The analysis confirms that temperature is the most significant environmental driver of household power usage, especially during colder days. While humidity plays a minor secondary role, precipitation shows minimal effect. he analysis confirms that temperature is the most significant environmental driver of household power usage, especially during colder days. While humidity plays a minor secondary role, precipitation shows minimal effect.

The use of rolling averages revealed that:

- Power usage tends to **increase** as **temperature drops**, particularly evident in colder months.
- Humidity has a **slightly positive** but less consistent relationship with power usage.
- Precipitation does **not** exhibit any noticeable pattern or correlation with energy consumption.

Lag analysis showed that **yesterday's temperature** has a slightly stronger influence on today's power consumption compared to same-day weather, suggesting **thermal lag effects** in energy behavior.

Clustering divided daily patterns into three interpretable weather-based groups:

- **Cold and Humid days** consistently led to the **highest energy usage**.
- **Mild and Rainy days** showed **moderate consumption**.
- **Warm and Dry days** were associated with the **lowest power demand**.

These insights confirm that weather, especially temperature, plays a predictive role in power consumption patterns. Understanding these relationships enables energy stakeholders to:

- Develop **smarter grid management systems** that anticipate demand during colder periods.
- Introduce **dynamic pricing models** that adjust based on expected environmental load.
- Support **home automation strategies** that optimize energy use based on short-term weather forecasts.

By combining statistical analysis with visual diagnostics and unsupervised learning, this study provides a replicable framework for uncovering environmental influences in energy data.

These insights can be leveraged in smart grid planning, dynamic pricing strategies, and home energy management systems.

References

[1] UCI Machine Learning Repository - Individual Household Electric Power Consumption.
[2] Open-Meteo API - Historical Weather Data.
[3] Scikit-learn Documentation - KMeans Clustering.
[4] pandas Documentation - Data Resampling and Aggregation.
[5] matplotlib, seaborn - Data Visualization Libraries.