SOLO
WORRIER

Why this dataset was chosen ?

- **High-frequency time sampling**

The data is recorded every 15 minutes, which allows detailed time-series analysis,
rolling windows, and reliable resampling to hourly or daily levels.

- **Long historical coverage**

The dataset spans multiple years 2015 -2022,
providing a large number of observations suitable for statistical analysis, trend detection,
and theory demonstrations such as LLN and CLT.

- **Availability of external influencing variables**

In addition to electricity consumption, the dataset includes weather variables
(temperature, humidity, solar radiation, wind, pressure),
enabling meaningful feature engineering and regression modeling.

- **Suitability for multiple analytical tasks**

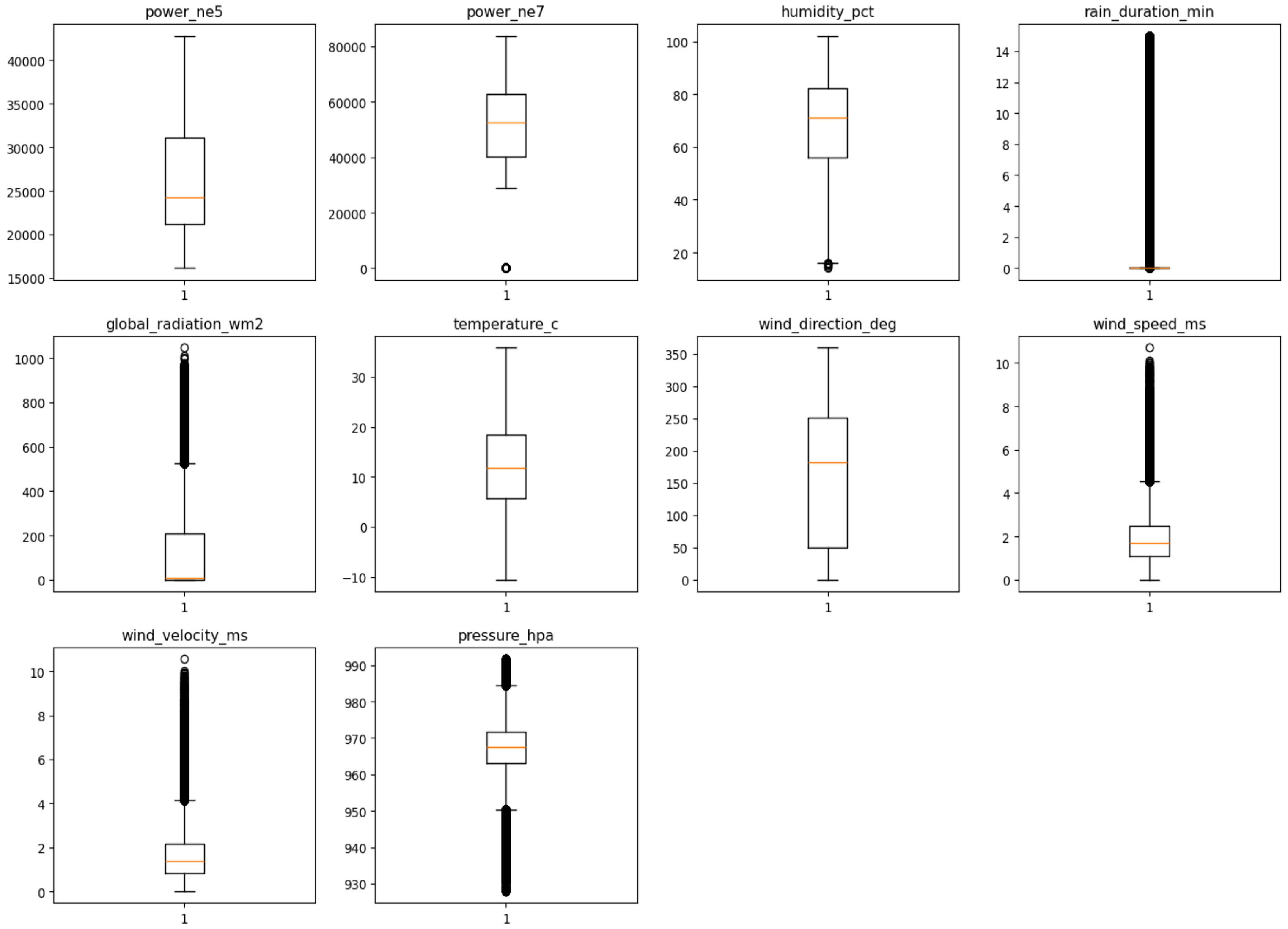
The dataset supports exploratory analysis, probability modeling,

- regression, and dimensionality reduction, making it well suited for all project requirements.

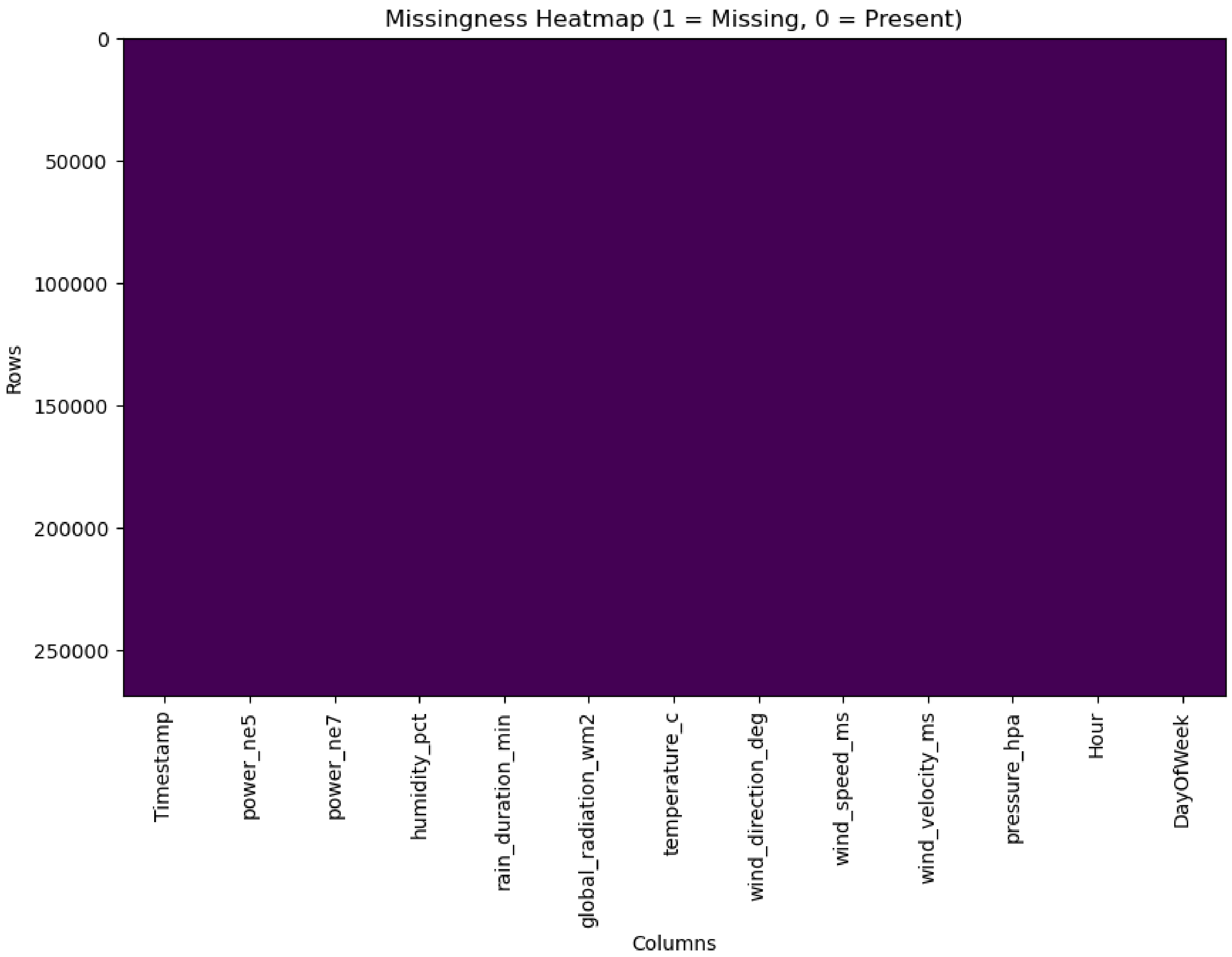
Dataset overview

Category	Details
Number of rows	268,705
Number of columns	11
Timestamp column	Timestamp
Data types	1 datetime variable, 10 numeric variables
Column names	Timestamp, power_ne5, power_ne7, humidity_pct, rain_duration_min, global_radiation_wm2, temperature_c, wind_direction_deg, wind_speed_ms, wind_velocity_ms, pressure_hpa
Time range	2015-01-01 00:00:00 → 2022-08-31 00:00:00
Sampling rate	15-minute intervals
Sampling consistency	Constant (no irregular gaps detected)
Missing values	No missing values in the raw dataset

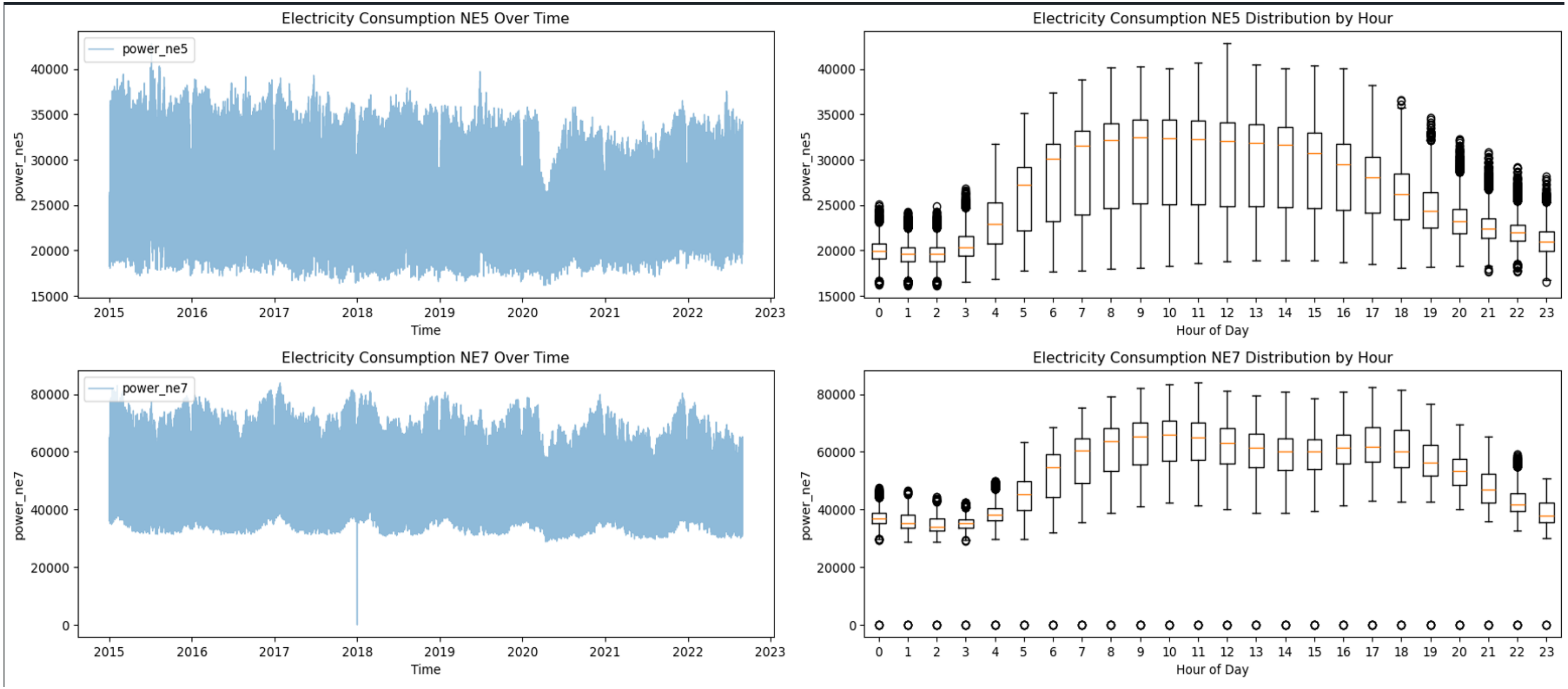
boxplot


power_ne5
15000
20000
25000
30000
35000
40000
1
power_ne7
0
20000
40000
60000
80000
1
humidity_pct
20
40
60
80
100
1
rain_duration_min
0
2
4
6
8
10
12
14
1
global_radiation_wm2
0
200
400
600
800
1000
1
temperature_c
−10
0
10
20
30
1
wind_direction_deg
0
50
100
150
200
250
300
350
1
wind_speed_ms
0
2
4
6
8
10
1
wind_velocity_ms
0
2
4
6
8
10
1
pressure_hpa
930
940
950
960
970
980
990
1

Heat map


Missingness Heatmap (1 = Missing, 0 = Present)
0
50000
100000
150000
200000
250000
Rows
Timestamp
power_ne5
power_ne7
humidity_pct
rain_duration_min
global_radiation_wm2
temperature_c
wind_direction_deg
wind_speed_ms
wind_velocity_ms
pressure_hpa
Hour
DayOfWeek
Columns

Time Series Overview



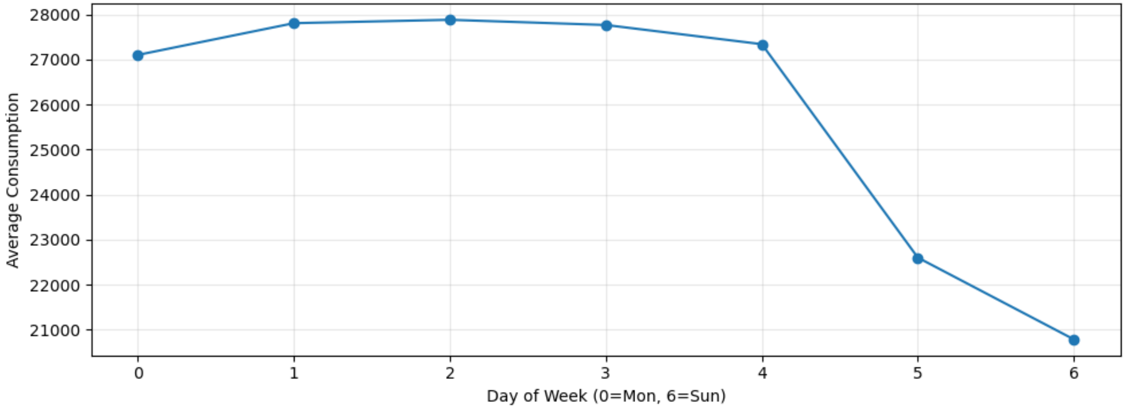
Electricity Consumption NE5 Over Time

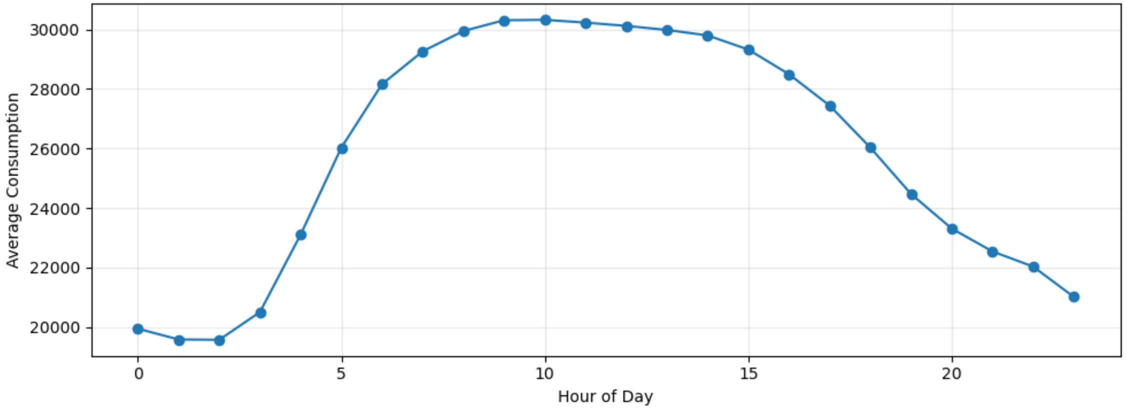
Electricity Consumption NE5 Distribution by Hour

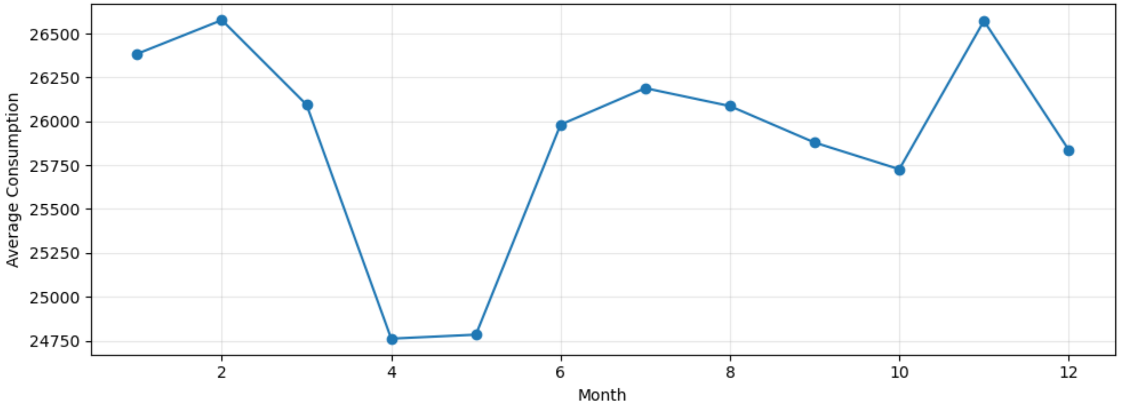
Electricity Consumption NE7 Over Time

Electricity Consumption NE7 Distribution by Hour

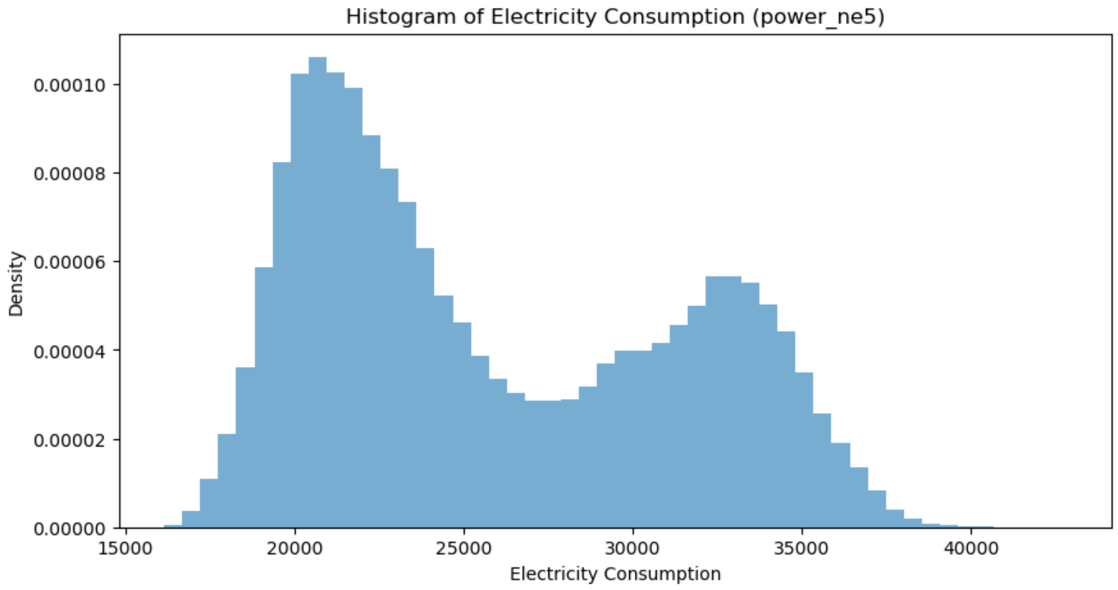
Time-series visualizations


Average Electricity Consumption by Day of Week

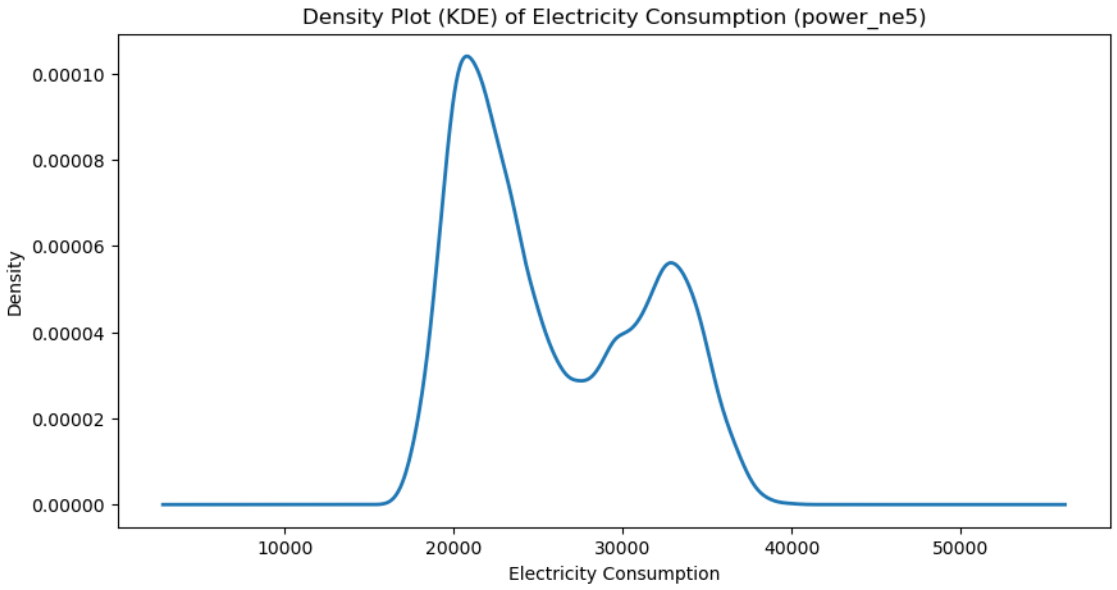

Average Electricity Consumption by Hour of Day


Average Electricity Consumption by Month

Distribution analysis

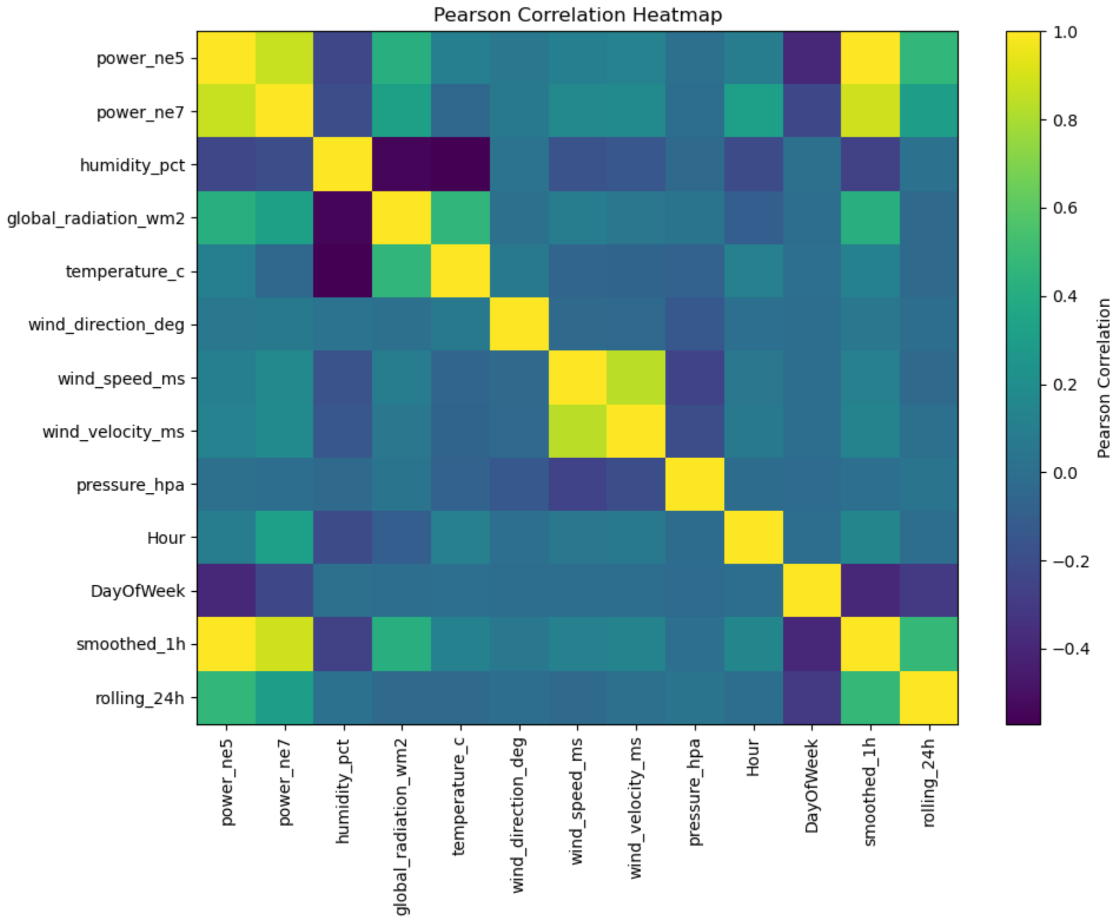


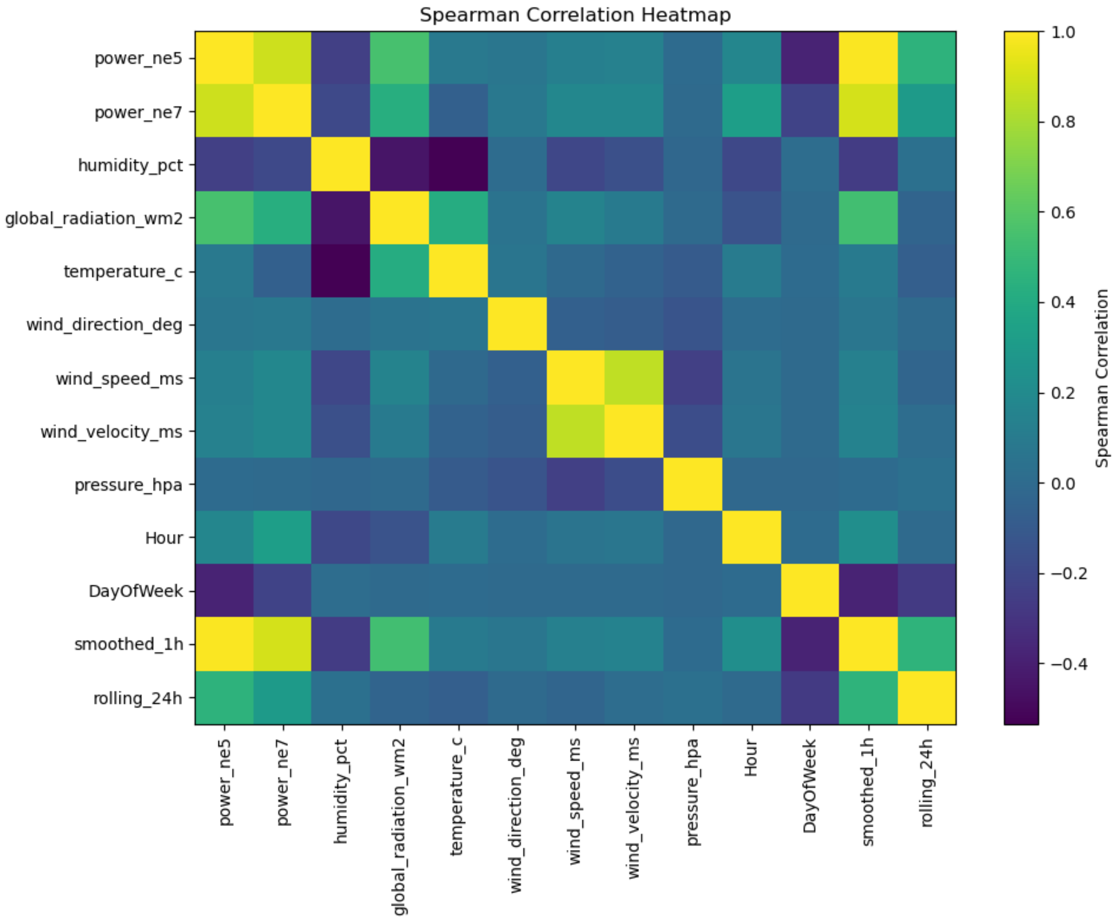
Histogram of Electricity Consumption (power_ne5)



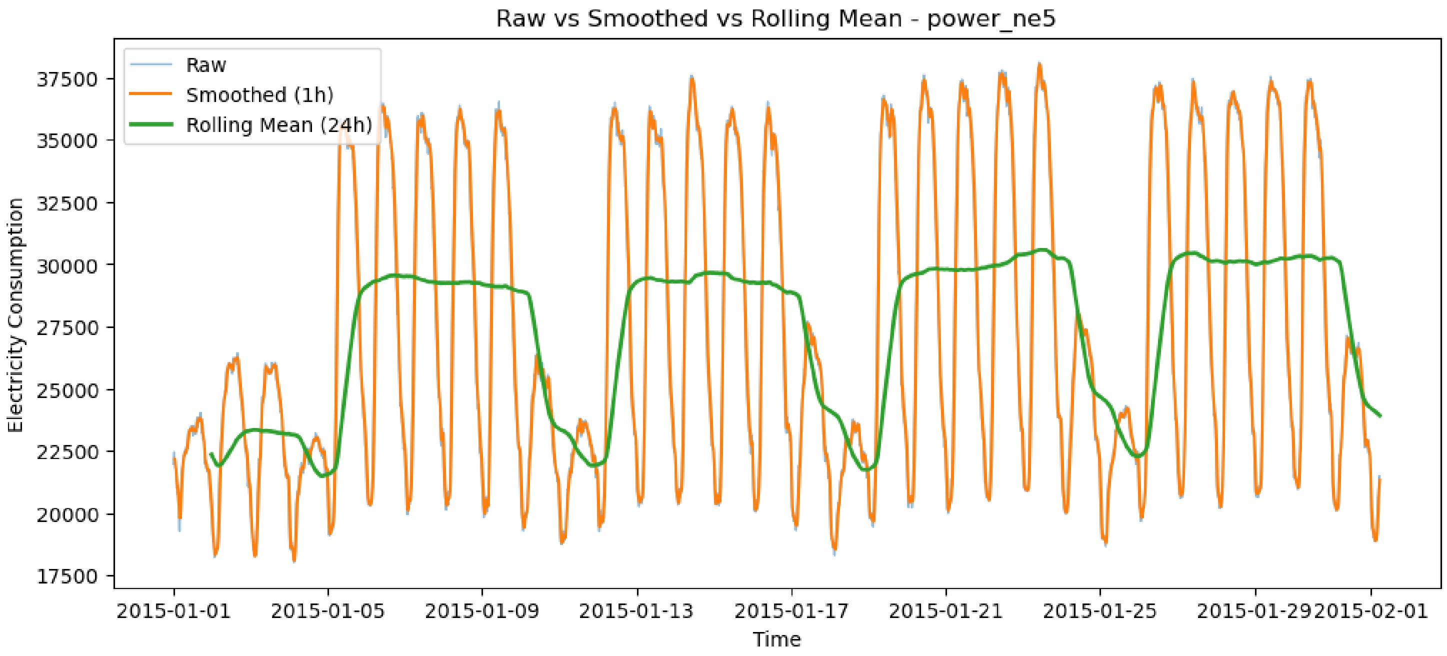
Density Plot (KDE) of Electricity Consumption (power_ne5)

Correlation analysis



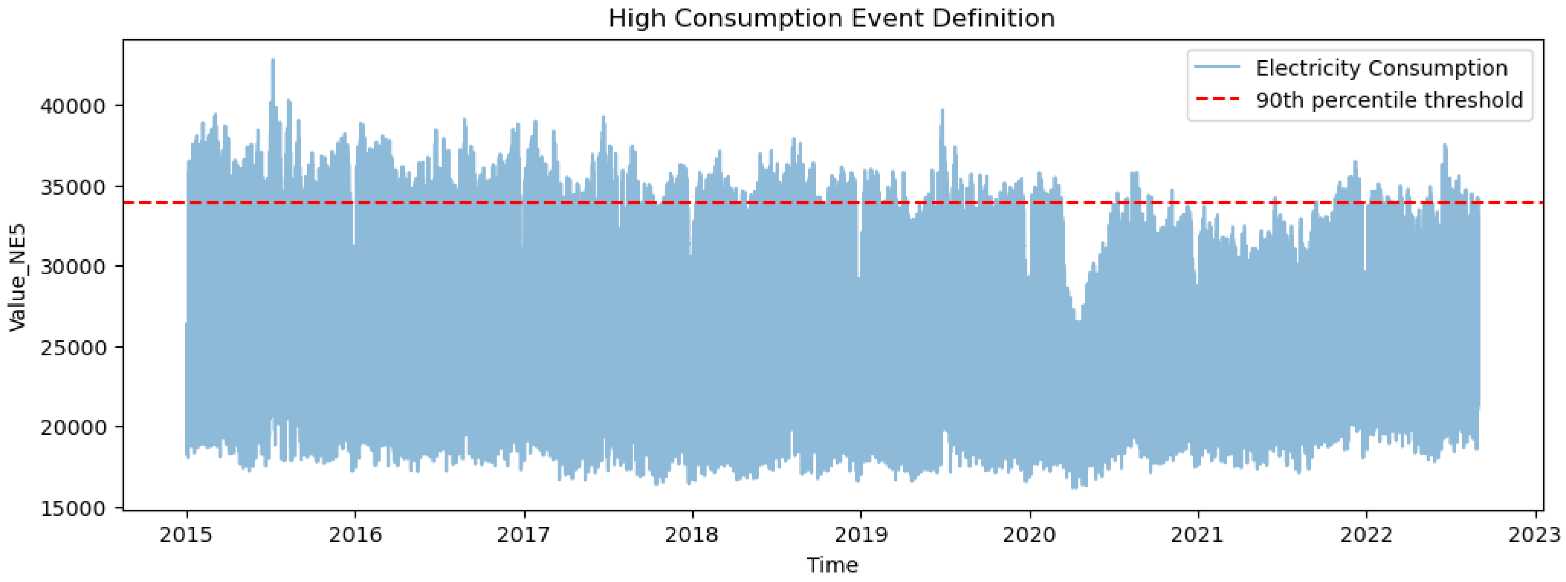


Daily pattern



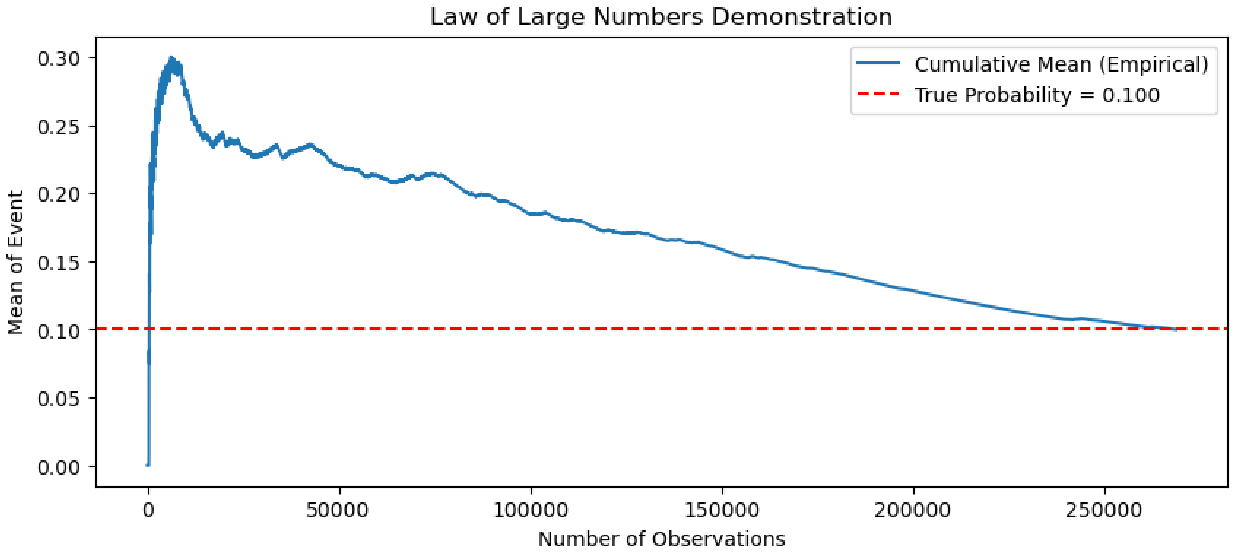
Threshold-Based Event Probability

P(High Consumption | Day) = 0.1838603072525902
P(High Consumption | Night) = 0.016144038465832546

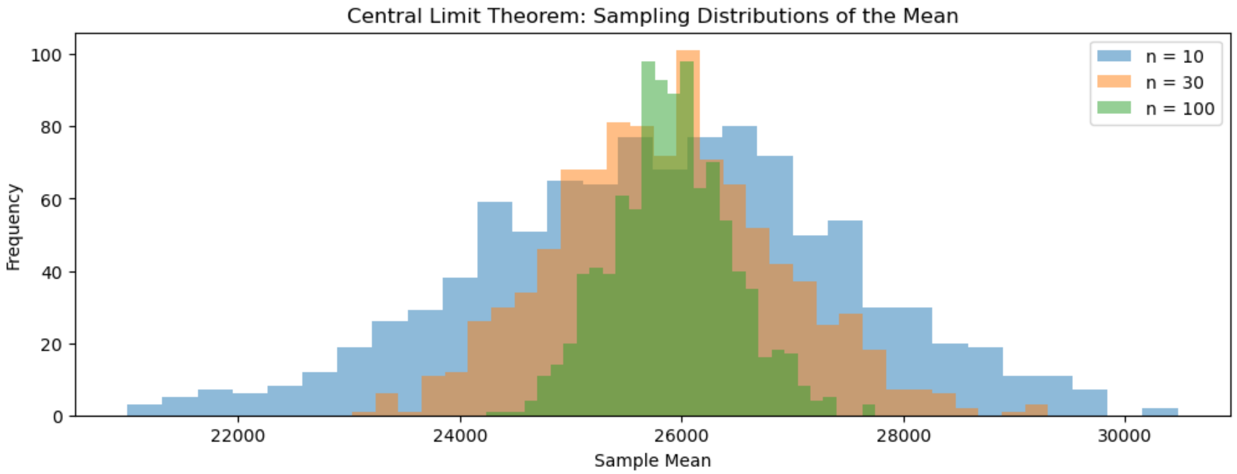


Law of Large Numbers (LLN): Cumulative Event Probability Converges

Central Limit Theorem (CLT): Sample Means Become Normal as n Increases



Blue = cumulative mean of the high-consumption event (consumption > 0.9). Red = overall event probability. As more data is added, the blue line becomes stable and approaches the red line.



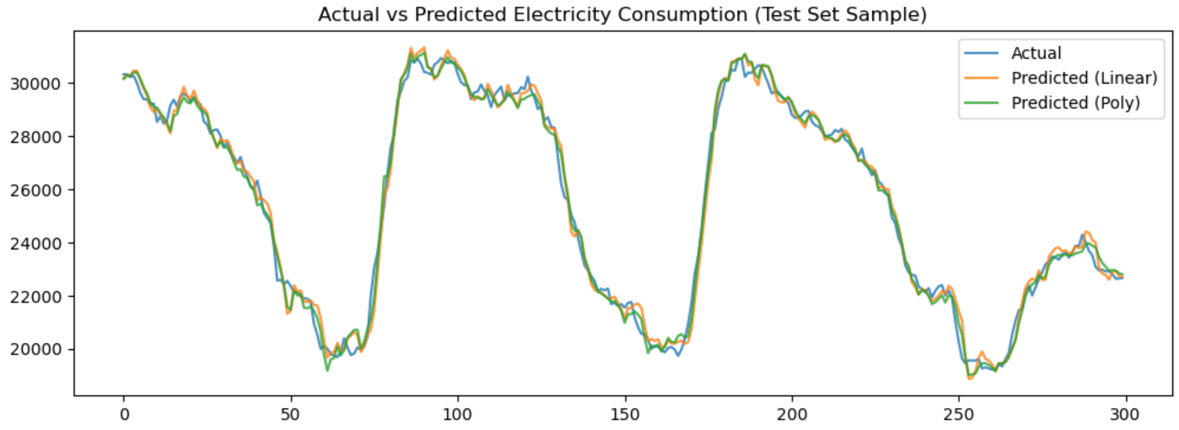
Histograms show the distribution of sample means for n = 10, 30, 100. With larger n, the shape becomes more bell-shaped and the spread becomes smaller, meaning the mean is more stable.

Model Selection: Linear vs Polynomial Regression (deg=2)

Model	RMSE	MAE
Linear Regression	429.579945	324.505781
Polynomial Regression (deg=2)	373.449804	285.163489

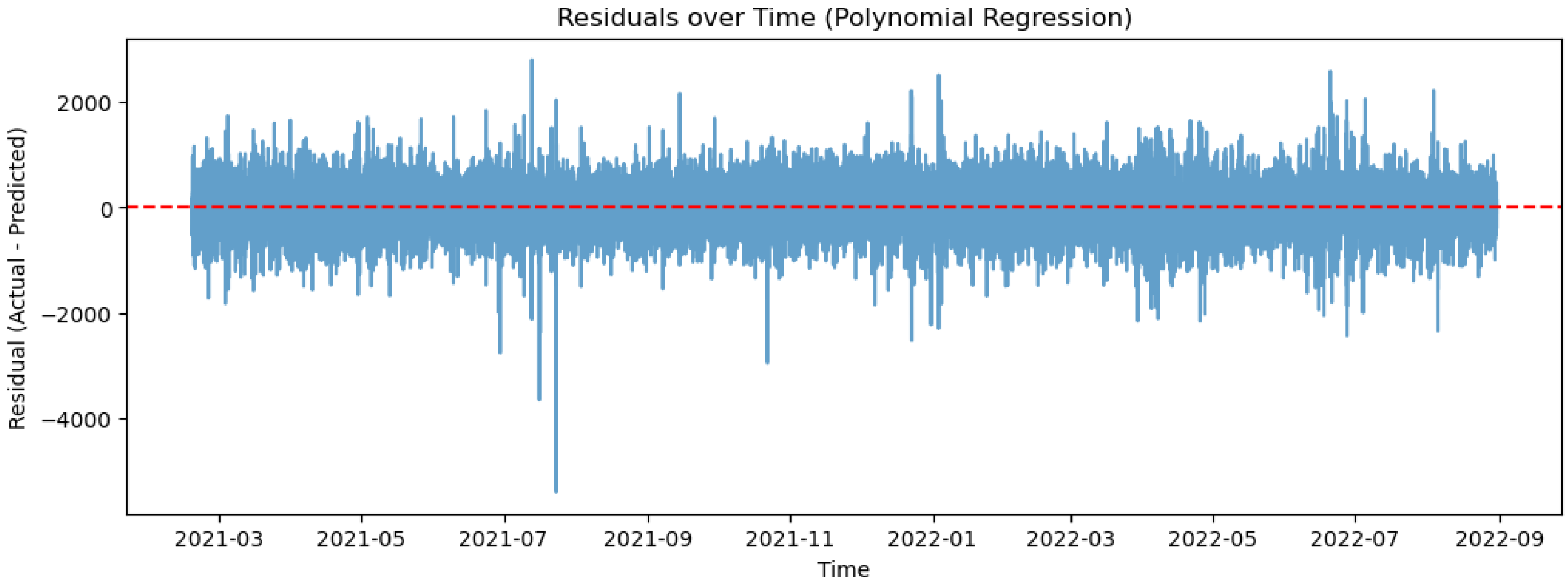
Polynomial (degree 2) gives lower errors than Linear (RMSE 373 < 429, MAE 285 < 324), so it fits the data better. This suggests a small non-linear relationship, so we choose Polynomial (deg=2).

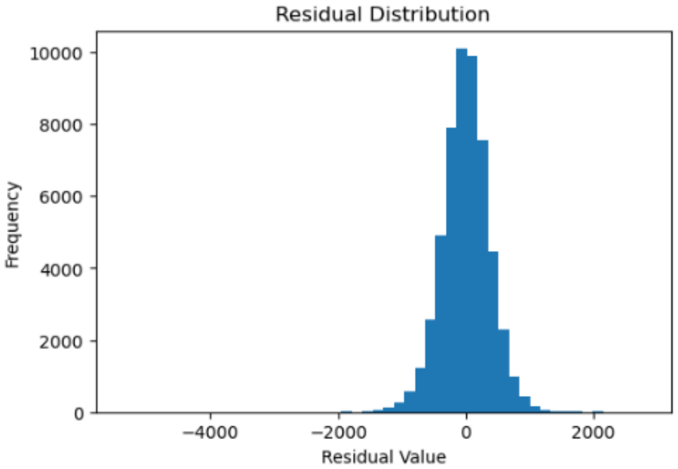
Actual vs Predicted (Test Set): Model Tracks the Main Pattern



Both models follow the real consumption trend well. The Polynomial curve is slightly closer in some parts, showing a small non-linear effect and better fit.

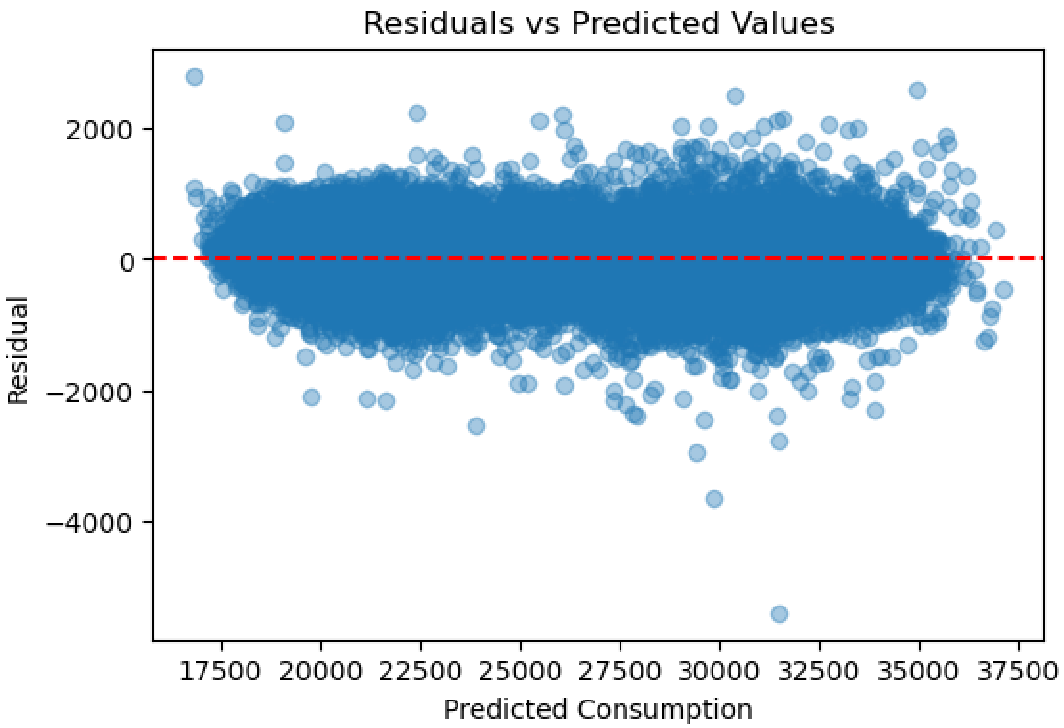
Residual Analysis (Polynomial): Mostly Small Errors, Few Large Spikes





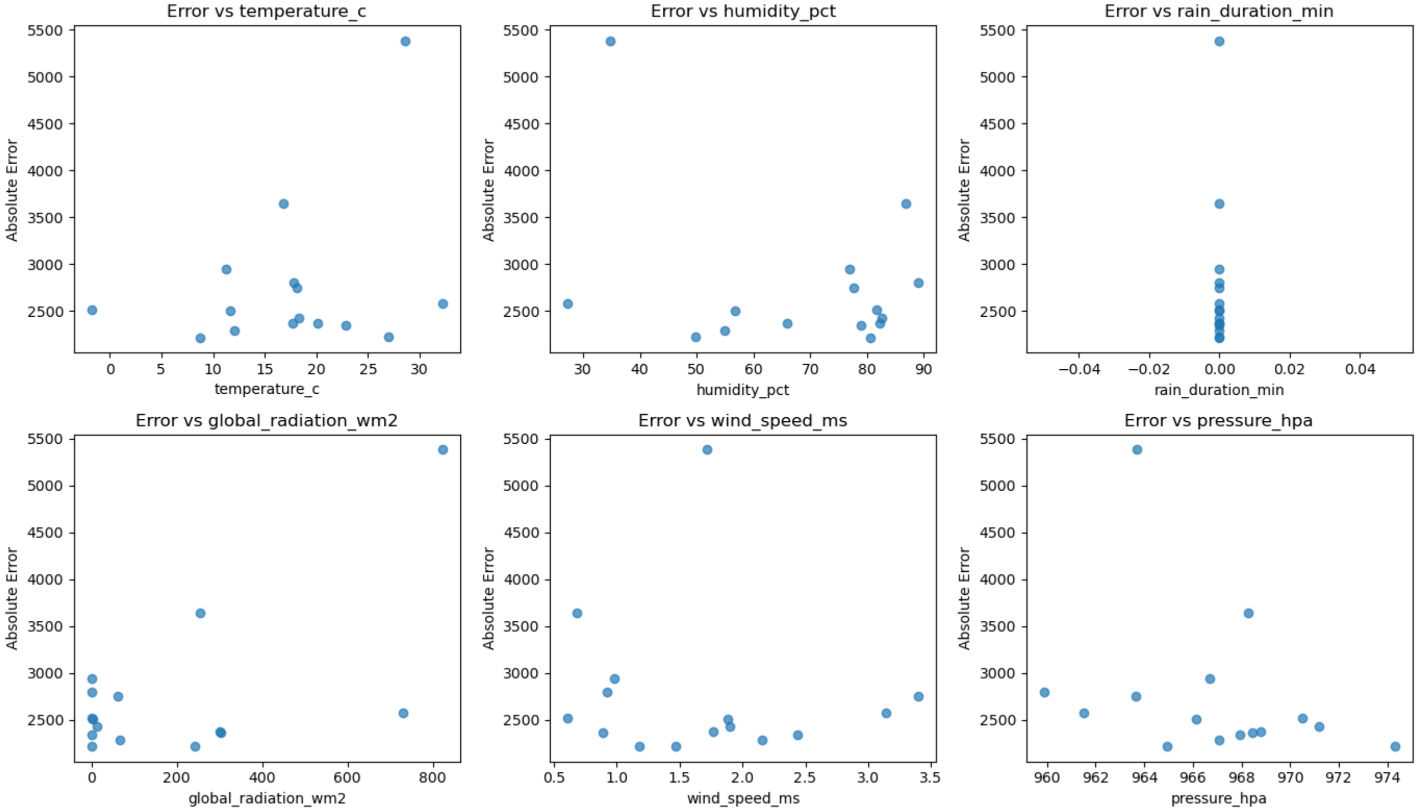
Residuals are centered near zero (no strong bias), but a few spikes and long tails show outliers—errors increase during sudden changes or rare events.

Residuals vs Predicted: Errors Increase at High Consumption



Residuals are mostly centered around zero (no clear bias), but the spread is larger at high predicted values, meaning the model is less accurate during high load and a few outliers remain.

Error vs Weather: Biggest Errors in Extreme Conditions



Most errors are scattered (no strong clear trend), but the largest errors appear in rare extreme weather—especially high humidity and very high solar radiation. Next, we should help the model learn these extreme cases better or use a stronger non-linear model.