Exploratory Data Analysis and Regression Model

Data Loading

Raw data loaded and saved as 'raw_merged_data.csv'.

Data Processing

Missing Values Summary:



Missing data is likely MAR (Missing at Random) or MNAR (Not Missing at Random)

Missing values handled using mode (categorical) or median (numerical)

Outliers detected in demand_mwh: 28496 rows

Outliers detected in temperature: 121 rows

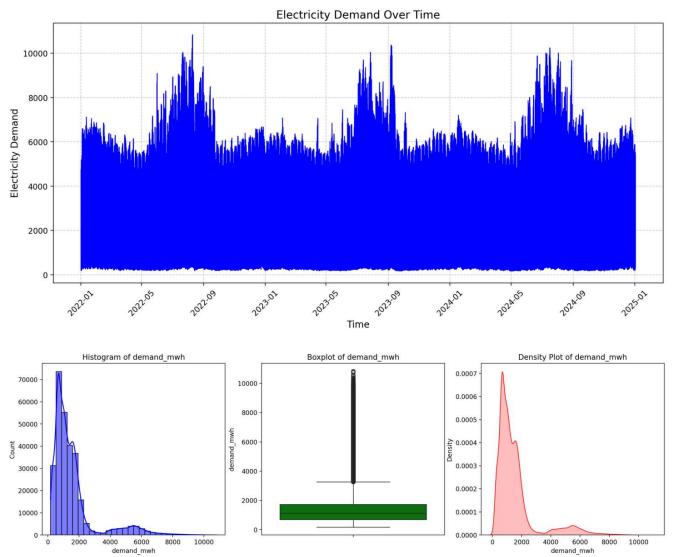
Processed data saved as 'processed_data.csv'.

Exploratory Data Analysis (EDA)

	Mean	Median	Standard Deviation	Variance	Skewness	Kurtosis	Min	Ma
demand_mwh	1,556.9376	1,120	1,455.5686	2,118,680.0336	2.4044	5.9275	168	1
temperature	11.3461	10.8085	8.0957	65.5407	0.1798	-0.6072	-10.4915	37
hour	11.5022	12	6.9217	47.9106	-0.0005	-1.2039	0	
day	15.7219	16	8.8077	77.5754	0.0067	-1.1948	1	
month	6.5176	7	3.4509	11.9089	-0.0082	-1.209	1	
year	2,023.0029	2,023	0.8186	0.67	0.0051	-1.4823	2,022	
day_of_week	2.9987	3	2.0017	4.0067	0.0016	-1.2524	0	
is_weekend	0.2861	0	0.4519	0.2043	0.9465	-1.1042	0	

localhost:8502 1/11

02/03/2025, 22:33



main

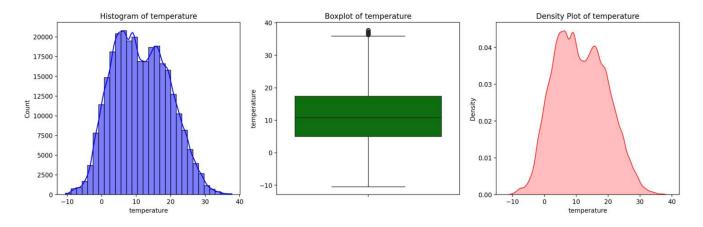
Statistical Summary for demand_mwh:

	demand_mwh
count	289,531
mean	1,556.9376
std	1,455.5686
min	168
25%	696
50%	1,120
75%	1,725
max	10,830

Skewness: 2.404388716273257

Kurtosis: 5.9275798229111425

localhost:8502 2/11

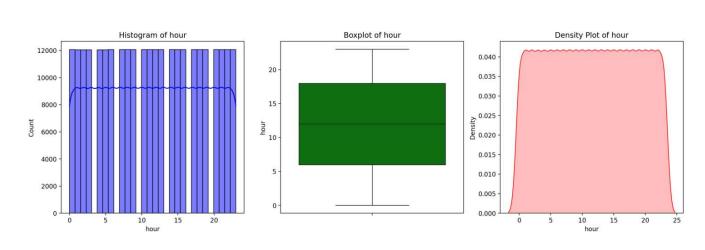


Statistical Summary for temperature:

	temperature
count	289,531
mean	11.3461
std	8.0957
min	-10.4915
25%	5.0085
50%	10.8085
75%	17.4085
max	37.7085

Skewness: 0.17975472735212175

Kurtosis: -0.6071541002206962

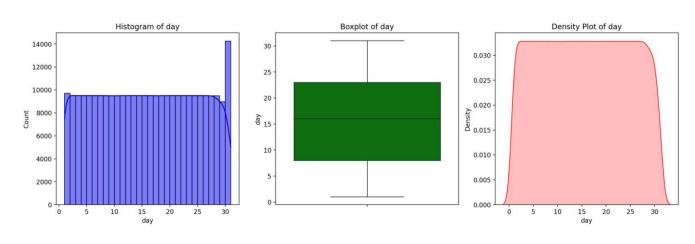


Statistical Summary for hour:

	hour
count	289,531
mean	11.5022
std	6.9217
min	0
25%	6
50%	12
75%	18
max	23

Skewness: -0.00046360804247514456

Kurtosis: -1.2038709206037737

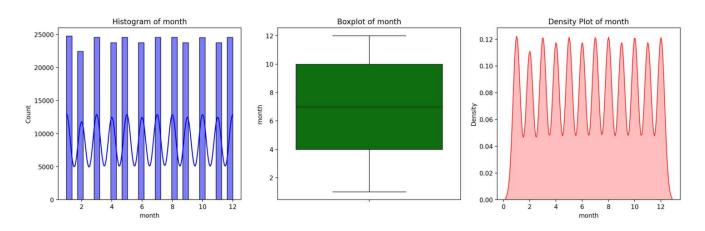


• Statistical Summary for day:

	day
count	289,531
mean	15.7219
std	8.8077
min	1
25%	8
50%	16
75%	23
max	31

Skewness: 0.0066659479749006795

Kurtosis: -1.1947627911996819



Statistical Summary for month:

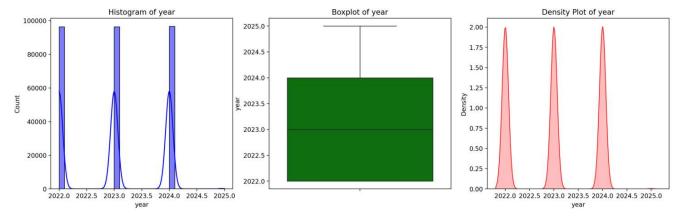
	month
count	289,531
mean	6.5176
std	3.4509
min	1
25%	4
50%	7
75%	10
max	12

Skewness: -0.0082259912991205

Kurtosis: -1.208967106110422

localhost:8502 5/11



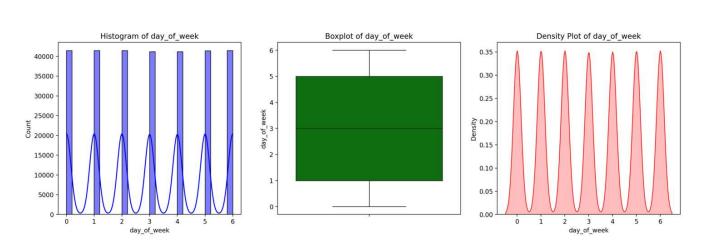


Statistical Summary for year:

	year
count	289,531
mean	2,023.0029
std	0.8186
min	2,022
25%	2,022
50%	2,023
75%	2,024
max	2,025

Skewness: 0.005073054080100816

Kurtosis: -1.4822670264756326



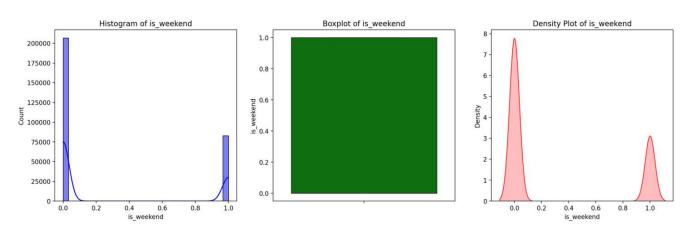
Statistical Summary for day_of_week:

localhost:8502 6/11

	day_of_week
count	289,531
mean	2.9987
std	2.0017
min	0
25%	1
50%	3
75%	5
max	6

Skewness: 0.001632828515679579

Kurtosis: -1.2524354837870715

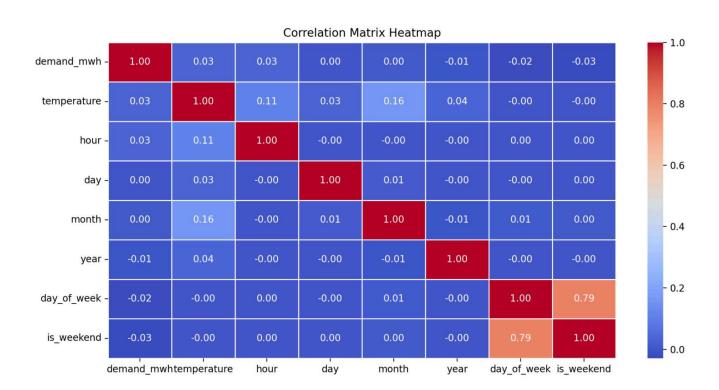


Statistical Summary for is_weekend:

	is_weekend
count	289,531
mean	0.2861
std	0.4519
min	0
25%	0
50%	0
75%	1
max	1

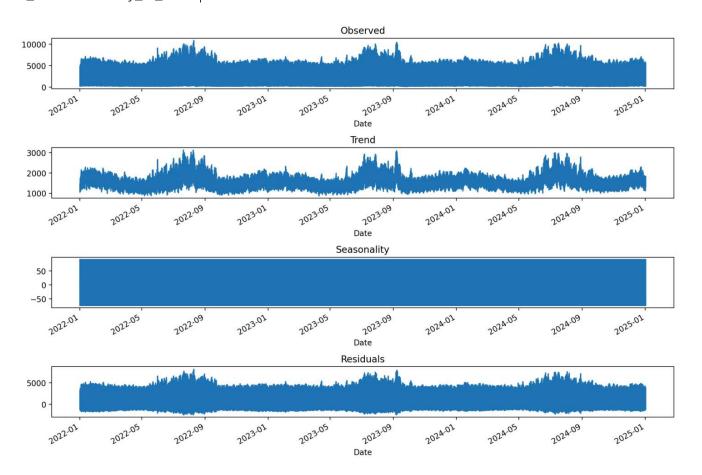
Skewness: 0.9464819917733677

Kurtosis: -1.1041794666625326



• Highly Correlated Feature Pairs (|correlation| > 0.75):

is_weekend ↔ day_of_week | Correlation: 0.79



Augmented Dickey-Fuller Test Results:

	0
Test Statistic	-36.1998
p-value	0
# Lags Used	89
# Observations Used	289,441
Critical Value (1%)	-3.4304
Critical Value (5%)	-2.8615
Critical Value (10%)	-2.5668

✓ The time series is stationary (p-value < 0.05).

EDA performed on processed data.

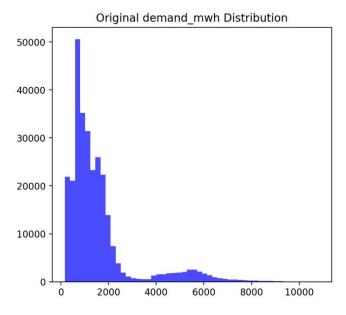
Outlier Detection

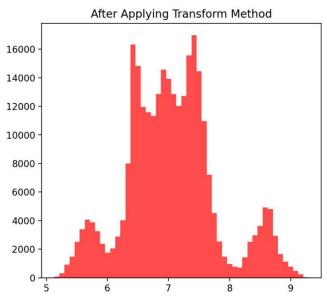
IQR Outliers (demand_mwh): 28496

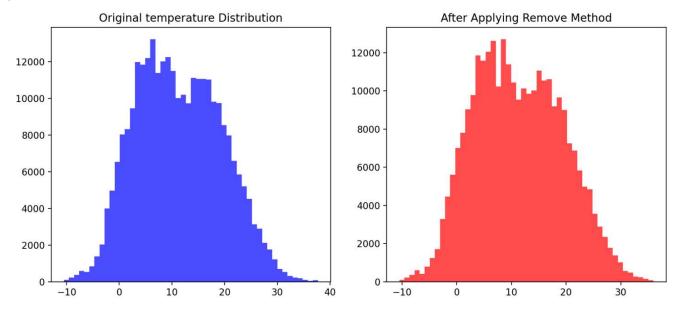
Z-score Outliers (demand_mwh): 8287

IQR Outliers (temperature): 121

Z-score Outliers (temperature): 132







Outliers detected and cleaned data saved as 'cleaned_data.csv'.

Normalization

Normalized data saved as 'normalized_data.csv'.

Regression Model

Model Performance Metrics:

MSE: 0.55

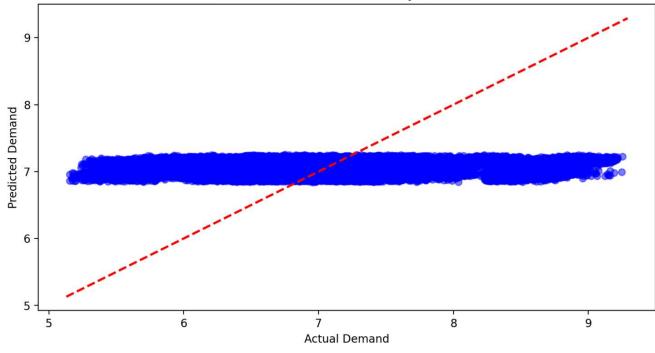
RMSE: 0.74

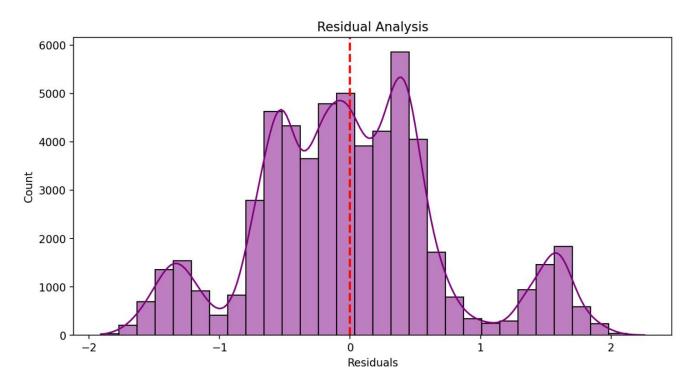
R² Score: 0.02

Model Score: 0.02

localhost:8502 10/11

Actual vs Predicted Electricity Demand





Regression model built and evaluated.