# Household Electricity Consumption Data Analysis Report

## Overview

This report presents a detailed analysis of household electricity consumption data, focusing on anomaly detection, correlation analysis, and consumption pattern analysis. The analysis was performed on Week 4 of the dataset, examining various electrical measurements including power consumption, voltage, and current intensity.

## 1. Anomaly Detection and Data Preprocessing

### 1.1 Data Preprocessing

The initial phase involved handling missing values in the dataset through linear interpolation. This method was chosen because:

- It preserves the temporal relationship between data points
- It provides reasonable estimates for missing values based on adjacent known values
- It maintains the overall trend of the time series data

The implementation used R's approx() function within a custom interpolate_na() function that handles both numeric and non-numeric columns appropriately.

### 1.2 Anomaly Detection Methodology

Point anomalies were detected using Z-score analysis, which measures how many standard deviations a data point is from the mean. The process involved:

- Calculating Z-scores for each numeric feature using the scale() function
- Identifying points where |Z-score| > 3 as anomalies
- Computing anomaly percentages per feature and for the entire dataset

### 1.3 Results

The following results summarize the total anomalies detected for each feature and their percentage of total data points:

| Feature | Anomalies Detected | Percentage of Anomalies |
|---|---|---|
| Global Active Power | 8,718 | 1.66% |
| Global Reactive Power | 5,606 | 1.07% |
| Voltage | 2,890 | 0.55% |
| Global Intensity | 9,623 | 1.83% |
| Sub-Metering 1 | 14,925 | 2.84% |
| Sub-Metering 2 | 14,868 | 2.83% |
| Sub-Metering 3 | 23 | 0.004% |

Complete Dataset Anomalies

- Total anomalies detected across all features: 56,653
- Overall anomaly percentage (across all features and observations): 1.539818% = 1.54% (Approximately)

## 2. Pearson's Correlation Calculation

### 2.1 Methodology

Pearson correlation coefficients were calculated for all pairs of variables using the cor() function with method="pearson". This analysis focused on Week 4 data to understand relationships between different electrical measurements.

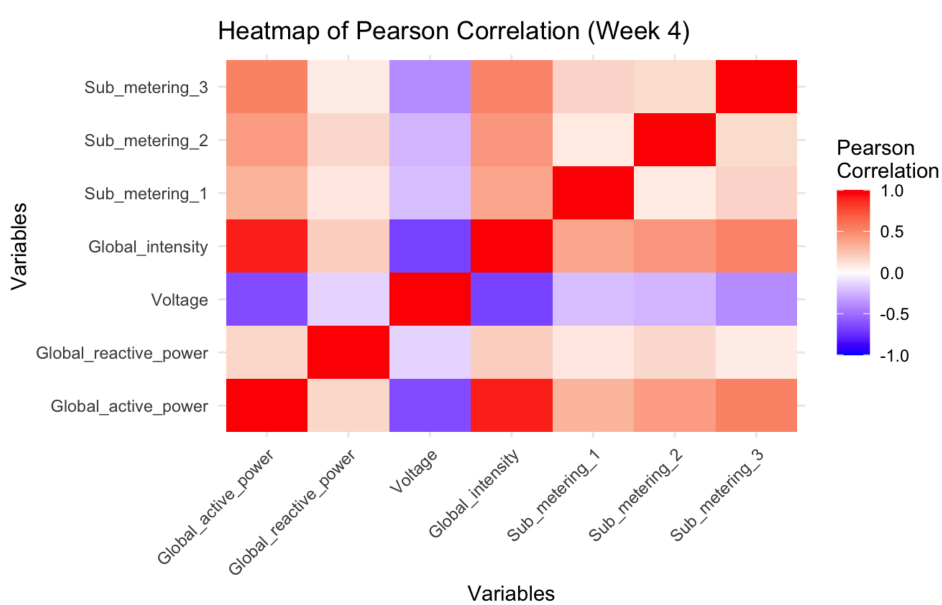
### 2.2 Key Findings

The correlation analysis helps identify relationships between variables such as:

- Global active power and global intensity
- Sub-metering relationships
- Voltage relationships with other measurements

There is a strong positive correlation between **Global Intensity** and **Global Active Power**, indicating that as power consumption increases, so does the current drawn.

Additionally, there is a strong negative correlation between **Voltage** and both **Global Active Power** and **Global Intensity**, suggesting that higher power consumption tends to cause a drop in voltage levels.



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | 1.00 | 0.17 | -0.65 | 0.91 | 0.32 | 0.41 | 0.53 |
| B | 0.17 | 1.00 | -0.14 | 0.20 | 0.09 | 0.15 | 0.07 |
| C | -0.65 | -0.14 | 1.00 | -0.68 | -0.21 | -0.25 | -0.39 |
| D | 0.91 | 0.20 | -0.68 | 1.00 | 0.38 | 0.44 | 0.52 |
| E | 0.32 | 0.09 | -0.21 | 0.38 | 1.00 | 0.08 | 0.17 |
| F | 0.41 | 0.15 | -0.25 | 0.44 | 0.08 | 1.00 | 0.14 |
| G | 0.53 | 0.07 | -0.39 | 0.52 | 0.17 | 0.14 | 1.00 |

## 3. Global Intensity Pattern Analysis

### 3.1 Time Window Analysis

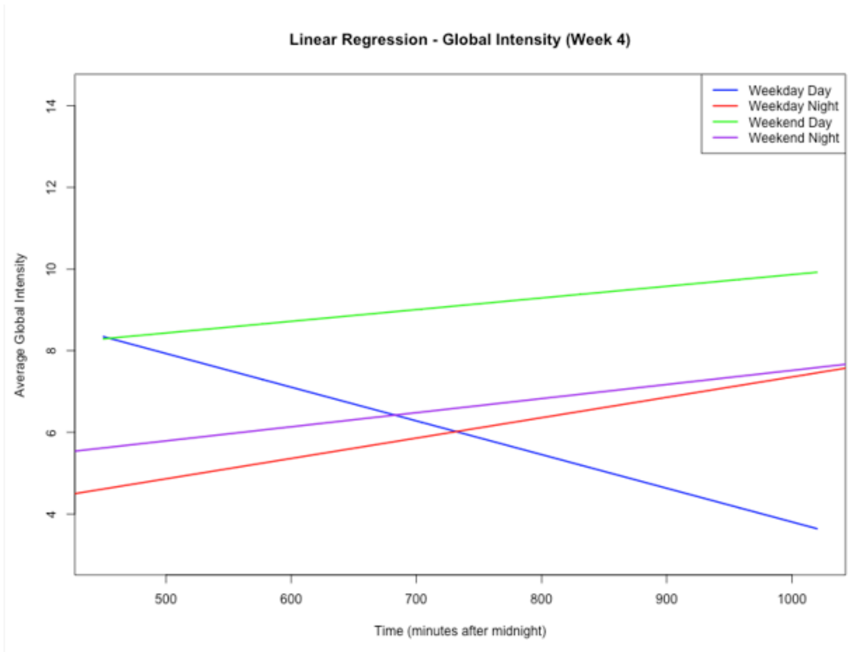
The analysis separated data into four distinct categories:

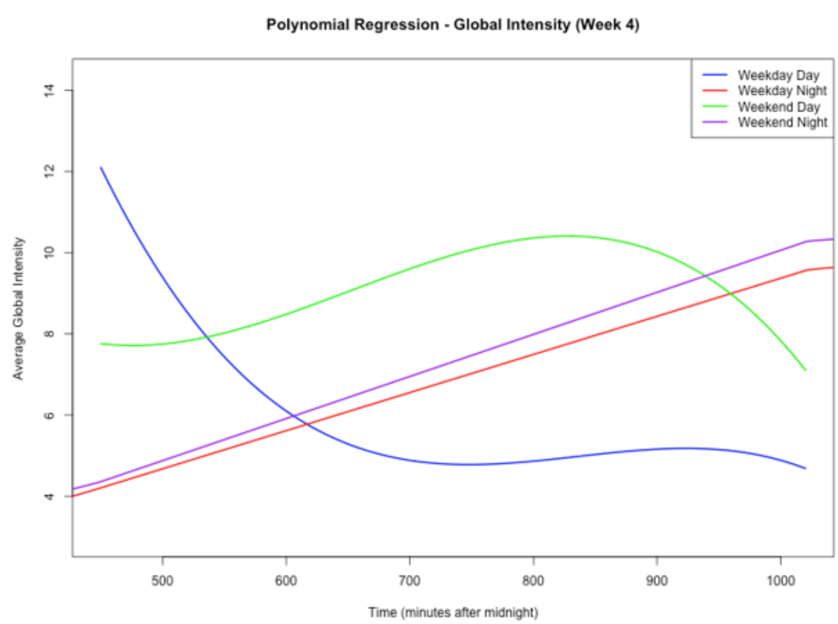
- Weekday Daytime (7:30 AM - 5:00 PM)
- Weekday Nighttime (6:00 PM - 12:00 AM)
- Weekend Daytime (7:30 AM - 5:00 PM)
- Weekend Nighttime (6:00 PM - 12:00 AM)

### 3.2 Regression Analysis

Two types of regression were performed:

- Linear Regression (First-Order)
- Polynomial Regression (Third-Order)




## 3.3 Pattern Analysis Results

### Linear Regression Findings

From the linear regression plot:

- Weekday daytime shows a declining trend in intensity
- Weekend daytime shows an increasing trend
- Both nighttime periods show moderate upward trends

- Weekend consumption generally shows higher intensity levels

**Polynomial Regression Findings**

The polynomial regression reveals more complex patterns:

- Weekday daytime shows a curved decline with some stabilization
- Weekend daytime shows a pronounced peak in the middle of the day
- Nighttime patterns show more nuanced variations compared to linear regression
- The curves better capture the natural fluctuations in power usage

### 3.4 Interpretation

The regression analysis reveals several important patterns:

- Clear differences between weekday and weekend consumption
- Different patterns between day and night usage
- Non-linear relationships in actual usage patterns
- Higher variability during daytime hours
- More stable patterns during nighttime hours

## 4. Conclusions

The analysis reveals several key insights about household electricity consumption:

- The data contains identifiable anomalies that may represent unusual usage patterns
- Strong correlations exist between certain electrical measurements
- Clear distinctions exist between weekday and weekend consumption patterns
- Polynomial regression better captures the nuanced patterns of electricity usage
- Time-of-day significantly influences consumption patterns

## 5. Technical Implementation Notes

The analysis was implemented in R using several key functions and packages:

- Data manipulation: base R function
- Statistical analysis: cor(), lm()
- Visualization: base R plotting functions
- Time series handling: POSIXct and POSIXlt classes