Delivery Bot Socket Analysis Report

Huidi “Judy” Yang

June 13, 2025

Statement of Purpose

The purpose of this report is to analyze the activity logs of smart plugs deployed in a service robot charging station and a public PC terminal, with the goal of characterizing their operational behaviors and identifying usage patterns over time. By applying time series processing, clustering techniques, and temporal profiling, we aim to uncover the structure underlying state transitions and usage routines across both devices. For the service robot, the focus is on understanding transitions between charging and non-charging states, while for the PC, we investigate more nuanced and overlapping patterns of power consumption and system activity. A comparative analysis highlights the differences in clustering interpretability, temporal regularity, and device-specific behavior. These insights can support scheduling, energy optimization, anomaly detection, and predictive maintenance in the broader context of smart facility management.

Executive Summary

This report presents an in-depth analysis of smart plug usage data from two distinct environments: a hotel delivery robot and a public computer station. The primary objective was to uncover operational patterns, energy usage behaviors, and inform potential improvements for facility-wide smart energy management.

Key methods included time series feature extraction, PCA-based dimensionality reduction, clustering (K-Means, DBSCAN, and manual), and temporal usage profiling. While clustering revealed distinct activity states for the public computer, the delivery bot exhibited more polarized, routine behavior that resisted algorithmic clustering, requiring manual segmentation.

Temporal analysis showed consistent low-power states for the PC during late night and early morning, and highly regimented charging patterns for the robot centered around specific daily intervals. The comparison highlights how different device roles shape their interaction with power infrastructure.

The insights gained can inform practical scheduling, energy optimization, and anomaly detection strategies. Moreover, the analytical framework demonstrated here—combining feature clustering and temporal profiling—can be scaled to other smart devices for real-time monitoring and predictive management across a facility.

Contents

Part I

Introduction

All datasets, product documentation, and analysis scripts referenced in this report are available in the following GitHub repository. This report is organized into five parts. Part I provides an overview of the project and describes the datasets used. Part II presents an in-depth analysis of the smart socket associated with the delivery robot. Part III focuses on the analysis of the socket connected to the hotel's public computer. Part IV offers a comparative evaluation of the two devices, highlighting differences and similarities in their usage patterns. Finally, Part V summarizes the key findings and outlines actionable insights for operational improvements.

1 About the Data

The datasets were provided by Bi-Fan Yin of 住友酒店, at the request of Alex Hon. It contains measurements recorded from two smart sockets over time. Each record includes four variables: `equipmentName`, `functionType`, `time`, and `value`.

1.1 equipmentName

This column contains two unique identifiers:

- `Zprime_Socket_01`: corresponds to the smart socket used by the autonomous delivery robot.
- `Zprime_Socket_02`: corresponds to the smart socket connected to a public PC.

1.2 functionType

There are 12 unique function types measured by the sockets. Each corresponds to a specific electrical or environmental property, as defined in the device protocol documentation (安驿电管家系列产品协议文档-*V1.07 1.pdf*):

- **CSQ**: Signal strength percentage (0.00%–100.00%, higher values indicate stronger signal)

- **Current**: Electrical current, in amperes (A)
- **LastHourPF**: Energy consumed in the previous hour, in kilowatt-hours (kWh)
- **Leakage**: Leakage current, in milliamperes (mA)
- **PartPF**: Energy consumed on the current day, in kilowatt-hours (kWh)
- **PhaseAngle**: Phase angle, in degrees
- **Power**: Active power, in watts (W)
- **PowerFactor**: Power factor (unitless)
- **RelayStatus**: Relay status (0 = open, 1 = closed)
- **Temperature**: Temperature, in degrees Celsius (°C)
- **TotalPF**: Total energy consumed, in kilowatt-hours (kWh)
- **Voltage**: Voltage, in volts (V)

1.3 time

The timestamp at which the measurement was recorded, formatted in ISO 8601 with a UTC offset.

1.4 value

The numeric reading corresponding to the specified `functionType` at the recorded time.

2 Data Preprocessing

Before beginning the analysis, I carried out a series of preprocessing steps to clean and structure the raw socket data for time series analysis and clustering.

2.1 Timestamp Parsing

The original dataset recorded timestamps in ISO 8601 format with UTC offsets. I converted these strings into `datetime` objects using `pandas.to_datetime()`, which allowed me to extract relevant temporal features such as the hour of day, weekday, and categorical time blocks. Furthermore, I converted the time to Beijing time, making it contextually appropriate and intuitive for this project.

2.2 Data Reshaping

To facilitate analysis, I reshaped the data from a long format into a wide format, creating one row per timestamp and one column per function type. This resulted in a minute-level time series covering the following features: `CSQ`, `Current`, `LastHourPF`, `Leakage`, `PartPF`, `PhaseAngle`, `Power`, `PowerFactor`, `RelayStatus`, `Temperature`, `TotalPF`, and `Voltage`.

2.3 Filtering for Relevant Equipment

The dataset included data from two devices. I filtered the dataset based on equipmentName (`Zprime_Socket_01` and `Zprime_Socket_02`) for the two parts of the analyses.

2.4 Data Cleaning

I began by checking for and removing duplicate rows. I also verified the dataset for missing values and confirmed that there were none. This ensured that all downstream analyses, such as PCA and clustering, could proceed on complete cases without the need for imputation.

2.5 Derived Features

To support clustering and behavioral profiling, I engineered several additional variables:

- `duration_min`: The amount of time the bot remained in a particular inferred state, computed after state assignment.
- `time_of_day`: A categorical variable representing broad temporal bins (Morning, Afternoon, Evening, Night).

- `2hr_bin`: A finer-grained time grouping based on two-hour intervals (e.g., `08:00--10:00`).

These preprocessing steps provided the structure necessary to extract meaningful patterns from the socket data.

Part II

Delivery Bot Analysis

3 Exploratory Data Analysis (EDA)

After preprocessing the data, I conducted EDA to get a general sense of the structure and the distribution of the data.

3.1 Correlation Analysis

To identify potential redundancy among the socket features, I computed the Pearson correlation matrix across all `functionType` variables. The resulting heatmap is shown in Figure 1. Several strong positive correlations are evident—for example, `Current` exhibits high correlations with `PhaseAngle`, `Power`, and `PowerFactor`. These three functions are also strongly correlated with each other, suggesting they may be measuring overlapping aspects of power consumption. This multicollinearity motivated the use of dimensionality reduction techniques in the subsequent analysis.

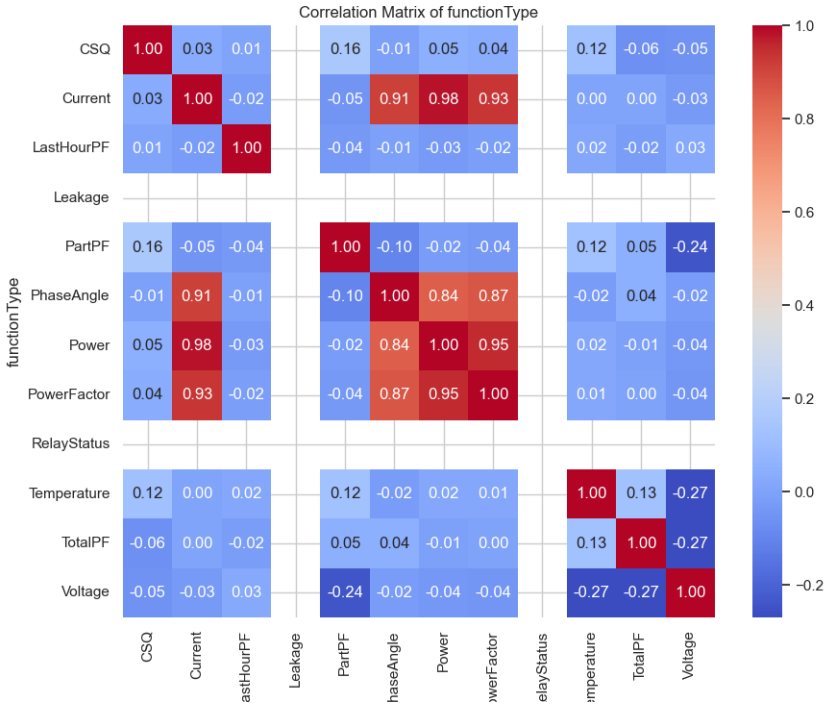

Figure 1: Heatmap of the correlation between the functions. Warmer color suggests a larger positive correlation, while cooler color indicates a small correlation.

3.2 Distribution of the Functions

To better understand the behavior of the smart socket associated with the delivery bot, I visualized the distribution of each recorded electrical function. These functions correspond to key electrical and signal metrics, many of which may reveal patterns in energy usage, device state transitions, or operational anomalies.

3.2.1 Signal Strength (CSQ)

`CSQ` represents the quality of the cellular signal, expressed as a percentage. This metric is essential for evaluating the reliability of wireless communication,

particularly for remote devices like delivery bots that depend on mobile networks to report telemetry or receive instructions.

Figure 2 shows the distribution of signal strength readings. The values form a multimodal distribution, primarily centered between 65% and 75%, suggesting that the device typically operates in areas with moderate to strong signal quality. Peaks in this range may correspond to regular stopping points (e.g., docking stations), while the lack of extreme values implies generally stable connectivity.

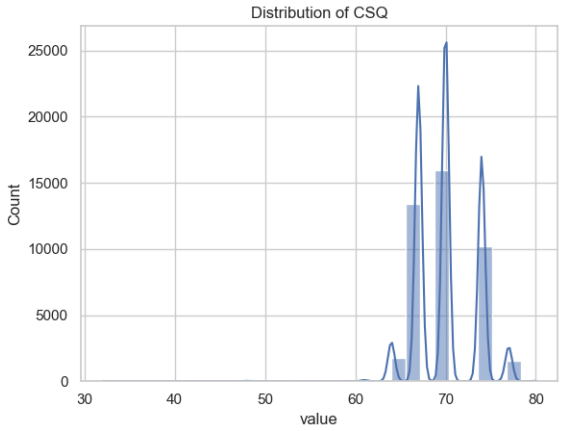

Figure 2: Distribution of signal strength (`CSQ`) values.

3.2.2 Current

`Current` measures the flow of electrical charge drawn by the socket, expressed in amperes (A). This feature is directly related to the bot's power consumption and can be used to infer periods of activity (e.g., charging or task execution).

As shown in Figure 3, the distribution is highly skewed, with most values near zero. This indicates long periods of inactivity or low power draw. A smaller cluster near 0.6 A suggests periodic charging events, aligning with expected robot behavior during idle docked states.

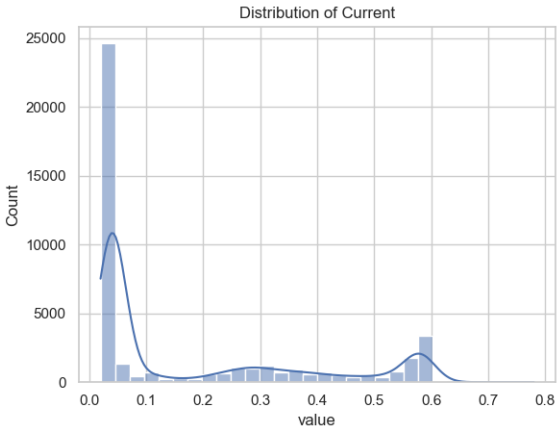

Figure 3: Distribution of electrical current (`Current`) drawn by the socket.

3.2.3 Last Hour Power Factor

`LastHourPF` reflects the power factor (i.e., the efficiency of energy usage) averaged over the past hour. A low value suggests reactive or inefficient power usage, while a value closer to 1 indicates more efficient usage.

In Figure 4, most values lie between 0.02 and 0.06, suggesting generally low power draw over time. This aligns with the bot' s role, which likely involves short bursts of activity interspersed with longer idle periods. The rhythmic peaks in the density plot imply regular operational patterns, possibly aligned with delivery schedules.

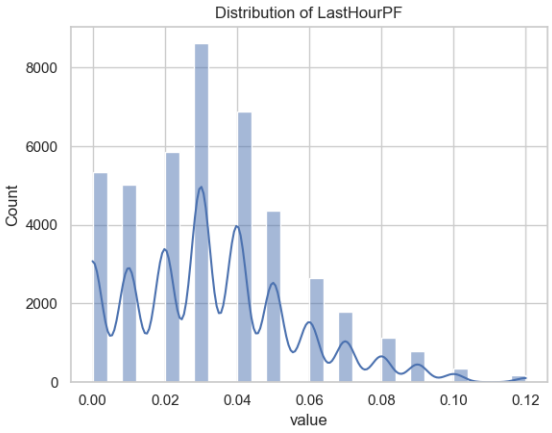

Figure 4: Distribution of energy consumption in the previous hour (`LastHourPF`).

3.2.4 Leakage

`Leakage` measures the presence of unwanted current escaping the intended circuit path - typically a safety concern. Persistent leakage may signal insulation faults or equipment degradation.

Figure 5 shows that all values of `Leakage` are zero. This suggests either that the socket consistently operated within safe electrical limits or that leakage monitoring was disabled or not applicable in this setup.

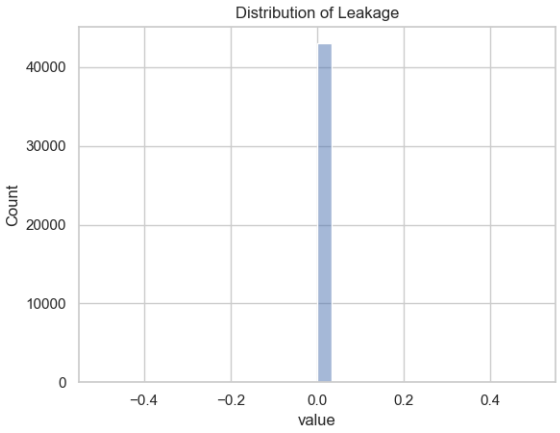

Figure 5: Distribution of `Leakage`: All values are zero.

3.2.5 Partial Power Factor (PartPF)

`PartPF` is a partial or momentary estimate of power factor, typically computed over a shorter window than the full-hour version. It provides a more granular snapshot of energy efficiency in real time.

Figure 6 shows a relatively broad and flat distribution from 0 to 0.8, with a slight peak near 0.2. This indicates variability in load characteristics throughout the day, possibly reflecting transient states during transitions between charging and idle.

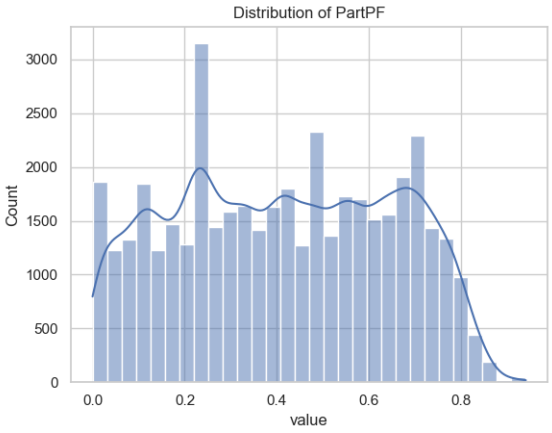

Figure 6: Distribution of `PartPF`.

3.2.6 Phase Angle

`PhaseAngle` measures the angular difference between voltage and current waveforms. It' s closely related to the type of electrical load; large shifts can indicate transitions between resistive and inductive loads (e.g., idle vs. charging states).

Figure 7 reveals a strong bimodal distribution with peaks around 115° and 175°, implying the socket alternates between two dominant states—likely corresponding to idle and active usage. This makes `PhaseAngle` a valuable signal for state classification in later analyses.

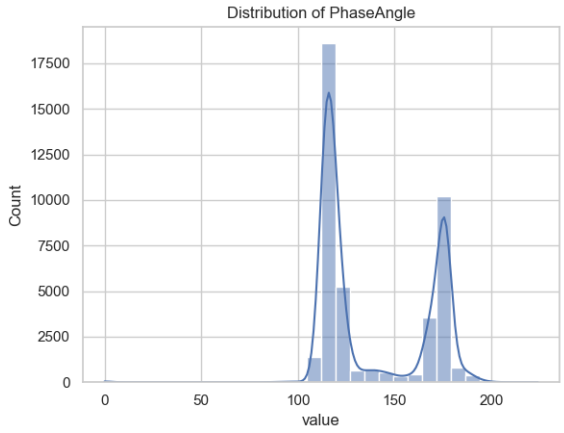

Figure 7: Distribution of `PhaseAngle`.

3.2.7 Power

Figure 8 shows the distribution of `Power`, which measures real power consumption in watts. Most readings are near zero, suggesting the socket is idle or consuming minimal energy for a majority of the time. However, a secondary cluster appears near 130 W, likely indicating active charging sessions.

`Power` is a critical indicator of when the bot is engaged in charging behavior, and its distribution supports our broader classification into "charging" and "activity" states.

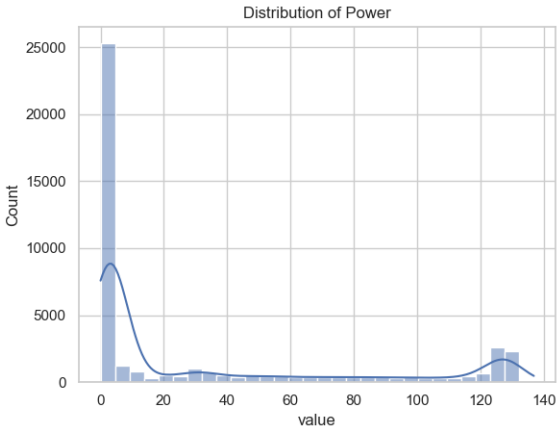

Figure 8: Distribution of real power usage (`Power`).

3.2.8 Power Factor

Figure 9 presents the distribution of the `PowerFactor`, defined as the ratio of real power to apparent power in an alternating current (AC) system. It ranges from 0 (all reactive power) to 1 (purely resistive load). The data show a clear bimodal pattern, with one peak near 0.3 and another near 0.95.

This duality may correspond to the two operational modes: one with inefficient power usage (e.g., standby or irregular loads), and one with efficient power consumption, likely during active charging.

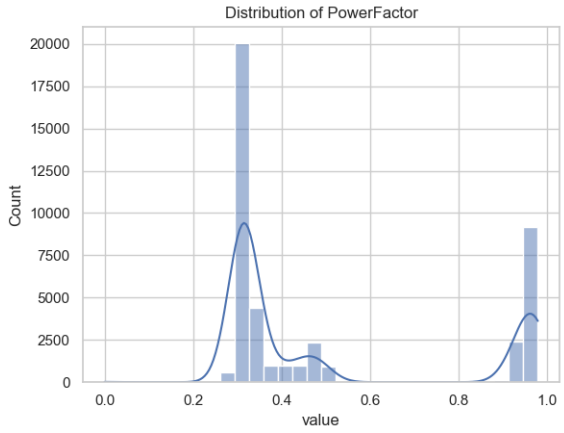

Figure 9: Distribution of `PowerFactor`.

3.2.9 Relay Status

Figure 10 shows the `RelayStatus`, a binary indicator of whether the relay switch in the socket was on (1) or off (0). All values are 1, suggesting the relay remained continuously engaged during the observation window.

This lack of variability suggests that power was always permitted to flow through the socket, further supporting the notion that the bot was either charging or idle with the relay active, but never fully powered down.

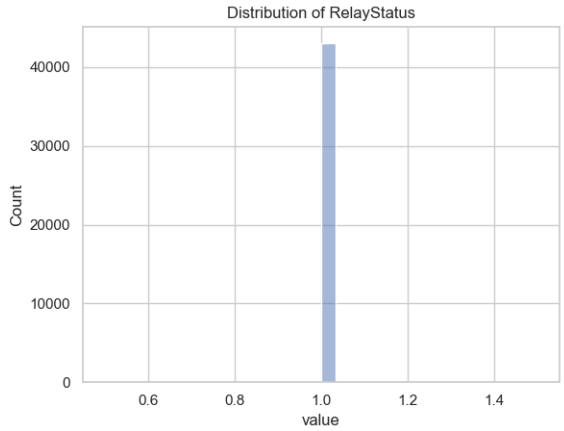

Figure 10: Distribution of relay status (`RelayStatus`).

3.2.10 Temperature

Figure 11 illustrates the distribution of internal socket temperature readings. The values cluster around typical ambient indoor temperatures (35–38°C), with multiple sharp peaks, potentially reflecting periodic measurement artifacts or rounding in sensor precision.

Monitoring `Temperature` can help identify overheating risks or detect anomalies in socket usage patterns. However, the consistency here suggests the environment was stable throughout the monitoring period.

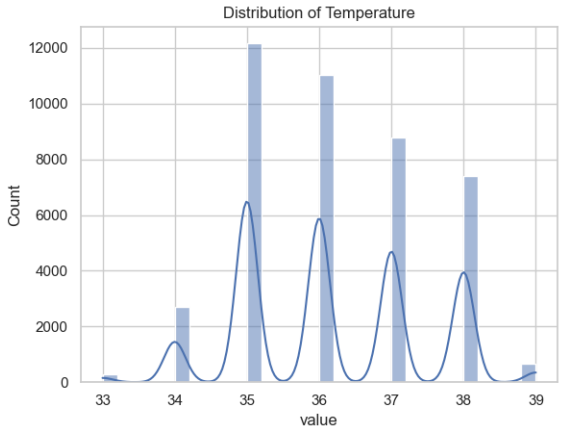

Figure 11: Distribution of internal temperature (`Temperature`).

3.2.11 Total Power Factor

Figure 12 presents the `TotalPF` distribution, which may represent a smoothed or averaged power factor metric. The values range from 50 to 75 and are distributed relatively uniformly, with no strong peaks or skew.

This could reflect aggregated socket performance across intervals or different phases of usage, although the units or context for this feature are not well documented. It may be less immediately informative without further clarification.

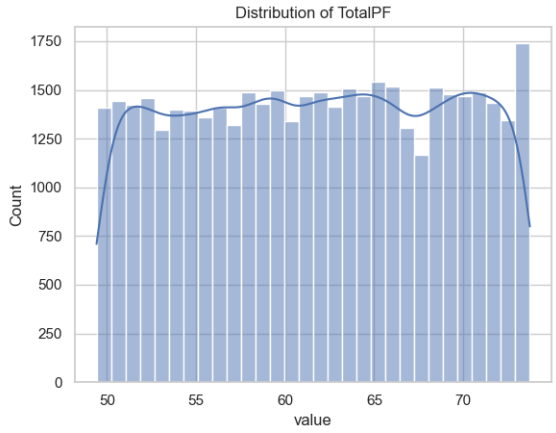

Figure 12: Distribution of total power factor (`TotalPF`).

3.2.12 Voltage

Figure 13 shows the distribution of voltage across the socket. The values form a symmetric, bell-shaped curve centered around 225 V, which is typical for AC voltage in many industrial or commercial environments.

Stable `Voltage` readings confirm consistent power delivery to the socket and suggest that voltage fluctuations are unlikely to play a major role in driving usage patterns.

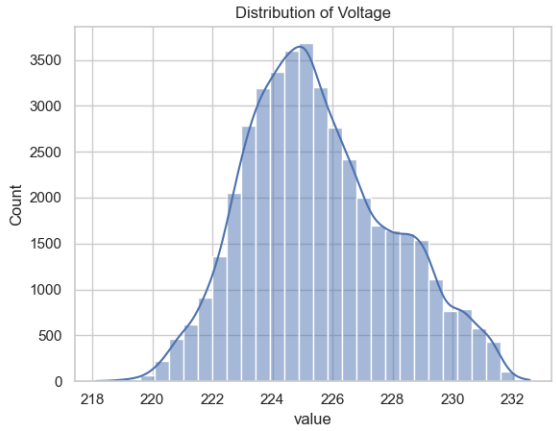

Figure 13: Distribution of line voltage (`Voltage`).

3.3 Time Series Analysis

To better understand the temporal behavior of each function recorded by the smart socket, I conducted a time series analysis across the entire observation window. While static distributions provide insights into typical ranges and patterns, they do not capture fluctuations, periodicity, or trends that evolve over time. Since the smart socket interacts with a mobile robot that follows task-based routines, it is essential to examine how electrical and environmental parameters change over the course of days and weeks.

Figure 14 displays time series plots for each function. Several patterns stand out:

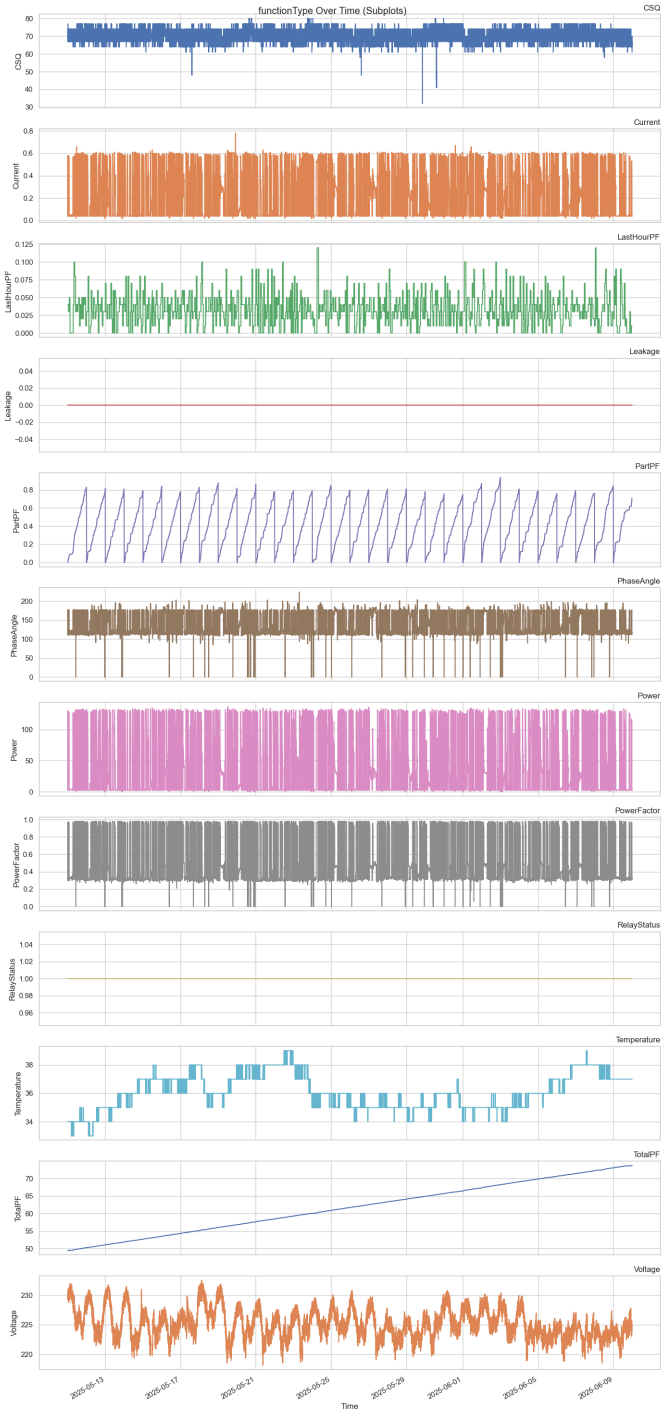

Figure 14: Time series plots of all function values over the data collection period.

- **CSQ (Signal Strength)** remains relatively stable but exhibits slight fluctuations, likely reflecting changes in network quality as the bot moves through different areas.
- **Current, Power, PowerFactor**, and **PhaseAngle** show synchronized periods of increased activity, indicating coordinated charging or high-load periods. These coincide with near-zero values in other periods, reflecting idle or standby states.
- **LastHourPF** appears noisy but trends downward in later days, potentially indicating shorter or fewer charging sessions.
- **Leakage** and **RelayStatus** are flat throughout, confirming again that leakage was always zero and the relay remained active.
- **PartPF** follows a clear sawtooth pattern, suggesting a reset behavior or capped integration that recurs regularly.
- **Temperature** exhibits gradual warming over time, possibly due to seasonal changes or ambient heat accumulation in the robot's storage environment.
- **TotalPF** increases steadily, which may reflect changes in cumulative usage or improved efficiency over time.
- **Voltage** oscillates with a daily rhythm, likely tied to grid-level fluctuations or diurnal patterns in facility energy demand.

This temporal overview not only validates the presence of cyclical behavior (e.g., daily charge patterns) but also highlights the need for dynamic features or temporal aggregation when constructing models or performing anomaly detection.

4 Clustering Analysis with K-Means and PCA

In order to uncover distinct patterns in socket usage behavior, I applied K-means clustering to the cleaned time series feature data. Clustering can help identify underlying usage states or operational modes - such as idle, charging, and/or transitional states - based on multivariate sensor readings. Since the features span different scales and dimensions, clustering enables a data-driven categorization of socket activity over time.

Before clustering, I reduced the feature space using Principal Component Analysis (PCA). This dimensionality reduction technique transforms the data into a new coordinate system such that the first few principal components retain the majority of the variance. By projecting the data into 2D with PCA, I was able to visualize the clusters more intuitively and observe their spatial separation.

4.1 Principal Component Loadings and Interpretation

To better understand the meaning of each principal component, I examined the loadings matrix, which describes how much each original feature contributes to the principal components. Table 1 presents the loadings for PC1 and PC2.

Table 1: PCA loadings for the first two principal components.

Feature	**PC1**	**PC2**
Current	0.5101	0.0071
Power	0.5046	-0.0099
PowerFactor	0.5003	-0.0070
PhaseAngle	0.4824	0.0286
CSQ	0.0177	-0.2294
TotalPF	0.0085	-0.3852
Temperature	0.0044	-0.4884
RelayStatus	0.0000	0.0000
Leakage	0.0000	0.0000
LastHourPF	-0.0130	0.0486
Voltage	-0.0227	0.5975
PartPF	-0.0329	-0.4473

Interpretation of PC1 as Activity Level The first principal component (PC1) accounts for the largest portion of variance in the data and is strongly and positively associated with four key features: `Current`, `Power`, `PowerFactor`, and `PhaseAngle`. These variables are all directly linked to electrical activity and power consumption.

This concentration of high positive loadings suggests that PC1 can be interpreted as a proxy for the socket' s overall activity level:

- Higher PC1 values correspond to time points when the bot is actively charging or operating.

- Lower PC1 values reflect idle or standby periods with minimal current or power draw.

The low or near-zero loadings for unrelated features (e.g., `Leakage`, `RelayStatus`, `CSQ`) support this interpretation by showing that PC1 is not influenced by static or irrelevant sensor dimensions.

Thus, clustering based on PC1 primarily separates the dataset by operational intensity - an insight that aligns with the results of the K-means clustering and time-series inspection.

4.2 Silhouette Coefficients

To determine the optimal number of clusters, I calculated Silhouette Scores for a range of cluster counts from 2 to 10. The silhouette score measures how similar each point is to its assigned cluster compared to other clusters; higher values indicate better-defined and more cohesive clusters.

Figure 15 shows that the highest silhouette score occurs at $k = 2$, with diminishing returns for larger values of k. This suggests that a two-cluster solution provides the most natural partitioning of the data in terms of compactness and separation. The result aligns with the trend we discovered through distributional EDA conducted on the function types. While the drop in silhouette score between $k = 2$ and $k = 3$ is relatively minor - indicating that a three-cluster solution may still offer interpretive value - we ultimately decided to proceed with the two-cluster solution for its superior cohesion and clearer separation.

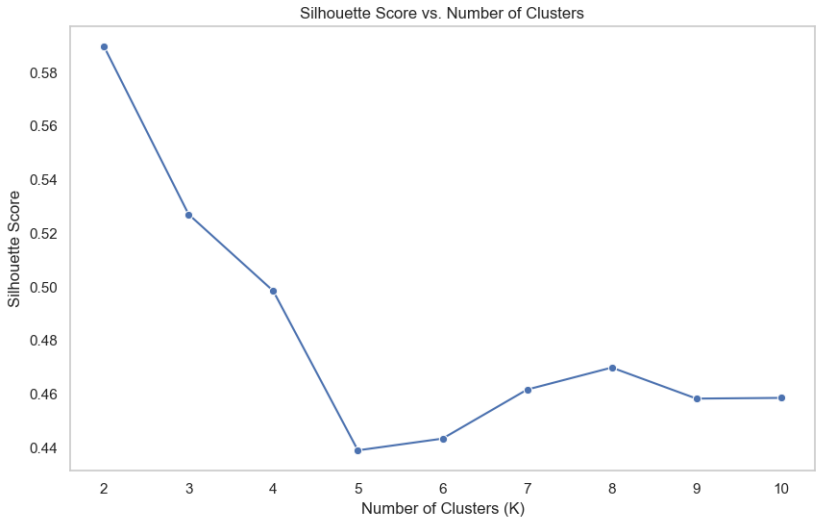

Figure 15: Silhouette Score as a function of the number of clusters k.

4.3 K-Means Clustering Results

Figure 16 shows the K-means clustering results for $k = 2$. With two clusters (left subplot), the data are clearly divided into two broad groups. This dichotomy may reflect an overarching split between "charging"and "non-charging" states, supported by patterns seen in Current, Power, and PowerFactor.

When increasing the number of clusters to three (right subplot), an intermediate region emerges, suggesting the presence of a transitional or mixed state - such as brief wake or idle periods between full charging and full inactivity. This added nuance could be beneficial for behavior classification or anomaly detection in the bot' s operation.

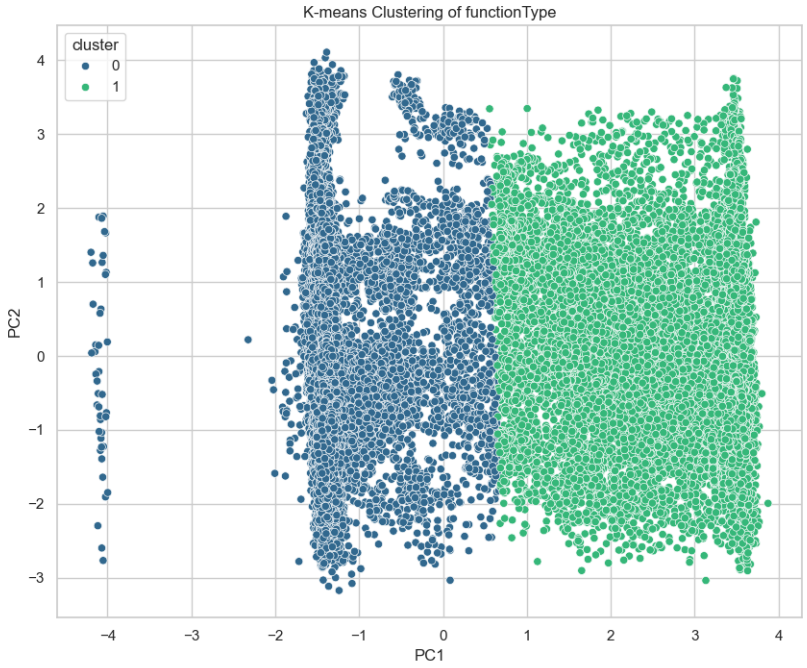

Figure 16: PCA projection of K-means clustering for $k = 2$.

4.4 State Distributions

After clustering, we assigned each timestamp to one of two operational states: charging or activity. These states reflect whether the smart socket was drawing significant current (charging) or not (activity), providing insight into the delivery bot' s behavior patterns.

Figure 17 shows that the socket spent approximately 66.2% of the recorded time in the charging state, and the remaining 33.8% in activity. This suggests the bot was stationary or docked for most of its lifecycle, consistent with its role in periodic delivery and return cycles.

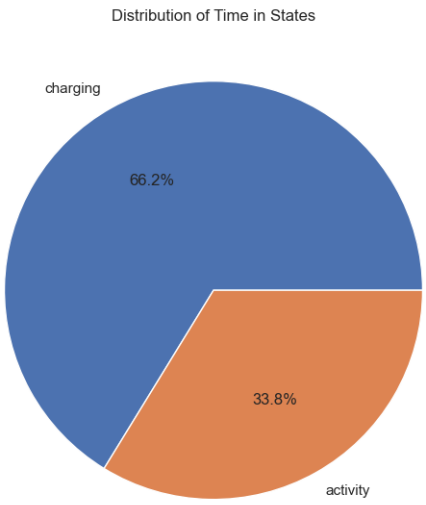

Figure 17: Distribution of time spent in each operational state.

We also examined how state usage varied by weekday (Figure 18). While the charging state consistently dominates across all days, there is a slight increase in activity on Sundays and Mondays, suggesting more frequent deployment at the start and end of the week. Tuesday through Friday show relatively stable distributions, possibly reflecting regular daily routines.

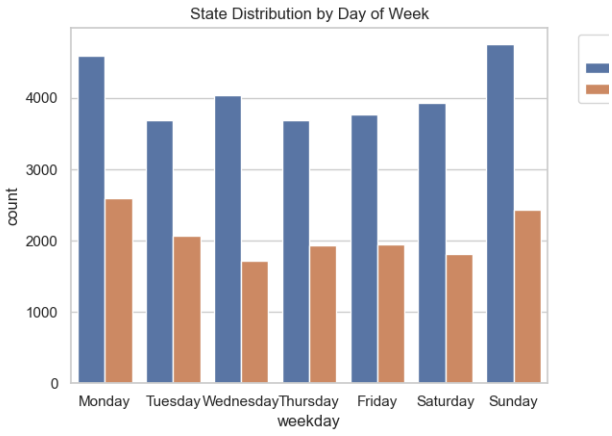

Figure 18: State distribution by day of the week.

Additional exploratory plots included:

- State distribution by hour of day
- Total time spent in each state per day
- Histogram of state durations
- State durations over time

However, these did not yield particularly meaningful insights due to noise and lack of clear patterns. Thus, we focus here on the more stable and interpretable weekday and overall time distributions.

4.5 State Transitions via Markov Chain

To better understand the dynamics between the two operational states - charging and activity - we modeled the system as a first-order Markov chain. This framework captures the transition probabilities between states over time, assuming that future states depend only on the current state.

Figure 19 displays the transition matrix derived from the observed sequence of states. The matrix reveals strong self-transition probabilities, especially for the charging state, suggesting that the system tends to remain in its current state for extended periods before switching.

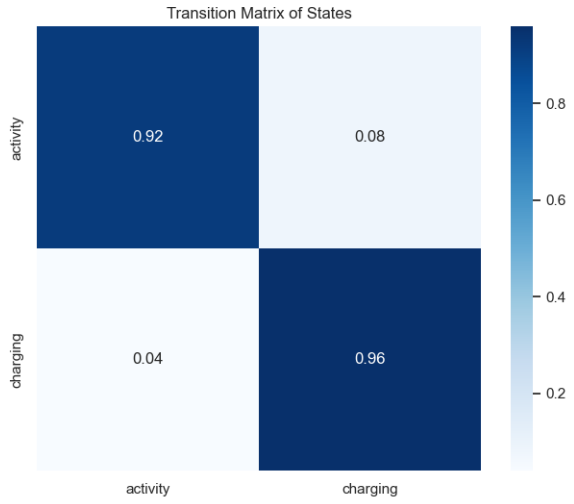

Figure 19: State transition matrix modeled as a Markov chain.

We also computed the steady-state distribution - the long-run proportion of time the system is expected to spend in each state. The results confirm that the system stabilizes with approximately 66.2% of time spent in the charging state and 33.8% in the activity state:

State	**Steady-State Probability**
Charging	0.662
Activity	0.338

These results are consistent with earlier visualizations (Figure 17), and reinforce the interpretation that the delivery bot primarily resides in a low-power or docked state, punctuated by shorter bursts of activity.

5 Temporal Usage Patterns

To better understand when the delivery bot is active versus charging, we analyzed state transitions across different time segments, correcting for timezone to reflect local behavior accurately. We segmented time both coarsely and finely to capture daily rhythms and operational windows.

5.1 Usage by Time of Day

Figure 20 shows the proportion of time the bot spends in *activity* and *charging* states across four broad time categories: Morning, Afternoon, Evening, and Night. The bot spends the majority of its time in the charging state across all periods, with the **Evening** showing the highest proportion of activity (~41%). This indicates that the bot is most engaged in tasks during evening hours, potentially aligning with delivery demand or scheduled operational shifts.

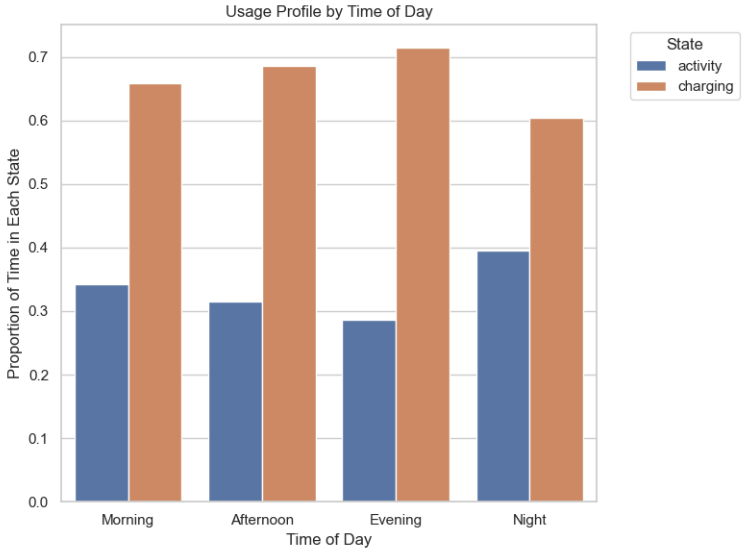

Figure 20: Proportion of time in each state across time-of-day categories.

5.2 Usage by 2-Hour Interval

To capture finer temporal dynamics, Figure 21 presents state distributions across 2-hour intervals. Bot activity fluctuates noticeably over the course of the day. **The lowest levels of activity** occur during midday (e.g., 14:00–16:00), while **evening hours (20:00–22:00)** show a clear spike in operational time. The early morning hours (02:00–06:00) also show moderate activity levels, which could indicate off-peak or preparatory functions.

In nearly all intervals, the charging state dominates, accounting for over 60% of the bot's time. This reinforces the energy-intensive nature of the system and highlights the importance of well-scheduled charging periods relative to

operational demand.

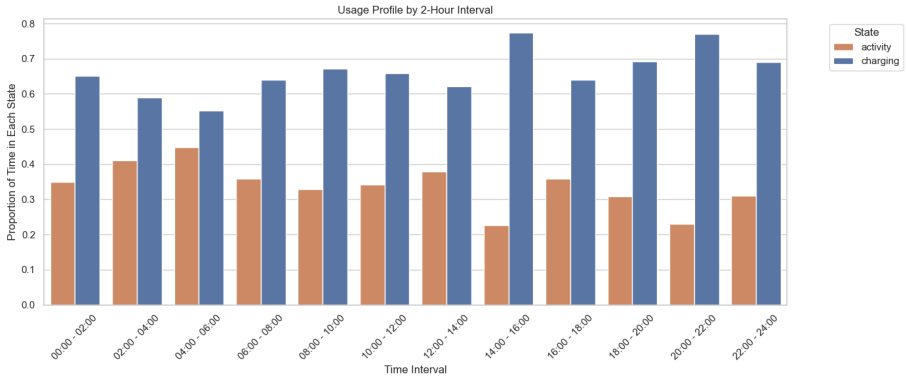

Figure 21: Usage profile by 2-hour time intervals.

Part III

Public PC Analysis

6 EDA

6.1 Correlational Analysis

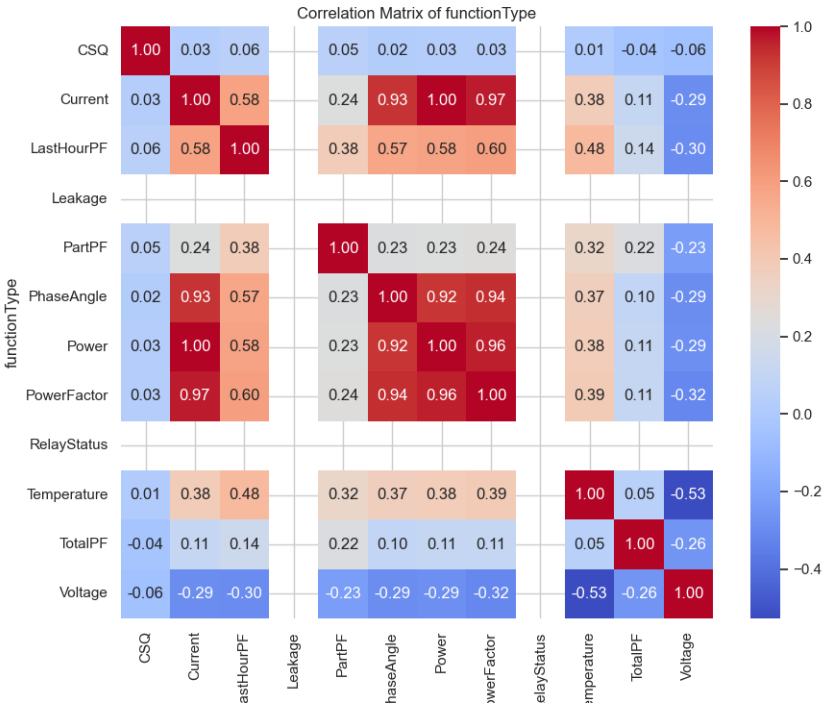

Figure 22: Correlation Matrix of `functionType`

Figure 22 presents the Pearson correlation coefficients among all `functionType` variables collected from the public PC. Several noteworthy patterns emerge:

- A strong positive correlation is observed between `Current`, `Power`, and `PowerFactor` (all above 0.9), suggesting these variables tend to co-vary during active device usage.

- `PhaseAngle` also exhibits high correlation with `Power` and `PowerFactor`, supporting the notion that phase shifts are tightly linked to power delivery characteristics.
- `LastHourPF` correlates moderately with both `Current` and `PowerFactor`, indicating some temporal consistency in power factor trends.
- `Temperature` is moderately correlated with `Power`, `Current`, and `PowerFactor`, which is consistent with expected thermal buildup from increased electrical activity.
- `Voltage` is weakly or negatively correlated with most other features, possibly reflecting stability in supply voltage regardless of usage intensity.
- `CSQ`, `Leakage`, and `RelayStatus` show minimal correlation with the rest of the variables, likely due to their static or near-constant behavior.

Overall, the matrix reveals distinct clusters of co-varying signals tied to energy consumption and operational states, reinforcing the structural relationships among current, power, and thermal outputs in the system.

6.2 Distribution of the Functions

6.2.1 Signal Strength (CSQ)

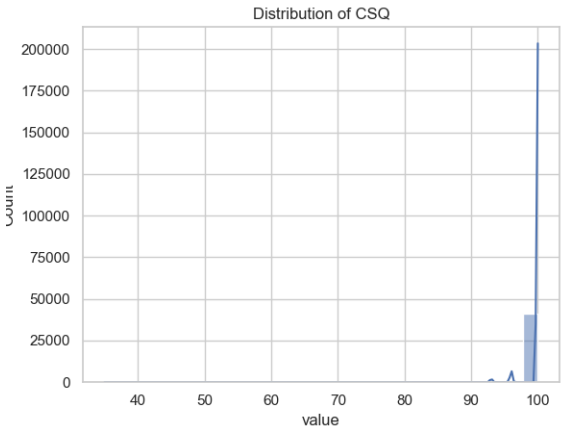

Figure 23: Distribution of `CSQ` Values

Figure 23 displays the distribution of `CSQ` (signal quality) values recorded from the public PC. The histogram reveals a sharply peaked distribution near the maximum value of 100, with negligible variation across the range. This suggests that the signal quality is consistently excellent throughout the observation period, with very few instances of degradation. The lack of variance implies `CSQ` is not a useful feature for distinguishing between different operational states or behaviors in this dataset.

6.2.2 Current

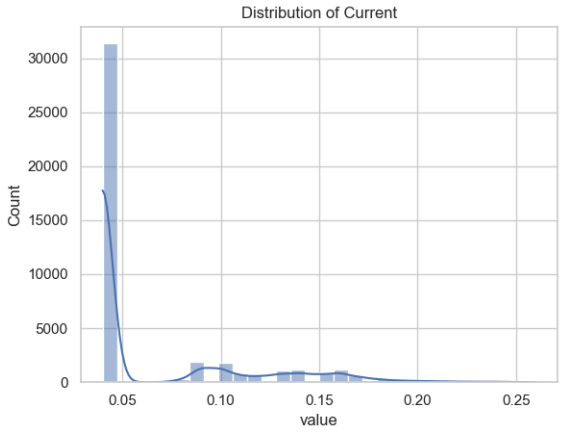

Figure 24: Distribution of `Current` Values

Figure 24 shows the distribution of `Current` values for the public PC. The distribution is heavily skewed towards a baseline value around 0.05, with smaller peaks occurring at higher intervals such as 0.1, 0.15, and 0.2. This multi-modal pattern suggests the presence of distinct operational levels or load states, possibly corresponding to different usage patterns or task intensities. The long tail also indicates occasional higher current draw, but these instances are relatively rare.

6.2.3 Last Hour Power Factor

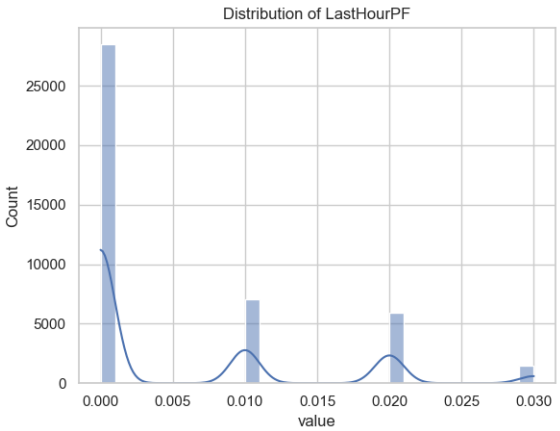

Figure 25: Distribution of `LastHourPF` Values

Figure 25 presents the distribution of `LastHourPF` values for the public PC. The values are tightly clustered near zero, with distinct spikes at regular intervals (approximately 0.01, 0.02, etc.). This suggests a discretized or quantized behavior in the power factor readings, likely corresponding to predefined operational states or rounding in measurement. The strong peak at zero may indicate long durations of inactivity or minimal power factor due to idling or light usage.

6.2.4 Leakage

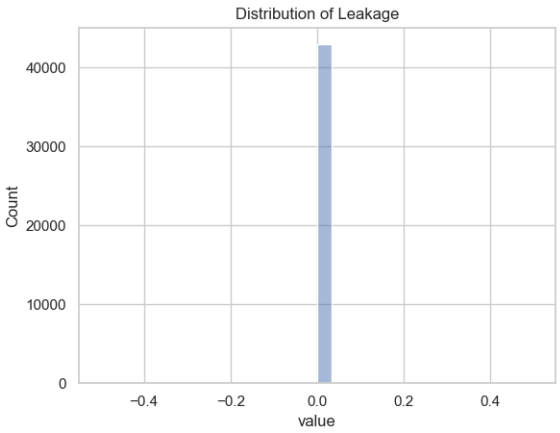

Figure 26: Distribution of `Leakage` Values

Figure 26 shows the distribution of `Leakage` values for the public PC. The data exhibits a single spike at zero, indicating that virtually all recorded values are zero. This suggests that leakage was either non-existent, below sensor resolution, or not actively monitored during the observed period.

6.2.5 Partial Power Factor

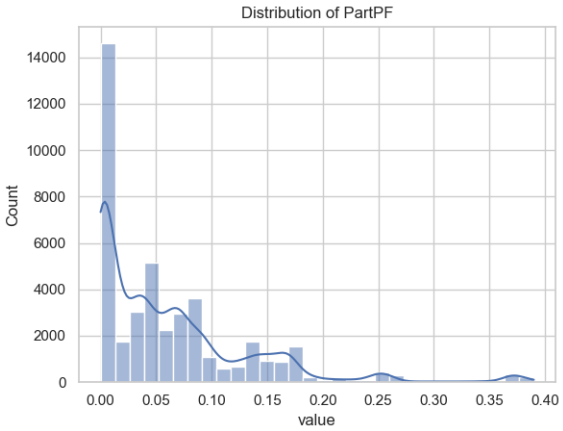

Figure 27: Distribution of `PartPF` Values

Figure 27 displays the distribution of `PartPF` values. The majority of the data is concentrated below 0.1, with a sharp peak at zero and a long tail extending toward 0.4. This right-skewed distribution suggests frequent low power factor conditions, possibly indicating idle or lightly loaded electrical devices.

6.2.6 Phase Angle

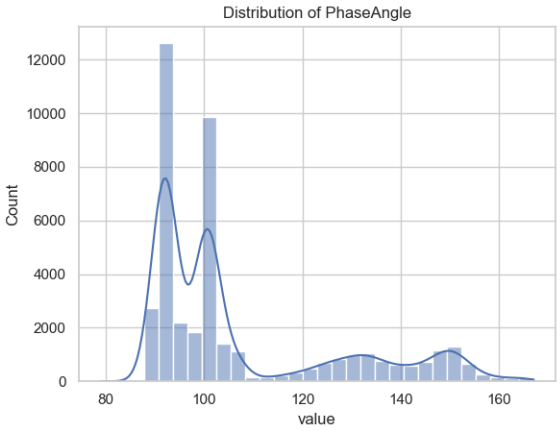

Figure 28: Distribution of `PhaseAngle` Values

Figure 28 shows the distribution of `PhaseAngle` values. The plot reveals multiple distinct peaks, with two dominant modes around 90 and 100 degrees, and smaller humps near 140 and 150 degrees. This multimodal structure indicates the presence of different operational states or device types with distinct phase behaviors.

6.2.7 Power

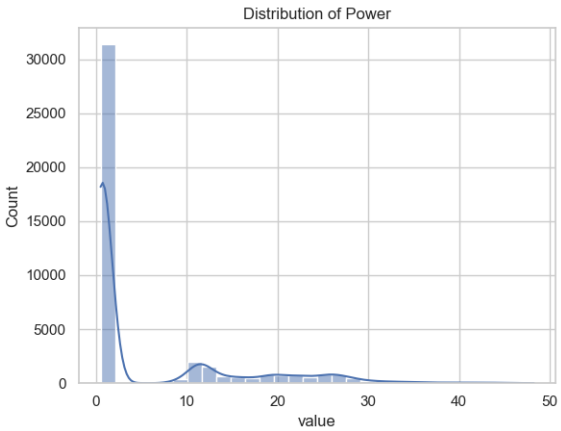

Figure 29: Distribution of `Power` Values

Figure 29 presents the distribution of `Power` values. The data exhibits a sharp peak near zero, followed by a long right tail that includes several smaller modes around 10, 20, and 30. This right-skewed distribution suggests that low power usage is most common, but higher usage levels occur in distinct clusters, possibly reflecting different device types or operational states.

6.2.8 Power Factor

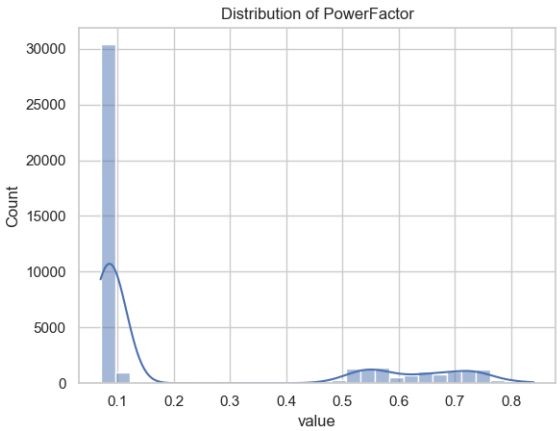

Figure 30: Distribution of `PowerFactor` Values

Figure 30 displays the distribution of `PowerFactor` values. The distribution is strongly right-skewed, with a pronounced peak near 0.1 and additional modes around 0.6 and 0.7. This suggests that the power factor is often low, possibly due to reactive or inefficient power use, but occasionally improves, perhaps when specific equipment is active.

6.2.9 Relay Status

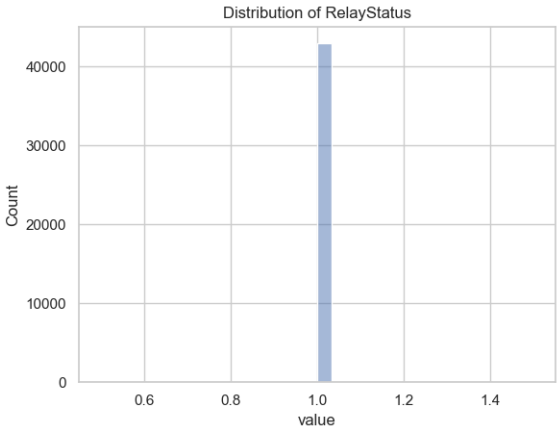

Figure 31: Distribution of `RelayStatus` Values

Figure 31 shows the distribution of `RelayStatus` values. The data is concentrated at 1.0, indicating that the relay is almost always in the "on" or active state. This could suggest a continuously powered or rarely toggled relay device.

6.2.10 Temperature

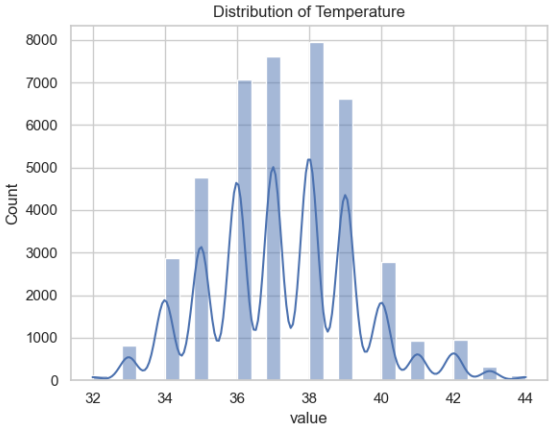

Figure 32: Distribution of `Temperature` Values

Figure 32 illustrates the distribution of `Temperature` values. The data follows a roughly normal distribution centered around 37–38°C, with regular oscillations in the density. The fluctuations likely reflect a periodic heating-cooling cycle or thermostat-controlled regulation.

6.2.11 Total Power Factor

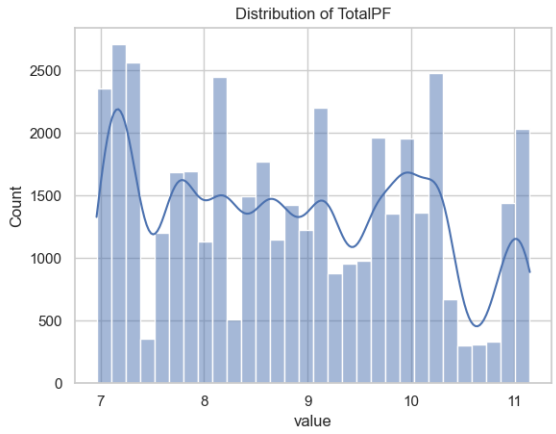

Figure 33: Distribution of `TotalPF` Values

Figure 33 shows the distribution of `TotalPF` values. The values range from 7 to 11, with a roughly uniform or slightly multimodal spread. This spread suggests variability in the total power factor, possibly influenced by load diversity or time-varying usage patterns.

6.2.12 Voltage

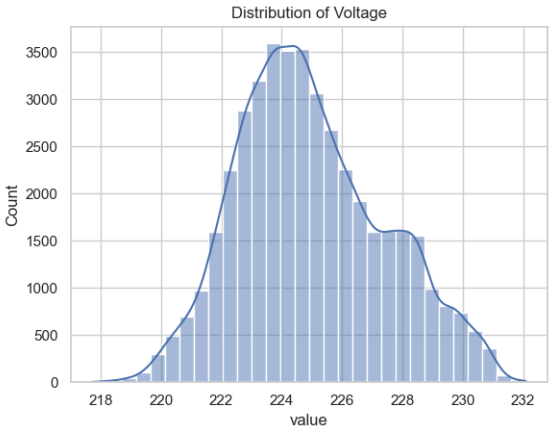

Figure 34: Distribution of `Voltage` Values

Figure 34 depicts the distribution of `Voltage` values. The distribution is approximately normal, centered around 224 V, with a standard deviation of a few volts. This pattern suggests stable voltage supply with minor fluctuations, typical of well-regulated electrical systems.

6.3 Time Series Analysis

Figure 35 presents the time series of various `functionType` variables, plotted as individual subplots for visual clarity. Each subplot tracks the behavior of a specific variable over time:

- **CSQ** remains mostly stable at high values, with occasional drops indicating signal quality interruptions.
- **Current**, **Power**, and **PowerFactor** all exhibit highly intermittent patterns, suggesting bursty or cyclical device usage.
- **Leakage** and **RelayStatus** are essentially constant, implying either static configurations or binary states that rarely change.
- **Temperature** shows gradual fluctuations, likely tied to environmental or internal heat dynamics.

- **TotalPF** increases steadily over time, possibly indicating accumulated power factor metrics.
- **Voltage** shows a strong periodic oscillation, typical of AC power supply behavior with possible diurnal influence.

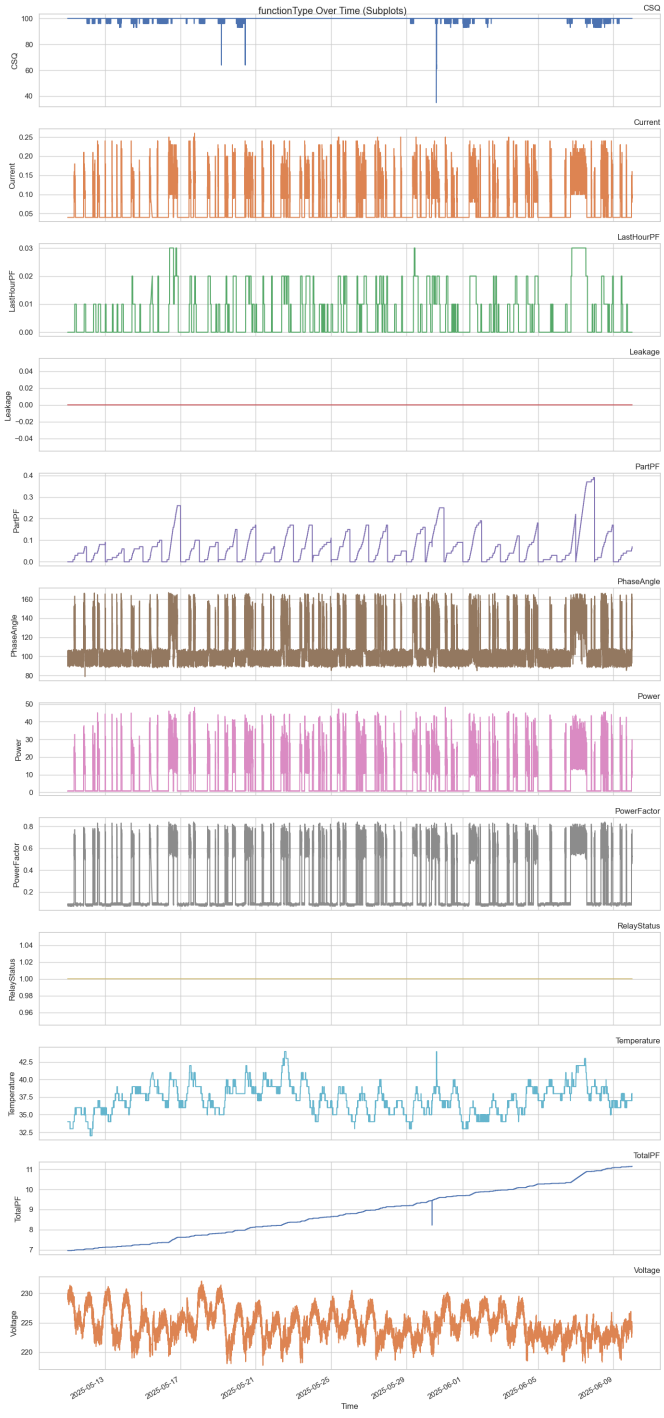

Figure 35: Time Series of `functionType` Variables Displayed as Subplots

7 Clustering Analysis

To explore natural groupings in the functional behavior of devices, we performed a clustering analysis on the preprocessed dataset.

7.1 Dimensionality Reduction via PCA

We began with Principal Component Analysis (PCA) to reduce the high-dimensional feature space into two principal components, enabling visual interpretation. The loading matrix revealed that `PC1` is primarily associated with active power-related variables, such as `Current`, `Power`, and `PowerFactor`, whereas `PC2` is influenced more by `Voltage`, `Temperature`, and `PartPF`.

7.2 Determining Optimal Clusters with Silhouette Analysis

To select the optimal number of clusters, we computed the silhouette score for K ranging from 2 to 10 (Figure 36). The highest silhouette score was observed at $k = 2$, indicating the most well-separated clusters under this setting.

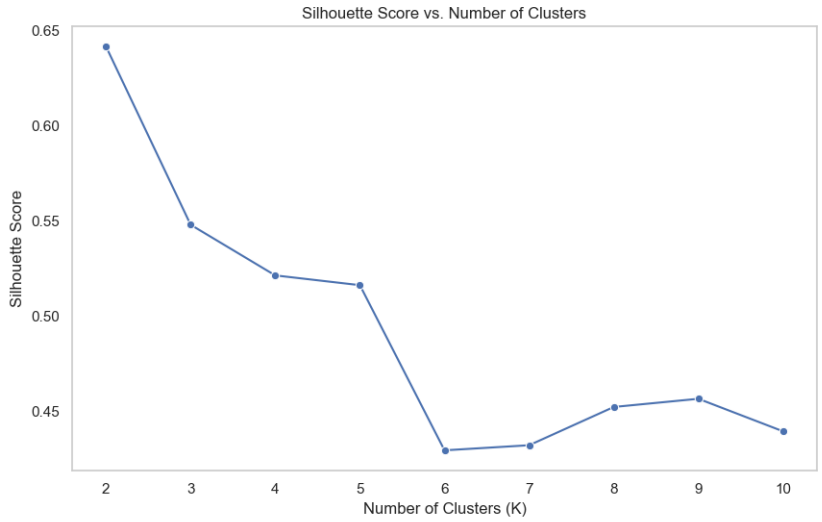

Figure 36: Silhouette Score vs. Number of Clusters

7.3 K-Means Clustering

Using $k = 2$, we applied the K-means algorithm and visualized the resulting clusters in PCA space (Figure 37). However, visual inspection revealed that the generated clusters do not align with clear or intuitive separations in the data. The decision boundary appears arbitrary, and K-means fails to capture the visible structure suggested by the PCA projection.

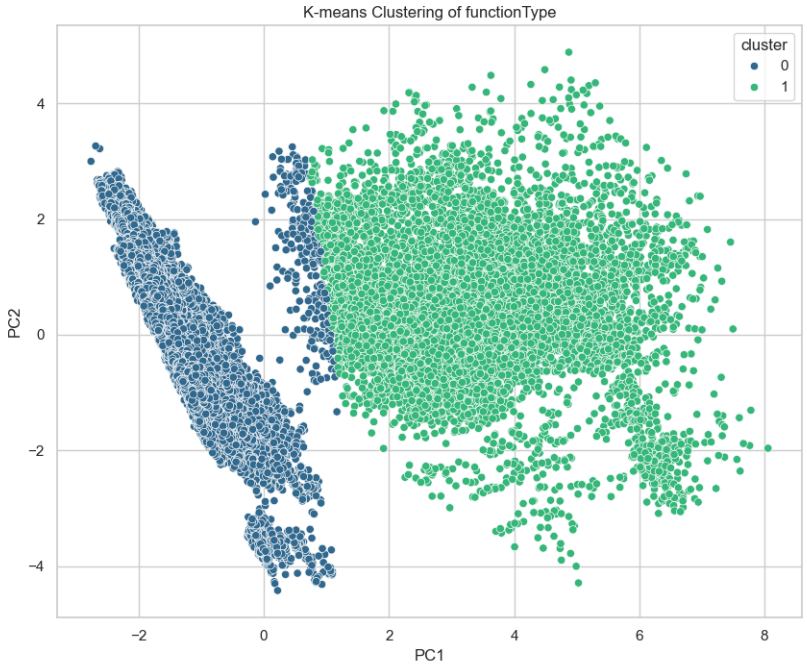

Figure 37: K-means Clustering in PCA space ($k = 2$)

7.4 DBSCAN and Density-Based Clustering

After K-means failed to produce intuitively meaningful clusters - despite a high silhouette score - we turned to DBSCAN, a density-based clustering algorithm. Unlike K-means, which assumes spherical clusters and partitions the space into equal-sized regions, DBSCAN can identify arbitrarily shaped clusters and is more robust to noise. This made it a promising alternative for

discovering natural groupings in PCA space that might not conform to K-means' assumptions.

To determine a suitable value for DBSCAN' s key parameter, ϵ (the neighborhood radius), we used the k-nearest neighbor (k-NN) distance method. By plotting the sorted distances to each point' s k^{th} nearest neighbor (typically with $k = \texttt{min\_samples}$), we identified an "elbow" point in the curve where distances sharply increase. This elbow suggests a threshold that separates dense regions (potential clusters) from sparse regions (noise), and is commonly used as a heuristic to choose ϵ. However, even after tuning ϵ based on this method, DBSCAN consistently identified only a single cluster (Figure 39), suggesting that the density variation in our dataset was insufficient for the algorithm to separate meaningful subgroups.

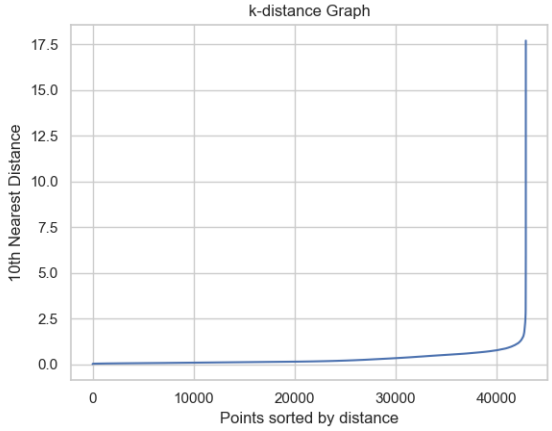

Figure 38: k-distance Graph for DBSCAN parameter tuning.

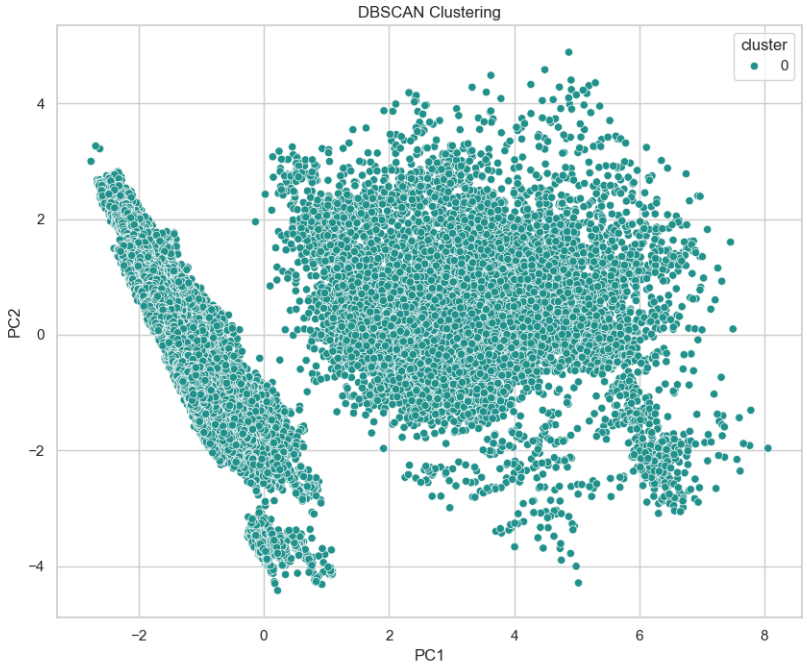

Figure 39: DBSCAN Clustering Result

7.5 Manual Clustering

Given the inadequacy of both K-Means and DBSCAN to produce intuitive clusters, we turned to manual clustering by defining a linear boundary in the PCA space. The decision boundary was chosen as a line with slope $m = -1.5$ and an adjustable intercept b:

$$y = -1.5x + b \tag{1}$$

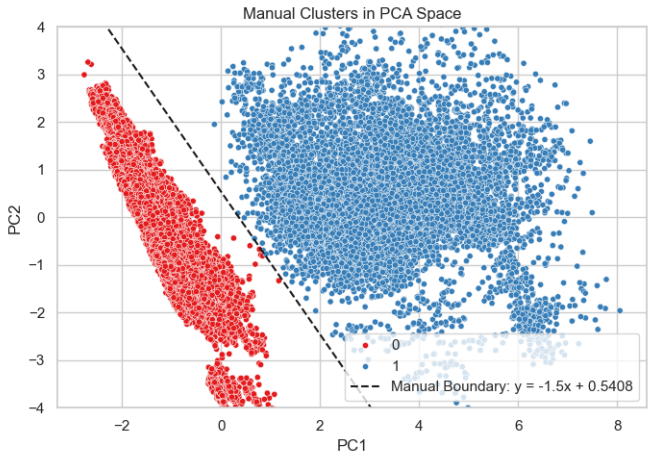

Figure 40: Manual clustering in PCA space using a custom linear decision boundary defined by $y = -1.5x + 0.5408$. Data points are split into two groups based on whether they fall above or below this boundary. This approach enables interpretable separation of operational states based on their projection along the boundary-normal vector.

Initially, the boundary was placed with $b = 0$ (Figure 40), then later adjusted to $b = 0.3$ based on visual inspection of the projection of points onto the normal vector of the boundary. This shifting allowed a cleaner separation between two visually distinct clouds in PCA space.

To determine this boundary more precisely, we projected each data point onto the normalized decision vector $\vec{v} = [1, -1.5]^\top / |[1, -1.5]^\top|$ and plotted a histogram of the projection values (Figure 41). Based on the histogram's valley between two peaks, we selected a threshold value of -0.45 to define the boundary:

$$y = -1.5x + 0.3 \tag{2}$$

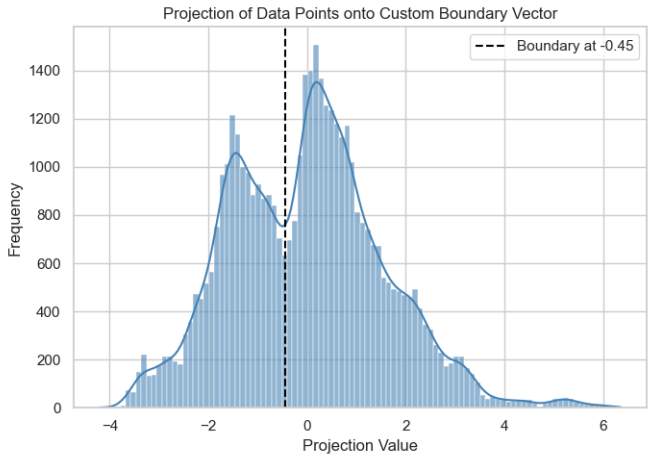

Figure 41: Projection of data points onto a custom boundary vector, defined along the direction orthogonal to the manual decision boundary. The dashed vertical line at -0.45 represents the adjusted threshold used to separate the two clusters. The clear bimodal structure supports the choice of manual clustering, as the projection reveals two well-separated groups.

Points below the boundary were labeled as Cluster 0, and those above as Cluster 1.

7.6 Cluster Interpretation

The interpretation of the manually defined clusters is supported by the PCA loadings matrix:

- **PC1** is strongly influenced by power-related variables: `Current`, `Power`, `PowerFactor`, and `PhaseAngle`. This suggests PC1 reflects overall energy demand and device operation intensity.
- **PC2** is driven more by contextual or environmental factors: `Voltage`, `Temperature`, and `PartPF`. This component likely captures external conditions or ambient influences.

functionType	**PC1**	**PC2**
Current	0.510	0.007
Power	0.505	-0.010
PowerFactor	0.500	-0.007
PhaseAngle	0.482	0.029
CSQ	0.018	-0.229
TotalPF	0.008	-0.385
Temperature	0.004	-0.488
RelayStatus	0.000	≈0.0
Leakage	≈0.0	≈0.0
LastHourPF	-0.013	0.049
Voltage	-0.023	0.598
PartPF	-0.033	-0.447

Table 2: PCA Loading Matrix

Because the decision boundary is not orthogonal to either PC1 or PC2, the clusters separate data points based on a combination of energy consumption and contextual variation. Specifically:

- **Cluster 0**: Contains points with lower PC1 and PC2 scores, likely representing low-power or standby device states.
- **Cluster 1**: Contains higher PC1 scores and mixed PC2 values, corresponding to active operation or higher-load periods.

As a result, I decided to label cluster 0 as idle, and cluster 1 as operating. In other words, cluster 0 indicates times where the public PC was on standby, while cluster 1 indicates time where the PC was being used actively.

7.7 Temporal Usage Pattern

Following the manual clustering analysis, which revealed two distinct operational states (interpreted as *idle* and *operating*), we investigated the temporal distribution of these states to better understand device behavior across different time scales. This section explores how the device transitions between idle and operating modes depending on the time of day and the day of the week.

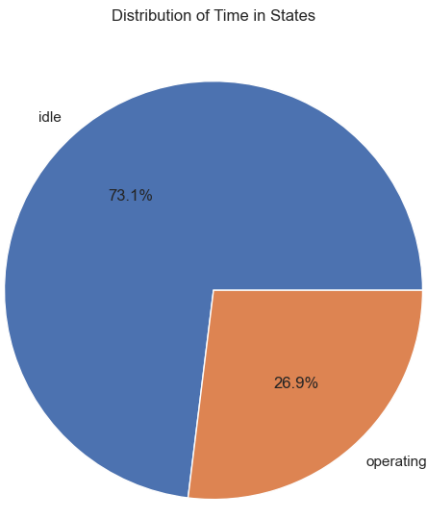

Figure 42: Overall distribution of time spent in each state. The device was idle for the majority of the time (73.1%) and operating for 26.9% of the time.

Overall Usage Proportion. As shown in Figure 42, the device remains idle for most of the time. This suggests that periods of active usage are relatively infrequent, possibly corresponding to scheduled or externally triggered operations.

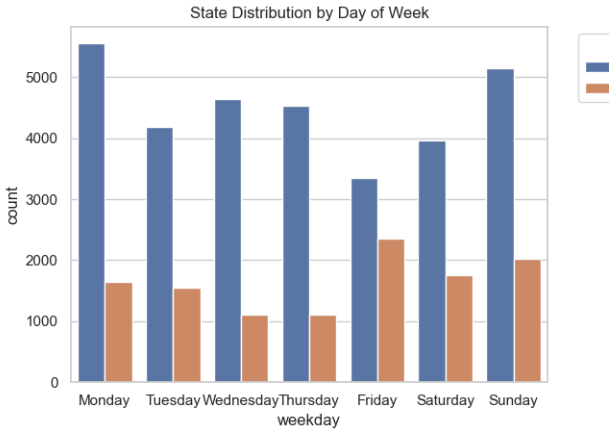

Figure 43: State distribution by day of week. Device usage is reduced during midweek and increases again toward the weekend.

Weekday Trends. In Figure 43, we observe a noticeable dip in activity on Tuesday through Thursday, with slightly elevated activity levels on Friday through Sunday. The peak on Monday may indicate the device is used more heavily at the start of the work week, possibly for diagnostic or maintenance purposes.

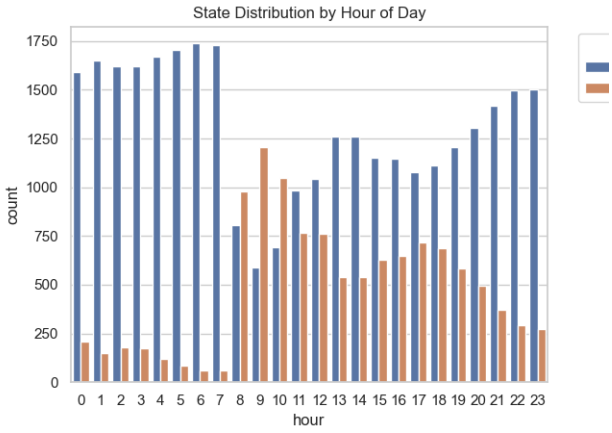

Figure 44: Hourly state distribution. The device is largely idle during early morning and night hours, with higher activity between 08:00 and 18:00.

Hourly Behavior. Hourly breakdown in Figure 44 reveals a strong diurnal pattern: activity rises sharply starting at 08:00, peaks around late morning to early afternoon, and declines steadily after 18:00. This usage pattern aligns with typical business or operational hours.

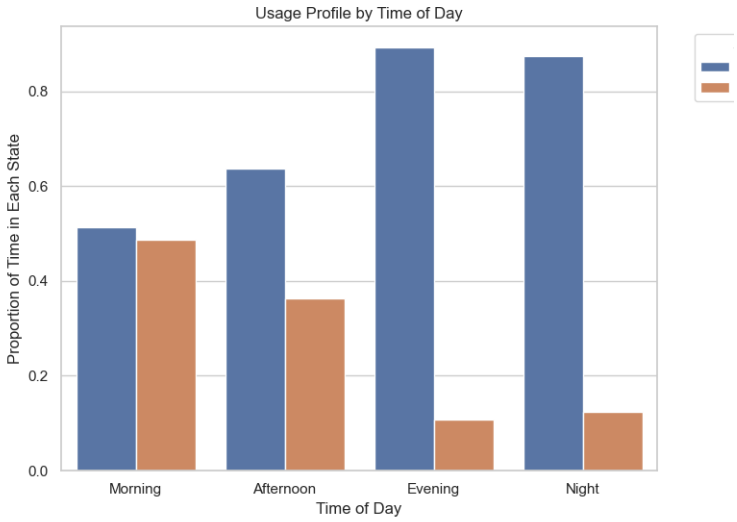

Figure 45: Proportion of time in each state across four major time blocks. Device activity is highest in the afternoon and lowest during the evening and night.

Time of Day Profile. Figure 45 simplifies the hourly trends into four main blocks: morning, afternoon, evening, and night. Activity is highest in the afternoon, confirming that the device operates predominantly during standard working hours.

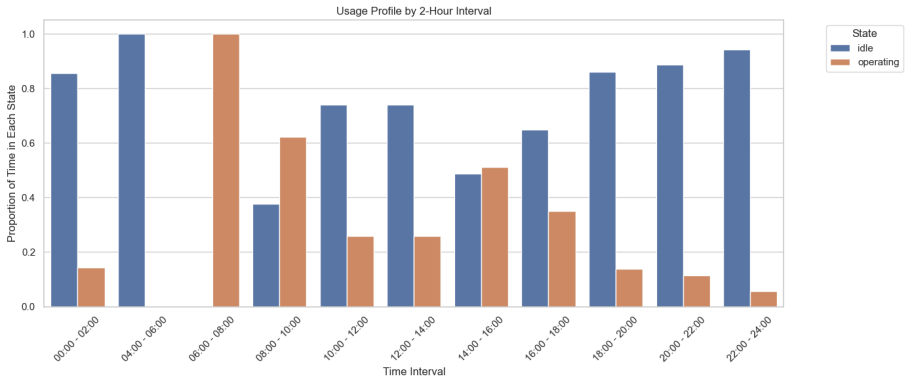

Figure 46: Usage profile broken into 2-hour intervals. Peak activity occurs between 06:00 and 10:00, while nighttime hours show near-zero operation.

Finer-Grained Usage Profile. To capture more nuanced patterns, Figure 46 presents usage distributions over 2-hour blocks. The sharp increase in operational activity between 06:00 and 10:00 reflects a consistent morning activation period. This is followed by a drop during midday and a rise again around 16:00–18:00, possibly reflecting split operational windows.

Summary. Overall, these temporal usage patterns suggest the device has a predictable daily rhythm with short operational windows mostly concentrated in the daytime. Idle time dominates overall, making it critical to identify the triggers or events associated with the transition into active use. These insights may inform optimization of scheduling, maintenance, or energy efficiency strategies.

Part IV

Comparative Analysis: Delivery Bot vs. Public PC

To synthesize insights across different smart socket deployments, we compared the usage patterns and clustering behaviors observed in the delivery bot versus the public PC.

8 State Clustering Comparison

Both devices underwent unsupervised clustering based on function-type data (e.g., power, voltage, current). For the delivery bot, the clusters derived from K-Means and PCA projections aligned well with expected operational modes (i.e., *activity* vs. *charging*), reflecting well-defined and cyclical usage. In contrast, the PC's usage pattern was less clearly separable using standard clustering methods. K-Means failed to capture intuitive groupings, and DBSCAN returned a single dense cluster regardless of parameter tuning. This ambiguity likely stems from the PC' s variable load profile and more continuous usage.

To address this, a manual boundary was applied in PCA space, yielding two interpretable clusters. Analysis of loadings showed that PC1 captured general power usage and device operation, while PC2 captured environmental or system state variations (e.g., voltage and temperature). The manually defined boundary (slope = -1.5) created a clear operational divide, with one cluster representing low-power or idle states, and the other reflecting high-load activity. This manual intervention proved necessary to account for the nuanced behavior of the PC.

9 Temporal Patterns Comparison

Temporal usage further highlighted the behavioral divergence between devices. The delivery bot exhibited strong periodicity, with clear distinctions in state distribution across times of day and weekdays. Activity levels spiked in the evenings, with low usage during late-night and early-morning periods, consistent with task-based or energy-availability constraints.

The PC, by contrast, demonstrated a more continuous but still patterned profile. While low-usage intervals were still visible (e.g., early morning), the transitions between states were more frequent and evenly distributed. Day-of-week analysis showed the PC to be in use across all days, with less pronounced dips on weekends, suggesting a mix of automated or unattended tasks in addition to human-driven usage.

10 Interpretability and Intervention

Overall, the delivery bot's usage could be inferred with minimal manual intervention, while the PC required a combination of dimensionality reduction, projection analysis, and domain-informed threshold tuning to extract meaningful operational states. This contrast underscores the value of flexible analytic approaches when dealing with heterogeneous devices in smart infrastructure systems.

Key Takeaways:

- The delivery bot displays high regularity and clustering clarity due to predictable routines.
- The PC exhibits overlapping behavioral states, necessitating more interpretive and manual analysis.
- Temporal profiles reflect use-case differences: time-bound delivery vs. persistent public access.

Part V

Conclusion

This analysis has uncovered distinct operational patterns in both the delivery robot and public PC systems, offering actionable insights for facility management. The temporal routines observed - such as the robot's concentrated evening activity and the PC's multi-state power usage - highlight opportunities for intelligent scheduling and targeted energy conservation. Understanding when and how these devices are used allows for load balancing, peak shaving, and potential automation of shutdown or charging protocols during off-peak hours.

Moreover, the structured data pipeline established through this analysis provides a scalable foundation for future development. By modeling transitions and identifying device states with high fidelity, we can design real-time monitoring dashboards, trigger-based alert systems, and predictive models for anomaly detection or preventative maintenance. These tools can be generalized across similar IoT deployments, enabling smarter, data-informed facility operations that align with sustainability and efficiency goals.