# Problem Statement

**Goal: To extract meaning insights and patterns from the file 2019Floor2**

## CRISP-DM

**CRISP-DM (Cross Industry Standard Process for Data Mining)** is a widely used methodology for structuring data science and machine learning projects. It provides a systematic approach to planning and executing data mining (or data analytics) projects, ensuring that the process is repeatable and understandable.

CRISP-DM consists of **6 phases**:

1. **Business Understanding**
2. **Data Understanding**
3. **Data Preparation**
4. **Modeling**
5. **Evaluation**
6. **Deployment**

## **Phase 1:** **Business Understanding**

### Step 1: Identify the Business Objective

***Business Objective: Predict and optimize electricity consumption (especially AC units) in an office building to reduce energy costs and improve sustainability.***

### Step 2 :Determine Project Goals

| **Business Goal** | **Data Science Goal** |
| --- | --- |
| Reduce AC usage cost | Predict AC consumption in next hour/day |
| Spot unexpected high usage | Detect anomalies in electricity usage |
| Optimize comfort and energy | Build relationship between temperature and AC usage |

➡️ Let's choose a data science goal for now:

❗**Data Science Goal:** Forecast AC unit power consumption in each zone for the next hour using historical sensor and power data.

## **Phase 2: Data Understanding**

The csv file 2019Floor2 has 4 zones

The dataset includes:

* Power consumption of AC, lighting, plug loads.
* Temperature, humidity, light (lux).
* Data collected every minute for 12months
* Some zones don't have all types of data

**Attribute representation**

* **z1\_Light(kW)**: Power consumption of lighting in Zone 1, measured in kilowatts.
* **z1\_AC1(kw) :** Represents the power consumption of zone\_1 AC\_1 in kilowatts
* **z1\_Plug(kW)**: Power consumption from plug-in devices in Zone 1, in kilowatts.
* **z1\_S1(degC)**: Temperature in Zone 1 recorded by sensor S1, in degrees Celsius.
* **z1\_S1(RH%)**: Relative humidity in Zone 1 recorded by sensor S1, in percentage.
* **z1\_S1(lux)**: Illuminance (light level) in Zone 1 recorded by sensor S1, in lux.

### File – 2019Floor\_2

* **Zone 1** :

Ac1 consumes an avg of 40 kw per min (An Industrial AC)  
Light consumes an avg of 8 - 9.8 kw per min

Plug consumes an avg of 4 kw per min

* **Zone 2 :**

Ac1 consumes an avg of 28 kw per min

Ac2 to Ac10 consumes an avg of 0.5 – 1.5 kw per min  
Ac11 and Ac14 is completely not used   
Ac12 and Ac13 is rarely used on an avg power consumption of 1.5 kw per min   
light consumes an avg of 2.8 kw per min

Plug consumes an avg of 1 – 1.5 kw per min

* **Zone 3:**

No AC in zone 3

Light consumes light consumes an avg of 1.25 to 1.5 kw per min

Plug consumes less than 1.8 kw per min

* **Zone 4 :**

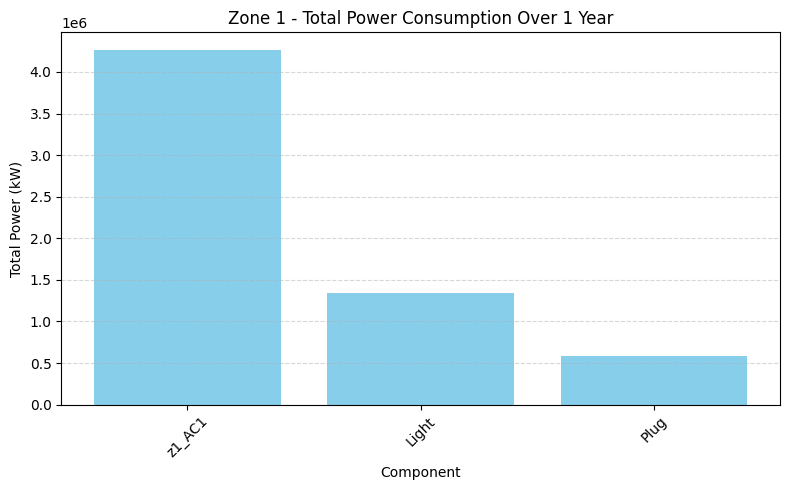
Ac1 consume an avg of 10kw per min

Light consumes an avg of 2kw per min

Plug consumes a very less amount of power which is 0.01 to 0.03 kw per min

* In all the 4 zones the temperature defers only in a small amount like 0.5 or 1 C at same time   
  All the sensor data for Temperature, Humidity ,Illuminance has missing values from January to March 1st week
* Mostly in the nighttime the power consumption is zero for all the systems

A graph of a graph with blue squares

AI-generated content may be incorrect.

A graph of a diagram

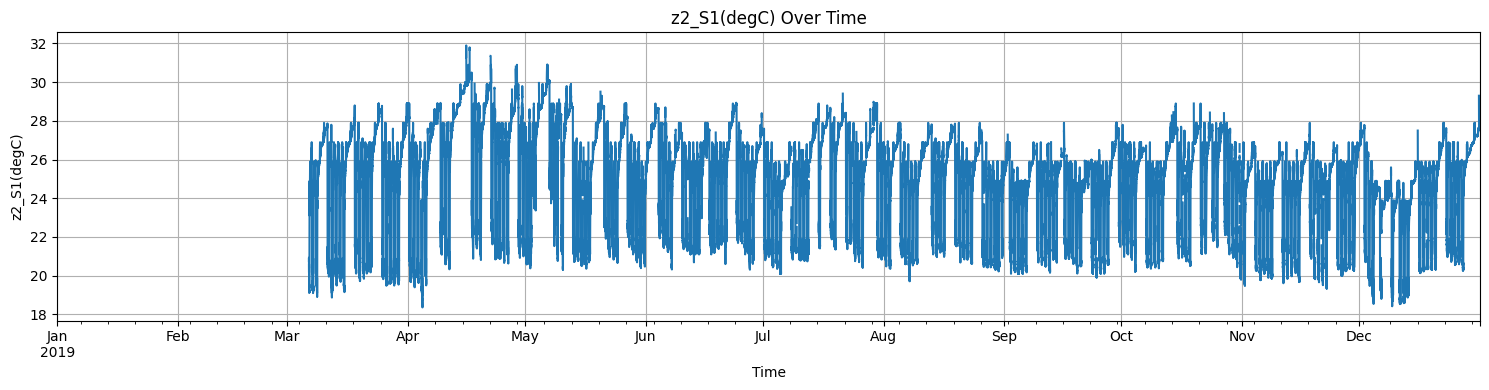
AI-generated content may be incorrect.A graph of a diagram

AI-generated content may be incorrect. **1e6 = 1 x 10^6**

### Phase 3 : Data preparation for 2019Floor2.csv

1. Since in all the 4 zones the sensor data for Temperature, Humidity ,Illuminance has missing values from January to March 1st week which is a big series of data cannot be filled . I **decides to remove all the entities that are missing**

A graph showing a number of columns

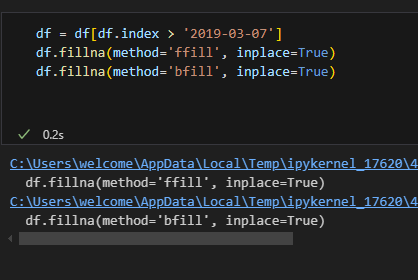
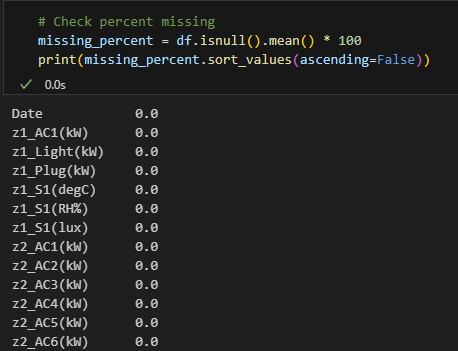
AI-generated content may be incorrect.Illuminance

A graph showing a blue line

AI-generated content may be incorrect.Temperature

Humidity

1. And use forward or backward fill to data that are missing in between entities



Hence there is no missing values here so we can start build our model

we can build models based on our desired output   
**1. Prediction (Regression)**

You estimate **continuous values**, e.g.:

* **Predict indoor temperature** (z1\_S1(degC), etc.) based on energy usage and time of day.
* **Forecast future energy consumption** (AC, lights, plug) using historical usage and sensor data.
* **Predict illuminance** using time, light usage, and weather (if external data is available).

**Common models**: Linear Regression, Random Forest Regressor, XGBoost, LSTM (for time series)

**2. Classification**

You assign labels to data. Examples:

* **Classify occupancy status** (occupied vs. unoccupied) based on plug, light, and temperature usage.
* **Detect overuse** of energy by classifying periods as normal vs. excessive usage.

You'd need to **create labels**, e.g.,:

* Add a binary "occupied" label based on rules like light > 0.5 kW, plug > 0.3 kW
* Classify "high usage" if total power > threshold

**Common models**: Logistic Regression, Decision Trees, Random Forests, SVM

**3. Clustering (Unsupervised Learning)**

You find **patterns or groupings** in the data without labels:

* **Cluster zones or time periods** based on similar energy consumption behavior.
* **Group daily usage profiles** into clusters like "weekday office hours", "weekend idle", etc.

**Common models**: K-Means, DBSCAN, Hierarchical Clustering

**4. Anomaly Detection**

You find **unusual or abnormal patterns** in the data:

* **Detect faulty sensors** or missing data that wasn’t caught in cleaning.
* **Identify unexpected energy usage** patterns (e.g., AC on at night).

**Common methods**: Isolation Forest, One-Class SVM, Autoencoders

### Phase 4 : Model (2019Floor2.csv)

**Weekday and Weekend Clustering Analysis**

**Objective:**

To understand patterns in energy usage (AC, lights, plugs) separately for weekdays and weekends, by grouping similar daily profiles using unsupervised learning (clustering). This helps in:

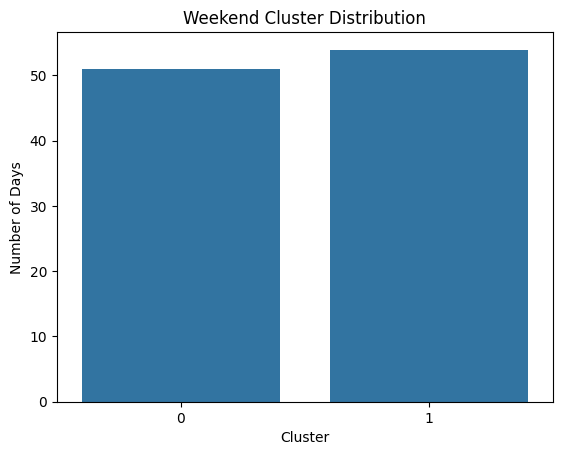
* Identifying typical operational patterns on working days vs non-working days.
* Detecting inefficient or unexpected behavior (e.g., high weekend usage).
* Supporting better energy scheduling, maintenance planning, or anomaly detection.

**Data Separation**  
The dataset was split into:

* weekday\_data: All Monday–Friday days.
* weekend\_data: All Saturday–Sunday days.

**Clustering Methodology**

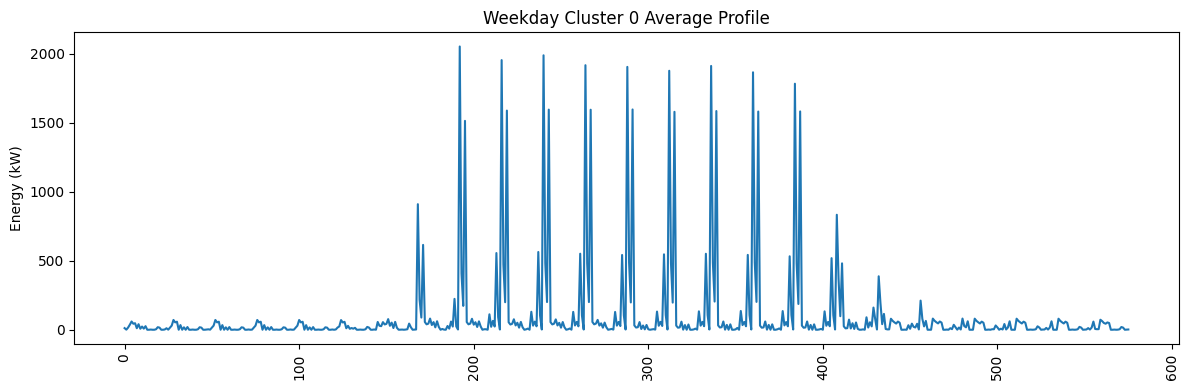
**K-Means Clustering** was applied **separately** to weekdays and weekends:

1. **Standardization**:  
   Each feature was scaled using StandardScaler() to remove bias from high-usage sources.
2. **KMeans Algorithm**: The number of clusters was chosen based on experimentation
3. **Cluster Labels** were added back to the data for interpretation.
4. A graph of a cluster distribution

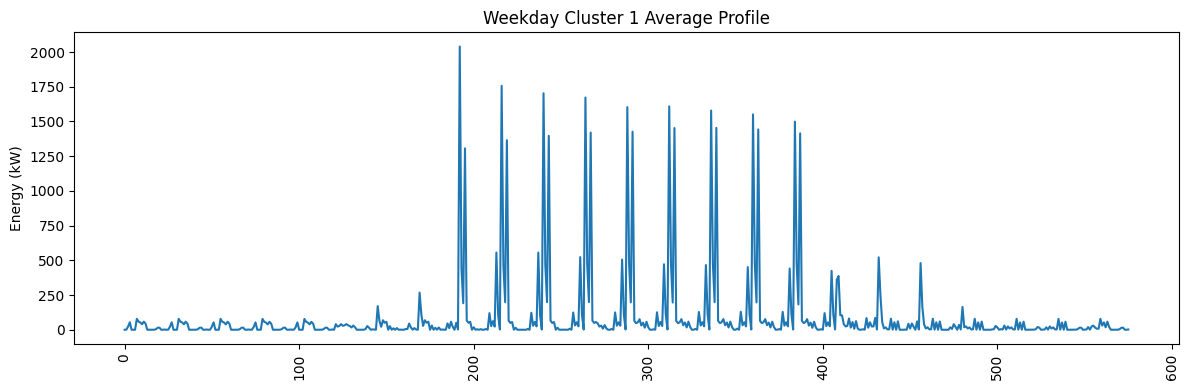
   AI-generated content may be incorrect.

**Insights & Interpretation**

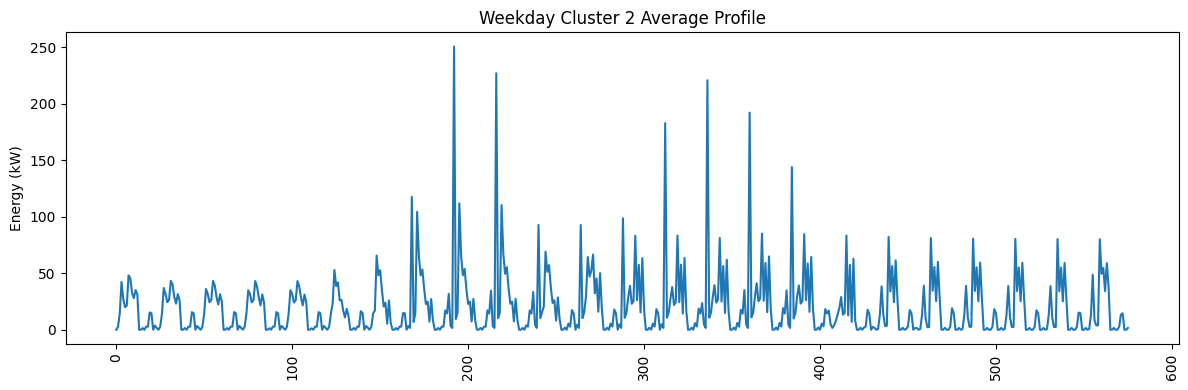
* **Weekday Clusters** showed:
  + **Cluster 0**: Typical working day pattern with peak usage between 9:00 AM – 5:00 PM.



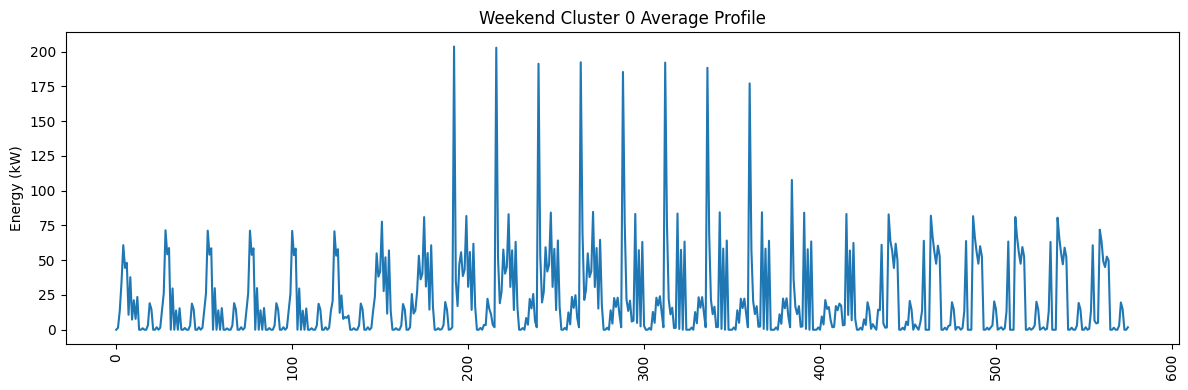
* + **Cluster 1**: Slightly lower usage, possibly half-staff days or partial holidays.



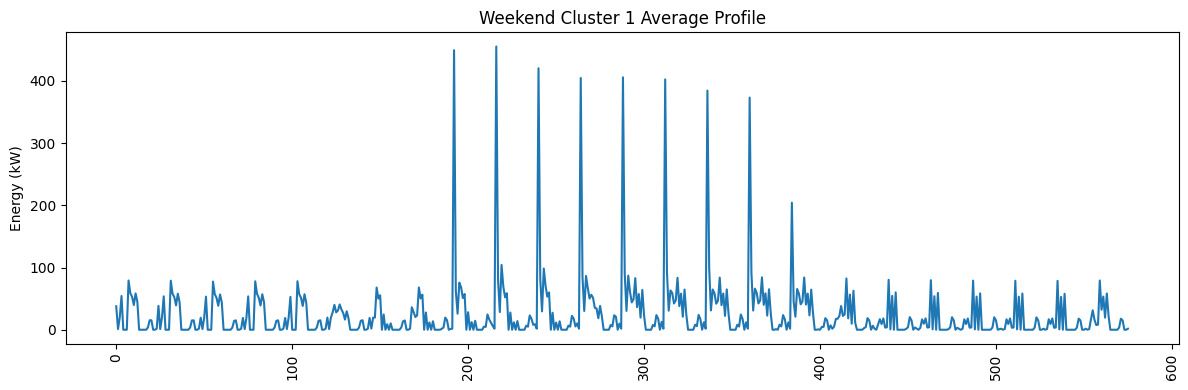
* + **Cluster 2**: Anomalies or energy-saving days (very low activity)



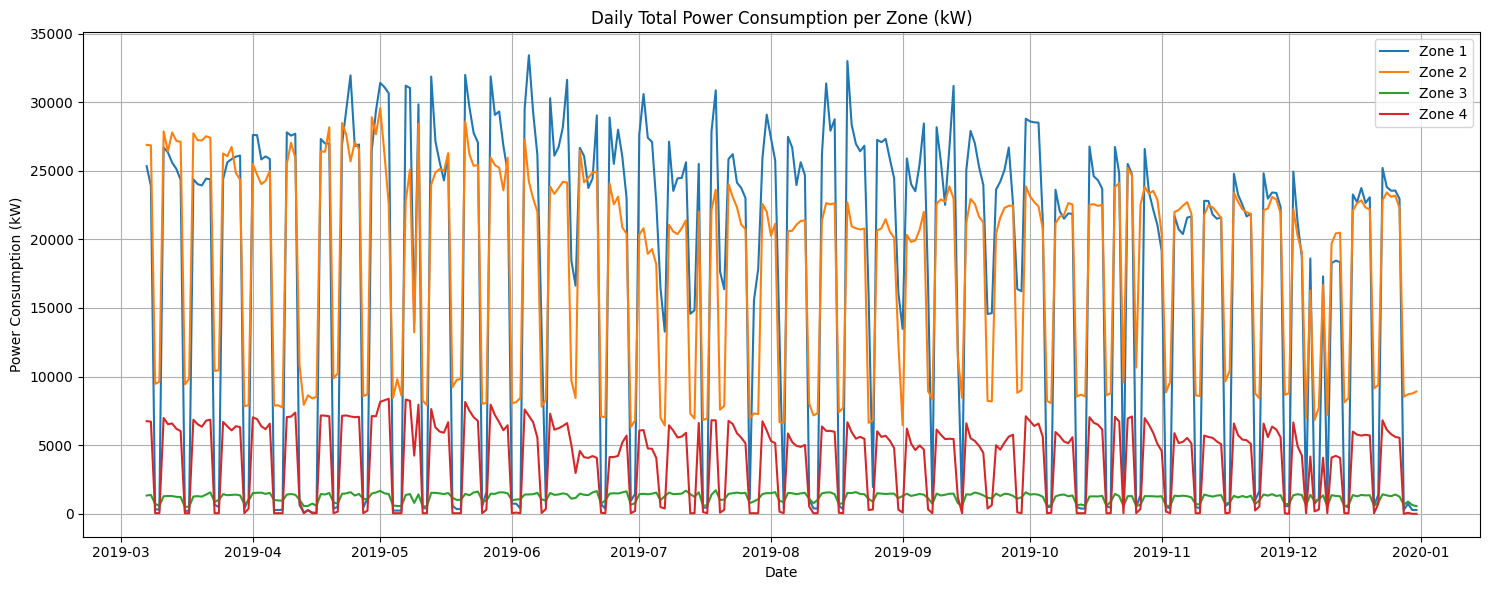
* **Weekend Clusters** showed:
  + **Cluster 0**: Mostly idle days with almost no activity.



* + **Cluster 1**: Unexpected high usage on some weekend days (possible maintenance, overtime).



**Zone wise total power consumption**



**🔹 Zone 1 (Blue):**

* **Highest consumption** overall (up to 34,000 kW).
* **Large daily fluctuations**, consistent high usage.
* Slight drop toward end of year.

**🟠 Zone 2 (Orange):**

* **Second highest**, peaks around **28,000 kW**.
* Regular up-down pattern, possibly workday-related.
* Slight mid-year and year-end decline.

**🟢 Zone 3 (Green):**

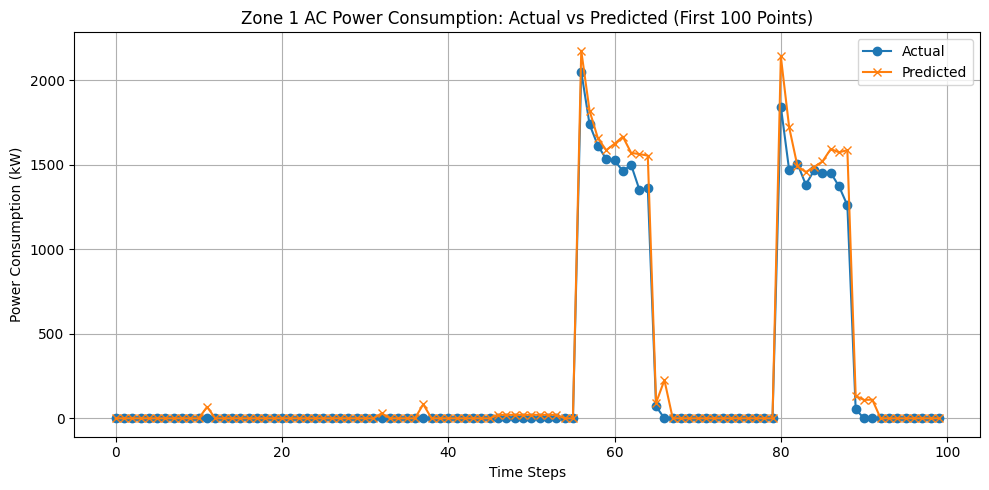
* **Lowest usage** (500–2000 kW).
* Very stable and minimal variation.
* Likely a low-demand area.

**🔴 Zone 4 (Red):**

* **Moderate usage** (1,000–8,000 kW).
* Regular pattern with some drops.
* More stable than Zones 1 & 2, but higher than Zone 3.

.  
Prediction model

**What we’re predicting:**  
The **Zone 1 AC power consumption** for the **next hour**.

**How we do it:**  
We use current hour’s **time features** (hour, day of week, month, weekend) and **sensor data** (temperature, humidity, illuminance) as inputs. We train a regression model on past data to learn the relationship and then predict the next hour’s AC load. Before training and predicting, we scale the features for better accuracy.  
  
**Evaluation  
**✅ Model Evaluation:

RMSE: 231.336 kW

MAE: 109.175 kW

R² Score: 0.886

**A screen shot of a computer program

AI-generated content may be incorrect.**

This can predict the next hour Zone\_1 AC\_1 consumption in Floor2 based on the give dataset