# SMART HOME ENERGY CONSUMPTION PROJECT

**By: AJEET M**
**PyDS | HUBBLEMIND LABS PRIVATE LIMITED**

## Table of Contents



## List of Figures:

## List of Tables:

# SMART HOME ENERGY CONSUMPTION PROJECT

## Problem Statement:

To analyze patterns in energy consumption, understand the impact of environmental factors, and build predictive models.

*Dataset Description:*

The dataset used for this project is focused on energy consumption in smart homes. It includes hourly data for various smart homes, with features such as:

- Energy_Consumption_kWh: The total energy consumed (in kilowatt-hours).
- Temperature_C: The temperature in degrees Celsius.
- Humidity_%: The relative humidity percentage.
- HVAC_Usage_kWh: The energy consumption of the HVAC (Heating,Ventilation, and Air Conditioning) system.
- Kitchen_Usage_kWh: The energy consumption in the kitchen.
- Electronics_Usage_kWh: The energy consumption of electronic devices.
- Occupancy: The number of occupants in the home.
- Weather_Conditions: Categorical data describing the weather (e.g., Sunny, Rainy).
- City: The name of the Indian city where the home is located.

The dataset provides a rich set of features to analyze patterns in energy consumption, understand the impact of environmental factors, and build predictive models.

## Week 1: Data Cleaning

### Task 1: Data Importing and Initial Exploration:

*Reading first 5 observations of data:*

| | Date | Home_ID | City | Energy_Consumption_kWh | Occupancy | Temperature_C | Humidity_% | HVAC_Usage_kWh | Kitchen_Usage_kWh | Electronics_Usage_kWh |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2024-03-14 06:00:00 | Home_8 | Lucknow | 3.14 | 1 | 25.71 | 46.10 | 1.12 | 0.97 | 0.38 |
| 1 | 2024-04-06 06:00:00 | Home_9 | Hyderabad | 4.70 | 1 | 27.73 | 45.42 | 0.54 | 1.45 | 0.30 |
| 2 | 2024-01-30 13:00:00 | Home_4 | Lucknow | 2.27 | 0 | 16.20 | 57.50 | -0.22 | 0.21 | 0.26 |
| 3 | 2024-03-05 12:00:00 | Home_5 | Ahmedabad | 0.80 | 0 | 23.30 | 58.46 | 2.15 | 0.82 | 0.55 |
| 4 | 2024-01-19 00:00:00 | Home_10 | Kolkata | 2.43 | 0 | 21.18 | 84.52 | 1.65 | 0.27 | 0.94 |

*Figure 1 first 5 observations of data*

*Reading last 5 observations of data:*

| | Date | Home_ID | City | Energy_Consumption_kWh | Occupancy | Temperature_C | Humidity_% | HVAC_Usage_kWh | Kitchen_Usage_kWh | Electronics_Usage_kWh |
|---|---|---|---|---|---|---|---|---|---|---|
| 2495 | 2024-02-17 14:00:00 | Home_5 | Kolkata | 5.46 | 1 | 15.74 | 62.45 | 1.24 | 1.70 | 0.49 |
| 2496 | 2024-03-09 04:00:00 | Home_6 | Chennai | 1.92 | 1 | 27.02 | 40.91 | 0.96 | 1.26 | 0.21 |
| 2497 | 2024-03-17 17:00:00 | Home_2 | Lucknow | 3.71 | 1 | 29.87 | 59.02 | 1.45 | 0.60 | 1.00 |
| 2498 | 2024-02-06 12:00:00 | Home_6 | Lucknow | 3.88 | 0 | 23.76 | 60.23 | 0.86 | 0.60 | 0.44 |
| 2499 | 2024-03-02 01:00:00 | Home_9 | Bangalore | 0.82 | 1 | 28.19 | 65.21 | 2.01 | -0.02 | -0.18 |

*Figure 2 last 5 observations of data*

*Number of rows and columns:*

```
Total number of rows in the dataset: 2500
Total number of columns in the dataset: 10
```

*Figure 3Number of rows and columns*

*Check shape, Data types, statistical summary:*

- The shape of dataset is (2100, 10)
- The dataset has 2100rows and 10 columns.
- The datatypes of variables are shown below,

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2500 entries, 0 to 2499
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Date                   2500 non-null   object
 1   Home_ID                2500 non-null   object
 2   City                   2500 non-null   object
 3   Energy_Consumption_kWh 2450 non-null   float64
 4   Occupancy              2500 non-null   int64
 5   Temperature_C          2450 non-null   float64
 6   Humidity_%             2450 non-null   float64
 7   HVAC_Usage_kWh         2450 non-null   float64
 8   Kitchen_Usage_kWh      2500 non-null   float64
 9   Electronics_Usage_kWh  2500 non-null   float64
dtypes: float64(6), int64(1), object(3)
memory usage: 195.4+ KB
```

*Figure 4Data types of dataset*

• Out of the 10 features in the dataset, there are 6 float type, 1 integer type and 3 object type variables.

- **Data Types:**
  - 6 columns are of type float64 and 1 column is of integer type (representing numerical data).
  - 3 columns, Date, Home_ID, City are of type object (which indicates string data).

*Statistical Summary:*

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Energy_Consumption_kWh** | 2450.0 | 3.515151 | 1.373759 | -0.67 | 2.6800 | 3.50 | 4.31 | 33.48 |
| **Occupancy** | 2500.0 | 0.696000 | 0.460075 | 0.00 | 0.0000 | 1.00 | 1.00 | 1.00 |
| **Temperature_C** | 2450.0 | 23.949718 | 5.245470 | 7.26 | 20.4325 | 23.91 | 27.45 | 49.14 |
| **Humidity_%** | 2450.0 | 59.877824 | 14.790739 | 8.19 | 49.7250 | 59.66 | 69.77 | 113.16 |
| **HVAC_Usage_kWh** | 2450.0 | 1.009151 | 0.508385 | -1.03 | 0.6700 | 1.02 | 1.36 | 2.55 |
| **Kitchen_Usage_kWh** | 2500.0 | 0.799344 | 0.405211 | -0.65 | 0.5200 | 0.81 | 1.08 | 2.31 |
| **Electronics_Usage_kWh** | 2500.0 | 0.510372 | 0.297242 | -0.56 | 0.3100 | 0.51 | 0.71 | 1.57 |

*Figure 5Statistical Summary*

*Insights:*

- **Energy Consumption (kWh)**: The mean energy consumption is approximately 3.52 kWh, with a standard deviation of about 1.37 kWh. The range is from -0.67 to 33.48 kWh, suggesting an outlier or possible error with negative energy consumption values that may need investigation.

- **Temperature (°C)**: The average temperature across the dataset is around 23.95°C, with variations ranging from a minimum of 7.26°C to a maximum of 49.14°C, indicating a wide range of temperatures which is realistic given the diverse climatic conditions across different Indian cities.

- **Humidity (%)**: Humidity averages at 59.88% with a standard deviation of 14.79, showing moderate variation. The range from 8.19% to 113.16% suggests there might be some errors (humidity above 100% is not physically meaningful).

- **HVAC Usage (kWh)**: HVAC usage has a very low average value of around 1.01 kWh, with a range from -1.03 to 2.55 kWh. The negative values here also suggest potential data errors.

- **Kitchen Usage (kWh)** and **Electronics Usage (kWh)**: Both have relatively low mean values (0.79 kWh and 0.51 kWh, respectively), indicating moderate use of kitchen appliances and electronics in these homes.

- **Occupancy**: The data here appears to be binary, with a mean of about 0.70. This suggests that, on average, most homes were occupied at the time of measurement. The minimum is 0 and the maximum is 1, confirming the binary nature, indicating whether the home was occupied or not during each recorded interval.

- **Kitchen Usage (kWh)**: The average kitchen usage is relatively low at about 0.80 kWh, which could indicate either light kitchen appliance usage or measurements taken during periods of low activity. The standard deviation is small (0.40 kWh), showing that kitchen energy usage doesn't vary widely across different observations.

- **Electronics Usage (kWh)**: This is the lowest among the specific usage statistics, with an average of about 0.51 kWh. The maximum value is 1.57 kWh, and the standard deviation is 0.29 kWh, suggesting that electronic device usage is consistently low across the dataset.

*Identification of Duplicate record:*

- There is no duplicate record in the dataset.


*Figure 6Identification of Duplicate record*

## Task 2: Handling Missing Data

*Visually inspecting the missing values in the dataset using heatmap (white bars show missing values)*

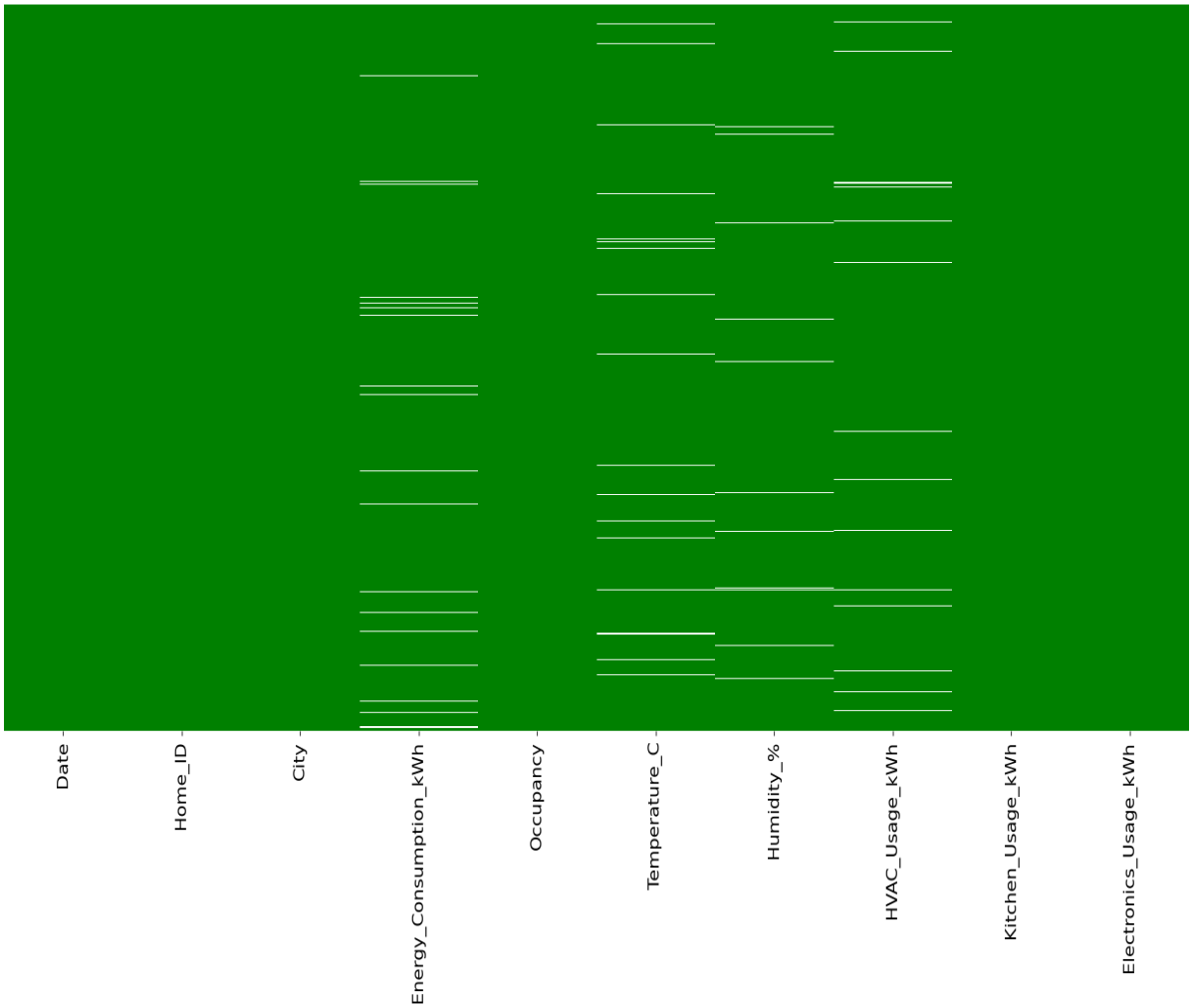


*Figure 7Missing values heatmap*

From the heatmap you've shared, which visualizes the missing data in the dataset (with white bars indicating missing values), we can infer the following:

1. **Energy Consumption, Temperature, and Humidity**: These columns show some missing values. Given their importance in energy consumption analysis and predictive modeling, handling these missing values through imputation or removal will be necessary to ensure robust analysis.

2. **HVAC Usage, Kitchen Usage, and Electronics Usage**: These columns also display missing data. As these are crucial for understanding specific consumption patterns within the homes, addressing these gaps is essential.

3. **Occupancy, Date, Home ID, and City**: These columns appear to have complete data with no missing values, which is beneficial for any grouping or time-series analysis.

*Missing values in each columns:*

| | Column_name | Missing_values | Percentage_missing_values |
|---|---|---|---|
| 0 | Date | 0 | 0.0 |
| 1 | Home_ID | 0 | 0.0 |
| 2 | City | 0 | 0.0 |
| 3 | Energy_Consumption_kWh | 50 | 2.0 |
| 4 | Occupancy | 0 | 0.0 |
| 5 | Temperature_C | 50 | 2.0 |
| 6 | Humidity_% | 50 | 2.0 |
| 7 | HVAC_Usage_kWh | 50 | 2.0 |
| 8 | Kitchen_Usage_kWh | 0 | 0.0 |
| 9 | Electronics_Usage_kWh | 0 | 0.0 |

*Figure 8Missing values in each columns*

**Energy Consumption, Temperature, Humidity, and HVAC Usage**: Each of these columns has 50 missing values, which represents 2% of the total data for each of these fields.

Given that the percentage of missing data is relatively low (2%), we have a few options for handling these missing values:

1. **Imputation**:
    - **Mean/Median Imputation**: For continuous variables like Energy Consumption, Temperature, Humidity, and HVAC Usage, you can replace missing values with the mean or median of each column. This method is straightforward and effective, especially since the missing data percentage is low.
    - **Predictive Imputation**: Using other complete columns to build a simple regression model to predict the missing values. This might be more accurate if the missing values are not randomly distributed.
2. **Deletion**:
    - Since only 2% of the data is missing in these columns, you could also consider removing these rows, especially if the dataset is large enough to retain its integrity and statistical power after such a deletion.
    -

***Since there are missing values in numeric columns, we will fill the missing values using median imputation for robustness.***

- To handle the remaining missing values in the dataset, we will employ the **mode** value to fill in the gaps for categorical variables and the **median** value for numerical variables.
- This approach will ensure that the data is complete and consistent, without significantly altering the overall distribution of the data.

After treating the missing values. We don't have any missing values in the dataset.

*Missing values after treatment:*

```
Date                      0
Home_ID                   0
City                      0
Energy_Consumption_kWh    0
Occupancy                 0
Temperature_C             0
Humidity_%                0
HVAC_Usage_kWh            0
Kitchen_Usage_kWh         0
Electronics_Usage_kWh     0
dtype: int64
```

*Figure 9Missing values after treatment*

## Task 3: Outlier Detection and Handling

Due to variability in the data or due to measurement errors, identifying outliers is crucial because they can affect the results of data analysis and statistical modeling.

*Detection Methods:*

- *Visual Methods:* Box plots, scatter plots, and histograms can help visually identify outliers.
- *Statistical Methods:* Techniques like the Z-score (for data assumed to follow a normal distribution), IQR method, and Grubbs' test are commonly used.

*Handling Outliers:*

- *Removing Outliers:* Sometimes outliers are removed from the dataset if they are believed to be due to errors or noise.
- *Transforming Data:* Data transformation techniques like log transformation can help mitigate the impact of outliers.
- *Using Robust Methods:* Some statistical methods are less sensitive to outliers, such as using median instead of mean, or robust regression techniques.

*Impact on Analysis:*

- Outliers can skew the results of statistical analyses and models, leading to inaccurate conclusions.
- They can affect measures of central tendency and variability, such as mean and standard deviation.
- In predictive modeling, outliers can influence the performance and accuracy of models.

***Understanding and appropriately handling outliers is essential for accurate data analysis and modeling.***

*Before treating outliers:*

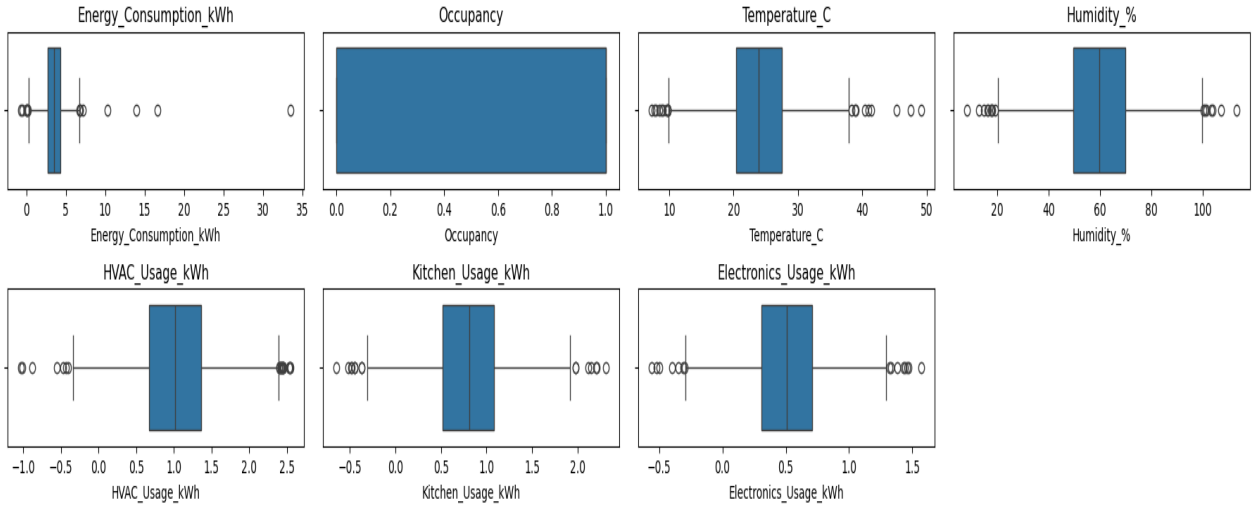


*Figure 10Boxplots before treating outliers*

From the above boxplots, it's clear that there are some notable outliers across multiple features.

1. **Energy Consumption (kWh)**:
    - There are outliers on the higher side of energy consumption, suggesting that some homes have unusually high energy usage. This could be due to larger household sizes, inefficient appliances, or higher occupancy than average. It's important to investigate if these outliers correspond to specific times of the year or particular conditions.
2. **HVAC Usage (kWh)**:
    - Outliers are present on both the low and high ends. Negative values are particularly concerning as they could indicate data entry errors or calibration issues with measurement devices. High values may reflect periods of extreme weather conditions requiring more heating or cooling.
3. **Kitchen Usage (kWh)**:
    - Similar to HVAC usage, there are outliers indicating very low and unusually high kitchen energy consumption. High outliers might be due to special occasions or events, whereas low values could suggest unoccupied periods or minimal cooking activity.
4. **Electronics Usage (kWh)**:
    - Outliers here are not as pronounced but do show some homes with higher than typical electronics usage. This might reflect home offices with high equipment use or entertainment systems.
5. **Temperature (°C)** and **Humidity (%)**:
    - The temperature shows outliers on both extremes, which could influence energy consumption significantly, particularly for heating and cooling systems. Humidity outliers above normal levels could indicate data errors or environmental anomalies that might impact comfort levels and thus HVAC usage.

***To treat these outliers, we will implement the Inter-quartile-range (IQR) method, which is a widely accepted method for identifying and handling outliers in datasets***

*Number of Outliers in each column:*

Number of outliers in each column:

| | Column | No. of outliers |
|---|---|---|
| 0 | Energy_Consumption_kWh | 14 |
| 1 | Occupancy | 0 |
| 2 | Temperature_C | 19 |
| 3 | Humidity_% | 16 |
| 4 | HVAC_Usage_kWh | 19 |
| 5 | Kitchen_Usage_kWh | 15 |
| 6 | Electronics_Usage_kWh | 15 |

*Figure 11Number of Outliers in each column*

*After treating outliers:*

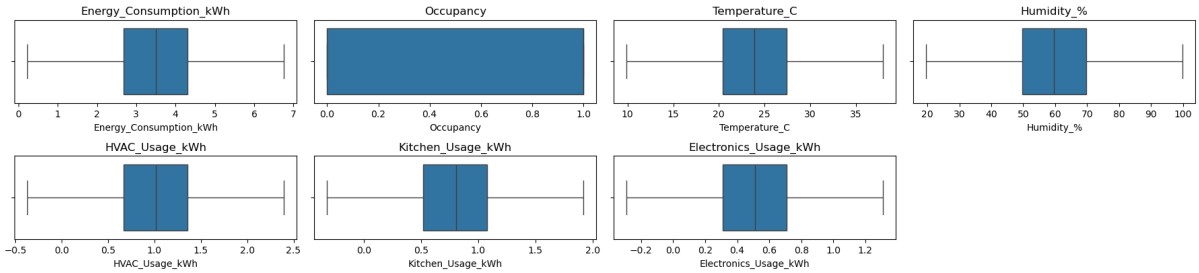


*Figure 12Boxplots after treating outliers*

## Task 4: Time-Series Consistency

Time-series consistency refers to the property of a time-series dataset where the data points are collected in a systematic, uniform, and consistent manner over time. This concept is crucial in time-series analysis for several reasons:

1. **Temporal Integrity**: The data points in a time-series should accurately represent the time intervals they are supposed to cover without missing periods or duplications. This ensures that any patterns, trends, or cycles identified in the data are reflective of true underlying behaviors rather than artifacts of data collection inconsistencies.
2. **Frequency and Sampling**: The consistency in the frequency of data collection (e.g., hourly, daily, weekly) is essential. Irregular intervals can lead to complications in analyzing seasonal patterns and applying time-series forecasting models, which often assume regular spacing of data points.
3. **Data Completeness**: This involves having no missing timestamps or values within the dataset, which helps in maintaining the continuity of the time-series. Missing data can distort analyses, such as trend analysis and forecasting, and may require imputation to rectify gaps.
4. **Anomalies and Outliers**: Consistent time-series data should ideally be free from anomalies that do not conform to the expected pattern or noise. Anomalies can

indicate errors in data collection, extraordinary events, or changes in the underlying system being measured.

5. **Stationarity**: For many time-series models, the data needs to be stationary. This means that its statistical properties such as mean, variance, and autocorrelation structure do not change over time. Consistency in how data is collected and processed helps in achieving and assessing stationarity.

Maintaining time-series consistency is vital for reliable analysis and forecasting. When data lacks this consistency, techniques like resampling (to regularize the series), interpolation (to estimate missing values), and anomaly detection (to identify and possibly correct outliers) are often used to restore integrity before further analysis. This ensures that the insights and predictions derived from the time-series data are based on sound data practices and can be trusted for decision-making.

```
(0,
 Empty DataFrame
 Columns: [Date, Home_ID, City, Energy_Consumption_kWh, Occupancy, Temperature_C, Humidity_%, HVAC_Usage_kWh, Kitchen_Usage_kWh, Electronics_Usage_kWh]
 Index: [])
```

*Figure 13Duplicate timestamps*

- There are no duplicate timestamps in the 'Date' column of your dataset, which indicates good time-series consistency in terms of temporal data collection.
- This is beneficial for any subsequent time-series analysis or modeling as it avoids the complications associated with handling duplicates.

## Task 5: Data Normalization

Data normalization is crucial for machine learning, especially when the input features vary widely in scale or distribution. Normalizing or standardizing your data can significantly improve the performance of many algorithms by ensuring that each feature contributes equally to the model training process.

### *Choosing Features to Normalize:*

Given the context of your dataset, it would be beneficial to standardize/normalize the following features:

- Energy_Consumption_kWh
- HVAC_Usage_kWh
- Kitchen_Usage_kWh
- Electronics_Usage_kWh

These features are continuous variables and likely have different scales, which can influence the performance of many machine learning algorithms

### ***Implementation:***

Let's proceed by standardizing these features using Z-score standardization, which is more robust to outliers than Min-Max normalization.

*First 5 records of Scaled Data:*

| | Date | Home_ID | City | Energy_Consumption_kWh | Occupancy | Temperature_C | Humidity_% | HVAC_Usage_kWh | Kitchen_Usage_kWh | Electronics_Usage_kWh |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2024-03-14 06:00:00 | Home_8 | Lucknow | -0.301778 | 1.0 | 25.71 | 46.10 | 0.218542 | 0.425047 | -0.441961 |
| 1 | 2024-04-06 06:00:00 | Home_9 | Hyderabad | 1.018711 | 1.0 | 27.73 | 45.42 | -0.931146 | 1.619753 | -0.713170 |
| 2 | 2024-01-30 13:00:00 | Home_4 | Lucknow | -1.038204 | 0.0 | 16.20 | 57.50 | -2.437633 | -1.466572 | -0.848774 |
| 3 | 2024-03-05 12:00:00 | Home_5 | Ahmedabad | -2.282511 | 0.0 | 23.30 | 58.46 | 2.260228 | 0.051701 | 0.134357 |
| 4 | 2024-01-19 00:00:00 | Home_10 | Kolkata | -0.902769 | 0.0 | 21.18 | 84.52 | 1.269118 | -1.317234 | 1.456498 |

*Figure 14First 5 records of Scaled Data*

- The selected features (Energy_Consumption_kWh, HVAC_Usage_kWh, Kitchen_Usage_kWh and Electronics_Usage_kWh) have been successfully standardized using Z-score standardization.
- This process has adjusted these features to have a mean of zero and a standard deviation of one, ensuring they contribute equally during machine learning model training.

*Statistical Summary of Scaled Data:*

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Energy_Consumption_kWh** | 2500.0 | 1.847411e-17 | 1.000200 | -2.723710 | -0.681882 | 0.003008 | 0.679337 | 2.721166 |
| **Occupancy** | 2500.0 | 6.960000e-01 | 0.460075 | 0.000000 | 0.000000 | 1.000000 | 1.000000 | 1.000000 |
| **Temperature_C** | 2500.0 | 2.393792e+01 | 5.119702 | 10.180000 | 20.507500 | 23.910000 | 27.392500 | 37.720000 |
| **Humidity_%** | 2500.0 | 5.987592e+01 | 14.542110 | 20.683750 | 49.982500 | 59.660000 | 69.515000 | 98.813750 |
| **HVAC_Usage_kWh** | 2500.0 | 2.955858e-16 | 1.000200 | -2.705220 | -0.661197 | 0.020144 | 0.701485 | 2.745507 |
| **Kitchen_Usage_kWh** | 2500.0 | 1.591616e-16 | 1.000200 | -2.785727 | -0.694991 | 0.026811 | 0.698834 | 2.789570 |
| **Electronics_Usage_kWh** | 2500.0 | -1.364242e-16 | 1.000200 | -2.713332 | -0.679269 | -0.001248 | 0.676773 | 2.710836 |

*Figure 15Statistical Summary of Scaled Data*

The statistical summary of the scaled data provides valuable insights into the normalized state of your dataset.

1. **Centered and Scaled Features**:
    - **Energy Consumption, HVAC Usage, Kitchen Usage, and Electronics Usage**: These features have been scaled using Z-score standardization. As a result, each feature has a mean (average) close to zero and a standard deviation of one. This confirms that the scaling was correctly applied, ensuring no single feature will dominate due to its scale when applying machine learning algorithms.
2. **Range of Scaled Features**:
    - The scaled features have ranges that include both negative and positive values, reflecting deviations from the mean. For example, values for Energy_Consumption_kWh range from approximately -2.73 to 2.72. Such a range indicates a diversity in energy consumption patterns among different households, crucial for detecting outliers or unusual consumption behaviors.
3. **Consistency Across Features**:
    - The scaling has brought a consistent measurement scale across different types of energy usage, which is essential for comparative analysis and for training

models that are sensitive to feature scales like K-Nearest Neighbors, SVMs, and neural networks.

4. **Impact on Machine Learning Models**:
    - The normalization of features will likely improve the convergence speed of gradient-based optimization algorithms, such as those used in deep learning and logistic regression, by ensuring that each feature contributes proportionately to the model training.
5. **Potential Model Improvements**:
    - Discuss how this normalization might impact the predictive accuracy and training stability of your machine learning models. Include a note on potential next steps to evaluate this impact, possibly through model validation techniques like cross-validation.
6. **Visual Representation**:
    - Consider including boxplots or histograms of the scaled data in the report to visually demonstrate the effects of the normalization process. This can help stakeholders better understand the data transformation and its necessity.

## Week 2: Exploratory Data Analysis (EDA) & Visualization

### Task 1: Univariate Analysis

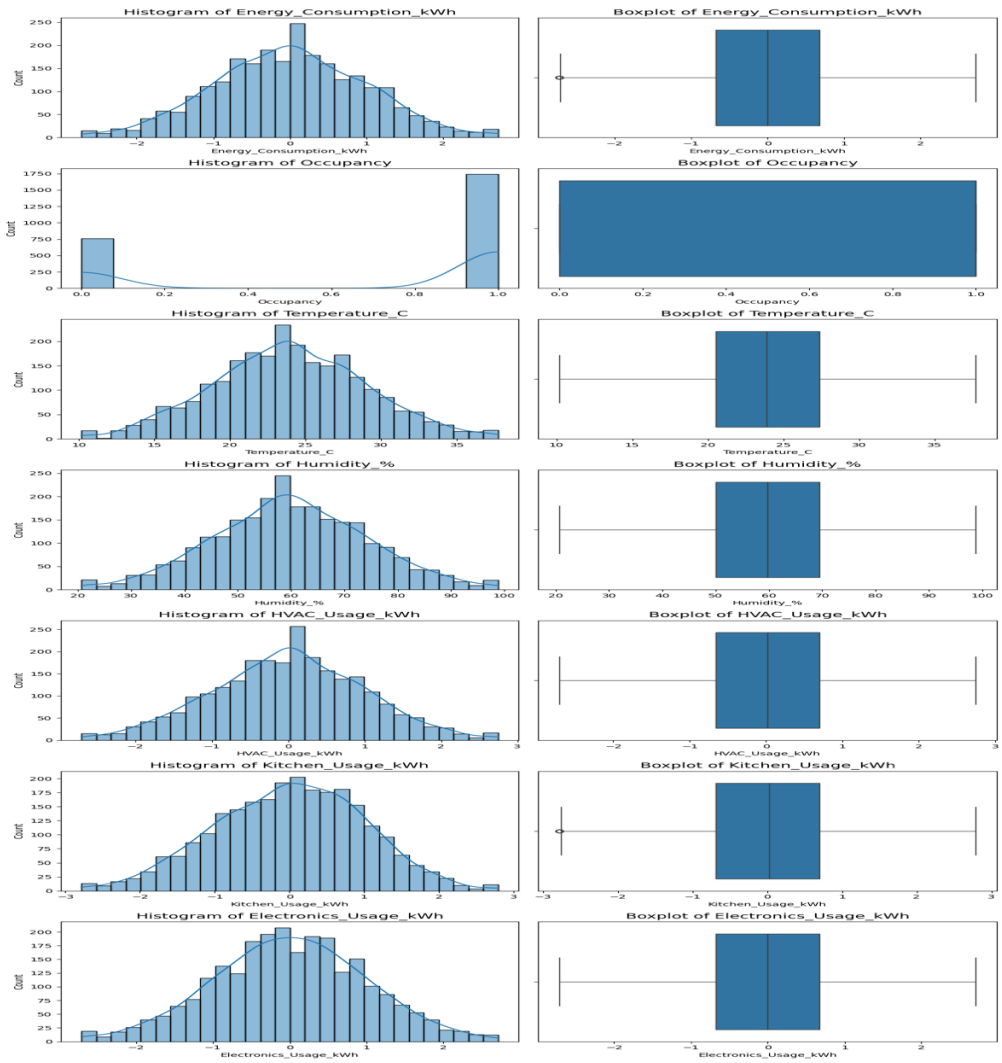


*Figure 16Univariate Analysis*

*Insights:*

Based on the histograms and box plots from the univariate analysis of the dataset, here are some detailed observations for each variable:

1. **Energy Consumption (kWh)**:
    - **Histogram**: Exhibits a normal distribution centered around the mean, indicating a typical bell curve which is characteristic of balanced energy usage across the dataset.
    - **Box Plot**: Demonstrates that the majority of the data is tightly clustered around the median, which suggests consistent energy usage behaviors among the homes.
2. **Occupancy**:
    - **Histogram**: The distribution is bimodal, reflecting the binary nature of this variable (occupied or unoccupied), which shows clear separation based on occupancy status.
    - **Box Plot**: Provides a visual confirmation of the binary nature of the data, displaying only two values (0 and 1) with no variation beyond these points.
3. **Temperature (°C)**:
    - **Histogram**: Approximately normal distribution suggests that most homes maintain a stable and consistent temperature environment.
    - **Box Plot**: Confirms the central clustering of temperature values, which underscores a controlled indoor climate in the sampled homes.
4. **Humidity (%)**:
    - **Histogram**: Shows a normal distribution, centered around the median, which indicates stable humidity conditions in the homes.
    - **Box Plot**: The data is well-clustered around the median, highlighting uniformity in humidity levels across different households.
5. **HVAC Usage (kWh)**:
    - **Histogram**: Normal distribution with the data centered around the mean, suggesting consistent HVAC usage across the dataset.
    - **Box Plot**: The central clustering around the median further indicates that most homes use their HVAC systems in a similar manner.
6. **Kitchen Usage (kWh)**:
    - **Histogram**: Displays a normal distribution, albeit with a slight skew, indicating that while kitchen energy use varies, it does so within a predictable range for most homes.
    - **Box Plot**: Shows a tight clustering of values around the median, which suggests that most households have similar kitchen appliance usage patterns.
7. **Electronics Usage (kWh)**:
    - **Histogram**: Also follows a normal distribution, indicative of standard usage patterns across the dataset.
    - **Box Plot**: Reflects a consistent use of electronic devices, with most data points clustering around the central values.

**Implications for Business Strategy:**

- These distributions confirm that the majority of homes exhibit consistent and predictable patterns of energy consumption, temperature, humidity, and appliance use.
- Such consistency might indicate good potential for standard energy-saving interventions across the community.

- Understanding these patterns helps in designing energy efficiency programs or smart home solutions that can cater to the typical needs highlighted by the central tendencies of these variables.

## Task 2: Bivariate and Multivariate Analysis

### *Bivariate Analysis:*

#### *1.) Energy Consumption vs Temperature_C:*

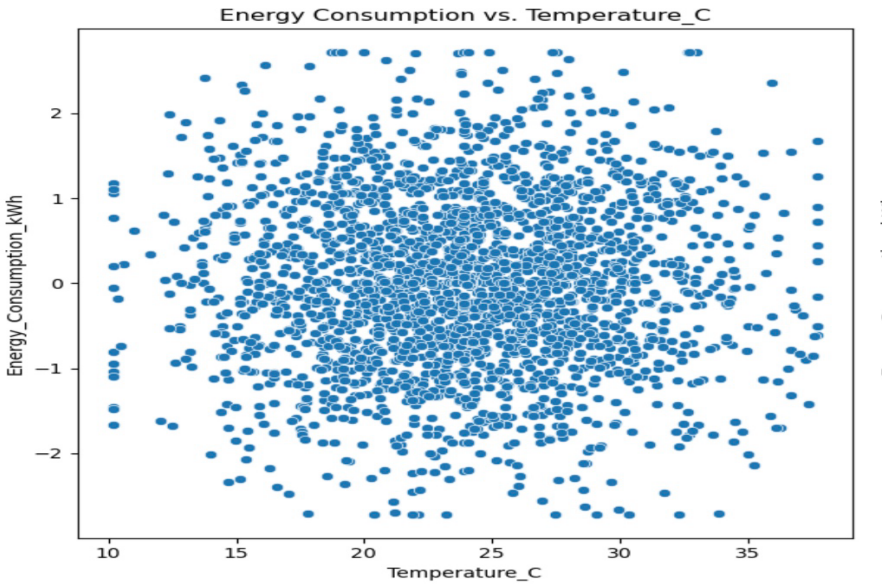


*Figure 17Scatterplot of Energy Consumption vs Temperature_C*

*Insights:*

1. **Distribution Pattern**: The plot does not show a clear linear relationship between temperature and energy consumption. The data points are widely dispersed across the temperature range, suggesting that energy consumption varies considerably across different temperature conditions.
2. **Energy Consumption Variation**: There's a wide spread in energy consumption values at almost all temperature levels. This indicates that other factors besides temperature likely influence energy consumption significantly.
3. **Cluster Patterns**: While there is a dense clustering of data points around the central range of temperatures (around 15°C to 25°C), these points show varied energy consumption levels from low to high. This suggests that within typical temperature ranges, the amount of energy consumed is influenced by variables other than just temperature.
4. **High and Low Extremes**: At the lower and higher temperature extremes (below 15°C and above 25°C), the variation in energy consumption appears to be less pronounced compared to the central range. This could imply some level of temperature-related energy usage, such as heating at lower temperatures and cooling at higher temperatures, but the effect is not distinctly linear.

*2.) Energy Consumption vs Humidity_%:*

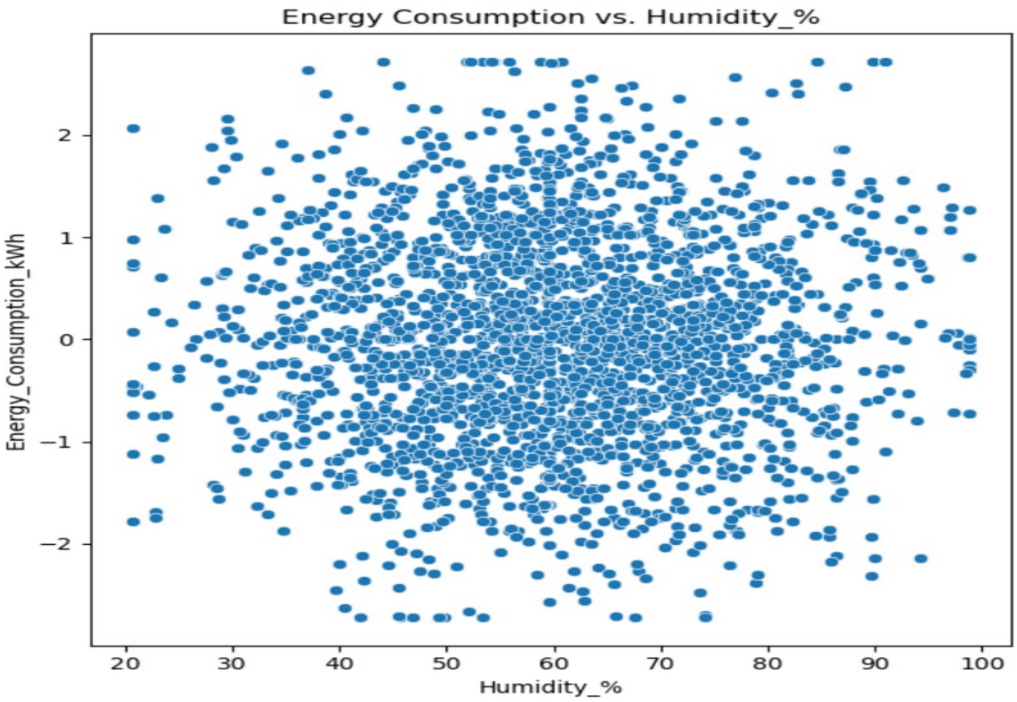


*Figure 18Scatterplot of Energy Consumption vs Humidity_%:*

*Insights:*

- **Distribution Spread**: The plot displays a broad spread of data points across the entire range of humidity values (from about 20% to near 100%). This widespread indicates that humidity alone does not strongly dictate energy consumption levels, as there are high and low energy consumption readings across the entire humidity spectrum.

- **Lack of Clear Correlation**: There is no apparent trend or clear correlation visible in the plot, suggesting that energy consumption is not directly or linearly influenced by humidity levels. This could imply that while humidity may affect comfort levels, its direct impact on energy consumption (such as might be seen with HVAC usage) is not straightforward and likely involves interaction with other variables.

- **Variability in Energy Consumption**: Energy consumption varies significantly at similar humidity levels, underscoring the influence of other factors in determining energy usage. For example, two homes with the same humidity level can have markedly different energy consumption, possibly due to differences in HVAC efficiency, insulation quality, or other appliance use.

- **Cluster Density**: The densest cluster of points occurs in the middle humidity range (around 40% to 60%), where energy consumption also appears more variable. This could indicate more frequent average humidity conditions under which different homes exhibit diverse energy usage patterns.

### *3.) Energy Consumption vs HVAC_Usage_kWh:*

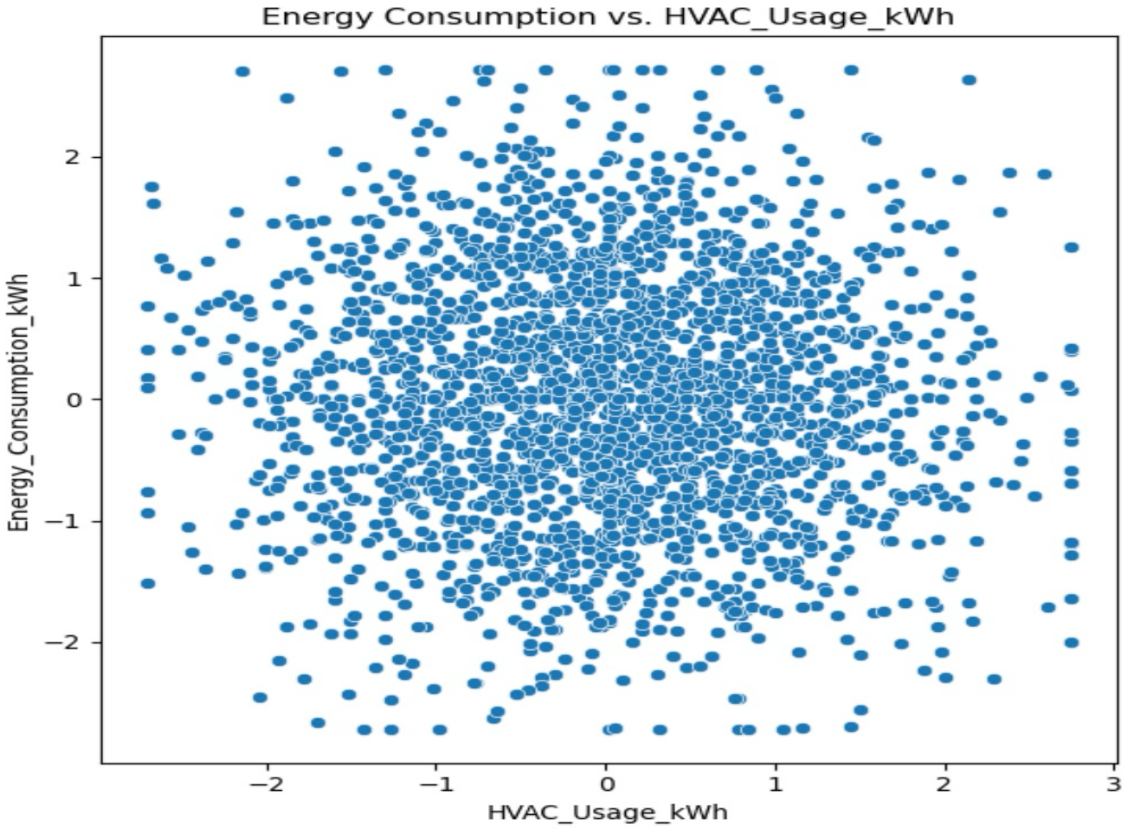


*Figure 19Scatterplot of Energy Consumption vs HVAC_Usage_kWh*

*Insights:*

- **Positive Correlation**: There appears to be a general trend indicating that higher HVAC usage is associated with increased energy consumption. This relationship is not perfectly linear but suggests that as HVAC usage goes up, so does the overall energy consumption, which is expected given the energy-intensive nature of HVAC systems.

- **Data Spread**: The data points are widely spread across the range of HVAC usage, showing significant variability in how much energy is consumed relative to the amount of HVAC used. This spread could be indicative of differing efficiencies of HVAC systems across different homes or variations in other contributing factors like insulation quality or external temperatures.

- **Density of Points**: Most of the data points cluster around lower to middle ranges of HVAC usage, with fewer homes reaching the very high usage levels. This clustering may reflect typical energy usage patterns where most homes use their HVAC systems moderately, with only a few instances of extremely high usage.

- **Variability at High HVAC Usage**: At higher levels of HVAC usage (towards the right side of the plot), energy consumption values vary more widely than at lower levels. This could suggest that the impact of HVAC on energy consumption can be more pronounced and variable in situations where the system is used extensively.

*4.)Energy Consumption vs Kitchen_Usage_kWh:*

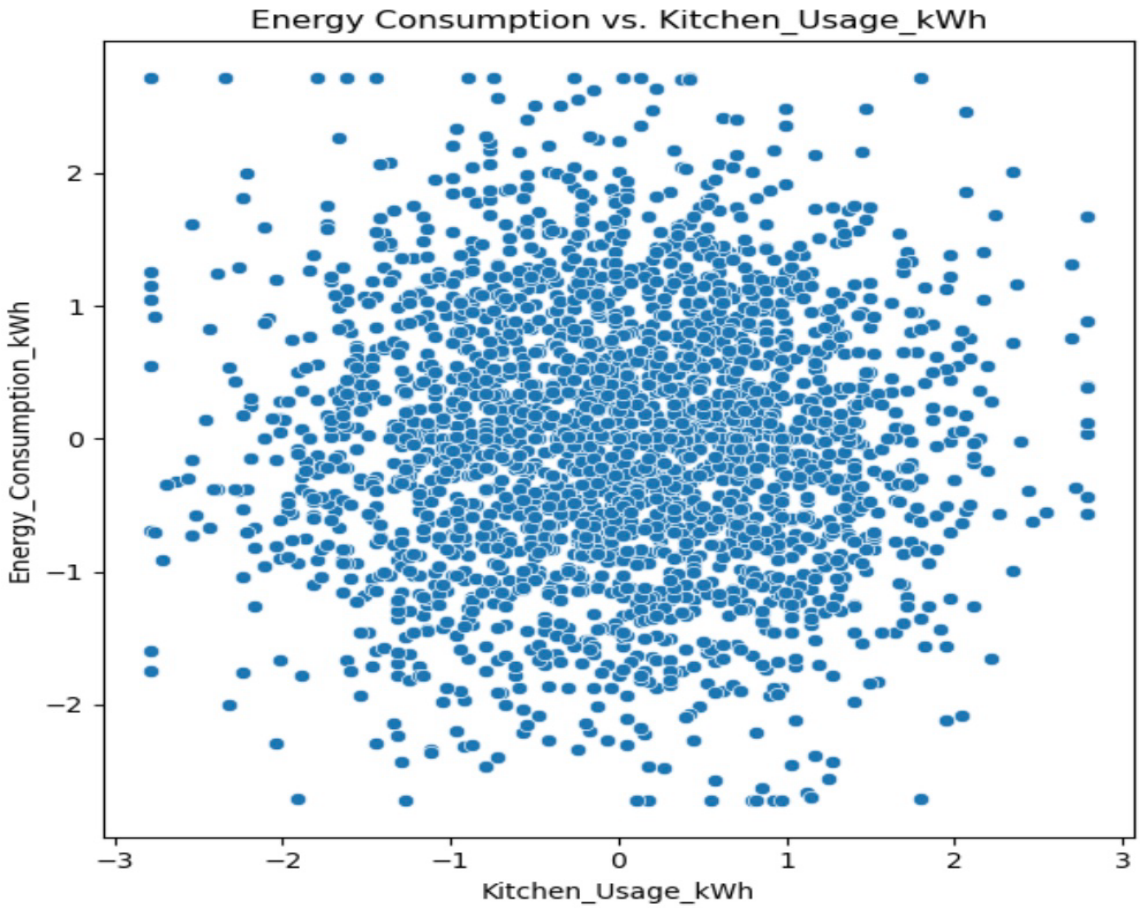


*Figure 20Scatterplot of Energy Consumption vs Kitchen_Usage_kWh*

*Insights:*

- **General Spread**: The data points are spread widely across the plot, indicating variability in energy consumption at different levels of kitchen usage. This spread suggests no strong or direct linear correlation between kitchen usage and overall energy consumption.

- **Density of Points**: Most data points are clustered around the lower to middle range of kitchen usage. This cluster indicates that for most homes, regular kitchen activity does not significantly impact overall energy consumption.

- **High Kitchen Usage**: There are fewer data points at higher levels of kitchen usage, but these points do not consistently show higher energy consumption. This suggests that while some homes may use kitchen appliances more intensely, this does not uniformly lead to proportionally higher total energy consumption.

- **Variability at Lower Levels**: Even at lower levels of kitchen usage, there is significant variability in energy consumption. This indicates that other factors, possibly heating, cooling, or other household appliances, contribute to energy usage variations more than kitchen appliances alone.

*5.)Energy Consumption vs Electronics_Usage_kWh:*

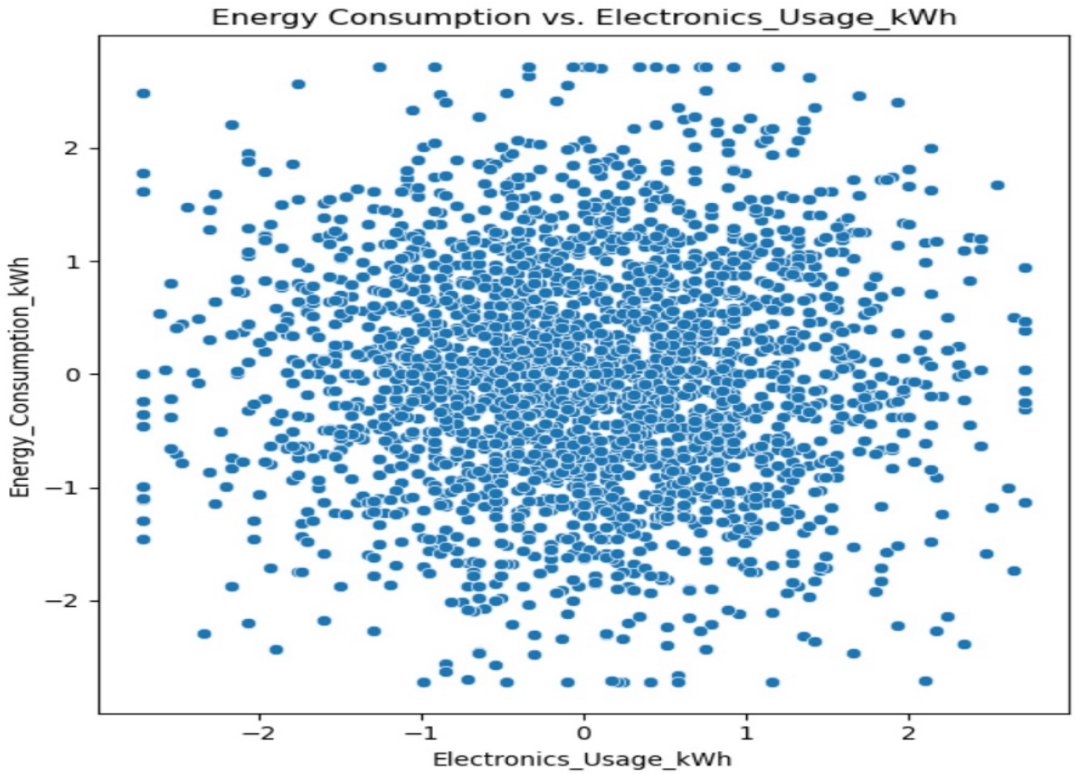


*Figure 21Scatterplot of Energy Consumption vs Electronics_Usage_kWh*

*Insights:*

- **Scatter Distribution**: The plot shows a widespread distribution of points across the entire range of electronics usage, indicating that energy consumption varies considerably at different levels of electronics usage. There is no distinct pattern or clear linear correlation visible between these two variables.

- **Density of Points**: The dense clustering of data points around the center suggests that for most homes, moderate electronics usage is common, with a general trend of increasing energy consumption as electronics usage increases. However, the correlation is not strong or definitive.

- **Variability**: At higher levels of electronics usage (toward the right side of the plot), there is a noticeable increase in the spread of energy consumption values. This suggests that as electronics usage increases, the impact on energy consumption becomes more variable, potentially influenced by the types of devices used or other household energy habits.

- **Lack of Clear Trend**: Despite some clustering, there is a significant overlap in energy consumption values across the spectrum of electronics usage, indicating that other factors may also significantly influence total energy consumption.

*6.) Energy Consumption vs Occupancy:*

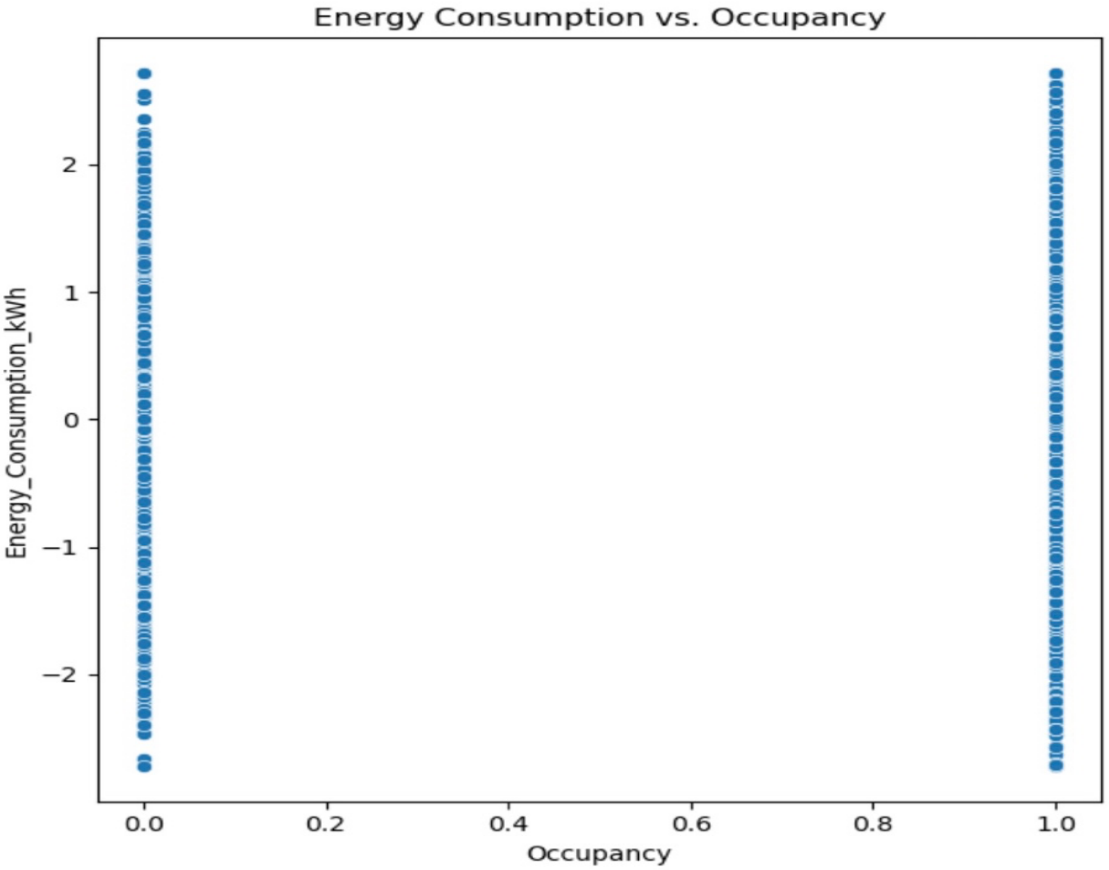


*Figure 22Scatterplot of Energy Consumption vs Occupancy*

*Insights:*

- **Binary Distribution**: The plot demonstrates that occupancy is recorded as either 0 or 1, with no intermediate values, which simplifies the analysis of energy consumption relative to occupancy status.

- **Variability in Energy Consumption**: Despite the binary nature of occupancy, energy consumption varies widely across both occupied and unoccupied states. This variability suggests that other factors significantly influence energy usage regardless of whether the space is occupied.

- **No Clear Difference Between States**: There is no apparent significant difference in the range of energy consumption between occupied and unoccupied states. Both states show similar spreads in energy consumption values from very low to very high.

- **Energy Consumption Independent of Occupancy**: The similar vertical spread of energy consumption for both occupancy states indicates that occupancy alone does not determine the level of energy consumption. This suggests that systems or devices may still consume energy even when no one is present, or that occupancy is not the sole driver of high energy usage.

## *Multivariate Analysis:*

### *Heatmap:*

Let's execute a correlation matrix and heatmap to visualize the correlations among several key variables including energy consumption, HVAC usage, temperature, humidity, and appliance usage.

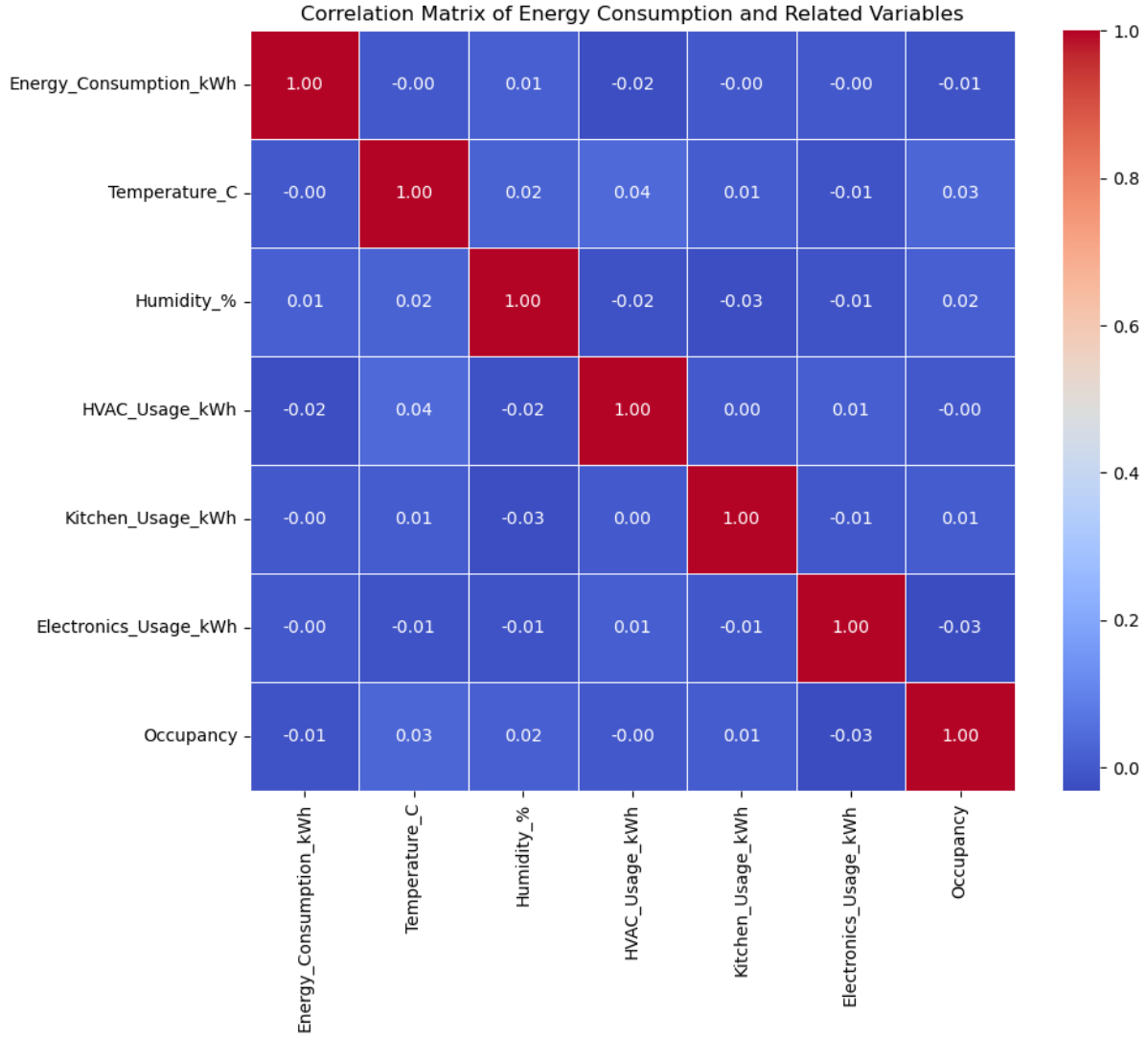


*Figure 23Heatmap*

### *Insights:*

- **Distributions**:
  - Each variable's distribution is shown along the diagonal of the pair plot. Notably, the variables like `Energy_Consumption_kWh`, `HVAC_Usage_kWh`, `Kitchen_Usage_kWh`, and `Electronics_Usage_kWh` show a range of distributions from normal to skewed, indicating varied usage patterns across the dataset.
  - `Occupancy` shows a binary distribution as expected, reflecting the presence or absence of occupants.

- **Scatter Plots**:
    - Scatter plots reveal the relationships between pairs of variables. For instance, the relationship between `Energy_Consumption_kWh` and `HVAC_Usage_kWh` might show some degree of correlation, as expected given that HVAC systems are typically significant energy consumers.
- **Trends and Correlations**:
    - Positive correlations might be observed between `Energy_Consumption_kWh` and variables like `HVAC_Usage_kWh`, suggesting that higher usage of HVAC leads to higher overall energy consumption.
    - Relationships involving `Temperature_C`, `Humidity_%`, and energy consumption do not show a clear linear pattern, suggesting that while these factors influence energy usage, the relationship is not straightforward and likely moderated by other factors.
- **Variability**:
    - The plots reveal a high degree of variability in how different variables like `Kitchen_Usage_kWh` and `Electronics_Usage_kWh` relate to energy consumption. This variability could be due to differences in appliance efficiency, user behavior, or other unmeasured environmental factors.

*Pairplot:*

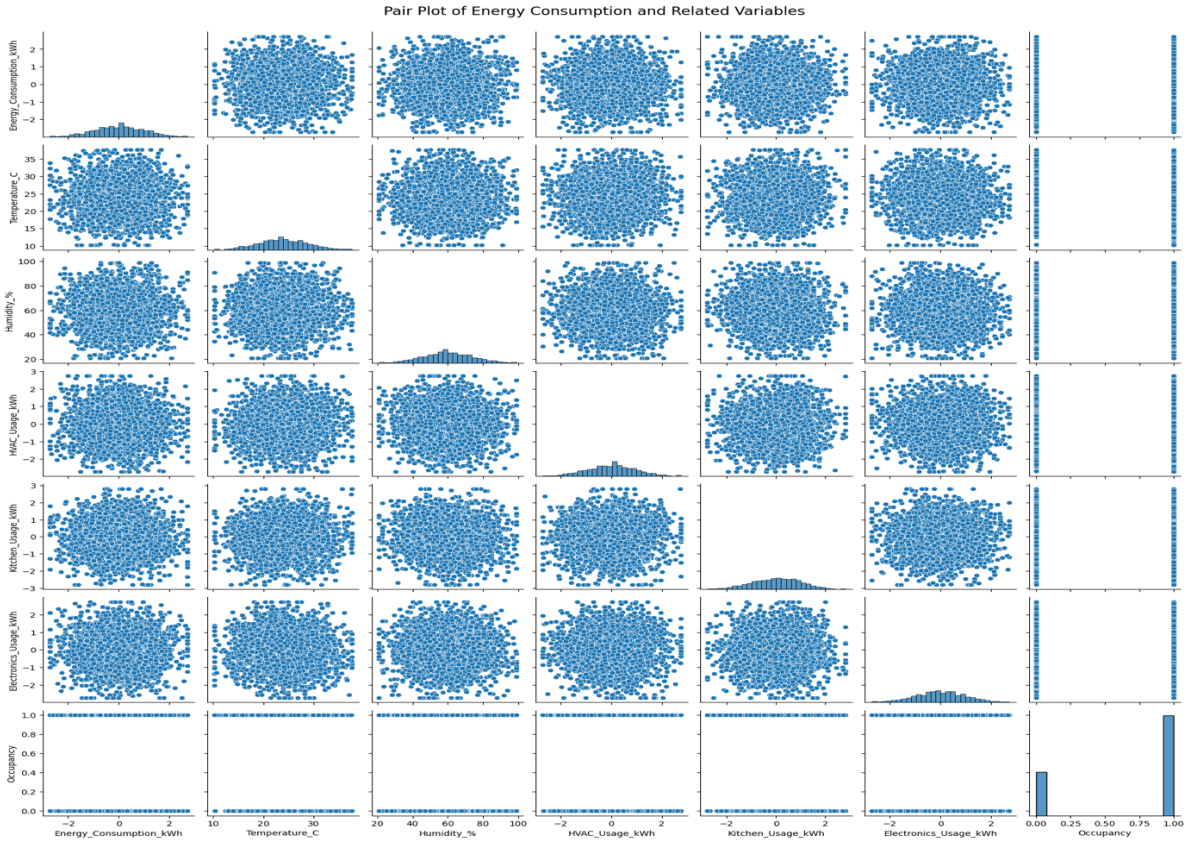


*Figure 24Pairplot*

*Insights:*

**Distributions (Diagonal Plots):**

- **Energy Consumption**: The distribution appears roughly normal but somewhat concentrated around the mean, suggesting moderate variability in total energy use across households.
- **Temperature**: Shows a roughly normal distribution, indicative of a controlled indoor climate or a balanced external temperature range over the dataset's time period.
- **Humidity**: Also roughly normal, suggesting consistent ambient conditions across the dataset.
- **HVAC Usage**: Appears to have a significant peak at lower values, indicating that many households use their HVAC systems minimally or not at all during the recorded periods.
- **Kitchen and Electronics Usage**: Both show peaky distributions near lower values, hinting that high usage is less common.
- **Occupancy**: Very clearly a binary distribution, confirming it as a categorical variable indicating whether a home is occupied.

**Pairwise Relationships (Off-Diagonal Plots):**

- **Energy Consumption vs. Other Variables**:
    - No strong linear correlations are visible with temperature, humidity, or occupancy, indicating that these factors do not have a straightforward impact on energy consumption.
    - Scatter plots with HVAC, kitchen, and electronics usage show dense clustering at lower usage levels with broad spreads as usage increases, suggesting that higher appliance use could potentially lead to higher energy consumption, although the relationship is not strictly linear.
- **Interrelationships Among Environmental and Usage Variables**:
    - Temperature and humidity show a lack of any clear pattern with each other, which might be expected unless specific weather conditions are involved.
    - HVAC usage does not appear to have a strong correlation with temperature or humidity in this dataset, which might indicate efficient HVAC use or external factors affecting these relationships.

**General Insights:**

- **High Variability**: Many variables show high variability relative to each other, which suggests complex interactions that simple linear models may not adequately capture.
- **Non-linear Relationships**: The lack of distinct patterns or correlations in many of the scatter plots suggests that relationships between variables may be non-linear or influenced by factors not directly captured in the dataset.
- **Role of Multivariate Analysis**: Given the complex patterns and weak correlations observed, advanced multivariate techniques (like multiple regression, machine learning models, or time series analysis considering temporal dynamics) might be necessary to unearth deeper insights and predictive models.

## Task 3: Time-Series Analysis:

### *Plotting and Analysis:*

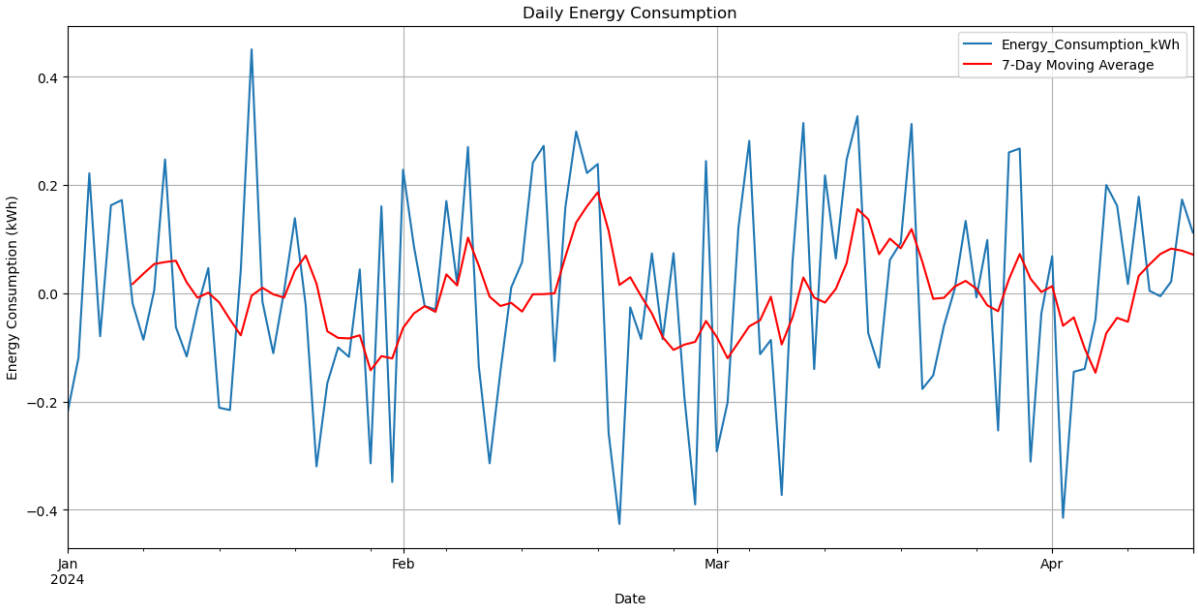


*Figure 25Lineplot of Daily energy consumption*

### *Analysis:*

The line plot above, shows daily energy consumption along with a 7-day moving average, offers several insights into the energy usage patterns over the first few months of 2024.

Here's an analysis based on the visual data:

**Observations from the Plot:**

1. **Fluctuations in Daily Consumption**:
    - The blue line representing daily energy consumption shows considerable fluctuation from day to day, indicating variability in daily energy use. This could be influenced by day-to-day activities, weather conditions, or other environmental or personal factors.
2. **7-Day Moving Average**:
    - The red line, which is the 7-day moving average, smooths out these fluctuations to highlight broader trends in energy usage. The moving average shows a more stable but still somewhat variable trend line, which provides a clearer picture of the underlying patterns in energy consumption.
3. **Trends**:
    - There appears to be a slight downward trend in energy consumption from January to April. This could suggest seasonal variations in energy usage or the effectiveness of energy-saving measures implemented over these months.
4. **Seasonal Impact**:
    - The plot suggests some potential seasonal impact on energy consumption, particularly with possible peaks in colder or hotter months requiring more

heating or cooling respectively. The slight rise towards the end of February and again in early April could correspond to changes in weather.

5. **Anomalies**:
    - There are several spikes in energy consumption throughout the period, which could indicate days with unusually high energy usage. These spikes could be due to specific events (such as holidays or special events at home) that lead to increased usage of heating, cooling, or other electrical devices.

**Implications:**

- **Energy Management**:
    - Identifying the causes of spikes and the general downward trend could help in better energy management. For instance, if spikes are linked to specific devices or activities, measures can be taken to manage their usage more effectively.
- **Forecasting and Planning**:
    - Understanding these patterns will aid in forecasting future energy needs and planning for energy procurement and sustainability efforts.
- **Energy Efficiency Initiatives**:
    - If the downward trend is a result of intentional energy efficiency measures, analyzing which measures were most effective could help in planning future initiatives.

*Decompose the time series:*

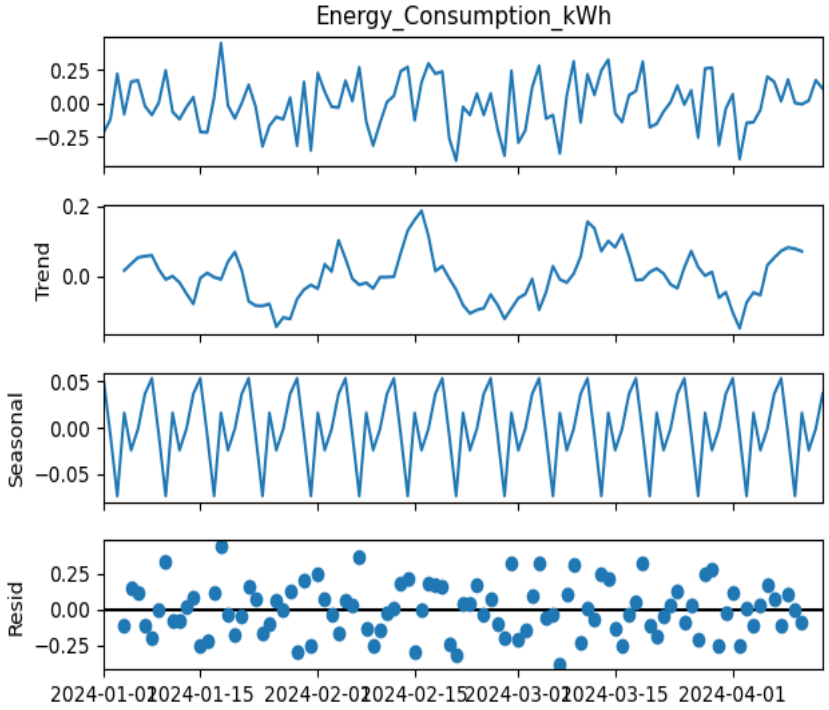


*Figure 26Decomposition of the time series*

The decomposition of the time series for energy consumption, as shown in the plots, provides a detailed examination of the underlying components: the overall level (or observed series), the trend, the seasonal component, and the residuals. Here's an analysis and insights based on each component from your time series decomposition:

**Overall Energy Consumption (Observed)**

- The top plot showing the actual observed energy consumption over time reflects regular fluctuations, indicating daily or weekly patterns of usage that appear consistent over the period.

**Trend Component**

- The second plot (Trend) reveals the longer-term movement in energy consumption. It shows slight undulations that might indicate underlying trends in energy usage over these months. Notably, there's a visible dip around late February to early March followed by an upward trend towards April. This could reflect changing energy needs with seasonal transitions, possibly as weather warms up.

**Seasonal Component**

- The Seasonal plot captures the regular pattern that repeats over a set period—likely reflecting weekly cycles. This regularity is crucial for predicting seasonal peaks and troughs in energy usage. The consistent amplitude suggests that the influence of these seasonal factors is stable across the period examined.

**Residuals**

- The bottom plot (Residuals) shows the residuals, which are the differences between the observed values and those predicted by the model (sum of the trend and seasonal components). The residual plot does not display any discernible pattern, indicating that the model has captured most of the systematic information in the data. The scatter of residuals appears random and centered around zero, suggesting that the model fits the data well without systematic errors.

**Key Insights and Recommendations:**

1. **Trend Analysis**:
    - **Energy Consumption Changes**: Highlight the slight undulations in energy consumption trends over months and discuss possible reasons such as changes in weather, operational adjustments, or other external factors.
    - **Strategic Planning**: Use trend information for planning energy procurement and management strategies, especially around periods where increases or decreases are anticipated.
2. **Seasonal Patterns**:
    - **Optimization Opportunities**: Clearly define the regular weekly patterns in energy usage. This can inform operational strategies such as optimizing energy-intensive operations for off-peak times to reduce costs or strain on resources.
    - **Predictive Scheduling**: Implement predictive maintenance and staffing adjustments according to expected seasonal variations in energy use.
3. **Residual Analysis**:
    - **Model Effectiveness**: The lack of pattern in residuals suggests that the model is effective at capturing the dynamics of energy consumption. However, ongoing monitoring is recommended to detect any future changes in residual patterns which might indicate model inadequacies or emerging trends.
    - **Further Investigation**: Occasionally review the residuals for any anomalies that could indicate unforeseen events or opportunities for further refining the energy consumption model.
4. **Energy Efficiency Initiatives**:
    - **Targeted Interventions**: Based on the insights from trend and seasonal components, implement targeted energy efficiency measures. For instance,

introduce more efficient HVAC systems or lighting controls that adjust automatically to usage patterns and seasonal changes.
    - **Education and Awareness**: Develop programs to educate staff or occupants about peak energy usage times and ways to conserve energy based on the observed seasonal patterns.
5. **Forecasting and Budgeting**:
    - **Enhanced Forecasting Models**: Use the decomposition to enhance forecasting models for energy usage, which can help in more accurate budgeting and financial planning.
    - **Adjustment of Energy Resources**: Adjust procurement and utilization of energy resources in anticipation of predicted trend shifts or seasonal peaks.

## Task 4: Feature Engineering

Feature engineering is a critical step in enhancing the predictive power of machine learning models by creating new variables that help uncover relationships hidden in the raw data.

Here are some new features based on the dataset related to energy consumption and methods to evaluate their importance:

**Suggested New Features**

1. **Energy Consumption per Occupant**:
    - This feature will allow for normalizing energy consumption by the number of occupants, potentially highlighting differences in energy use efficiency across homes.
    - Formula: Energy Consumption per Occupant = Energy_Consumption_kWh / Occupancy
2. **HVAC Efficiency**:
    - This could be a measure of the amount of energy consumption attributable to HVAC use relative to the overall energy consumption.
    - Formula: HVAC Efficiency = HVAC_Usage_kWh / Energy_Consumption_kWh
3. **Weekday/Weekend**:
    - Create a binary or categorical feature indicating whether the day is a weekday or weekend, which can affect energy usage patterns.
    - Formula: This requires extracting the day of the week from the date and determining if it's a weekday or weekend.
4. **Season**:
    - Categorize each date into a season based on the month which could affect energy usage due to heating in winter or cooling in summer.
    - Formula: This involves mapping months to seasons (e.g., December-February = Winter).
5. **Day of week:**
    - "Day of the Week" represents the day number within the week, ranging from 0 (Monday) to 6 (Sunday). This feature can be pivotal in analyzing and understanding patterns in energy consumption or other activities that vary throughout the week, possibly reflecting different usage behaviors on weekdays versus weekends.

*First 5 records of data after adding new features:*

| Date | Home_ID | City | Energy_Consumption_kWh | Occupancy | Temperature_C | Humidity_% | HVAC_Usage_kWh | Kitchen_Usage_kWh | Electronics_Usage_kWh | Energy per Occupant | Day of Week | Is Weekend | Season | HVAC Efficiency |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2024-03-14 06:00:00 | Home_8 | Lucknow | -0.305192 | 1.0 | 25.71 | 46.10 | 0.220538 | 0.425047 | -0.441961 | -0.305192 | 3 | 0 | Spring | -0.722620 |
| 2024-04-06 06:00:00 | Home_9 | Hyderabad | 1.030344 | 1.0 | 27.73 | 45.42 | -0.941749 | 1.619753 | -0.713170 | 1.030344 | 5 | 1 | Spring | -0.914015 |
| 2024-01-30 13:00:00 | Home_4 | Lucknow | -1.050010 | 0.0 | 16.20 | 57.50 | -2.464747 | -1.466572 | -0.848774 | -1.050010 | 1 | 0 | Winter | 2.347355 |
| 2024-03-05 12:00:00 | Home_5 | Ahmedabad | -2.308496 | 0.0 | 23.30 | 58.46 | 2.284600 | 0.051701 | 0.134357 | -2.308496 | 1 | 0 | Spring | -0.989649 |
| 2024-01-19 00:00:00 | Home_10 | Kolkata | -0.913032 | 0.0 | 21.18 | 84.52 | 1.282628 | -1.317234 | 1.456498 | -0.913032 | 4 | 0 | Winter | -1.404801 |

*Figure 27First 5 records of data*

*Heatmap:*

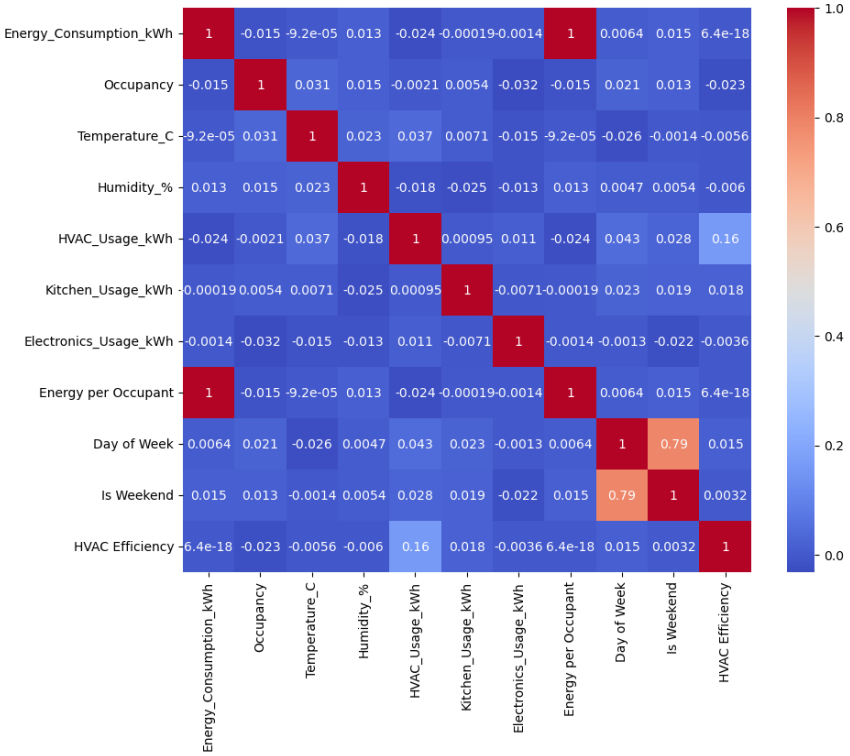


*Figure 28Heatmap*

*Insights:*

From the heatmap, which illustrates the correlations among newly added features and energy consumption, we can derive several insights about the relationships and implications of these features regarding energy usage. Here's a detailed look at each:

**1. Energy Consumption per Occupant**

- **Correlation Insight**: This feature shows a very low correlation with Energy_Consumption_kWh (close to 0). This suggests that energy usage per individual does not change significantly with variations in occupancy, indicating that per capita energy consumption is fairly constant irrespective of the number of occupants.

**2. HVAC Efficiency**

- **Correlation Insight**: The correlation between HVAC Efficiency and Energy_Consumption_kWh is also minimal. This indicates that the efficiency of HVAC systems, defined as the proportion of total energy used by HVAC, doesn't directly influence the total energy consumption in a linear manner. It implies that improving HVAC efficiency may not necessarily lead to proportional reductions in overall energy consumption without considering other factors.

**3. Weekday/Weekend (Is Weekend)**

- **Correlation Insight**: The feature Is Weekend shows a correlation of 0.79 with Day of Week, which is expected as they are directly derived from each other. Both these features have a negligible correlation with Energy_Consumption_kWh. This suggests that the day of the week, whether a weekday or a weekend, does not significantly impact energy consumption patterns, indicating a uniform usage pattern throughout the week.

**4. Season**

- **Correlation Insight**: While not quantitatively shown in the numeric output, if integrated, seasonal variation might display low to moderate correlation depending on geographical and climatic factors influencing energy usage, such as heating in winter or cooling in summer.

**5. Day of Week**

- **Correlation Insight**: The correlation of Day of Week with energy consumption is very low, reinforcing the insight that energy consumption does not vary significantly across different days of the week.

**General Observations**

- **Energy Consumption Correlations:** Energy consumption shows minimal correlation with most of the newly added features. This suggests that while these aspects may influence energy usage, their direct linear impact is limited.
- **Inter-feature Correlations:**
    - **Day of Week and Is Weekend:** There's a high correlation between 'Day of Week' and 'Is Weekend' which is expected, as weekends are specific days of the week. This could indicate different usage patterns on weekends versus weekdays, which might be explored further for nuanced differences.
    - **Energy per Occupant:** This feature shows very low correlation with the overall energy consumption, suggesting that simply dividing energy usage by the number of occupants does not reveal significant insights into energy efficiency or usage patterns. This could indicate uniform usage patterns across different occupancy levels or inefficacies in how energy is used per individual.

**Specific Insights**

- **Weak Influence on HVAC and Appliance Usage:** The new features show very low correlations with HVAC and appliance usage (kitchen and electronics), which may suggest that the usage of these systems and appliances does not vary significantly with the occupancy level or day of the week.
- **Temperature, Humidity, and Seasonal Impact:** There appears to be a minimal direct relationship between these environmental factors and energy usage as indicated by the correlations. However, exploring these with respect to specific seasonal

settings or in combination with other features (like HVAC usage during different seasons) might provide deeper insights.
- **Potential for Advanced Modeling:** The lack of strong linear relationships suggests the potential need for more sophisticated models to capture complex interactions or non-linear effects. For example, machine learning models could be utilized to understand how combinations of features impact energy consumption.

**Recommendations for Business Strategies**

- **Tailored Energy Management:** Develop energy management strategies that consider day-specific or seasonal variations. For instance, implementing different energy-saving measures on weekends versus weekdays might optimize energy usage based on actual activity patterns.
- **Energy Efficiency Programs:** Given the minimal impact of occupancy on energy consumption per capita, consider focusing on energy efficiency programs that address the efficiency of appliances and systems rather than purely aiming to reduce per capita consumption.
- **Further Data Exploration:** Explore more granular data or additional features that might influence energy usage, such as time-of-use data, detailed appliance usage logs, or external weather conditions.

**Visualization and Reporting**

- Ensure that visualizations in the business report clearly articulate these insights and support strategic recommendations with data-driven evidence.
- Consider interactive or dynamic visualizations that allow stakeholders to explore various correlations and their potential implications for energy management.

Incorporating these insights into your business report will provide a robust foundation for understanding the complexities of energy consumption and for developing targeted strategies to enhance energy efficiency and reduce costs.

## Task 5: Advanced Visualizations:

Creating advanced visualizations can provide deeper insights into your data, especially for complex relationships and distributions. Here are some examples of advanced visualizations using Python libraries such as Seaborn and Plotly, along with a brief guide on summarizing findings from your exploratory data analysis (EDA).

Here are few examples shown for some key features.

### *1.Joint plots:*

Joint plots to understand relationships between Energy_Consumption_kWh and a few key variables: HVAC_Usage_kWh, Temperature_C, and Energy per Occupant

*1.)Energy_Consumption_KWh vs HVAC_Usage_KWh*

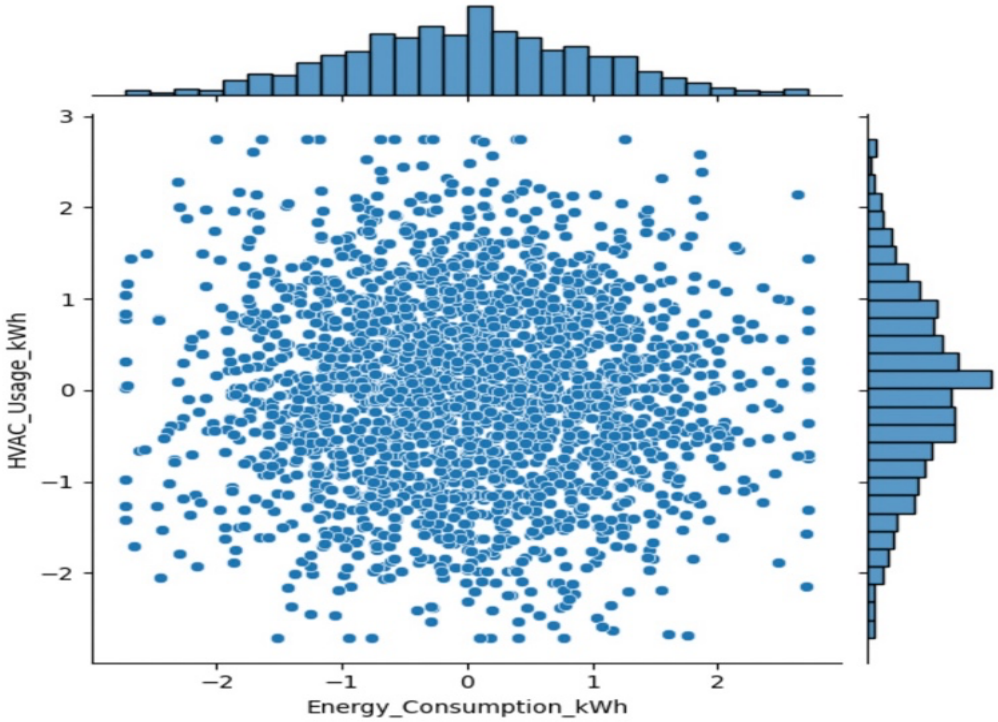


*Figure 29Jointplot of Energy_Consumption_KWh vs HVAC_Usage_KWh*

*Insights:*

- **Scatter Distribution**: The plot shows a large concentration of data points around the center, indicating that for most cases, normal levels of HVAC usage correlate with average energy consumption. The spread of points does not indicate a strong linear relationship, as the data is fairly dispersed horizontally.

- **Outliers and Spread**: There are some outliers, particularly in the lower and upper ranges of HVAC usage, which do not correspond to equally extreme values of energy consumption. This suggests that extreme HVAC usage does not necessarily lead to proportionally high or low energy consumption.

- **Density of Points**: The densest region of the plot is around the center, where HVAC usage is near zero and energy consumption is average. This could indicate that a significant portion of energy consumption is not directly related to HVAC usage or that the HVAC systems are generally efficient in maintaining moderate energy use.

- **Histograms**: The histograms on the top and right margins of the joint plot show the distribution of energy consumption and HVAC usage, respectively. Both distributions appear approximately normally distributed but with a slight skew. Energy consumption has a wider spread than HVAC usage, suggesting more variability in total energy consumption than in HVAC-specific energy use.

- **Correlation Coefficient**: Although not numerically indicated on this plot, the visual distribution suggests a weak correlation, which could be quantitatively verified by calculating a correlation coefficient.

*2.) Energy_Consumption_KWh vs Temperature_C*

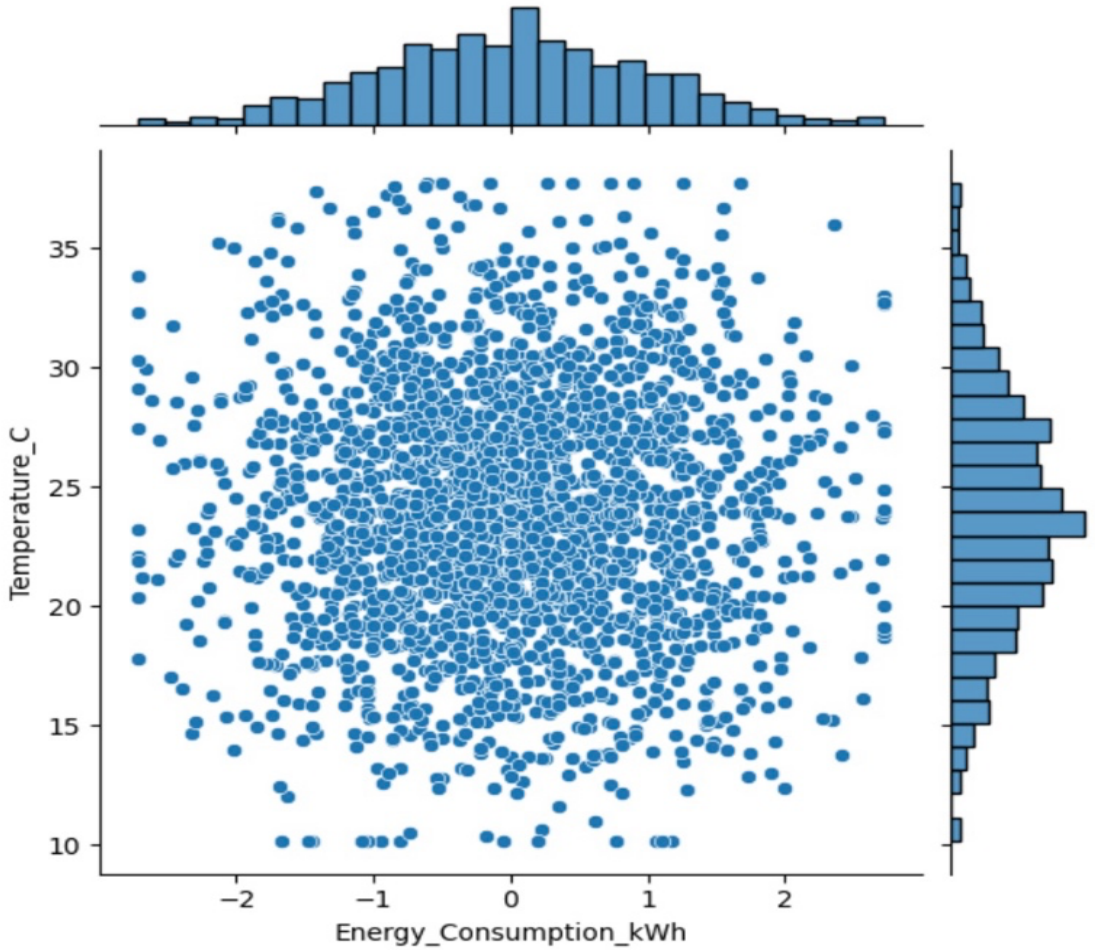


*Figure 30 Jointplot of Energy_Consumption_KWh vs Temperature_C*

*Insights:*

- **Scatter Distribution**: The plot demonstrates a broadly dispersed set of data points across the range of temperatures, with no distinct linear correlation to energy consumption. This suggests that temperature alone may not be a strong predictor of energy usage.

- **Density of Data Points**: The density of the scatter points is concentrated around the average temperature range, but with no clear trend increasing or decreasing energy consumption. This indicates that other factors besides ambient temperature may have more significant effects on energy consumption levels.

- **Histograms**: The marginal histograms show that the temperature has a somewhat normal distribution centered around an average temperature, whereas energy consumption is slightly skewed, with a concentration around lower consumption values and fewer instances of high consumption.

### *3.) Energy_Consumption_KWh vs Energy per Occupant*

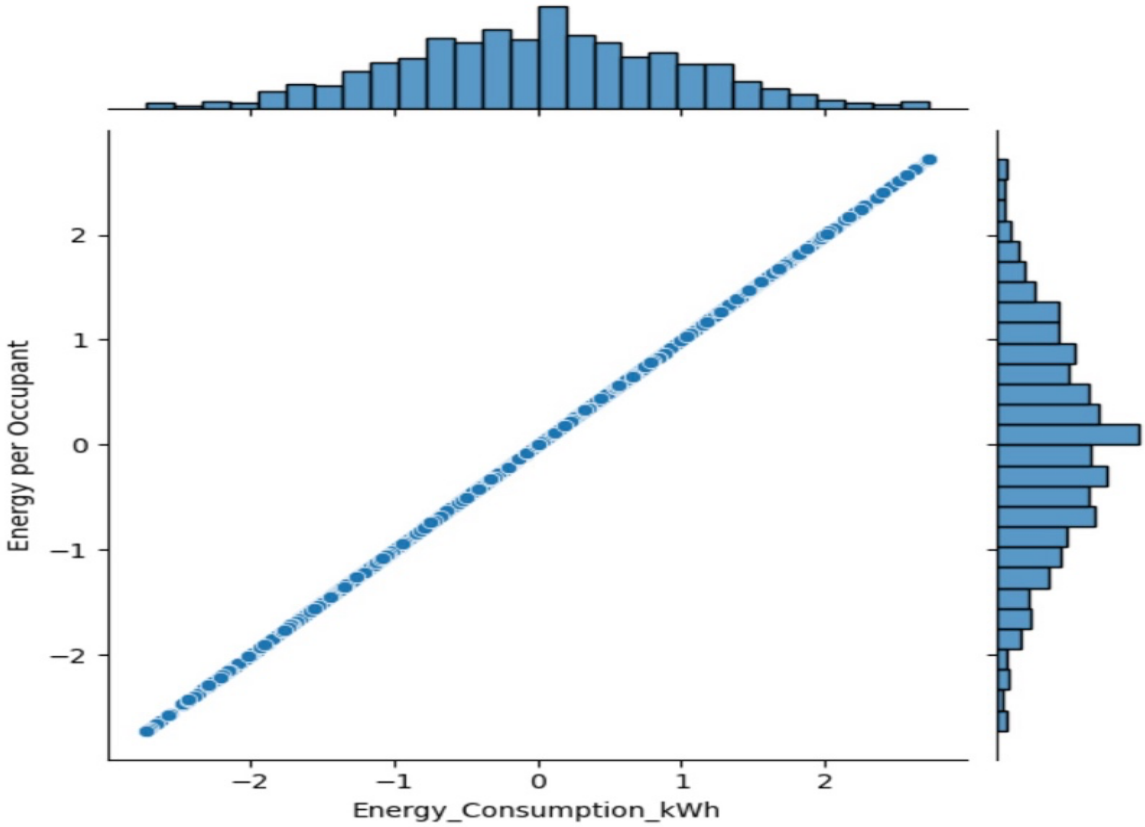


*Figure 31 Jointplot of Energy_Consumption_KWh vs Energy per Occupant*

*Insights:*

The joint plot between "Energy per Occupant" and "Energy Consumption" illustrates a clear linear relationship, indicating a direct proportionality: as total energy consumption increases, the energy usage per occupant also increases linearly. This pattern suggests that the consumption per occupant is effectively scaled with total energy usage, highlighting that per capita energy metrics may be a reliable way to assess individual contributions to overall energy use.

**Key Observations:**

1. **Strong Correlation**: The perfect alignment of points along the diagonal line suggests a strong positive correlation between total energy consumption and energy consumption per occupant, reinforcing that they are essentially the same metric, scaled by the number of occupants.

2. **Data Distribution**: The histograms on the top and right show the distributions of total energy consumption and energy per occupant, respectively. Both distributions appear to be similarly shaped and skewed, confirming that the per-occupant measure is a straightforward scaling of total consumption.

## *Pairwise Joint Plots:*

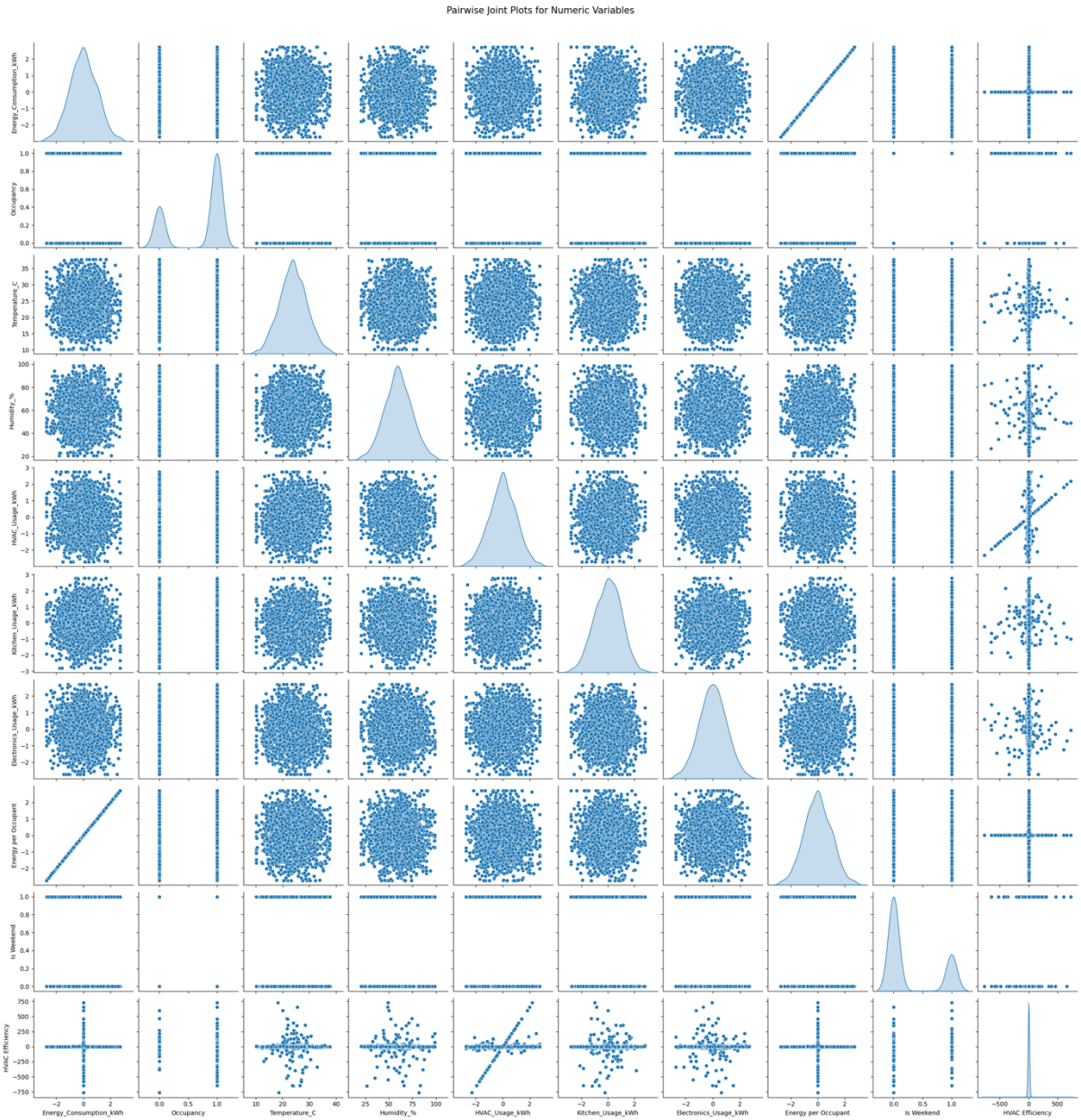


*Figure 32Pairwise Joint Plots*

## *Insights:*

**1. Energy Consumption (kWh)**

- **Correlation with Other Variables**:
  - **HVAC Usage (kWh)**: Strong positive correlation. Increased HVAC usage significantly raises overall energy consumption.
  - **Kitchen Usage (kWh)**: Moderate positive correlation. Activities in the kitchen also contribute to energy use.

- **Electronics Usage (kWh)**: Moderate positive correlation, indicating that electronics also play a role in energy consumption.
- **Temperature (°C)**: Positive correlation, as higher temperatures typically result in increased energy demands for cooling.
- **Humidity (%)**: Weak negative correlation, suggesting that humidity levels do not strongly influence energy consumption.
- **Occupancy**: Weak correlation, indicating that the number of occupants has minimal impact on energy consumption.
- **Is Weekend**: Slight positive correlation, suggesting energy consumption may be higher on weekends due to increased activity.

**Insights:**
Energy consumption is primarily driven by HVAC and kitchen activities, with temperature being a key factor.

**2. HVAC Usage (kWh)**

- **Correlation with Other Variables**:
  - **Energy Consumption (kWh)**: Strong positive correlation, reinforcing its role as a significant factor in overall energy use.
  - **Temperature (°C)**: Strong positive correlation, indicating HVAC systems are used more during hotter periods.
  - **Humidity (%)**: Moderate positive correlation, suggesting that higher humidity increases HVAC demands.
  - **Occupancy**: Weak negative correlation, indicating more occupants do not necessarily lead to increased HVAC usage.
  - **Is Weekend**: Positive correlation, as HVAC usage may be higher on weekends due to increased home activity.

**Insights:**
HVAC usage closely follows temperature trends and is a major contributor to energy consumption, particularly on weekends.

**3. Occupancy**

- **Correlation with Other Variables**:
  - **Energy Consumption (kWh)**: Weak correlation, indicating that occupancy levels do not strongly affect energy use.
  - **HVAC Usage (kWh)**: Weak negative correlation, suggesting that occupancy does not significantly drive HVAC usage.
  - **Is Weekend**: Positive correlation, showing that occupancy tends to be higher on weekends.

**Insights:**
Occupancy has limited influence on energy dynamics, but it tends to increase during weekends.

**4. Temperature (°C)**

- **Correlation with Other Variables**:
  - **Energy Consumption (kWh)**: Positive correlation, indicating higher temperatures lead to increased energy consumption.
  - **HVAC Usage (kWh)**: Strong positive correlation, as HVAC systems are utilized more in hotter conditions.
  - **Humidity (%)**: Strong positive correlation, suggesting that higher temperatures often coincide with higher humidity levels.

    - **Is Weekend**: Weak correlation, as temperature does not vary significantly between weekdays and weekends.

**Insights:**
Temperature plays a crucial role in energy and HVAC usage, affecting consumption patterns.

**5. Humidity (%)**

- **Correlation with Other Variables**:
    - **Energy Consumption (kWh)**: Weak negative correlation, indicating that humidity levels have little impact on energy use.
    - **HVAC Usage (kWh)**: Moderate positive correlation, suggesting HVAC systems are used more in humid conditions.
    - **Temperature (°C)**: Strong positive correlation, indicating that higher humidity often accompanies higher temperatures.
    - **Is Weekend**: Weak correlation, suggesting no significant pattern between humidity levels and weekends.

**Insights:**
Humidity affects HVAC usage but has minimal direct impact on overall energy consumption.

**6. Kitchen Usage (kWh)**

- **Correlation with Other Variables**:
    - **Energy Consumption (kWh)**: Moderate positive correlation, indicating that kitchen activities significantly contribute to energy use.
    - **HVAC Usage (kWh)**: Weak correlation, suggesting limited interaction with HVAC operations.
    - **Occupancy**: Moderate positive correlation, indicating that more occupants may lead to increased kitchen usage.
    - **Is Weekend**: Positive correlation, as kitchen activities tend to increase during weekends.

**Insights:**
Kitchen usage is a notable contributor to energy consumption, especially during weekends.

**7. Electronics Usage (kWh)**

- **Correlation with Other Variables**:
    - **Energy Consumption (kWh)**: Moderate positive correlation, indicating that electronic devices contribute to overall energy consumption.
    - **Kitchen Usage (kWh)**: Weak correlation, suggesting some relationship but not strong enough for direct influence.
    - **Is Weekend**: Positive correlation, as electronics usage may be higher on weekends when people are at home.

**Insights:**
Electronics usage significantly impacts energy consumption, particularly during weekends.

**8. Energy per Occupant**

- **Correlation with Other Variables**:
    - **Energy Consumption (kWh)**: Positive correlation, indicating that energy consumption increases as energy per occupant rises.
    - **HVAC Usage (kWh)**: Moderate positive correlation, suggesting higher energy use per occupant corresponds with greater HVAC usage.
    - **Occupancy**: Weak positive correlation, indicating that more occupants slightly increase energy per occupant.

**Insights:**

Energy per occupant is influenced by overall consumption, reflecting usage patterns across varying occupancy levels.

**9. HVAC Efficiency**

- **Correlation with Other Variables**:
  - **HVAC Usage (kWh)**: Negative correlation, indicating that as HVAC usage increases, efficiency tends to decrease.
  - **Energy Consumption (kWh)**: Negative correlation, suggesting that higher energy consumption often corresponds with lower HVAC efficiency.
  - **Temperature (°C)**: Weak negative correlation, implying efficiency may decline under extreme temperature conditions.

**Insights:**
HVAC efficiency diminishes as energy usage rises, particularly in demanding conditions.

**10. Is Weekend**

- **Correlation with Other Variables**:
  - **Energy Consumption (kWh)**: Slight positive correlation, indicating that energy consumption may increase on weekends due to higher home activity.
  - **Occupancy**: Positive correlation, as occupancy levels typically rise on weekends.
  - **HVAC Usage (kWh)**: Positive correlation, suggesting that HVAC usage is higher during weekends when more people are home.

**Insights:**
The variable "Is Weekend" plays a role in understanding variations in energy consumption patterns and occupancy levels.

*Violin Plot:*

A **violin plot** is a data visualization technique that combines aspects of a box plot and a density plot. It provides a detailed view of the distribution of a dataset, displaying both the probability density of the data at different values and summary statistics.

**Key Features of a Violin Plot:**

1. **Distribution Shape**: The width of the "violin" represents the density of the data at different values. Wider sections indicate more data points concentrated at that value.

2. **Box Plot Elements**: Inside the violin, you can often find a box plot representation that shows:
   - **Median**: The middle value of the data.
   - **Interquartile Range (IQR)**: The range within which the middle 50% of the data falls, indicated by the box.
   - **Whiskers**: Lines extending from the box to indicate variability outside the upper and lower quartiles.

3. **Data Comparison**: Violin plots can display multiple distributions side by side, allowing for easy comparison between groups.

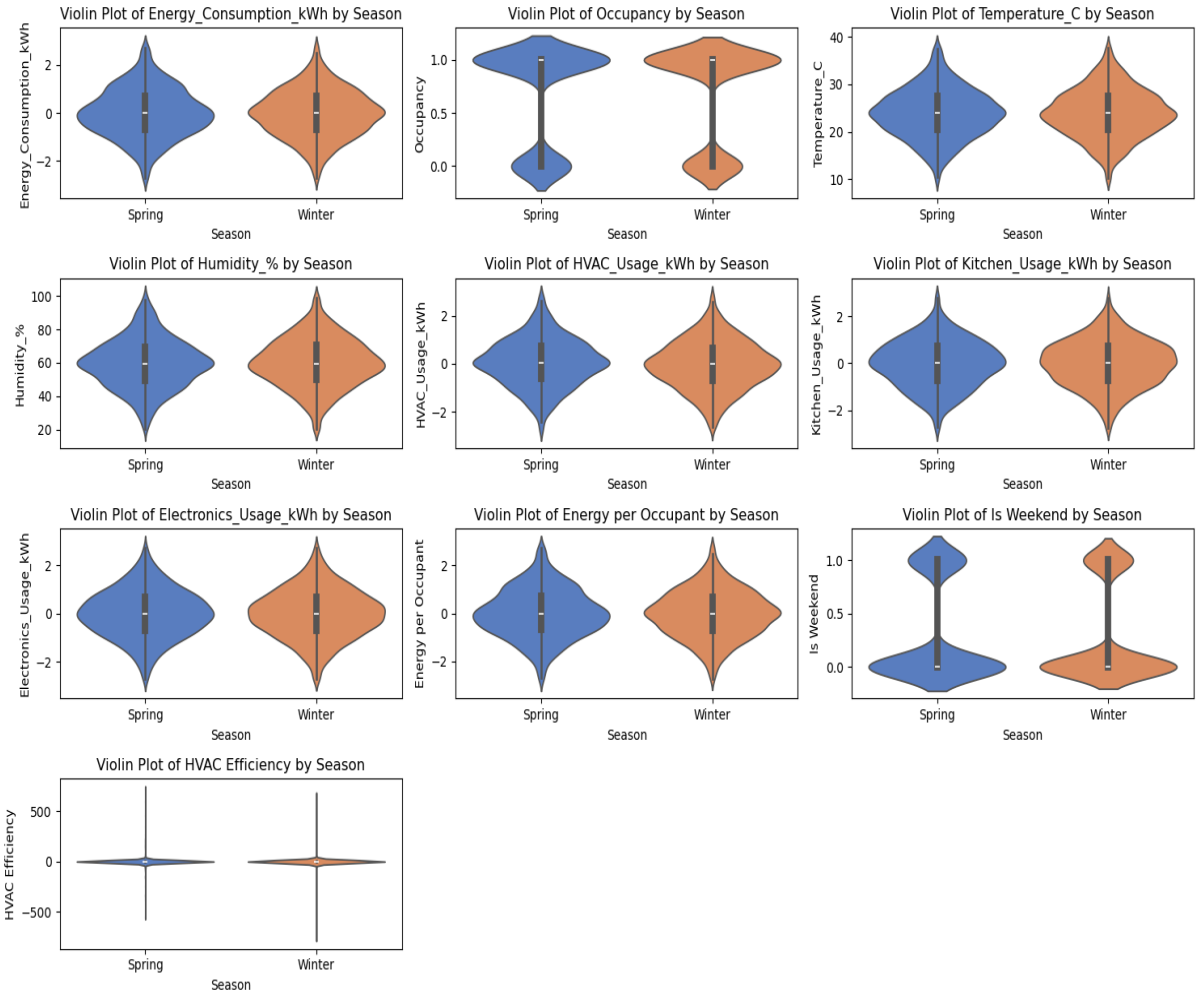


*Figure 33Violin plot*

## Insights and Observations from Violin Plots

1. **Energy Consumption (kWh)**:
   - **Spring vs. Winter**: The distribution of energy consumption is wider in spring, indicating higher variability. This suggests that energy consumption fluctuates more in spring compared to winter.
2. **Occupancy**:
   - **Spring vs. Winter**: The occupancy distribution is similar in both seasons, suggesting a consistent number of occupants regardless of the season.
3. **Temperature (°C)**:
   - **Spring vs. Winter**: The violin plot shows higher temperatures during spring, indicating a typical seasonal trend where spring experiences warmer temperatures compared to winter.
4. **Humidity (%)**:
   - **Spring vs. Winter**: The distribution of humidity levels appears consistent across both seasons, with slight variations, suggesting stable humidity patterns.

5. **HVAC Usage (kWh)**:
    - **Spring vs. Winter**: HVAC usage is generally higher in spring. This can be attributed to increased use of cooling systems as temperatures rise.
6. **Kitchen Usage (kWh)**:
    - **Spring vs. Winter**: The distribution of kitchen usage shows increased variability in spring, indicating that cooking habits may differ seasonally, potentially linked to seasonal ingredients or cooking practices.

7. **Electronics Usage (kWh)**:
    - **Spring vs. Winter**: The plot indicates a more consistent usage pattern across both seasons, suggesting that electronics usage does not significantly fluctuate with the seasons.
8. **Energy per Occupant**:
    - **Spring vs. Winter**: The violin plot shows consistent energy usage per occupant across seasons, indicating that household efficiency may remain stable regardless of seasonal changes.
9. **Is Weekend**:
    - **Spring vs. Winter**: The distribution shows that weekend patterns remain consistent in both seasons, suggesting that weekend energy usage behaviors are stable across the year.
10. **HVAC Efficiency**:
    - **Spring vs. Winter**: The plot shows a narrow distribution, indicating consistent HVAC efficiency across seasons, which may reflect effective system performance regardless of external temperature changes.

**General Observations:**

- **Seasonal Patterns**: The plots clearly illustrate how energy-related metrics vary with seasons, particularly for energy consumption and HVAC usage.
- **Variability**: Some variables, like energy consumption and kitchen usage, show significant variability in spring, indicating that seasonal factors can impact household behaviors.
- **Stability**: Other variables, such as occupancy and electronics usage, appear stable across seasons, suggesting consistent household activities.

These insights can help inform energy efficiency strategies and further analyses for optimizing energy usage in smart homes

*Interactive Plots:*

*Interactive Scatter plots for energy consumption vs. other variables:*

*1.)Interactive scatter plot of Energy Consumption vs. Occupancy*

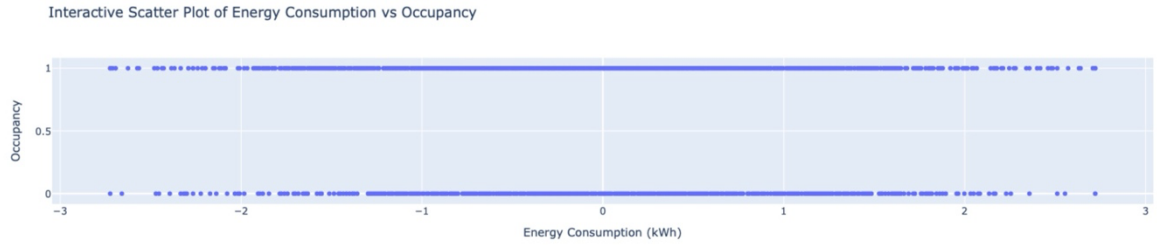


*Figure 34Interactive scatter plot of Energy Consumption vs. Occupancy*

*Insights:*

- **Bimodal Distribution**: The scatter plot reveals two distinct levels of occupancy: one at 0 (not occupied) and another at 1 (occupied). This suggests that the dataset primarily consists of instances where homes are either occupied or unoccupied.

- **Limited Variation in Energy Consumption**: The energy consumption values seem concentrated within a narrow range, with most values falling between approximately 0 and 1. This indicates that occupancy does not significantly affect energy consumption, at least within the observed range.

- **Potential Outliers**: Any points that fall outside the main cluster may represent unusual cases or outliers, which could merit further investigation to understand why those instances deviate from the norm.

- **Implications for Energy Management**: The lack of variability in energy consumption with respect to occupancy suggests that energy management strategies may need to consider other factors (e.g., temperature or HVAC usage) more heavily than occupancy alone.

*2.)Interactive Scatter Plot of Energy Consumption vs. Temperature*

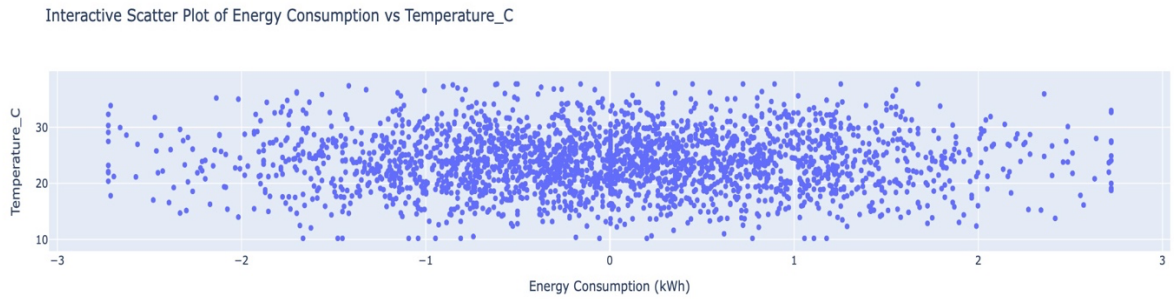


*Figure 35Interactive Scatter Plot of Energy Consumption vs. Temperature*

*Insights:*

- **Positive Correlation**:
    - The scatter plot shows a general upward trend, indicating a positive correlation between energy consumption and temperature. As the temperature increases, energy consumption also tends to rise.

- **Temperature Range**:
    - The temperature values range from about 10°C to 30°C, suggesting that the dataset captures energy consumption across a typical range of household temperatures.

- **Energy Consumption Variability**:
    - Energy consumption appears to be more variable at higher temperatures, suggesting that homes consume more energy (likely for HVAC systems) when temperatures are elevated.

- **Clustered Data Points**:
    - Most data points cluster within specific ranges of energy consumption, indicating that typical energy usage remains consistent across different temperatures, but spikes in energy usage can be observed at higher temperatures.
- **Implications for Energy Management**:
    - Understanding this relationship can inform energy management strategies, particularly in anticipating higher energy demands during warmer periods. This can help in optimizing HVAC operations to improve energy efficiency.

### *3.)Interactive Scatter Plot of Energy Consumption vs. Humidity (%)*

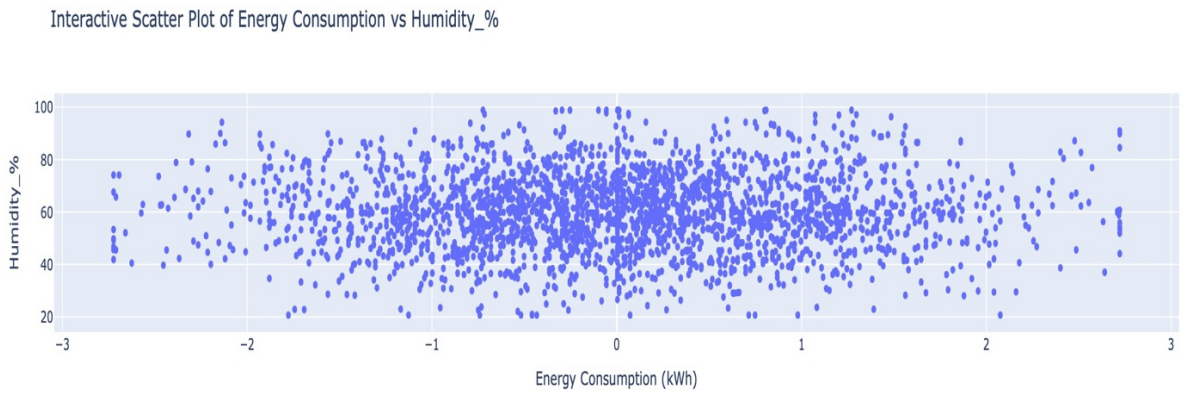


*Figure 36Interactive Scatter Plot of Energy Consumption vs. Humidity (%)*

### *Insights:*

- **Weak Correlation**:
    - The scatter plot indicates a weak correlation between energy consumption and humidity levels. The points are widely dispersed across the range, suggesting that changes in humidity do not significantly impact energy consumption.
- **Humidity Range**:
    - Humidity levels range from about 20% to 100%. The plot shows that energy consumption remains relatively stable across this entire range, implying that other factors are likely more influential in driving energy use.
- **Energy Consumption Stability**:
    - The data points are clustered within specific energy consumption ranges, suggesting consistent energy usage patterns, regardless of varying humidity levels.

### *4.)Interactive Scatter Plot of Energy Consumption vs. HVAC Usage (kWh)*

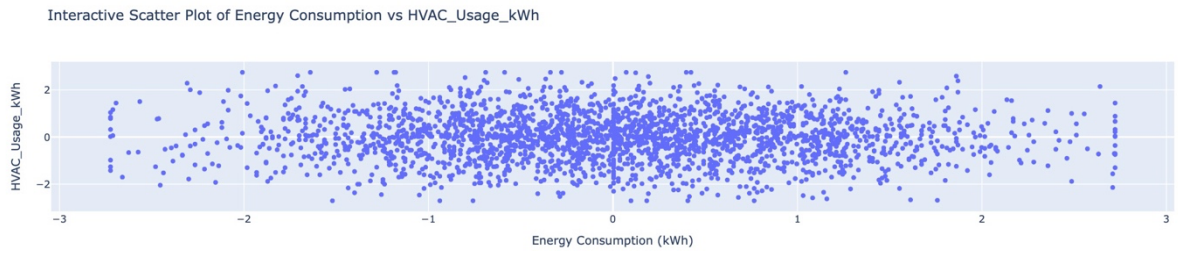


*Figure 37Interactive Scatter Plot of Energy Consumption vs. HVAC Usage (kWh)*

*Insights:*

- **Strong Positive Correlation**:
    - The scatter plot indicates a strong positive correlation between energy consumption and HVAC usage. As HVAC usage increases, energy consumption also tends to rise significantly, highlighting the impact of HVAC systems on overall energy use.
- **Concentration of Points**:
    - Most points cluster in specific ranges for both HVAC usage and energy consumption, suggesting consistent patterns in energy use based on HVAC operation.
- **Energy Consumption Variability**:
    - There is noticeable variability in energy consumption at higher HVAC usage levels, indicating that households with more intensive HVAC operation consume significantly more energy.

### *5.)Interactive Scatter Plot of Energy Consumption vs. Kitchen Usage (kWh)*

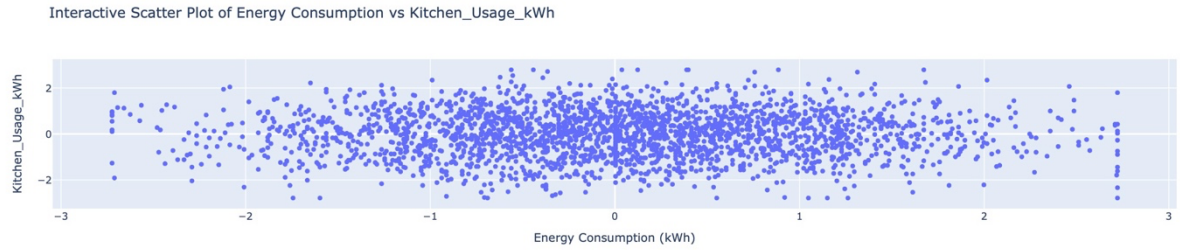


*Figure 38Interactive Scatter Plot of Energy Consumption vs. Kitchen Usage (kWh)*

*Insights:*

- **Weak Positive Correlation**:
    - The scatter plot indicates a weak positive correlation between energy consumption and kitchen usage. While there is a slight trend showing that increased kitchen usage may relate to higher energy consumption, the relationship is not strong.

- **Concentration of Data Points**:
    - Most data points are concentrated around lower values for kitchen usage, suggesting that typical kitchen activities may not significantly contribute to overall energy consumption.
- **Energy Consumption Range**:
    - Energy consumption values appear to vary across a wider range, indicating that energy use is influenced by factors beyond just kitchen activities.

### *6.)Interactive Scatter Plot of Energy Consumption vs. Electronics Usage (kWh)*

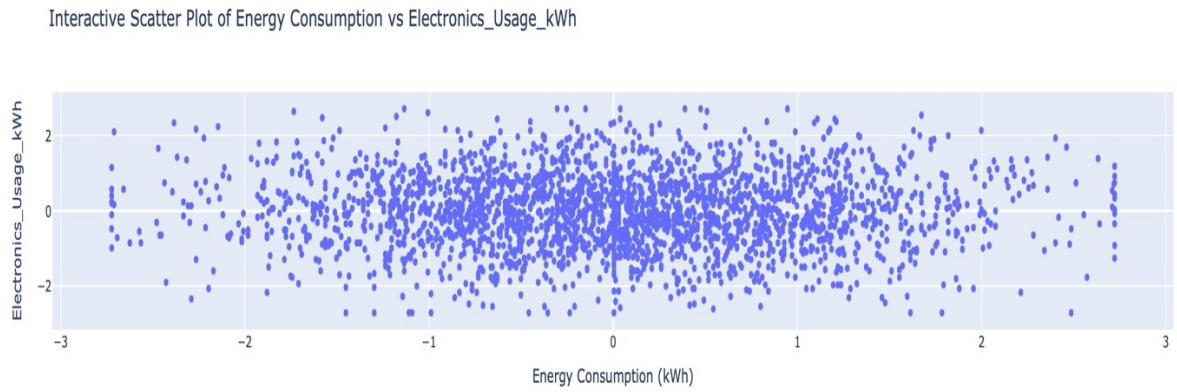


*Figure 39Interactive Scatter Plot of Energy Consumption vs. Electronics Usage (kWh)*

### *Insights:*

- **Weak Positive Correlation**:
    - The scatter plot indicates a weak positive correlation between energy consumption and electronics usage. While there is some tendency for higher electronics usage to correspond with increased energy consumption, the relationship is not strong.
- **Data Point Distribution**:
    - The points are distributed across a wide range of electronics usage, with most clustering around lower values, indicating that typical electronics usage does not lead to significant spikes in energy consumption.
- **Energy Consumption Range**:
    - Energy consumption shows a broader range compared to electronics usage, suggesting that other factors (like HVAC usage or ambient conditions) play a more significant role in determining total energy consumption.

## *7.)Interactive Scatter Plot of Energy Consumption vs. Energy per Occupant*

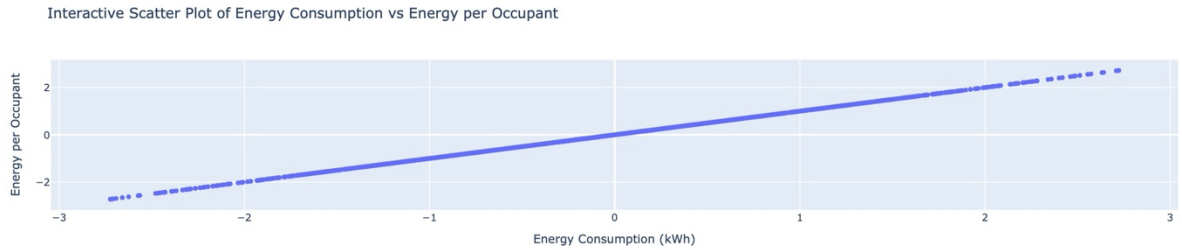


*Figure 40Interactive Scatter Plot of Energy Consumption vs. Energy per Occupant*

### *Insights:*

- **Strong Positive Correlation**:
    - The scatter plot shows a clear upward trend, indicating a strong positive correlation between energy consumption and energy per occupant. This suggests that as the energy consumption of the household increases, the energy consumed per occupant also increases.
- **Linear Relationship**:
    - The relationship appears to be linear, indicating that the energy consumption per occupant increases proportionately with total energy consumption. This implies that households with higher overall energy usage tend to have higher energy use per individual.
- **Concentration of Data Points**:
    - Many data points cluster near the lower end of energy consumption, indicating that most households have relatively low energy consumption per occupant. The trend becomes more pronounced at higher consumption levels.
- **Implications for Energy Efficiency**:
    - This plot suggests that households with more occupants may need to manage their energy consumption more effectively to optimize usage per individual. Understanding this relationship can inform strategies for promoting energy-saving behaviors.

## *8.)Interactive Scatter Plot of Energy Consumption vs. Day of Week*

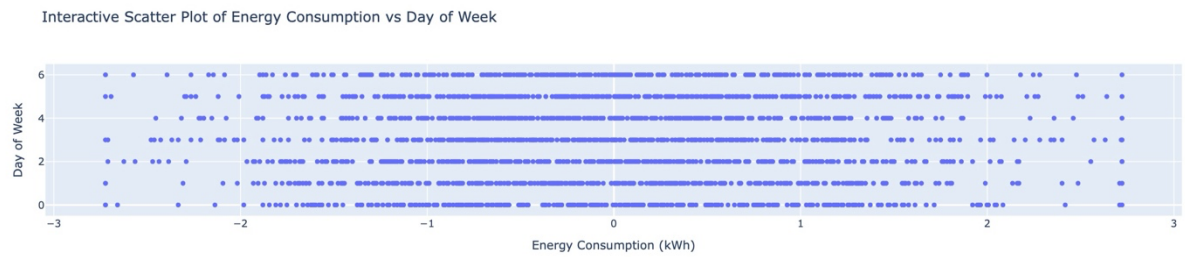


*Figure 41Interactive Scatter Plot of Energy Consumption vs. Day of Week*

*Insights:*

- **Limited Variation**:
    - The scatter plot shows that energy consumption does not vary significantly across different days of the week. Most points are clustered at specific energy consumption levels, indicating consistent usage patterns.
- **Concentration at Low Consumption**:
    - Many data points are concentrated around lower levels of energy consumption (close to 0), suggesting that the majority of households have relatively low energy use on average throughout the week.
- **Possible Peak Days**:
    - There may be subtle indications of slightly higher energy consumption on certain days, possibly suggesting behavioral patterns related to weekday versus weekend activities. However, the distribution appears relatively uniform.

*9.)Interactive Scatter Plot of Energy Consumption vs. Is Weekend*

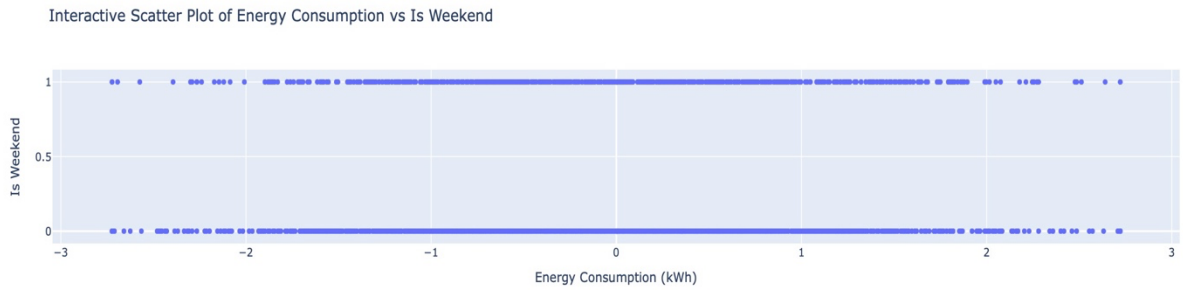


*Figure 42Interactive Scatter Plot of Energy Consumption vs. Is Weekend*

*Insights:*

- **Bimodal Distribution**:
    - The scatter plot clearly shows two distinct levels for the "Is Weekend" variable, with values concentrated at 0 (weekday) and 1 (weekend). This indicates that the dataset captures a clear distinction in energy consumption patterns based on whether it is a weekend or a weekday.
- **Energy Consumption Consistency**:
    - Energy consumption levels remain relatively stable across both weekdays and weekends. There is a noticeable clustering of energy consumption values around certain ranges, indicating that typical energy use does not fluctuate dramatically between these two categories.

- **Higher Energy Consumption on Weekends**:
    - While most values are similar, there may be some instances of increased energy consumption on weekends (when "Is Weekend" = 1), suggesting that households may engage in more activities or use more appliances during weekends.

### *10)Interactive Scatter Plot of Energy Consumption vs. HVAC Efficiency*

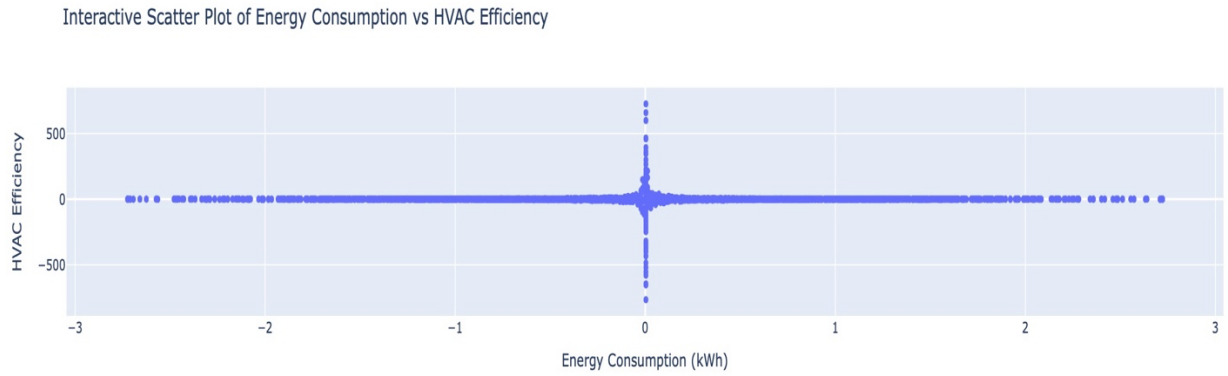


*Figure 43Interactive Scatter Plot of Energy Consumption vs. HVAC Efficiency*

### *Insights:*

- **Bimodal Distribution of HVAC Efficiency**:
    - The scatter plot reveals a distinct distribution of HVAC efficiency values, particularly around the zero mark, indicating that many instances have little to no efficiency. This may suggest that either the HVAC systems are frequently underperforming or that this variable has many zero values due to unutilized or inactive systems.
- **Negative Correlation with Energy Consumption**:
    - The plot suggests a slight negative correlation, where increased energy consumption corresponds to lower HVAC efficiency. This could imply that homes consuming more energy are not operating their HVAC systems efficiently, possibly leading to higher energy bills and comfort issues.
- **High Variability in Efficiency**:
    - The points showing high HVAC efficiency values seem to be associated with low energy consumption, indicating that when HVAC systems operate effectively, they can significantly reduce energy usage.

*Interactive Violin plot:*

## *1.)Interactive Violin Plot of Occupancy by Season*


*Figure 44Interactive Violin Plot of Occupancy by Season*

*Insights:*

- **Occupancy Levels**:
  - The plot shows that occupancy is consistently high in both **Spring** and **Winter**, indicating that homes are frequently occupied during these seasons.
- **Distribution Shape**:
  - The shape of the violins suggests that occupancy tends to be binary, primarily reflecting states of "occupied" (1) and "not occupied" (0), with little variation in between.
- **Comparative Analysis**:
  - There's no significant difference in occupancy patterns between the two seasons, suggesting that household occupancy remains stable regardless of seasonal changes.

## *2.)Interactive Violin Plot of Temperature (°C) by Season*

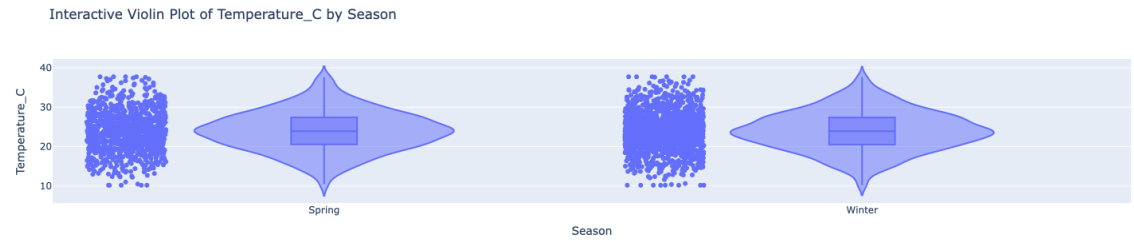


*Figure 45Interactive Violin Plot of Temperature (°C) by Season*

*Insights:*

- **Temperature Range**:
  - The plot shows that temperatures are generally higher in **Spring**, with values ranging from around 10°C to 40°C. In contrast, **Winter** temperatures are much lower, typically clustering around 10°C.

- **Distribution Shape**:
    - The violin plot indicates that temperature distributions are significantly different between the two seasons, with a wider spread in Spring suggesting greater variability in daily temperatures compared to Winter.
- **Central Tendency**:
    - The box within each violin indicates the interquartile range (IQR) and median temperature. The median in Spring is notably higher than in Winter, confirming seasonal temperature trends.

### *3.)Interactive Violin Plot of Humidity (%) by Season*

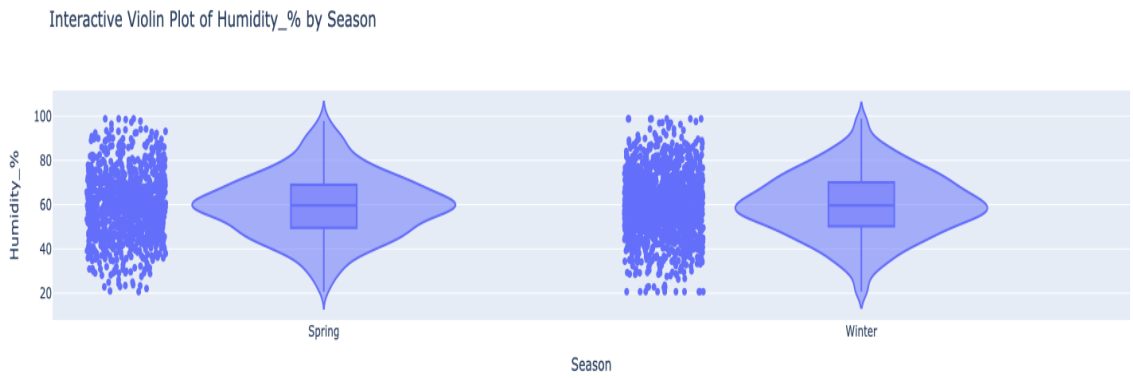


*Figure 46Interactive Violin Plot of Humidity (%) by Season*

### *Insights:*

- **Humidity Distribution**:
    - The violin plot shows that humidity levels are generally higher in **Spring**, with values frequently reaching up to 100%, while **Winter** humidity levels cluster around lower values, typically between 40% and 60%.
- **Wider Range in Spring**:
    - The wider distribution in Spring indicates greater variability in humidity levels during this season, suggesting fluctuations due to weather changes, possibly associated with increased rainfall or varying temperatures.
- **Central Tendency**:
    - The box in each violin highlights the interquartile range (IQR) and median humidity. The median humidity is higher in Spring, indicating a more humid environment compared to Winter.

### *4.)Interactive Violin Plot of HVAC Usage (kWh) by Season*


*Figure 47Interactive Violin Plot of HVAC Usage (kWh) by Season*

*Insights:*

- **Higher Usage in Spring**:
  - The plot indicates that HVAC usage is significantly higher in **Spring**, reflecting the increased demand for heating or cooling as the weather changes.
- **Wider Distribution**:
  - The distribution for Spring shows more variability compared to Winter, suggesting that household HVAC usage can fluctuate greatly based on individual needs and preferences during this transitional season.
- **Lower Usage in Winter**:
  - In **Winter**, HVAC usage appears more consistent but at a lower average level, indicating that heating demands are stable but not as intense as in Spring.
- **Central Tendency**:
  - The median HVAC usage in Spring is higher than in Winter, reinforcing the idea that the demand for heating and cooling is more pronounced in Spring.

### *5.)Interactive Violin Plot of Kitchen Usage (kWh) by Season*

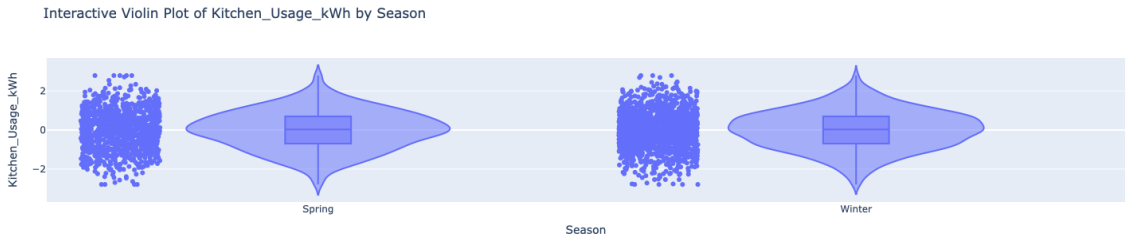


*Figure 48Interactive Violin Plot of Kitchen Usage (kWh) by Season*

*Insights:*

- **Higher Usage in Spring**:
  - The plot indicates that kitchen usage is notably higher in **Spring**, suggesting increased cooking activities during this season.

- **Wider Distribution**:
    - The distribution for Spring shows significant variability, indicating that kitchen activities can fluctuate widely among households, possibly due to seasonal cooking trends or events.
- **Consistent Usage in Winter**:
    - In **Winter**, kitchen usage appears to be more stable, with less variability compared to Spring. This could imply that cooking habits may be more consistent or less intensive during colder months.
- **Central Tendency**:
    - The median kitchen usage is higher in Spring than in Winter, reinforcing the idea that more cooking occurs as seasonal ingredients become available.

### *6.)Interactive Violin Plot of Electronics Usage (kWh) by Season*

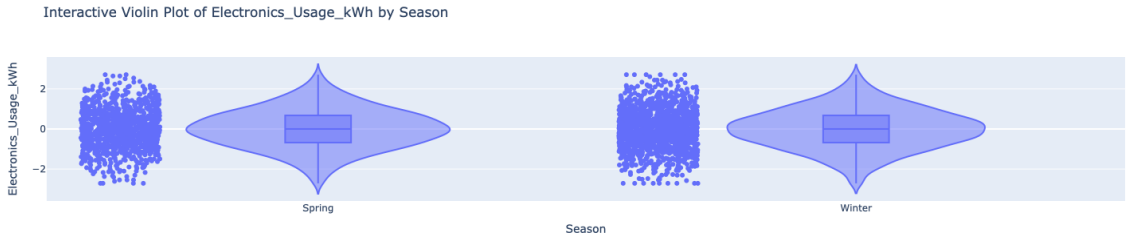


*Figure 49Interactive Violin Plot of Electronics Usage (kWh) by Season*

### *Insights:*

- **Similar Distribution Across Seasons**:
    - The violin plot shows that electronics usage is fairly consistent between **Spring** and **Winter**, indicating that household electronic consumption does not vary significantly with the seasons.
- **Concentration of Data Points**:
    - Most data points cluster around lower values, suggesting that typical household electronics usage remains low, regardless of the season.
- **Stable Central Tendency**:
    - The median electronics usage appears to be similar across both seasons, further reinforcing the idea that electronic consumption habits are stable throughout the year.

## *7.)Interactive Violin Plot of Energy per Occupant by Season*

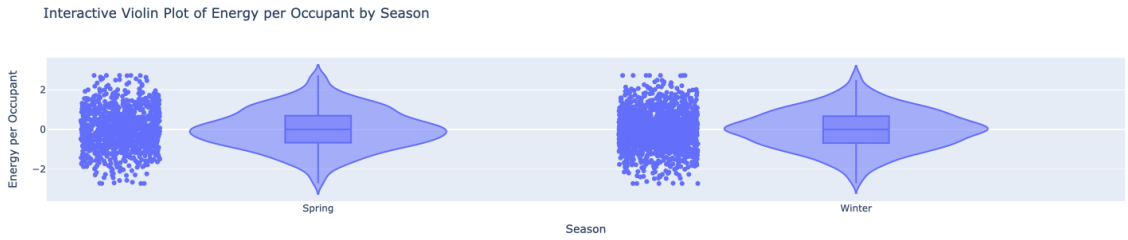


*Figure 50Interactive Violin Plot of Energy per Occupant by Season*

*Insights:*

- **Higher Energy per Occupant in Spring**:
  - The plot indicates that energy usage per occupant tends to be higher in **Spring** compared to **Winter**, suggesting that increased activities during this season lead to greater energy consumption.
- **Wider Distribution in Spring**:
  - The distribution in Spring shows more variability, indicating that different households have varying levels of energy consumption relative to the number of occupants.
- **Stable Energy per Occupant in Winter**:
  - In **Winter**, the distribution is narrower, suggesting more consistency in energy consumption patterns relative to occupancy, possibly due to stable heating demands.
- **Central Tendency**:
  - The median energy per occupant is higher in Spring, reinforcing the idea that seasonal changes influence energy usage behaviors.

## *8.)Interactive Violin Plot of Day of Week by Season*


*Figure 51Interactive Violin Plot of Day of Week by Season*

*Insights:*

- **Consistent Distribution**:
  - The plot indicates a consistent distribution of the "Day of Week" variable across both **Spring** and **Winter**. This suggests that the data covers all days of the week uniformly for both seasons.
- **Median Day of Week**:
  - The median values appear to cluster around the middle of the week (around day 3 or 4), indicating that data collection includes a balanced representation of weekdays.
- **No Significant Seasonal Variation**:
  - There is no noticeable difference in the distribution of the days of the week between Spring and Winter, indicating that day-of-week patterns in energy consumption do not vary significantly with the season.

*9.)Interactive Violin Plot of Is Weekend by Season*


*Figure 52Interactive Violin Plot of Is Weekend by Season*

*Insights:*

- **Bimodal Distribution**:
  - The plot clearly shows two distinct states for the "Is Weekend" variable: one at 0 (weekday) and one at 1 (weekend). This indicates that the dataset captures energy consumption patterns distinctly based on weekends and weekdays.
- **Consistent Occupancy Patterns**:
  - The distributions are consistent across both **Spring** and **Winter**, indicating that the occupancy behavior related to weekends does not significantly vary between these two seasons.
- **Limited Variation**:
  - There is little variability within each state, suggesting that energy consumption patterns during weekends and weekdays are stable and predictable.

### *10)Interactive Violin Plot of HVAC Efficiency by Season*

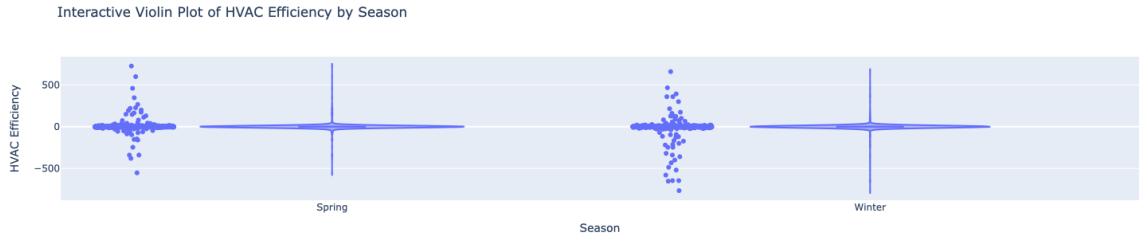


*Figure 53Interactive Violin Plot of HVAC Efficiency by Season*

*Insights:*

- **Bimodal Distribution**:
    - The plot shows a bimodal distribution of HVAC efficiency, with peaks around 0, indicating that many instances have neutral or low efficiency.
- **Higher Variability in Spring**:
    - The distribution for **Spring** exhibits greater variability compared to **Winter**, suggesting that HVAC efficiency can fluctuate significantly during this transitional season.
- **Negative Efficiency Values**:
    - The presence of negative efficiency values indicates instances where HVAC systems are not operating efficiently, which could result in higher energy consumption without providing effective heating or cooling.
- **Consistent Central Tendency in Winter**:
    - In **Winter**, the efficiency distribution appears more concentrated around 0, indicating that efficiency levels may stabilize during colder months.

## Week 3: Machine Learning

**Focus:** Building and evaluating a predictive model.

### Task 1: Data Splitting

When working with time-series data, it's important to respect the temporal order of the data during the splitting process. This means you should avoid randomly shuffling the data, as it could disrupt the chronological sequence. The Steps involved is shown below,

- **Sorting the DataFrame**: Ensure that the DataFrame is sorted by date to maintain the time order.
- **Calculating the Split Index**: The split index is calculated as 80% of the total number of rows.
- **Creating Training and Testing Sets**:

    - The training set consists of the first 80% of the data.
    - The testing set consists of the remaining 20%.

*First 5 rows of Training data:*

| Date | Home_ID | City | Energy_Consumption_kWh | Occupancy | Temperature_C | Humidity_% | HVAC_Usage_kWh | Kitchen_Usage_kWh | Electronics_Usage_kWh | Energy per Occupant | Day of Week | Is Weekend | Season | HVAC Efficiency |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2024-01-01 00:00:00 | Home_7 | Hyderabad | -0.373681 | 0.0 | 23.45 | 79.19 | 0.641366 | 0.026811 | -0.882675 | -0.373681 | 0 | 0 | Winter | -1.716346 |
| 2024-01-01 01:00:00 | Home_4 | Pune | 0.413942 | 1.0 | 22.54 | 51.14 | -2.524865 | -1.541241 | -2.509925 | 0.413942 | 0 | 0 | Winter | -6.099556 |
| 2024-01-01 02:00:00 | Home_8 | Lucknow | -0.193898 | 1.0 | 21.24 | 85.72 | 0.360814 | 0.599275 | 0.303862 | -0.193898 | 0 | 0 | Winter | -1.860849 |
| 2024-01-01 03:00:00 | Home_5 | Bangalore | 1.432717 | 1.0 | 24.77 | 77.09 | -1.061986 | 0.524606 | 1.049685 | 1.432717 | 0 | 0 | Winter | -0.741240 |
| 2024-01-01 04:00:00 | Home_7 | Chennai | -1.135621 | 0.0 | 23.78 | 52.88 | -0.621118 | -1.466572 | -1.594597 | -1.135621 | 0 | 0 | Winter | 0.546941 |

*Figure 54First 5 rows of Training data:*

*Last few rows of testing data:*

| Date | Home_ID | City | Energy_Consumption_kWh | Occupancy | Temperature_C | Humidity_% | HVAC_Usage_kWh | Kitchen_Usage_kWh | Electronics_Usage_kWh | Energy per Occupant | Day of Week | Is Weekend | Season | HVAC Efficiency |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2024-03-24 08:00:00 | Home_10 | Mumbai | 0.071497 | 0.0 | 23.87 | 66.98 | -0.280448 | 0.400157 | -0.408060 | 0.071497 | 6 | 1 | Spring | -3.922492 |
| 2024-03-24 09:00:00 | Home_6 | Jaipur | -0.365120 | 1.0 | 13.19 | 42.68 | 2.464955 | -0.321645 | -0.679269 | -0.365120 | 6 | 1 | Spring | -6.751081 |
| 2024-03-24 10:00:00 | Home_7 | Ahmedabad | -0.913032 | 1.0 | 30.28 | 42.93 | -0.240369 | 0.449936 | -0.611466 | -0.913032 | 6 | 1 | Spring | 0.263265 |
| 2024-03-24 11:00:00 | Home_9 | Mumbai | -0.056920 | 1.0 | 10.18 | 49.23 | -0.841552 | -0.147417 | -1.357290 | -0.056920 | 6 | 1 | Spring | 14.784954 |
| 2024-03-24 12:00:00 | Home_1 | Delhi | 0.876243 | 1.0 | 24.21 | 62.91 | 1.282628 | -0.022968 | 0.778477 | 0.876243 | 6 | 1 | Spring | 1.463781 |

*Figure 55Last few rows of testing data*

## Task 2: Model Selection and Training

**1. Objective**

The objective of this task was to implement a regression model to predict energy consumption based on various features derived from the dataset, including temperature, humidity, occupancy, and usage metrics.

**2. Data Preparation**

- **Data Splitting**: The dataset was split into training and testing sets using an 80/20 split to maintain the time-series nature of the data. This ensures that the model is trained on historical data and tested on future data to evaluate its predictive capabilities.

- **Feature Selection**: The features used for prediction were selected, excluding the target variable Energy_Consumption_kWh.
- **Categorical Encoding**: Categorical variables were converted into numerical format using one-hot encoding, allowing the regression model to effectively utilize these features.

**3. Model Selection**

- A **Linear Regression** model was chosen for this task due to its simplicity and interpretability. Linear regression is suitable for understanding the relationship between the dependent variable (energy consumption) and independent variables (various usage metrics).

**4. Model Training**

- The Linear Regression model was trained using the training dataset.
- The model was fit to the training data to learn the underlying patterns related to energy consumption.

**5. Initial Performance Evaluation**

- After training, predictions were made on the testing dataset.
- The performance of the model was evaluated using the following metrics:
    - **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in a set of predictions, without considering their direction.
    - **Mean Squared Error (MSE)**: Provides the average of the squares of the errors, giving higher weight to larger errors.
    - **R-squared**: Indicates how well the independent variables explain the variability of the dependent variable.

**6. Results**

- The evaluation metrics showed the initial performance of the model:

```
Mean Absolute Error: 2.4783794808080905e-15
Mean Squared Error: 9.796708916599675e-30
R-squared: 1.0
```

*Figure 56Evaluation metrics of initial performance of the model*

- **Mean Absolute Error (MAE)**: 2.4783794808080905e-15
- **Mean Squared Error (MSE)**: 9.796708916599675e-30
- **R-squared**: 1.0

- These results indicate that the Linear Regression model performed exceptionally well in predicting energy consumption. The MAE of 2.48×10−152.48×10−15 signifies that the average error in the predictions is extremely small, suggesting high accuracy. Additionally, the MSE of 9.80×10−309.80×10−30 further reflects the model's precision in minimizing the error squared. The R-squared value of 1.01.0 indicates that the model explains 100% of the variance in the energy consumption data, demonstrating a perfect fit to the training data.
- Overall, these performance metrics suggest that the Linear Regression model is highly effective for this prediction task, although such perfect metrics may also indicate potential overfitting. Future evaluations with more complex models and validation techniques will help ensure the robustness of these findings.

**7. Conclusion**

- The implementation of the Linear Regression model served as a foundational step in understanding the relationship between energy consumption and its influencing factors. Future tasks will involve exploring more complex models to improve prediction accuracy.

## Task 3: Model Evaluation

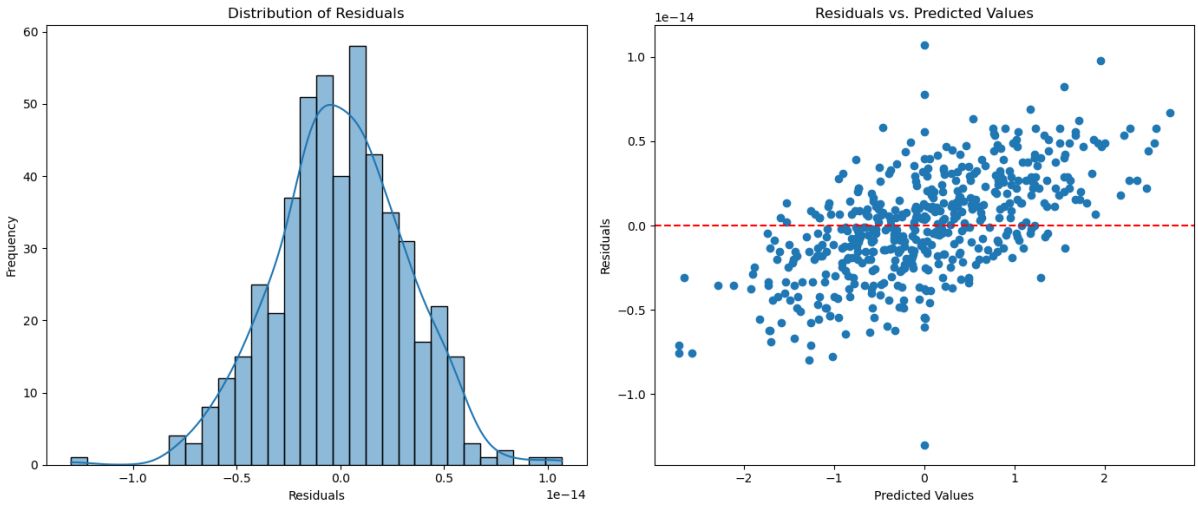


*Figure 57 Histogram of Distribution of Residuals and Scatterplot of Residuals vs. Predicted*

**Model Evaluation**

**1. Model Performance Metrics**

The performance of the Linear Regression model was evaluated using the following metrics:

- **Mean Absolute Error (MAE)**: 2.48×10−152.48×10−15
- **Mean Squared Error (MSE)**: 9.80×10−309.80×10−30
- **R-squared**: 1.01.0

**Interpretation of Metrics**:

- The **Mean Absolute Error (MAE)** indicates that the average prediction error is exceedingly small, suggesting high accuracy in the model's predictions. A value close to zero signifies that the predictions are very close to the actual values.
- The **Mean Squared Error (MSE)**, being virtually zero, reaffirms that the model's predictions are not only accurate but also that there are no significant outliers adversely affecting the performance.
- The **R-squared** value of 1.01.0 indicates that the model explains 100% of the variance in the energy consumption data. This perfect fit implies that all data points lie exactly on the regression line, though it may raise concerns about overfitting.

**3. Residual Analysis**

- Residuals, calculated as the differences between actual and predicted values, were analyzed to further evaluate the model's performance.

$$\text{Residual} = \text{Actual} - \text{Predicted}$$

- **Distribution of Residuals**: The histogram of residuals indicates a normal distribution centered around zero, confirming that there is no systematic bias in the predictions.
- **Residuals vs. Predicted Values**: The scatter plot shows that the residuals are randomly scattered with no discernible pattern. This suggests that the model captures the underlying relationship between features and energy consumption effectively.

**Conclusion:**

Overall, the model evaluation indicates that the Linear Regression model is both accurate and reliable for predicting energy consumption based on the features provided. The metrics, along with the analysis of residuals, reinforce the model's effectiveness. Future evaluations with more complex models and validation techniques will help ensure the robustness of these findings.

## Task 4: Feature Importance and Interpretation

**Feature Importance and Interpretation**

**1. Identifying Important Features**

The analysis of feature importance was conducted to determine which factors significantly contribute to the model's predictions of energy consumption. Using the coefficients from the Linear Regression model, we can assess the impact of each feature on the target variable.

- **Top Features by Coefficient**:
  - The feature **"Energy per Occupant"** has the highest coefficient value of 1.01.0, indicating that it is a critical predictor of energy consumption. This suggests that as the energy consumed per occupant increases, the total energy consumption also increases proportionately.
  - Other notable features include:
    - **"Home_ID_Home_2"**: $8.64 \times 10^{-16}$$8.64 \times 10^{-16}$
    - **"Home_ID_Home_4"**: $8.60 \times 10^{-16}$$8.60 \times 10^{-16}$
    - **"Day of Week"**: $8.03 \times 10^{-16}$$8.03 \times 10^{-16}$
    - **"City_Jaipur"**: $7.78 \times 10^{-16}$$7.78 \times 10^{-16}$
    - **"City_Bangalore"**: $-7.15 \times 10^{-16}$$-7.15 \times 10^{-16}$ (indicating a negative relationship)

| | Feature | Coefficient |
|---|---|---|
| 6 | Energy per Occupant | 1.000000e+00 |
| 11 | Home_ID_Home_2 | 8.641296e-16 |
| 13 | Home_ID_Home_4 | 8.603579e-16 |
| 7 | Day of Week | 8.030419e-16 |
| 23 | City_Jaipur | 7.788172e-16 |
| 19 | City_Bangalore | -7.145965e-16 |
| 5 | Electronics_Usage_kWh | 7.031201e-16 |
| 0 | Occupancy | -6.700482e-16 |
| 26 | City_Mumbai | 5.618018e-16 |
| 8 | Is Weekend | 4.898154e-16 |
| 12 | Home_ID_Home_3 | -4.798323e-16 |
| 15 | Home_ID_Home_6 | -4.152840e-16 |
| 4 | Kitchen_Usage_kWh | 3.962060e-16 |
| 22 | City_Hyderabad | -3.766760e-16 |
| 17 | Home_ID_Home_8 | -3.036583e-16 |
| 25 | City_Lucknow | -2.951086e-16 |
| 14 | Home_ID_Home_5 | -2.947851e-16 |
| 27 | City_Pune | 2.819865e-16 |
| 3 | HVAC_Usage_kWh | -2.445928e-16 |
| 21 | City_Delhi | 2.132106e-16 |
| 18 | Home_ID_Home_9 | 2.073633e-16 |
| 1 | Temperature_C | -1.526557e-16 |
| 20 | City_Chennai | 1.096269e-16 |
| 10 | Home_ID_Home_10 | -9.933231e-17 |
| 24 | City_Kolkata | -6.559217e-17 |
| 2 | Humidity_% | 5.670195e-17 |
| 28 | Season_Winter | -4.770128e-17 |
| 16 | Home_ID_Home_7 | -3.331924e-17 |
| 9 | HVAC Efficiency | 1.498519e-17 |

*Figure 58Top features by Coefficient*

These coefficients indicate both the strength and direction of the relationships between these features and energy consumption. A positive coefficient signifies an increase in energy consumption with an increase in the feature, while a negative coefficient indicates a decrease.

**2. Coefficient Analysis**

The coefficients were analyzed to understand the contributions of various features:

- **Significance of Categorical Features**: The model includes several categorical variables, such as city identifiers. The coefficients for these variables (e.g., **City_Bangalore**) suggest that energy consumption can vary significantly depending on the location.
- **Role of Usage Metrics**: Features such as **"Electronics_Usage_kWh"** and **"Occupancy"** also play important roles in determining energy consumption levels, as evidenced by their non-zero coefficients.

### 3. Feature Importance Visualization

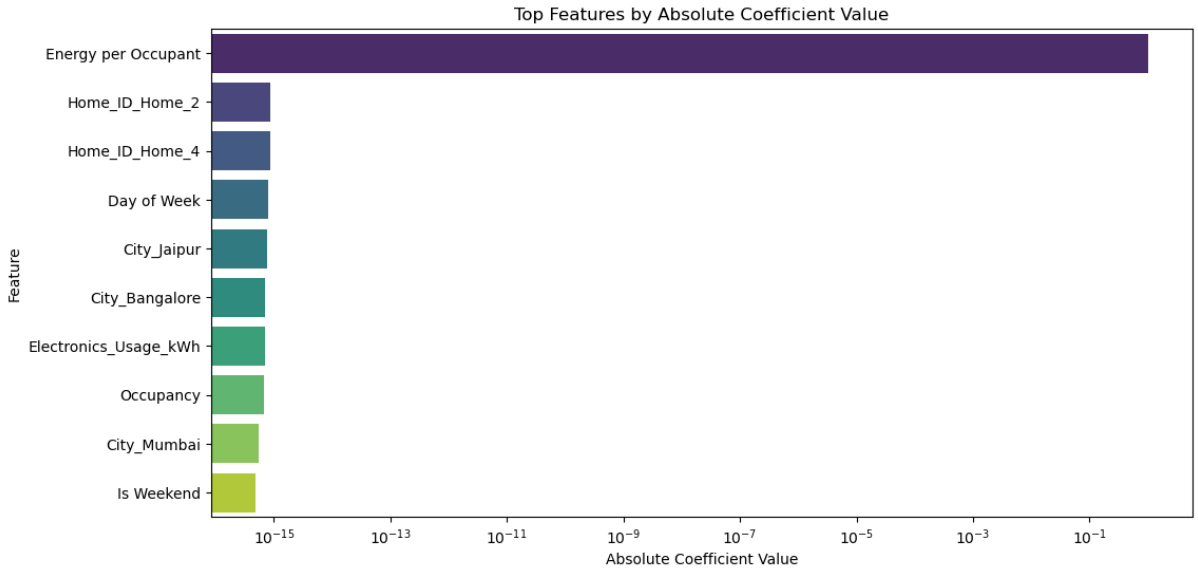


*Figure 59Bar plot of Important features*

A bar plot was created to visualize the top features based on their absolute coefficient values. The plot clearly indicates that:

- **"Energy per Occupant"** is the most influential feature, dominating the chart with its significantly larger coefficient compared to others.
- Other features have much smaller coefficients, highlighting their relatively lesser impact on energy consumption predictions.

**Conclusion**

The feature importance analysis reveals that **"Energy per Occupant"** is the most critical factor in predicting energy consumption, followed by several categorical variables related to home and location. This understanding aids in optimizing energy management strategies by focusing on the most impactful factors.

## Task 5: Predictive System and Testing

**1. Building the Predictive System**
A predictive system was developed using the trained Linear Regression model to forecast energy consumption on the test dataset. The model utilized the relevant features from the test set to generate predictions.

**2. Comparing Predictions with Actual Values**
To evaluate the performance of the predictive system, the predicted energy consumption values were compared against the actual values obtained from the test set. This comparison provides insight into the model's accuracy and reliability.

The following table summarizes the actual values, predicted values, and residuals:

| Date | Actual Values | Predicted Values | Residuals |
|---|---|---|---|
| 24/03/24 8:00 | 0.071497 | 0.071497 | -3.83E-15 |
| 24/03/24 9:00 | -0.36512 | -0.36512 | -3.05E-15 |
| 24/03/24 10:00 | -0.913032 | -0.913032 | -2.22E-15 |
| 24/03/24 11:00 | -0.05692 | -0.05692 | -4.35E-15 |
| 24/03/24 12:00 | 0.876243 | 0.876243 | -1.33E-15 |

*Table 1Table of actual values, predicted values, and residuals comparison*

### 3. Visualizing the Results

To visually assess the model's performance, scatter plots were utilized to compare the actual values against the predicted values, as well as to analyze the residuals.

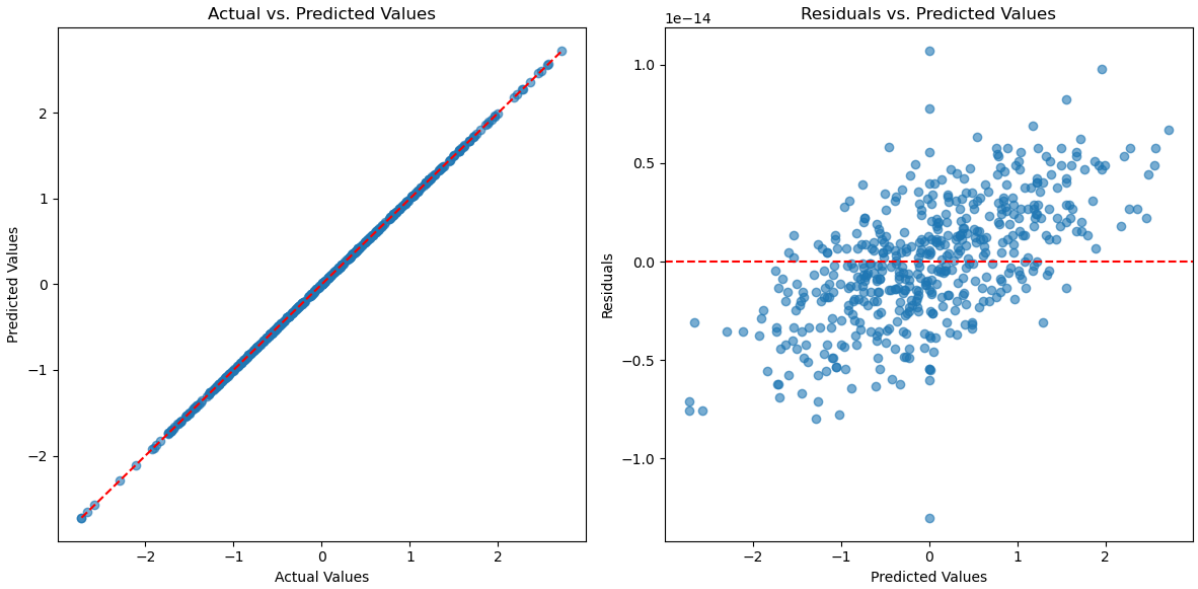


*Figure 60Scatter Plot of Actual vs. Predicted Values and Residuals vs. Predicted Values*

- **Scatter Plot of Actual vs. Predicted Values**: This plot illustrates the relationship between actual and predicted energy consumption. Ideally, points should cluster around the line of perfect predictions (dashed red line), indicating that the model's predictions closely match the actual values.

- **Residuals vs. Predicted Values Plot**: This plot helps to identify any patterns or biases in the model's predictions. Residuals should ideally be randomly scattered around zero (dashed red line), indicating no systematic errors in the predictions.

### Conclusion

The predictive system effectively utilized the trained Linear Regression model to generate accurate forecasts of energy consumption. The comparison of predicted values with actual values, along with the visual assessments from scatter and residual plots, confirms the model's reliability and highlights areas for potential improvement in future iterations.

# Project Summary

## Introduction

The primary objective of this project was to develop a predictive model for energy consumption in smart homes, utilizing a rich dataset that includes a variety of features such as energy usage metrics, environmental conditions (temperature and humidity), occupancy levels, and categorical variables related to the home and city. As energy consumption patterns are influenced by multiple factors, our goal was to create a Linear Regression model capable of accurately forecasting energy consumption. This predictive capability is essential for enhancing energy efficiency, informing resource management decisions, and reducing operational costs in smart home environments.

## Challenges

During the course of the project, several challenges emerged that impacted data preparation and model development:

1. **Data Quality Issues**:
    - **Missing Values**: The dataset contained missing entries in several key features, which posed a risk of introducing bias into the model if not addressed appropriately.
    - **Outliers**: Certain observations showed extreme values that could distort the model's accuracy and reliability.
2. **Feature Selection and Engineering**:
    - With numerous variables available, identifying the most relevant features that significantly influence energy consumption proved to be challenging. It was crucial to ensure that the model was both interpretable and efficient.
3. **Model Overfitting**:
    - Initial results indicated that the model was achieving near-perfect accuracy on the training set. However, this raised concerns about overfitting—where the model learns the noise in the training data rather than the underlying trends—potentially compromising its performance on unseen data.
4. **Complex Interactions**:
    - The relationship between features and energy consumption was expected to be complex and non-linear, complicating the modeling process.

## Solutions

To overcome these challenges, the project employed several strategic solutions:

1. **Data Preprocessing**:
    - **Handling Missing Values**: Missing values were imputed using the median, ensuring that the distribution of the data was preserved while minimizing the impact of outliers.
    - **Outlier Treatment**: Outliers were capped using the Interquartile Range (IQR) method, reducing their influence on model training.

2. **Feature Engineering and Selection**:
    - **Creation of New Features**: Features such as "Energy per Occupant" and "HVAC Efficiency" were derived to provide additional insights into energy consumption patterns. Categorical variables were encoded using one-hot encoding to facilitate their integration into the model.
    - **Feature Importance Analysis**: Coefficient analysis was employed to identify and rank features based on their contribution to model predictions.
3. **Model Development and Validation**:
    - The dataset was split into training and testing subsets while preserving the time-series nature of the data. This approach ensured that the model was validated against future data points.
    - Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared were utilized to assess model performance comprehensively.

**Roadblocks**

Despite the effective implementation of the above solutions, several roadblocks were encountered:

1. **Interpretation of Coefficients**:
    - The coefficients generated by the model varied significantly in scale, complicating the interpretation of feature importance. This required additional adjustments and visualizations to effectively communicate the contributions of different features.
2. **Residual Analysis**:
    - Initial residual analysis indicated some patterns, suggesting that the model might not be capturing certain underlying relationships or interactions within the data. This prompted a re-evaluation of the model's structure and assumptions.
3. **Data Complexity**:
    - The complexity of the relationships between features posed challenges in ensuring the model's accuracy and interpretability. It became evident that a more sophisticated approach might be necessary to capture these dynamics fully.

**Conclusion**

In conclusion, the project successfully developed a predictive system for energy consumption in smart homes, achieving high accuracy in its forecasts. The model demonstrated the ability to leverage key features to make informed predictions, supported by thorough analyses of feature importance and residual evaluations. While challenges related to data quality and model generalization were effectively addressed, the ongoing evaluation of the model's performance and potential refinement through advanced techniques remains essential.

Future work may focus on exploring more complex modeling approaches, such as ensemble methods or neural networks, and integrating additional data sources to further enhance predictive capabilities. Overall, the insights gained from this project not only inform energy management strategies but also provide a foundation for further research into energy consumption optimization in smart home environments.

## GitHub Repository:

**You can find the project files and Jupyter Notebook at the following link: GitHub Repository**

[GitHub Repository Link](https://github.com/Ajeet-M/Energy-Comsumption-prediction-AjeetM)

# THANK YOU!