Appliances Energy Consumption Prediction

**Problem statement:**

This project focuses on developing a machine learning model to predict **household appliance energy consumption** based on a combination of **indoor environmental conditions** (such as temperature and humidity across various rooms) and **external weather data** (including wind speed, visibility, and dew point temperature).

**Dataset:**

The dataset includes time-stamped measurements collected over several weeks, offering a robust foundation for time series analysis, regression modeling, and feature engineering.

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| recorded\_timestamp | The exact date and time when the measurement was taken. |
| total\_energy\_use | Total electrical energy consumed by household appliances (in watt-hours). |
| lighting\_energy\_use | Amount of energy consumed specifically by lighting systems. |
| temp\_sensor\_1 | Temperature reading from sensor 1 in the building. |
| humidity\_sensor\_1 | Relative humidity percentage recorded by sensor 1. |
| temp\_sensor\_2 | Temperature reading from sensor 2 in the building. |
| humidity\_sensor\_2 | Relative humidity percentage recorded by sensor 2. |
| temp\_sensor\_3 | Temperature reading from sensor 3 in the building. |
| humidity\_sensor\_3 | Relative humidity percentage recorded by sensor 3. |
| temp\_sensor\_4 | Temperature reading from sensor 4 in the building. |
| humidity\_sensor\_4 | Relative humidity percentage recorded by sensor 4. |
| temp\_sensor\_5 | Temperature reading from sensor 5 in the building. |
| humidity\_sensor\_5 | Relative humidity percentage recorded by sensor 5. |
| temp\_sensor\_6 | Temperature reading from sensor 6 in the building. |
| humidity\_sensor\_6 | Relative humidity percentage recorded by sensor 6. |
| temp\_sensor\_7 | Temperature reading from sensor 7 in the building. |
| humidity\_sensor\_7 | Relative humidity percentage recorded by sensor 7. |
| temp\_sensor\_8 | Temperature reading from sensor 8 in the building. |
| humidity\_sensor\_8 | Relative humidity percentage recorded by sensor 8. |
| temp\_sensor\_9 | Temperature reading from sensor 9 in the building. |
| humidity\_sensor\_9 | Relative humidity percentage recorded by sensor 9. |
| external\_temp | Outdoor temperature measured by an external sensor. |
| atmospheric\_pressure | Barometric pressure recorded in millimeters of mercury. |
| external\_humidity | Percentage of humidity outside the building. |
| wind\_speed\_mps | Speed of the wind measured in meters per second. |
| visibility\_km | Distance visibility measured in kilometers. |
| dew\_point\_temp | Temperature at which dew begins to form. |

**Problem statement analysing and dataset:**

The problem is need to find the household appliance energy consumption based on the given data and the data in the regression base, so according I need to analyse and train a model, and predict on it, based on the dataset the target variable will be total\_energy\_use, to analyse and create a model and predict on it

**Abstract:**

The model need to predict the household appliance energy consumption, Firstly need do the preprocessing like descriptive Eda and using descriptive analysis need to work on data, like handling null values, correcting data types, removing unwanted columns and unknown columns, after Visual Eda which is used to know about the relationships by correlation between the target variable and also using correlation with target we can find the best feature for model and after correlation, Outliers analysis using boxplot or violinplot visuals and after need to find the patterns like time series on the data whether which month more getting according to the target variable

Scaling the data get best understanding for the model and then splitting data into two parts using target variable, using two parts x, y making it into four x\_train,y\_train, x\_test,y\_test by 80,20 scale using train\_test\_split in sklearn, now using regression model like linear regression, radomforestregressor, decsiontreeregressor , GradientBoostingregressor

Fitting each of them to the training set , predicting the test set using trained model,then evaluating with metrics MSE,MAE,R2 then cross validation of the model based the results selecting the model if required need to do the hyperparameter tunning.

**Steps Taken:**

**Step1:**

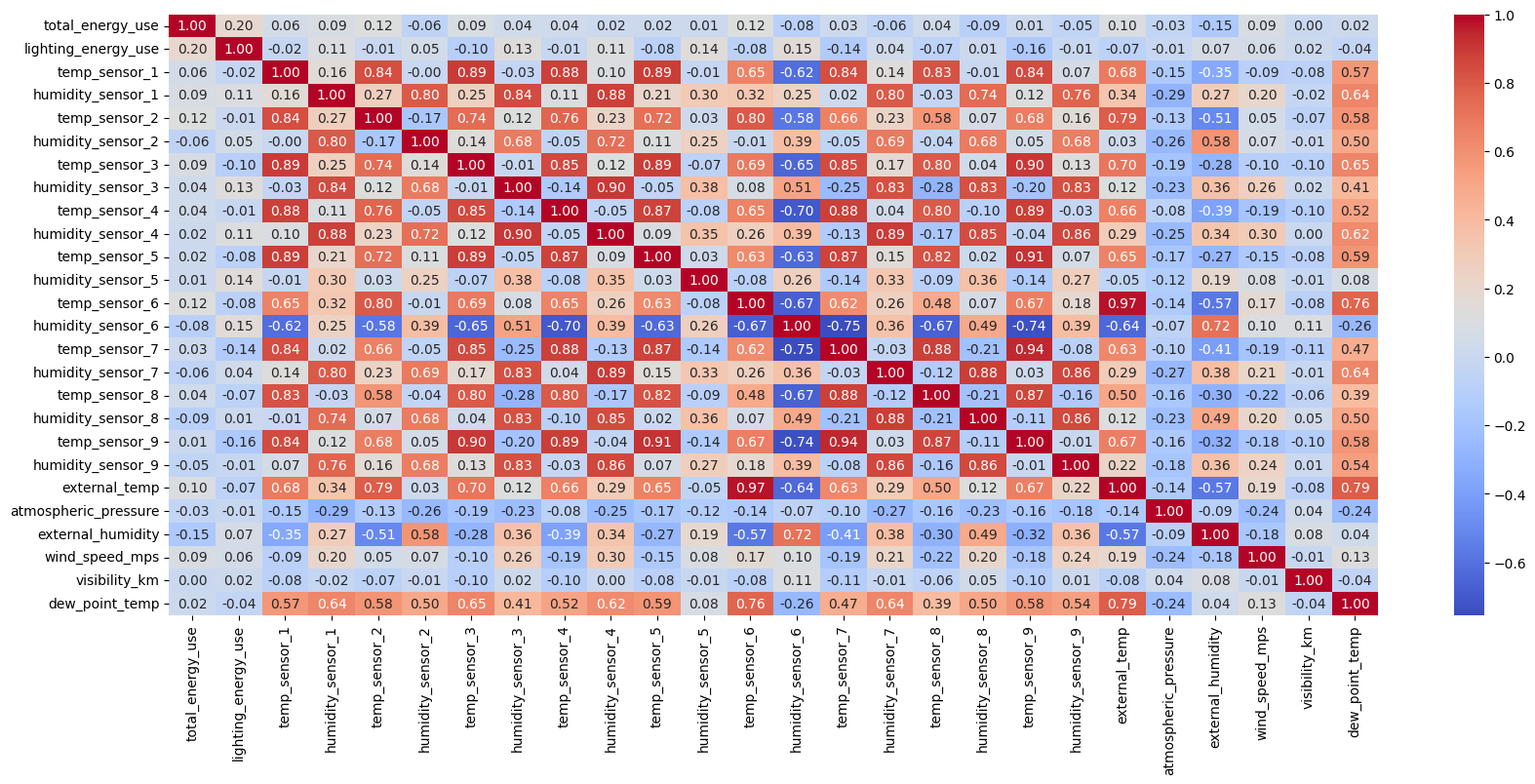
* Installed and imported required libraries

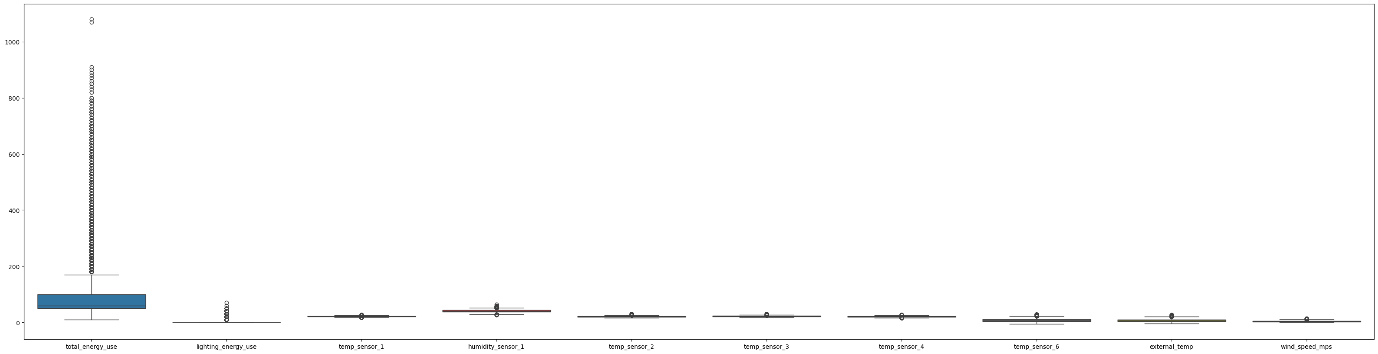
**Step2:**

* Knowing about the data like shape, null values, data types, knowing the unwanted columns like a descriptive EDA
* Step3:
* Removing the unwanted column’s , converting the incorrect data type correct data type for datetime column🡪 recorded\_timestamp

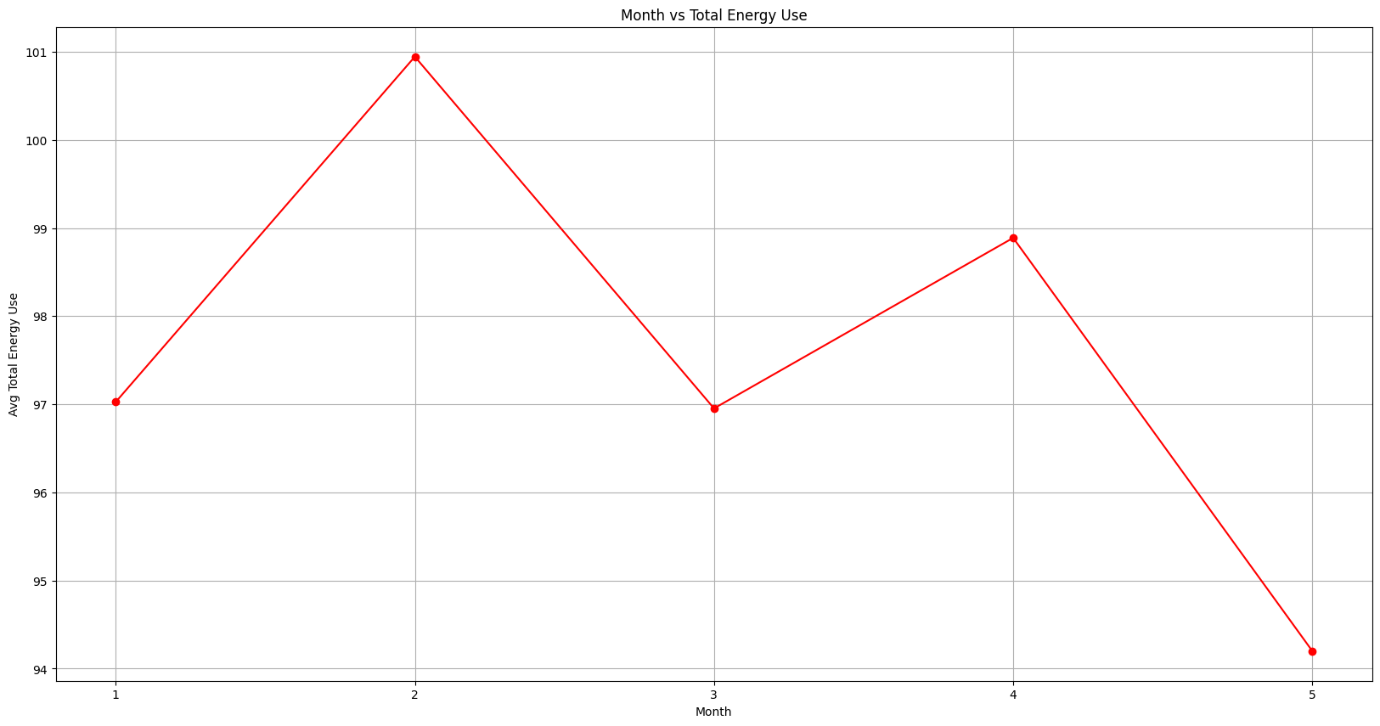
**Step3:**

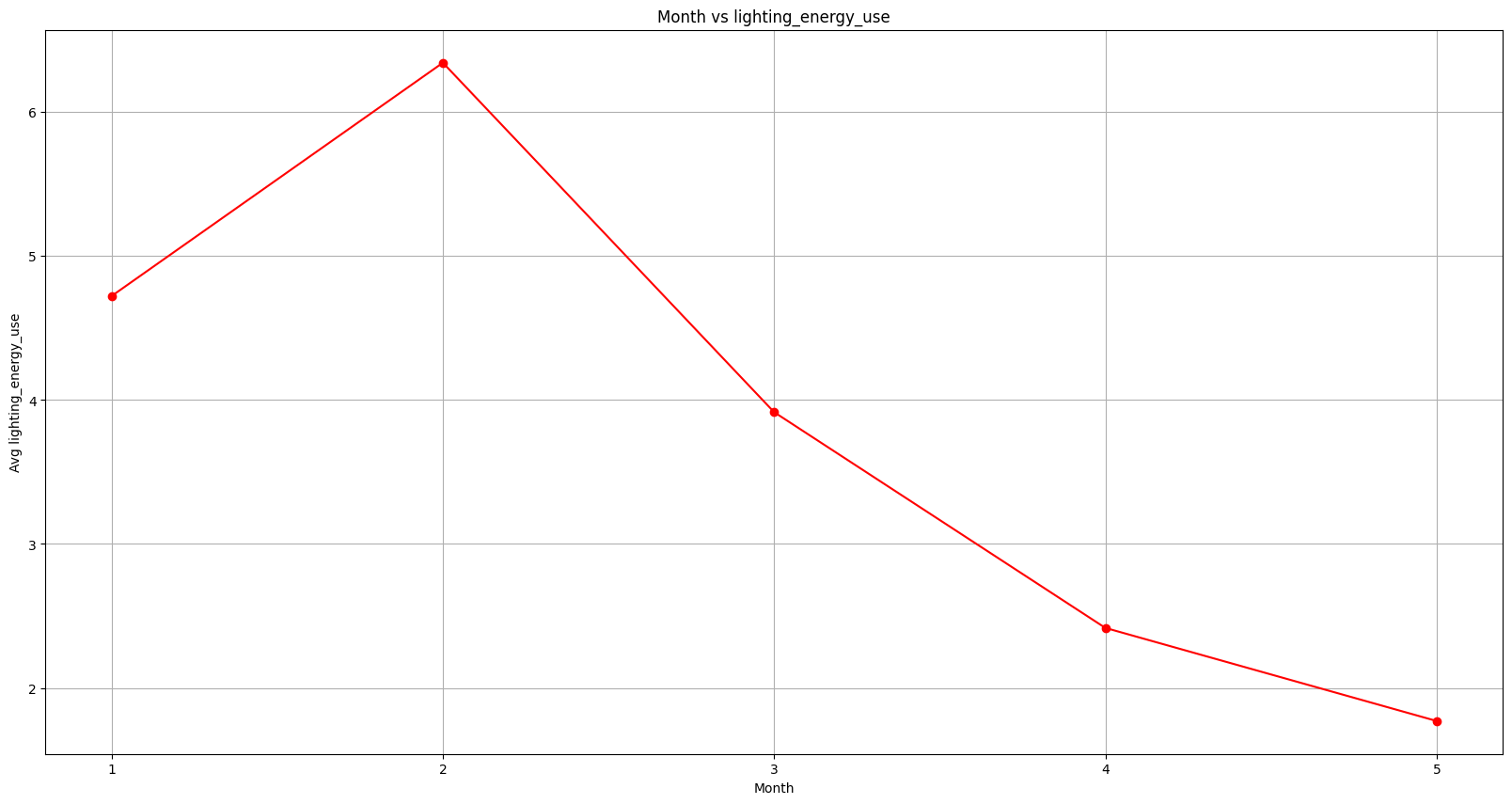
* Visual EDA for know the patterns and relationship in the data.
* Using heatmap for know the correlation blw the data columns and knowing the best feature for the model here we done with feature selection
* Removing the unwanted features from the data but I kept the recorded\_timestamp

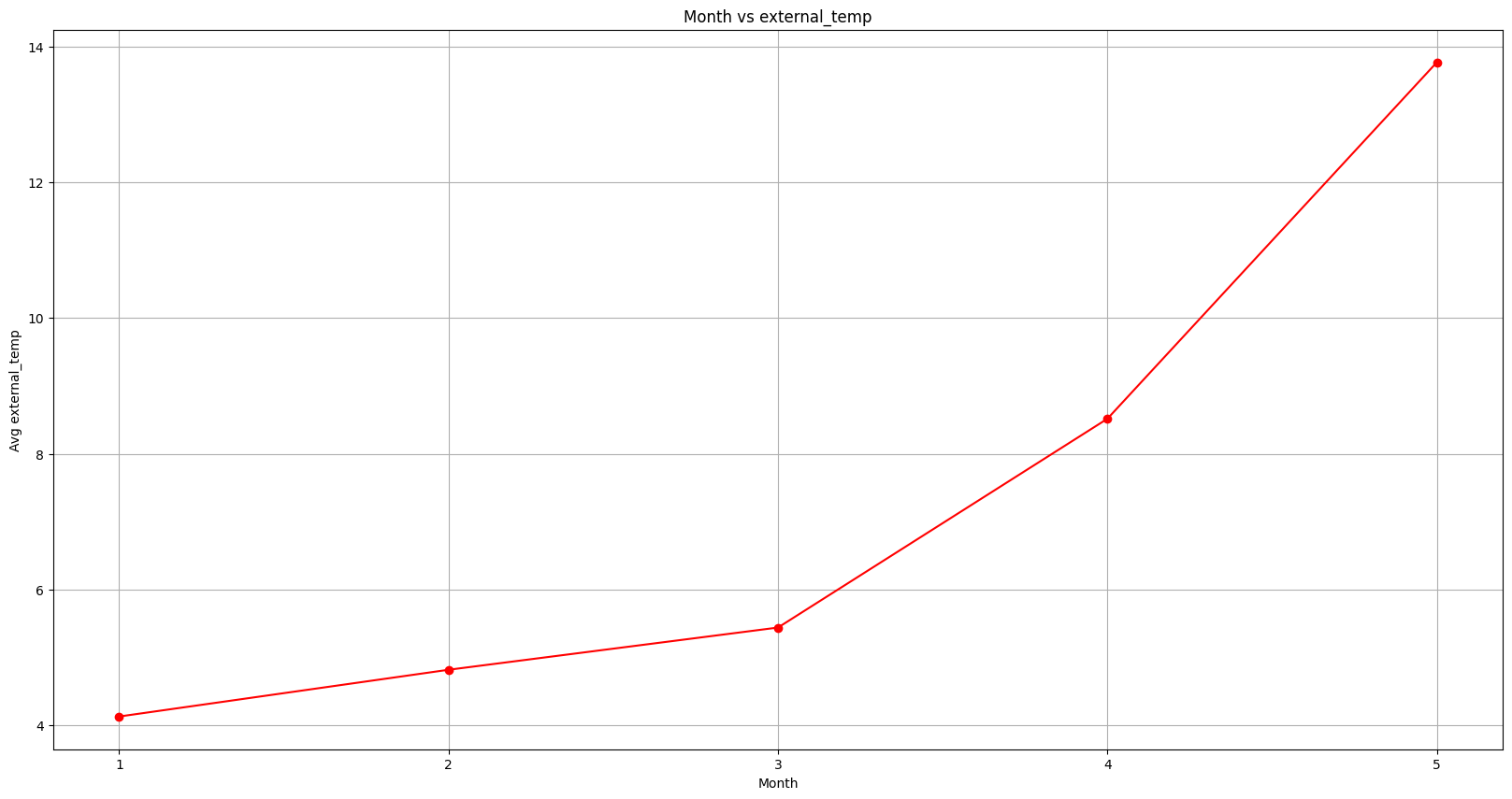


* Now knowing about the outliers using boxplot
* Noticed a lot outliers in each columns
* Needed work on using isolation forest because according the corr the data is non linear so, we need to use the isolation forest

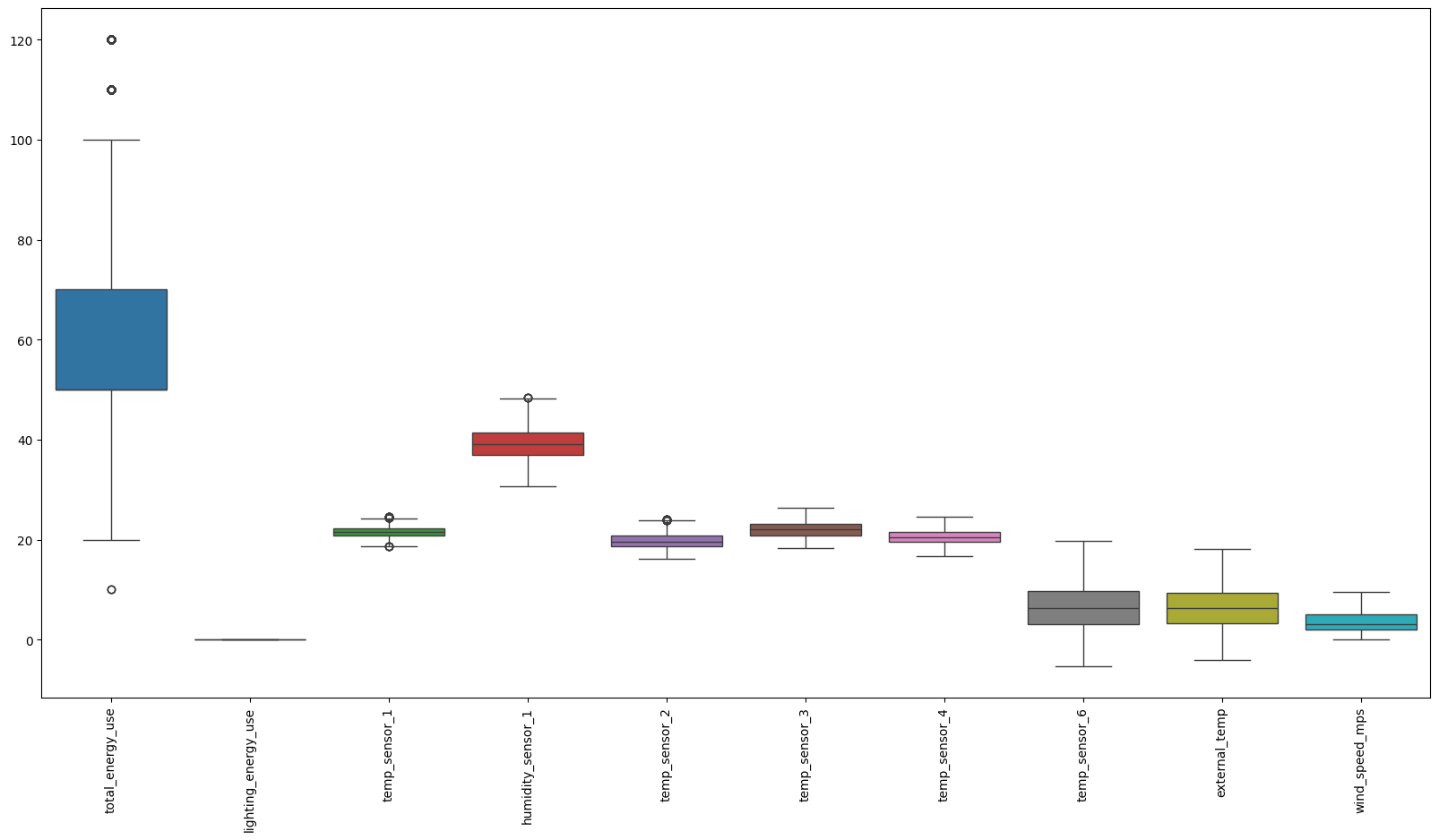
**Step5:**

* Time series analysis on the month because in the dataset I had only one data
* I chosen 3 columns to analyse the trend of the consumption
* Based on the 2-month getting consumption more and the while temp increase the consumption drops





**Step 6:**

* Handling the outliers using isolation forest with 0.3 means 30%, still there many more outliers have, so I need handle using mixed method to remove the outliers
* The second IQR which make the 75%, 25% for box then getting inter quantile using that making the lower and upper bound to detect and remove the outliers
* So using like this I remove the most of the outliers

**Step7:**

* Scaling the data using standardscaler which make data more understanding for model

**Step8:**

* Spliting the data into x,y using target variable

**Step9:**

* Now splitting the data into train and test of both x,y
* With stratify because of the data is non-linear

**Step10:**

* Fitting each model to the train sets
* Compare the metrics selecting the best

**Linear regression():**

Linear Regression Model Performance:

Mean Absolute Error: 0.7246000752060987

Mean Squared Error: 0.9057442030285817

R-squared: 0.09433162546702556

Cross-validated MSE: 0.8929502624521428

**RandomForestRegressor():**

Random Forest Model Performance:

Mean Absolute Error: 0.42566017712296145

Mean Squared Error: 0.3325177527183799

R-squared: 0.6675100855176994

Cross-validated MSE for Random Forest: 0.3762314898566887

**DecisionTreeRegressor():**

Decision Tree Model Performance:

Mean Absolute Error: 0.5280650236538458

Mean Squared Error: 0.6094836210206739

R-squared: 0.3905674046728108

Cross-validated MSE for Decision Tree: 0.7087531386417835

**GradientBoostingRegressor():**

Gradient Boosting Model Performance:

Mean Absolute Error: 0.6355838001096322

Mean Squared Error: 0.7121651366330188

R-squared: 0.28789448551057795

Cross-validated MSE for Gradient Boosting: 0.7280520201265104

**Step11:**

* By the results we get to know that RandomForestRegressor() perform well
* With low mse and high r2 compared to other models

**Step12:**

* Need to do the hyperparameter tuning to get more good results for the model
* Using gridsearchcv with parameter of n\_estimators,max\_depth,min\_sampls\_split,min\_samples\_leaf
* I worked with different parameter but still not getting good result compared to model before tuning
* Hence, I need to stick with before tuning model for good outcome

Best Hyperparameters: {'max\_depth': 15, 'min\_samples\_leaf': 1, 'min\_samples\_split': 10, 'n\_estimators': 100}

Best Random Forest Model Performance:

Mean Absolute Error: 0.4557229326910484

Mean Squared Error: 0.37652046488902674

R-squared: 0.6235110572342422

Cross-validated MSE for Best Random Forest: 0.42015585640533193