# DATA ANALYTICS IN BUILDING SCIENCE ( AR32203 )

## TIME-SERIES FORECASTING MODEL FOR ENERGY USAGE AND OCCUPANCY PREDICTION USING IAQ AND ENERGY DATA IN AN OFFICE ROOM

UNDER THE GUIDANCE OF : PROF. PRASHANT ANAND

Submitted by: GROUP-4

A Gautham Lakshmanan 19AR10001
Aragya Gupta 21AR10005
Aryan Jaiswal 21AR10006
Kumar Shivam 21AR10019
Mohit Gwal 21AR10021

Sneh Patel 21AR10023
Rupangshu Banik 21AR10029
Sumeet Dubey 21AR10034
Tushar Mondal 21AR10037

# TABLE OF CONTENTS

# SUMMARY

The main topic of our conversation revolves around the development of a predictive model for occupancy prediction in an office environment using IAQ (Indoor Air Quality) and energy data. This topic is crucial in the field of building management systems and occupant well-being, as accurate occupancy prediction can optimize energy usage, enhance indoor air quality, and improve occupant comfort.

Our discussion addressed the research gap in the existing literature concerning the integration of IAQ and energy data for occupancy prediction. By leveraging machine learning techniques and time-series forecasting models, we aimed to bridge this gap and provide a comprehensive solution for building managers and researchers.

The primary research question guiding our conversation was how to effectively utilize IAQ and energy data to predict occupancy levels in office spaces. We adopted a systematic approach, starting with data preprocessing steps, including handling missing values and feature engineering, followed by model selection, training, and evaluation.

Key findings from our exploration include the identification of relevant features for occupancy prediction, the implementation of predictive models such as random forest classifiers, and the evaluation of model performance using metrics like accuracy and classification reports.

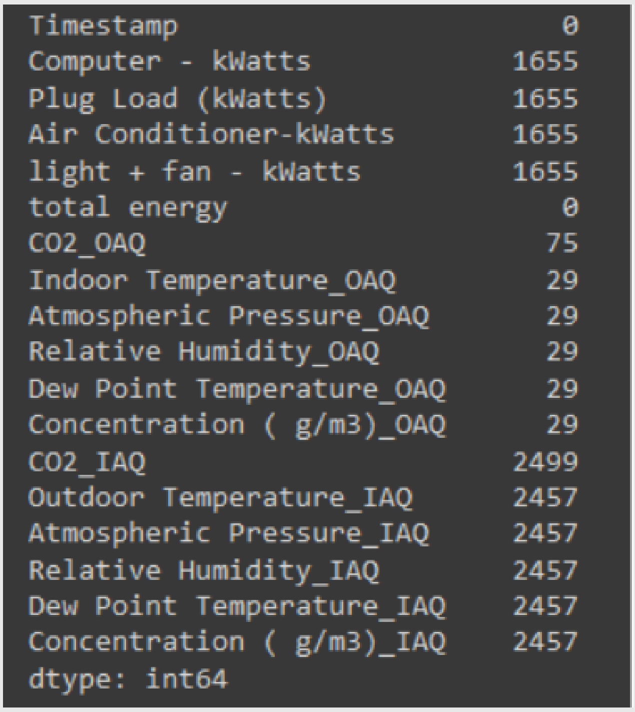
Our findings underscore the importance of integrating IAQ and energy data for occupancy prediction, offering insights into energy usage patterns, indoor air quality dynamics, and occupant behavior. By accurately predicting occupancy levels, building managers can optimize resource allocation, improve building sustainability, and enhance occupant comfort and productivity.

In conclusion, our conversation contributes to the field by offering a comprehensive approach to occupancy prediction, leveraging IAQ and energy data and machine learning techniques. Our findings highlight the potential of predictive analytics in building management systems and underscore the importance of considering IAQ alongside energy usage for optimal building operation and occupant well-being.

**GROUP-4**

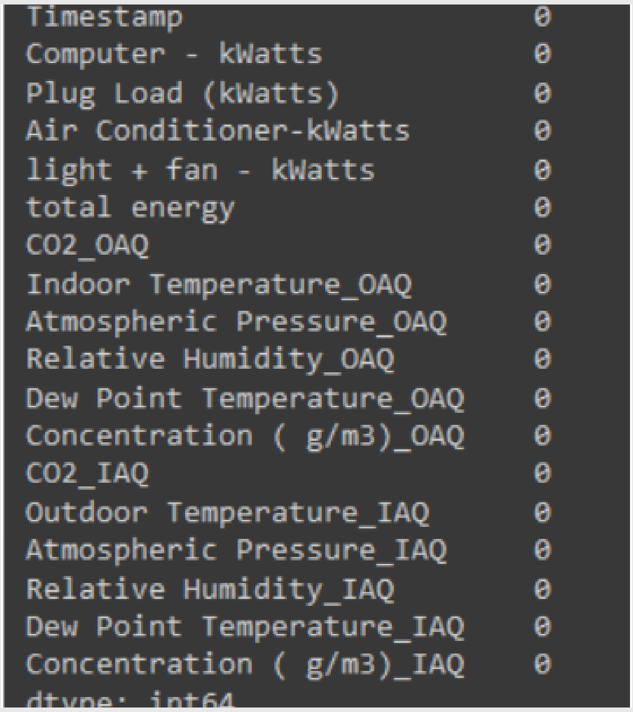
## 1.1 DATA CLEANING AND FORMATTING

The combined version of three distinct datasets—ENERGY DATA, OUTDOOR AIR QUALITY (OAQ), and INDOOR AIR QUALITY (IAQ)—makes up the final integrated time series data. The common timestamp shared by all three of the datasets is the final timestamp. a number of different data preprocessing methods, including mean imputation to replace NaN (Not a Number) values and formatting timestamps in the widely recognized HH:MM:SS YY:MM:DD format. To ensure that the entire dataset produces forecasts and predictions of the highest caliber, duplicate timestamps have been eliminated.



```
Timestamp                      0
Computer - kWatts           1655
Plug Load (kWatts)          1655
Air Conditioner-kWatts      1655
light + fan - kWatts        1655
total energy                   0
CO2_OAQ                       75
Indoor Temperature_OAQ        29
Atmospheric Pressure_OAQ      29
Relative Humidity_OAQ         29
Dew Point Temperature_OAQ     29
Concentration ( g/m3)_OAQ     29
CO2_IAQ                     2499
Outdoor Temperature_IAQ     2457
Atmospheric Pressure_IAQ    2457
Relative Humidity_IAQ       2457
Dew Point Temperature_IAQ   2457
Concentration ( g/m3)_IAQ   2457
dtype: int64
```

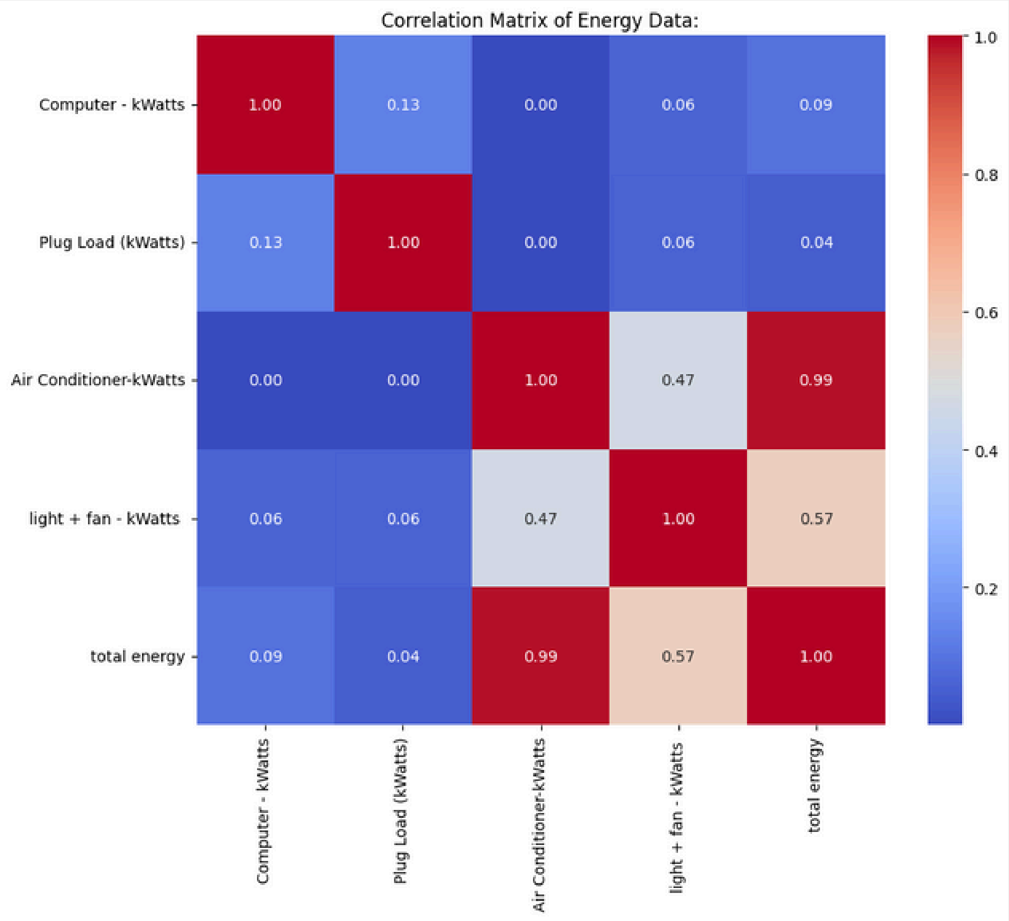
Number of NaN values before



```
Timestamp                    0
Computer - kWatts            0
Plug Load (kWatts)           0
Air Conditioner-kWatts       0
light + fan - kWatts         0
total energy                 0
CO2_OAQ                      0
Indoor Temperature_OAQ       0
Atmospheric Pressure_OAQ     0
Relative Humidity_OAQ        0
Dew Point Temperature_OAQ    0
Concentration ( g/m3)_OAQ    0
CO2_IAQ                      0
Outdoor Temperature_IAQ      0
Atmospheric Pressure_IAQ     0
Relative Humidity_IAQ        0
Dew Point Temperature_IAQ    0
Concentration ( g/m3)_IAQ    0
dtype: int64
```

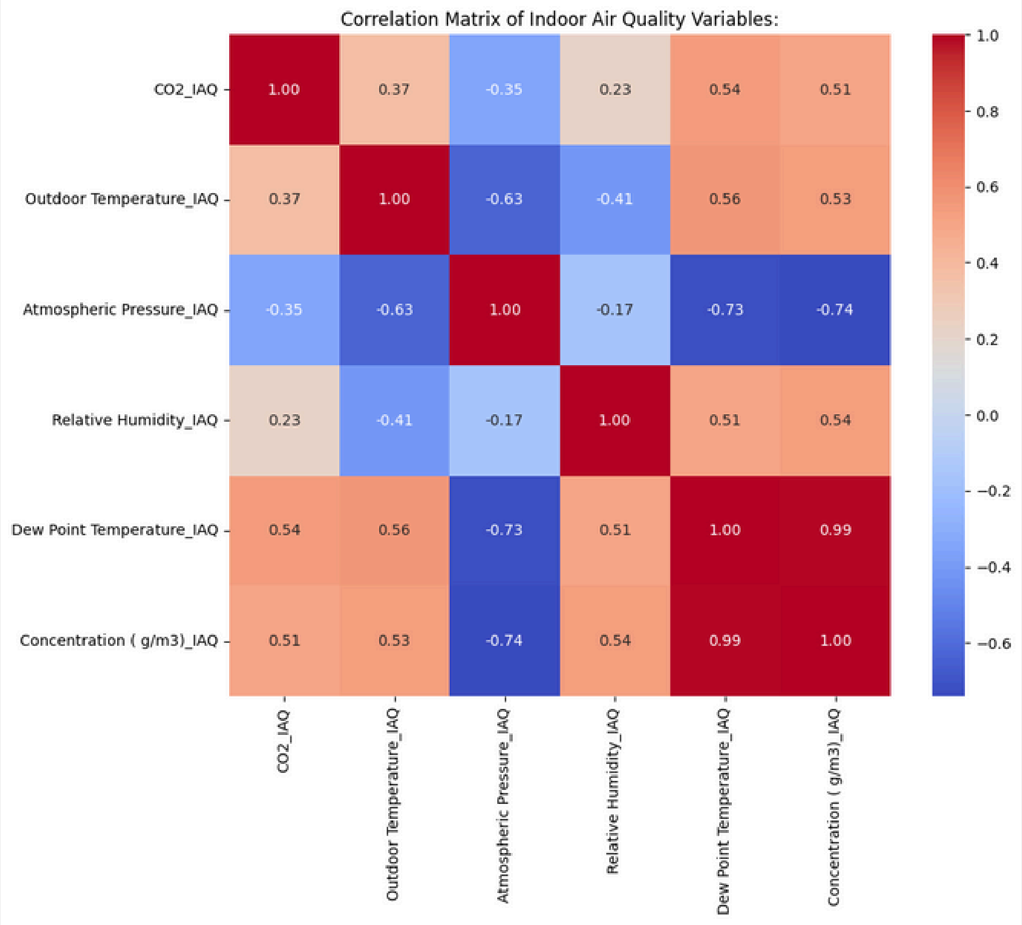
Number of NaN values after

# 1.2 CORELATION ANALYSIS



Correlation matrix of Energy Data

In the room where the measurement was made, there is a nearly perfect correlation (0.99) between AC load and total energy, indicating that AC is the primary source of energy use overall.
When the two variables have a 0.99 correlation coefficient, it means that they typically move in the same direction. The tendency is for both variables to rise in tandem with an increase in one.
The relationship between the light+fan load and total energy is modest (0.57) linear. It suggests that there is a correlation between these factors that is discernible but not very strong. A correlation value of 0.57 suggests that there is a tendency for the two variables to move in tandem. The tendency is for both variables to rise in tandem with an increase in one.



| | CO2_IAQ | Outdoor Temperature_IAQ | Atmospheric Pressure_IAQ | Relative Humidity_IAQ | Dew Point Temperature_IAQ | Concentration ( g/m3)_IAQ |
|---|---|---|---|---|---|---|
| CO2_IAQ | 1.00 | 0.37 | -0.35 | 0.23 | 0.54 | 0.51 |
| Outdoor Temperature_IAQ | 0.37 | 1.00 | -0.63 | -0.41 | 0.56 | 0.53 |
| Atmospheric Pressure_IAQ | -0.35 | -0.63 | 1.00 | -0.17 | -0.73 | -0.74 |
| Relative Humidity_IAQ | 0.23 | -0.41 | -0.17 | 1.00 | 0.51 | 0.54 |
| Dew Point Temperature_IAQ | 0.54 | 0.56 | -0.73 | 0.51 | 1.00 | 0.99 |
| Concentration ( g/m3)_IAQ | 0.51 | 0.53 | -0.74 | 0.54 | 0.99 | 1.00 |

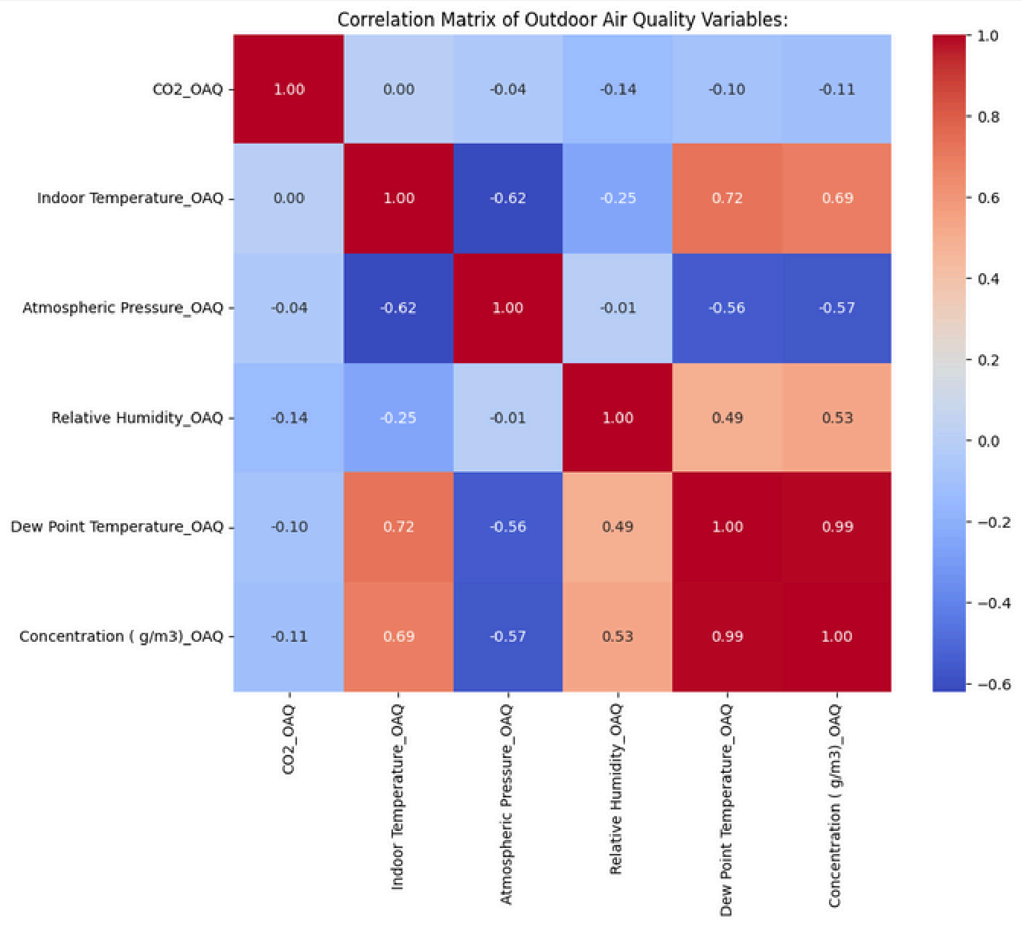
Correlation matrix of Indoor Air Quality

The association between CO2_IAQ and Dew Point Temperature_IAQ is strongly positive (0.54), suggesting that higher indoor CO2 levels are linked to higher dew point temperatures.

The relationship between Outdoor Temperature_IAQ and Atmospheric Pressure_IAQ is moderately negative (-0.63), indicating that greater outdoor temperatures are typically associated with lower atmospheric pressure.

There is a somewhat positive association (0.51 and 0.54) between Relative Humidity_IAQ and Concentration (g/m3)_IAQ, respectively, and Dew Point Temperature_IAQ. This suggests that greater concentrations of air contaminants and higher dew point temperatures are related to higher relative humidity.

Dew Point Temperature_IAQ and Atmospheric Pressure_IAQ have a strong negative connection (-0.73), meaning that greater dew point temperatures are linked to lower atmospheric pressure.



Correlation matrix of Outdoor Air Quality

FThe fact that CO2_OAQ and Indoor Temperature_OAQ have no connection (0.00) indicates that indoor and outdoor CO2 concentrations are unrelated. greater inside temperatures are correlated with greater outdoor dew point temperatures, as shown by the strong positive correlation (0.72) between inside Temperature_OAQ and Dew Point Temperature_OAQ.

There is a moderately negative association (-0.56 and -0.57) between Concentration (g/m3)_OAQ and Dew Point Temperature_OAQ and Atmospheric Pressure_OAQ, respectively. This suggests a relationship between higher dew point temperatures and higher outdoor air pollution concentrations and lower atmospheric pressure.

Pollutants and Dew Point Temperature_OAQ exhibit a somewhat positive connection (0.49 and 0.53) with Relative Humidity_OAQ.

Concentration (g/m3)_OAQ and Dew Point Temperature_OAQ have a very strong positive correlation (0.99), suggesting that they are closely associated and have a tendency to rise or fall together.

## 1.3 MODEL DEVELOPMENT

The whole process involves following steps :-

**Importing Libraries:** The code imports necessary libraries from scikit-learn, including functions for splitting data (train_test_split), preprocessing (StandardScaler), building a random forest classifier (RandomForestClassifier), and evaluating model performance (accuracy_score, classification_report).

**Defining Occupancy Label:** An occupancy label is defined based on existing features in the dataset. In this example, occupancy is determined by whether the total energy consumption exceeds a specified threshold (0.5). The 'Occupancy' column in the dataset is created, where a value of 1 indicates occupancy and 0 indicates non-occupancy.

**Selecting Features and Target Variable:** Relevant features related to energy usage and IAQ are selected for model training, while the target variable 'Occupancy' is defined.

**Splitting Data:** The dataset is split into training and testing sets using the train_test_split function. The training set comprises 80% of the data, while the testing set comprises the remaining 20%.

**Feature Scaling:** The selected features are standardized using StandardScaler to ensure that all features have a mean of 0 and a standard deviation of 1. This step is crucial for models that rely on distance-based calculations.

**Choosing and Training the Model:** A random forest classifier is chosen as the classification model. The model is trained on the scaled training data using the fit method.

**Making Predictions:** The trained model is used to make predictions on the scaled testing data using the predict method.

**Evaluating the Model:** The accuracy of the model is calculated using the accuracy_score function, which compares the predicted labels with the actual labels from the test set. Additionally, a detailed classification report is generated using the classification_report function, providing metrics such as precision, recall, and F1-score for each class.

### 1.3.1 TIMESERIES FORECASTING OF ENERGY

The whole process involves following steps :-

The code reads an Excel file containing the energy consumption data and performs necessary preprocessing steps, such as renaming columns, handling missing values, and converting the timestamp column to datetime format.

The code visualizes the energy consumption time series plot to understand the patterns and trends in the data.

It performs the Augmented Dickey-Fuller (ADF) test to check for stationarity of the time series. Stationarity is an important assumption for many time series models, including ARIMA.

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are generated to help determine the appropriate orders for the ARIMA model (p, d, q).

Based on the ACF and PACF plots, the code selects the orders (p=3, d=0, q=6) for the ARIMA model.

The ARIMA model is fitted to the training data using the specified orders.

The fitted ARIMA model is used to forecast the energy consumption values for the testing set.

The Root Mean Squared Error (RMSE) is calculated between the actual and forecasted values for the testing set, providing an evaluation metric for the model's performance.

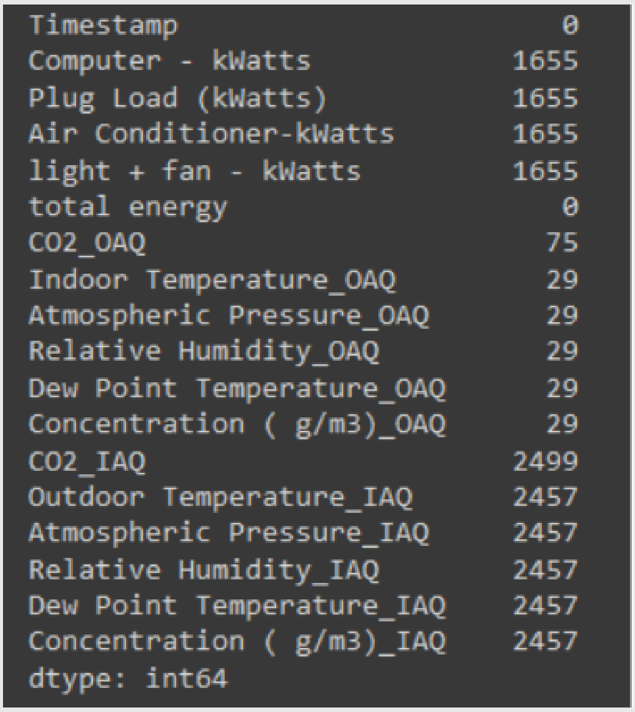
## 1.3.2 OCCUPANCY PREDICTION

The whole process involves following steps :-

- **Importing Libraries:** The code imports the scikit-learn libraries required, such as train_test_split, StandardScaler for preprocessing, RandomForestClassifier for creating a random forest classifier, and Accuracy Score and Classification Report for assessing model performance.
- **Defining Occupancy Label:** Based on attributes already present in the dataset, an occupancy label is defined. In this instance, occupancy is ascertained by determining if the overall energy usage surpasses a designated cutoff point (0.5). A value of 1 denotes occupancy, while a value of 0 denotes non-occupation. This is the 'Occupancy' column that is created in the dataset.
- **Selecting Features and Target Variable:** For model training, pertinent characteristics pertaining to energy consumption and indoor air quality are chosen, and the target variable 'Occupancy' is specified.
- **Splitting Data:** The train_test_split function divides the dataset into training and testing sets. Eighty percent of the data are from the training set and the remaining twenty percent are from the testing set.
- **Feature Scaling:** To guarantee that every feature has a mean of 0 and a standard deviation of 1, the chosen features are standardized using StandardScaler. For models that depend on computations based on distance, this phase is essential.
- **Choosing and Training the Model:** The selection of the classification model is a random forest classifier. Using the fit method, the model is trained on the scaled training data.
- **Making Predictions:** Using the predict approach, predictions are made on the scaled testing data using the trained model.
- **Evaluating the Model:** The accuracy_score function, which contrasts the predicted labels with the actual labels from the test set, is used to determine the model's accuracy. Furthermore, the classification_report function generates a comprehensive classification report that includes metrics for each class, including F1-score, precision, and recall.

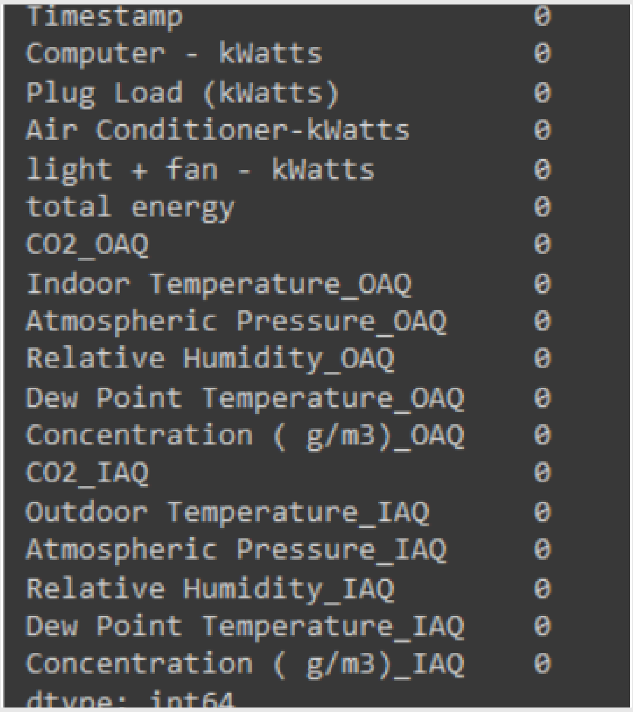
## 2.1 OVERVIEW OF TIME SERIES DATA AND GENERAL INFERENCES

The final merged time series data is a combined version of three different datasets, namely **ENERGY DATA**, **OUTDOOR AIR QUALITY (OAQ)**, and **INDOOR AIR QUALITY (IAQ).** The final timestamp is the common timestamp present in each three individual datasets. Various data preprocessing techniques such as **Mean Imputation** for replacing NaN (Not a Number) values, Changing format of timestamp to internationally accepted format of **YY:MM:DD HH:MM:SS.** Removing duplicate timestamps have been done to make the overall dataset produce high quality forecasting and predictions.



```
Timestamp                         0
Computer - kWatts              1655
Plug Load (kWatts)             1655
Air Conditioner-kWatts         1655
light + fan - kWatts           1655
total energy                      0
CO2_OAQ                          75
Indoor Temperature_OAQ           29
Atmospheric Pressure_OAQ         29
Relative Humidity_OAQ            29
Dew Point Temperature_OAQ        29
Concentration ( g/m3)_OAQ        29
CO2_IAQ                        2499
Outdoor Temperature_IAQ        2457
Atmospheric Pressure_IAQ       2457
Relative Humidity_IAQ          2457
Dew Point Temperature_IAQ      2457
Concentration ( g/m3)_IAQ      2457
dtype: int64
```

**Number of NaN values before**

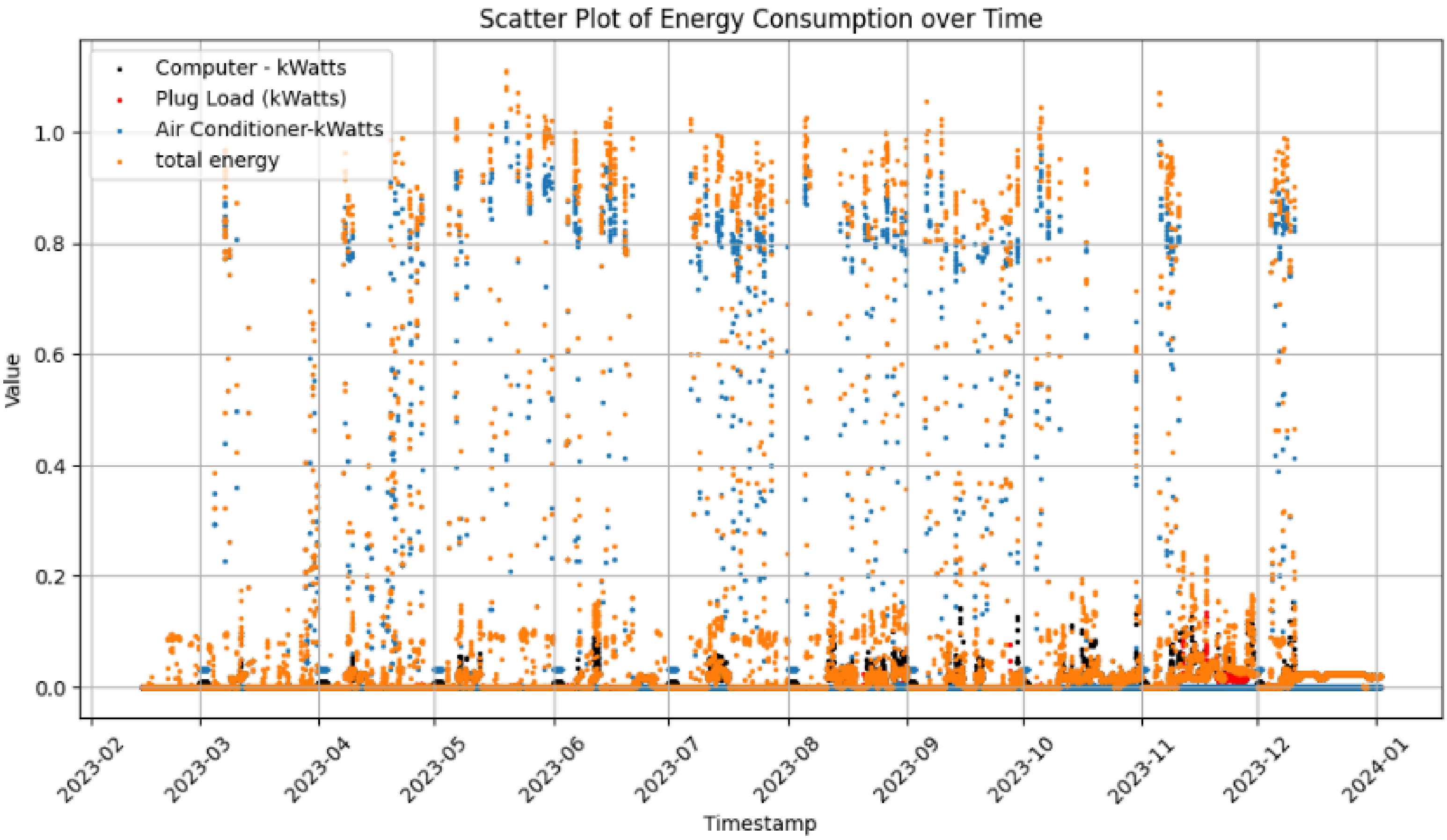


```
Timestamp                    0
Computer - kWatts            0
Plug Load (kWatts)           0
Air Conditioner-kWatts       0
light + fan - kWatts         0
total energy                 0
CO2_OAQ                      0
Indoor Temperature_OAQ       0
Atmospheric Pressure_OAQ     0
Relative Humidity_OAQ        0
Dew Point Temperature_OAQ    0
Concentration ( g/m3)_OAQ    0
CO2_IAQ                      0
Outdoor Temperature_IAQ      0
Atmospheric Pressure_IAQ     0
Relative Humidity_IAQ        0
Dew Point Temperature_IAQ    0
Concentration ( g/m3)_IAQ    0
dtype: int64
```

**Number of NaN values after**

Data Visualization techniques like using Boxplots, Scatterplot have been used to discover trends and relationships between dependent variables. Each of these techniques are used individually on ENERGY DATA, OUTDOOR AIR QUALITY (OAQ), and INDOOR AIR QUALITY (IAQ), taking “Timestamp” label as X axis in all, and dependent label as Y axis.

## Below is the Scatter Plot for **ENERGY DATA** vs **Timestamp** :-



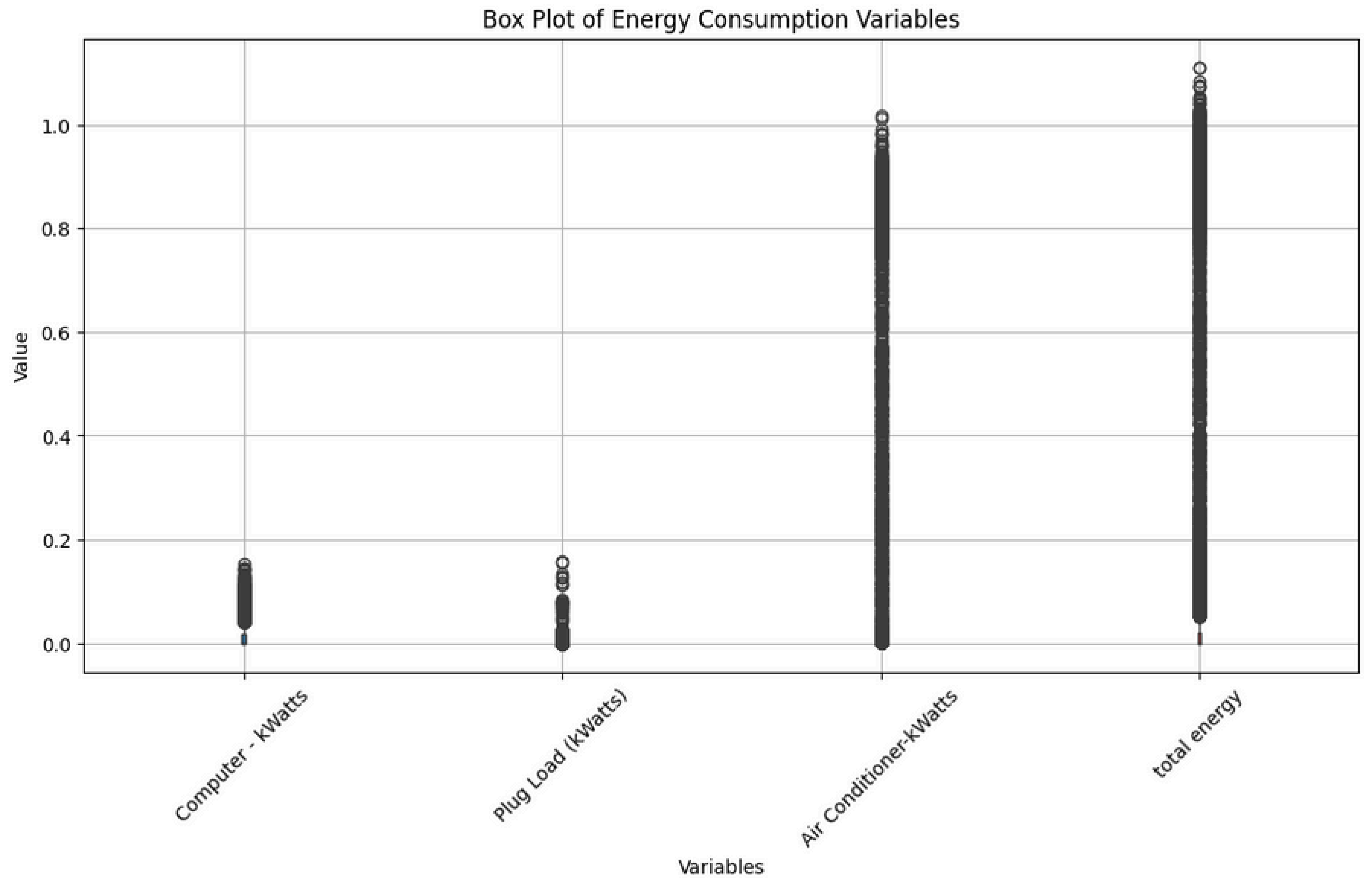
We have taken 4 dependent labels, namely - “Computer - kWatts”, “Plug Load (kWatts)”, “Air Conditioner-kWatts”, “total energy” as Y axis and Timestamp from February 2023 to January 2024 as X axis. The colors of individual labels are different for easy understanding.
Inferences -

- **Computer Load** - Marked with black, the values of computer load stays within the range of 0-0.2 KWatt per timestamp. Computer Load is an important parameter as it gives us direct connection to occupancy in the room. Values higher than 0 means an Occupant is present in the room.

- **Plug load** - Marked with red, the values of plug load varies in a systematic way per month. Plug Load isn’t an important parameter for us as it is the sum of total plug load consumed per timestamp. It may happen that occupants may or may not be present in the room, but Plug Load is present and greater than 0.

- **AC Load -** Marked with blue, the values of AC load fluctuates between 0-1 KWatt per timestamp, depending on various factors like season, outdoor temperature, time of the day. AC Load is an important parameter as it gives us overall total connection to occupancy in the room, as well as total energy spend. The AC load values as seen from the graph typically increases during mid year or summer season and decreases in winter season. This is self explanatory, due to outdoor temperature. Also AC load values are slightly higher in mid-day time, suggesting increasing indoor temperature during mid-day time of a typical day.

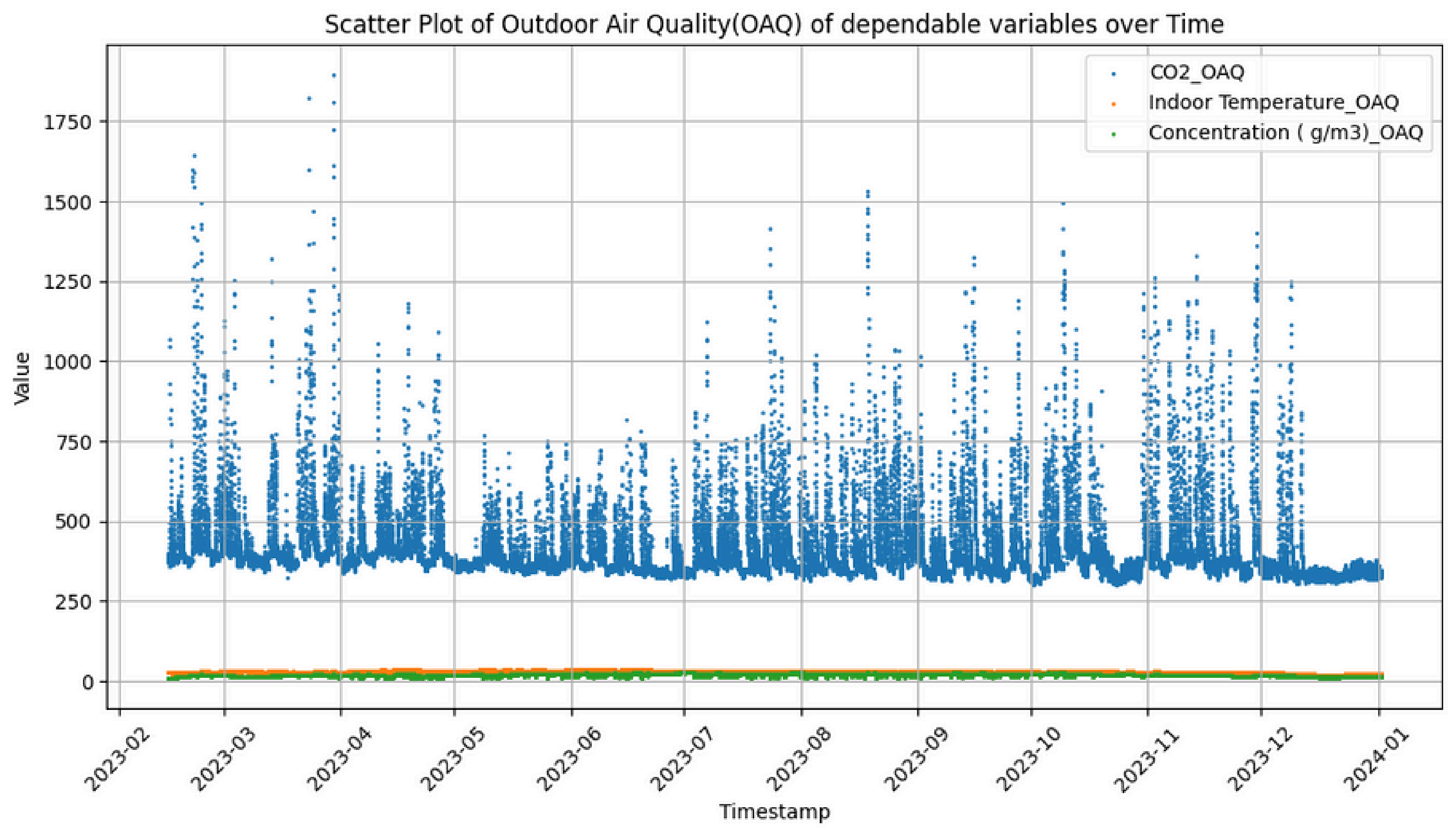
- **Total Energy-** Marked with orange, the Total Energy is the sum of all 3 dependent variables. Total Energy is the direct measure of occupancy levels in a room. A value higher than 0.5, possibly tells us that atleast one person is present in the room from where measurement is taken.



BoxPlot for Energy Consumption Variables - Each circle signifies each value

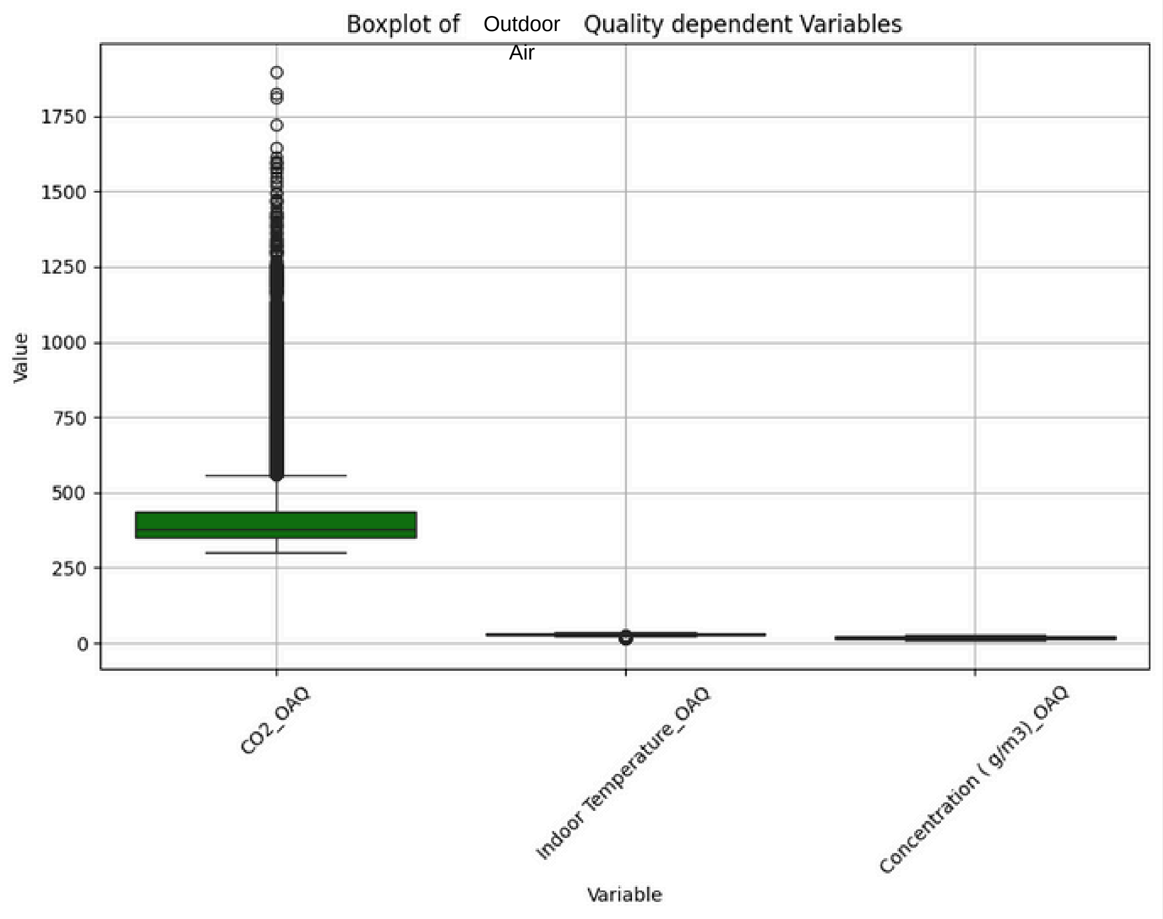
### From the BoxPlot we can conclude that -

- The value of Computer load ranges from 0-0.2 KW, which is merely 10 - 15% of total energy, thereby signifying it has less contribution to total energy consumption,
- The range of Plug load is almost similar to Computer load, thus giving us a clear idea that it also contributes less to the total energy
- The high range of values of AC clearly signifies the wide range of indoor temperature that was there in the room when values were taken. Also, the values are uniform from 0-1 , giving us a sign that maybe the AC was turned on throughout the occupancy time.
- 90% of the value of total energy is contributed by AC load.
- Total energy consumption gives us direct idea when the room was occupied and when not.



### From the above Scatter Plot we can conclude that -

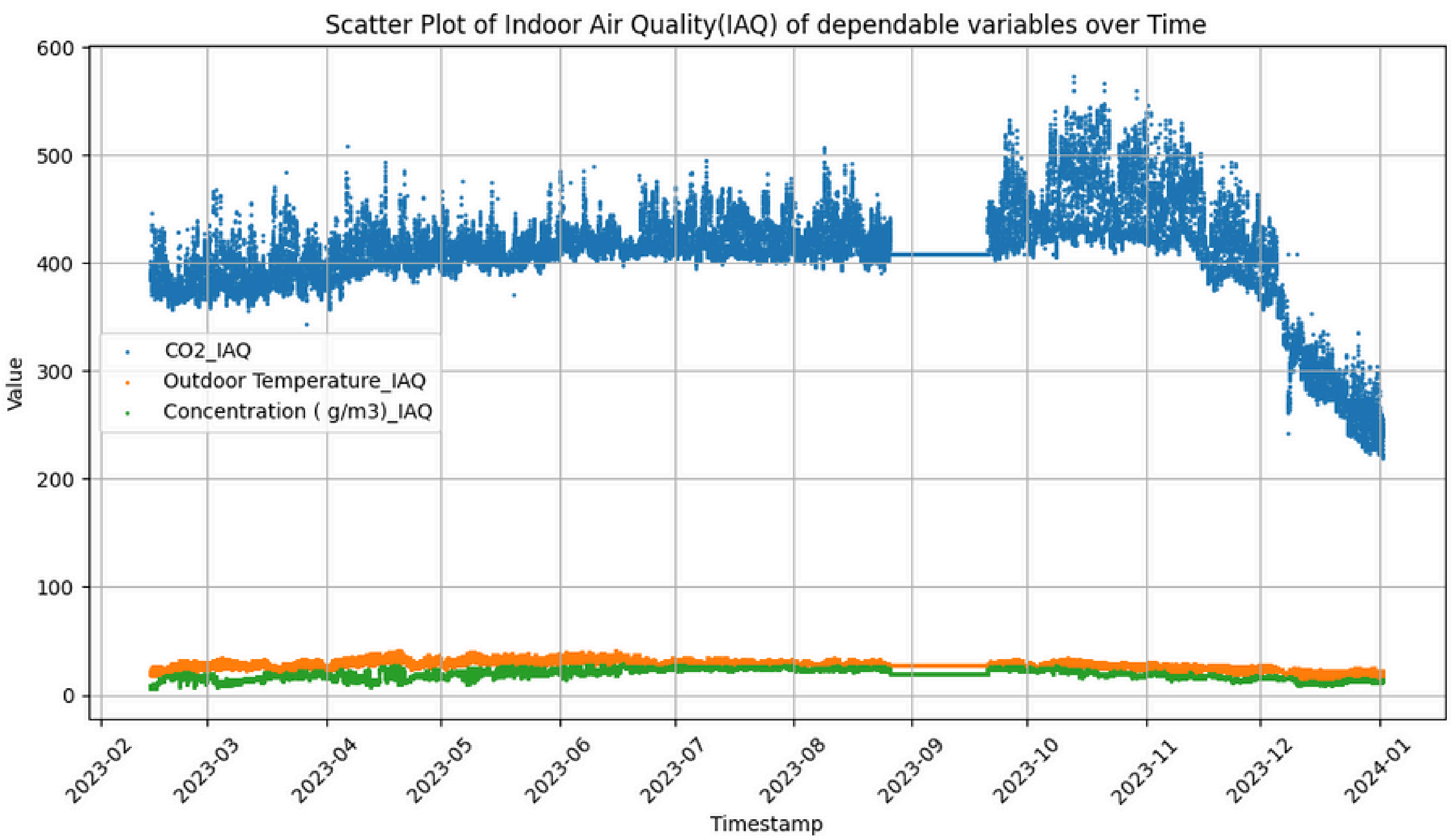
- The value of indoor CO2 levels varied very much during an interval of days or hours.
- Variation in Concentration isnt as high as that of CO2 levels.
- There are certain days where there isnt much variation of values, and the curve is constant ones (ex- 05-2023, 11-2023, 12-2023 to 01-2024). This says us that those values were imputed with mean bvalues, and original data were absent in the dataset.



## Box Plot for OAQ vs **Timestamp**

The above statements can be clearly proved from box plot as shown in the above graph. There is negligible to no outliers in Indoor Temperature and Concentration, but there are significant outliers in CO2 levels. This clearly shows us that there was wide variation of CO2 levels in some days or time frame. The reasons for these high fluctuations are :-

- Some days there have been some group meetings in the room where data was taken. Since all members of the building participated, this led to high CO2 levels during that Timestamp.
- High number of electrical equipment which emit CO2 were in use in some days, which elevated CO2 levels.

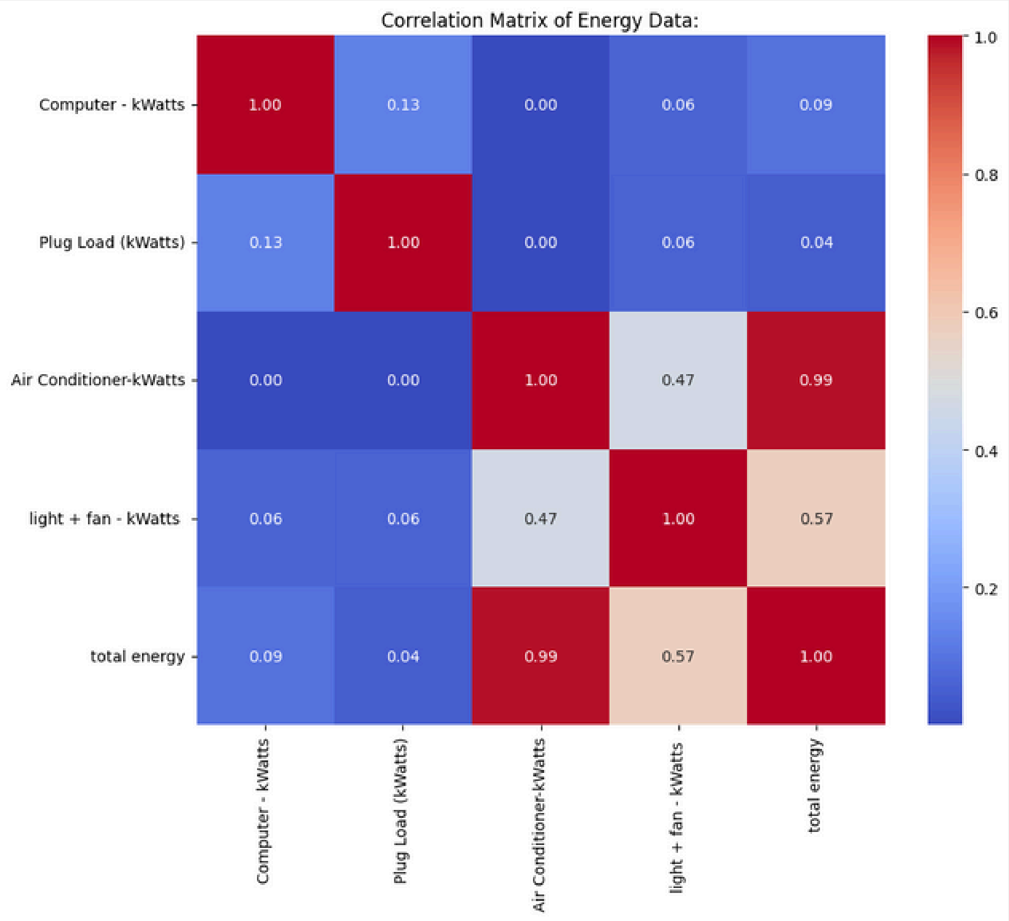


## Scatter Plot for **IAQ** vs **Timestamp**

Whereas a shift in trend is noticed in this scatterplot. Here the values dont have as high variation as OAQ, and stays within range of 350-550. Also, a staright line can be seen between 2023-09 and 2023-10. That is the timestamp, where NaN values were replaced by Mean imputation method. Also we can see a sharp dip in CO2 values from 11-2023 to 01-2024. This sharp dip can be caused due to these reasons -

- During winter, you might be opening windows more frequently to bring in fresh air. This influx of fresh air with lower CO2 concentration would dilute the CO2 levels in the room, causing a sharp dip.
- Winter months might lead to people spending less time in the room. Fewer occupants breathing and releasing CO2 would result in a lower overall CO2 concentration.
- Since AC wasn’t used in winter season, so there may be less CO2 levels, as AC is a major contributor of high electricity consumption, thereby increasing CO2 levels. So not using it infact decreased CO2 levels drastically.

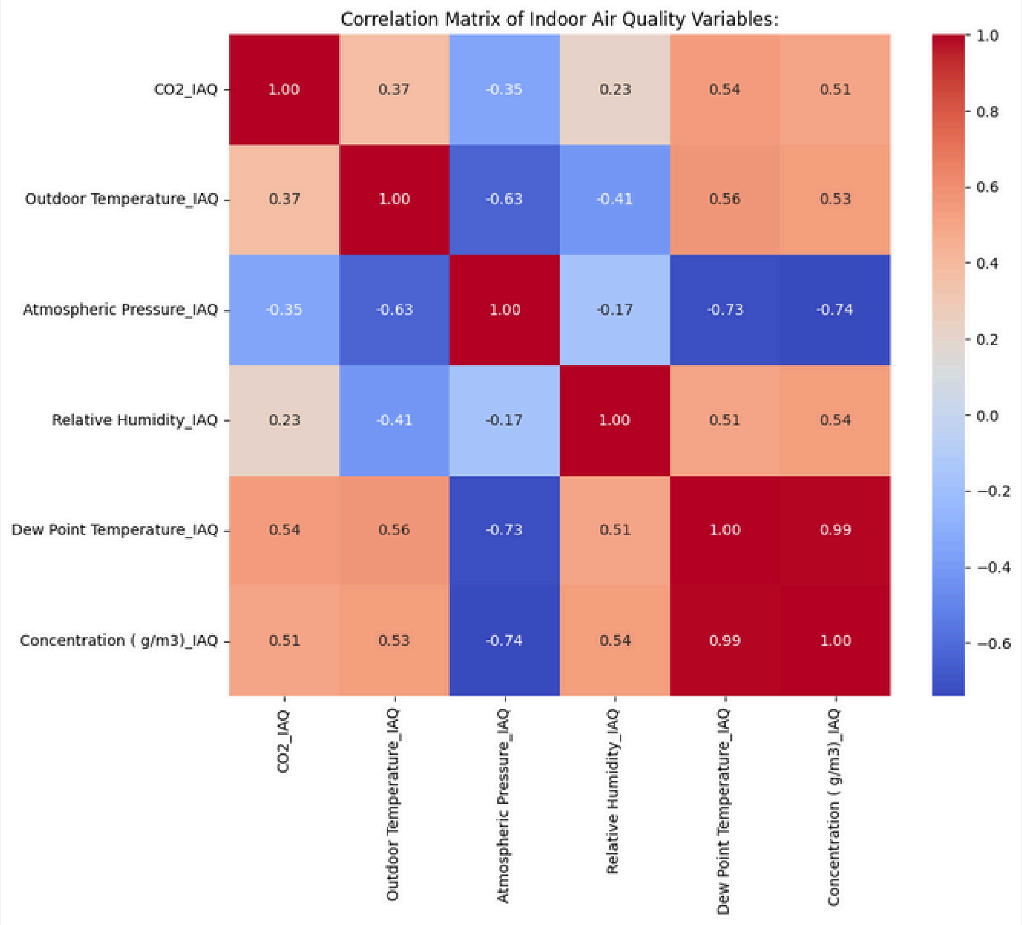
## 2.2 CORRELATION MATRIX



Correlation matrix of Energy Data

From the correlation matrix of Energy Data, we can come to these conclusions :-

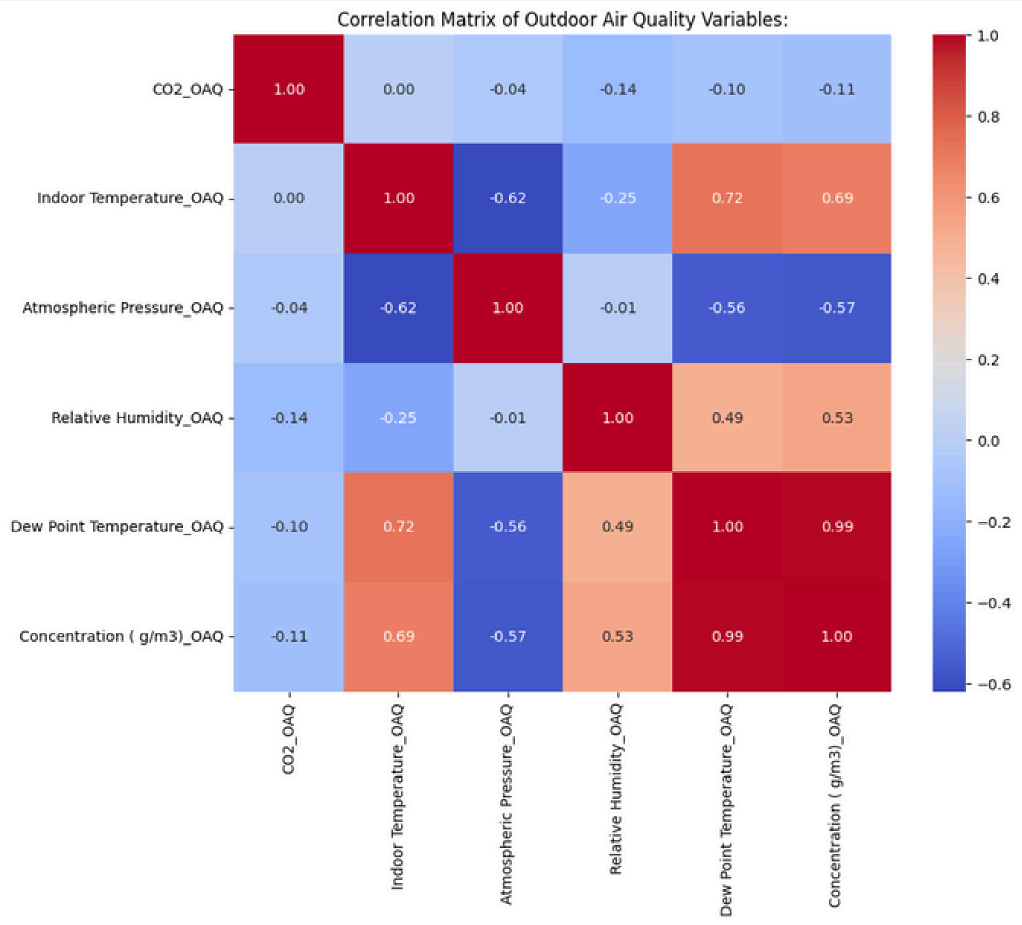
- There is near to perfect correlation (0.99) between AC Load and Total Energy, thus saying us that AC is the major contributor of total energy consumption, in the room the measurement was taken.
- A correlation coefficient of 0.99 indicates that the two variables tend to move together in the same direction. As one variable increases, the other variable also tends to increase, and vice versa.
- There is moderate (0.57) linear relationship between light+fan load and total energy. It implies that there is a noticeable but not exceptionally strong correlation between these variables. A correlation coefficient of 0.57 indicates that the two variables tend to move together in the same direction. As one variable increases, the other variable also tends to increase, and vice versa.
- No other significant positive or negative relationship was seen apart from these.



## Correlation matrix of Indoor Air Quality

From the correlation matrix of Indoor Air Quality, we can come to these conclusions :-

- CO2_IAQ has a strong positive correlation (0.54) with Dew Point Temperature_IAQ, indicating that higher CO2 levels indoors are associated with higher dew point temperatures.
- Outdoor Temperature_IAQ has a moderate negative correlation (-0.63) with Atmospheric Pressure_IAQ, suggesting that higher outdoor temperatures tend to occur with lower atmospheric pressure.
- Relative Humidity_IAQ has a moderate positive correlation (0.51 and 0.54) with Dew Point Temperature_IAQ and Concentration (g/m3)_IAQ, respectively. This implies that higher relative humidity is linked to higher dew point temperatures and higher concentrations of air pollutants.
- Atmospheric Pressure_IAQ has a strong negative correlation (-0.73) with Dew Point Temperature_IAQ, indicating that lower atmospheric pressure is associated with higher dew point temperatures.
- Dew Point Temperature_IAQ and Concentration (g/m3)_IAQ have a very strong positive correlation (0.99), suggesting that they are highly related and tend to increase or decrease together.



Correlation matrix of Outdoor Air Quality

From the correlation matrix of Outdoor Air Quality, we can come to these conclusions :-

1. CO2_OAQ has no correlation (0.00) with Indoor Temperature_OAQ, suggesting that outdoor CO2 levels are independent of indoor temperatures.
2. Indoor Temperature_OAQ has a strong positive correlation (0.72) with Dew Point Temperature_OAQ, indicating that higher indoor temperatures are associated with higher dew point temperatures outdoors.
3. Atmospheric Pressure_OAQ has a moderate negative correlation (-0.56 and -0.57) with Dew Point Temperature_OAQ and Concentration (g/m3)_OAQ, respectively. This implies that lower atmospheric pressure is linked to higher dew point temperatures and higher concentrations of outdoor air pollutants.
4. Relative Humidity_OAQ has a moderate positive correlation (0.49 and 0.53) with Dew Point Temperature_OAQ and Concentration (g/m3)_OAQ, respectively. This suggests that higher relative humidity is associated with higher dew point temperatures and higher concentrations of outdoor air pollutants.
5. Dew Point Temperature_OAQ and Concentration (g/m3)_OAQ have a very strong positive correlation (0.99), indicating that they are highly related and tend to increase or decrease together.

Based on these observations, we can infer that outdoor air quality variables, such as atmospheric pressure, relative humidity, dew point temperature, and pollutant concentrations, are interrelated and can influence each other. For example, lower atmospheric pressure can lead to higher dew point temperatures and higher pollutant concentrations. Additionally, indoor temperatures seem to have a strong influence on outdoor dew point temperatures, suggesting that indoor and outdoor environments are interconnected in terms of air quality.
These insights can be valuable for understanding the factors affecting outdoor air quality and developing strategies for monitoring and improving it, while also considering the potential impact of indoor environments.

## 2.3 INDEPENDENT AND DEPENDENT VARIABLES

### Dependent Variables:

- **CO2 levels:** The concentration of carbon dioxide in the air can be influenced by human activities such as ventilation practices, increased occupancy, using large number of electrical appliances for long times.
- **Indoor Temperature:** The temperature inside a building can be adjusted by humans through heating, cooling, and insulation systems.

### Independent Variables:

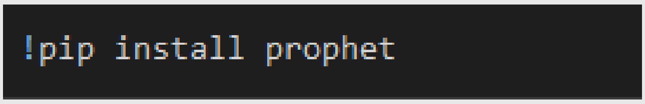
- **Atmospheric Pressure:** Atmospheric pressure is not directly controlled by human actions in a typical setting.
- **Relative Humidity:** While humidity levels can be influenced by human actions such as using humidifiers or dehumidifiers, they are often considered environmental factors that are not directly controlled by humans.
- **Dew Point Temperature:** The dew point temperature is a measure of humidity and is influenced by atmospheric conditions, but it is not directly controlled by human actions.
- **Concentration (g/m3):** This variable is not specific enough to determine its relationship with human control. It could represent various substances or pollutants, some of which may be influenced by human activities while others may not.

## 2.4 TIMESERIES FORECASTING OF ENERGY

For Timeseries forecasting, we have used Meta’s Prophet model. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

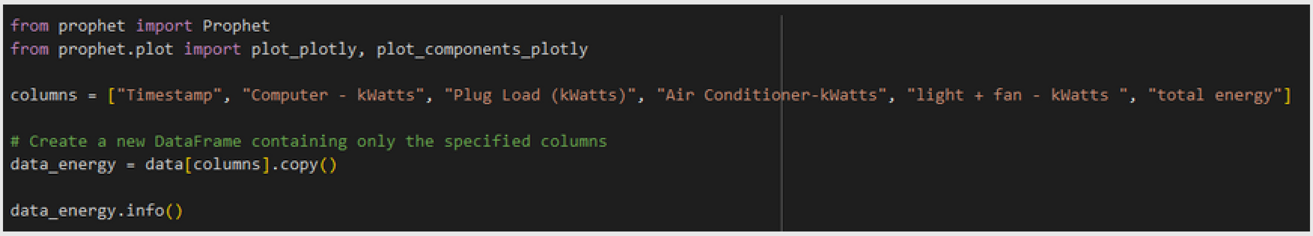
Prophet follows the sklearn model API. We create an instance of the Prophet class and then call its fit and predict methods.

The input to Prophet is always a dataframe with two columns: ds and y. The ds (datestamp) column should be of a format expected by Pandas, ideally YYYY-MM-DD for a date or YYYY-MM-DD HH:MM:SS for a timestamp. The y column must be numeric, and represents the measurement we wish to forecast.



```
!pip install prophet
```

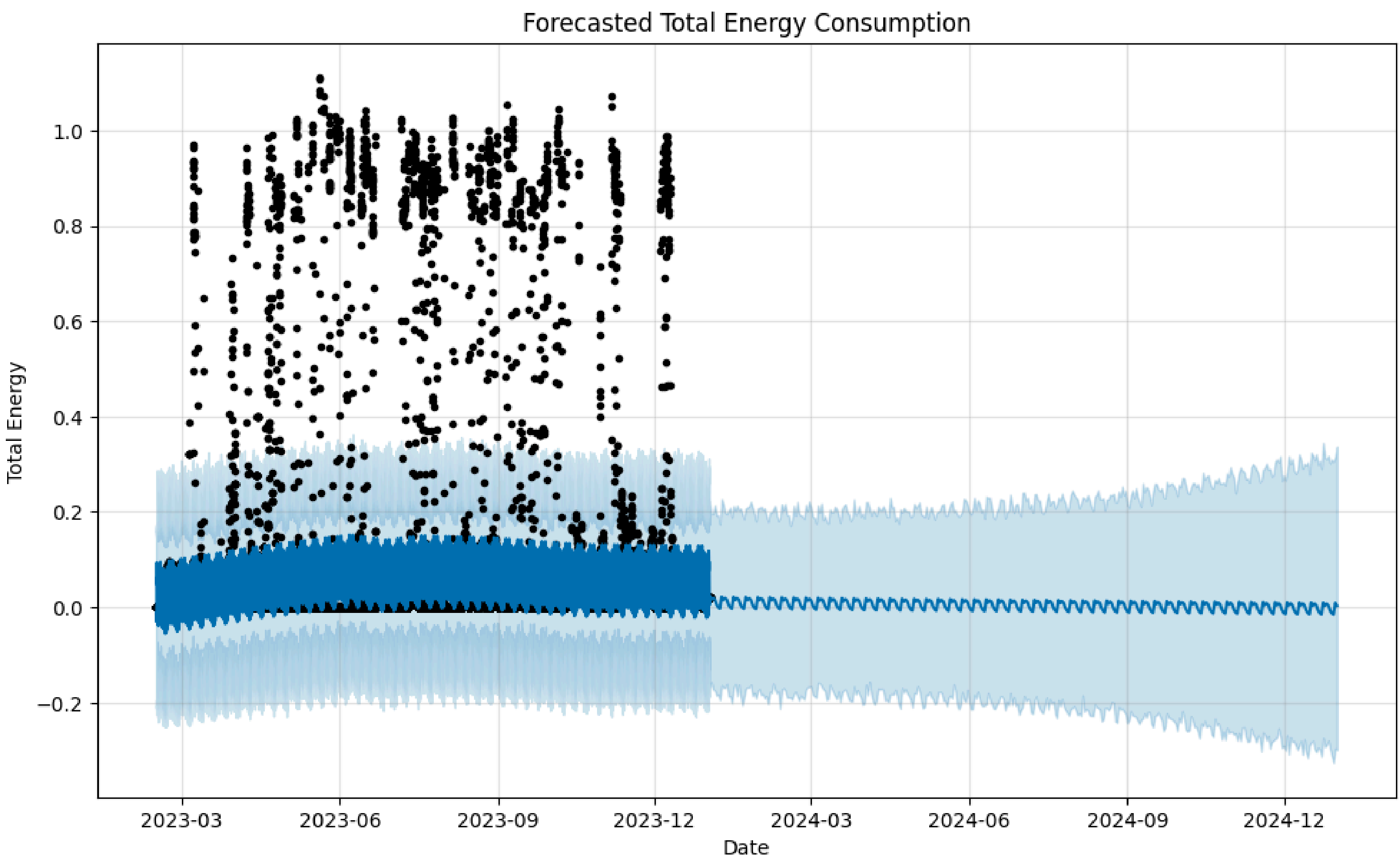
First we install prophet in our system using pip function



```
from prophet import Prophet
from prophet.plot import plot_plotly, plot_components_plotly

columns = ["Timestamp", "Computer - kWatts", "Plug Load (kWatts)", "Air Conditioner-kWatts", "light + fan - kWatts ", "total energy"]

# Create a new DataFrame containing only the specified columns
data_energy = data[columns].copy()

data_energy.info()
```

Then we start with importing the necessary components from the Prophet and Plotly libraries. Then, specify a list of column names related to energy consumption, such as timestamps, computer power usage, plug loads, air conditioner power consumption, lighting and fan power usage, and total energy consumption. Using this list of columns, we create a new DataFrame called data_energy by selecting only the relevant columns from the original data and making a copy of the selected data to ensure any modifications won't affect the original dataset. Finally, we print a summary of the data_energy DataFrame, providing information about the data types, non-null values, and memory usage.

# 2.4 INFERENCES SPECIFIC TO TIMESERIES FORECASTING



## Timeseries forecast of Total Energy vs Timestamp

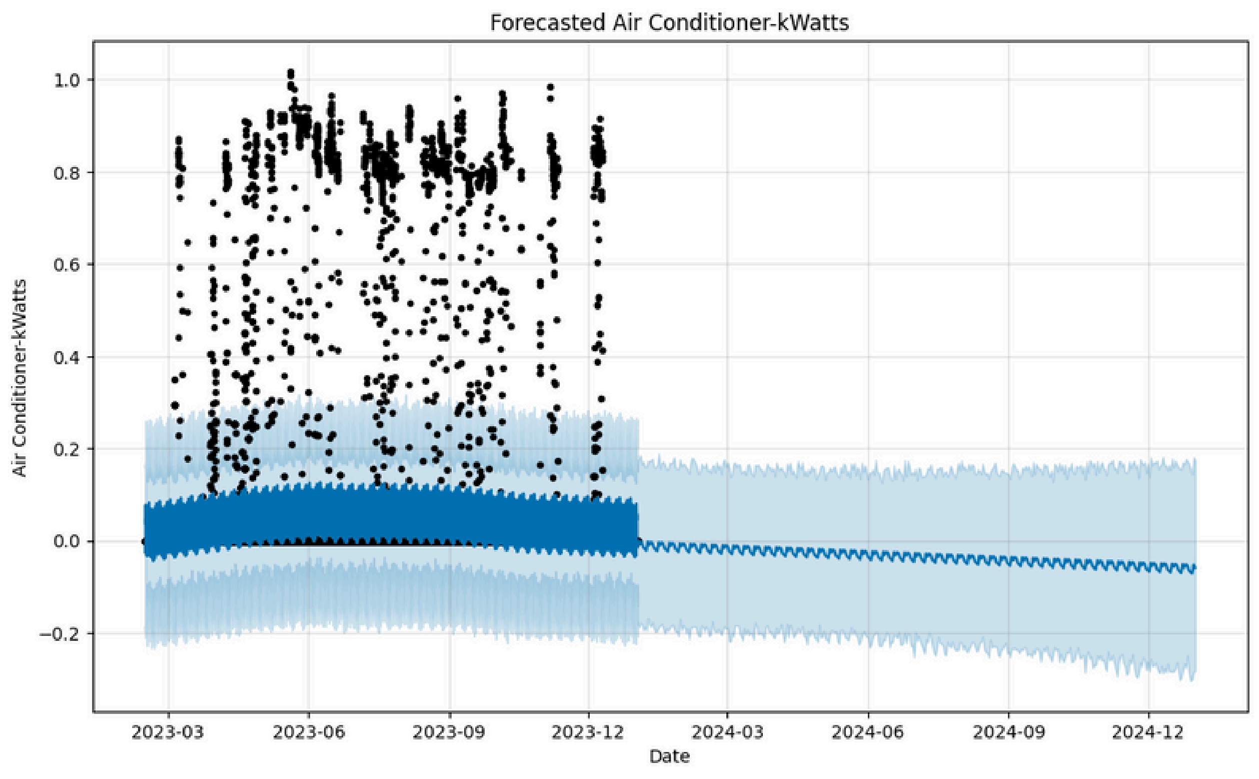
Here the x-axis represents the Timestamp (Date), ranging from March 2023 to December 2024, while the y-axis represents the total energy consumption.

The scattered black dots represent the historical data points or observations of total energy consumption. These data points seem to exhibit some seasonal patterns and fluctuations.

The blue shaded area represents the forecast of total energy consumption made by the Prophet model. The model has captured the seasonal patterns and trends present in the historical data and has projected them into the future. The blue line in the center of the shaded area represents the most likely forecast, while the shaded area around it represents the uncertainty or confidence intervals of the forecast.

Based on the forecast, it appears that the total energy consumption is expected to follow a cyclical pattern, with higher consumption during certain periods (likely warmer months) and lower consumption during other periods (likely cooler months). This cyclical pattern is consistent with the seasonal variations often observed in energy consumption due to factors like heating and cooling demands.

It's important to note that the forecast becomes more uncertain as it extends further into the future, as indicated by the widening of the shaded area (confidence interval) towards the end of the forecast period.



We have also done timeseries forecasting on AC Load, since it has almost perfect correlation with Total Energy. Based on the forecast, it is evident that air conditioner power consumption is expected to follow a cyclical pattern, with higher consumption during the warmer months (peaking around summer) and lower consumption during the cooler months. This pattern aligns with the typical behavior of air conditioning systems, which require more energy to maintain comfortable indoor temperatures during hotter periods.
The forecast becomes more uncertain as it extends further into the future, as indicated by the widening of the shaded area (confidence interval) towards the end of the forecast period.

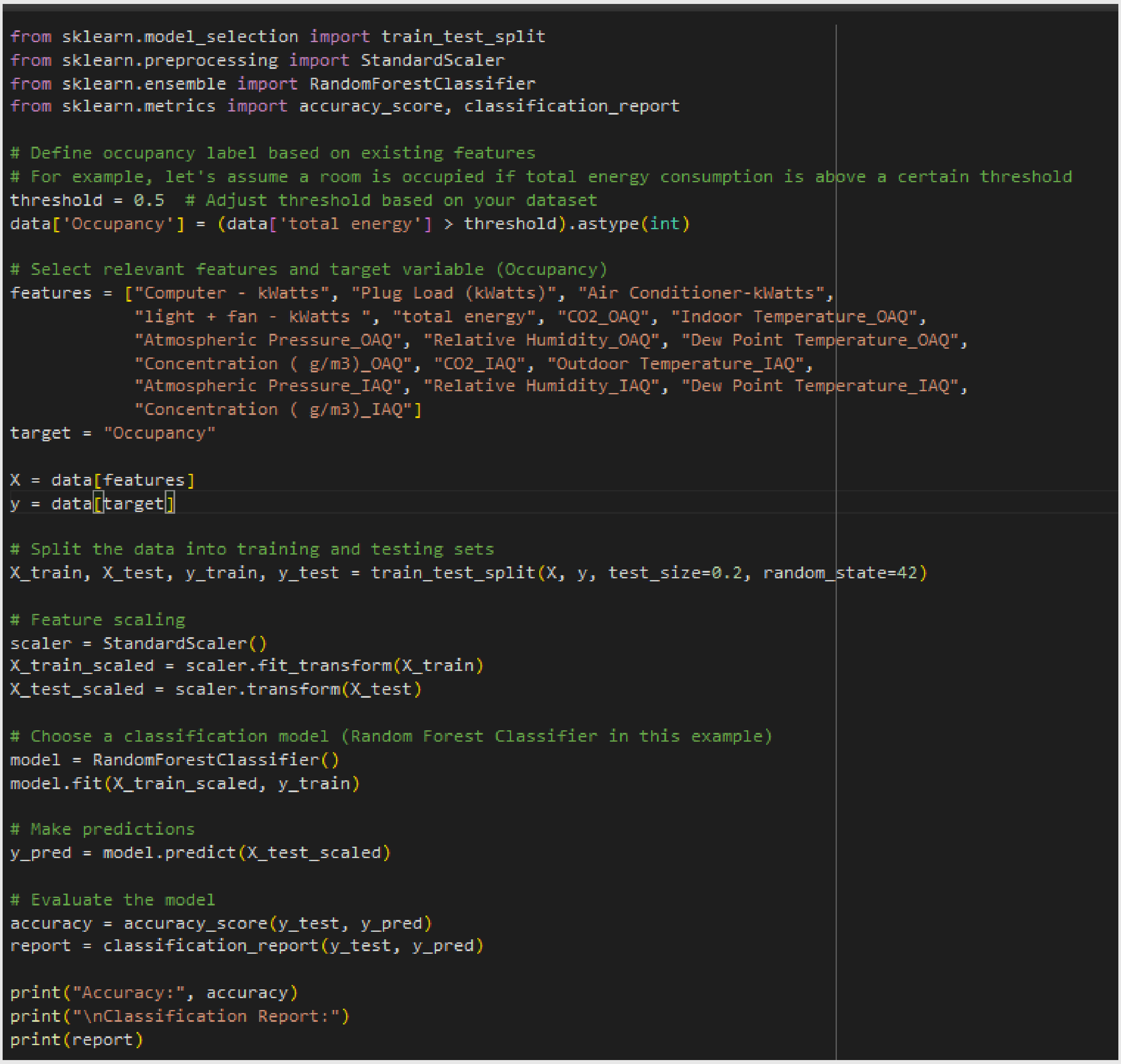
## 2.5 OCCUPANCY PREDICTION

For predicting occupancy, we have decided to take most important variable Total Energy. We have presumed if threshold is greater than 0.5, then there is a chance of Occupancy in the building. Since this is a classification problem, we will use standard classifier like Random Forest Classifier for predicting occupancy or not. 1 means Occupancy is present, and 0 means no occupancy.

### 2.5.1 INFERENCES SPECIFIC TO OCCUPANCY PREDICTION

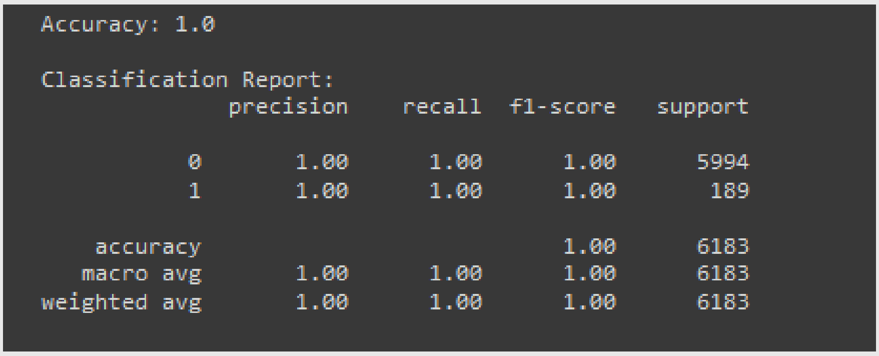
The whole process involves following steps :-

- **Importing Libraries:** The code imports necessary libraries from scikit-learn, including functions for splitting data (train_test_split), preprocessing (StandardScaler), building a random forest classifier (RandomForestClassifier), and evaluating model performance (accuracy_score, classification_report).
- **Defining Occupancy Label:** An occupancy label is defined based on existing features in the dataset. In this example, occupancy is determined by whether the total energy consumption exceeds a specified threshold (0.5). The 'Occupancy' column in the dataset is created, where a value of 1 indicates occupancy and 0 indicates non-occupancy.
- **Selecting Features and Target Variable:** Relevant features related to energy usage and IAQ are selected for model training, while the target variable 'Occupancy' is defined.
- **Splitting Data:** The dataset is split into training and testing sets using the train_test_split function. The training set comprises 80% of the data, while the testing set comprises the remaining 20%.
- **Feature Scaling:** The selected features are standardized using StandardScaler to ensure that all features have a mean of 0 and a standard deviation of 1. This step is crucial for models that rely on distance-based calculations.
- **Choosing and Training the Model:** A random forest classifier is chosen as the classification model. The model is trained on the scaled training data using the fit method.
- **Making Predictions:** The trained model is used to make predictions on the scaled testing data using the predict method.
- **Evaluating the Model:** The accuracy of the model is calculated using the accuracy_score function, which compares the predicted labels with the actual labels from the test set. Additionally, a detailed classification report is generated using the classification_report function, providing metrics such as precision, recall, and F1-score for each class.



```python
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report

# Define occupancy label based on existing features
# For example, let's assume a room is occupied if total energy consumption is above a certain threshold
threshold = 0.5  # Adjust threshold based on your dataset
data['Occupancy'] = (data['total energy'] > threshold).astype(int)

# Select relevant features and target variable (Occupancy)
features = ["Computer - kWatts", "Plug Load (kWatts)", "Air Conditioner-kWatts",
            "light + fan - kWatts ", "total energy", "CO2_OAQ", "Indoor Temperature_OAQ",
            "Atmospheric Pressure_OAQ", "Relative Humidity_OAQ", "Dew Point Temperature_OAQ",
            "Concentration ( g/m3)_OAQ", "CO2_IAQ", "Outdoor Temperature_IAQ",
            "Atmospheric Pressure_IAQ", "Relative Humidity_IAQ", "Dew Point Temperature_IAQ",
            "Concentration ( g/m3)_IAQ"]
target = "Occupancy"

X = data[features]
y = data[target]

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Feature scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Choose a classification model (Random Forest Classifier in this example)
model = RandomForestClassifier()
model.fit(X_train_scaled, y_train)

# Make predictions
y_pred = model.predict(X_test_scaled)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

print("Accuracy:", accuracy)
print("\nClassification Report:")
print(report)
```

## 2.6 MODEL ACCURACY



```
Accuracy: 1.0

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      5994
           1       1.00      1.00      1.00       189

    accuracy                           1.00      6183
   macro avg       1.00      1.00      1.00      6183
weighted avg       1.00      1.00      1.00      6183
```

Random Forest Classifier has done an excellent work in predicting occupancy, It has 100% value in all important parameters like Accuracy, Precision Score, Recall score, F1-score. I talso has correctly predicted when occupant is there (1) and when no occupant is there (0). Overall, we are very satisfies with the result.

# 3 LIMITATIONS WITH STUDY AND SCOPE OF FUTURE IMPROVEMENTS

## Limitations:

- **Data Quality and Availability:** The quality and availability of data can significantly impact the performance of the forecasting model. Incomplete, inconsistent, or noisy data may lead to inaccurate predictions.
- **Feature Engineering:** The selection and engineering of relevant features play a crucial role in model performance. Limited domain knowledge or oversight in feature selection may result in suboptimal models.
- Model Complexity: More complex models may lead to overfitting, especially when dealing with limited data. Balancing model complexity with generalization capability is essential for reliable predictions.
- **Interpretability:** Some advanced models, such as ensemble methods or deep learning models, may lack interpretability, making it challenging to understand the underlying factors driving the predictions.
- **Generalization:** Models trained on data from a specific office room may not generalize well to other environments with different characteristics. Robustness across various office settings is essential for broader applicability.
- **External Factors:** External factors such as seasonal variations, holidays, or unexpected events (e.g., equipment malfunction) may influence energy usage and occupancy patterns but are not explicitly captured in the model.

## Future Scope of Improvement:

- **Incorporating External Data:** Integrating additional external data sources (e.g., weather data, building occupancy schedules) can enhance the model's predictive accuracy and robustness.
- **Advanced Feature Engineering:** Exploring advanced feature engineering techniques, such as time-series decomposition, lag features, or rolling statistics, can capture complex temporal patterns and dependencies more effectively.
- **Ensemble Methods:** Employing ensemble methods, such as model stacking or boosting, can combine the strengths of multiple models to improve overall performance and stability.
- **Anomaly Detection:** Integrating anomaly detection algorithms can identify abnormal energy usage or occupancy patterns, enabling proactive maintenance or intervention.

- **Real-time Monitoring and Feedback:** Implementing a real-time monitoring system that continuously updates the model with new data and provides feedback can ensure adaptability to changing conditions and improve forecasting accuracy.
- **User Feedback and Collaboration:** Incorporating feedback from end-users or facility managers can provide valuable insights into model performance and guide iterative improvements.
- **Model Explainability:** Investing in techniques for model explainability, such as SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations), can enhance transparency and trust in the forecasting results.

Addressing these limitations and exploring the suggested areas for improvement can lead to more accurate and robust time-series forecasting models for energy usage and occupancy prediction in office environments.

## 4 CONCLUSION

In this ASSIGNMENT, we embarked on a journey exploring the development of a predictive model for occupancy prediction in an office environment, leveraging a diverse array of techniques and methodologies. Our exploration began with data preprocessing steps, including handling missing values, converting categorical data, and feature engineering. We then delved into the construction of a time-series forecasting model using IAQ and energy data, emphasizing the significance of feature selection, model selection, and evaluation metrics.

Throughout our journey, we encountered various challenges and opportunities for improvement. We addressed issues such as data quality, model interpretability, and the balance between model complexity and generalization. We explored techniques for handling imbalanced datasets, standardizing features, and selecting appropriate evaluation metrics to gauge model performance accurately.

Furthermore, we discussed the implications of correlation coefficients in understanding the relationships between variables and the factors influencing CO2 levels in office spaces. We examined the limitations of our approach, including data availability, model interpretability, and the potential for overfitting. Additionally, we identified future avenues for improvement, such as incorporating external data sources, enhancing feature engineering techniques, and implementing real-time monitoring systems.

## REFERENCES

https://colab.research.google.com/drive/1ruiftu6roX4wtlPm7y_ylige8mAXbZGZ?usp=sharing

https://facebook.github.io/prophet/docs/seasonality,_holiday_effects,_and_regressors.html#modeling-holidays-and-special-events

https://www.tableau.com/learn/articles/what-is-data-cleaning

https://www.onlinemanipal.com/blogs/popular-regression-algorithms-in-machine-learning

https://youtu.be/2vF2xTUXJwM?si=L7ZhYCuvt8hWS1Pv

https://youtu.be/DxtDh0cvTTk?si=s4VWCdENyskoX3Pv