Improve Building Sustainability with data

A Dissertation submitted for the degree of Master of Science in

Data Analysis for Business Intelligence

By

Srijeet Mohan Nair

229001998

Supervisor: Dr. Grechuk Bogdan

School of Computing & Mathematical Sciences

University of Leicester

England

October 2023

Abstract

This study investigates how to fully harness the potential of data in the quest of sustainable urban development. The effort, which focuses on constructing sustainability, operates inside the framework of the 'SmartViz' platform, a complete web based platform perfected over a two-decade period.

Finding useful insights that advance building sustainability is the main objective of this study. Temperature, Occupancy, CO2 Emissions, and Humidity are the four primary factors that the inquiry focuses on. By identifying substantial correlations between these variables for building data and its most frequently utilised rooms, which are heavily combined with all four metrics, we want to pave the way for a more sustainable urban landscape.

Step by step, the project is developed. It starts by carefully examining the correlation matrix with the goal of discovering undiscovered connections and interactions between the specified components in the "geometry" the most frequently utilised rooms of the building. The groundwork for the following stage, in which we explore the field of time series forecasting, is laid out by this preliminary inquiry. To obtain the desired information and conclusions, we apply forecasting models, which are well known for their ability to predict future trends. With this the target stakeholders will be better able to allocate resources, make informed decisions, and encourage environmentally responsible behaviour with the use of this information.

Finally, this study contributes to the growth of practical construction processes. The findings of the project provides a detailed information about the correlations between the parameters or metrices and optimal forecasting models for each of the metrics so that sustainable planning can be done in a better way in future. Also, to enhance the understandability interactive visualisation graphs are created along with an interactive dashboard-like report. All these would help us to lead the way to a future in which buildings are more sustainable when CO2 emissions, temperature, occupancy, and humidity are considered. With the use of data analysis techniques the urban environment’s development can be improved in more efficient, sustainable, and environmentally friendly way in the future.

Keywords: sustainability, metrics, geometry, correlations, forecasting

Acknowledgements

In this last phase of my master's programme in Data Analysis for Business Intelligence, I would like to express my gratitude to everyone who contributed to the accomplishment of my final research project, "Improving Building Sustainability with data". It has proved really rewarding. My entry into the field of real-time tracked and filtered data analysis, and forecasting has been defined by it, and I am really appreciative of the chance.

I'm grateful to “University of Leicester” and “SmartViz Ltd” for giving me this opportunity. I would want to express my sincere gratitude to Mr. Vishal Sharma. This study attempt has been greatly influenced by your advice, encouragement, and faith in my talents.

I also want to express my heartfelt thank you to my academic supervisor Dr. Grechuk Bogdan. This master's thesis has been improved because of your advice, well-informed recommendations, and dedication to perfection.

Dr. Andrey Morozov, thank you for allowing me to collaborate with “SmartViz Ltd” on this project work. The execution and outcome of the initiative has depended heavily on your assistance in creating this relationship.

A heartfelt 'thank you' to Mrs. Ambika V (my mother), Mrs. Sridevi Nambiar (my sister) and my friends for their support and motivational encouragement, even though they were thousands of miles away. Your belief in me has always motivated me. My success in the field of data-driven programming is directly attributed to them, as well as my lecturers and personal tutors, for their dedication to teaching and fostering a great learning atmosphere.

Declaration

I hereby declare that the contents written in this dissertation work are based on my own knowledge, understanding, and analysis. Where other ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity, and I have not misrepresented, fabricated, or falsified any idea, data, fact, or source in my submission. I understand that any violation of the above will be cause for disciplinary action by the university..

Name: Srijeet Mohan Nair

Student ID: 229001998

Signature: 

Date: 1st October 2023

Table of Contents

List of Figures

List of Tables

1. Introduction

Environmental preservation is essential in today's society. Buildings require a lot of electricity, which uses up pollutants and is bad for the environment. Structures, however, could also support our attempts to live more sustainably.

The main focus of this thesis is on using data and knowledge to improve structures over time. This study's primary objective is to employ a unique tool called SmartViz, which has been created over a 20-year period. Similar to a clever computer programme, SmartViz can improve the efficiency of towns, schools, buildings, and even entire cities. In order to ensure that everything proceeds smoothly and effectively, it accomplishes this by analysing data, generating forecasts, and keeping a close check on events as they unfold.

1.1 Contextual Background

The worldwide push for sustainability has forced a fundamental reassessment regarding how we design, build, and run buildings. Besides just being made of concrete and brick, the present-day structures are evolving ecosystems that require ongoing monitoring and adjustment in order to achieve sustainability goals. Stakeholders can negotiate this challenging environment using the ever-evolving platform provided by SmartViz (Figure 1), a monument to human creative thinking and technical innovation. Planners, developers, and operators may all benefit from its user-friendly interface, which includes engaging 3D visualisations and gives them crucial insights into how well their constructed surroundings are doing. SmartViz has developed into an essential tool in the quest of sustainable building practises by enabling the capacity to examine how buildings operate under various situations.[1]

Figure 1 SmartViz Make Buildings Smart

1.2 Problem Statement

How can data analytics be used to light the road to sustainable building is the fundamental issue at the centre of this project. The issue description is brief but significant. We want to identify interesting correlations between four crucial parameters or metrics temperature, occupancy, CO2 emissions, and humidity within the boundaries of each of the buildings represented in our dataset. Additionally, we make an effort to identify the rooms in these buildings that are most often utilised and match the requirements for these four metrics. With these findings, we want to build machine learning models that can predict these metrics accurately enough to direct sustainable enhancements.

In order to assure the integrity and quality of our dataset, the first stages of this project require thorough data cleansing. In order to reveal the complex interactions between temperature, occupancy, CO2 emissions, and humidity, we then build a correlation matrix. This early study can, however, point up anomalies and a lack of significant relationships, which would call for advance or deep data cleaning. After deep cleaning and correlation analysis for each building's geometry, we separate our dataset between weekday and weekend data in order to conduct a more thorough examination. To understand the context of my analysis within the scope of actual work patterns, we adhere our partitioning approach to the working hours that majority of IT companies in UK follow.

1.3 Objectives

Several major goals serve as the foundation for my thesis:

- To pinpoint the internal rooms or zones or structures of each buildings that meet the required criteria of being associated or dealing with the 4 parameters or metrices ‘temperature, occupancy, CO2 emissions, and humidity.’
- To identify meaningful relationships between 4 metrics ‘temperature, occupancy, CO2 emissions, and humidity’ for each of the building’s geometry (internal rooms/zones/structures)

- To create forecasting models for each of the four metrics 'CO2, temperature, occupancy, and humidity', within the geometry of each building while taking into consideration the time-series structure of the data.
- To assess and contrast the effectiveness of several time series forecasting models in order to choose the best strategy for precise forecasts.

1.4 Significance

This project's importance cannot be emphasised. Buildings have an indisputable role as energy users and greenhouse gas emitters in a society that struggles with the urgent need to slow down global warming. The research done in the context of data analytics has the potential to completely change how we think about sustainable construction. This study may directly influence decisions about building design, CO2 release and occupancy patterns, by revealing beneficial findings through correlation analysis and future forecasting.

The results of this study go beyond the realm of academics. They have the potential to direct current procedures in city development, managing buildings, and construction. The data-driven approaches described in this research can help sustainability projects, whether in new construction or retrofitting old facilities. Additionally, the decrease in energy usage and greenhouse gas emissions that results from making educated decisions may considerably support global sustainability goals.

1.5 Methodology

This analysis work follow the following processes;

- Data collection and preprocessing:

  To guarantee the accuracy and dependability of the data, we carefully gather and preprocess it. Procedures for cleaning data are used to get rid of inconsistencies in data.
- Correlation Analysis:

  For each of the building's geometry, to investigate the connections between the parameters that we have, we use methods of statistical analysis. To represent and analyse these correlations, a correlation matrix is created.
- Dataset segmentation:

In order to better understand connection patterns between the 4 parameters for each of the building's geometry in scrutinized manner, we divide the dataset into weekday and weekend data. This segmentation is in line with the actual working hours in UK.

- Development of Machine Learning Models:

  Taking into account the data which we have segregated from the original dataset we design machine learning techniques for predicting or forecasting the four critical metrics within the geometry of each building and also as a whole building.

- Model Assessment:

  To determine the most precise and optimal forecasting model for predicting the four critical metrics within the geometry of each building and also as a whole building, we thoroughly assess the performance of our forecasting models.

1.6 Scope

Our research explores how urban growth, technological advancement, and sustainability are all connected, with a specific focus on data analytics and predictive modelling for sustainable building. Our discoveries have an influence on cities and institutions in addition to buildings, despite the fact that our main concentration is on buildings. Our objective is to use data analytics and modelling to promote sustainable urban development. We want to offer useful insights for increasing sustainability in the framework of construction via thorough study.

2. Literature Review

2.1. Machine learning Time series models

Machine learning is a domain in computer science where automated computers are given the capacity to learn from data, recognise and follow trends, and form predictions or judgements. For these jobs, no further programming is needed.

In machine learning, a time series dataset can be forecasted using time series forecasting models, which are specialised methods where temporal dependencies and data trends are taken into consideration. The frequently used models are elaborated below. In order to create forecasts based on past data patterns, forecasting based on time is essential. This technique supports sustainable strategy plan and well informed decision making.

- **ARIMA** (Autoregressive Integrated Moving Average)

  The time series forecasting technique known as "Autoregressive Integrated Moving Average, or ARIMA, is used to identify and forecast trends in data that has been acquired over a period of time."[1][2] "ARIMA takes into account the associations between lagged data values and residual error values from moving average models"[1] in order to forecast future values based on historical data patterns. "ARIMA (p, d, q) is a commonly used abbreviation for ARIMA models[1][2]".

  " $\Phi(B)\nabla^{\mathrm{d}} y_t = \Theta(B)\varepsilon_t$ "

  Figure 2 ARIMA(p,d,q) equation
  Source "Application of SARIMA model in forecasting and analysing inpatient cases of acute mountain sickness from BMC Public Health" [3]

- **SARIMA** (Seasonal Autoregressive Integrated Moving Average)

  "A variation of the ARIMA model called SARIMA, or Seasonal Autoregressive Integrated Moving Average, was developed to handle time series data with both trends and seasonality"[1].When working with time series data that displays regular patterns at precise intervals, such as daily, monthly, or yearly seasonality, this approach is especially helpful with an aim to make the predictions accurately. "The typical designation for

ARIMA models is a combination of non-seasonal ARIMA (p, d, q) and seasonal ARIMA(P, D, Q)S"[3].

$$\nabla^{\mathrm{d}}\nabla_S^D y_t = \frac{\Theta(B)\times\Theta_S(B)}{\Phi(B)\times\Phi_S(B)}\varepsilon_t$$

Figure 3 SARIMA(P, D, Q, p, d, q, S) equation
Source "Application of SARIMA model in forecasting and analysing inpatient cases of acute mountain sickness from BMC Public Health" [3]

- **LSTM** (Long Short Term Memory)

In the field of time series analysis, the LSTM model, is an excellent representation of "deep learning"[1]. Although it has its foundations in the "design of recurrent neural networks (RNNs)" [1], it also "incorporates a crucial innovation: a compact memory component" [1]. "An input gate, an output gate, a forget gate, and a cell make up the basic building blocks of an LSTM unit" [1]. The below Figure 4, Figure 5, Figure 6 are the "formula or equation for the gates:[4]"

$$\text{Forgetting value } f_t = \sigma\left[W_f(h_t-1, X_t) + b_f\right]$$

Figure 4 Forget gate value equation for LSTM

$$\text{Input gate value } i_t = \sigma\left(W_i \cdot [h_t-1, X_t] + bi\right)$$

Figure 5 Input gate value equation for LSTM

$$\text{Updated value } C_t = \tanh\left(WC \cdot [h_t-1, X_t] + bC\right)$$

$$C_t = f_t[C_t-1+i_t]e$$

Figure 6 Obtained cell equation for LSTM
Source for the above 3 formulas "Time Series Forecasting using LSTM and ARIMA from (IJACSA) International Journal of Advanced Computer Science and Applications" [4]

The outstanding ability of LSTM to acquire and store information over lengthy periods sets it apart in the field of time series. This distinctive feature makes LSTM a good option for analysing, categorising, and predicting time series data. The strength of LSTM comes in its capacity to identify both short-term variations and long-term patterns within data that is sequential, making it a vital resource for time series modelling and

analysis whether used for financial market forecasts, forecasting the weather, or medical diagnostics.

2.2. Machine learning Regression models

Regression analysis identifies relationships between the target variable and variables that are independent, in order to forecast outcomes for continuous variables. The number, form of the independent variables, and also the shape of the regression line, change depending on the type of regression technique used. One method that is frequently used is:

- **Linear regression**

  "A statistical technique used to examine the connection between a dependent variable (Y) and one or more independent variables (X)" [5][6]. The key step in this technique is "calculating the error term e which requires fitting an even straight line" [5] to the data being analysed. The "Least Square Method" [5] is frequently used to attain the "goal of finding the line that fits the data the best"[5].

  Equation "$Y = \beta_0 + \beta_1 X + e$ describes a simple linear connection between the defined variables. Here, $\beta_0 =$ intercept, $\beta_1 =$ slope and $e =$ error term" [5][6]. The aforementioned data may be used to construct metrics like the MSE (Mean Square Error), MAE (Mean Absolute Error), and R squared value. However, the "R squared value of a linear regression analysis is more useful than the MSE and MAE values."[7] "The optimum value for a R squared value is 1.0, and a R squared score always falls between 0 and 1"[8] but depending on the data that is being worked on the range criteria is not much of a concern.

"$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{m} (X_i - Y_i)^2}{\sum_{i=1}^{m} (\bar{Y} - Y_i)^2}$$"

Figure 7 R squared score formula Linear regression analysis

Source "The coefficient of determination R squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation from National Library of Medicine" [7]

2.3. Performance Metrics

The word "performance metrics" is crucial and is frequently used when discussing machine learning models. By taking into consideration both anticipated values and recorded values, it is primarily used to examine the generated model's accuracy and precision in terms of prediction. When calculating the predicting functionality performance of a model for time series data, a number of widely used metrics may be taken into account. The following are the most frequently utilised or taken into consideration:

- "Mean Absolute Error (MAE)" [9]
- "Mean Square Error (MSE)" [9]
- "Root Mean Square Error (RMSE)" [9][10]
- "R squared score"[8][11]

2.4. Sustainable buildings

Worldwide, the proportion of people who "live in urban regions is thought to be greater than 50%"[12]. Even though they are buzzing enterprises hub, these towns and localities put an unnecessary burden on the environment. "With the constant growth of carbon dioxide emissions, largely due to human activity, the world's environmental issues have gotten worse" [12][13]. In the year 2012, nations such as "India, the United States of America, China, and the European Union collectively produced a staggering 35.6 billion tonnes of carbon dioxide emissions" [12] from buildings. The buildings are notable since they "account for 40% of worldwide energy demand in metropolitan areas and are the largest man-made structures"[14]. In order to lessen their negative consequences on the environment, financial markets, and society around the world, reducing carbon dioxide emissions and optimising energy use in buildings are crucial challenges that must be tackled. When the aforementioned challenges are met and a sustainable development plan is implemented, the following benefits can be observed.

- Environmental aspect

Sustainable buildings, often known as "green buildings" [12], are created with the intention of "reducing the amount of carbon dioxide emissions and optimising energy consumption in buildings. This aids with mitigating climate change and protecting the environment"[15].

- Economic factor

  Commercially speaking, green buildings are naturally more valuable due to their ability of reducing the amount of carbon dioxide emissions and optimising energy consumption in buildings, and these give a lasting financial advantages over the course of their life cycles" [16].

- Efficient resources utilisation

  Sustainable buildings promote the "use of recyclable building materials that are easily accessible locally, which lowers costs and promotes more sustainable resource use"[17].

- Advantages of Remodelling

  Energy usage and the carbon dioxide emissions could be "significantly handled and decreased by implementing green principles while renovating the existing structures"[12][18].

With the incentive to "overcome the constraints caused by urbanisation, sustainable building development is critical as it contributes to the formation of more resilient cities and communities"[12][17][18].

3. Methodology

In order to increase sustainability across each buildings and its internal room or zone or section it owns, the methodology used here will use techniques from data analysis to discover insights and evaluate the best predictive model for each parameter or metric CO2, Occupancy, Humidity, and Temperature, for each building's geometry room. These extensive methods are used in this project's workflow in Figure 8, which follows a pattern in sequential order:

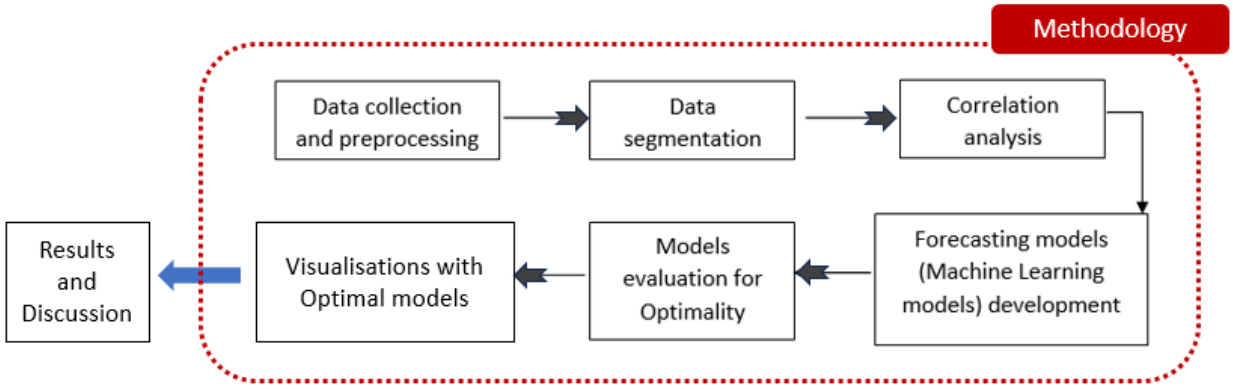


Figure 8 Methodology workflow

This methodological procedure allows for the collection of significant insights from the detailed sensor data as well as the development of optimised models for the forecasting of important parameters. Utilising advanced statistical analysis for assisting programmes centred around data which optimise the utilisation of energy, lower CO2 emissions, and support sustainable development is made possible by following this sequential procedure. The precise order of the processes is intended to provide a thorough examination that progresses from an understanding of fundamental concepts to authorization and evaluations for sustainable activities.

3.1. Data collection and preprocessing

For a successful analysis the main prerequisites are the datasets. These datasets are formed by data collection which means gathering the data and transforming it into a usable and readable format. In our case of data collection, SmartViz provided us with the dataset (.xlsx file) initially which contained the real-time IoT censored data. In granularity, it contained information about building names, metric names (parameters), time, metric values

recorded during that time period and the geometry name (internal room structure of the building) which is the internal room or zone or section this data record was observed. Refer Figure 9 for initial buildings dataset's structure.

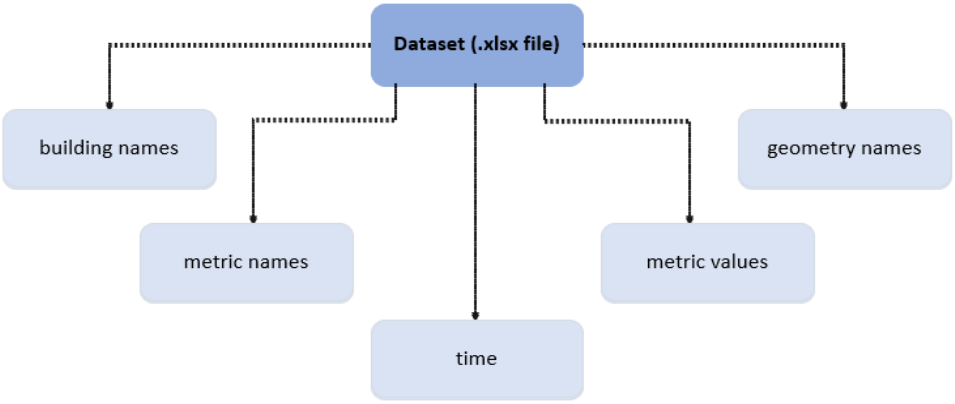


Figure 9 Buildings dataset initial structure

The next step after having the dataset in our hand is data preprocessing. In the overall process of data analysis and machine learning, data preprocessing is an important step which involves data cleaning, data splitting, data transforming, organizing the data in proper format etc., in order to make our data suitable for analysis and predictive modelling. Out of all these, the most important or crucial step is data cleaning.

- Data cleaning:

  The process of finding and fixing or eliminating errors, inconsistencies, and inaccuracies in a dataset is known as data cleaning. Initially in the dataset provided there were some duplicate rows where the building records were exactly the same and also some missing value rows where one or more than one particular column's value were missing from the row. So, to handle such inconsistencies, cleansing the data is vital. After initially cleaning the data (after initial clean nomenclature for the datasets), still some discrepancies related to the metric or parameter values and its order were observed due to which deep cleaning was performed (before deep clean nomenclature for the datasets) and again, after that the metric benchmark values were provided, the row's with the metric values excluding the metric benchmark value were removed (After deep clean nomenclature for the datasets). Henceforth, continuous, and rigorous data cleaning process was performed. With deep data cleaning the data quality and integrity

is maintained with which the performance of the analysis or forecasting models would be better. Refer Figure 10 for buildings dataset's hierarchy after data cleaning phase.

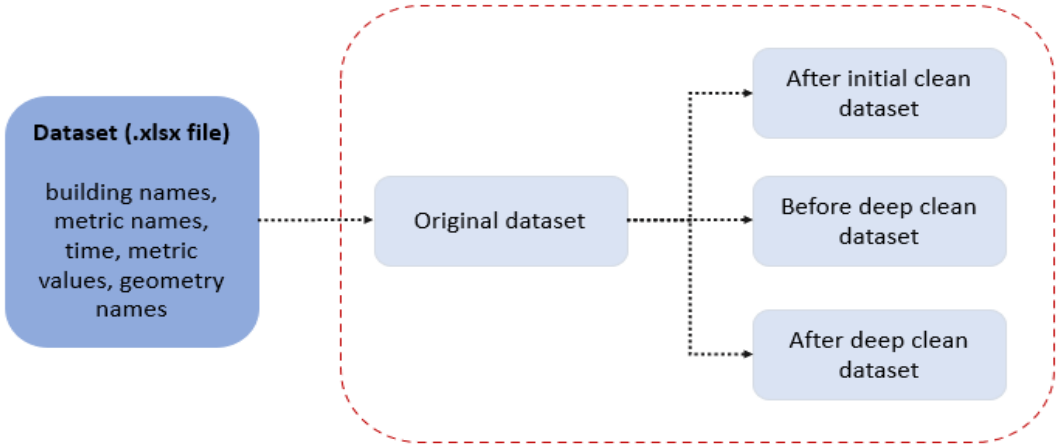


Figure 10 Dataset hierarchy after data cleaning process

3.2. Data segmentation

A key component of our process, data segmentation, is the rigorous segmentation of a dataset into different subsets according to predetermined criteria and qualities. Our project data underwent a methodical segmentation procedure after a rigorous round of data cleaning (initial clean, before deep clean(complete clean), after deep clean) to assure data correctness and consistency. With this, the datasets comprises of 4 categories based on cleansing, original dataset(1), after initial clean dataset(2), before deep clean dataset(3) and after deep clean dataset(4). So, whatever segmentation is carried out of the building's data it will have these cleansed form of datasets (total 4). Keeping this in mind the project data was segmented or partitioned or divided into;

- Individual buildings segmentation:

  Data divided into many parts with each segment matching to a different building name. We were able to focus in on the particular features of each specific building in this stage.

- Building's geometry room segmentation:

  We deepened our research by exploring the details of the dataset for each building. Here, we separated the data into groups based on the distinctive geometry designations, which stand in for the names of interior building rooms. This segmentation

was very important since the rooms included in it had all the important metrics or parameters supplied in the dataset considered.

- Weekday and Weekend hours segmentation:

  We added another segmentation to our exploratory study to further our understanding and get a more complete picture. This time, we concentrated on particular time periods, such as weekday peak hours, weekday off peak hours, and weekend hours. By analysing the relationships and interconnections between these metrics and parameters for the geometrical rooms of each building during various time periods, this method allowed us to broaden the area of our analytical investigation. The ultimate objective of our methodical process was to derive deeper and more significant correlations, which would act as a basis for the succeeding phase of methodological workflow.

Refer Figure 11 for buildings dataset's hierarchy after data segmentation phase.

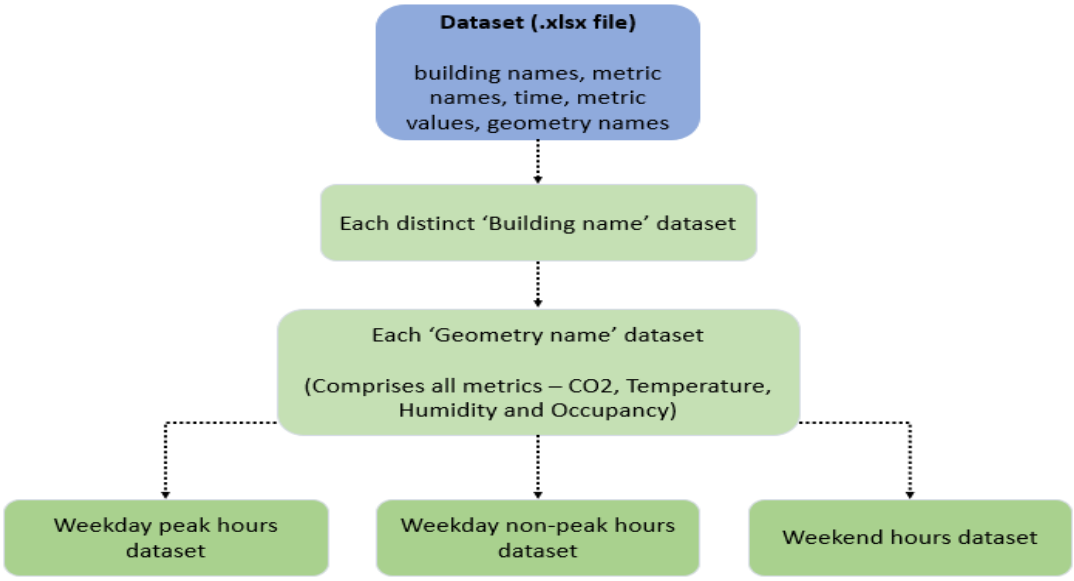


Figure 11 Dataset hierarchy after data segmentation process

3.3. Correlation Analysis

A statistical method called correlation analysis is used to measure the strength and direction of the linear relationship existing between two or more variables in a dataset. It is crucial because it help in understanding the connections between the variables, that would help in terms of analysis or modelling and making well informed decisions.

- Correlation matrix:

We perform a thorough correlation analysis by creating correlation matrix across various datasets, including the original and the cleansed datasets for each building's geometry rooms as well as datasets specific to weekday peak hours, weekday non-peak hours, and weekend hours, for improving building sustainability through data analysis. Finding significant connections between the various metrics or parameters is the main goal. For each geometrical room in a building, there are numerous cleaned datasets that have been created (after initial clean, before deep clean and after deep clean). As a result, it might be difficult to choose which dataset provides the most exact and reliable correlation insights.

- Correlation significance check:

  We perform a correlation significance test with simple linear regression statistics. With this task we intend to choose the dataset that most accurately depicts the correlations between the parameters. Ultimately, we want to make sure that the selected datasets correctly represent and convey the correlation results. This activity is carried out for each of the building's geometrical rooms and each of its time period data. By using this approach of significance check decisions can be made that will optimise resource use, advance sustainability, and create more environmentally friendly structures.

3.4. Forecasting models development

Forecasting defined as speculating about future occurrences or trends based on analysis of past data, is essential to decision making. In order for businesses to make wise decisions, allocate resources effectively, plan for the future, reduce risks, and assess the effects of plans, forecasting models are crucial. In order to produce these models, which predict future values or trends for each metric defining the geometry of each buildings, mathematical, statistical, or machine learning approaches must be used. Forecasting models are crucial instruments for directing activities and maximising outcomes, whether in business, finance, or sustainability programmes. The main processes and factors for developing forecasting models are:

- Data Preparation:

After doing correlation analysis, make sure that the chosen dataset is properly preprocessed and cleaned for the geometry of each building as well as for the entire dataset. This entails managing missing and duplicate values, keeping metric or parameter records in the right sequence, eliminating inconsistent data, and ensuring that no records omit the metric's benchmark value.

- Model Selection and Development:

  Choosing the most suitable forecasting models based on the type of data. We built the most typical and commonly used time series forecasting models for each metric or parameter for each building's geometry because the data in our instance is time series data, with a brief description of the models provided in the **section 2.1.**

- Data splitting:

  Dividing the dataset into two separate subsets, the training set, and the test set. The training set was mainly utilised to develop our forecasting models and perform date forecasting in the future. The test set, on the other hand, made it possible to compare actual values with our model's predictions, which was a critical step in determining how accurate our models were.

Forecasting each statistic for future dates was done with the training set and the forecasting model that had been created. This procedure includes creating forecasts of each measure for each building's geometry as well as for the entire building for future dates using historical data, which is an essential step in our sustainability research. The test set was used to compare actual and predicted values for each and every metric for every section or zone or geometry of the building.

3.5. Forecasting models evaluation for optimality

In this process, various forecasting models are meticulously evaluated to determine which is best for each metric or parameter within each building's geometry and for the building as a whole. To achieve this, comparing and evaluating the predicted outcomes of the forecasting model to the test set's actual values in order to assess its predictive performance. The developed models are assessed using efficiency metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and a Composite Score.

- Mean Squared Error (MSE)

  "The average squared discrepancies between the actual values recorded ($Actual_{values}$) and the expected values ($Predicted_{values}$)."[9]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Predicted_{values} - Actual_{values})^2$$

- Mean Absolute Error (MAE)

  "The average of the absolute difference between the values that were predicted ($Predicted_{values}$) and the values that were actually recorded ($Actual_{values}$)." [9]

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(Predicted_{values} - Actual_{values})|$$

- Root Mean Squared Error (RMSE)

  "The square root of the MSE, or the root mean square error (RMSE), calculates the approximate value of the error's standard deviation." [9]

$$RMSE = \sqrt{MSE}$$

- Composite Score

  The Composite Score, which is a performance statistic taken into account for accuracy purpose, is a combination metric that takes into account the MSE, MAE, and RMSE values to offer an overall rating. Assigned their relative importance to one another, these metrics can be summed together and assigned various weights.

$$Comp_{score} = weight_1 * MSE + weight_2 * MAE + weight_3 * RMSE$$

  where, the weights for MSE is $weight_1$, MAE is $weight_2$ and RMSE is $weight_3$.

  The best model for each metric or parameter, both at the geometrical level of a particular building and for the entire building can be found by computing the aforementioned performance metrics for several forecasting models and comparing their values. The model that has the best forecasting precision for the targeted predicting task is often regarded as the most optimum one since it has the lowest MSE, MAE, RMSE, and Composite Score.

3.6. Data Visualisations with Optimal Forecasting Models

An important aspect of the project is data visualisation, where we integrated Power BI interactive reports with the best forecasting model for each parameter inside the geometry of each building and the building as an entire entity. With this we can visually observe the forecasts for each metric inside the geometry of each building as well as for the entire structure.

This data visualisation method has 3 sections:

- Actual vs. Predicted values graphs:

  These graphs display the actual versus predicted values for each of the metric. The graphs are plotted using the optimal forecasting model for that metric in that particular building. It also provides an easy way to evaluate the model's performance and analyse how accurate our predictions are.

- Future Date Forecasting Graphs:

  These graphs display future date forecasting which enables us to foresee trends and patterns. This may be then used as an effective resource for resource allocation in buildings and proactively planning for sustainability.

- Interactive report (dashboard like report):

  An interactive report is also created using Power BI to improve accessibility and user friendliness. This report gives customers the ability to easily choose the data for the particular building they want to forecast. Users can browse the report and select any metric or parameter or structure for which they wish to view forecasts. This adaptability includes forecasts for both the overall structure and each particular building's geometry, providing a holistic perspective of sustainable and environmentally friendly practises.

4. Data Collection and Preprocessing

The process of gathering the data for the analysis purpose and then transforming (if required) it for into a more readable format is called Data Collection.

4.1. Data Source

The initial data source was provided by SmartViz (the project stakeholders). The format of the data source is a Microsoft Excel file. It contains the real-time IoT censored data that was recorded for each of the parameters in each of the buildings.

The excel file data source (X Buildings temp, co2, humidity, Occupancy Raw Data until 20230712.xlsx) provided had 5 columns in it:

1. Building names (project_name)

   This column describes the name of the buildings that were recorded.

   4 distinct buildings were observed: XCudworth, XWestgate, XTownhall, and XWorsbrough.

2. Metric names (metric_name)

   This column describes the parameters or metrics that were recorded for each of the buildings.

   4 distinct metrics were observed: CO2 ,Humidity, Occupancy, and Temperature

3. Time (time)

   This column describes the time at which the data was collected and recorded.

4. Metric values (value)

   This column describes the observed value for the metric for that particular building

   Each metric had a particular benchmark values set. This cut-off values were also shared by SmartViz. Table 1 conveys the same.

Metrics	Metric value benchmark
CO2	0.999 → 2000
Humidity	0 → 100
Occupancy	0 → 100

Temperature	9.999 → 29.999

Table 1 Metrics benchmark values

5. Geometry Names (geometry_name)

   This column describes the building's internal room name or zone name or section name for which each metric value were recorded.

   The 4 distinct buildings which were observed had many geometry rooms in it but proceeding ahead with the analysis we only consider those geometry that deals with or is associated or involved with the 4 metrics or parameters that was considered.

   So, now the geometry room's name for each buildings is given in the below Table 2.

Building names	Geometry room names
XCudworth	1st Floor Meeting Room 3 1st Floor Open Office 2nd Floor Meeting Room 1 2nd Floor Open Office
XTownhall	1st Floor Meeting Room 1 1st Floor Meeting Room 2
XWorsbrough	1st Floor Office 1 1st Floor Office 2
XWestgate	Level 1 Open Office Level 2 Open Office Level 3 Open Office Level 4 Open Office Level 5 Open Office

Table 2 Building's Geometry rooms

The below Figure 12 is a sample of the aforementioned information about the dataset columns.

project_name	metric_name	time	value	geometry_name
XTownhall	co2	2023-06-14 00:00:16.346827+00	998	1st Floor Meeting Room 1_iaq
XTownhall	temp	2023-06-14 00:00:16.346827+00	23	1st Floor Meeting Room 1_iaq
XTownhall	humidity	2023-06-14 00:00:16.346827+00	41	1st Floor Meeting Room 1_iaq
XCudworth	co2	2023-06-14 00:00:20.953786+00	604	2nd Floor Meeting Room 1_iaq
XCudworth	temp	2023-06-14 00:00:20.953786+00	25	2nd Floor Meeting Room 1_iaq
XCudworth	humidity	2023-06-14 00:00:20.953786+00	41	2nd Floor Meeting Room 1_iaq
XWestgate	co2	2023-06-14 00:00:41.186002+00	711	Level 3 Open Office_iaq

Figure 12 Initial Dataset information overview

4.2. Data Cleaning

The next step to do in data analysis once the data is provided is to perform data preprocessing, with data cleaning being the most crucial out of them all. Data Cleaning is the activity performed to remove duplicate data and missing values, handling outliers to maintain data integrity i.e.., consistent and complete.

The initial data cleaning was performed on the original dataset using Python programming language. This had 1,617 rows removed from a total of 5,30,164 data records that brought the data records tally down to 5,28,547. With this we segregated 2 datasets – Original dataset(5,30,164 data records) and After initial clean dataset(5,28,547 data records).

But, after performing this cleaning and analysing the datasets, some inconsistencies in the data were still observed. These inconsistencies were:

- Redundant data – the data records that excludes the parameters value range benchmark (outliers detection)

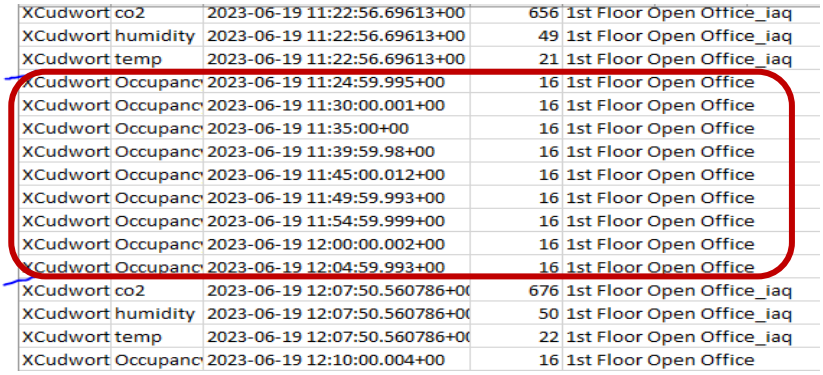
  Ex: Temperature would never be negative but there were few records where temperature within a building's geometry room (XCudworth – 1st Floor Open Office) was negative, so such anomalies needed to be handled

  SQL script was developed in SQL Server Management Studio (SSMS) to handle this data redundancy. *(Source code in **Appendix 1**)*
- Inconsistent data – the data records that experienced a sudden growth or spike after following a certain trend (outliers).

  Ex: the sudden spike or data value growth observed while moving in a steady rate like occupancy recorded 0 at one time and next 70 and then back to 0 which means the record that gave 70 might mislead us and affect our result.

Using mean average functionality in Microsoft Excel of the previous 2 data records value the spiked data value was replaced in-order to avoid or tackle this problem.

- Occupancy parameter that were repeating or duplicating which was making it redundant.



XCudwort	co2	2023-06-19 11:22:56.69613+00	656	1st Floor Open Office_iaq
XCudwort	humidity	2023-06-19 11:22:56.69613+00	49	1st Floor Open Office_iaq
XCudwort	temp	2023-06-19 11:22:56.69613+00	21	1st Floor Open Office_iaq
XCudwort	Occupanc	2023-06-19 11:24:59.995+00	16	1st Floor Open Office
XCudwort	Occupanc	2023-06-19 11:30:00.001+00	16	1st Floor Open Office
XCudwort	Occupanc	2023-06-19 11:35:00+00	16	1st Floor Open Office
XCudwort	Occupanc	2023-06-19 11:39:59.98+00	16	1st Floor Open Office
XCudwort	Occupanc	2023-06-19 11:45:00.012+00	16	1st Floor Open Office
XCudwort	Occupanc	2023-06-19 11:49:59.993+00	16	1st Floor Open Office
XCudwort	Occupanc	2023-06-19 11:54:59.999+00	16	1st Floor Open Office
XCudwort	Occupanc	2023-06-19 12:00:00.002+00	16	1st Floor Open Office
XCudwort	Occupanc	2023-06-19 12:04:59.993+00	16	1st Floor Open Office
XCudwort	co2	2023-06-19 12:07:50.560786+0(	676	1st Floor Open Office_iaq
XCudwort	humidity	2023-06-19 12:07:50.560786+0(	50	1st Floor Open Office_iaq
XCudwort	temp	2023-06-19 12:07:50.560786+0(	22	1st Floor Open Office_iaq
XCudwort	Occupanc	2023-06-19 12:10:00.004+00	16	1st Floor Open Office

Figure 13 Repeating data for Occupancy metric

Refer the Figure 13 above for the same, the building here is XCudworth and 1st Floor Open Office geometry room name.

- The reason for this inconsistency observed maybe due to the fact that the occupancy metric censored data timestamp would be different when compared with other metrics. As other metric observed censored data timestamp difference of 5mins throughout, the occupancy metric showed 1min, 2min and even 3min of censored data timestamp difference between 2 records for the considered buildings.
- The other reason for this would be because in the data provided for each building's meeting rooms geometry only occupancy metric are recorded and for the office rooms geometry both _iaq (humidity, temperature, and CO2) and occupancy parameters are recorded so when we consider the entire buildings data including all the geometry room names in it, the occupancy metric count is more.

SQL script was developed in SQL Server Management Studio (SSMS) to handle this. A lead function was used, that whenever a data for occupancy is recorded then only the data for CO2 can be recorded as any presence within the room only leads to CO2 emissions, and hence further humidity and temperature would be recorded. So, the data record order should be of Occupancy followed by other 3 parameters. *(Source code in **Appendix 2**)*

- The recorded data for the buildings are not in a particular 'metrics' order.

  If CO2, humidity, and temperature are recorded but occupancy is not recorded before it then there is no use of that data for us to achieve our result of better correlation between the parameters and therefore remove such records (Figure 14) as well. Here, the building is XCudworth and the geometry room name is 1st Floor Open Office.

XCudwort	co2	2023-06-14 09:16:44.598386+00	604	1st Floor Open Office_iaq
XCudwort	temp	2023-06-14 09:16:44.598386+00	23	1st Floor Open Office_iaq
XCudwort	humidity	2023-06-14 09:16:44.598386+00	44	1st Floor Open Office_iaq
XCudwort	co2	2023-06-14 13:56:49.894771+00	889	1st Floor Open Office_iaq
XCudwort	temp	2023-06-14 13:56:49.894771+00	24	1st Floor Open Office_iaq
XCudwort	humidity	2023-06-14 13:56:49.894771+00	39	1st Floor Open Office_iaq

Figure 14 Data records not in a proper metrics order.

  SQL script was developed in SQL Server Management Studio (SSMS) to handle this, where we used a lag function. *(Source code in **Appendix 3**)*

By carrying out deep cleaning and handling these additional inconsistencies or redundancies in the data after initial clean, 2 more datasets – Before deep clean dataset(3,62,612 data records) and After deep clean dataset(3,60,098 data records) were considered for the buildings correlation analysis between the parameters.

In Before deep clean datasets, the deep cleaning was carried out on the buildings data where the additional inconsistencies mentioned above was removed excluding one activity which is removing the data records that excludes the parameters value range records, whereas in After deep clean datasets, complete deep cleaning was carried out on the buildings data where the aforementioned all the data related inconsistencies were handled. Therefore, the sequential order of the cleansed data between the 4 datasets is Original → After initial clean → Before deep clean → After deep clean.

The below Table 3 gives an overview of the rigorous data cleaning process carried out on the buildings data as a whole and each generated dataset's data records count after cleaning.

Buildings data (All 4 buildings)	Datasets	Data records count	Data records removed
	Original	530164	-
	After initial clean	528547	1617

	Before deep clean	362612	165935
	After deep clean	360098	2514

Table 3 Summary of the buildings dataset's data record count

It can be seen from the below Figure 15, Figure 16, Figure 17, Figure 18 when each metric is taken into consideration, the buildings data after carrying out deep cleaning looks more clear, easy to understand, consistent and complete when compared to the original or actual buildings data.

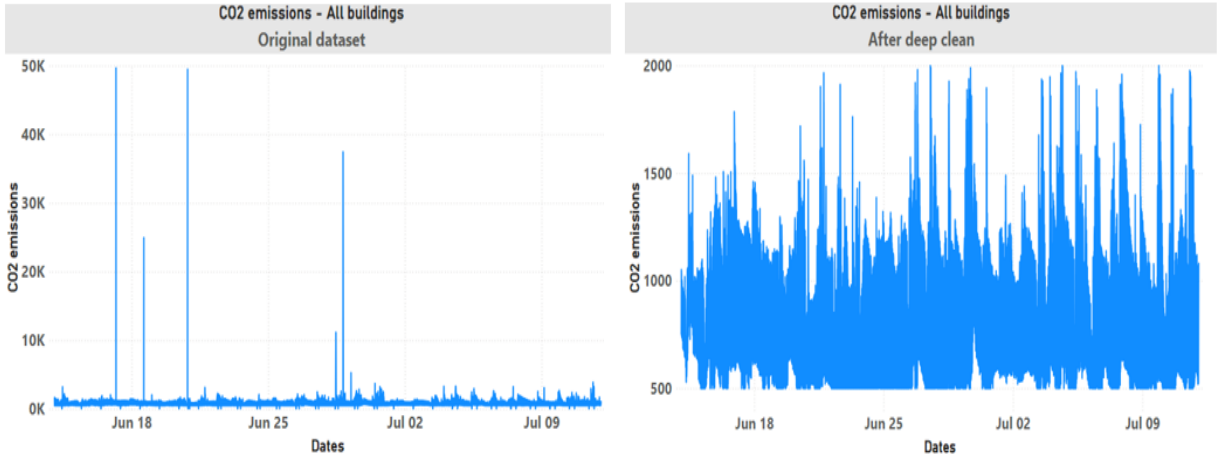


Figure 15 Original dataset VS After deep clean for Buildings CO2 emissions values

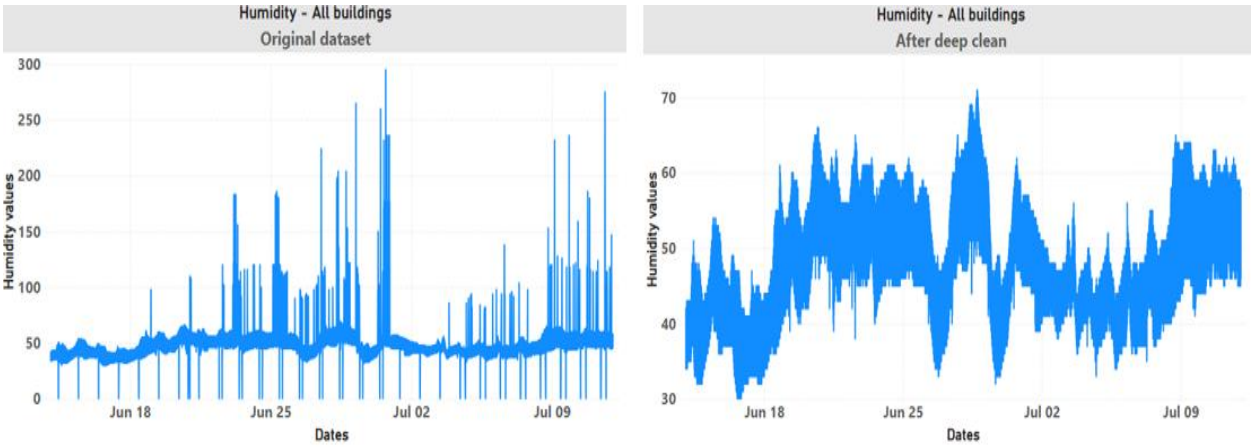


Figure 16 Original dataset VS After deep clean for Buildings Humidity values

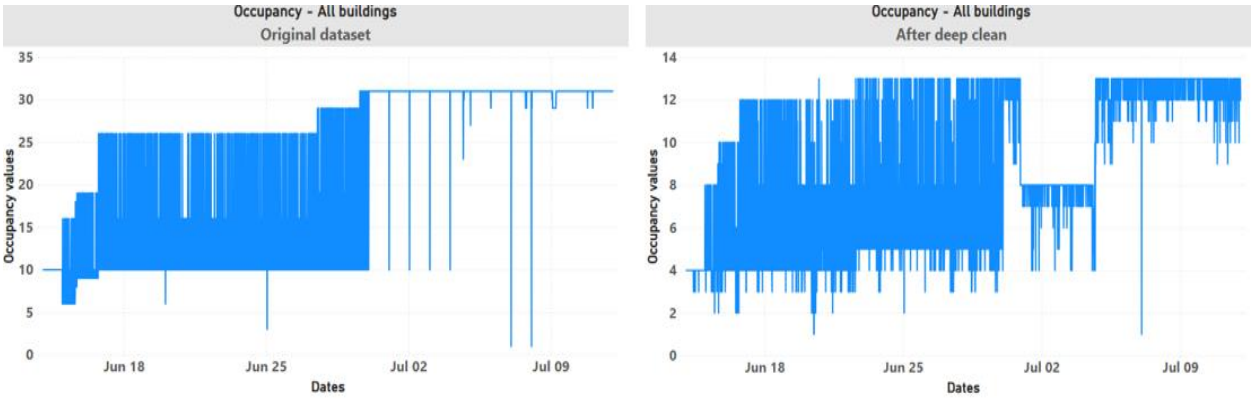


Figure 17 Original dataset VS After deep clean for Buildings Occupancy values

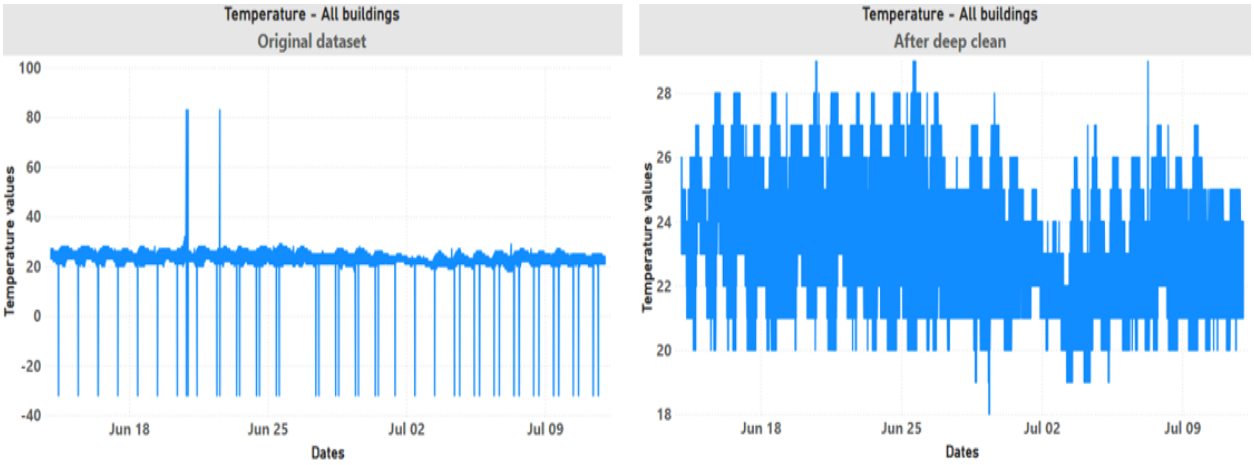


Figure 18 Original dataset VS After deep clean for Buildings Temperature values

Figure 19 portrays the dataset hierarchy after gathering the data and performing rigorous cleaning (preprocessing) in a more scrutinized manner.

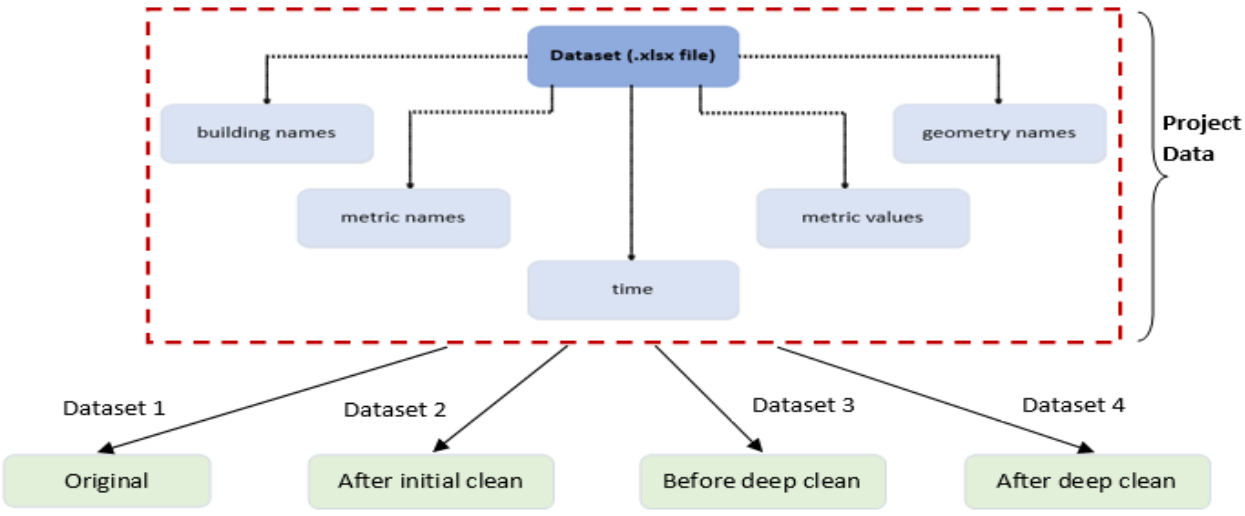


Figure 19. Project Buildings Dataset hierarchy after data collection and cleaning in more scrutinized way

5. Data Segmentation

Data segmentation is defined as the precise splitting of a dataset into numerous subsets considering predetermined criteria and qualities as the key components of the approach. Each segmented dataset in the project we were working on included the Original dataset, After initial clean dataset, Before deep clean dataset, and After deep clean dataset.

5.1. Individual buildings segmentation

The dataset is split into four segments, each of which corresponds to a distinct building name. The Original dataset, After initial clean dataset, Before deep clean dataset, and After deep clean dataset are all parts of the segmented or partitioned datasets, as was before indicated. This split is shown in Figure 20.

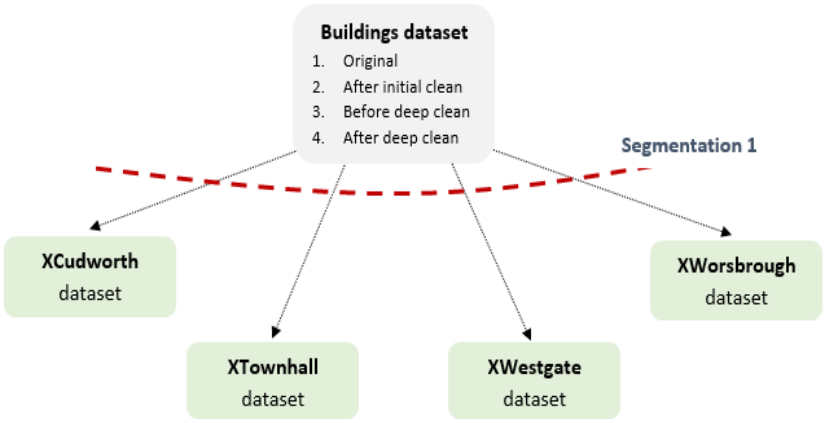


Figure 20 Datasets for each distinct building

5.2. Building's geometry room segmentation

By investigating each detail of the dataset for each building, we broadened our investigation. Here, as visible from Figure 21 we divided the data into categories based on the specific geometric designations, which serve as names for building interior rooms. Due to the fact that the segmentation's rooms took into account all the crucial metrics or parameters provided in the dataset, it was extremely vital.

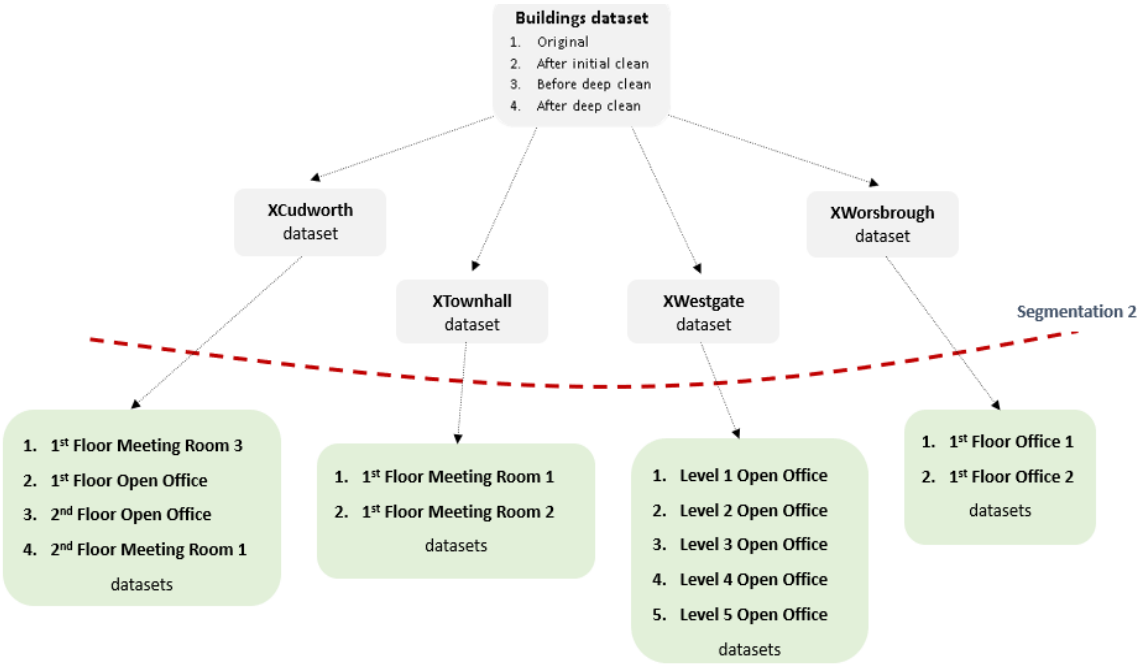


Figure 21 Datasets for each of the distinct building's geometry rooms

5.3. Weekday and Weekend hours segmentation

We added more segmentation to our exploratory study to broaden our perspective and present a more complete picture (Figure 22). Specifically, we concentrated on weekday peak hours, weekday off peak hours, and weekend hours this time. This strategy enabled us to expand the scope of our analytical inquiry by examining the associations and connections between the metrics CO2, Humidity, Occupancy and Temperature for each of the geometrical rooms of each building across various periods in time.

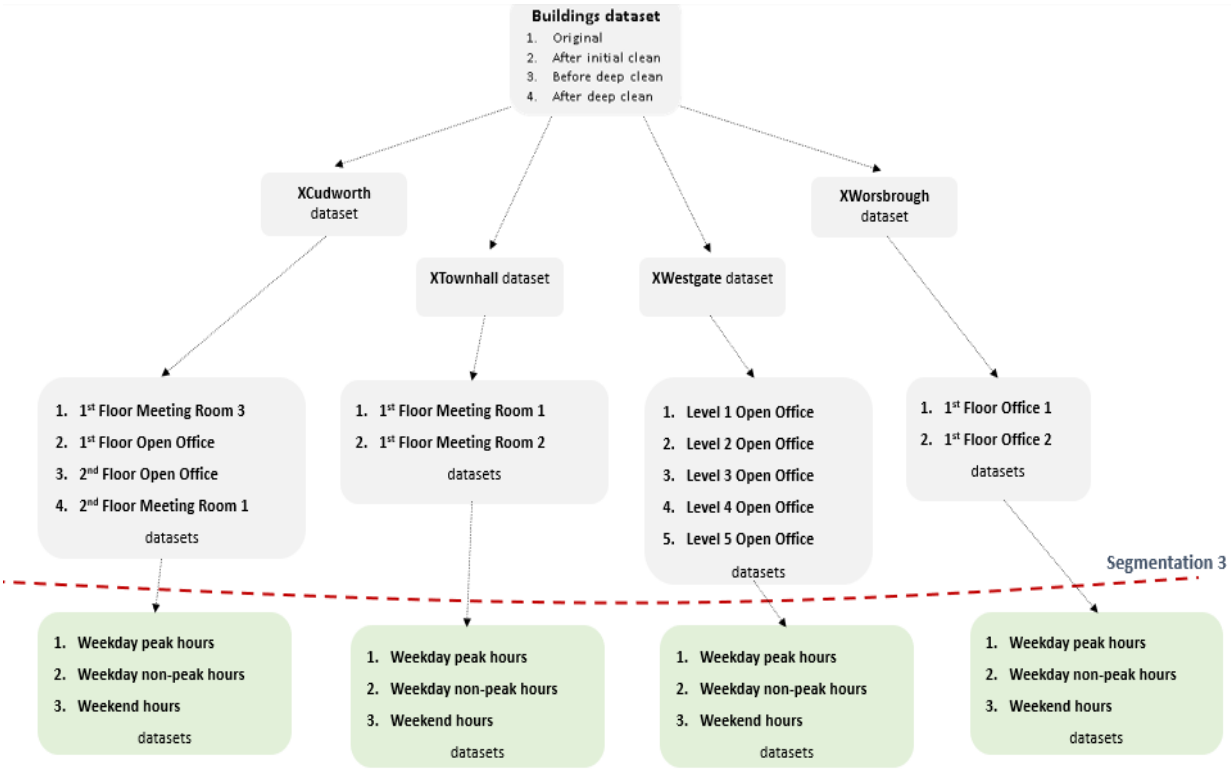


Figure 22 Weekday and Weekend hours datasets for each of the building's geometry room names

With the help of data segmentation, we delve into the smallest of details in the building's data and created several partitions with the goal of drawing more detailed and significant correlations among the metrics or the parameters CO2, Humidity, Occupancy, and Temperature within each of the geometrical rooms of each and every building.

6. Correlation Analysis

Correlation analysis is a process where we intend to find meaningful correlations between the metrics or parameters CO2, Humidity, Occupancy and Temperature(Figure 23). This is performed within each of the building's geometry room as a whole and also for its segregated weekday and weekend data outlined already in Figure 22.

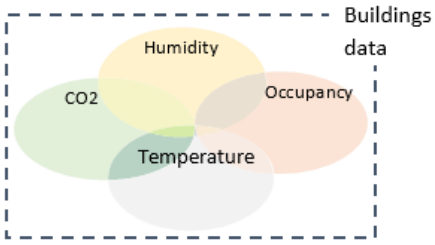


Figure 23 Metrics correlation analysis for buildings data

6.1. Correlation Matrix

"To express the correlation coefficient between two variables, parameters, or metrics (in our case), a square matrix known as a correlation matrix is used."[19] Using "Python programming language, correlation matrix is created to examine how strong the metrics CO2, Humidity, Occupancy and Temperature are associated and also whether it is positively or negatively associated." [20] For each of the building's geometry room dataset (Figure 21) and also each building's geometry room weekday and weekend hours dataset (Figure 22) which are weekday peak hours, weekday non-peak hours and weekend hours, correlation matrix is created for a significant analysis in order to identify the connections between the CO2, Humidity, Occupancy and Temperature metrics.

As we have several cleansed datasets (Original, After initial clean, Before deep clean, After deep clean) with us for each building, the correlation matrix is created for each one of them. One such example of correlation matrix for analysis is illustrated in Figure 24, Figure 25, Figure 26, Figure 27. The consider building is XCudworth with the geometry room name 1st Floor Open Office.

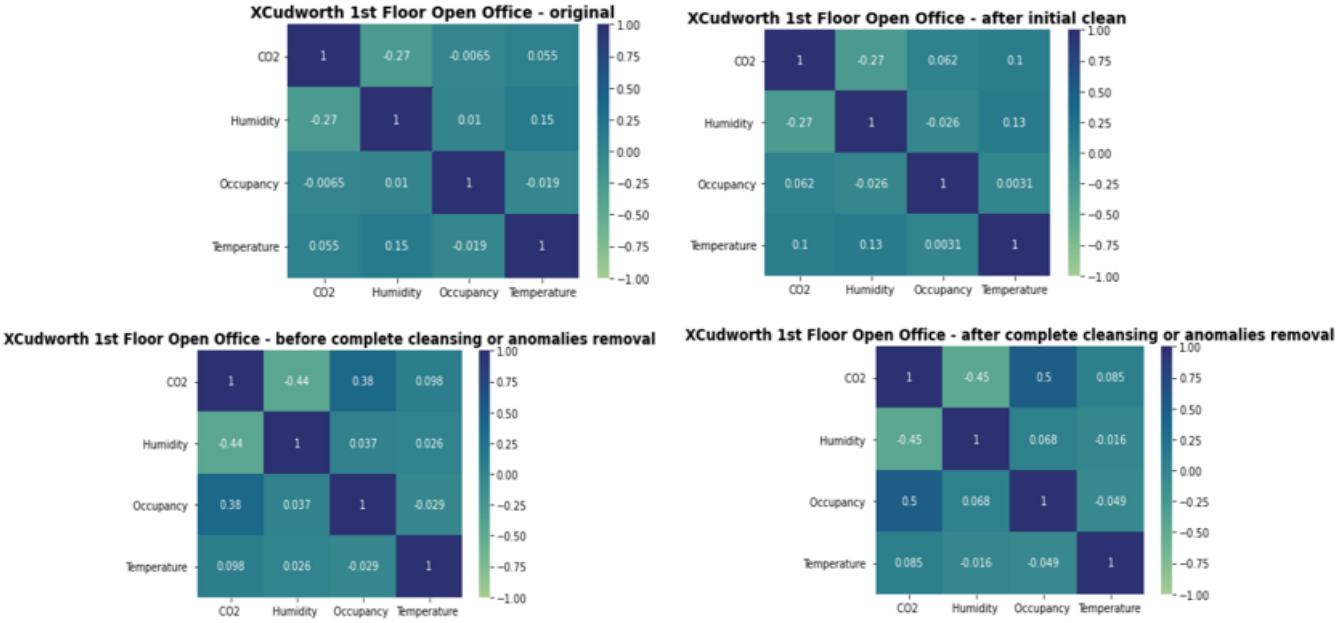


Figure 24 Correlation matrix for XCudworth 1st Floor Open Office

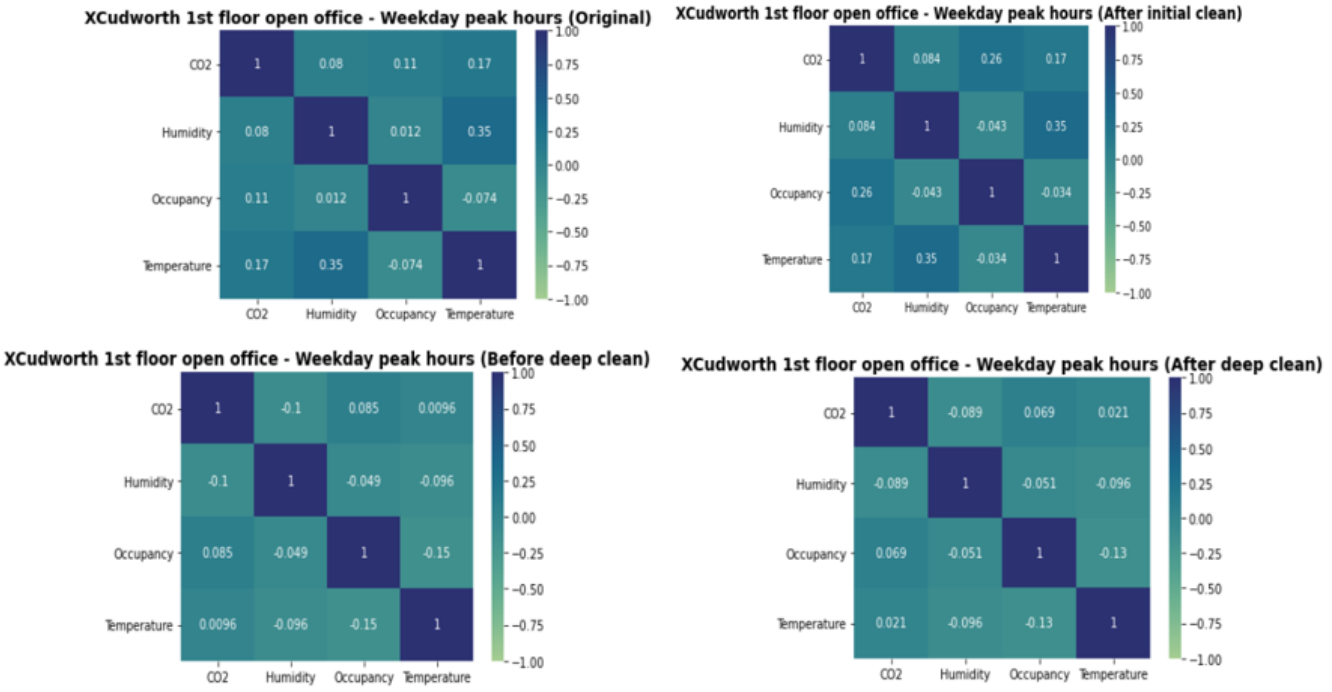


Figure 25 Correlation matrix for XCudworth 1st Floor Open Office Weekday Peak hours

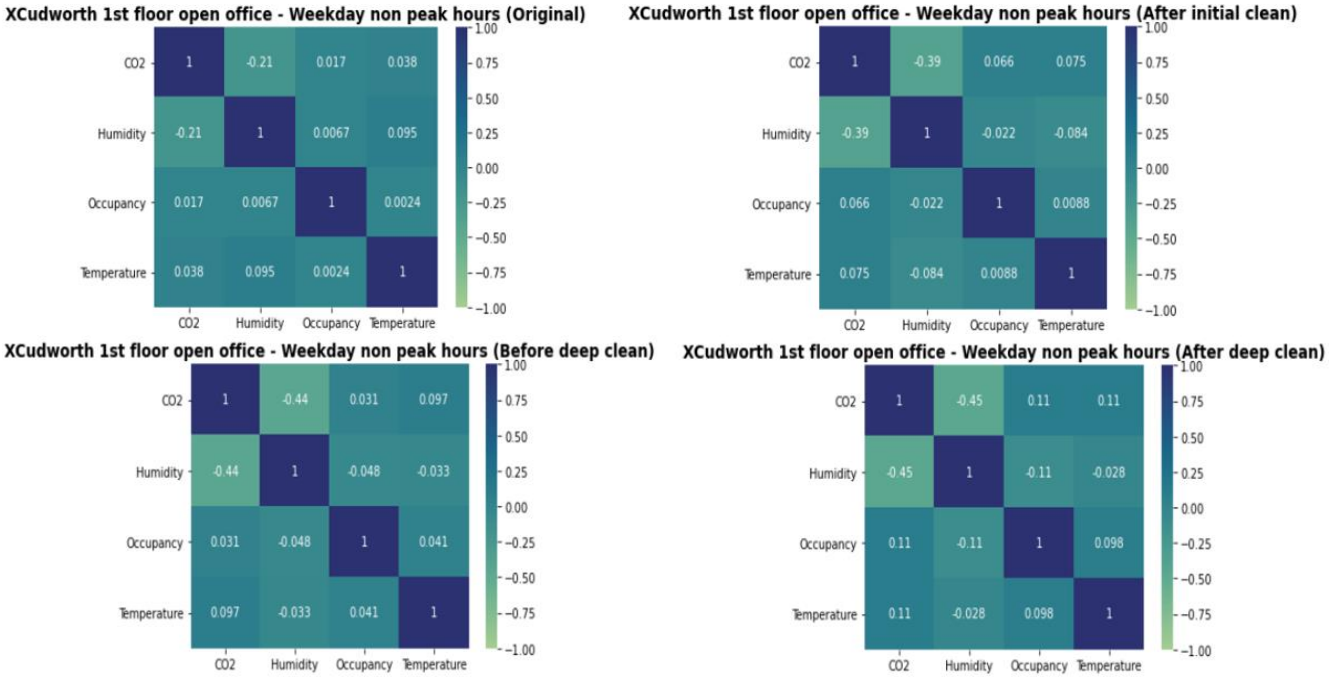


Figure 26 Correlation matrix for XCudworth 1st Floor Open Office Weekday Non Peak hours

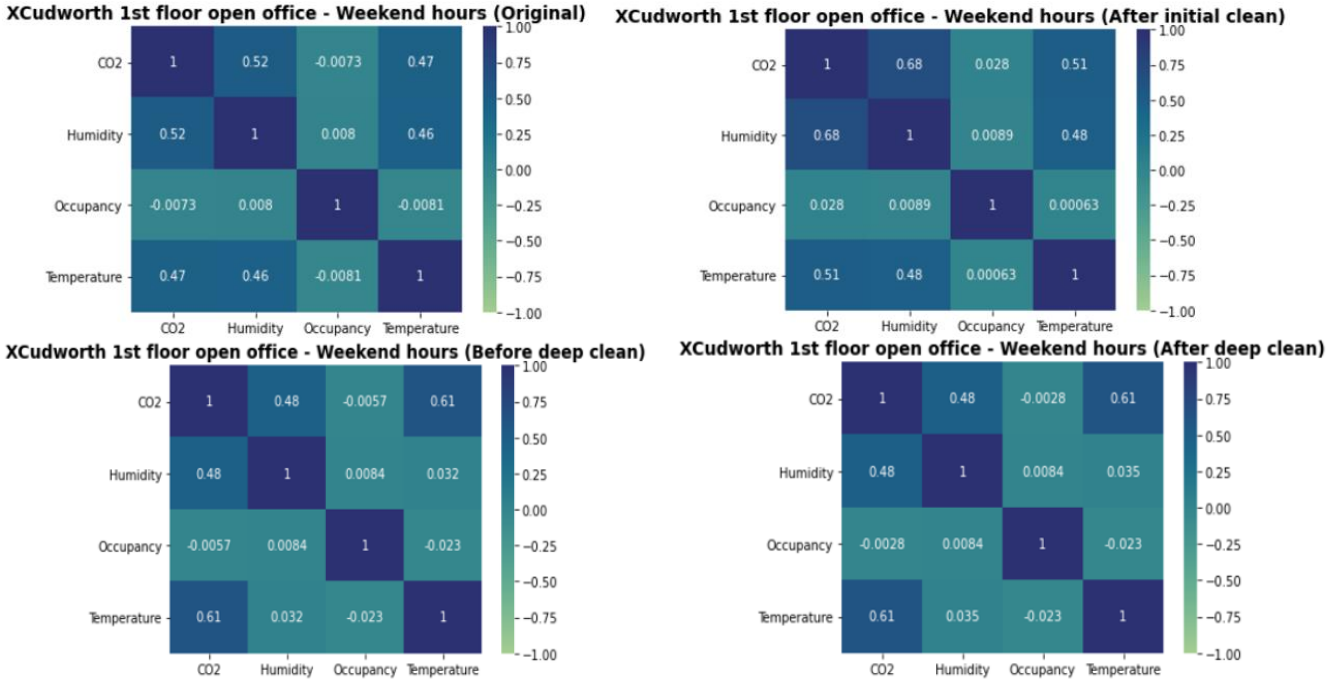


Figure 27 Correlation matrix for XCudworth 1st Floor Open Office Weekend hours

Like the above correlation analysis done for XCudworth's 1st Floor Open Office geometry room (Figure 24) along with its weekday (Figure 25, Figure 26)and weekend hours (Figure 27) datasets, all the other geometry rooms of XCudworth and also each and every geometry

rooms of other buildings like XWestgate, XWorsbrough, and XTownhall along with its weekday and weekend data were analysed for finding relationship among the 4 metrics.

But the main objective with performing this correlation analysis is to find the most accurate, precise, and meaningful correlations between the CO2, Humidity, Occupancy and Temperature metrics for each of the building's geometry room name and for each geometry room's weekday peak hours, weekday non-peak hours and weekend hours. As we have 4 datasets for a particular building's data, so, instead of highlighting all the correlation analysis results which would certainly be a huge count of correlation matrix and difficult to follow, hence we have to find a way to make a conclusion about which dataset's correlation analysis result is the most accurate and meaningful to explain the major relationships that the metrics hold for that particular building's geometry room.

6.2. Simple Linear Regression statistical check

To select which buildings dataset give the meaningful and accurate results in terms of correlations between the metrics, we use a correlation significance test with simple linear regression statistics. By evaluating the performance metric like Mean Square Error (MSE), Mean Absolute Error (MAE) and R-squared score it can be determined which dataset can be used to achieve our objective. Occupancy is considered as the explanatory variable and the other 3 metrics are the response variable because when a data for occupancy is recorded then the data for CO2 may be recorded as any presence within the room lead to CO2 emissions, followed by other 2 metrics. All this work done to correctly convey the accurate correlation results among the metrics, both for the unique building's geometrical rooms and their corresponding time period data of weekday (peak and non-peak) and weekend hours. The below Figure 28 narrates the same.

Buildings data Original

Regression Statistics	
Multiple R	0.580128
R Square	0.336549
Adjusted R Square	0.336376
Standard Error	0.193106
Observations	7663

ANOVA

	df	SS	MS	F	ignificance F
Regression	2	144.8975	72.44875	1942.843	0
Residual	7660	285.6419	0.03729		
Total	7662	430.5394			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	2.845118	0.062343	45.63652	0	2.722908	2.967327	2.722908	2.967327
X Variable 1	0.008992	0.00168	5.352955	8.9E-08	0.005699	0.012285	0.005699	0.012285
X Variable 2	0.571493	0.009383	60.90727	0	0.5531	0.589886	0.5531	0.589886

Coefficients	2.845118	0.008992	0.571493
MSE	31.64192		
MAE	0.08017		

Buildings data After initial clean

Regression Statistics	
Multiple R	0.145452
R Square	0.021156
Adjusted R Square	0.020901
Standard Error	0.311549
Observations	7668

ANOVA

	df	SS	MS	F	ignificance F
Regression	2	16.08019	8.040095	82.83417	0
Residual	7665	743.9843	0.097063		
Total	7667	760.0645			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	5.692485	0.0753	75.59753	0	5.544877	5.840094	5.544877	5.840094
X Variable 1	0.004128	0.002686	1.536893	0.124361	-0.00114	0.009392	-0.00114	0.009392
X Variable 2	0.144025	0.011303	12.7421	8.14E-37	0.121868	0.166182	0.121868	0.166182

Coefficients	5.692485	0.004128	0.144025
MSE	743.9843		
MAE	0.148344		

Buildings data Before deep clean

Regression Statistics	
Multiple R	0.9571769
R Square	0.91618762
Adjusted R Square	0.91616471
Standard Error	0.05489137
Observations	7320

ANOVA

	df	SS	MS	F	ignificance F
Regression	2	241.0003	120.5001	39992.57	0
Residual	7317	22.04658	0.003013		
Total	7319	263.0468			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	0.37295998	0.025433	14.66467	5.19E-48	0.323105	0.422815	0.323105	0.422815
X Variable 1	0.00400106	0.000553	7.237345	5.04E-13	0.002917	0.005085	0.002917	0.005085
X Variable 2	0.94349209	0.00385	245.0497	0	0.935945	0.95104	0.935945	0.95104

Coefficients	0.37295998	0.004001	0.943492
MSE	22.04658		
MAE	0.04062		

Buildings data After deep clean

Regression Statistics	
Multiple R	0.959898148
R Square	0.921404455
Adjusted R Square	0.92138296
Standard Error	0.053129578
Observations	7317

ANOVA

	df	SS	MS	F	ignificance F
Regression	2	242.003	121.0015	42866.49	0
Residual	7313	20.64279	0.002823		
Total	7315	262.6457			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	0.368219723	0.02466	14.93194	1.1E-49	0.319879	0.41656	0.319879	0.41656
X Variable 1	0.004595038	0.000537	8.559259	1.37E-17	0.003543	0.005647	0.003543	0.005647
X Variable 2	0.944111368	0.003733	252.8766	0	0.936793	0.95143	0.936793	0.95143

Coefficients	0.368219723	0.004595	0.944111
MSE	20.64278578		
MAE	0.040264446		

Figure 28 Simple Linear Regression Statistics on the buildings data

As it can be clearly understood from Figure 28 the dataset after deep cleaning is the one that should be considered for the most meaningful correlations between the CO2, Humidity, Occupancy and Temperature metrics for each of the building's geometry room name and for each geometry room's weekday peak hours, weekday non-peak hours and weekend hours.

Keeping all this in mind about after deep clean datasets for accurate correlation analysis between the metrics, a detailed correlation analysis result for all the building's geometry room name and for each geometry room's weekday peak hours, weekday non-peak hours and weekend hours is briefed in **section 10.1.**

7. Forecasting models development

The project's next phase will be to forecast each measure for each building utilising the after deep clean datasets that we currently have for each building and its geometrical rooms. Many forecasting models may be constructed for this purpose. However, because the data presented to us is time-series data from a real-time sensor, time series forecasting models are advised. So, before we go into the specifics of the constructed models, a precondition check must be performed.

7.1. Data preparation prerequisite

Prior to being utilised for forecasting, making sure that each geometry room dataset from the building has been appropriately preprocessed with careful cleaning. This simply implies that any data redundancies and inconsistencies should be eliminated from the datasets. The fact that the after deep clean datasets preserve data integrity so well is another justification for using them for each of the building's geometry rooms. Additionally, we need to forecast each of the four metrics within each of the building's geometry rooms. For this, we broke down our datasets(Figure 29) further into four with one for each of the metric.

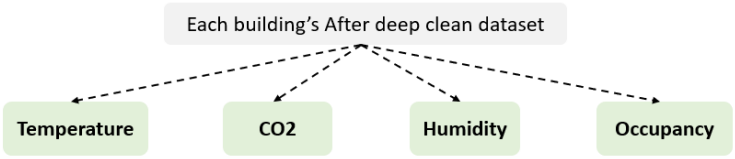


Figure 29 Each metric dataset breakdown

The geometry room dataset for each building is also taken into account for forecasting purposes rather of the datasets for the weekday peak hours, non-peak hours, and weekend hours. This is due to the fact that with these partitioned datasets, we will have less data rows, which might prevent the forecasting models from producing accurate findings. For time series forecasting models, more data yields better results.

Following the completion of this prerequisite check, the following time series forecasting models were developed and before that, each dataset underwent a split.

7.2. Data Splitting – Training and Test set

After performing dataset segregation as showed in Figure 29, each dataset records are then divided into training and test set.

- *Training set* comprising 70% of the metric based dataset records. It's mainly used to train the developed forecasting models and perform future date forecasting for each of the metric.
- *Test set* comprising the remaining 30% of the dataset records comprise. It's mainly used to determine how accurate our model's predictions were by comparing the anticipated values with the real values.

7.3. ARIMA

Autoregressive Integrated Moving Average (ARIMA) is a most common and popular time series forecasting model developed to analyse and predict trends in the datasets that we have for each metric for each building's geometry. The main functionality of this model is to analyse the data showing non-stationarity behaviour where the mean, median and variance keep on varying over the time. The past dates observation or the historical data trends are taken into account for forecasting future date values for each of the metrics. ARIMA model is defined as ARIMA(p, d, q) where;

- p "Non-Seasonal Autoregressive Order"[1][2] represents the number of "past time or 'lag' data"[2] considered for forecasting.
- d "Non-Seasonal Differencing Order"[1][2] represents the number of "data differencing steps needed to make the observations stationary"[2] where the mean, median and variance attain constant value over the time.
- q "Non-Seasonal Moving Average Order"[1][2] represents the number of "past time or 'lag' data forecast errors"[2] used. For forecasting how many seasonal 'lag' data errors should be considered is specified using parameter q.

7.4. SARIMA

Seasonal Autoregressive Integrated Moving Average (SARIMA) is an advanced version of ARIMA which is developed to handle the trends and the seasonality of the datasets that we have for each metric for each building's geometry, in a more efficient way. In addition to the ARIMA's non-seasonal p, d, q values, seasonal P, D, Q values along with the frequency s are considered in SARIMA where;

p, d, q are the non-seasonal parameters that were inherited from the ARIMA model.

- P "Seasonal Autoregressive Order"[1][3] represents the number of "seasonal past time or 'lag' data"[3] considered for forecasting.
- D "Seasonal Differencing Order"[1][3] represents the number of "seasonal data differencing steps needed to make the observations stationary"[3].
- Q "Seasonal Moving Average Order"[1][3] represents the number of "seasonal past time or 'lag' data forecast errors"[3] used. For forecasting how many seasonal 'lag' data errors should be considered is specified using parameter Q.
- s represents the seasonal frequency indicating "how frequently the seasonality within the data points repeat."[3]

The seasonal pattern for each building's internal room structure is hour-wise with frequency s considered as 12 because in 1 hour 12 data records are recorded since the timeframe difference between 2 records are 5mins. All this done to forecast accurate/precise values for each of the metric within each building's geometry/internal room.

Using python programming, ARIMA and SARIMA models were developed *(Source code in **Appendix 4, Appendix 5** respectively)* and then trained for each of the metric dataset within each building's geometry room to make the predictions for that metric or parameter.

7.5. Prophet

Prophet, a forecasting model initially introduced by Facebook is a powerful and widely used model nowadays that effectively deals with time series data with strong seasonal trends. It handles the missing values and outliers as well, all with an aim to provide accurate

predictions. Using python programming *(Source code in **Appendix 6**)* prophet model was developed where 2 important parameters were specified;

- *periods*

  It specifies the seasonality length in the data. Our focus is on the daily seasonality as each day the value for each of the metrics will change and since our data records time difference is 5mins, the corresponding periods value is '288' (24hrs × 60mins ÷ 5mins).

- *freq*

  It specifies the frequency of the data. We have used '5T' as the frequency since the data records time difference is 5mins, and with the same difference the forecasting is also carried out.

7.6. Other models

One more time series forecasting model was developed but not mainly considered further in the analysis due to inaccuracy of prediction and it was;

- LSTM (Long Short Term-Memory) model

  With LSTM model already studied in **section 2.1**, using python programming LSTM model was developed and trained initially *(Source code in **Appendix 7**)* with epoch = 100 and batch size = 24 (day-wise timeframe).

  With the developed LSTM model the forecasting was performed on one of the building XCudworth's 2nd floor open office geometry room and that too on the geometry's partitioned dataset of weekend hours (Figure 30).

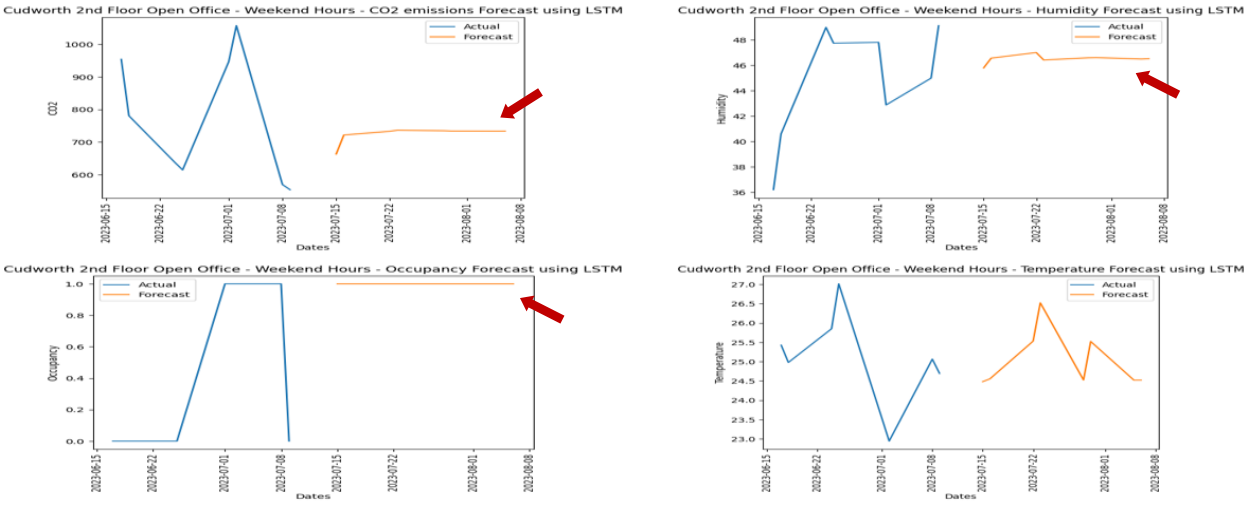


Figure 30 LSTM forecasting for each metric of XCudworth 2nd Floor Office during weekend hours

As it can be drawn, the forecasted values for CO2, Occupancy and even for Humidity metric are not good as a linear line is observed. This shows the inaccuracy of the predictions as linearity shouldn't be the case if the forecasting performed is based on the historical dates trend. The reason for this behaviour could you due to the day-wise timeframe considered for each metric and also due to the consideration of partitioned weekend hours data which gives less data rows as input to the forecasting model (Figure 31).

date	temp	predicted dates	predicted
6/17/2023	25.42756	7/15/2023	24.48211
6/18/2023	24.97834	7/16/2023	24.5548
6/24/2023	25.85098	7/22/2023	25.53042
6/25/2023	27.01439	7/23/2023	26.52291
7/1/2023	23.53091	7/29/2023	24.524
7/2/2023	22.94326	7/30/2023	25.52244
7/8/2023	25.06475	8/5/2023	24.522
7/9/2023	24.69366	8/6/2023	24.52191

Figure 31 Forecasting dataset sample

Because of this, the forecasting is carried out for each of the building's geometry as a whole rather than considering the segmented datasets of weekday and weekend hours and also using the 5min time difference for forecasting which is the same as to the original sensored dataset provided by SmartViz.

7.7. AIC score test

In **section 7.3** and **7.4** it was mentioned that ARIMA is defined with non-seasonal parameters p, d, q and SARIMA is defined with seasonal parameters p, d, q, P, D, Q, s. To

determine these parameter's value, Akaike Information Criterion (AIC) score test is carried out. Using python programming *(Source code in **Appendix 8**)*, AIC score test for ARIMA and SARIMA models for each of the respective metric dataset within each building's geometry is performed. By considering the lowest AIC score value, we obtain the optimal parameters value for ARIMA and SARIMA models which is then used to make accurate predictions for the metric CO2, temperature, Humidity and Occupancy respectively. The AIC score changes depending on the dataset we use due to its historical trend.

Figure 32 illustrates the overall time series forecasting models development flow from segregating the datasets for each of the metric within each of the building's geometry room to developing the aforementioned time series forecasting models and training them for each of the metric to predict future date values.

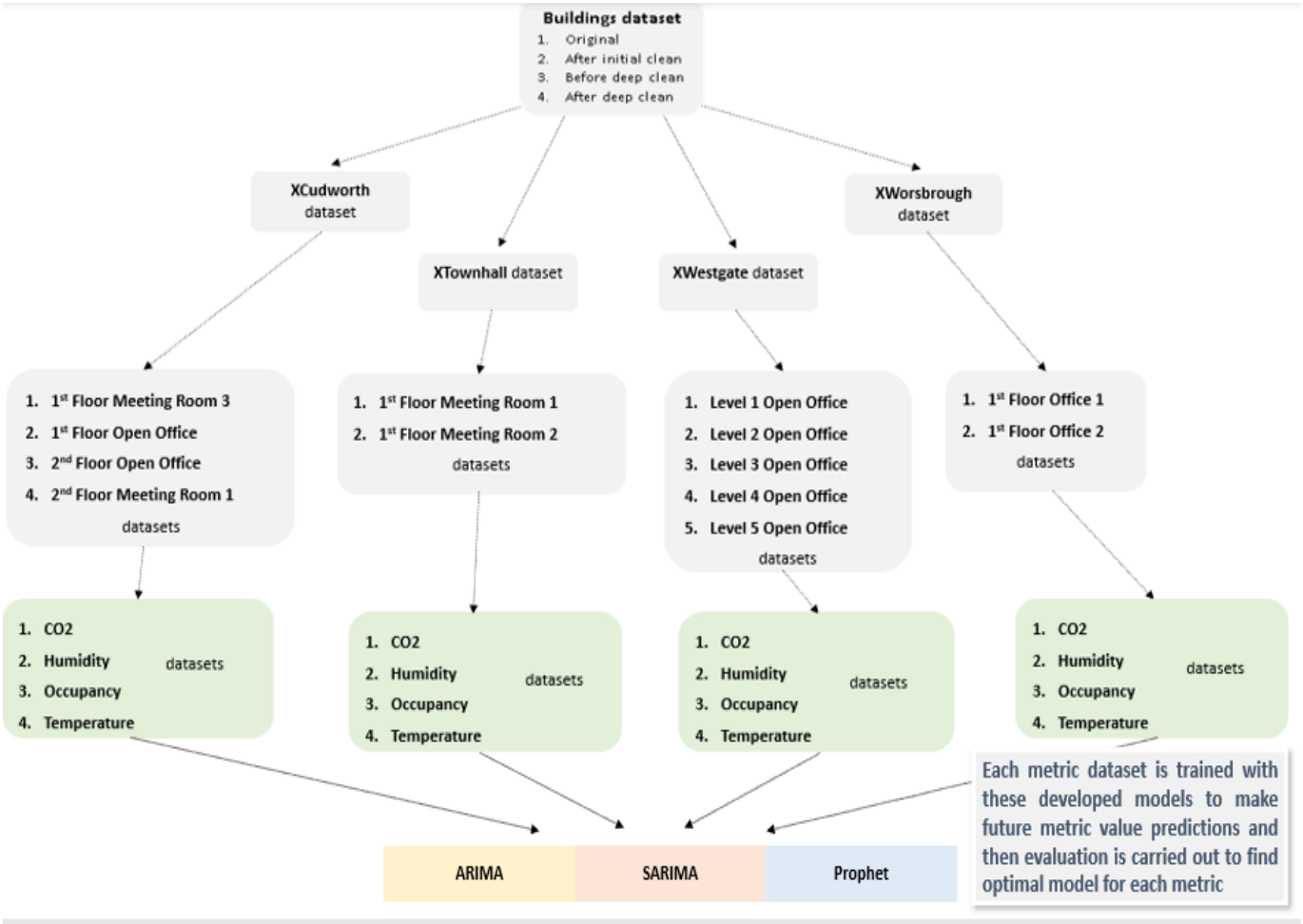


Figure 32 Forecasting models development chapter flow

8. Optimal Forecasting models

The objective for the second phase of this project is to determine the optimal models among the developed ones (ARIMA, SARIMA, Prophet) for metrics CO2, Humidity, Temperature and Occupancy within each of the building including its geometry rooms. For achieving this, performance metrics evaluation needs to be carried out for each of the developed forecasting models.

8.1. Evaluating models

The performance evaluation metric like MSE, RMSE, MAE and Composite Score (each of them explained in detail in **section 3.5**) are assessed using the test set by comparing the model's foreseen values with the originally observed values.

8.2. Comparing models

Table 4, Table 5 Table 6, Table 7 outlines the performance evaluation values for each metric within each building's internal room structure. The model with the least MSE, RMSE, MAE and Composite Score values is concluded to be that metric's optimal model within the considered building (highlighted in green in Table 4, Table 5, Table 6, Table 7)

The MSE, RMSE, MAE values does not have any criteria or interval. It totally depends on the parameter's historical values and trends for which the performance metrics are evaluated.

XCudworth geometry room name	Parameters	Models	MSE	RMSE	MAE	Composite Score
1st Floor Open Office	CO2	**ARIMA**	4420.20	66.48	55.62	1804.71
		SARIMA	6129.95	78.29	65.07	2494.99
		Prophet	24739.73	157.29	142.10	9985.71
	Humidity	ARIMA	`11.54	3.40	2.90	6.52
		SARIMA	1.16	1.08	0.81	1.03
		Prophet	30.93	5.56	4.66	15.44
	Occupancy	ARIMA	19.88	4.46	4.15	10.54
		SARIMA	0.32	0.57	0.39	0.41
		Prophet	17.32	4.16	2.84	9.03
	Temperature	ARIMA	5.18	2.28	2.15	3.40
		SARIMA	0.69	0.83	0.77	0.75
		Prophet	0.39	0.63	0.49	0.49
	CO2	ARIMA	14913.02	122.12	117.14	6036.99
		SARIMA	37081.05	192.56	167.67	14940.49
		Prophet	13457.50	116.01	99.47	5447.64

1st Floor Meeting Room 3	Humidity	ARIMA	4.78	2.19	1.90	3.14
		SARIMA	0.33	0.57	0.43	0.43
		Prophet	1.89	1.37	1.24	1.54
	Occupancy	ARIMA	0.30	0.50	0.51	0.44
		SARIMA	0.00014	0.01	0.01	0.006
		Prophet	0.08	0.29	0.13	0.16
	Temperature	ARIMA	3.47	1.86	1.77	2.48
		SARIMA	0.71	0.84	0.81	0.78
		Prophet	0.17	0.42	0.35	0.30
2nd Floor Open Office	CO2	ARIMA	153997.95	392.43	354.13	61823.15
		SARIMA	54461.91	233.37	184.72	21910.19
		Prophet	16643.29	129.01	94.94	6724.50
	Humidity	ARIMA	20.85	4.57	4.02	10.92
		SARIMA	3.55	1.88	1.45	2.42
		Prophet	6.14	2.48	1.94	3.78
	Occupancy	ARIMA	35.57	5.96	5.76	17.74
		SARIMA	26.83	5.18	2.47	13.02
		Prophet	9.77	3.13	1.85	5.40
	Temperature	ARIMA	3.36	1.83	1.63	2.38
		SARIMA	0.44	0.67	0.39	0.49
		Prophet	1.83	1.35	1.00	1.44
2nd Floor Meeting Room 1	CO2	**ARIMA**	9409.24	97.00	75.84	3815.55
		SARIMA	11139.00	105.54	89.75	4514.19
		Prophet	19784.10	140.66	116.43	7990.76
	Humidity	ARIMA	9.23	3.04	2.79	5.44
		SARIMA	7.68	2.77	2.55	4.67
		Prophet	15.18	3.90	2.93	8.12
	Occupancy	ARIMA	0.49	0.70	0.67	0.61
		SARIMA	0.00001	0.00257	0.00108	0.00110
		Prophet	0.42	0.65	0.43	0.49
	Temperature	ARIMA	3.97	1.99	1.75	2.71
		SARIMA	0.38	0.62	0.39	0.45
		Prophet	0.34	0.59	0.38	0.43

Table 4. Models evaluation for XCudworth

XTownhall geometry room name	Parameters	Models	MSE	RMSE	MAE	Composite Score
1st Floor Meeting Room 1	CO2	**ARIMA**	7395.83	86.00	64.41	3003.46
		SARIMA	8246.28	90.81	59.44	3343.58
		Prophet	31810.15	178.35	147.52	12821.82
	Humidity	ARIMA	7.14	2.67	2.30	4.35
		SARIMA	2.91	1.71	1.29	2.07
		Prophet	38.22	6.18	5.08	18.67
	Occupancy	ARIMA	0.51	0.71	0.69	0.63
		SARIMA	0.14	0.37	0.22	0.23
		Prophet	0.25	0.50	0.28	0.33

	Temperature	**ARIMA**	0.45	0.67	0.59	0.56
		SARIMA	0.71	0.84	0.77	0.77
		Prophet	0.74	0.86	0.79	0.79
1st Floor Meeting Room 2	CO2	ARIMA	17089.22	130.73	109.62	6907.79
		SARIMA	16428.97	128.18	108.74	6642.66
		Prophet	68976.32	262.63	245.01	27742.82
	Humidity	ARIMA	13.46	3.67	3.21	7.45
		SARIMA	2.38	1.54	1.14	1.76
		Prophet	23.78	4.88	3.82	12.12
	Occupancy	ARIMA	2.34	1.53	1.49	1.84
		SARIMA	0.01	0.11	0.09	0.06
		Prophet	0.88	0.94	0.46	0.77
	Temperature	ARIMA	0.54	0.73	0.64	0.63
		SARIMA	0.52	0.72	0.63	0.61
		Prophet	0.92	0.96	0.71	0.87

Table 5 Models evaluation for XTownhall

XWorsbrough geometry room name	Parameters	Models	MSE	RMSE	MAE	Composite Score
1st Floor Office 1	CO2	**ARIMA**	40850.79	202.12	166.56	16450.92
		SARIMA	76609.06	276.78	215.76	30791.39
		Prophet	76418.71	276.44	225.97	30718.21
	Humidity	ARIMA	22.38	4.73	3.81	11.52
		SARIMA	2.15	1.47	1.29	1.69
		Prophet	43.73	6.61	5.41	21.10
	Occupancy	**ARIMA**	4.44	2.11	1.80	2.95
		SARIMA	10.96	3.31	2.60	6.16
		Prophet	3.11	1.76	1.24	2.14
	Temperature	ARIMA	1.53	1.24	1.14	1.33
		SARIMA	0.38	0.62	0.40	0.46
		Prophet	0.93	0.96	0.85	0.92
1st Floor Office 2	CO2	**ARIMA**	52417.62	228.95	189.03	21092.44
		SARIMA	20921.15	457.47	383.76	83964.83
		Prophet	73231.92	270.61	218.02	29439.36
	Humidity	ARIMA	70.89	8.42	7.26	33.06
		SARIMA	83.14	9.12	7.92	38.37
		Prophet	53.27	7.30	6.57	25.47
	Occupancy	ARIMA	12.95	3.60	3.26	7.24
		SARIMA	15.90	3.99	2.95	8.44
		Prophet	2.08	1.44	0.81	1.51
	Temperature	**ARIMA**	3.95	1.99	1.79	2.71
		SARIMA	6.05	2.46	2.29	3.84
		Prophet	7.39	2.72	2.35	4.48

Table 6 Models evaluation for XWorsbrough

XWestgate geometry room name	Parameters	Models	MSE	RMSE	MAE	Composite Score
Level 1 Open Office	CO2	ARIMA	48432.24	220.07	155.75	19485.64
		SARIMA	48575.48	220.40	156.27	19543.19
		Prophet	28508.57	168.84	139.08	11495.81
	Humidity	**ARIMA**	68.31	8.27	7.56	32.07
		SARIMA	86.61	9.31	8.58	40.01
		Prophet	109.34	10.46	7.99	49.27
	Occupancy	**ARIMA**	301.49	17.36	12.74	129.63
		SARIMA	476.44	21.83	13.62	201.21
		Prophet	359.17	18.95	13.32	153.35
	Temperature	ARIMA	1.07	1.03	0.74	0.96
		SARIMA	0.95	0.98	0.63	0.86
		Prophet	0.85	0.92	0.62	0.80
Level 2 Open Office	CO2	**ARIMA**	82369.78	287.00	238.24	33105.49
		SARIMA	101199.42	318.12	266.79	40655.24
		Prophet	213692.75	462.27	371.69	85727.29
	Humidity	**ARIMA**	126.63	11.25	7.03	56.14
		SARIMA	160.76	12.68	12.53	71.87
		Prophet	127.98	11.31	7.83	56.94
	Occupancy	ARIMA	1157.51	34.02	28.90	481.88
		SARIMA	1684.08	41.04	26.61	692.73
		Prophet	1124.39	33.53	25.72	467.53
	Temperature	**ARIMA**	0.45	0.67	0.60	0.56
		SARIMA	0.78	0.88	0.72	0.79
		Prophet	1.21	1.10	0.93	1.09
Level 3 Open Office	CO2	**ARIMA**	31571.07	177.68	132.47	12721.47
		SARIMA	111141.91	333.38	292.24	44644.45
		Prophet	41658.69	204.10	171.75	16776.23
	Humidity	**ARIMA**	40.70	6.38	6.03	20.00
		SARIMA	91.13	9.55	9.38	42.13
		Prophet	87.10	9.33	7.14	39.78
	Occupancy	ARIMA	1113.79	33.37	26.02	463.33
		SARIMA	1532.55	39.15	20.10	630.79
		Prophet	915.24	30.25	17.45	380.41
	Temperature	ARIMA	5.35	2.31	2.24	3.50
		SARIMA	1.49	1.22	1.10	1.29
		Prophet	30.66	5.54	2.20	14.58
Level 4 Open Office	CO2	**ARIMA**	58521.72	241.91	207.56	23543.53
		SARIMA	99564.79	315.54	275.20	40003.14
		Prophet	320565.04	566.18	512.46	128549.61
	Humidity	ARIMA	33.47	5.79	5.27	16.71
		SARIMA	31.22	5.59	4.74	15.58
		Prophet	55.80	7.47	6.08	26.38
	Occupancy	**ARIMA**	478.54	21.88	18.14	203.42
		SARIMA	672.61	25.93	13.81	280.96

		Prophet	607.80	24.65	16.38	255.43
	Temperature	ARIMA	1.74	1.32	1.03	1.40
		SARIMA	1.73	1.31	1.04	1.40
		Prophet	0.72	0.85	0.66	0.74
Level 5 Open Office	CO2	ARIMA	14928.72	122.18	105.25	6039.72
		SARIMA	13526.74	116.30	92.52	5473.34
		Prophet	10367.22	101.82	83.04	4202.35
	Humidity	**ARIMA**	98.23	9.91	8.09	44.69
		SARIMA	196.55	14.02	8.72	85.44
		Prophet	87.50	9.35	8.12	40.24
	Occupancy	ARIMA	329.22	18.14	15.81	141.87
		SARIMA	329.34	18.15	15.82	141.92
		Prophet	328.21	18.12	9.25	139.49
	Temperature	**ARIMA**	0.46	0.68	0.58	0.56
		SARIMA	0.79	0.89	0.71	0.79
		Prophet	0.69	0.83	0.72	0.74

Table 7 Models evaluation for XWestgate

8.3. Hyperparameters tuning

The proper hyperparameters need to be adjusted in order to optimise the forecasting models (ARIMA, SARIMA, and Prophet). The p, d, q (ARIMA), P, D, Q, s (SARIMA), seasonality, and holiday prior scale (Prophet) are regarded as the hyperparameters for the constructed models. With the purpose of providing more informative sustainability indicators for buildings, the MSE, RMSE, and MAE values of the model may be reduced by adjusting the hyperparameter's value using the grid search approach. After performing this hyper-tuning on the best forecasting model for the metric under the considered building's section name, not much difference was observed in the performance as only a minute betterment (5%) was observed for metric forecasting. But this cross validates and confirms that the highlighted optimal model for each of the metrics in **section 8.2** is precise.

Section 10.2 displays an overall summarised view of the optimal forecasting model for each metric within each of the building's geometry room. Once the optimal model for each of the metric is determined, using the training set data and/or overall data of each metric dataset within the building as a whole and its geometry rooms, the metric values can be forecasted for future dates.

9. Visualisations with optimal models

“Data visualisation is the process of developing a graphical representation of the data for interactive display and efficient communication.”[21] There are several examples that may be utilised "to easily understand and interpret the data which is beneficial in terms of data analysis and decision-making process."[21]

In our analysis, interactive visualisations are made using determined optimal forecasting model for each metric in order to examine and comprehend the pattern or trend of each metric CO2, humidity, occupancy, and temperature that are associated within each building's geometry. Additionally, visualisations are a more effective and simple approach to communicate the efficacy of determined or chosen optimal predicting models for each of the parameter or metric. In relation to our goal, this would assist in making informed decisions and in planning and proactively implementing activities to increase a building's sustainability.

9.1. Actual VS Predicted values graph

Using the test set that was split in **section 7.2** for each of the metric dataset with each building’s geometry room, the actual versus prediction graphs were constructed for CO2, Humidity, Temperature and Occupancy respectively within each building’s geometry using its determined optimal model. Here, the metric’s observed or recorded values are displayed alongside the projected values for a better comparison in-order to convey what values could have been recorded if the metric’s optimal forecasting model was used within the building’s internal block.

9.2. Future date forecasting graph

Visualisations are created for CO2, Humidity, Temperature and Occupancy future dates forecasted values respectively that was resulted from executing the metric’s optimal model for the considered building’s geometry room. This was performed with the use of training set that was split in **section 7.2**. With this it is easy to understand and portray the model’s future forecasting performance for the metric or parameter interactively and more effectively.

Both the actual vs predicted values graph and future date forecasting graph for each of the CO2, Temperature, Occupancy, Humidity metric within each of the building's geometry room are illustrated in **section 10.3**.

9.3. Interactive forecasting report

In addition to the graphs that were discussed prior to this, an interactive power BI report is also developed. With this report, specific building ("XCudworth", "XWorsbrough", "XWestgate", "XTownhall") as a whole or its geometry room and metric or parameters ("CO2", "Humidity", "Temperature", "Occupancy") can be selected and forecasted using the determined optimal forecasting model for the selected metric within the selected building or building's internal room.

This interactive forecasting report is explained in **section 10.4** of the results.

10. Results and Discussion

Keeping the main objectives in our mind, this chapter explains the results of correlations and optimal forecasting models for each metric within each building along with meaningful visualisations.

10.1. Correlation analysis

Each building's geometrical room's name, weekday peak hours, weekday non-peak hours, and weekend hours are analysed in a complete correlation study for the entire buildings data by considering After deep clean datasets for the respective structure as the most accurate and precise dataset to generate meaningful correlations between the 4 metrics.

1. XCudworth building

- 1st Floor Open Office (geometry room name)

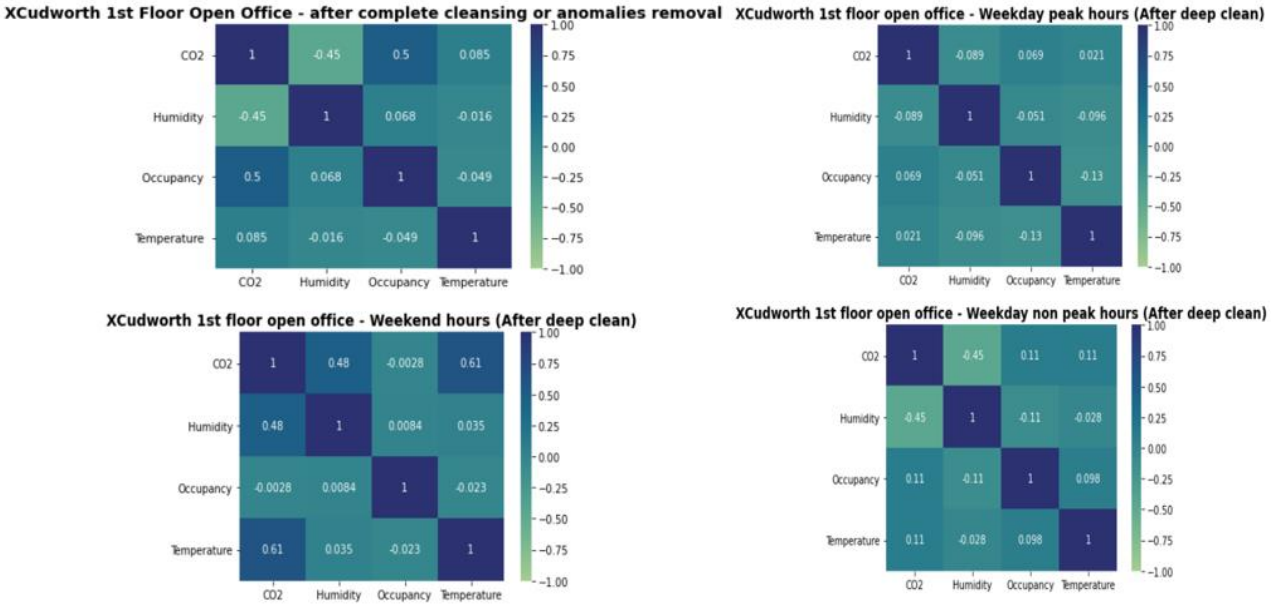


Figure 33 Correlation matrix for XCudworth 1st Floor Open Office

From the above Figure 33 correlation matrix the major conclusion that can be made for XCudworth 1st Floor Open Office are: Strong positive correlation (0.5) between metrics CO2 and Occupancy and a strong negative correlation (-0.45) between Humidity and CO2. The weekday peak hours and weekday non-peak

hours follow this same correlation trend but for weekend hours CO2 and Occupancy are negatively correlated whereas Humidity and CO2 are positively correlated.

- 1st Floor Meeting Room 3 (geometry room name)

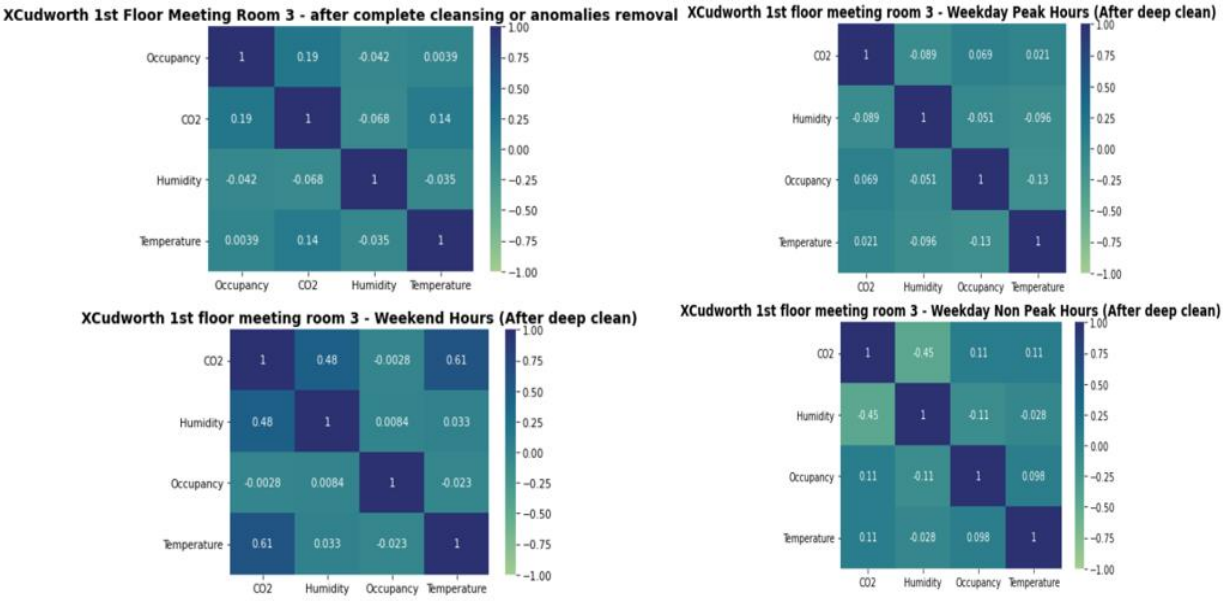


Figure 34 Correlation matrix for XCudworth 1st Floor Meeting Room 3

With Figure 34 correlation matrix the major conclusion that can be made for XCudworth 1st Floor meeting room 3 are: Positive correlation between metrics CO2 and Occupancy (0.19). Same positive correlation is observed during weekday peak (0.069) and non-peak hours (0.11) while weekend hours observe negative correlation (-0.0028). Positive correlation between CO2 and temperature (0.14). Same correlation trend is observed for weekday(0.021 peak, 0.11 off-peak) and weekend hours (0.61 very strong). A strong positive correlation between Humidity and CO2 in weekend hours (0.48) whereas negative correlation is observed during weekday (-0.089 peak, -0.45 off peak) and for the geometry as a whole (-0.068).

- 2nd Floor Open Office (geometry room name)

From the below Figure 35 correlation matrix the major conclusion that can be made for XCudworth 2nd Floor Open Office are: Strong negative correlation

observed between metrics CO2 and Humidity (-0.44). Same strong negative correlation trend is observed for weekday(-0.42 peak, -0.46 off-peak) and weekend hours (-0.54 very strong). Good positive correlation between CO2 and Occupancy with 0.18 for geometry as a whole, 0.32 for weekday peak hours, 0.23 weekday non peak hours and 0.28 for weekend hours. Negative correlation between metrics CO2 and Temperature with -0.14 for geometry as a whole, -0.13 for weekday peak hours, -0.078 weekday non peak hours and -0.46 (strong) for weekend hours.

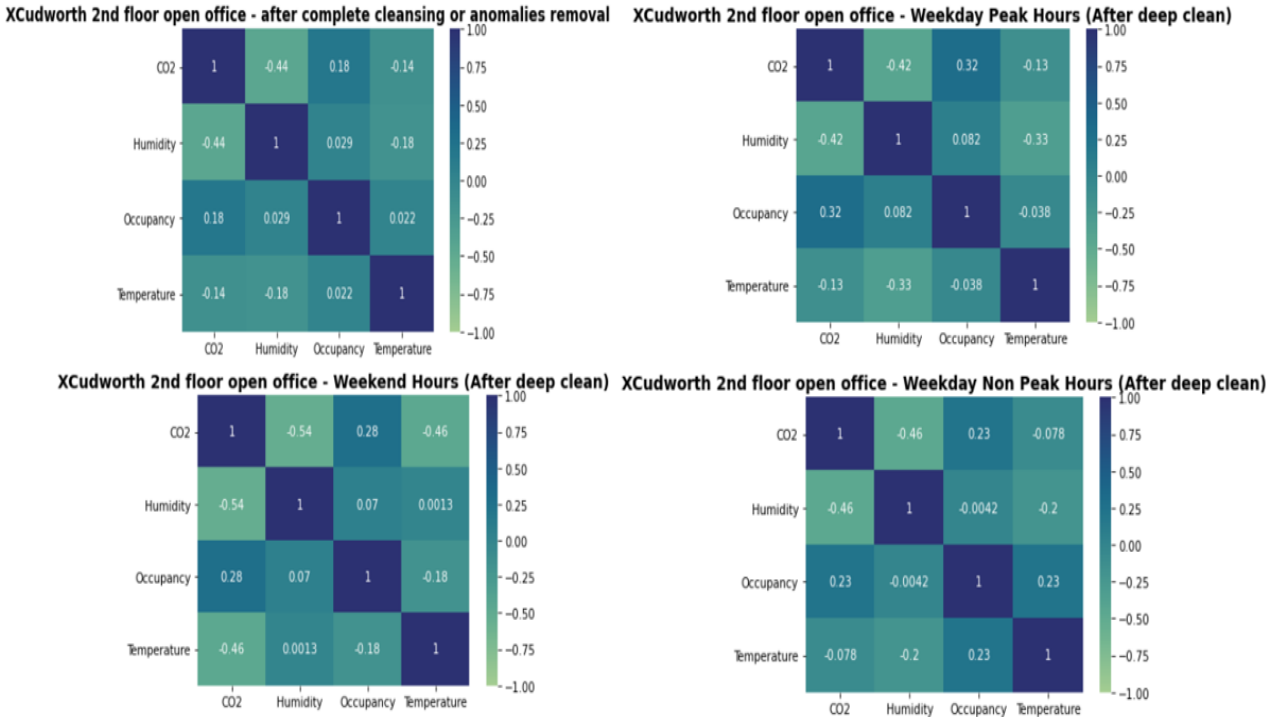


Figure 35 Correlation matrix for XCudworth 2nd Floor Open Office

- 2nd Floor Meeting Room 1 (geometry room name)

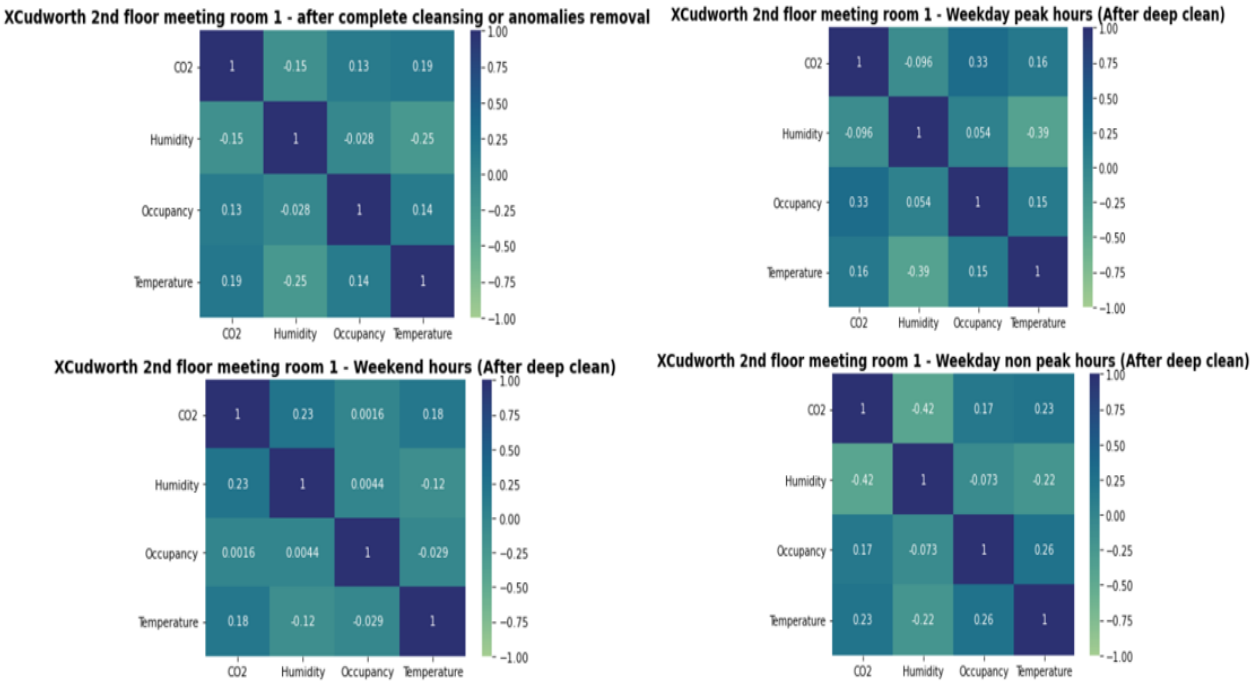


Figure 36 Correlation matrix for XCudworth 2nd Floor Meeting Room 1

With Figure 36 correlation matrix the major conclusion that can be made for XCudworth 2nd Floor Meeting Room 1 are: Good positive correlation between metrics CO2 and Occupancy with 0.13 for geometry as a whole, 0.33 for weekday peak hours, 0.17 weekday non peak hours and 0.0016 (not so good) for weekend hours. A good negative correlation is observed between Humidity and Temperature with -0.25 for geometry as a whole, -0.39 for weekday peak hours, -0.22 weekday non peak hours and -0.12 for weekend hours. Also, except for the weekend hours negative correlation is observed between CO2 and Humidity metrics with 0.23 for weekend hours, -0.15 for geometry as a whole, -0.096 for weekday peak hours and -0.42 (strong) for weekday off-peak hours.

2. XTownhall building

- 1st Floor Meeting Room 1 (geometry room name)

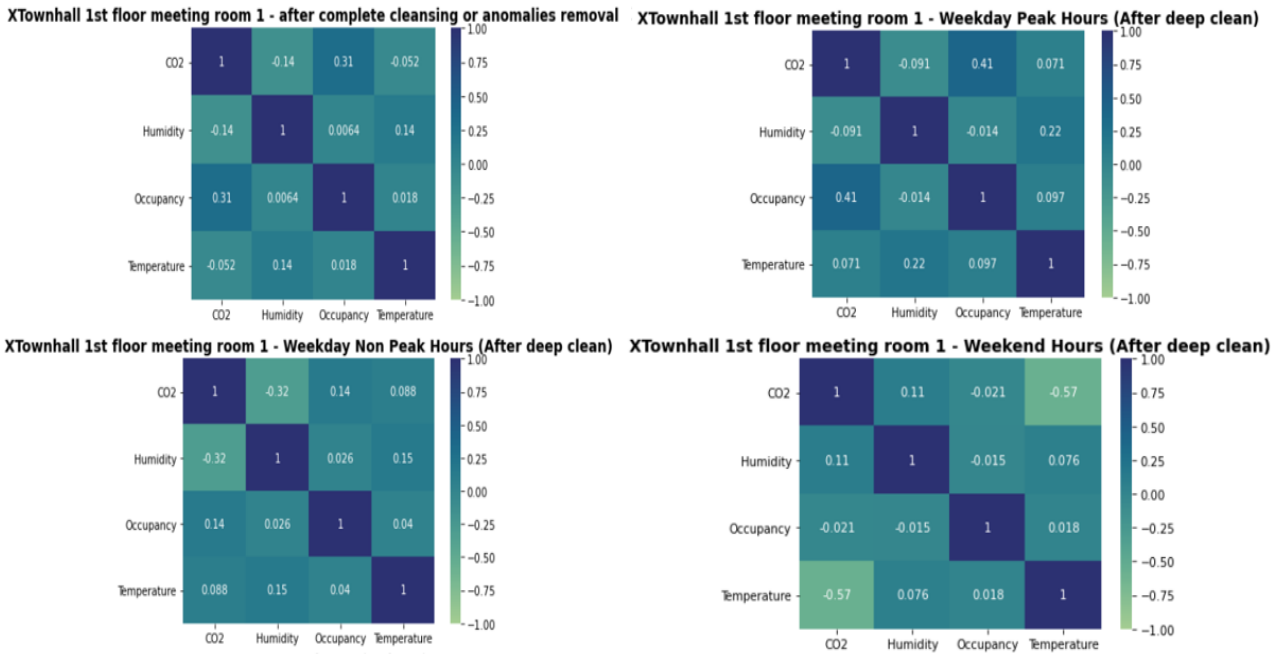


Figure 37 Correlation matrix for XTownhall 1st Floor Meeting Room 1

From the above Figure 37 correlation matrix the major conclusion that can be made for XTownhall 1st Floor Meeting Room 1 are: Strong positive correlation between metrics CO2 and Occupancy except for weekday non-peak hours having -0.021, geometry room as a whole having 0.31, 0.41 for weekday peak hours and 0.14 for weekend hours. A strong negative correlation between Temperature and CO2 for Weekday non-peak hours with -0.57. Also, except for the weekday non-peak hours negative correlation is observed between CO2 and Humidity metrics with -0.32 for weekend hours, -0.14 for geometry as a whole, -0.091 for weekday peak hours and 0.11 for weekday non-peak hours.

- 1st Floor Meeting Room 2 (geometry room name)

  With the below Figure 38 correlation matrix the major conclusion that can be made for XTownhall 1st Floor Meeting Room 2 are: Good negative correlation between metrics CO2 and Humidity with -0.29 for geometry room as a whole, -0.27 for weekday peak hours, -0.32 for weekday non-peak hours and -0.22 for weekend hours. Except for weekend hours, a good positive correlation between Occupancy and CO2 is observed with 0.24 for geometry as a whole, 0.5 (strong) for weekday peak hours, 0.1 for weekday non-peak hours and -0.13 for weekend

hours. Strong positive correlation between metrics CO2 and Temperature with 0.47 for geometry room as a whole, 0.23 for weekday peak hours, 0.43 for weekday non-peak hours and 0.64 (very strong) for weekend hours.

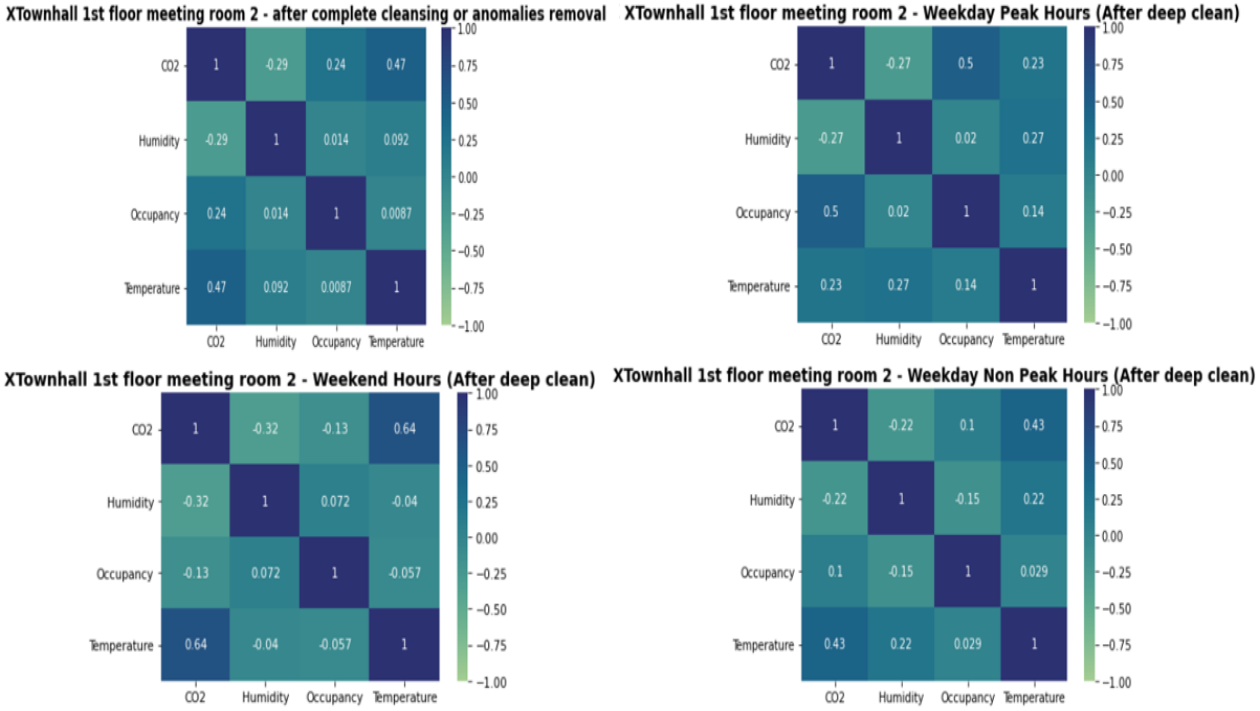


Figure 38 Correlation matrix for XTownhall 1st Floor Meeting Room 2

3. XWorsbrough building

- 1st Floor Office 1 (geometry room name)

  From the below Figure 39 correlation matrix the major conclusion that can be made for XWorsbrough 1st Floor Office 1 are: Negative correlation between metrics CO2 and Humidity with -0.13 for geometry room as a whole, -0.26 for weekday peak hours, -0.034 for weekday non-peak hours and -0.11 for weekend hours. Positive correlation between metrics CO2 and Occupancy except for weekend hours having -0.12, geometry room as a whole having 0.031, 0.24 (good) for weekday peak hours and 0.013 for weekday non-peak hours.

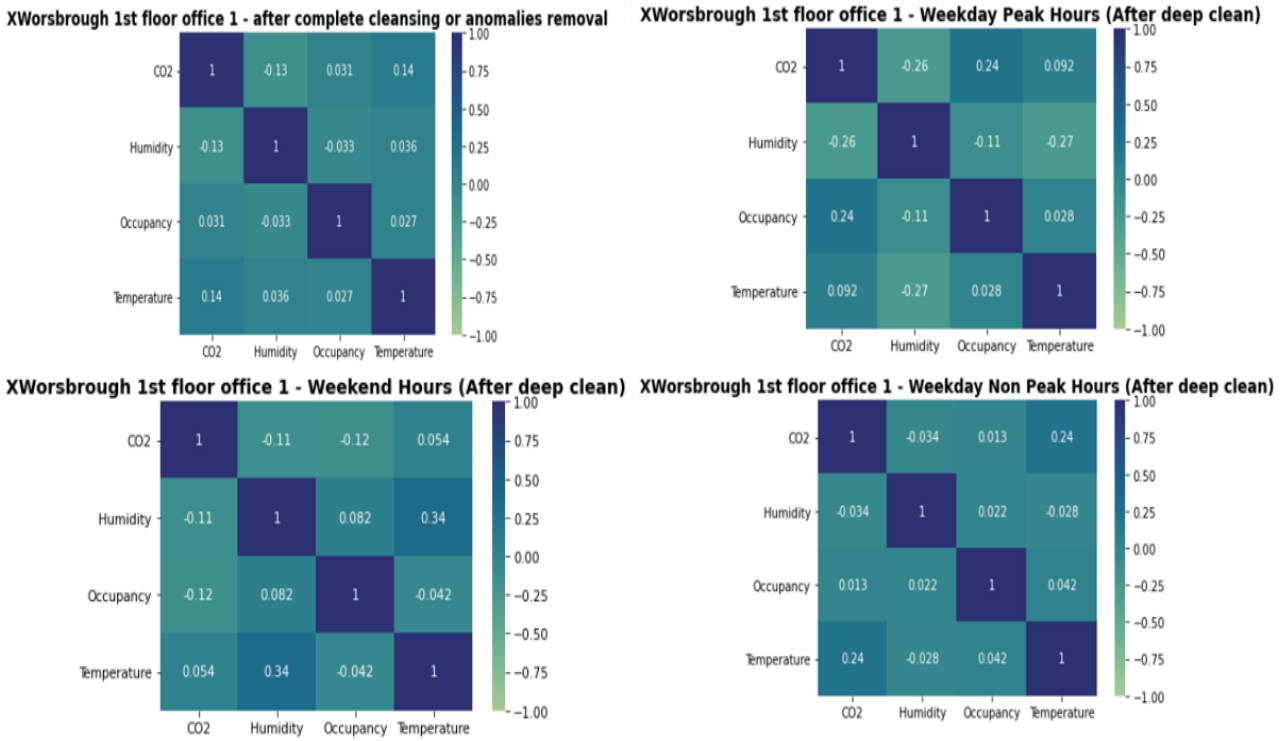


Figure 39 Correlation matrix for XWorsbrough 1st Floor Office 1

- 1st Floor Office 2 (geometry room name)

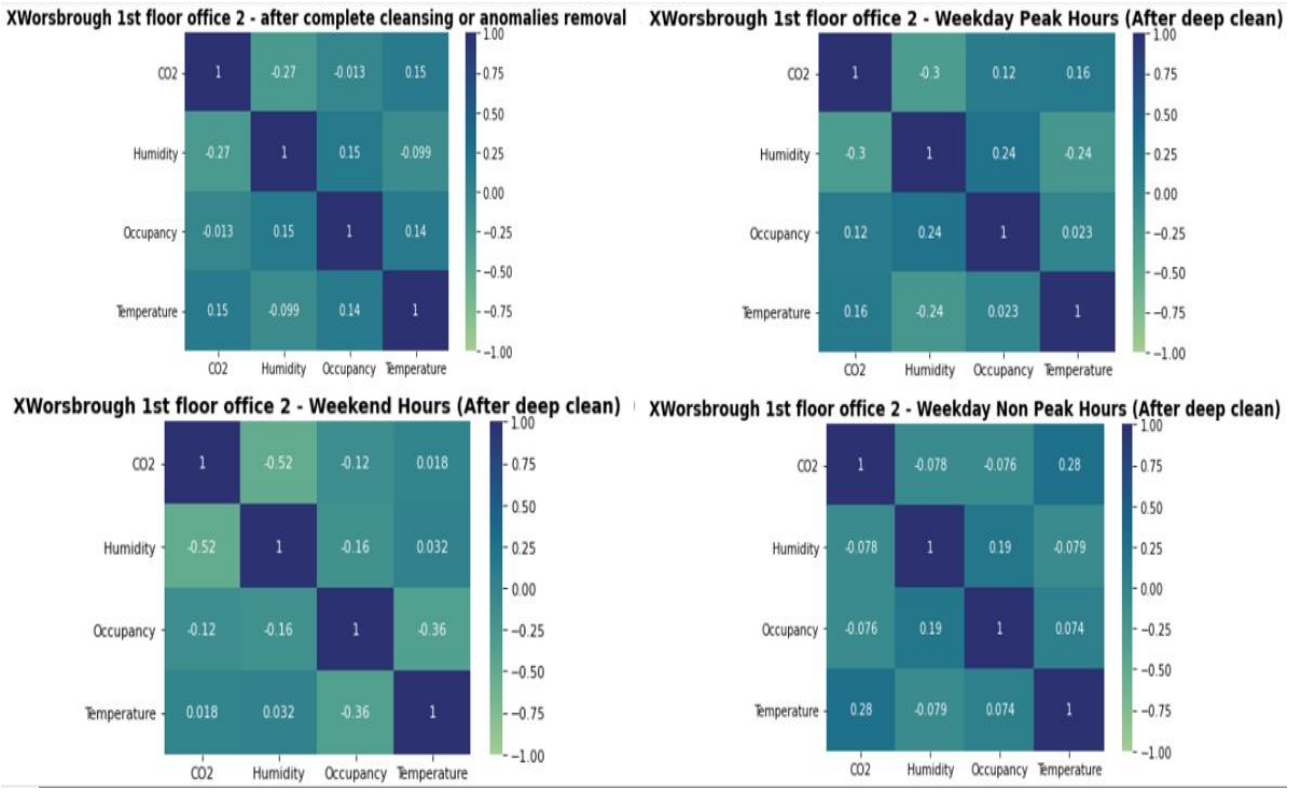


Figure 40 Correlation matrix for XWorsbrough 1st Floor Office 2

With the above Figure 40 correlation matrix the major conclusion that can be made for XWorsbrough 1st Floor Office 2 are: Good negative correlation between metrics CO2 and Humidity with -0.27 for geometry room as a whole, -0.3 for weekday peak hours, -0.078 for weekday non-peak hours and -0.52 (strong) for weekend hours. Except for weekday peak hours, a surprising negative correlation is observed between Occupancy and CO2 with -0.013 for geometry as a whole, -0.076 for weekday non peak hours, -0.12 for weekend hours and 0.12 for weekday non-peak hours. Good negative correlation between metrics Humidity and Temperature with -0.36 for weekend hours.

4. XWestgate building

- Level 1 Open Office (geometry room name)

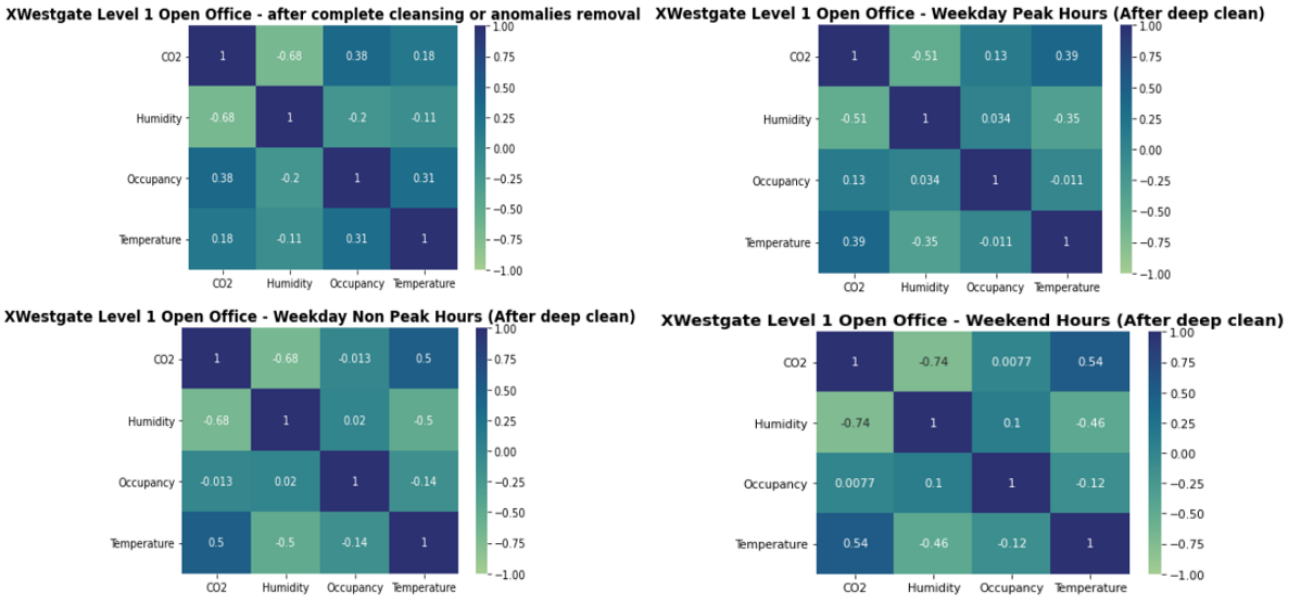


Figure 41 Correlation matrix for XWestgate Level 1 Open Office

From the Figure 41 correlation matrix the major conclusion that can be made for XWestgate Level 1 Open Office are: Very strong and high negative correlation between metrics CO2 and Humidity with -0.68 for geometry room as a whole, -0.51 for weekday peak hours, -0.74 for weekday non-peak hours and -0.68 for weekend hours. Good negative correlation between metrics Temperature and Humidity with -0.11 for geometry room as a whole, -0.35 for weekday peak hours, -0.46 (strong) for weekday non-peak hours and -0.5 (strong) for weekend

hours. Except for weekend hours, a positive correlation between Occupancy and CO2 is observed with 0.38 (good) for geometry as a whole, 0.13 for weekday peak hours, 0.0077 for weekday non-peak hours and -0.013 for weekend hours.

- Level 2 Open Office (geometry room name)

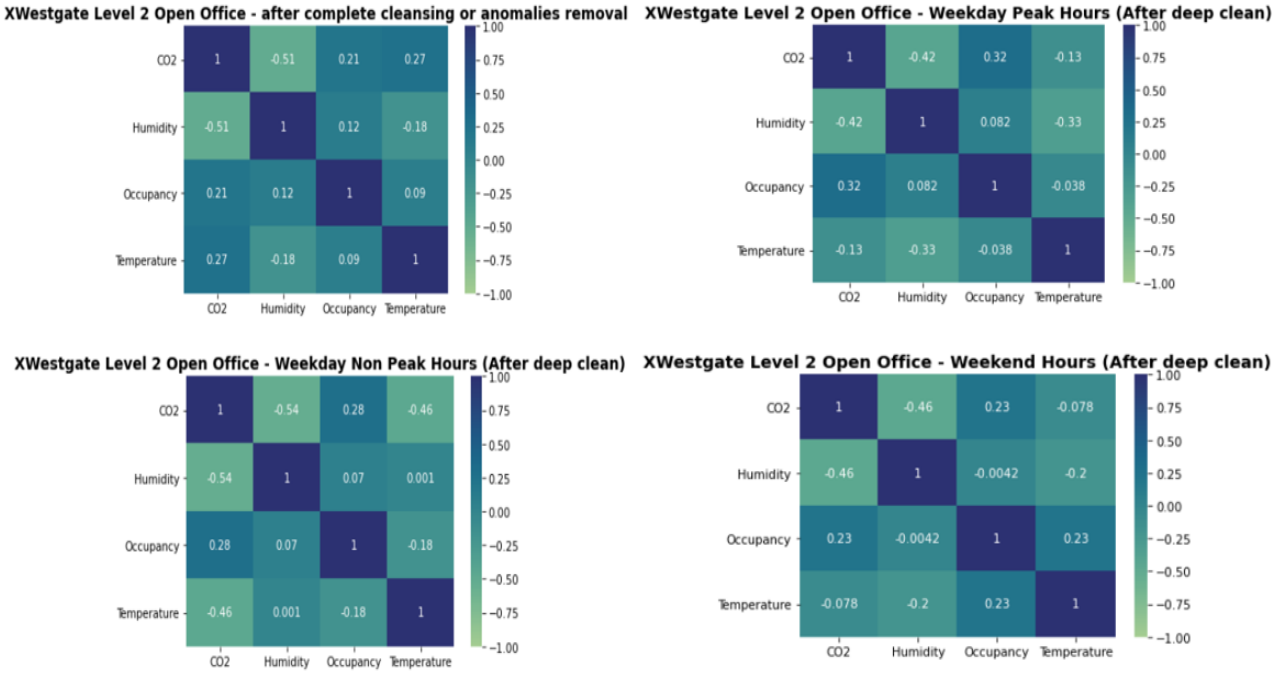


Figure 42 Correlation matrix for XWestgate Level 2 Open Office

With the Figure 42 correlation matrix the major conclusion that can be made for XWestgate Level 2 Open Office are: Strong negative correlation between metrics CO2 and Humidity with -0.51 for geometry room as a whole, -0.42 for weekday peak hours, -0.46 for weekday non-peak hours and -0.54 for weekend hours. Good positive correlation between metrics CO2 and Occupancy with 0.21 for geometry room as a whole, 0.32 for weekday peak hours, 0.23 for weekday non-peak hours and 0.28 for weekend hours. Strong negative correlation between CO2 and Temperature for weekend hours with -0.46. Good negative correlation between Humidity and Temperature for weekday peak hours with -0.33, -0.2 for weekday non-peak and -0.18 for the geometry room as a whole.

- Level 3 Open Office (geometry room name)

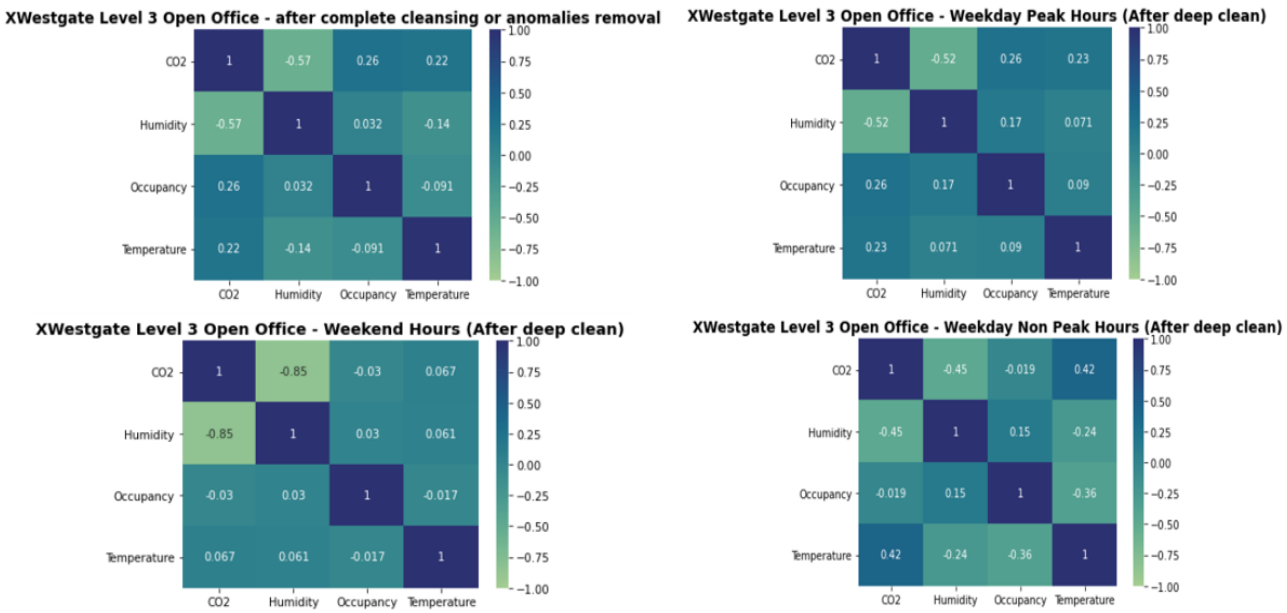


Figure 43 Correlation matrix for XWestgate Level 3 Open Office

From Figure 43 correlation matrix the major conclusion that can be made for XWestgate Level 3 Open Office are: Very strong negative correlation between metrics CO2 and Humidity with -0.57 for geometry room as a whole, -0.52 for weekday peak hours, -0.45 for weekday non-peak hours and -0.85 for weekend hours. In terms of CO2 and Occupancy metrics, XWestgate Level 3 Open Office geometry as a whole and weekday peak hours shared the same positive correlation coefficient 0.26. For weekday non-peak hours, Temperature is negatively correlated with Humidity and Occupancy with -0.24 and -0.36 respectively.

- Level 4 Open Office (geometry room name)

  With the below Figure 44 correlation matrix the major conclusion that can be made for XWestgate Level 4 Open Office are: Very much strong and high negative correlation between metrics CO2 and Humidity with -0.79 for geometry room as a whole, -0.75 for weekday peak hours, -0.75 for weekday non-peak hours and -0.92 for weekend hours. Good negative correlation between metrics Humidity and Temperature with -0.3 for geometry room as a whole, -0.36 for weekday peak hours, -0.33 for weekday non-peak hours and -0.22 for weekend hours. Also, positive correlations between metrics CO2 and Occupancy were

observed except for weekend hours but the coefficient values are not significant(very less association value).

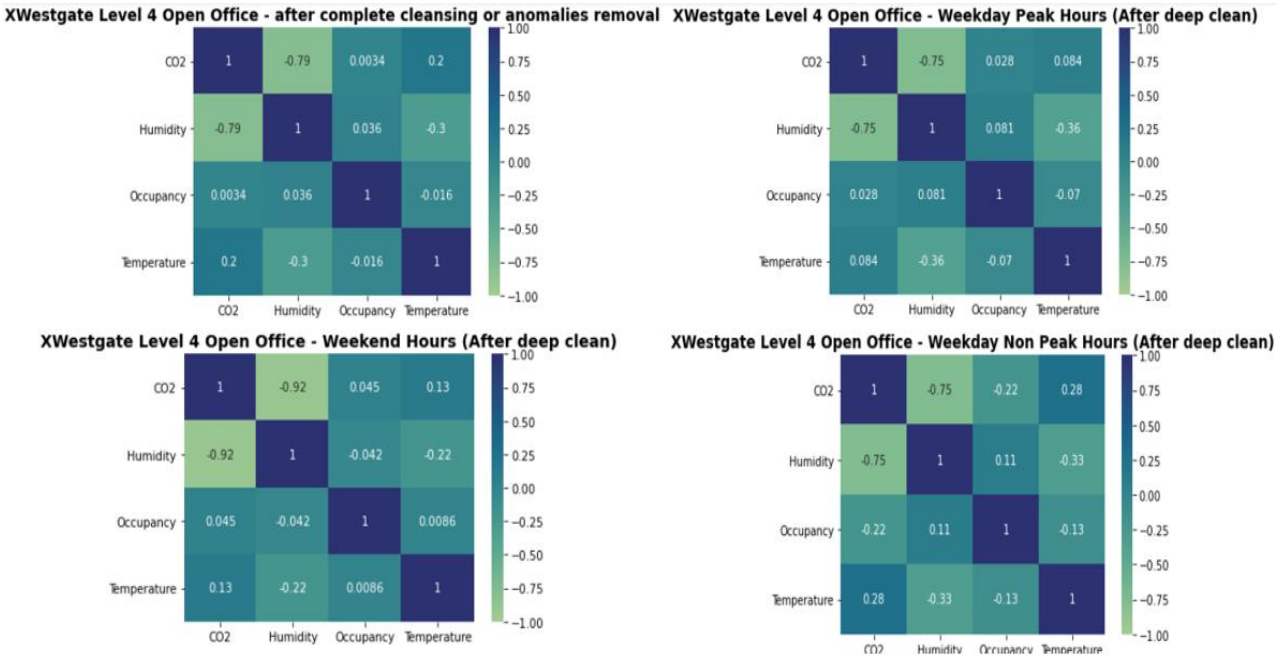


Figure 44 Correlation matrix for XWestgate Level 4 Open Office

- Level 5 Open Office (geometry room name)

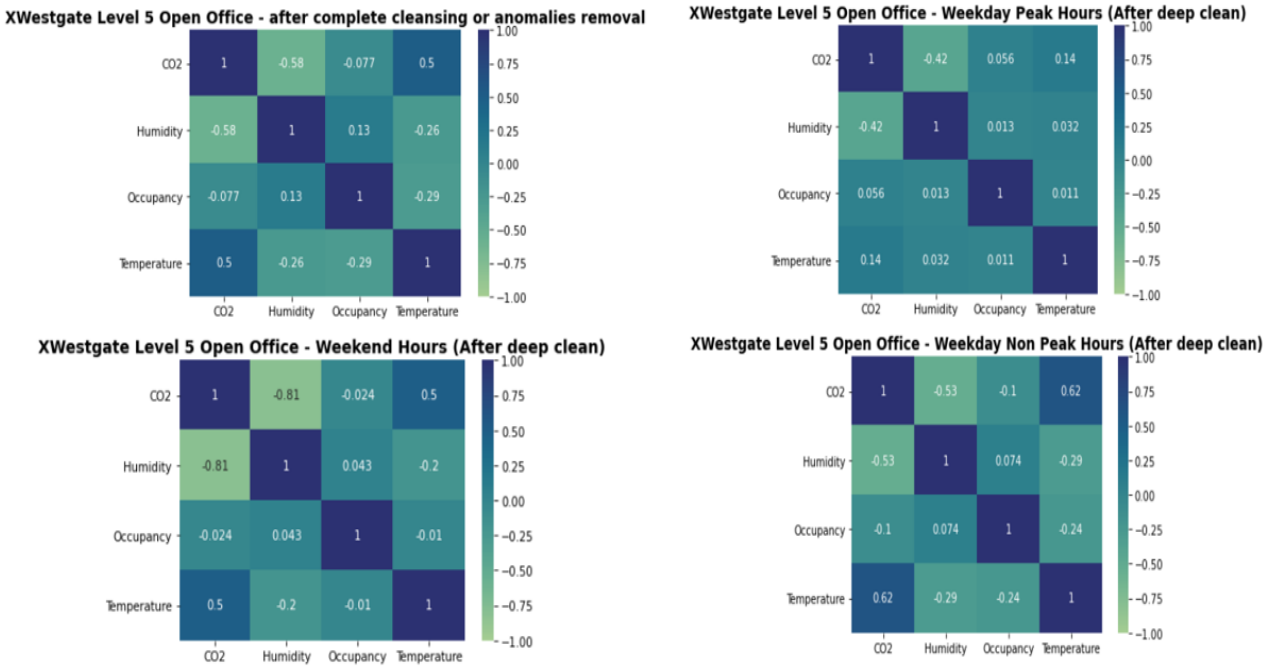


Figure 45 Correlation matrix for XWestgate Level 5 Open Office

From the Figure 45 correlation matrix the major conclusion that can be made for XWestgate Level 5 Open Office are: Very strong and high negative correlation between metrics CO2 and Humidity with -0.58 for geometry room as a whole, -0.42 for weekday peak hours, -0.53 for weekday non-peak hours and -0.81 for weekend hours. Temperature and Humidity metrics also observed a strong correlation or association between them with 0.5 for the geometry considered as a whole, 0.14 for weekday peak hours, 0.62 for weekday non-peak hours and 0.5 for weekend hours. Positive correlation between Humidity and Occupancy metrics were observed with 0.13 for the geometry as a whole and for weekday peak and non-peak, weekend hours but their correlation coefficient values were not that much noteworthy. And, except for weekday peak hours, negative correlations were observed between Occupancy and CO2 but the correlation coefficient values were not significant or notable (very less coefficient value). with -0.013 for geometry as a whole, -0.076 for weekday non peak hours, -0.12 for weekend hours and 0.12 for weekday non-peak hours.

With the help of this detailed or elaborative correlation analysis performed we examined the connections or correlations between the 4 metrics within each of building's geometry rooms dealing with these metrics. Also, for each building's geometry room names and, for its weekday peak and non-peak hours and weekend hours, well-informed decisions can be made based on the meaningful correlations discovered between the metrics CO2, temperature, humidity, and occupancy for each building's internal room structure associated with these 4 metrics along with its weekday hours and weekend hours time-period. This will optimise the use of resources, improve building's sustainability, and create more environmentally friendly structures.

10.2. Optimal Forecasting Model

1. **XCudworth** (Table 8)

	Geometry Names	CO2	Humidity	Temperature	Occupancy
Optimal Forecasting Model	1st Floor Meeting Room 3	Prophet	SARIMA	Prophet	SARIMA
	1st Floor Open Office	ARIMA	SARIMA	Prophet	SARIMA
	2nd Floor Meeting Room 1	ARIMA	SARIMA	Prophet	SARIMA
	2nd Floor Open Office	Prophet	SARIMA	SARIMA	Prophet

Table 8 Optimal forecasting model for each metric within XCudworth

2. **XTownhall** (Table 9)

	Geometry Names	CO2	Humidity	Temperature	Occupancy
Optimal Forecasting Model	1st Floor Meeting Room 1	ARIMA	SARIMA	ARIMA	SARIMA
	1st Floor Meeting Room 2	SARIMA	SARIMA	SARIMA	SARIMA

Table 9 Optimal forecasting model for each metric within XTownhall

3. **XWorsbrough** (Table 10)

	Geometry Names	CO2	Humidity	Temperature	Occupancy
Optimal Forecasting Model	1st Floor Office 1	ARIMA	SARIMA	SARIMA	Prophet
	1st Floor Office 2	ARIMA	Prophet	ARIMA	Prophet

Table 10 Optimal forecasting model for each metric within XWorsbrough

4. **XWestgate** (Table 11)

	Geometry Names	CO2	Humidity	Temperature	Occupancy
Optimal Forecasting Model	Level 1 Open Office	Prophet	ARIMA	Prophet	ARIMA
	Level 2 Open Office	ARIMA	ARIMA	ARIMA	Prophet
	Level 3 Open Office	ARIMA	ARIMA	SARIMA	Prophet
	Level 4 Open Office	ARIMA	SARIMA	Prophet	ARIMA
	Level 5 Open Office	Prophet	Prophet	Prophet	ARIMA

Table 11 Optimal forecasting model for each metric within XWestgate

By selecting the best forecasting model for each parameter resources like electric power (lighting), heat, and cooling may be distributed more effectively with reducing wastage. It can provide a green environment with lower carbon emissions and building emissions. Additionally, proactive sustainable planning may be implemented to raise building productivity and comfort.

10.3. Visualisation graphs

The actual vs predicted values graph and future date forecasting graph for each metric within each building's geometry is illustrated here;

1. XWestgate building

- Level 1 Open Office

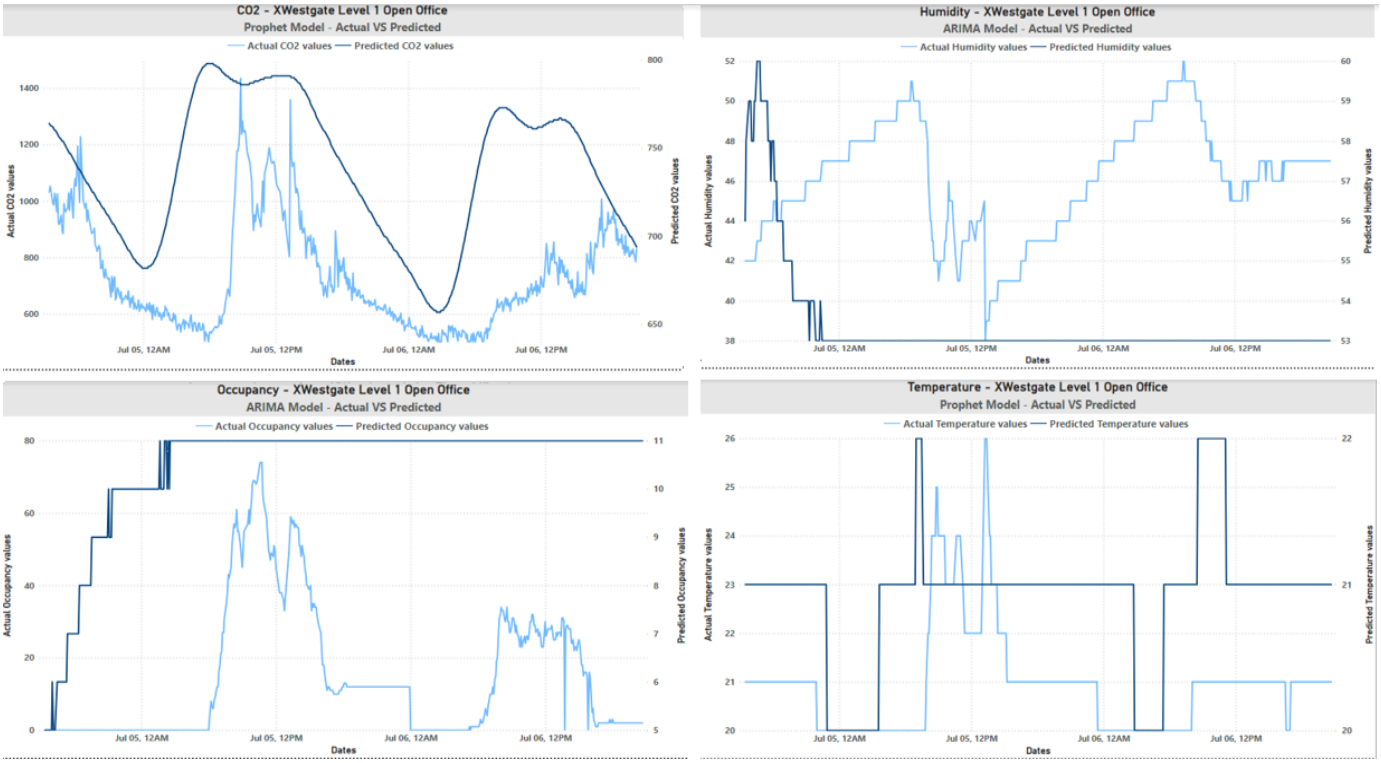


Figure 46 Actual VS Predicted graph for Level 1 Open Office

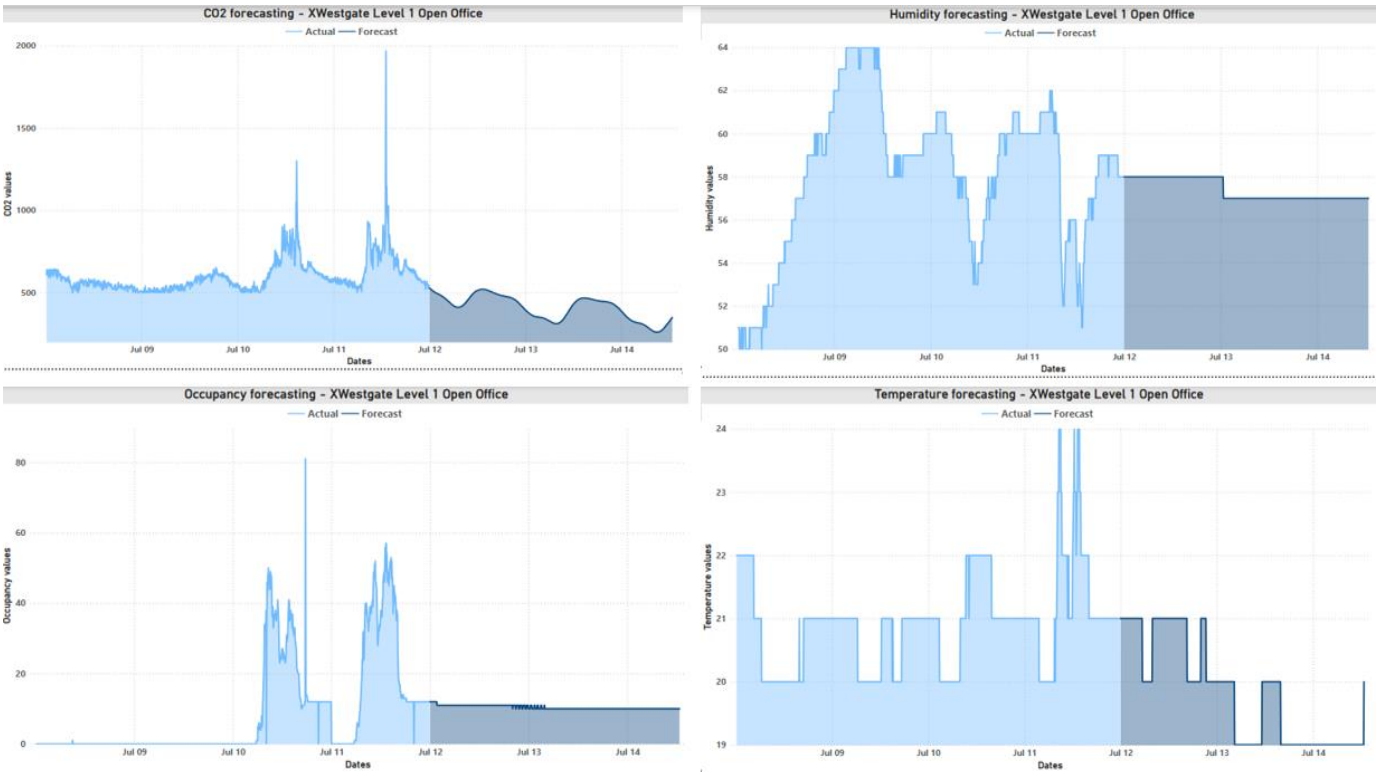


Figure 47 Forecasting graph for Level 1 Open Office

Figure 46, Figure 47 concludes that the determined optimal models for XWestgate Level 1 Open Office is accurate and efficient in terms of forecasting or prediction.

- Level 2 Open Office

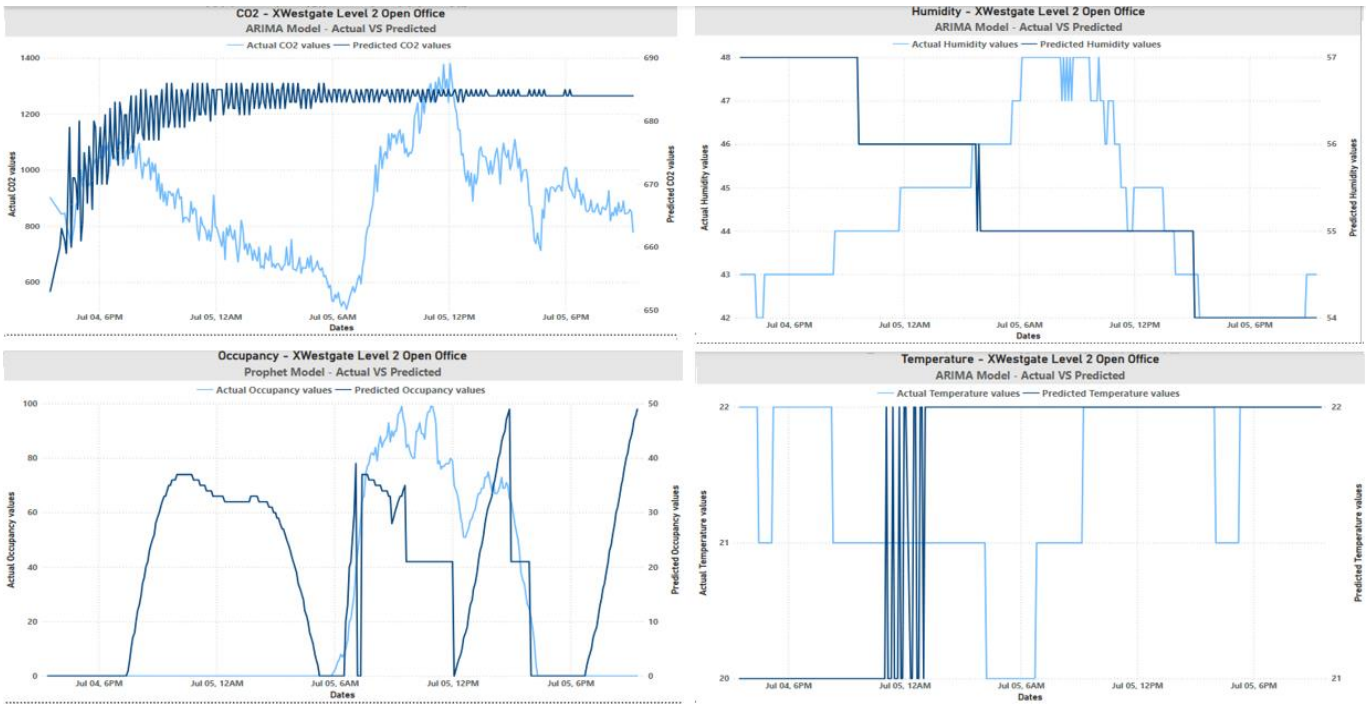


Figure 48 Actual VS Predicted graph for Level 2 Open Office

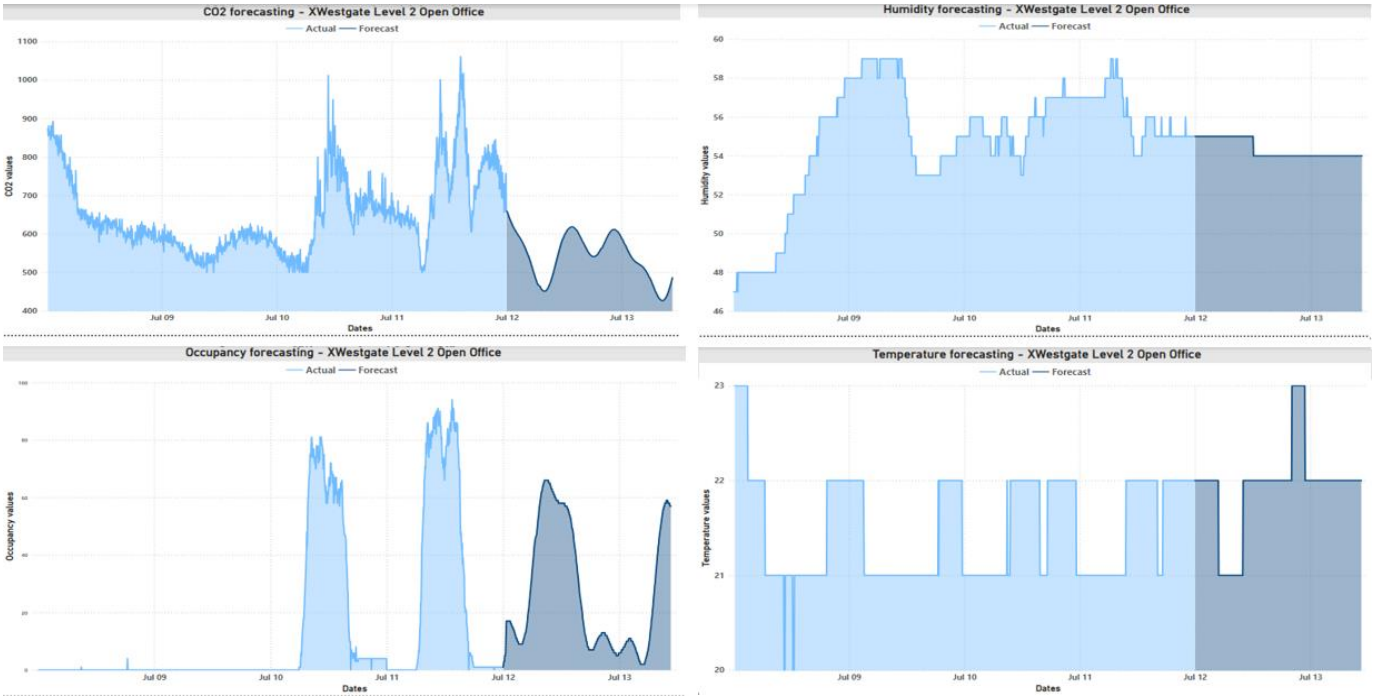


Figure 49 Forecasting graph for Level 2 Open Office

Figure 48, Figure 49 concludes that the determined optimal models for XWestgate Level 2 Open Office is accurate and efficient in terms of predictions.

- Level 3 Open Office

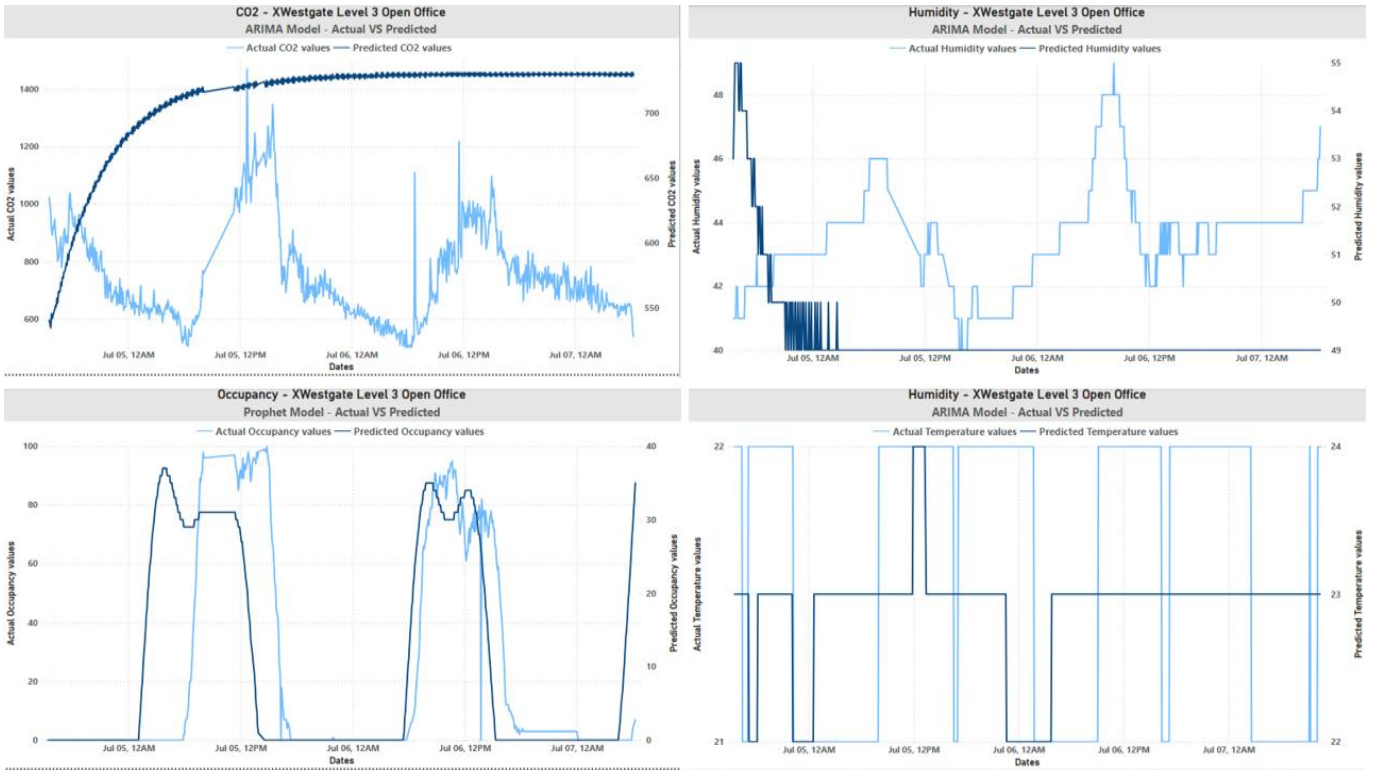


Figure 50 Actual VS Predicted graph for Level 3 Open Office

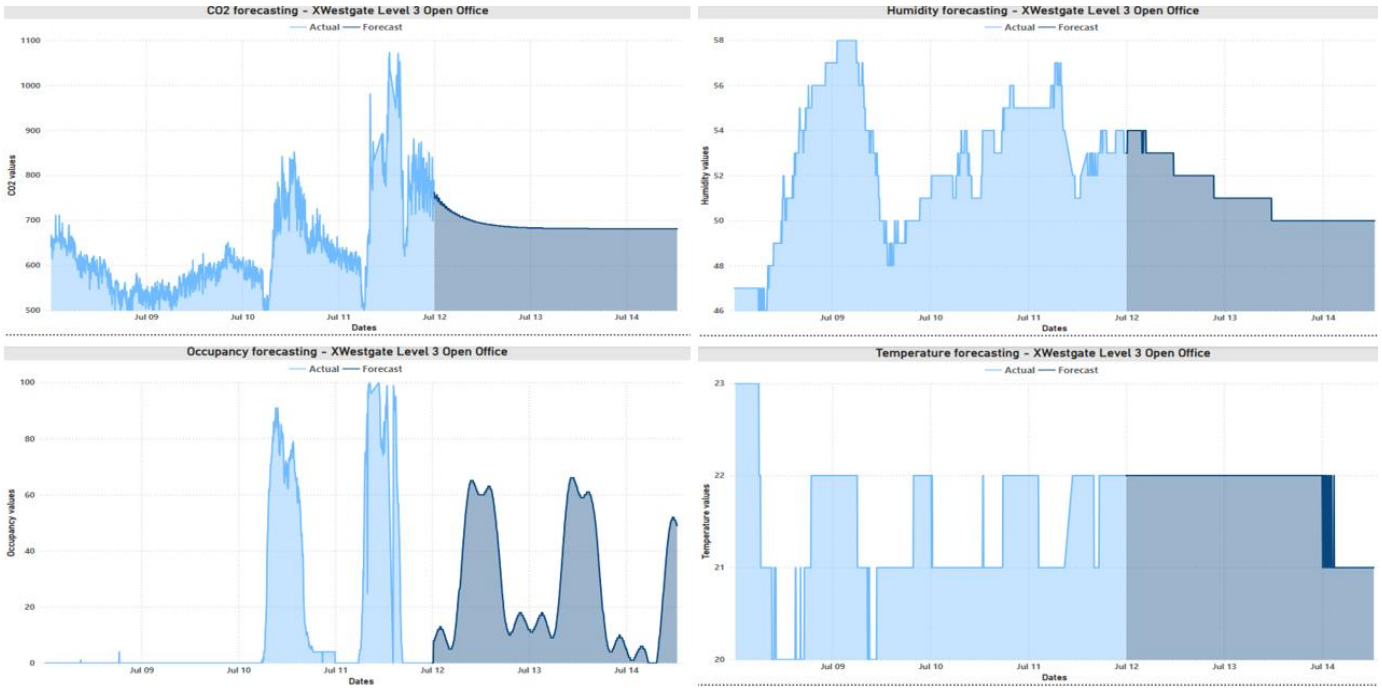


Figure 51 Forecasting graph for Level 3 Open Office

Figure 50, Figure 51 concludes that the determined optimal models for XWestgate Level 3 Open Office is accurate and efficient in terms of predictions.

- Level 4 Open Office

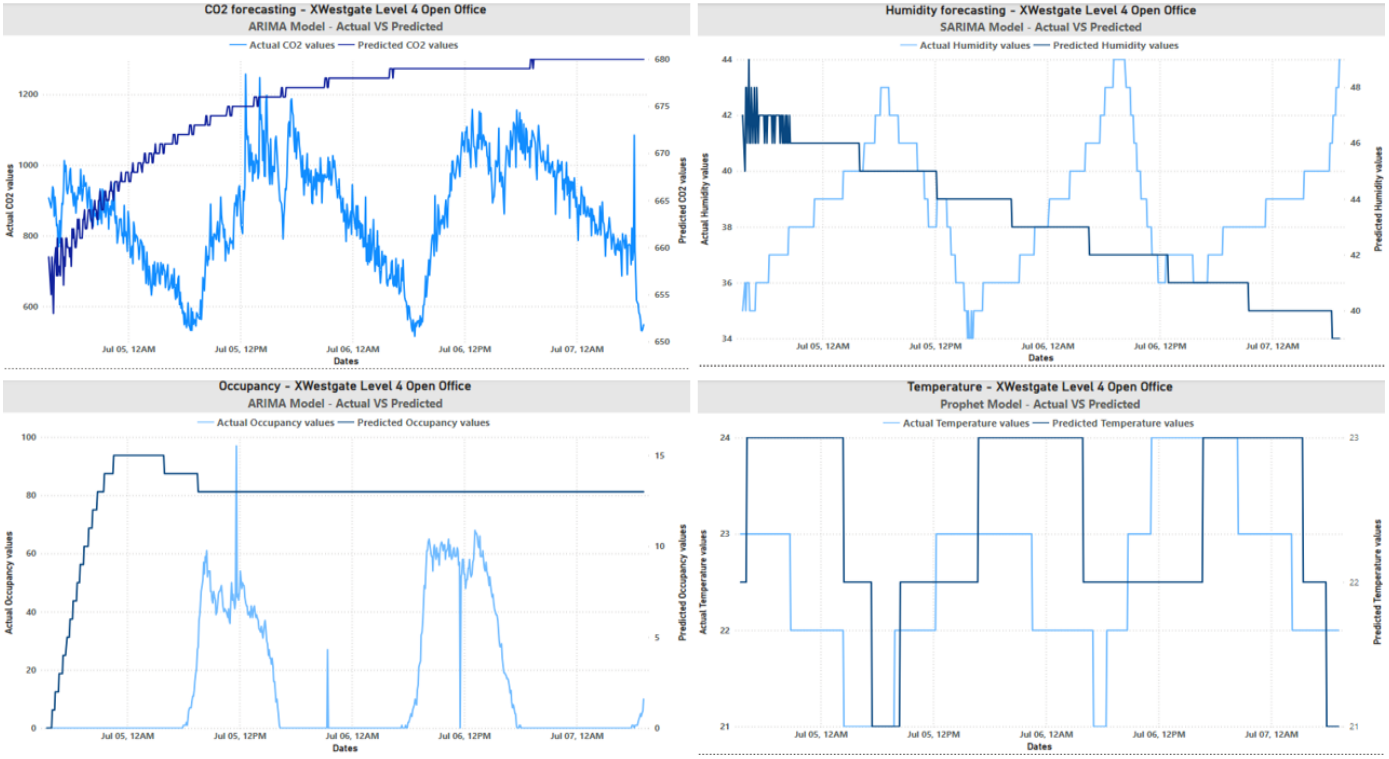


Figure 52 Actual VS Predicted graph for Level 4 Open Office

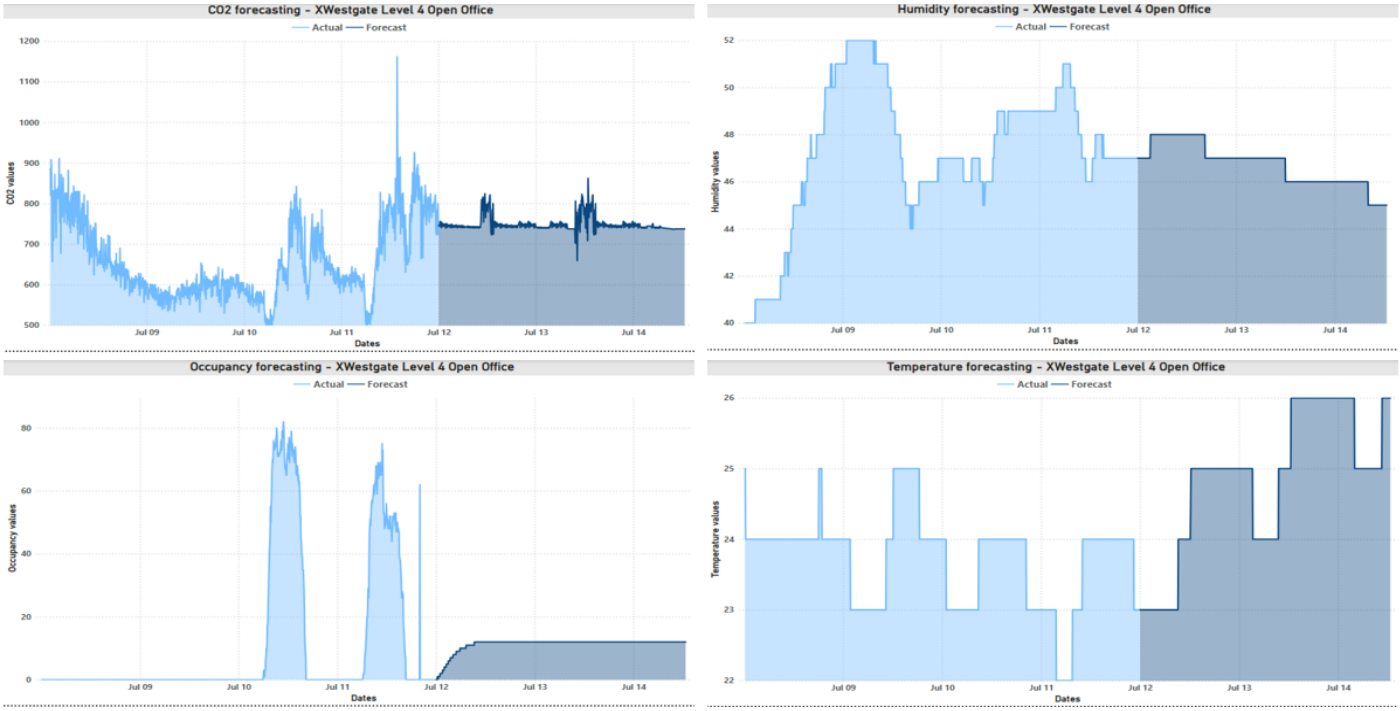


Figure 53 Forecasting graph for Level 4 Open Office

Figure 52, Figure 53 concludes that the determined optimal models for XWestgate Level 4 Open Office is accurate and efficient in terms of predictions.

- Level 5 Open Office

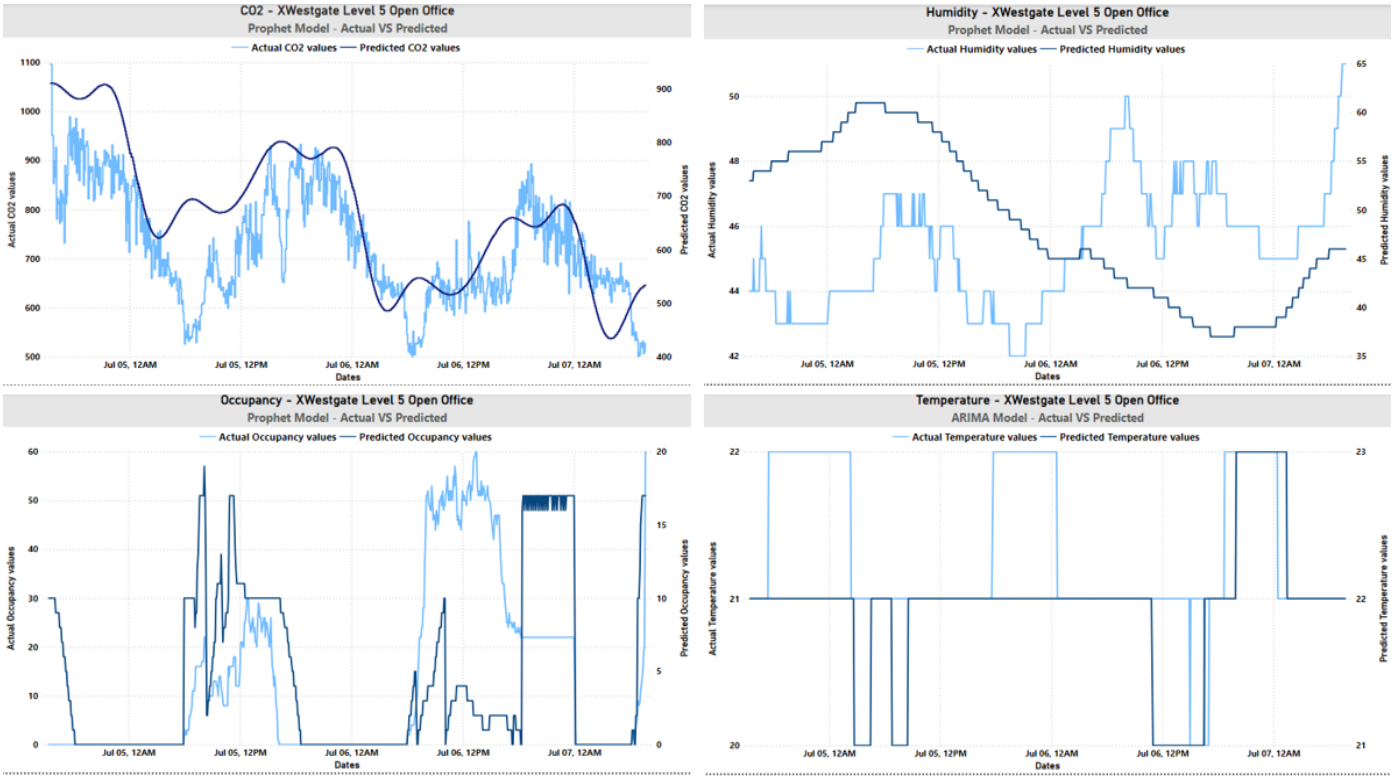


Figure 54 Actual VS Predicted graph for Level 5 Open Office

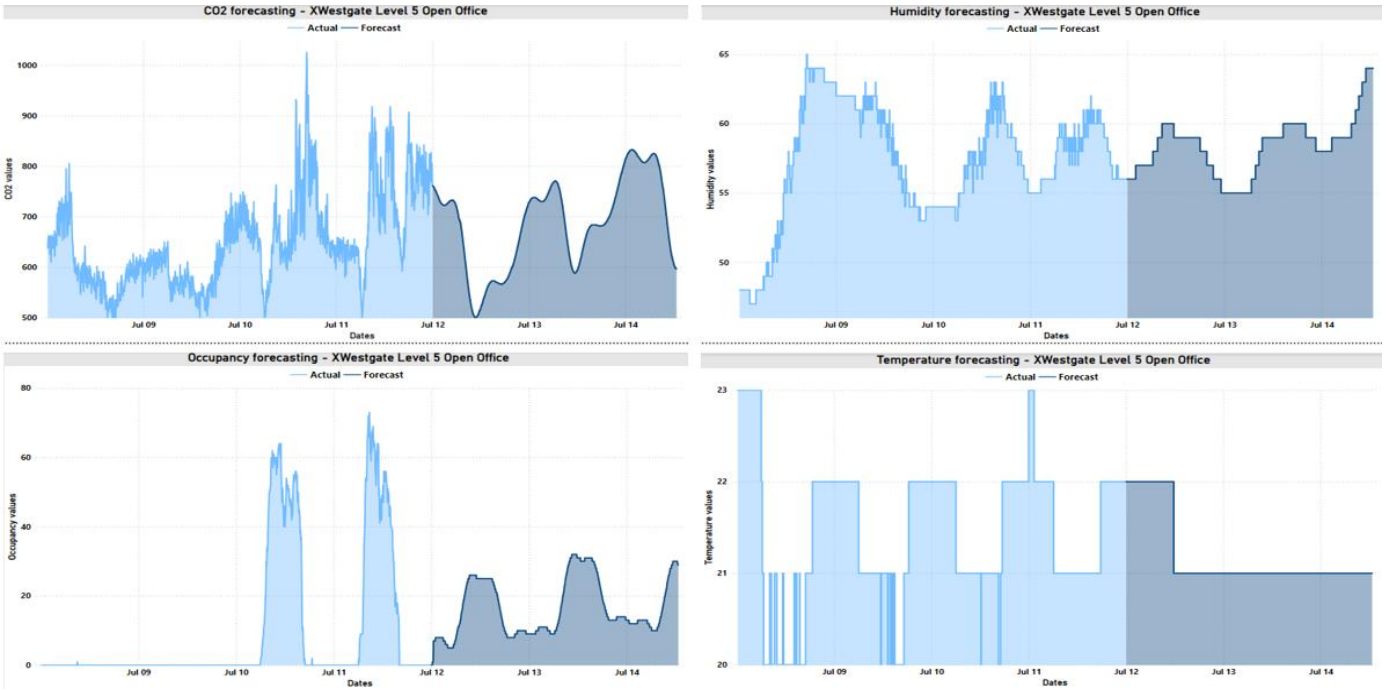


Figure 55 Forecasting graph for Level 5 Open Office

Figure 54, Figure 55 concludes that the determined optimal models for XWestgate Level 5 Open Office is accurate and efficient in terms of predictions.

2. XWorsbrough building

- 1st Floor Office 1

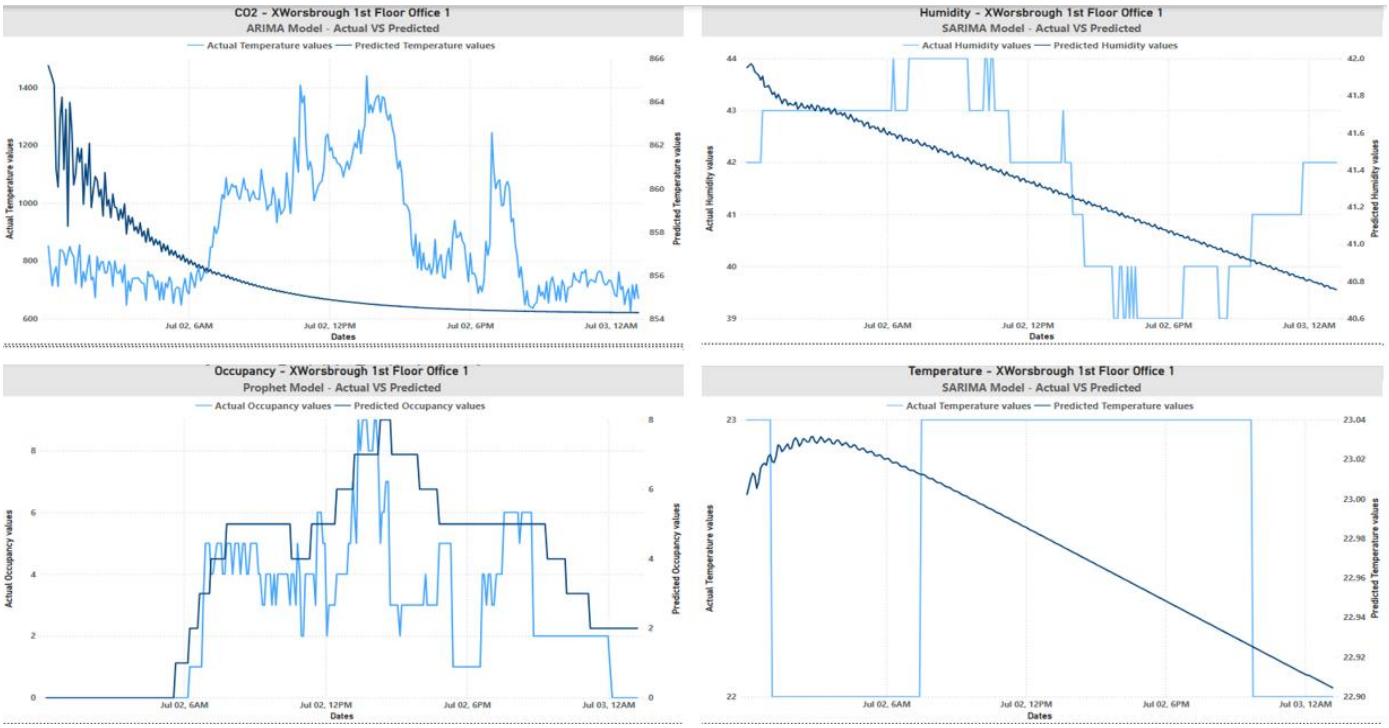


Figure 56 Actual VS Predicted graph for 1st Floor Office 1

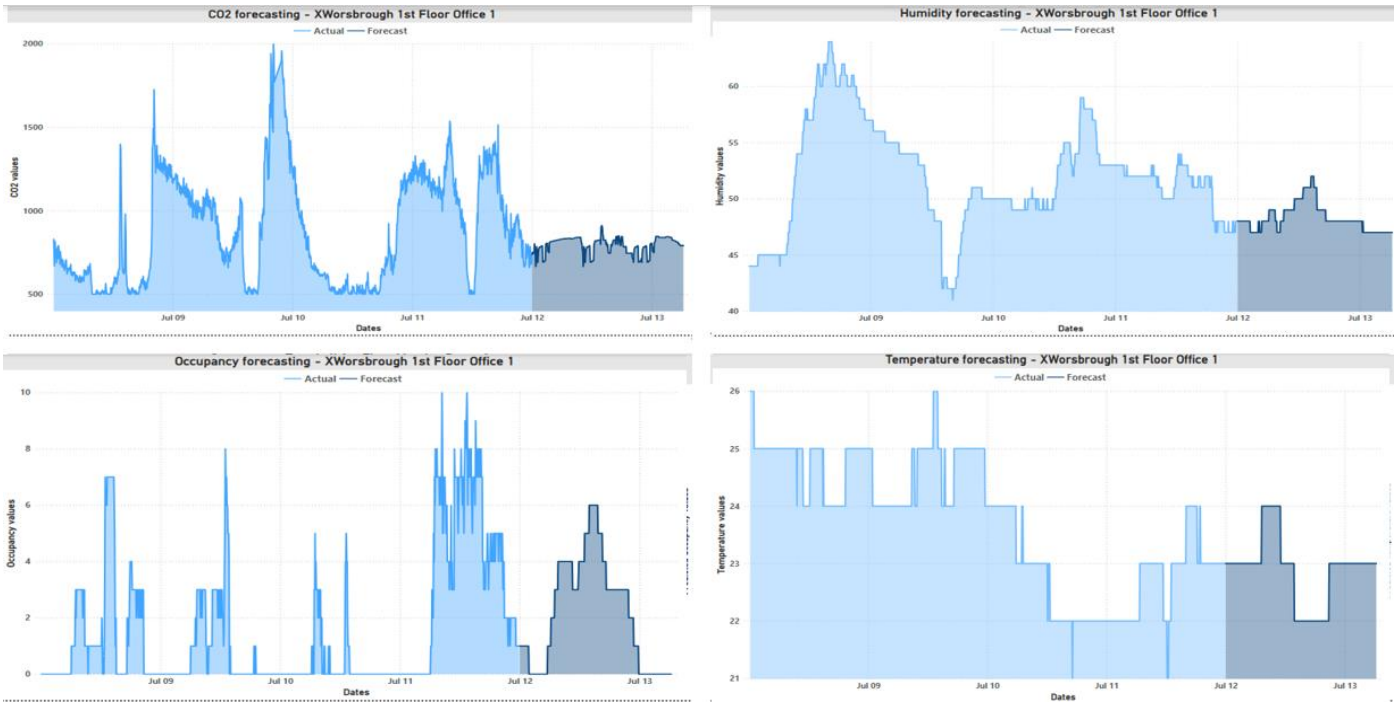


Figure 57 Forecasting graph for 1st Floor Office 1

Figure 56, Figure 57 concludes that the determined optimal models for XWorsbrough 1st Floor Office 1 is accurate and efficient in terms of predictions.

- 1st Floor Office 2

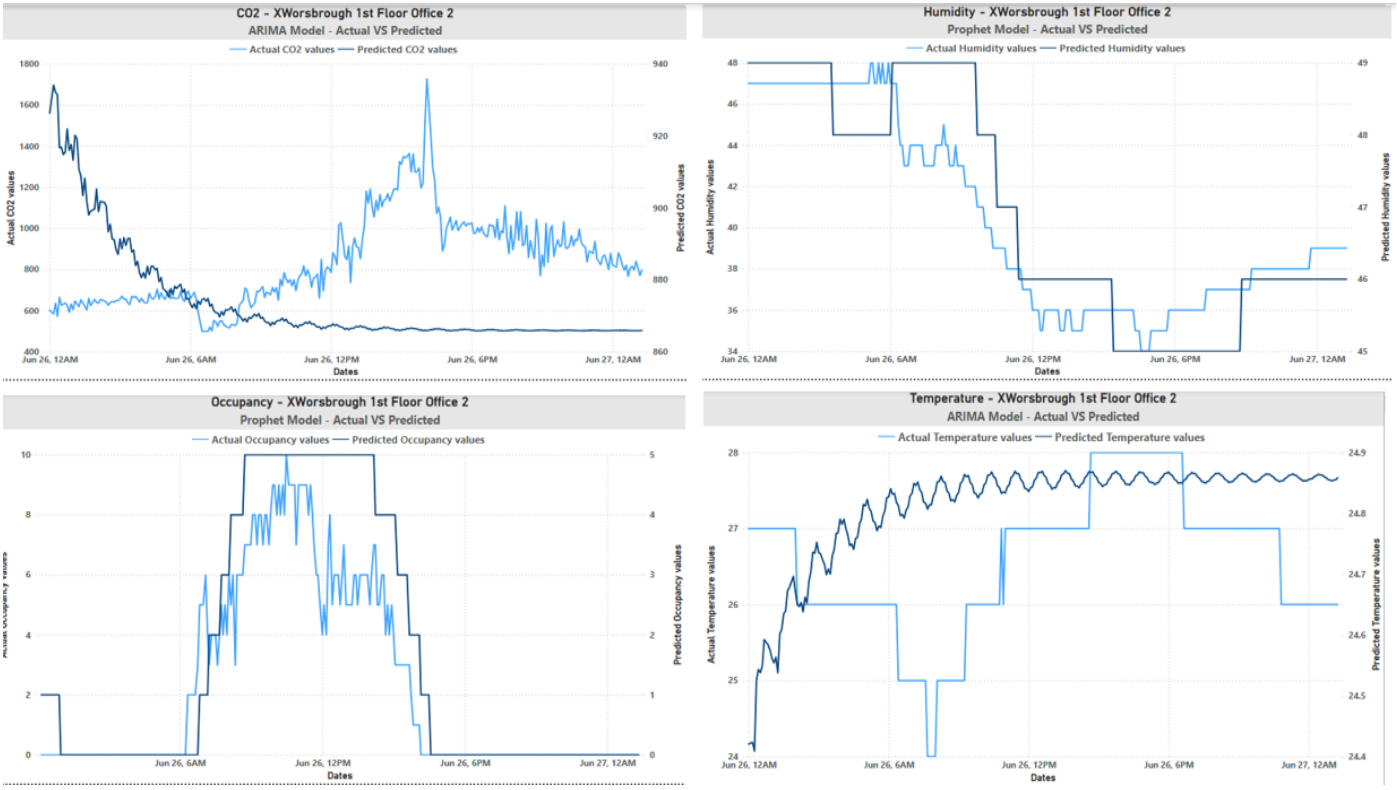


Figure 58 Actual VS Predicted graph for 1st Floor Office 2

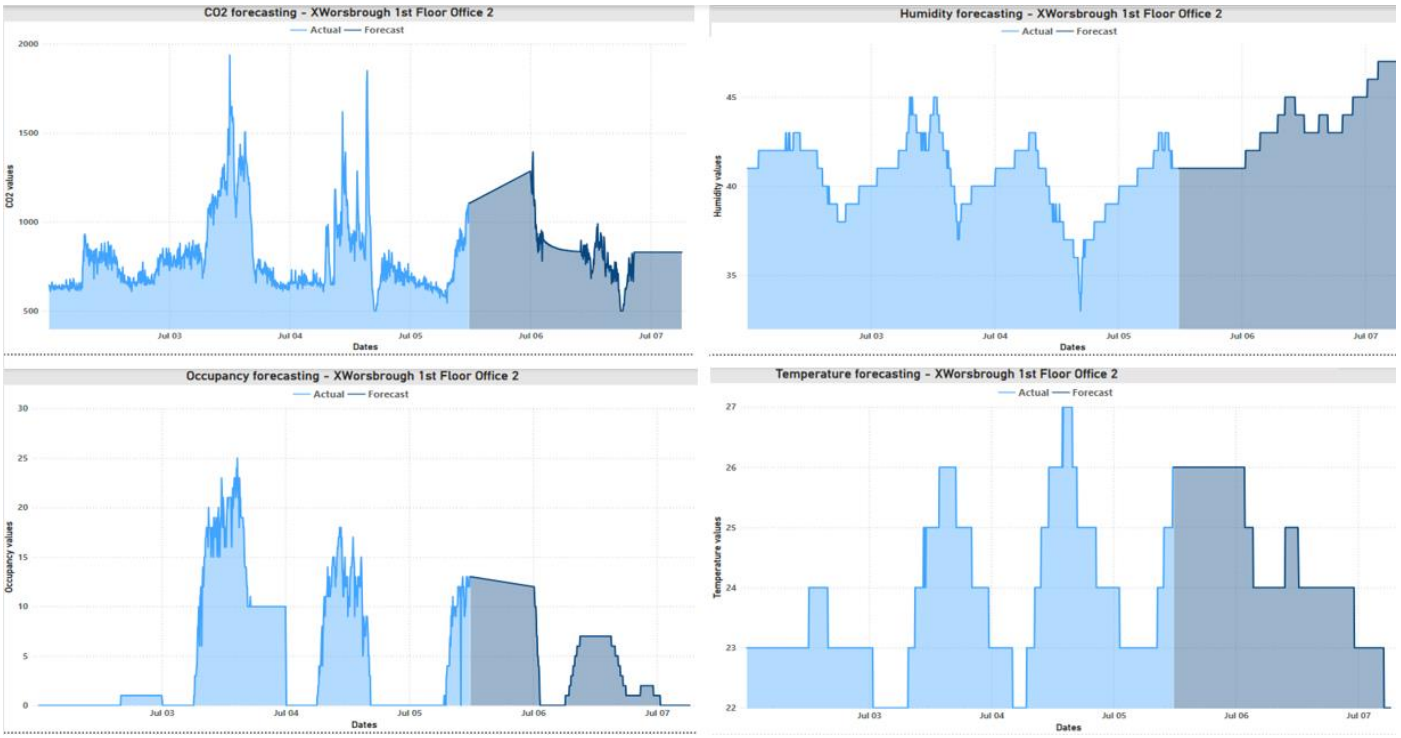


Figure 59 Forecasting graph for 1st Floor Office 2

Figure 58, Figure 59 concludes that the determined optimal models for XWorsbrough 1st Floor Office 2 is accurate and efficient in terms of predictions.

3. XCudworth building

- 1st Floor Open Office

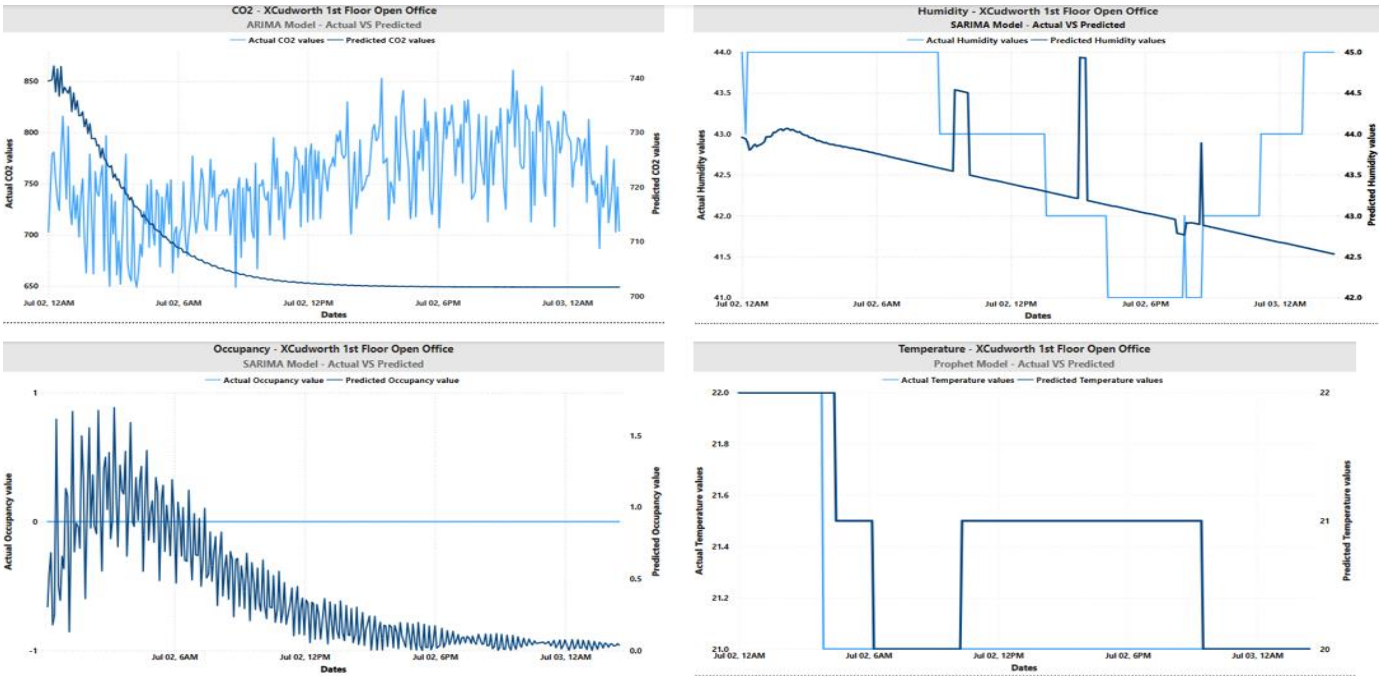


Figure 60 Actual VS Predicted graph for 1st Floor Open Office

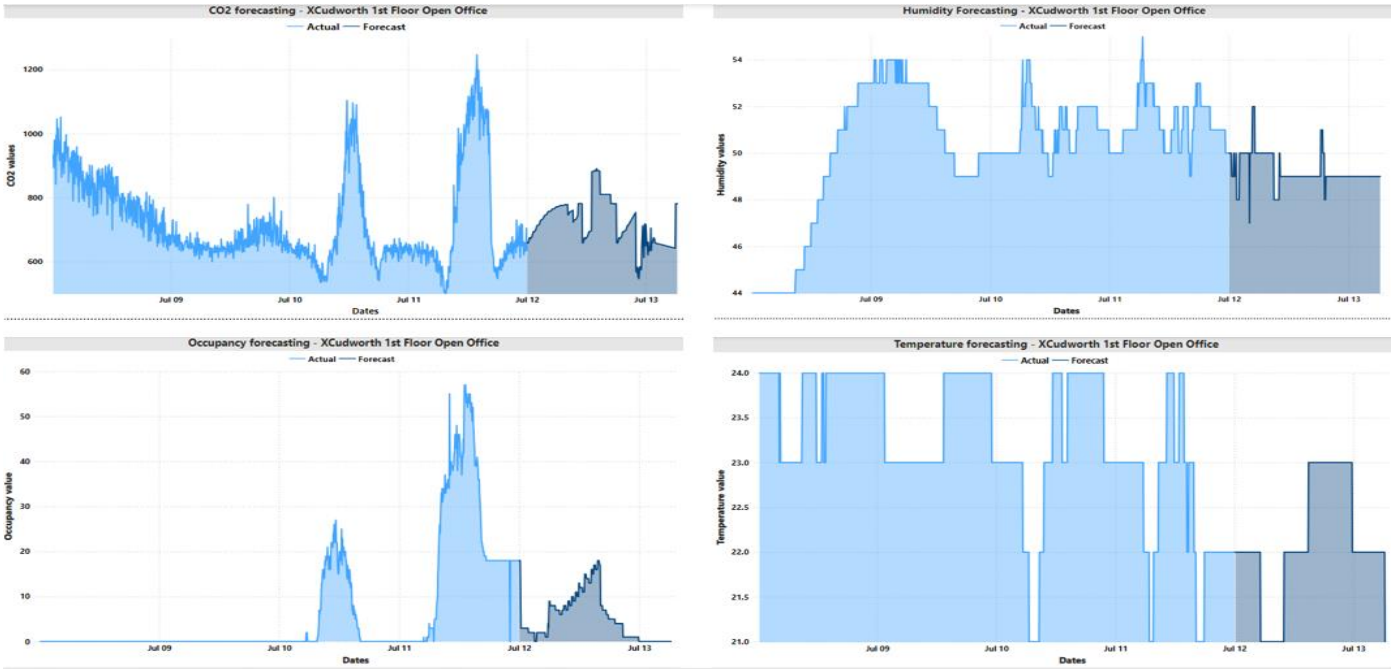


Figure 61 Forecasting graph for 1st Floor Open Office

Figure 60, Figure 61 concludes that the determined optimal models for XCudworth 1st Floor Open Office is accurate and efficient in terms of predictions.

- 1st Floor Meeting Room 3

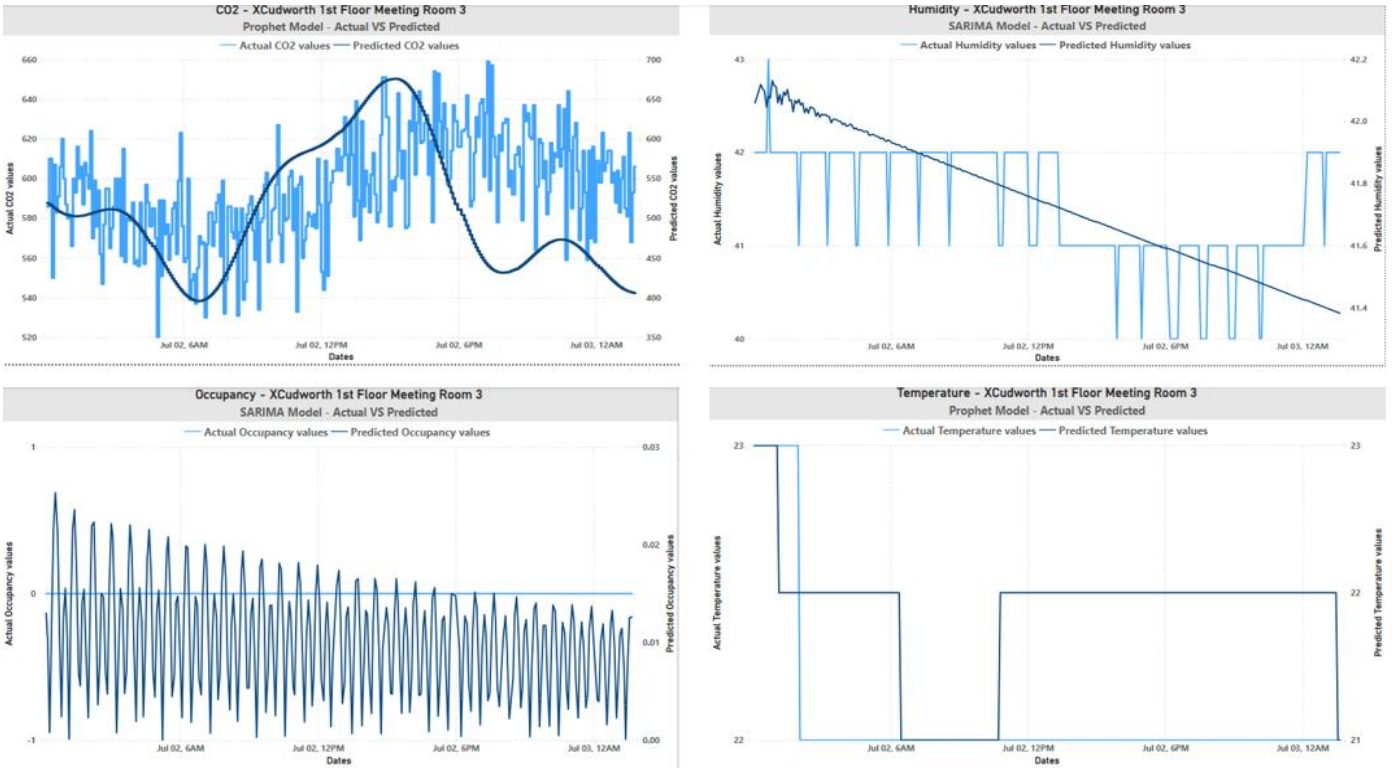


Figure 62 Actual VS Predicted graph for 1st Floor Meeting Room 3

Figure 63 Forecasting graph for 1st Floor Meeting Room 3

Figure 62, Figure 63 concludes that the determined optimal models for XCudworth 1st Floor Meeting Room 3 is accurate and efficient in terms of predictions.

- 2nd Floor Open Office

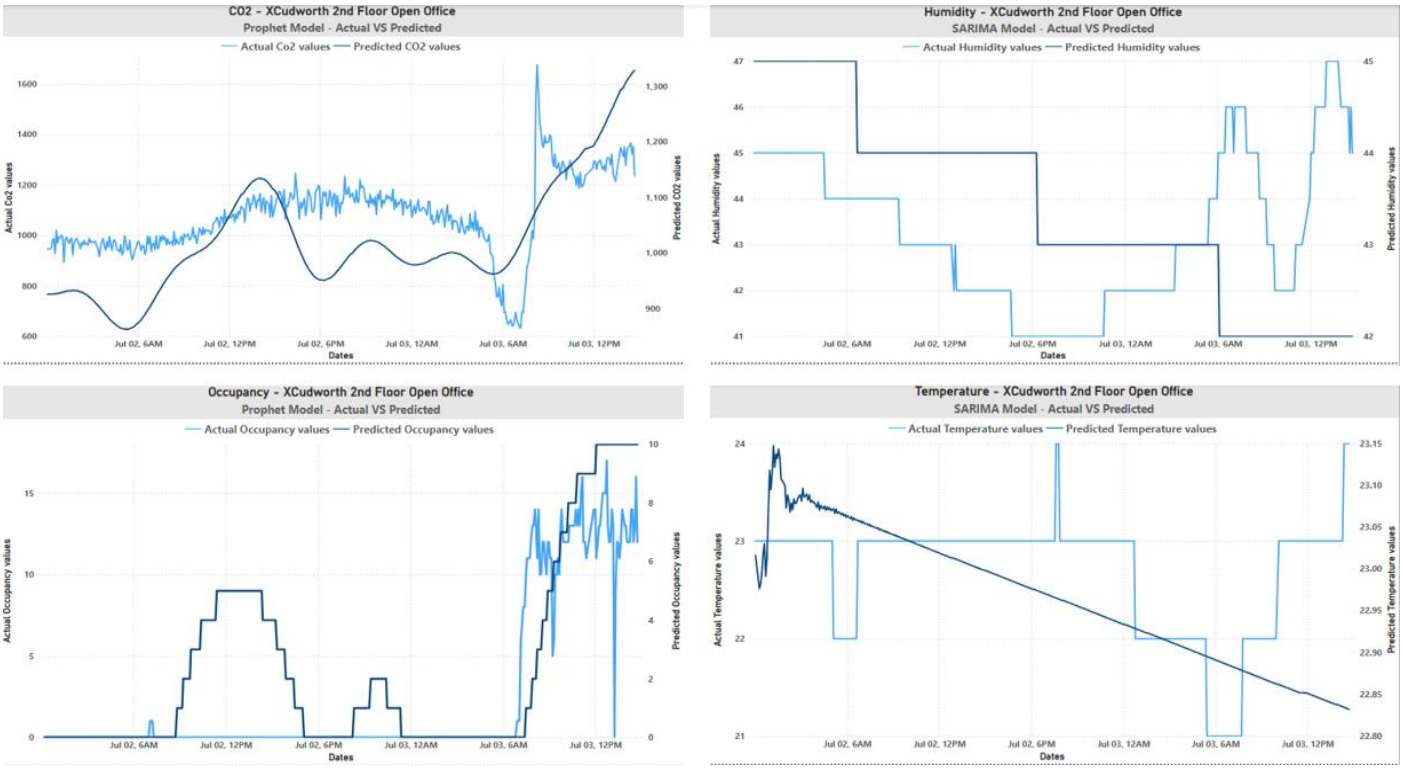


Figure 64 Actual VS Predicted graph for 2nd Floor Open Office

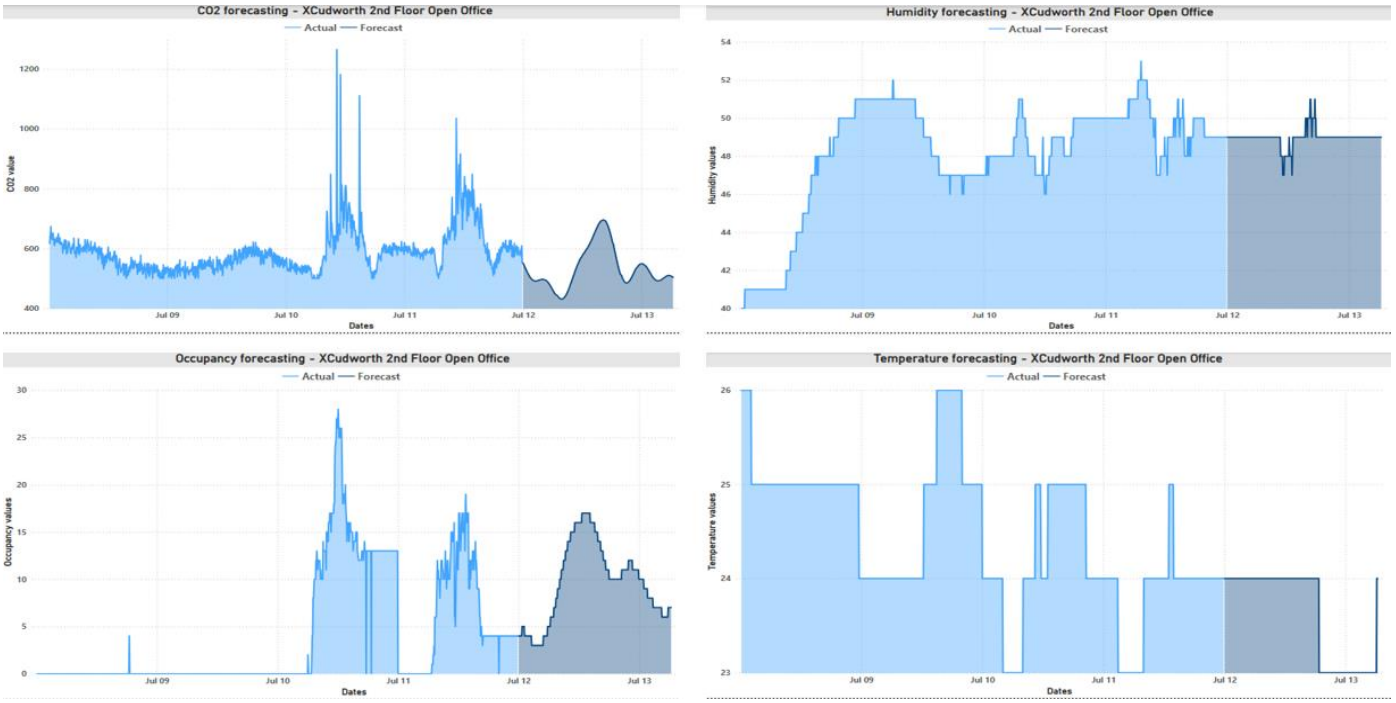


Figure 65 Forecasting graph for 2nd Floor Open Office

Figure 64, Figure 65 concludes that the determined optimal models for XCudworth 2nd Floor Open Office is accurate and efficient in terms of predictions.

- 2nd Floor Meeting Room 1

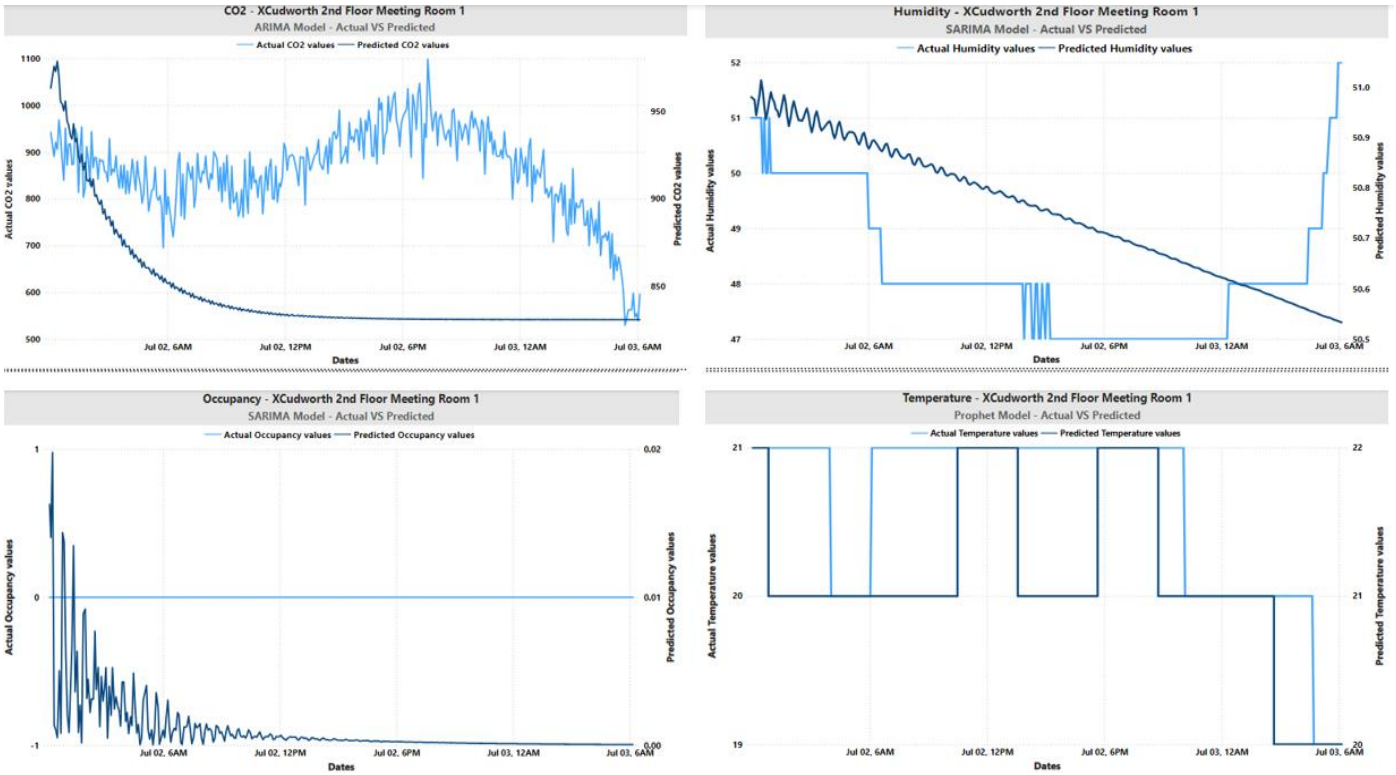


Figure 66 Actual VS Predicted graph for 2nd Floor Meeting Room 1

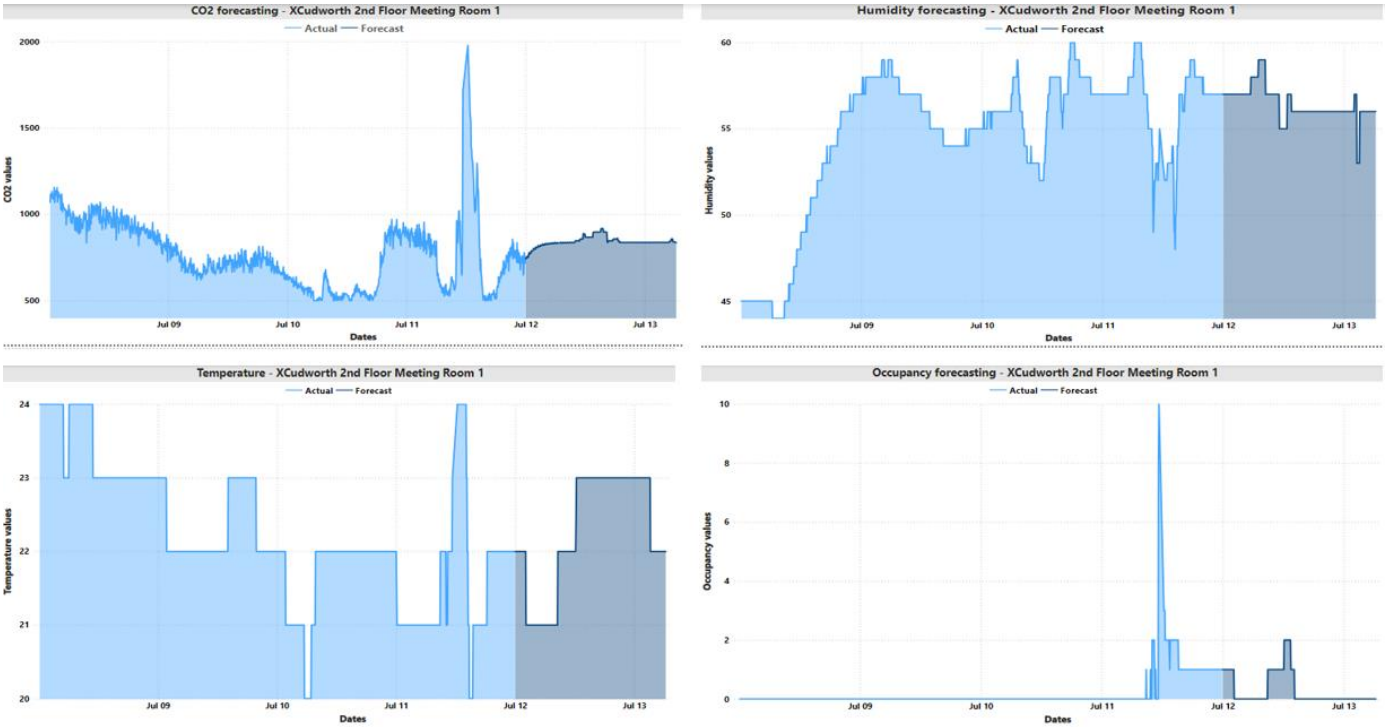


Figure 67 Forecasting graph for 2nd Floor Meeting Room 1

Figure 66, Figure 67 concludes that the determined optimal models for XCudworth 2nd Floor Meeting Room 1 is accurate and efficient in terms of predictions.

4. XTownhall building

- 1st Floor Meeting Room 1

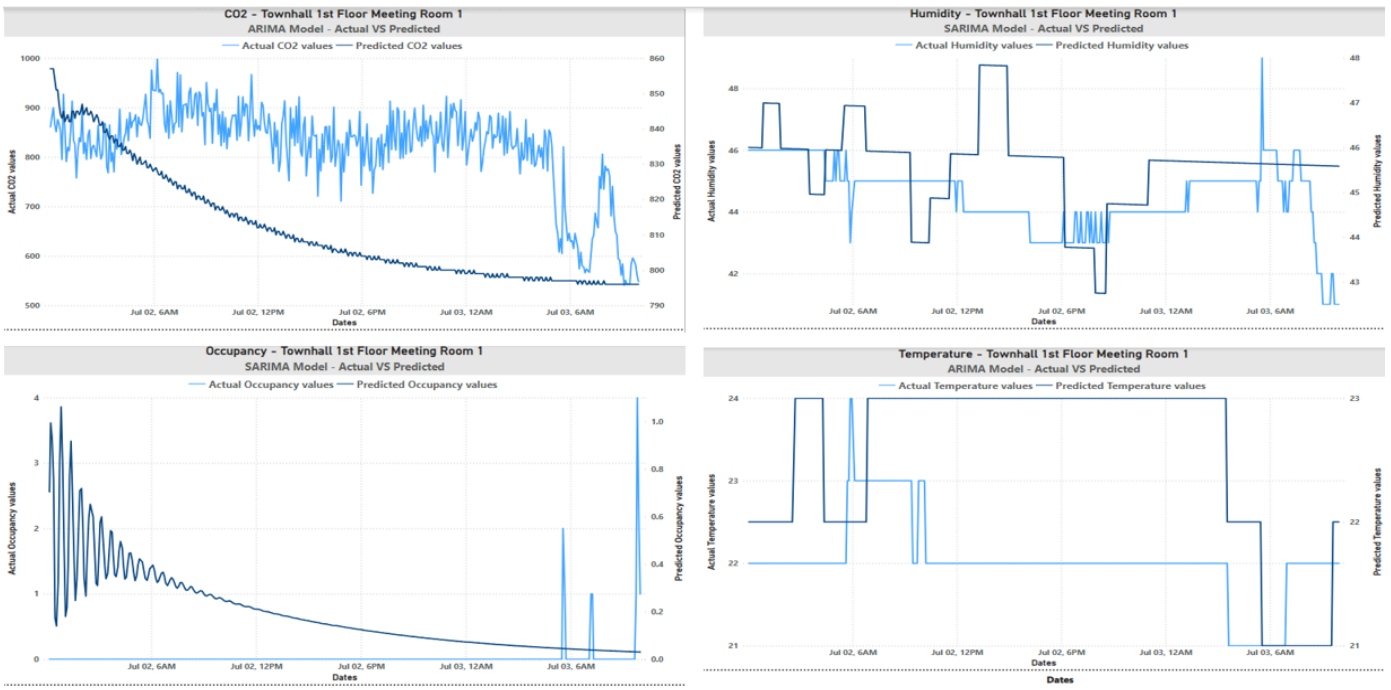


Figure 68 Actual VS Predicted graph for 1st Floor Meeting Room 1

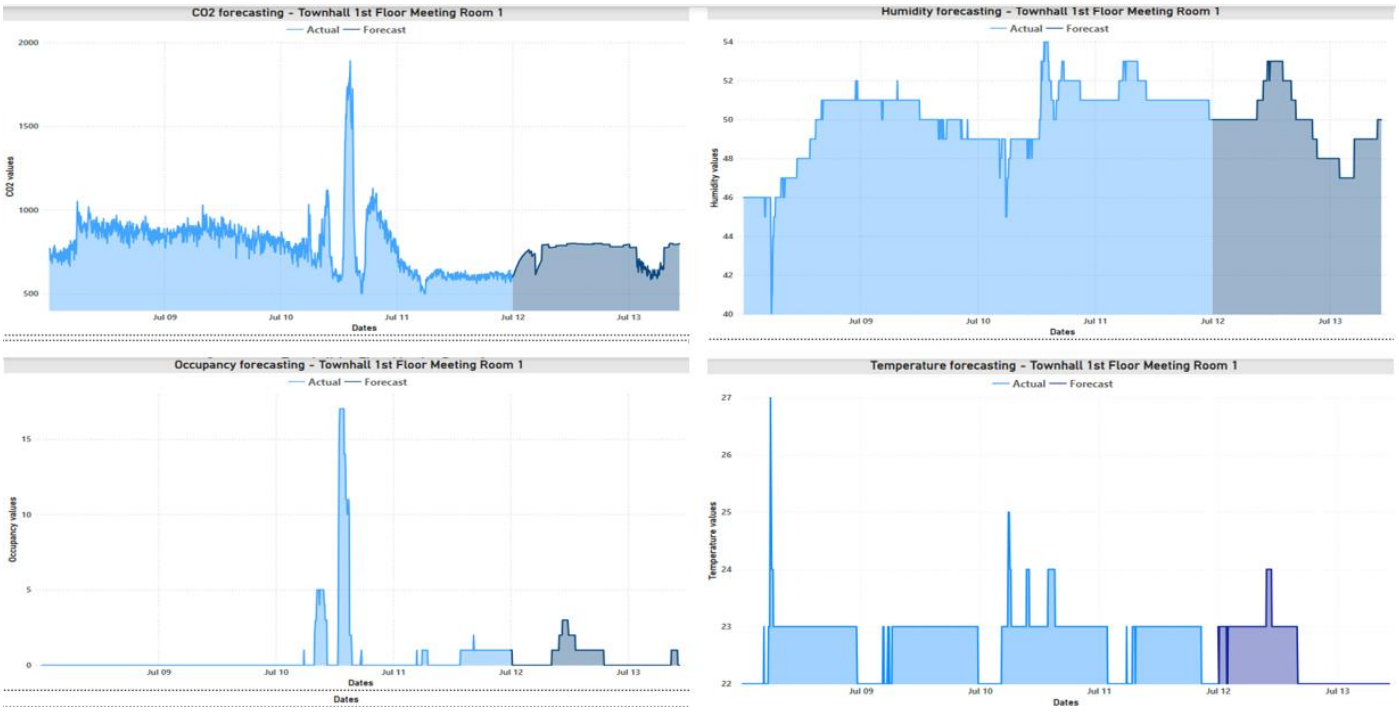


Figure 69 Forecasting graph for 1st Floor Meeting Room 1

Figure 68, Figure 69 concludes that the determined optimal models for XTownhall 1st Floor Meeting Room 1 is accurate and efficient in terms of predictions.

- 1st Floor Meeting Room 2

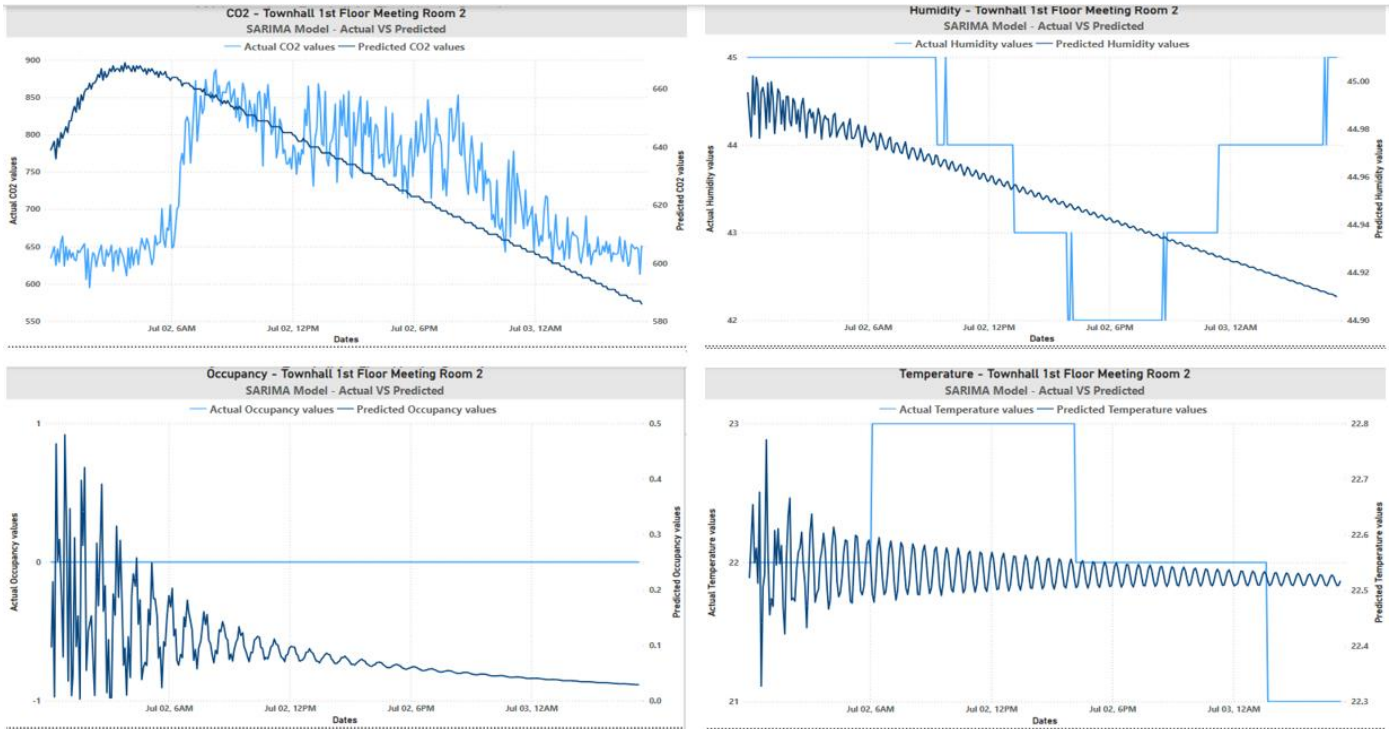


Figure 70 Actual VS Predicted graph for 1st Floor Meeting Room 2

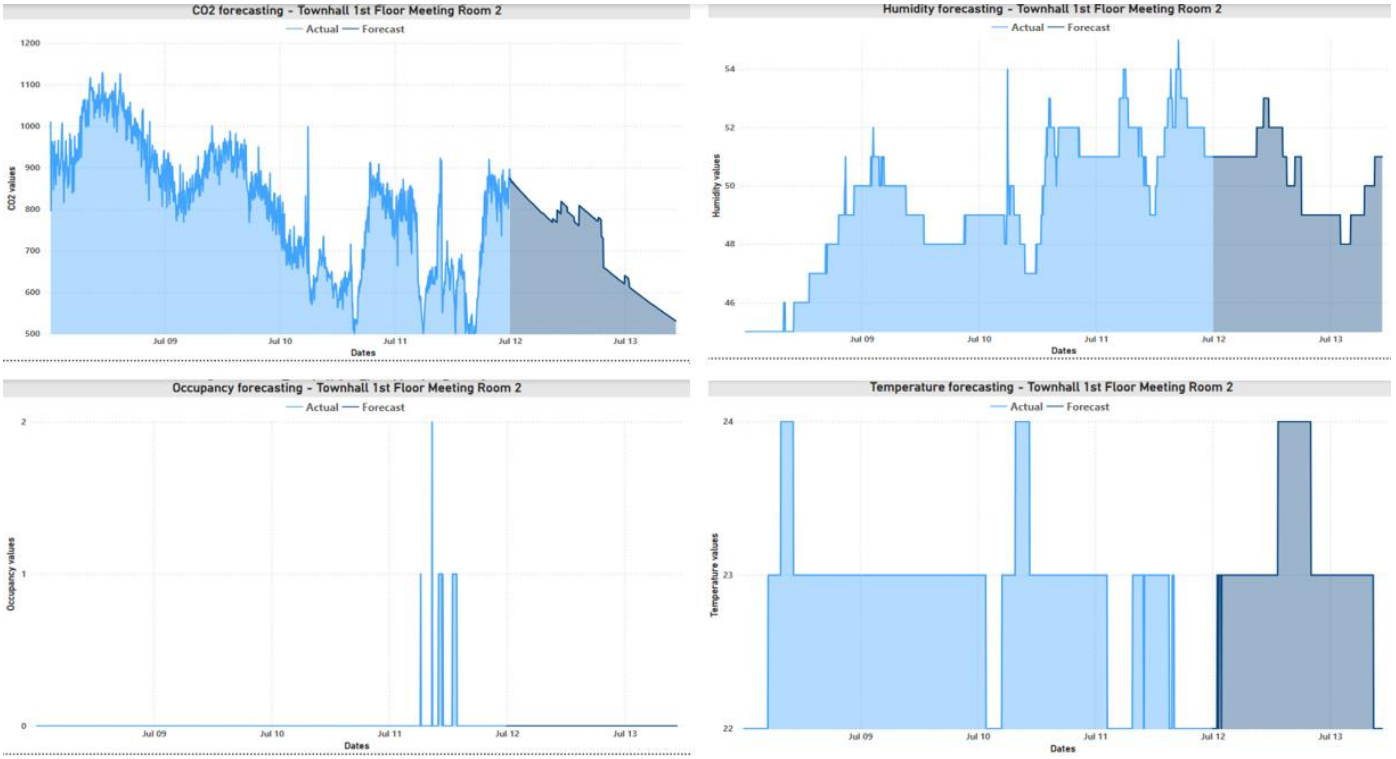


Figure 71 Forecasting graph for 1st Floor Meeting Room 2

Figure 70, Figure 71 concludes that the determined optimal models for XTownhall 1st Floor Meeting Room 2 is accurate and efficient in terms of predictions.

With the above created visualisation graphs mainly focussing on optimal forecasting models determined for each of the metric within each of the building's geometry (summarised in **section 10.2**), the efficient performance and the accuracy of predictions compared to the historical values and real observed values can be underlined.

10.4. Interactive forecasting report

This report gives the ability to easily choose the particular building which includes the entire structure as a whole entity and also the internal rooms of the building constituting the CO2, temperature, humidity, and occupancy metrics they want to forecast (Figure 72).

Figure 72 Building's structure selection dropdown

Also, the metric or parameter for which the forecasting graph needs to be viewed can be chosen (Figure 73), along with the choice of selecting actual values display and/or forecast values display (Figure 74).

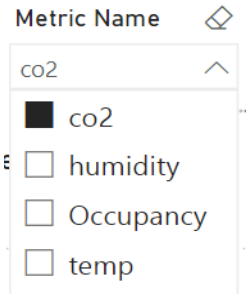


Figure 73 Metric selection dropdown

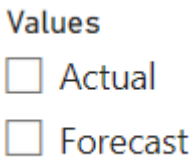


Figure 74 Actual and/or Forecast selection

The forecasting graphs displayed are of the optimal models determined for the metric for each building. Figure 75 gives an overview of the report created using Power BI mainly created to provide a holistic perspective of the building's sustainable and environmentally friendly practises.

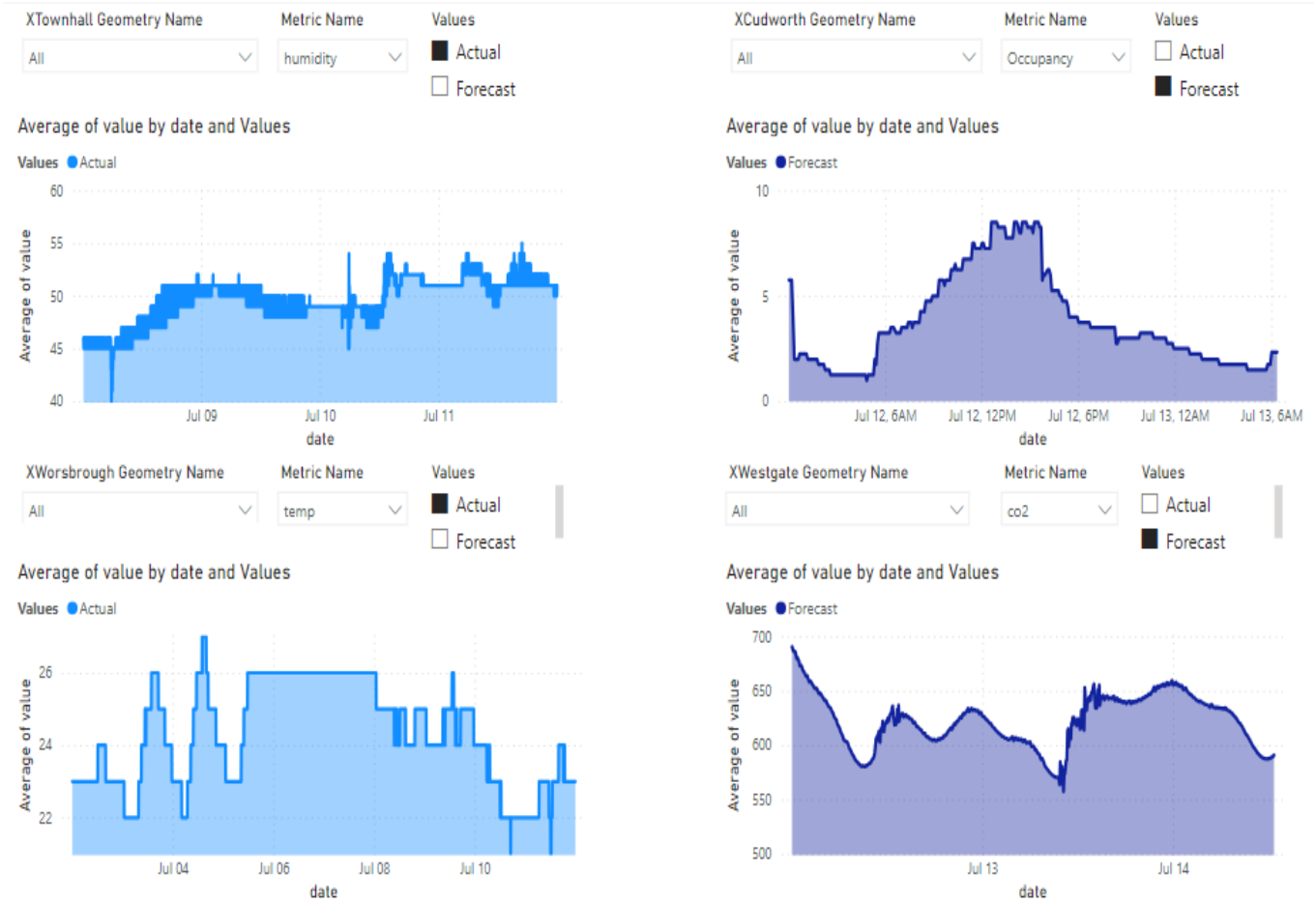


Figure 75 Interactive forecasting report for all the buildings

Conclusion

With the main goal of attaining insights for better sustainability for buildings, Figure 76 workflow was followed. It enhances the performance of each building and its geometry room with respect to CO2, Humidity, Occupancy and Temperature metrics.

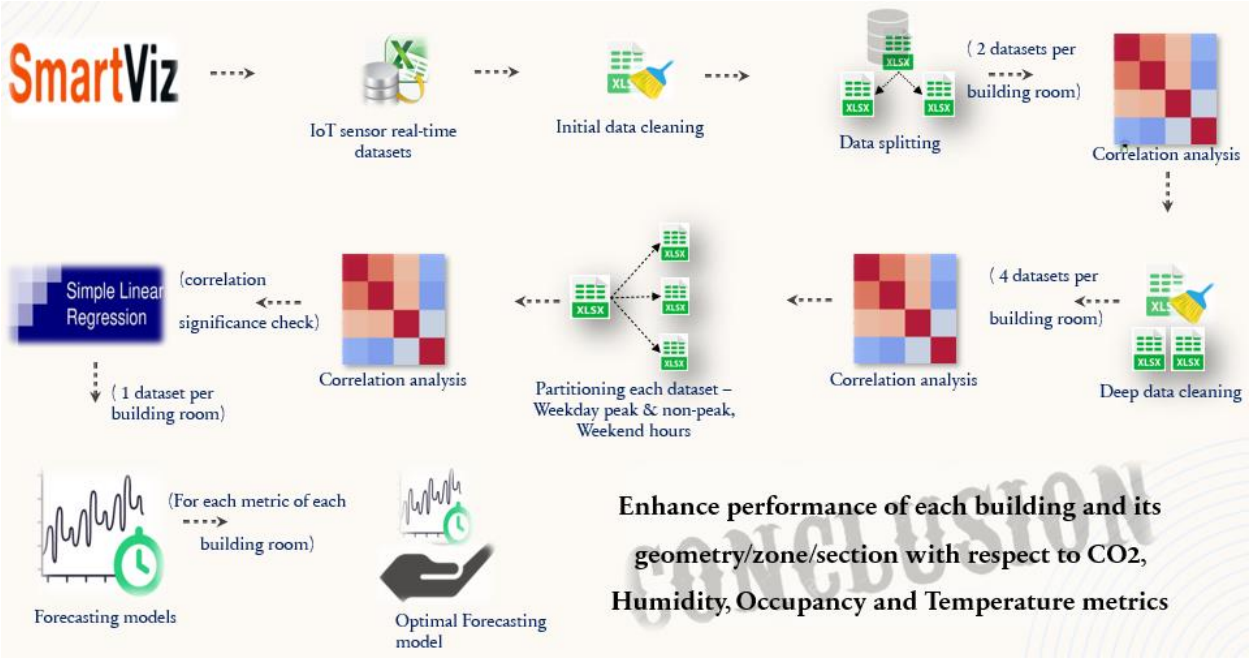


Figure 76 Project workflow

It empowers SmartViz with data driven insights to improve the building sustainability by successfully identifying meaningful correlations between metric CO2, Humidity, Occupancy and Temperature within each of the building and its geometry room. With this, more informed decisions can be made to utilise the building operations.

Furthermore, by determining the optimum forecasting model for each of the metric or parameter within any given building and providing accurate prediction or forecasting for CO2, Humidity, Occupancy and Temperature metrics independently, resources like electric power (lighting), heat, and cooling can be distributed more effectively and with reduction in waste across the buildings and its considered internal rooms. It can also provide a greener environment with lower carbon emissions and hence energy emissions. Additionally, proactive sustainable planning can be implemented to raise building productivity and comfort.

Overall, this thesis gives an upper hand over;

- More comprehensive sustainability insights

  Users of SmartViz may see in detail how temperature, occupancy, CO2, and humidity interact to affect the sustainability of a building thanks to the statistical analysis of correlation. This effort strengthens SmartViz's primary purpose of employing data analytics to optimise building efficiency by uncovering these important insights.

- Making decisions based on data

  Prioritising sustainability programmes and infrastructure upgrades is made possible by analysing the connections between the internal building's parameters involved. Instead of assuming, companies can allocate resources and expenses based on the relationships observed between the parameters for each buildings as shown by statistics in-order to maximise improvement in performance.

- Capabilities for effective prediction

  Having an exceptional ability to predict future values of important metrics like occupancy, temperature, CO2, and humidity, because of the time series forecasting models. Effectively relocating resources and optimising business operations to meet the targets for sustainability due to this predictive advantage.

- Building’s geometry rooms insights

  Providing customised insights at a precise room level of each building associated with the 4 metrics or parameters through the analysis of building's geometry name instead of analysing the building as a whole. By dive deeper in this analysis we got information on sustainability for particular places within the building and henceforth increasing the value of SmartViz's solutions.

- Optimal model determination for precise forecasting

  To guarantee that the time series forecasts take use of the most recent data science advancements, numerous sophisticated time series models are tested. Based on the performance metrics like MSE, MAE, RMSE, the best model is chosen for each metric for each building's geometry in-order to provide the most precise forecasts. With more assurance, proactive sustainability planning can be made, and resources can be managed when the accuracy is higher.

Future Work

- To collect more data for an in-depth view

  Including sensor data other than the basic temperature, CO2, humidity, and occupancy measurements gives a more complete view of energy consumption, sustainability issues, and efficiency improvements. Deeper connections and areas to focus on become clearly visible when data sources, such as air quality, lighting, weather patterns, water usage, and power consumption, are integrated.

- Make use of advanced and up-to-date forecasting methods

  Forecast accuracy for significant metrics can be increased by using increasingly complex predictive modelling approaches including neural networks, ensemble models, deep learning algorithms in addition to time series models like ARIMA, SARIMA and Prophet. This gives people a stronger belief to plan ahead and allocate resources to keep up sustainability objectives.

- Gaining Feedbacks

  Feedback from building occupants should be gathered to better understand their comfort and sustainability needs. Sustainability suggestions should take this information into account.

- Analysing a lifecycle

  To analyse the long-term sustainability of building materials and construction decisions, use lifecycle analysis data. Analyse sustainability from planning through demolition, taking into account the long-term effects on the environment.

- Resistance with climate changes

  Include planning for sustainability that addresses enhancing building’s resistance to climate change and extreme weather occurrences.

- Integrating Renewable Energy

  To incorporate renewable energy sources and how they effect a building's sustainability.

LinkedIn Posts

Post 1:

https://www.linkedin.com/posts/srijeetnair1811_buildingsustainability-dataanalytics-smartviz-activity-7108073719507341312-uNGc?utm_source=share&utm_medium=member_desktop

Post 2:

https://www.linkedin.com/posts/srijeetnair1811_dataanalysis-buildingsustainability-regressionanalysis-activity-7110403063735037952-sinV?utm_source=share&utm_medium=member_desktop

Post 3:

https://www.linkedin.com/posts/srijeetnair1811_sustainabilityinbuildings-environmentalfriendliness-activity-7111528269467787264-GaEY?utm_source=share&utm_medium=member_desktop

References

[1] Kumari, S. and Singh, S.K. (2022). Machine learning-based time series models for effective CO2 emission prediction in India. *Environmental Science and Pollution Research*. doi:https://doi.org/10.1007/s11356-022-21723-8.

[2] Fattah, J., Ezzine, L., Aman, Z., El Moussami, H. and Lachhab, A. (2018). Forecasting of demand using ARIMA model. *International Journal of Engineering Business Management*, 10, p.184797901880867. doi:https://doi.org/10.1177/1847979018808673.

[3] Liu, J., Yu, F. and Song, H. (2023). Application of SARIMA model in forecasting and analyzing inpatient cases of acute mountain sickness. *BMC Public Health*, 23(1). doi:https://doi.org/10.1186/s12889-023-14994-4.

[4] Albeladi, K., Zafar, B. and Mueen, A. (2023). Time Series Forecasting using LSTM and ARIMA. *International Journal of Advanced Computer Science and Applications*, 14(1). doi:https://doi.org/10.14569/ijacsa.2023.0140133.

[5] Altman, N. and Krzywinski, M. (2015). Simple linear regression. *Nature Methods*, 12(11), pp.999–1000. doi:https://doi.org/10.1038/nmeth.3627.

[6] Bangdiwala, S.I. (2018). Regression: simple linear. *International Journal of Injury Control and Safety Promotion*, 25(1), pp.113–115. doi:https://doi.org/10.1080/17457300.2018.1426702

[7] Chicco, D., Warrens, M.J. and Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, [online] 7(5), p.e623. doi:https://doi.org/10.7717/peerj-cs.623.

[8] Fairly Nerdy (2017). *What Is R Squared And Negative R Squared - Fairly Nerdy*. [online] Fairly Nerdy. Available at: http://www.fairlynerdy.com/what-is-r-squared/ [Accessed 30 Oct. 2019].

[9] Steurer, M., Hill, R.J. and Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, 38(2), pp.99–129. doi:https://doi.org/10.1080/09599916.2020.1858937.

[10] Plevris , V., Solorzano, G., Bakas , N. and Seghier, M.E.A.B. (2022). Investigation of performance metrics in regression analysis and machine learning-based prediction models. In: *ResearchGate*. 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2022).

[11] Wikipedia Contributors (2019). *Coefficient of determination*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Coefficient_of_determination.

[12] Liu, T., Chen, L., Yang, M., Sandanayake, M., Miao, P., Shi, Y. and Yap, P.-S. (2022). Sustainability Considerations of Green Buildings: A Detailed Overview on Current Advancements and Future Considerations. *Sustainability*, [online] 14(21), p.14393. doi:https://doi.org/10.3390/su142114393.

[13] Balaban, O. and Puppim de Oliveira, J.A. (2017). Sustainable buildings for healthier cities: assessing the co-benefits of green buildings in Japan. *Journal of Cleaner Production*, 163, pp.S68–S78. doi:https://doi.org/10.1016/j.jclepro.2016.01.086.

[14] Sandanayake, M., Zhang, G. and Setunge, S. (2019). Impediments affecting a comprehensive emission assessment at the construction stage of a building. *International Journal of Construction Management*, pp.1–11. doi:https://doi.org/10.1080/15623599.2019.1631977.

[15] Alwisy, A., BuHamdan, S. and Gül, M. (2019). Evidence-based ranking of green building design factors according to leading energy modelling tools. *Sustainable Cities and Society*, 47, p.101491. doi:https://doi.org/10.1016/j.scs.2019.101491.
[16] Li, F., Yan, T., Liu, J., Lai, Y., Uthes, S., Lu, Y. and Long, Y. (2014). Research on social and humanistic needs in planning and construction of green buildings. *Sustainable Cities and Society*, 12, pp.102–109. doi:https://doi.org/10.1016/j.scs.2014.03.003.
[17] Manso, M., Castro-Gomes, J., Paulo, B., Bentes, I. and Teixeira, C.A. (2018). Life cycle analysis of a new modular greening system. *Science of The Total Environment*, 627, pp.1146–1153. doi:https://doi.org/10.1016/j.scitotenv.2018.01.198.
[18] Vyas, G.S. and Jha, K.N. (2017). Benchmarking green building attributes to achieve cost effectiveness using a data envelopment analysis. *Sustainable Cities and Society*, 28, pp.127–134. doi:https://doi.org/10.1016/j.scs.2016.08.028.
[19] Wikipedia. (2021). *Correlation*. [online] Available at: https://en.wikipedia.org/wiki/Correlation.
[20] Python, R. (n.d.). *NumPy, SciPy, and Pandas: Correlation With Python – Real Python*. [online] realpython.com. Available at: https://realpython.com/numpy-scipy-pandas-correlation-python/.
[21] Unwin, A. (2020a). Why Is Data Visualization Important? What Is Important in Data Visualization? *Harvard Data Science Review*, 2(1). doi:https://doi.org/10.1162/99608f92.8ae4d525.

Appendix

SQL scripts

1. Appendix 1 Removing redundant data

```
/*add a new meaning column giving each data value a meaning*/

alter table [dbo].['1st floor office 2$']
add [meaning] nvarchar(50)
```

```
/*removing redundant data by removing values not in the specified benchmark*/

update [dbo].['1st floor office 2$']
set [meaning] =
    CASE
      WHEN metric_name = 'Occupancy' AND value  BETWEEN 36  AND 49 THEN 'FAIR'
      WHEN metric_name = 'Occupancy' AND value  BETWEEN 0 AND 36 THEN 'POOR'
      WHEN metric_name = 'Occupancy' AND value  BETWEEN 49 AND 64 THEN 'GOOD'
      WHEN metric_name = 'Occupancy' AND value  BETWEEN 64 AND 100 THEN 'EXCELLENT'

      WHEN metric_name = 'temp' AND value  BETWEEN 9.999  AND 18 THEN 'TOO COLD'
      WHEN metric_name = 'temp' AND value  BETWEEN 18 AND 22 THEN 'GOOD'
      WHEN metric_name = 'temp' AND value  BETWEEN 22 AND 29.999 THEN 'TOO HOT'

      WHEN metric_name = 'co2' AND value  BETWEEN 0.999 AND 400 THEN 'LOW/ERR'
      WHEN metric_name = 'co2' AND value  BETWEEN 400   AND 800 THEN 'GOOD'
      WHEN metric_name = 'co2' AND value  BETWEEN 800   AND 1000 THEN 'OK'
      WHEN metric_name = 'co2' AND value  BETWEEN 1000    AND 2000 THEN 'BAD/ERR'

      WHEN metric_name = 'humidity' AND value  BETWEEN 0  AND 30 THEN 'TOO DRY'
      WHEN metric_name = 'humidity' AND value  BETWEEN 30 AND 45 THEN 'OK'
      WHEN metric_name = 'humidity' AND value  BETWEEN 45 AND 55 THEN 'GOOD'
      WHEN metric_name = 'humidity' AND value  BETWEEN 55 AND 65 THEN 'OK'
      WHEN metric_name = 'humidity' AND value  BETWEEN 65 AND 100 THEN 'TOO HUMID'

      ELSE 'Range Not Found'
      END
```

2. Appendix 2 Use of Lead function

```
/*lead function to remove the occupancy repeatitions*/

DELETE FROM [dbo].['XCudworth 1st floor open off 1$']
WHERE [metric_name] = 'Occupancy' and [project_name] =  'XCudworth'
      AND [time] IN (
                        SELECT [time]
                        FROM
                        (
                            SELECT [time],
                                   Lead(metric_name) OVER (ORDER BY [time]) AS next_metric
                            FROM [dbo].['XCudworth 1st floor open off 1$'] where
                            geometry_name in ('1st Floor Open Office_iaq', '1st Floor Open Office')
                            and [project_name] =  'XCudworth'
                        ) AS nxt_metric
                        WHERE next_metric = 'Occupancy' and [project_name] =  'XCudworth'

                    )
```

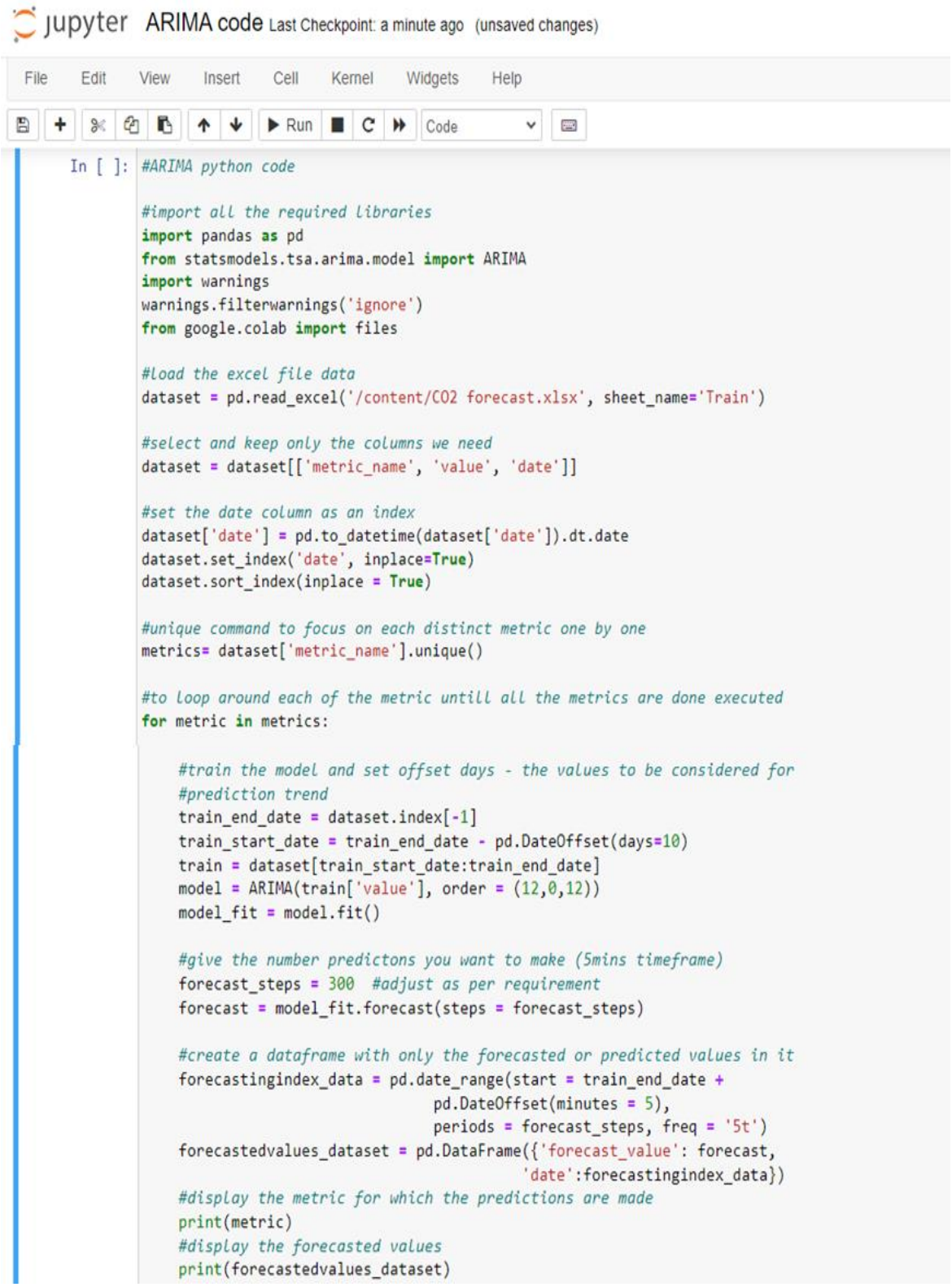
3. **Appendix 3 Use of Lag function**

```
/*lag function to remove the previous data not in an order to bring the
metric data in proper order*/

    WITH Lagfunction AS (
    SELECT
        metric_name,
        time,
        value,
        ROW_NUMBER() OVER (ORDER BY time) AS rn,
        LAG(metric_name, 1) OVER (ORDER BY time) AS next_metric,
        LAG(metric_name, 2) OVER (ORDER BY time) AS next_next_metric,
        LAG(metric_name, 3) OVER (ORDER BY time) AS next_next_next_metric
        ,geometry_name
        ,project_name
    FROM [dbo].['1st Floor Meeting Room 3$']
)
DELETE FROM Lagfunction
WHERE NOT EXISTS (
    SELECT 1
    FROM Lagfunction lf
    WHERE Lagfunction.rn = lf.rn
    AND (
        Lagfunction.metric_name = 'Occupancy'
        OR lf.next_metric = 'Occupancy'
        OR lf.next_next_metric = 'Occupancy'
        OR lf.next_next_next_metric = 'Occupancy'
    )
```

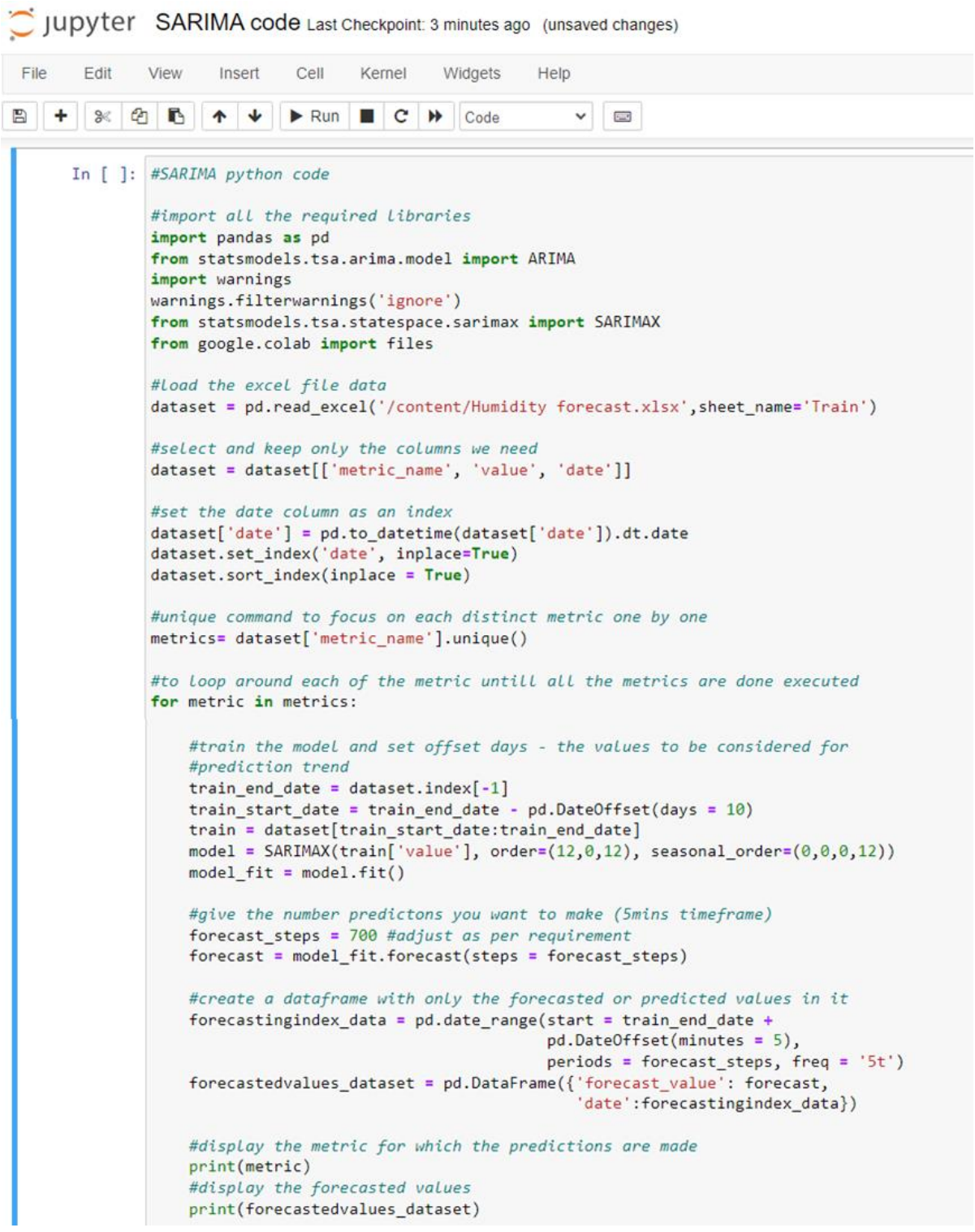
Python scripts

I used google colab for python scripts since the library installation process is way easier in colab than in jupyter but for the attachment purpose I used the same code in jupyter notebook for better understanding from the attachments to avoid black background that is seen in google colab codes.

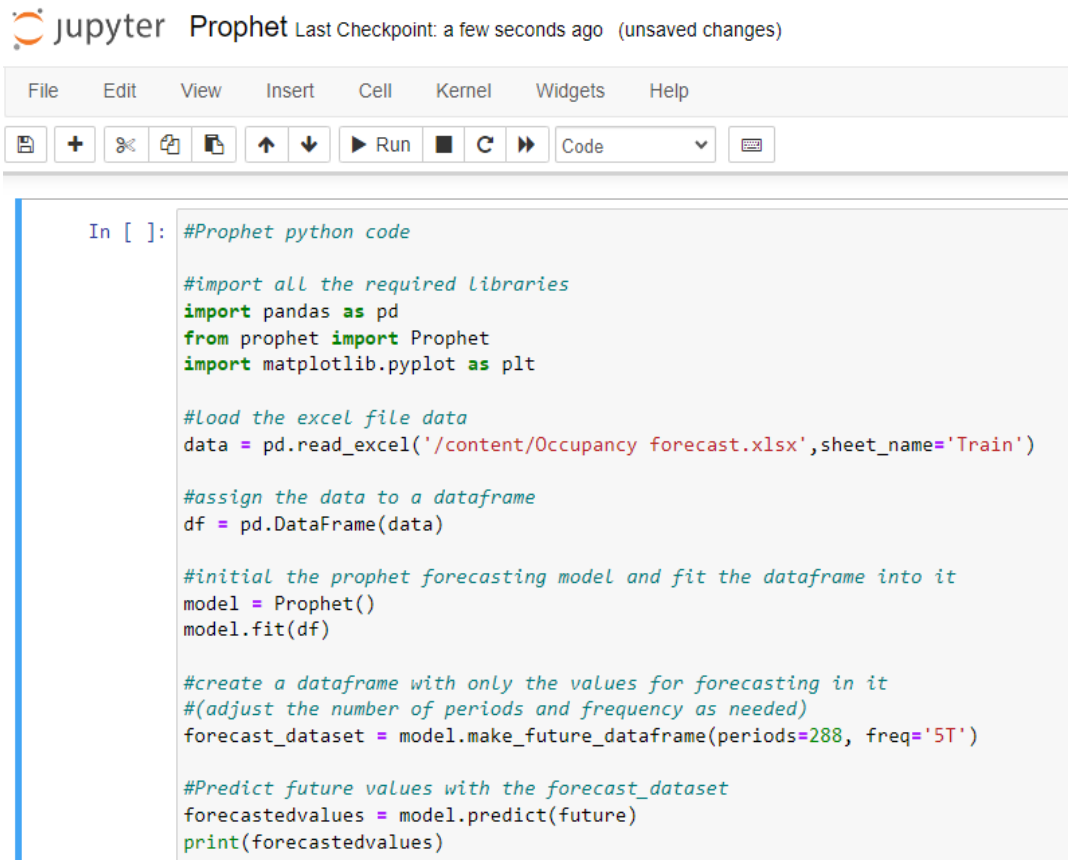
4. Appendix 4 ARIMA (Google colab and Jupyter notebook was used)



5. Appendix 5 SARIMA (Google colab and Jupyter notebook was used)



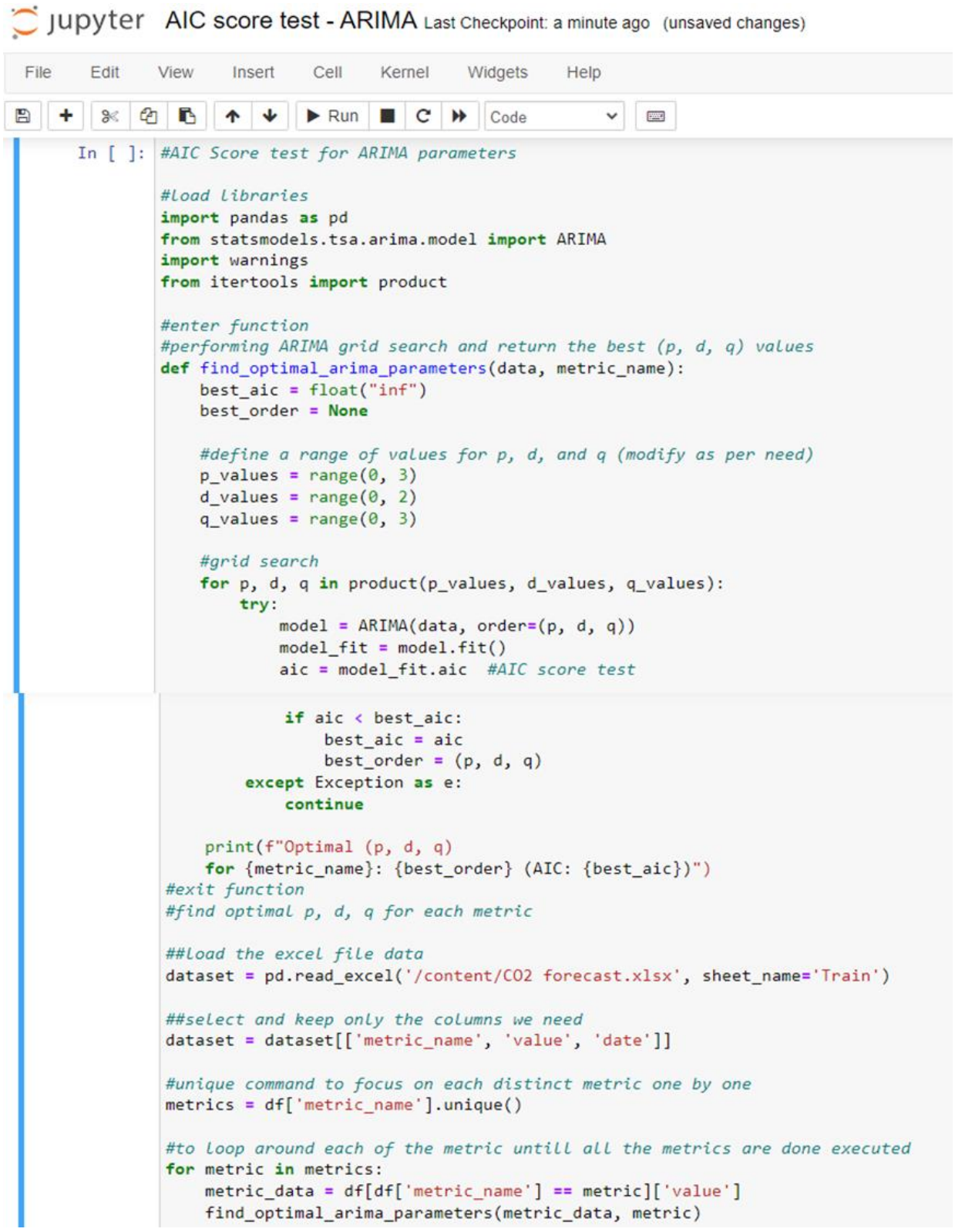
6. Appendix 6 Prophet (Google colab and Jupyter notebook was used)

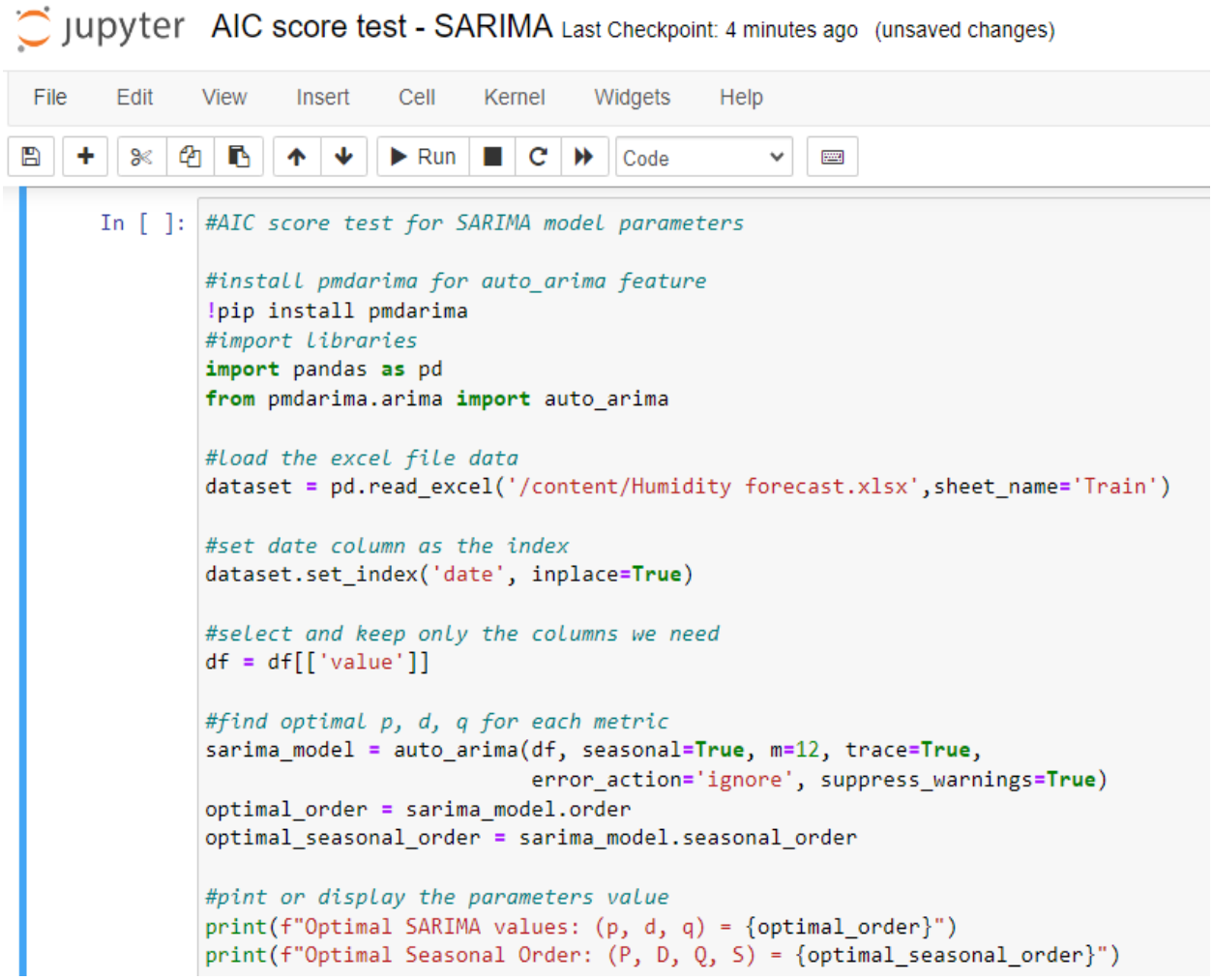



```python
#Prophet python code

#import all the required libraries
import pandas as pd
from prophet import Prophet
import matplotlib.pyplot as plt

#load the excel file data
data = pd.read_excel('/content/Occupancy forecast.xlsx',sheet_name='Train')

#assign the data to a dataframe
df = pd.DataFrame(data)

#initial the prophet forecasting model and fit the dataframe into it
model = Prophet()
model.fit(df)

#create a dataframe with only the values for forecasting in it
#(adjust the number of periods and frequency as needed)
forecast_dataset = model.make_future_dataframe(periods=288, freq='5T')

#Predict future values with the forecast_dataset
forecastedvalues = model.predict(future)
print(forecastedvalues)
```

7. Appendix 7 LSTM (Google colab and Jupyter notebook was used)

8. **Appendix 8 ARIMA and SARIMA AIC score test (Google colab and Jupyter notebook was used)**




jupyter AIC score test - SARIMA Last Checkpoint: 4 minutes ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
In [ ]:
#AIC score test for SARIMA model parameters
#install pmdarima for auto_arima feature
!pip install pmdarima
#import libraries
import pandas as pd
from pmdarima.arima import auto_arima
#load the excel file data
dataset = pd.read_excel('/content/Humidity forecast.xlsx',sheet_name='Train')
#set date column as the index
dataset.set_index('date', inplace=True)
#select and keep only the columns we need
df = df[['value']]
#find optimal p, d, q for each metric
sarima_model = auto_arima(df, seasonal=True, m=12, trace=True,
error_action='ignore', suppress_warnings=True)
optimal_order = sarima_model.order
optimal_seasonal_order = sarima_model.seasonal_order
#pint or display the parameters value
print(f"Optimal SARIMA values: (p, d, q) = {optimal_order}")
print(f"Optimal Seasonal Order: (P, D, Q, S) = {optimal_seasonal_order}")