TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Civil Engineering Institute of Construction Informatics

Processing of Sensor Data to improve Building Performance

by

Sujan Hossen

from

Rajbari, Bangladesh

A Project submitted to the Faculty of Civil Engineering, Institute of Construction Informatics of the University of Technology Dresden in partial fulfilment of the requirements for the degree of

Master of Science

Responsible Professors

Prof. Dr.-Ing. habil. Karsten Menzel

Prof. Dr.-Ing. John Grunewald

Advisor

Janakiram Karlapudi, M.Sc.

Dresden, November and 2022

TECHNISCHE UNIVERSITÄT DRESDEN

Fakultät Bauingenieurwesen Institut für Bauinformatik

Master Project Task Sheet

(Aufgabenstellung Projekt Masterprogramm ACCESS)

Name:	Hossen, Sujan	Matrikel Nr.: 4898436
Degree Program:	ACCESS	

Topic:	**Processing of Sensor Data to improve Building Performance**
	(Verarbeitung von Sensordaten zur Verbesserung der Gebäudeleistung)

Objective:

Modern or recently renovated buildings are increasingly equipped with a large number of sensors and actuators for monitoring, control and operation. The resulting data must be preprocessed and filtered for further design activities. The consistency and meaningfulness of the data must also be checked before conclusions can be drawn from it.

In the "Nürnberger Ei" office building a comprehensive home automation system has been installed. This system uses wireless data transmission techniques, which could be affected by interference with radio reception or failures in the power supply to individual devices. Checking the consistency of the received sensor data is therefore an important step in processing the collected data. In order to avoid relying solely on information technology solutions for transmission security, data from multiple sensors should be used for consistency checks.

The refined recorded data is not only used to verify the efficiency of the designed systems but also to adjust and update the operational efficiency of the systems. In the scenario of building simulations, this recorded data can be used as a input data for simulations and estimations of real consumptions originating from building operations.

Scope of Work:

The project work should address the following tasks:

1. State-of-the-art analysis on possible inconsistences of recorded data;
2. Verify the quality of sensor data recorded from multiple sensors installed in the living lab "Nürnberger Ei";
3. Identify the reasons for inconsistencies of recorded data, if any observed.
4. Propose an approach to improve the consistency of recorded data and to refine the data for building performance simulations.
5. Demonstrate the adopted approach with example data available from institute building.

Notes:

- The results of the project work, created source code as well as used data sources are also to be submitted.
- The documentation is to be submitted both as a PDF and as an original Word file. Graphics etc. are to be submitted in native format.

Responsible Parties and important Dates:

Responsible academic staff: Prof. Dr.-Ing. habil. Karsten Menzel

Second Examiner: Prof. Dr.-Ing. John Grunewald

Supervisor: Janakiram Karlapudi, M.Sc.

Topic handed over to student: 08/08/2022

Expected submission date: 29/11/2022

SUJAN
(Student Signature)

Declaration of originality

I confirm that this assignment is my own work and that I have not sought or used the inadmissible help of third parties to produce this work. I have fully referenced and used inverted commas for all text directly quoted from a source. Any indirect quotations have been duly marked as such.

This work has not yet been submitted to another examination institution – neither in Germany nor outside Germany – neither in the same nor in a similar way and has not yet been published.

Dresden, _____

Place, Date

_____

(Signature)

Acknowledgement

Firstly, I want to express my gratitude to Prof. Dr.-Ing. habil. Karsten Menzel and Prof. Dr.-Ing. John Grunewald for providing me the opportunity to work in this contemporary project topic which has broaden my thinking in smart building and future building technology including indoor environment and energy use.

Furthermore, I would love to thanks my advisor Mr. Janakiram Karlapudi, M.Sc. for the supervision, proper guideline and mental support throughout the period of my project work. Without whom I would not have made it through my master's project. The weekly meeting and the meeting with a very short notice with him really helped me immensely. In addition to that, I really thankful to those who attended in the intermediate meeting and made comments and gave me the suggestions to improve the quality of the work.

Finally, I must express my gratitude and love to my parents and siblings for providing me support and continuous encouragement throughout my years of study. This accomplishment would not have been possible without them. Their all form of supports helped me to reach this stage and hope they will be with me in my future achievement.

Abstract

A proper management of building can ensure the healthy indoor environment for the inhabitants which has a strong effect on productivity. With the time, addition of several technology (wireless sensor, IoT) has increased to improve the overall performance of building including optimization of energy consumption. Building sector (residential, industrial, commercial) is one of the biggest contributors of total energy consumption also producer of greenhouse gas therefore the reduction of energy use in this sector can have a large impact on global environment. Application of sensor technology in the building is convenient to observe the actual scenario in indoor environment including the major comfort parameters like temperature, humidity, CO2 concentration and brightness also the energy consumption. The study aims to analyze the sensor data of a building to find the inconsistency and discuss about the possible solution of it. The paper starts with an overview of human comfort including major comfort parameter and the global energy consumption record with some specific sector in Germany. Then it discusses about different sensor and state of art analysis of different inconsistency which can occur in sensor dataset. In previous study, very few discussed about the solution for the specific inconsistency and also the application of the method in actual dataset. Therefore, the paper introduces several sensor data from a Living Lab and implement the method to verify the quality of data (finding the inconsistency). In addition to that, a thorough analysis on the sensor data by comparing with outdoor environment also by comparing one sensor reading to another to check the data quality. Following this the noticeable inconsistencies are categorized and several methods applied to improve the consistency of the dataset. Lastly it describes the result and give the direction on future work which can be applied to get the more accurate solution in automated way. In this study, several abnormal situations in the sensor data are observed (outliers, unknowns, missing values) therefore it required the solution to update the corrected data to the sensor system.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

BIM	Building Information Modelling
BEM	Building Energy Modelling
HVAC	Heating Ventilation and Air Conditioning
IEQ	Indoor Environment Quality
WSN	Wireless Sensor Network
IoT	Internet of Things
IEC	International Electrotechnical Commission
LS	Lightning Start
REHVA	Representatives of European Heating and Ventilation Associations
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
ADF Test	Augmented Dickey Fuller test
rH	Relative Humidity
ISO	International Organization for Standardization
RNN	Recurrent Neural Network
ANN	Artificial Neural Network
PPM	Parts Per Million

1 Introduction

Modern building infrastructure are mostly equipped with WSN and IoT sensor for monitoring Temperature, Humidity, CO_2, Illuminance, Occupancy etc. to ensure comfort with optimize energy consumption. Here maintenance of these sensor is one of the challenging tasks to encounter. These sensors show a various type of inconsistency due to harsh environment, malfunction, rapid attrition, malicious attack, tempering, power supply, poor built in sensor etc. which produce some abnormal data (outliers, constant reading, bias, missing data etc.). As every sensor system produce a ton of data, it is difficult to categories the data and also pre-possessing to use these data in inconsistency detection method. A thorough investigation is required for sensor produced data and finding inconsistency. After that adapting the corrected data to the system can improve the overall building performance.

1.1 Motivation

1.1.1 Human Comfort

Indoor human comfort, which is often measured from four dimensions including thermal, visual, and respiratory comfort, has a direct relationship to the quality of the indoor environment. Indoor environmental quality and building energy efficiency are significantly impacted by the proper management of environmental building factors like temperature, humidity, light, etc. To connect the built environment with lighting and heating, ventilation, and air conditioning (HVAC) systems, such control often relies on a range of sensors. Additionally, the quality of the indoor environment has a significant impact on the occupant's productivity and health[1] .The indoor environment must be kept in the ideal comfort zone. For instance, if the indoor atmosphere is too cold, the inhabitant experiences discomfort and sleepiness, and the temperature is the cause of several health complications, which affects the occupant's working mind and eventually reduces productivity [2]. Therefore, it's essential to keep the buildings thermal environments in good range.

Thermal Comfort is used to describe “a condition of mind that expresses satisfaction with the thermal environment in which it is located” according to ISO Standard 7730 (1994) and ASHARE Standard 55[3]. ASHRAE develops standards and recommendations for thermal comfort. Building occupants place a higher value on thermal comfort than on visual, auditory, or respiratory comfort when comparing various aspects of comfort. It is said to have a bigger impact on occupants total IEQ satisfaction[4]. The fact that thermal comfort serves as the primary operating factor for HVAC (heating, ventilation, and air conditioning) systems in buildings is another factor that makes thermal comfort particularly significant.

Visual Comfort: The term "state of mind that reflects happiness with the visual environment" is referred to as visual comfort[5]. A good visual comfort guarantees that people have enough light for their activities or occupations without subjecting their eyes to lighter than their eyes can tolerate. Human visual discomfort will result from either much or insufficient lighting.

Respiratory comfort:

Respiratory Comfort is closely related to indoor air quality (IAQ) which depends on three factors including the quantity of pollutants, ventilation rate within the building and therefore the duration of the pollutants being trapped within the space [6]. Indoor air quality considers the subsequent parameters: temperature, humidity, carbon dioxide, PM2.5, ozone, formaldehyde, volatile organic chemicals, carbon monoxide gas etc.[7].

1.1.2 Optimization of energy

The rapidly growing world energy use has already raised concerns over supply difficulties, exhaustion of energy resources and heavy environmental impacts (ozone layer depletion, global warming, climate change, etc.)[8]. Buildings now account for a larger portion of the world's energy consumption than ever before, accounting for 20% to 40% of it in wealthy countries[8]. The growing trend in energy demand will continue in the future due to population growth, rising demands for comfort and building services, and an increase in the amount of time spent within buildings. Because of this, energy policy at the regional, national, and international levels now places a high priority on achieving energy efficiency in buildings. "In response to growing energy use, the depletion of energy resources, and major environmental effects has increased the concern throughout the world. Due to this, most nations have participated into international agreements for the benefit of society, such as the Paris Agreement in 2015. These sectors are among the most important to address because they account for 24% of global CO2 emissions and have an overall energy consumption of 41%"[9]. Most action plans are centered on improving Energy Efficiency (EE) through the promotion of renewable energies and the development of systems to reduce energy consumption. The following figure shows the worldwide energy consumption trend for last 50 years, the increasing trend is alarming for human being along with other creatures. Also, the figure depicts; industry, residential buildings, commercial and public service are the biggest contributor of total energy consumptions.

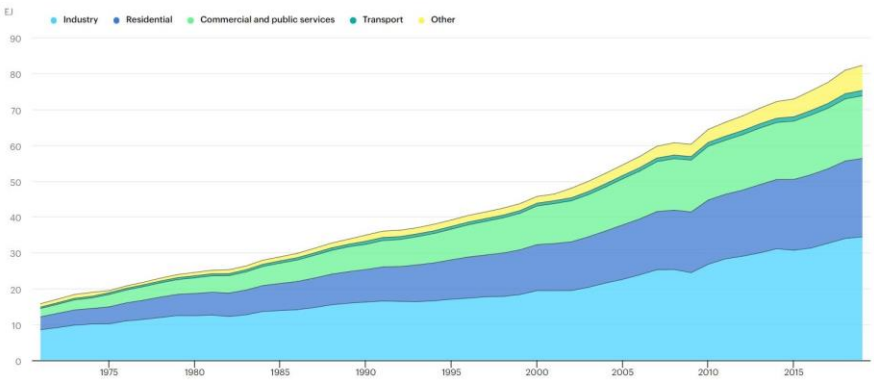


Figure 1: Worldwide Energy Consumption [10]

The energy consumption in Germany in the field of households, industry has a large contribution. According to "Final energy consumption by sector and energy source 2020" (Figure-2) along with the transportation sector, the industrial sector consuming 28.5% and household 28.6% of total energy. Though the generation and use of renewable energy has

increased in last three decades in Germany, still there are so many countries depends on the traditional oil, gas, coal-based energy system. The built environment consumes a significant amount of nonrenewable energy, that is what causes the major global warming gases like carbon dioxide, sulfur dioxide, and nitrogen dioxide. As a result, it's vital to cut back on energy use in the built environment.

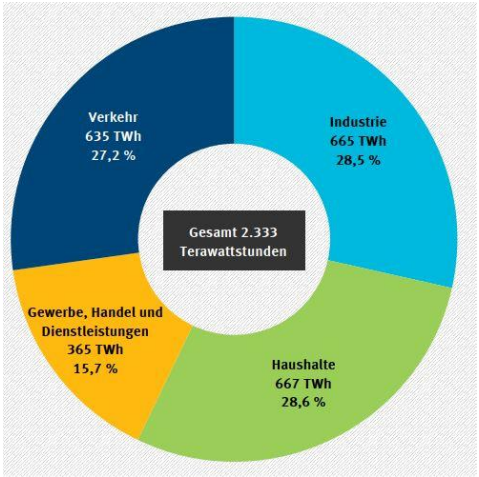


Figure 2: Energy consumption by sector in Germany [11]

1.1.3 Objectives

To ensure the above-mentioned human comfort and energy optimization it is necessary to analyze the generated data by the various sensor and checking the quality of the data. The major aims of this study include:

-Collection of data and preliminary cleaning of full dataset

-Initial decision on the data quality based on this further investigation can be made.

-Room wise data distribution and observe the inconsistency

-If any inconsistency observed, then finding the reason behind this inconsistency and find the possible solution

-Finally update the corrected data in the system

2 State of the art analysis on inconsistency

2.1 Sensor and it's limitation

In reaction to shifting physical conditions, sensors change their electrical properties. As a result, most artificial sensors collect, process, and send environmental data using electronic devices. Since these electronic systems function on the same principles as electrical circuits, the ability to control the flow of electrical energy is crucial. A sensor converts inputs like heat, light, sound, and motion into electrical signals. Before being sent to a computer for processing, these signals are routed through a device that converts them into a binary code. To control the flow of electric charges via the circuit, several sensors operate as switches. As they alter the state of the circuit, switches are an essential element of electronics. For example, a transistor works by using a small electrical current in one part of the circuit to switch on a large electrical current in another part of the circuit. According to [12] sensors show following challenges:

-In comparison to nodes in an ad hoc network, the number of sensor nodes in a sensor network might be several orders of magnitude higher.

-Dense deployment of sensor nodes

-Sensor nodes are susceptible to breakdowns.

-A sensor network's topology is subject to frequent alterations.

-Most ad hoc networks are built on point-to-point communications, whereas sensor nodes mostly use a broadcast communication paradigm.

-The processing, memory, and power of sensor nodes are constrained.

-Due to the high overhead and vast number of sensors, sensor nodes might not have global identification (ID).

2.2 Types of sensors considered in the study

2.2.1 Temperature sensor

An electronic device that measures the temperature of its environment and converts the input data into electronic data to record, monitor, or signal temperature changes [13]. There are different types of temperature sensors. Among these different temperature sensors, two major categories are contact and non-contact temperature sensor that measures in degree Celsius. Contact temperature sensors require direct contact with the object being monitored. Non-contact temperature sensors measure the temperature of an object indirectly. Non-contact temperature sensors usually use infrared radiation to detect the heat emitted by objects. This signal is then sent to a calibrated electronic circuit, which determines the object's temperature. According to REHVA[14], [15] the comfortable temperature range between (15°C-28 °C). A more specific range can be determined from the standard but depends on relative humidity, season, clothing worn, activity levels, and other factors. The

following figure shows the daily average outdoor temperature in Dresden. The maximum daily average temperature is around 25°C in the summer season and the minimum daily average temperature is about -3°C in the winter time.

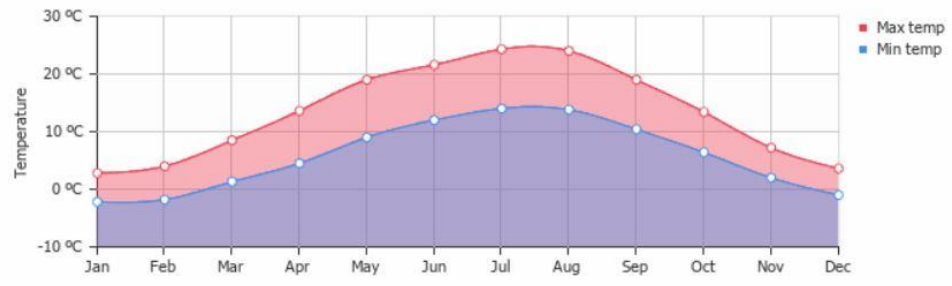


Figure 3: Daily average Temperature in Dresden [16]

2.2.2 Humidity Sensor

The use of humidity sensors in industrial and household applications has been progressing rapidly, with any fabrication technique being able to produce them. Among all the various humidity evaluation terms and units, absolute humidity and relative humidity are the most commonly used. Based on the units of measurement, humidity sensors are subsumed in two main classes: Relative Humidity (RH) and Absolute Humidity sensors [17]. In according to REHVA (Representatives of European Heating and Ventilation Associations) [18], in Germany the relative humidity level should not cross 12 g/kg to ensure comfort in indoor environment. The figure-4, explains, the humidity ranges with the different comfort level. In general, the comfortable range for humidity is 30-60% for indoor environment according to ASHRAE. It also mentioned the recommended humidity range which is 45-60% and high ranges between 55-80%.

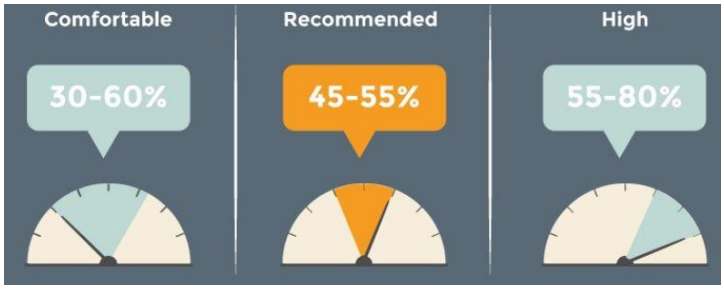


Figure 4: ASHRAE suggested Humidity Level [19]

2.2.3 CO2 Sensor

Modern buildings require carbon dioxide sensors to monitor air quality, in order to ensure the welfare of the occupants. The sensors adjust ventilation rates to match the occupants needs [20]. As CO2 levels increase, people will feel increasingly uncomfortable and unable to do as much as they used to at lower levels of activity. Many people will experience nausea, headaches, and poor sleep. Therefore, the CO2 concentration can be used as an indicator of indoor air quality. If there are people or animals in the building, ventilated air is needed to

limit the concentration of carbon dioxide and pollutants in the air (dust, smoke, volatile organic compounds, etc.) as well as to dilute odors and remove water vapor[21]. In according to the "German Committee on Indoor Air Guide Values" the indoor CO2 level must be below 2000 ppm. When the value of CO2 stays below 1000 ppm no action is necessary. If the value stays in between 1000-2000 ppm the ventilation system should be improved as it considered as hygienically noticeable. Finally, if the indoor CO2 goes above 2000 ppm that is hygienically unacceptable therefore necessary steps must be taken to improve indoor environment.

CO_2-concentration (ppm)	Hygienic assessment	Recommendation
< 1000	hygienically safe	no action
1000–2000	hygienically noticeable	Ventilation (outdoor air flow rates or rather increasing air change) proof of ventilation habits and improvement
> 2000	hygienically unacceptable	Proof for options of ventilation, proof for further measures

Source: German Environment Agency

Figure 5: CO2 level [22]

2.2.4 Light sensor/illuminance sensor

A Light Sensor generates an output signal indicating the intensity of light by measuring the radiant energy that exists in a very narrow range of frequencies basically called light, and which ranges in frequency from infra-red to visible up to ultraviolet light spectrum[23]. The light sensor is a passive device that transforms light energy from the visible and infrared portions of the spectrum into an output signal, which is an electrical signal and sensors are more usually referred to as photoelectric devices or photosensors since they convert light energy into electricity[24]. The photoelectric devices that generate electricity when lighted, such as photovoltaic or photo emissive devices, and those that change their electrical properties in some way, such as photo-resistors or photoconductors, are the two main groups of photoelectric devices.

The following table elaborates the different activities and the required level of light for that specific activity respectably. From the table, for the office work the lux value should be 250-500 lux. As the study is related to the office building therefore the lux value for the rooms should be around 250-500 lux for the occupant of the building for good visual comfort.

Activity	Illuminance (lx, lumen/m^2)
Public areas with dark surroundings	20 - 50
Simple orientation for short visits	50 - 100
Areas with traffic and corridors - stairways, escalators and travelators - lifts - storage spaces	100
Working areas where visual tasks are only occasionally performed	100 - 150
Warehouses, homes, theaters, archives, loading bays	150
Coffee break room, technical facilities, ball-mill areas, pulp plants, waiting rooms,	200
Easy office work	250
Class rooms	300
Normal office work, PC work, study library, groceries, show rooms, laboratories, check-out areas, kitchens, auditoriums	500
Supermarkets, mechanical workshops, office landscapes	750
Normal drawing work, detailed mechanical workshops, operation theaters	1000
Detailed drawing work, very detailed mechanical works, electronic workshops, testing and adjustments	1500 - 2000
Performance of visual tasks of low contrast and very small size for prolonged periods of time	2000 - 5000
Performance of very prolonged and exacting visual tasks	5000 - 10000
Performance of very special visual tasks of extremely low contrast and small size	10000 - 20000

Table 1: Lux level for different activity [25]

2.3 Most Common type of Inconsistency in sensor data

2.3.1 Outliers/anomalies

The term outlier, also known as anomaly, originally stems from the field of statistics. The two classical definitions of outliers are: Hawkins: "an outlier is an observation, which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"[26]. Barnett and Lewis: "an outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data"[26][27]. Outliers can be very revealing about the subject and the data collection process. It is essential to understand how outliers occur and whether they can occur again as a normal part of the process or study area. outliers can be classified as point outliers and collective outliers based on the number of data instances involved in the concept of outliers.

Point anomalies: when an individual data point is different from the rest of the data[28]. "A single outlying instance in a given group of data instances is referred to as a point outlier. The majority of existing outlier detection techniques concentrate on this sort of outliers because it is the easiest. When a data point exhibits outlier-ness on its own, rather than in connection with other data points, it is identified as an outlier"[29]. The following figure describes the point outlier. The sample dataset ranges between 0.6 to 1.4 but a single datapoint showing a reading 2.0 so, it considered as point outlier.

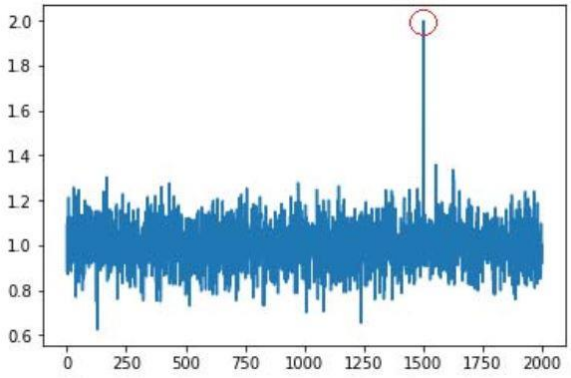


Figure 6: Point anomaly

Collective anomalies: when a group of related data points is anomalous compared to the dataset; the individual data points could represent normality, while it is their actual sequence that represents an anomaly[28]. The figure-7 elaborates the dataset has two major distribution N1 and N2. Here some datapoints named as O3 has variation to other major datapoints, therefore it considered as collective anomalies by definition.

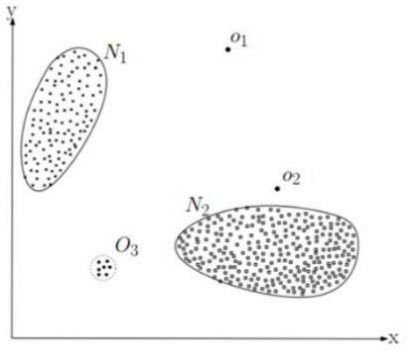


Figure 7: Collective anomalies [30]

A large number of tiny, inexpensive sensor nodes with sensing, processing, and short-range wireless communication capabilities make up wireless sensors [31]. Due to a variety of factors explained in [32], including the following, wireless sensor networks are consequently prone to outliers: (1) WSNs report the monitored data from the real world using imperfect sensing devices; (2) such devices are battery-powered, so their performance tends to decline as power is depleted; (3) since these networks may include a large number of sensors, this number may reach an extremely high value that can reach to million nodes depending on the application, therefore the chance of error is greater than that in traditional networks. As a result of their unique requirements, dynamic nature, and resource constraints, classic outlier detection approaches cannot be used in wireless sensor networks [33]. Several methods applied for detection and correction of outliers in the following studies.

Distance based method [34]. The feature points are utilized in this manner to represent the series. After that, the unequal split of series is realized using the second order regression model. The aberrant subsequence scores are computed based on the dynamic time warping distance. Then, decide if the anomalous score is an outlier by choosing the biggest k numbers. N. Chugh, M. Chugh, and A. Agarwal[35] stated that any form of data for which a

similarity or distance measure is available can be used to design a distance-based anomaly detection technique, and that this technique does not necessitate a thorough grasp of the application area. Because they make no assumptions about data distribution, distance-based anomaly detection techniques can be used to data streams. According to Ji Zhang [36], density-based anomaly detection is more accurate than distance-based anomaly detection, but it is also more expensive and complex because it takes into account the density of both neighbors and points. Density-based anomaly approaches are ineffective for high dimensional datasets because as dimensionality rises, estimation accuracy of density declines[37].Clustering based method[38].The data set is initially divided into various clusters using the approach, and the data points that do not fit into any cluster are considered outliers. Clustering technology is employed in the field of anomaly detection for both unsupervised and semi-supervised detection. However, the clustering process typically produces anomaly detection as a byproduct.

2.3.2 Sensor Drift

Sensor drift defied as "A slight temporal fluctuation of the sensor response subjected to the same analyte under equal conditions"[39]. Causes of sensor drift include environmental variables (variations in temperature, humidity, and ambient pressure), sensor surface poisoning, and sensors aging due to thermochemical fatigue following repeated gas exposures[40]. Sensor drift is a common problem that can lead to inaccurate data measurement readings. Figure-8 shows the difference between original data and the measured data by the sensor. Due to drift the temperature reading goes from 5 to 40 or above though the original reading never goes above 10. It can be caused by several factors including environmental contamination, vibration or extreme temperature fluctuations. Because this drift causes the measurement error to get worse over time, it's not possible to calibrate out the error. Traditionally, reducing inaccurate measurements caused by sensor drift has meant undertaking a regular and time-consuming preventative maintenance calibration program.

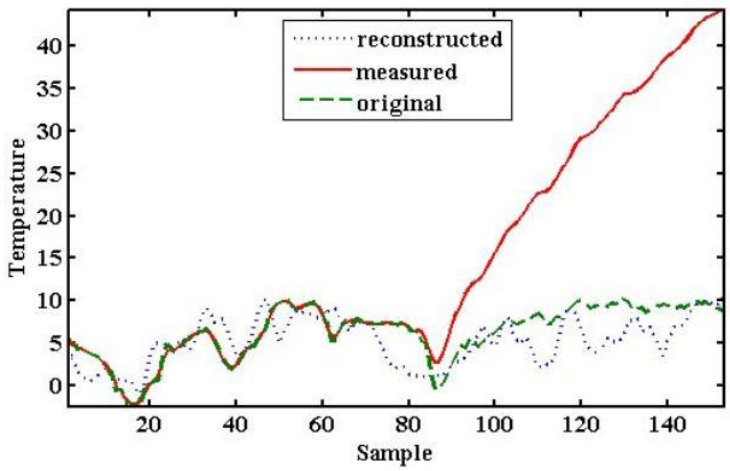


Figure 8: Sensor data drift [41]

There are many factors that can cause data to drift one key factor is the time dimension. If the sensor is placed in harsh environment also it's unattended for a long time, then sensor data shows drift. Several other factors can also cause drift like errors in data collection,

seasonality etc. The following studies have discussed about the sensor drift the reason behind the drift also the solution for the drift detection and correction.

G. von Arx, M. Dobbertin, and M. Rebetez [42] propose a method to remove sensor drift in high-frequency data series. In the study Visual Basic for Application (VBA) used to initial observation then they run a test for the homogeneity test and finally LR to compare with another dataset. S. Munirathinam suggested[43], The central sensor database, known as SCADA (Supervisory Control AND Data Acquisition), stores and regulates all sensor operations. The system functions and communicates within the operating environment. This method proposes three drift detection approaches that can identify drift, send out early warnings of drift, and assist manufacturing in taking proactive action to address the problems. For sensor networks used in general-purpose monitoring,[44] suggest a blind online drift calibration framework based on subspace projection and recovery. For data stream learning in sensor networks, a brand-new idea drift detection method is suggested. When there are fewer than 20% of sensors that have drifted, the suggested method can detect and recover the sensor drift; when there are 40% of sensors that have drifted, the recovery rate decreased at 80%.

2.3.3 Constant Value

Constant data is ovserable in various sensor which may the reason for sensor self prediction or sensor faults. And for this reason sensor produce some abonormal data in a certain period of time, it may continue from single data to thousands or more. The dataset recorded by the sensor gave a reading about 25 for every instance (figure-9). This type of inconsistency in data is absence in the previous study. It is found in a sensor dataset which published in Zenodo [45] by the heading “BIM4EEB ITALIAN BUILDING SENSORS MEASUREMENTS DATASET”.

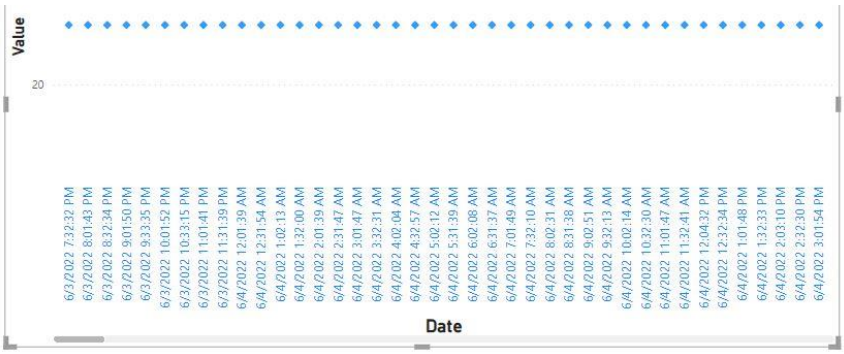


Figure 9: Constant data

2.3.4 Missing Data

According to [46], information lacking a meaningful response, such as "Don't know," "Refused," and "Unintelligible," might be regarded as missing information because they must be ignored. This is explained in terms of subjective human consciousness. Because the data gathered by sensors has its predefined and obvious meaning, it occasionally happens frequently in a questionnaire but seldom shows up in a monitoring system. According to the

definition that is most frequently used, "if one or more attribute values of a record in a data source are null, the record is called incomplete data or data with missing value," missing data is generally understood to refer to incomplete data entries. Based on this concept, the majority of the missing data phenomenon in monitoring systems is characterized. Here, in the figure-10, it can be observed that, the sensor has missing value for a certain time therefore the reading is showing empty.

Figure 10: Missing Data

Missing sensor data is inevitable to occur because of many reasons such as communication failure, hardware damage, security attacks, connection error, sensor faults and run out of power. In addition to that, the missing observations tend to occur in intervals where a sensor may stop functioning for several days in a particular place before it is restarted. This missing sensor values can lead to several data analysis problems if it not handled properly. Several studies have done for the missing value detection and imputation including the use of Machine learning and Artificial Neural Network.

The study of M. Caselli, L. Trizio, G. de Gennaro, and P. Ielpo, [47] was to realize and to compare two support decision system (neural networks and multivariate regression model) that, correlating the air quality data with the meteorological information, are able to predict the critical pollution events. The key point in using SVR for forecasting is how to determine the appropriate parameters. The study[48] proposed SVM (Support Vector machine) and linear regression for predicting electricity price. To their study they took the historical dataset at day level distribution. D. A. Guastella, G. Marcillaud, and C. Valenti explained[49] proposed a method, during the data collection phase, a unique mechanism for missing data imputation was devised. According to their method, which was based on the edge computing paradigm, the authors distributed computation across a variety of stationary, enabling the network to scale horizontally and increase the number of sensing devices while decreasing the impact of missing values caused by sensing errors. RNN (Recurrent Neural Network) is being used in the study[50] which can do the AR and MA at the same time. The study result (prediction) compared with the double seasonal time series in ARIMA model.

3 Data collection and finding inconsistency

Most used method to find inconsistency in data is visualization. By plotting the dataset in different types of charts like as bar chart, line chart and scatter chart. For checking the recorded data in different sensor, the bar chart has been used. Through this chart, it has counted the number of data has recorded by each sensor. And for inconsistency in a specific sensor data, the line chart has been used. For outlier detection, the study [51] discussed a method which gave the algorithms distinguish between outlying sensors and the event border by using the correlation of readings among nearby sensor nodes. Each node calculates the difference between its own reading and the median reading from its nearby readings as part of the technique for locating outlying sensors. Then it standardizes all variations from its locality. If a node's absolute reading deviation degree is sufficiently greater than a chosen threshold, it is regarded as an outlier reading.

In addition to that the gaussian distribution also gave the overview of data that how the data points are co-related. It can be easily calculated by the following formula. $z = (x-\mu)/\sigma$, here z is the value of z score, x is the individual sensor reading, μ is the mean value of the sample dataset and σ is the standard deviation of the dataset. If the score value is in between (-3 to +3) then it can be said that the data is well distributed, if it goes beyond the range that can be considered as an outlier (figure-11).

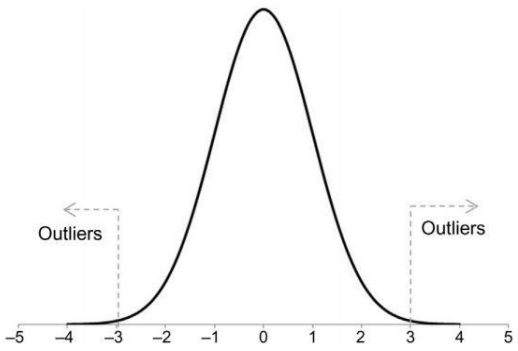


Figure 11: z-score range [52]

For visualizing and analysis purposes, several software's and methods are being used. These softwares are made up of various applications and the connector, which can generate massive visual observations and insights and can also be analyzed using the interconnected programming language. Furthermore, it allows the user to import the dataset from various files; in our case, we use these softwares to import a large dataset, filter it into different categories, and then visualize and analyze the data.

3.1 Available data

A number of sensors are installed in the second floor of Nürnberger-Ei (figure-12) in order to measure environmental components like Temperature, Relative Humidity, Carbon-dioxide and also brightness. A yearlong dataset started from 8th October 2021 and ended at 9th September 2022 is considered for the analysis. The data is collected from five different

rooms and twenty different sensors. The dataset is collected from the 'Home assistant system' by setting some query to avoid the unnecessary data for the study. Then the data is downloaded from the system as a CSV file.

Figure 12: Nürnberger-Ei

3.1.1 Sensor information

Sensor type 1: This sensor is used for measuring the brightness of the room and also for movement detection.

Figure 13: Brightness and motion sensor

Technical details:

- Measured values: brightness, movement
- Radio technology: EnOcean (IEC 14543-3-10)
- Frequency: 868 MHz
- Power supply: solar cell, internal super cap, backup battery LS14250 (3.6 V)
- Measuring range light 0 -1020 lux

Sensor type 2: This type of sensor is being used to get the data for three different environmental parameter called Temperature, Humidity and the Carbon-dioxide.

Figure 14: Temperature, Humidity and Brightness sensor

Technical details:

- Measured values: CO2, temperature, humidity
- Radio technology: EnOcean (IEC 14543-3-10)
- Frequency: 868 MHz
- Power supply: solar cell, internal super cap, support battery LS14250 (3.6 V)
- Temperature range: 0 to + 51°C
- Measuring range humidity: 0 to 100% rH without condensation
- Measuring range CO2: 0 to 2550 ppm
- Accuracy in temperature: ±1°C of measuring range
- Accuracy in humidity: ±3% between 20 to 80% rH
- Accuracy in CO2: ±75 ppm, above 750 ppm: ±10% variation

Room	Sensor reference	Sensor name	Unit
Room 201	sensor.sr04_01_temp	Temperature sensor	°C
Room 201	sensor.sr04_01_hum	Humidity	%
Room 201	sensor.sr04_01_co2	CO2	Ppm
Room 201	sensor.mds_01_brt	Brightness	lux
Room 202	sensor.sr04_02_temp	Temperature sensor	°C
Room 202	sensor.sr04_02_hum	Humidity	%
Room 202	sensor.sr04_02_co2	CO2	Ppm
Room 202	sensor.mds_02_brt	Brightness	lux
Room 204	sensor.sr04_03_temp	Temperature sensor	°C

Room	Sensor reference	Sensor name	Unit
Room 204	sensor.sr04_03_hum	Humidity	%
Room 204	sensor.sr04_03_co2	CO2	Ppm
Room 204	sensor.mds_03_brt	Brightness	lux
Room 210	sensor.sr04_04_temp	Temperature sensor	°C
Room 210	sensor.sr04_04_hum	Humidity	%
Room 210	sensor.sr04_04_co2	CO2	Ppm
Room 210	sensor.mds_06_brt	Brightness	lux
Room 213	sensor.sr04_05_temp	Temperature sensor	°C
Room 213	sensor.sr04_05_hum	Humidity	%
Room 213	sensor.sr04_05_co2	CO2	Ppm
Room 213	sensor.mds_07_brt	Brightness	lux

Table 2: Sensor list with the room number and references

3.2 Weather data for analysis

The outdoor temperature has a great influence in indoor temperature therefore outdoor temperature analysis also important for this research. In the study, two different weather station data is taken to compare the weather component (temperature) with the institute data. First set of data is taken from wetterzentrle.de, here the data is collected from the weather station in Dresden-Strhlen which is the nearest weather station to the Nürnberger-Ei. The figure below shows the daily temperature variation from October 2021 to September 2022. There are four different aspect is taken for everyday temperature as normal temperature as gray line, average temperature as black line, the red line for the maximum and blue line for the minimum. The temperature ranges in between -8°C to 37°C.

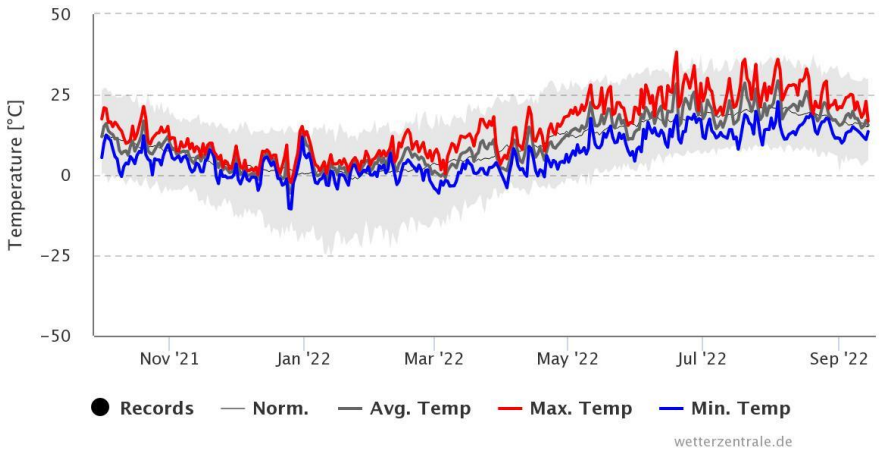


Figure 15: Temperature (Weather Station: Dresden-Strehlen)

The second set of data is collected from another website named “visual crossing”. Here it’s presented the weather data of Dresden. This dataset also taken from October 2021 to September 2022. Here the minimum daily average temperature is about -6°C and the maximum is about 30°C.

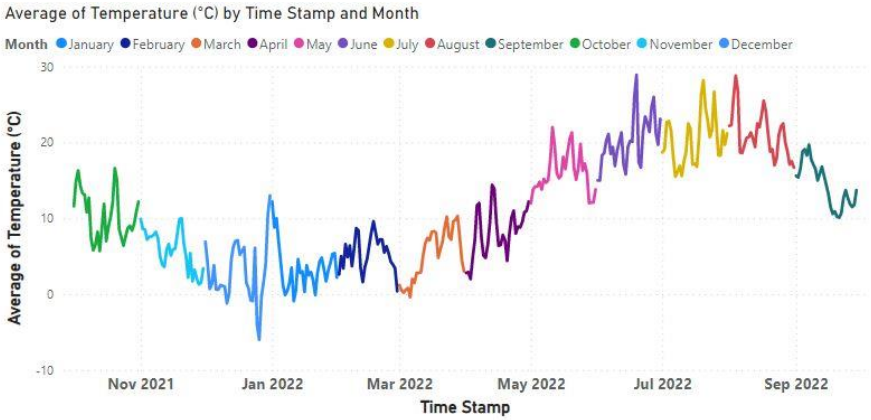


Figure 16: Temperature from visual crossing

3.2.1 Day level comparison of the weather data

The temperature reading from two different day of August 2022 is considered and checked how much is the difference in temperature in two different datasets. In 10th of August the temperature showed in the Strehlen weather station 20.8°C (right) and in visual crossing (left) it is 20.7°C.

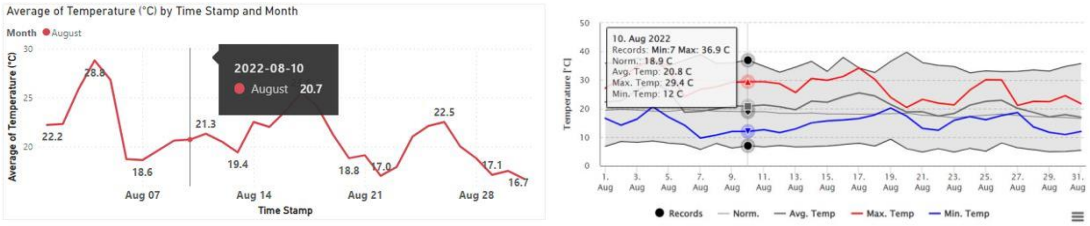


Figure 17: Temperature comparison (1)

Also, in 20th August it shows the same temperature (18.8°C) in both datasets (figure-18). So, the temperature dataset is being considered for comparison with the sensor reading (temperature) of institute building.

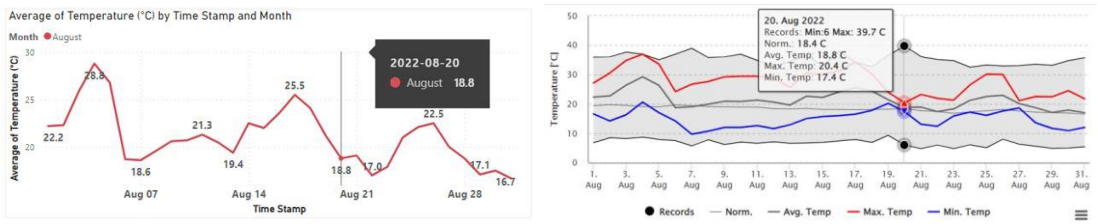


Figure 18: Temperature comparison (2)

3.3 Room wise analysis

The sensors are placed in five different rooms and each room has four types of sensors. The room size, uses, occupancy and the orientation are different from each other. Therefore, room-wise analysis has been done in the following. The following figure shows the considered rooms and their orientation in the building.

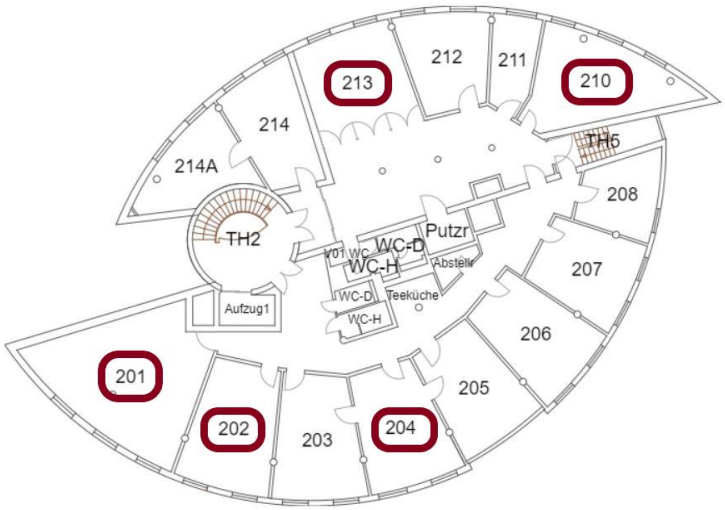


Figure 19: Room orientation

3.3.1 Room 201

Temperature

The sensor generated 7791 readings on an average of 23 reading per day. Among the reading, 44 readings are shown as unknown. The maximum temperature recorded 32.4°C on 28th of June 2022 and the minimum is recorded on 12th August 2022 is 18.6°C (figure-20).

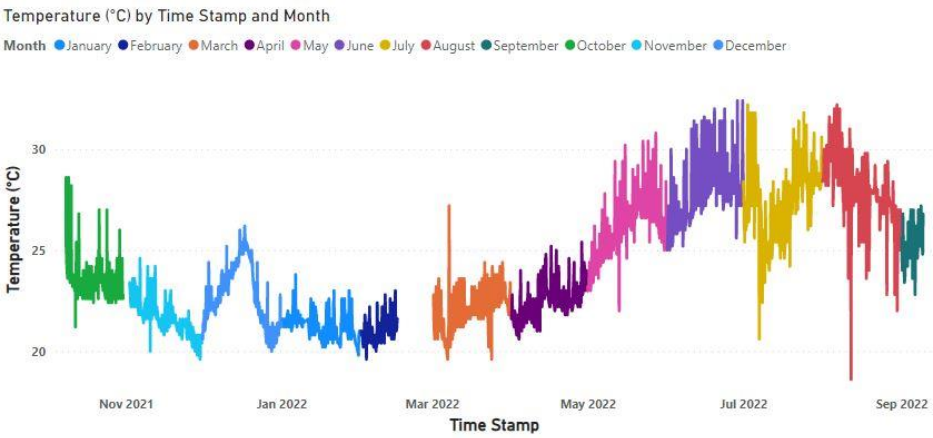


Figure 20: Temperature at Room-201

Here it is noticeable that the number of recorded readings varies with the month. From the data distribution it is assumed that, in some month's temperature variation is very less compared to other months, it's due to the less change of temperature on that specific month (figure-21-left). In case of outlier checking, this dataset is free from outliers, though the chart shows three rapid changes (figure-20). The sudden rise and drop didn't cross the limit

of the outliers. In addition to that, if we consider the comfort level for temperature then it can observe that in the summer the temperature goes above 28°C in June, July and August. Here in the following figure-21 (left) showing the total number of readings by months. Figure-21 (right) showing the number of unknowns by months. In the sensor dataset for January, six readings are recorded as unknown. The following readings are also recorded as unknown: February: 1, March: 4, April: 4, June: 8, July: 5, August: 8, October: 6, and November: 2. The maximum number of unknown readings is observed in June and August, and the minimum number of unknowns is in February.

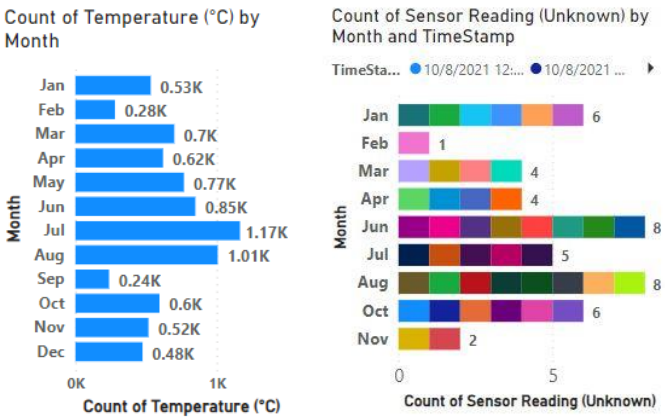


Figure 21: Count of data and number of unknown at Room-201

From the available data, it is observed from 30th October 2021 to 2nd November 2021 no data has been recorded. One more missing dataset observed from 15th February 2022 to 1st March 2022 (figure-22). This similar situation happened to all 20 sensors.

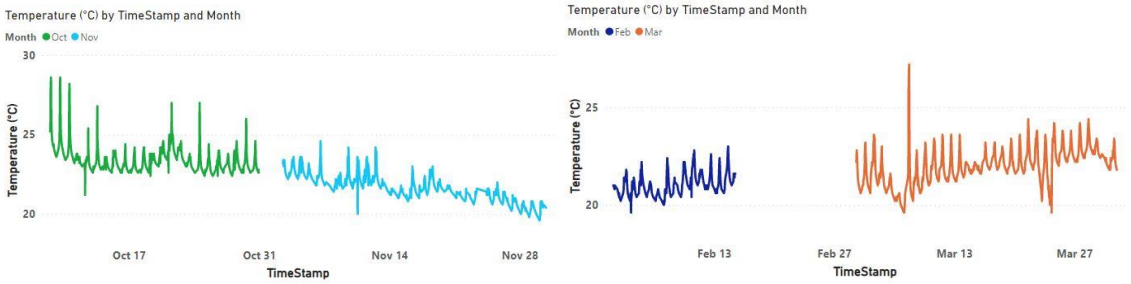


Figure 22: Data missing at Room-201

Humidity

The Humidity sensor received 6721 reading among these 44 reading are unknown. The humidity ranges between 14% to 53.5% and the average humidity 29.13% (figure-23). Several studies have done on temperature and humidity correlations. According to [53], [54] these two environmental parameters are closely related. The sensor for temperature and humidity in this room also shows the similar number of readings for the time being.

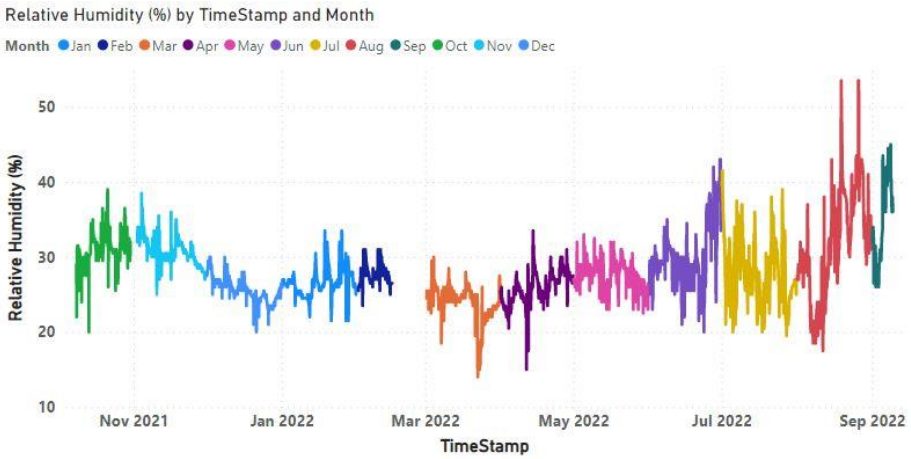


Figure 23: Humidity (%) at Room-201

The indoor humidity range should be 30%-60% and the recommended humidity range is 40-60% for healthy indoor environment. If we observe in our dataset (figure-23), it's showing the humidity level is below 30% most of the time and very few days it goes in between the standard range. The bar chart below shows the total number of sensor readings by month and also shows the number of unknown by month. Maximum number of sensor reading is recorded in July and the minimum number of sensor reading is observed in February. Here, unknown reading distribution is the same of temperature sensor.

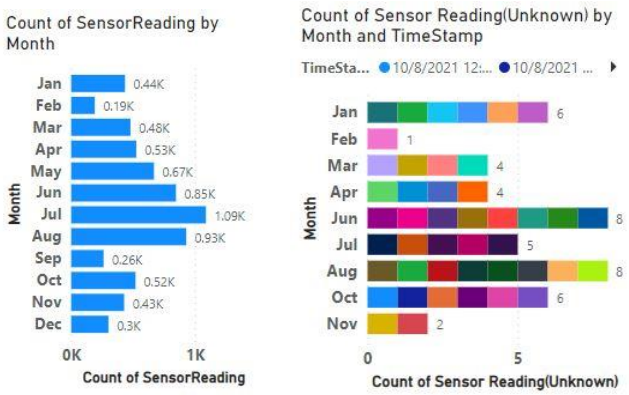


Figure 24: Reading and the unknown data at Room-201

CO2

The CO2 sensor recorded 16422 readings. In this case also 44 reading is recorded as unknown. The data recording in CO2 sensor is more than twice compared to temperature sensor and humidity sensor. This indicates the change of CO2 in air has changed more rapidly compared to other two weather parameters. Figure-2 shows the maximum CO2 reading is 2550 ppm which is the sensor maximum range. This indicates that the CO2 reading could be more than that the upper threshold. The minimum amount CO2 has recorded 340 ppm. And it is also noticeable that the CO2 data shows the distribution throughout the time is irregular as it strongly depends on the occupancy in the room.

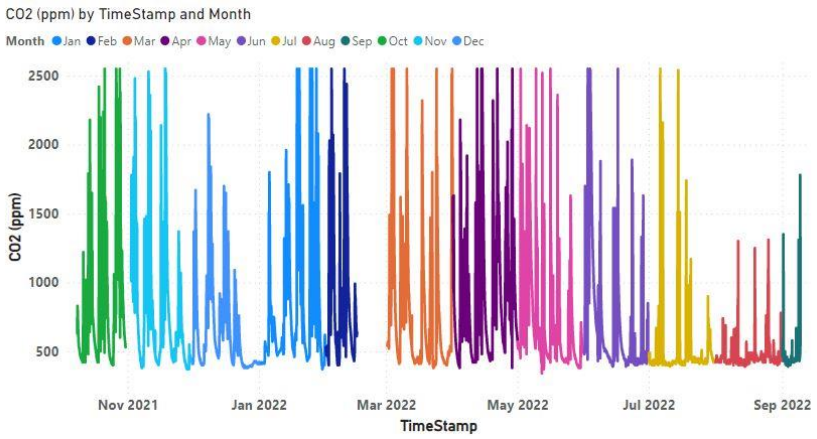


Figure 25: CO2 (PPM) at Room-201

The figure-25 indicates the dataset for CO2 sensor is also well distributed though the sensor reading varies every month. There is no inconsistency except the missing value and the unknowns. The following bar charts are showing the number of sensor reading by month and the number of unknown reading by month for room-201.

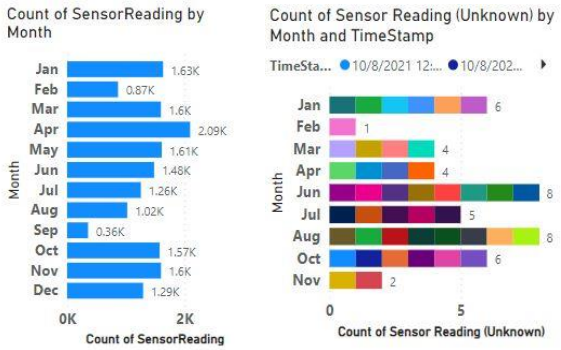


Figure 26: Reading and the unknown data at Room-201

The recording of CO2 sensor goes above 2000 ppm (figure-25) for 898 times and also reached threshold values 2550 ppm for 35 times. It indicates the reading may goes above 2550 ppm. According to the German standard of CO2 for indoor environment, the CO2 should not cross 2000 ppm which considered hygienically unacceptable. Moreover, it's recommended to keep it below 1000 ppm and step required if it shows the range between 1000-2000 ppm. The following figure is showing the CO2 distribution above 2000 ppm. After observing the dataset in day level in two different days in two different months, CO2 concentrations above 2000 ppm mostly happened on Tuesday and Thursday.

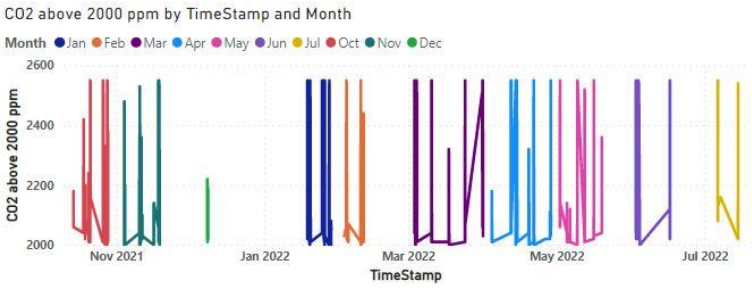


Figure 27: CO2 reading above 2000 ppm

Brightness

The sensor recorded 37699 reading which is maximum reading among all four sensors in the room. The figure-28 is showing maximum reading for this sensor is 1020 lux, which is the upper threshold for this sensor. The 1020 repeated many times from the beginning till the end. And the minimum reading for the sensor goes to 0 at night time mostly. The number of total readings for this sensor is 2-4 times higher than the other three sensors in this room.

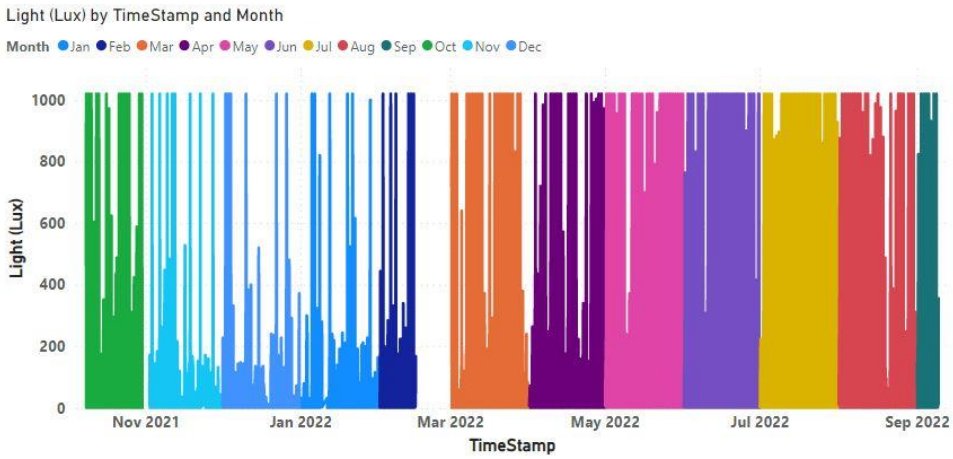


Figure 28: Brightness (Lux) at Room-201

Here the number of unknown data is also 44 though the sensor has recorded the maximum reading in this room. The data recording rate is higher in the summer compare to winter as light varies more in the summer time. The following bar chart is showing the total number of sensor readings by month. From April to August the sensor recoded around 5000 data in each month. And in winter time the recording is around 1500, which is very less compare to summer time. The sensor data also shows a similar distribution of unknowns with other sensor installed in this room.

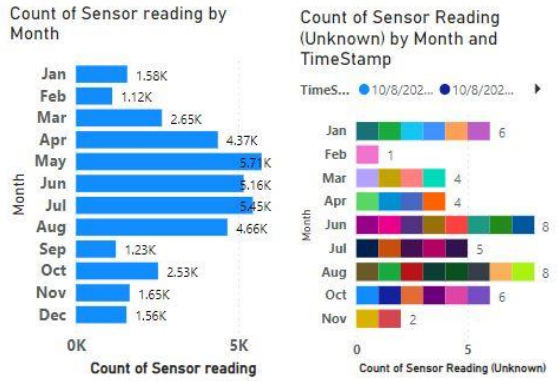


Figure 29: Reading and the unknown data at Room-201

3.3.2 Room 202

Temperature

In this room the total number of the reading for the temperature sensor is 10495, which is higher compare to the room 201. The temperature range between 16°C and 32.2°C (figure-30). The room is more exposed to the sunlight which may occur the higher number of the sensor reading. The following figure also shows some reading for temperature fall down rapidly. But it cannot be considered as outlier because it has not exceeded the given range.

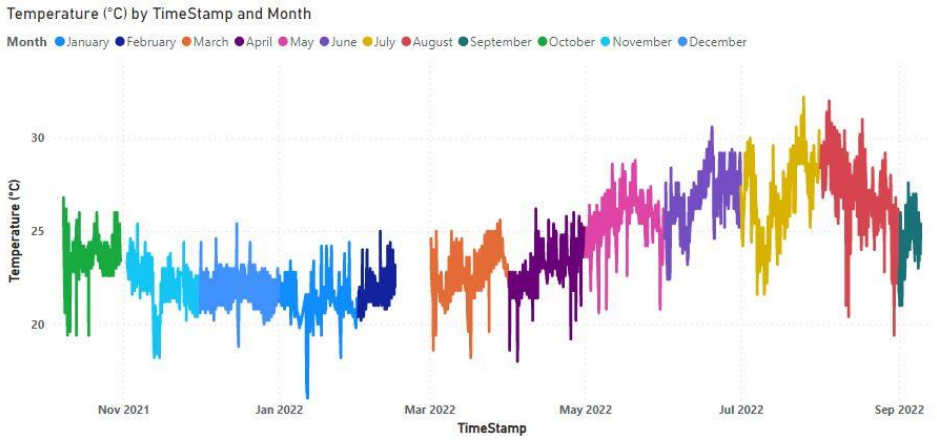


Figure 30: Temperature (°C) at Room-202

The temperature of this room exceeds the comfort level. Like the room 201, this room temperature also goes above 28°C but this happened fewer times compare to room 201. The recording of the temperature reading mostly similar in every month (figure-31). The unknown for this sensor is also 44 and the distribution of this unknown similar to the previous sensors.

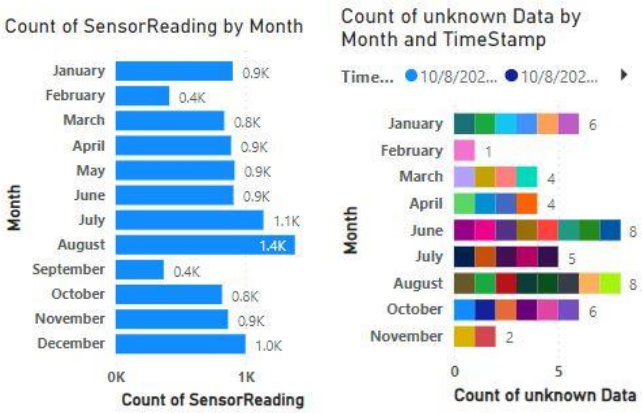


Figure 31: Reading and the unknown data at Room-202

Humidity

The sensor reading for humidity sensor is 7869. The minimum humidity reading is 13% and the maximum is 55% (figure-32). The difference between temperature and humidity reading is higher in this room. The relative humdity of this room is also below 30% for the time being. The figure also shows that, only in summer time the huimidity level goes above 40% which is recommended for the human comfort.

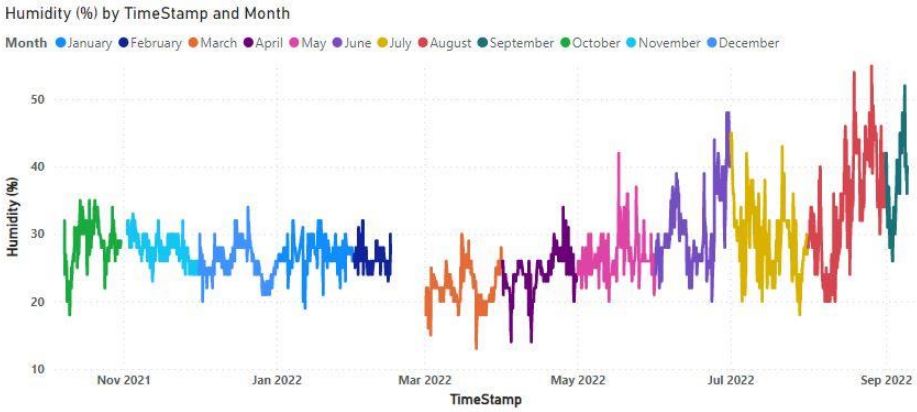


Figure 32: Humidity (%) at Room-202

The total reading of humidity in this room is higher than the room 201 and the unknown reading is same as previous sensors also the monthly count of it (figure-33). This sensor data is consistent and there is no outlier in it but missing value like previous sensors. The monthly count of the sensor reading also varies with the time.

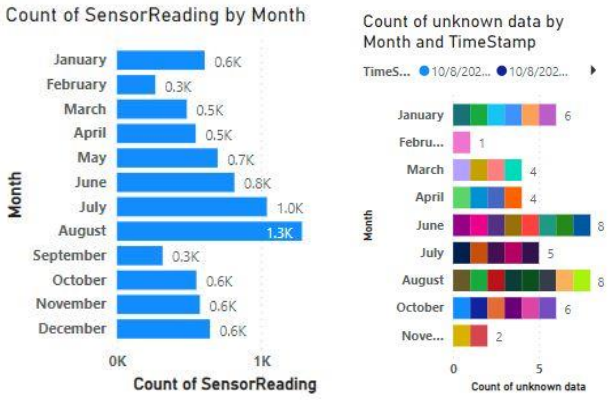


Figure 33: Reading and the unknown data at Room-202

CO2

Total reading is 15751 for this sensor. The maximum CO2 reading is 2550 ppm and the minimum is 360 ppm. From June to September the CO2 concentration in this room is in acceptable range (<2000 ppm).

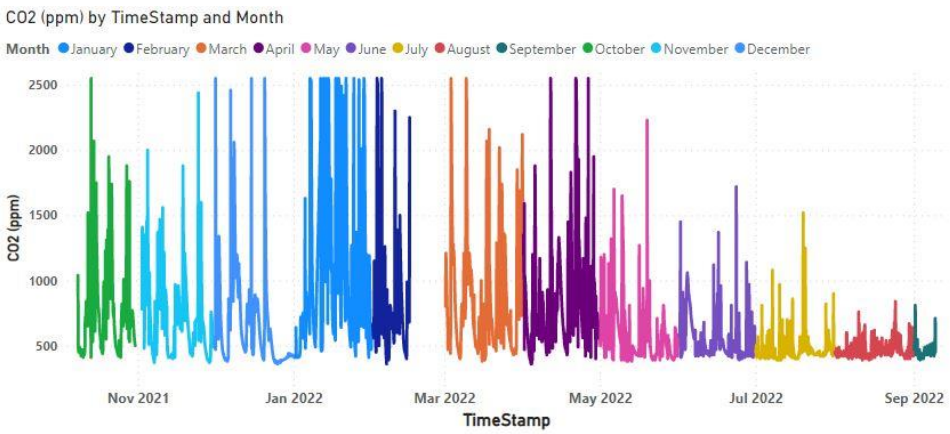


Figure 34: CO2(PPM) at Room-202

The reading above 2000 ppm in this room also noticeable also the threshold value of the sensor. A total of 554 reading goes above 2000 ppm among these 22 reading is 2550 ppm. This sensor shows that the maximum data has been recorded in the month of January as the total CO2 concentration is higher in this month compare to others. This sensor data also well distributed and there is no outlier in this dataset. This dataset also has missing value and the unknowns which considered as inconsistency.

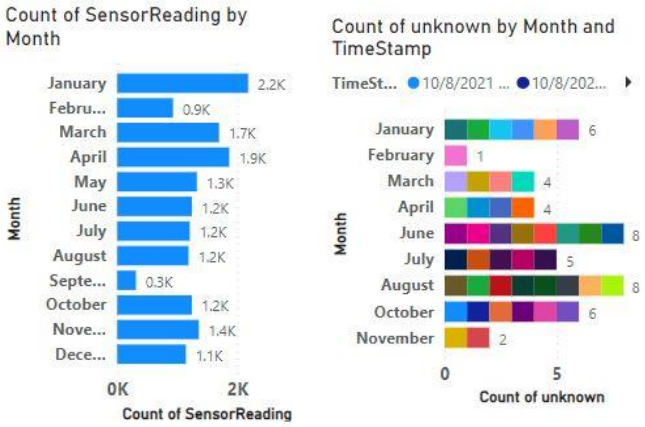


Figure 35: Reading and the unknown data at Room-202

Brightness

Brightness sensor in this room also recoded highest reading which is 38857. Maximum value for the brightness is 1020 lux and minimum is 0 lux. The following figure shows that, the higher lux value in winter season in this room, that is the difference of this room compare to previous rooms.

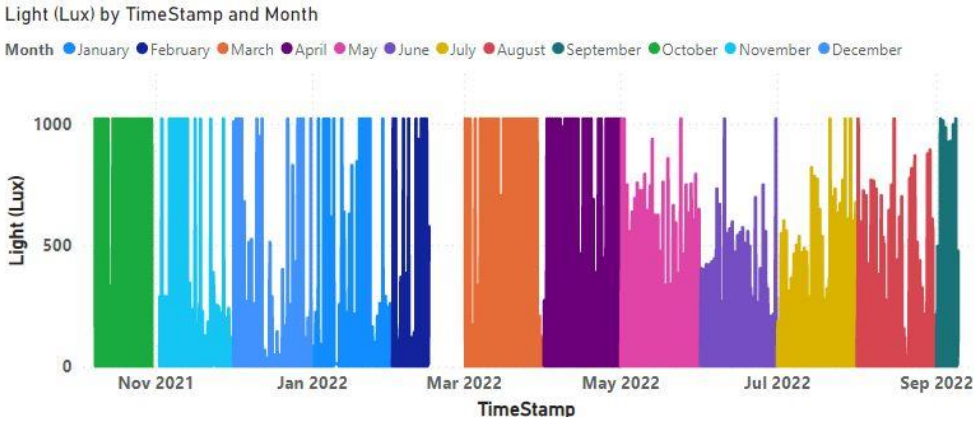


Figure 36: Brightness (Lux) at Room-202

The sensor readings for brightness as usual the highest. The sensor recorded most reading in the summer times and the brightness is less compare to the winter season. One more thing in this sensor is noticeable that the month of February and September has the minimum sensor reading because of missing value and incomplete data. The unknown value for this sensor is also 44 and the distribution of the unknows are similar to the previous sensor (figure-37).

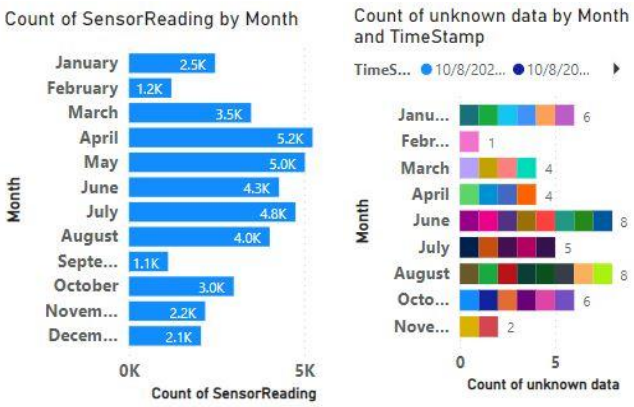


Figure 37: Reading and the unknown data at Room-202

3.3.3 Room 204

Temperature

Total reading for this sensor 3845. This sensor has recorded very less reading compare to room 201 and 202. The temperature range between 15°C and 30°C which is almost in the acceptable range. The temperature reading for this sensor also follows the pattern of the previous sensor. From the following figure it can be observed that, in this dataset there is no rapid increase or decrease of temperature.

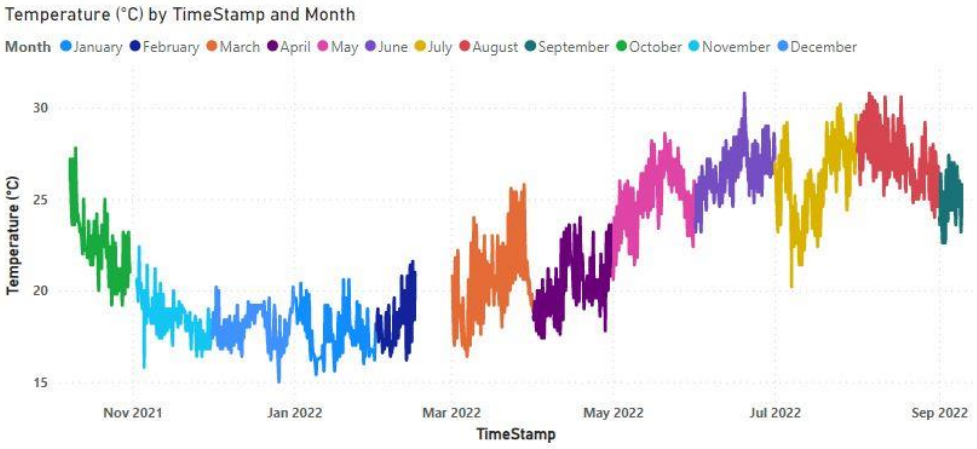


Figure 38: Temperature (°C) at Room-204

The recorded data for this sensor has similar distribution in the whole months except the month of February and September. This is one of the least data recorded sensor but the unknown value for this sensor also 44. There is no other inconsistency has found except the missing data and the unknown values.

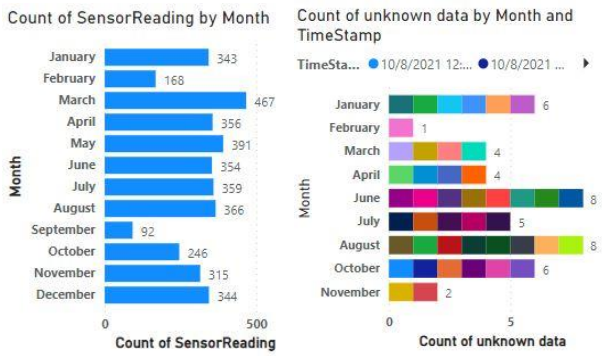


Figure 39: Reading and the unknown data at Room-204

Humidity

Total reading 3592. The range humditiy is in between 9% and 49%. This room the humdity level is very low also the sensor reading. The data distribution in the chart also shows that, the dataset of the sensor is not varies that much. The humdity level of this room is not in the comfortable range. The following chart indicates that, the humdity stayed below 30% most of the time.

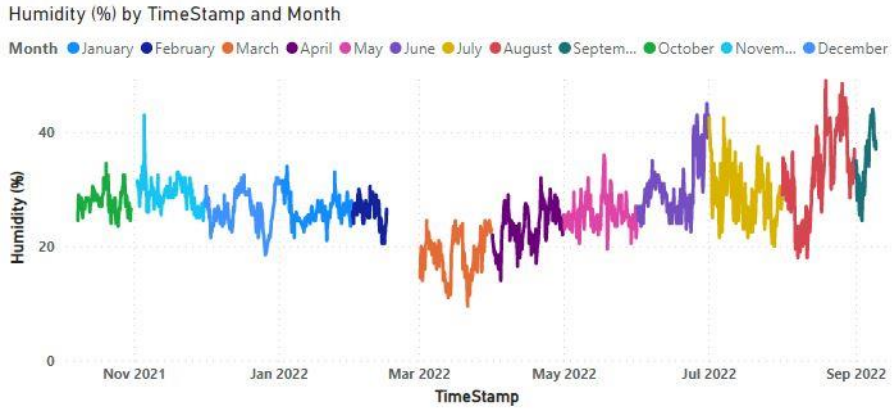


Figure 40: Humidity (%) at Room-204

The sensor of humidity recorded the reading throughout the time has similarity although the recording rate is very low. This sensor also has the unknowns and the missing values (figure-41).

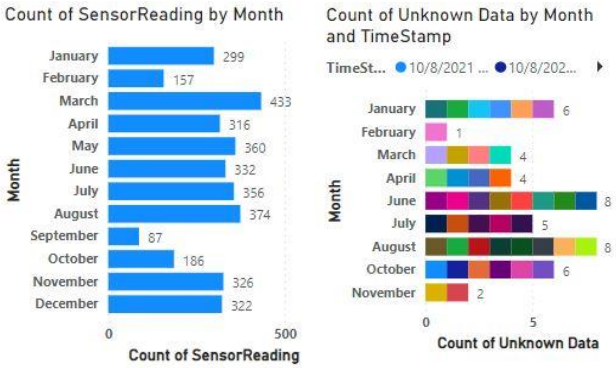


Figure 41: Reading and the unknown data at Room-204

CO2

Reading for the CO2 concentration is 4827. Here the maximum reading is 1540 ppm and the minmium reading is 240 ppm. Like the other two sensor the CO2 sensor reading is also very less in this room. Here in the following figure, it's noticeable that the sensor reading has't gone beyond the acceptable limit. The CO2 concentration in this room mostly stayed below 1000 ppm, which is recommended for good hygiene.

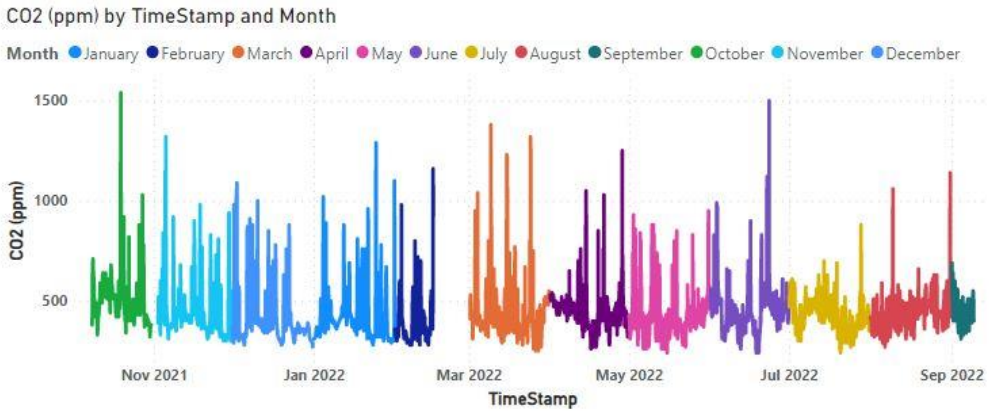


Figure 42: CO2(PPM) at Room-204

In this room the reading for the CO2 sensor showed the similar distribution. As the concentration of CO2 is very low, the reading of the sensor is very less. This sensor has also the unknowns and the missing values which considered as inconsistency of the dataset (figure-43)

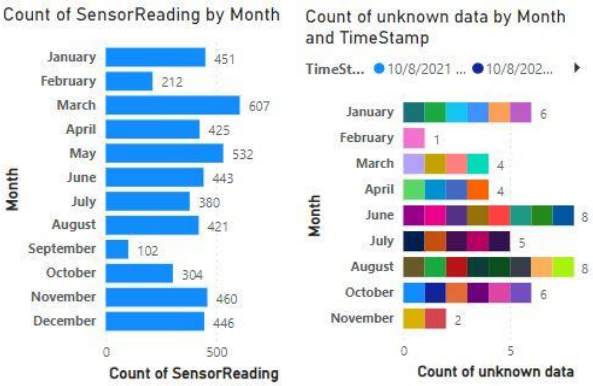


Figure 43: Reading and the unknown data at Room-204

Brightness

The sensor recorded 42101 reading and the maximum reading is 1020 lux and the minimum reading is 0. The lux value mostly higher in comparison with other sensor. From the plan it can be observed that this room is exposed to the sun most of the time of the year.

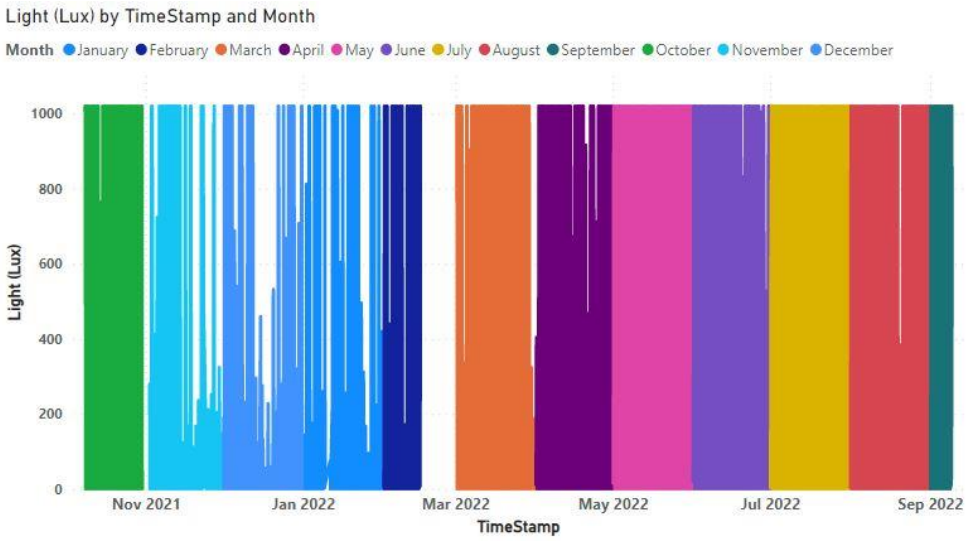


Figure 44: Brightness (Lux) at Room-204

The lux value also reached the upper threshold for many times in this room. From following bar chart, it can be observed that, the record of the sensor reading is higher in the summer months compared to the winter. This sensor has also 44 unknowns.

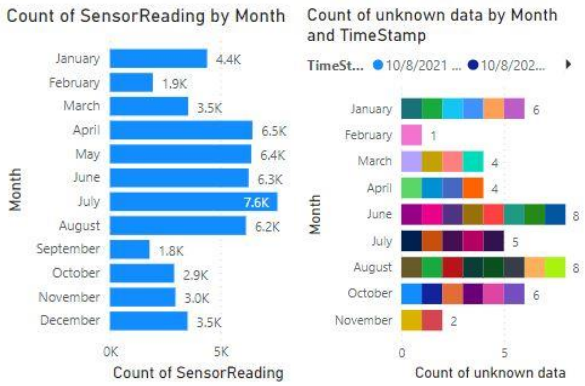


Figure 45: Reading and the unknown data at Room-204

3.3.4 Room 210

Temperature

Total reading for this sensor is 2839. The maximum temperature is recorded 29.8°C and the minimum temperature is 15°C (figure-46). This sensor data distribution with times shows a good comfort range excepts the summer times. There are some rapid changes can be observed in the chart but it's in the range and the consecutive data has also have the rapid change.

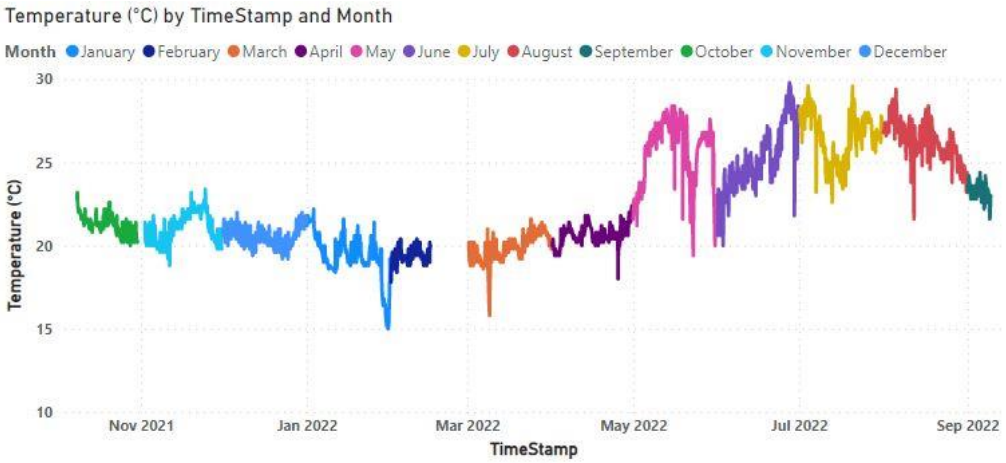


Figure 46: Temperature (°C) at Room-210

The rate of change of temperature in this room is very low as it recorded least amount of reading. The following bar chart shows, the collection range is almost similar except February due to missing data and September because only nine days data has been collected. This dataset is free from outlier. The number of unknowns in this sensor is 44, and the distribution of unknowns is similar to the other sensor (figure-47).

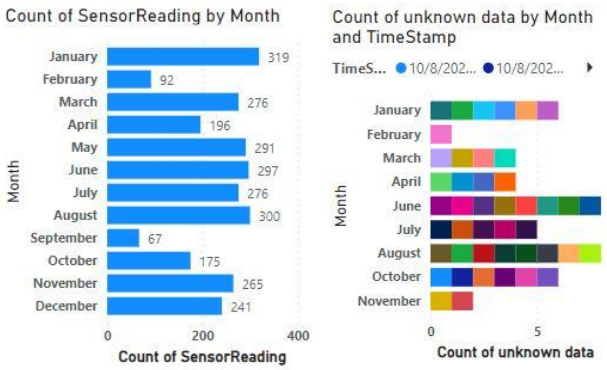


Figure 47: Reading and the unknown data at Room-210

Humidity

This sensor has recoreded a total 3387 reading. The total reading of this sensor is also very low compare to other sensor. The maximum humdity recorded 54.5% and the minimum humidity in this room is 13.5% (figure-48). The humdity level in this room is not in the comfortable range. The change of humdity in the room has not exceded the given range and there is no noticable abnormal situation obsereved in the dataset.

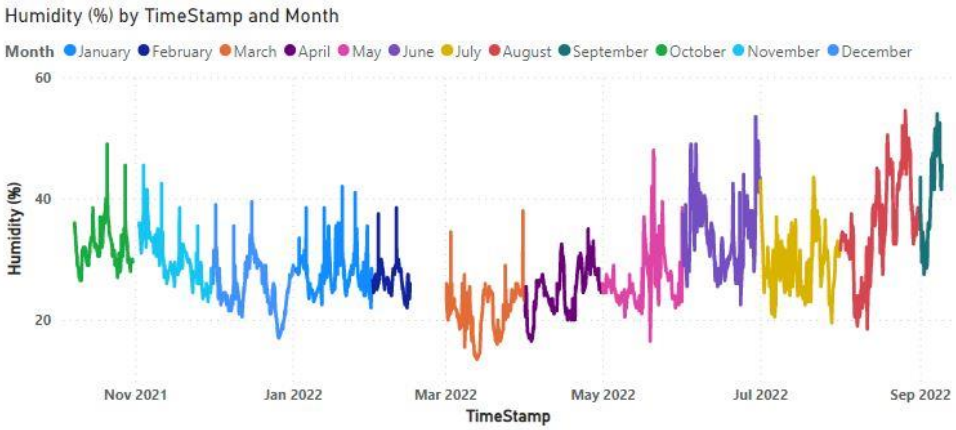


Figure 48: Humidity (%) at Room-210

Here the following bar chart is showing the number of readings recorded by the sensor by month. Also, the unknown reading by month. The total number of unknowns in this sensor is 44.

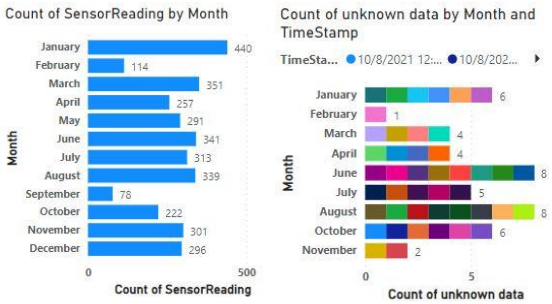


Figure 49: Reading and the unknown data at Room-210

CO2

Total reading for this sensor is 3942. The maximum CO2 concentration for this room is 2090 ppm and the minimum is 390 ppm (figure-50). This room has very few changes in CO2 concentration compare to other rooms specially room 201 and 202. Besides that the CO2 distributtion range stayed in acceptable range, except few reading which goes beyond 2000 ppm.

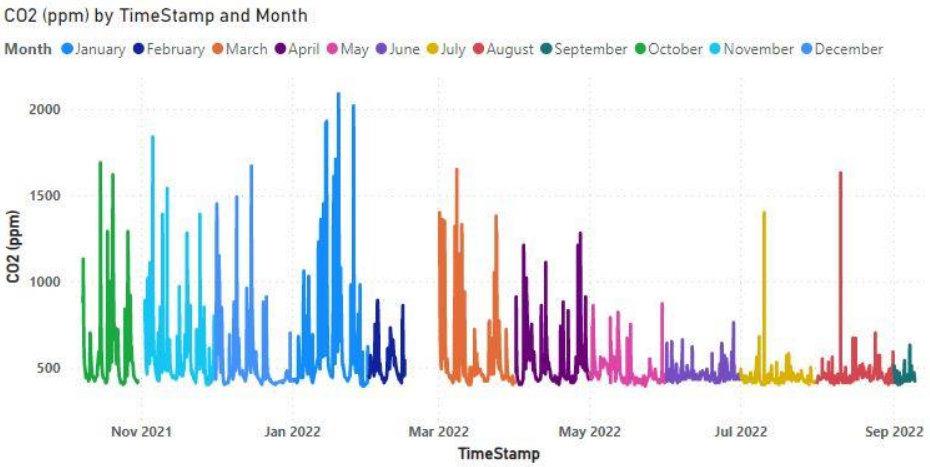


Figure 50: CO2(PPM) at Room-210

During the winter the CO2 concentration is higher compare to summer season. The dataset contains with the missing values and the unknowns. There are some sudden increases of the reading can be observed in the chart but it cannot be considered as outlier. The reason behind the sudden increase or fall could be the reason of occupancy increase or decrease in the room. The following chart is showing the normal and unknown reading distribution by months.

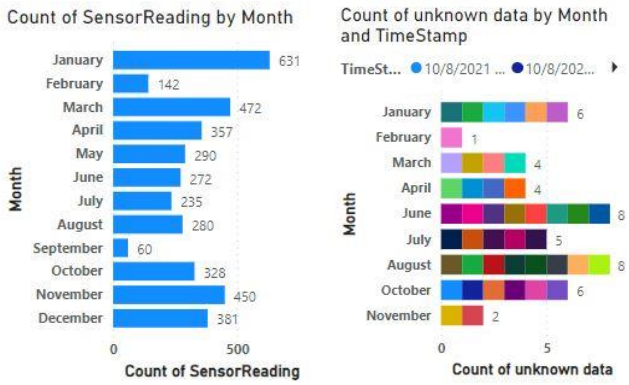


Figure 51: Reading and the unknown data at Room-210

Brightness

The total number of reading is 53985. Here the maximum reading for brightness is 1020 lux and minimum brightness is 0 (figure-210). The upper threshold value 1020 lux came 937 times. As usual, the brightness sensor recorded the maximum number of the reading in this room. The reading shows that it is in comfortable range though the lux value exceeds the required brightness for the office work.

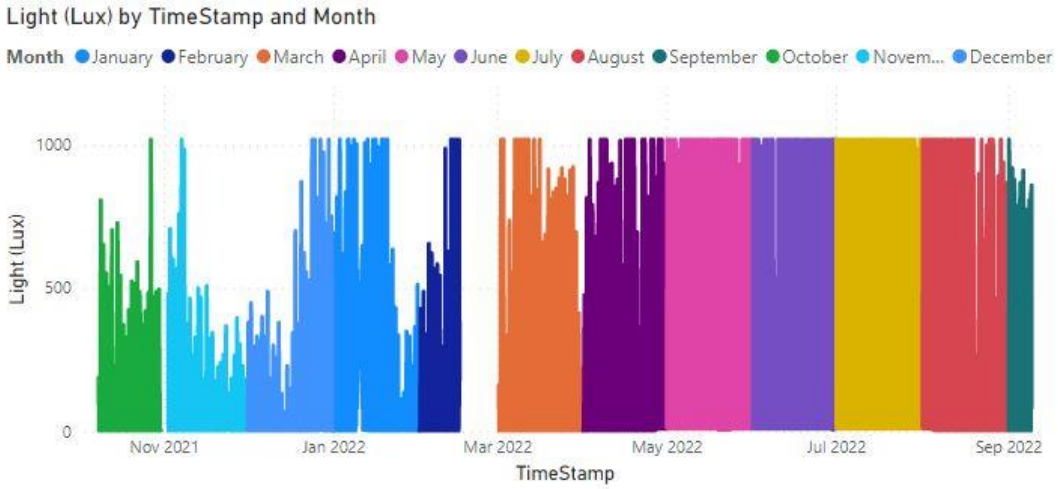


Figure 52: Brightness (Lux) at Room-210

This dataset is free of outliers. The missing values and unknows of the dataset are observed as previous sensors. The following chart is showing the normal and unknown reading distribution by months.

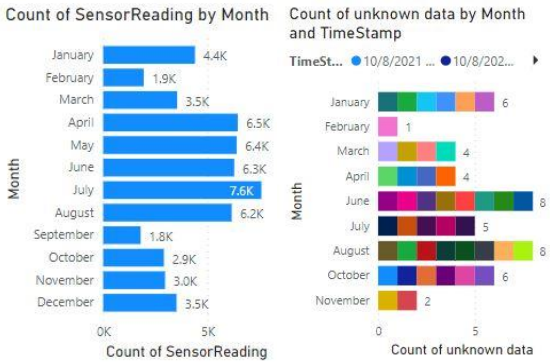


Figure 53: Reading and the unknown data at Room-210

3.3.5 Room 213

Temperature

The sensor has recorded a total number of 5032 reading. The maximum temperature has recorded 51°C and the minimum temperature has recorded as 16.4°C. Here a noticeable temperature reading has observed which is 51°C (figure-54).

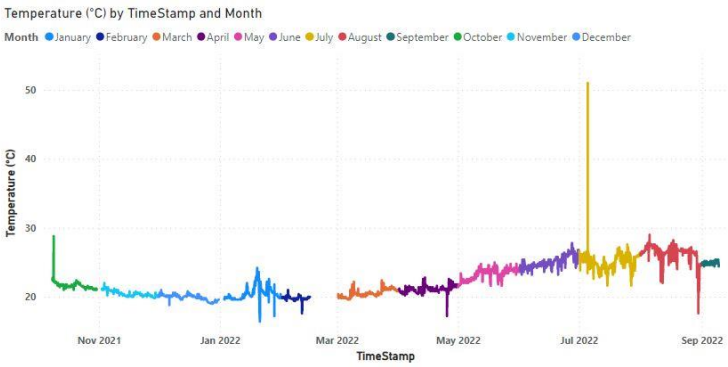


Figure 54: Temperature (°C) at Room-213

From the following bar chart, it can be observed that the data recording varied a lot with the time. The noticeable difference of data recoding can be observed between the summer and winter season. Here the orientation of the room has played a very important role in the monthly recording. As the room is not exposed to the sun therefore change of temperature is very less in the winter season. The unknowns for this sensor is also 44 and the distribution is similar to other sensors.

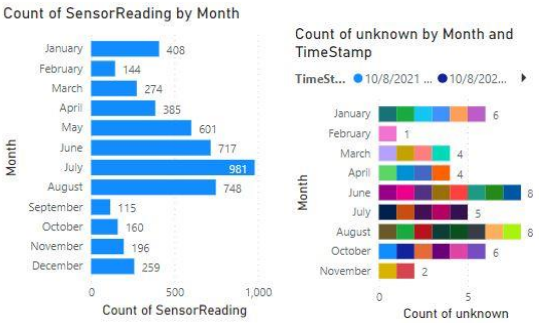


Figure 55: Reading and the unknown data at Room-213

On July 5th the sensor has recorded the temperature 51°C which is the maximum threshold for the sensor (figure-56). The other readings of this sensor stayed below 30°C. So, it is considered as a point outlier.

Figure 56: Point outlier in Temperature dataset at Room-213

Humidity

Total reading for this sensor 8622. The maximum reading for humidity is 84% and the minimum humidity is 23% (figure-57). This room humidity is in the range of comfort mostly.

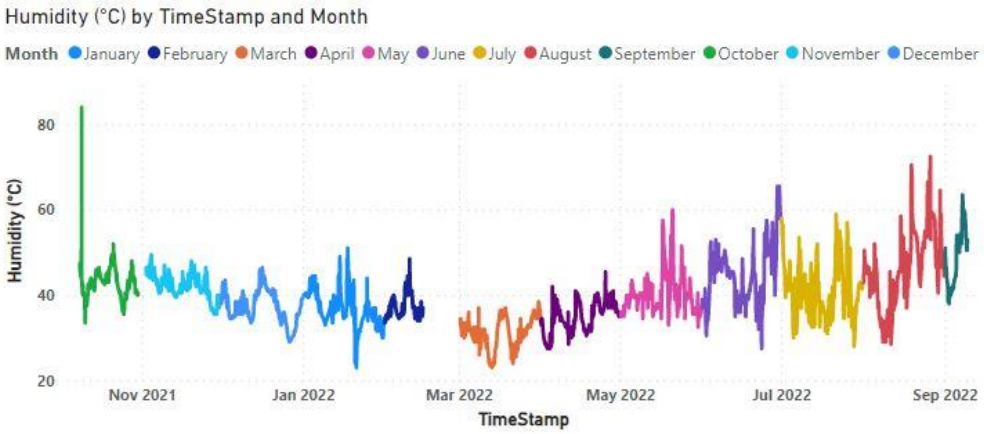


Figure 57: Humidity (%) at Room-213

This sensor has recorded a good amount of reading compare to the previous room. The sensor data is well distributed except one data point (figure-58). This sensor has an outlier also the missing values and the unknown data.

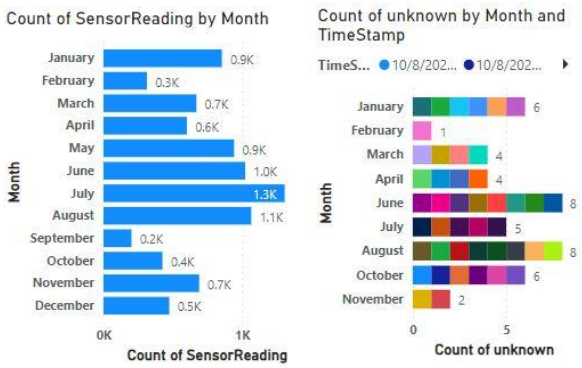


Figure 58: Reading and the unknown data at Room-213

Here in the following figure, it is showing on 10th October the humidity reading is 84% and the consecutive data below 45%. The outside humidity at the same time was 74.84%. So, it considered as outlier in the dataset.

Figure 59: Humidity comparison with weather data

CO2

Total recorded data for this sensor is 15849. The maximum value for CO2 sensor is 2050 ppm and the minimum value is 410 ppm (figure-60). The following figure shows that, the sensor recoded the higher value of CO2 concentration in the winter time compare to summer. The data has a good distribution. It hasn't gone beyond the acceptable limit except few days.

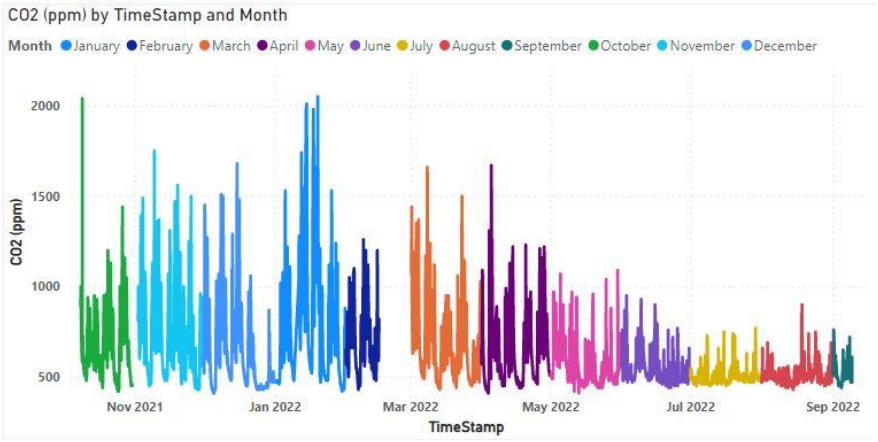


Figure 60: CO2(PPM) at Room-213

The sensor recorded higher number of readings compare to the previous room. This sensor has the missing values and also the similar number of unknowns as previous (figure-61). The unknown value distribution in every month has also the similarity with the other sensors.

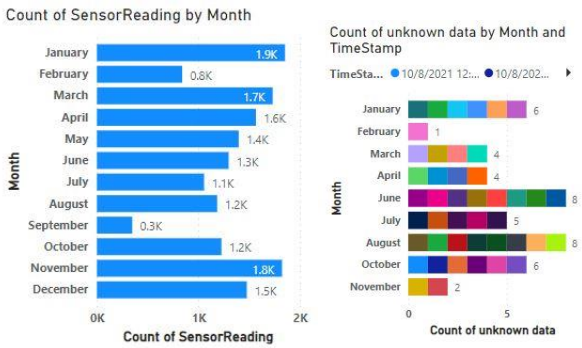


Figure 61: Reading and the unknown data at Room-213

Brightness

Total reading for this sensor is 36935. The maximum value is 1020 lux and the minimum for this sensor is 0 (figure-62). The sensor data shows, it reached the maximum threshold value (1020 lux) for 84 times. In the winter time the lux value were almost half compare to the summer times.

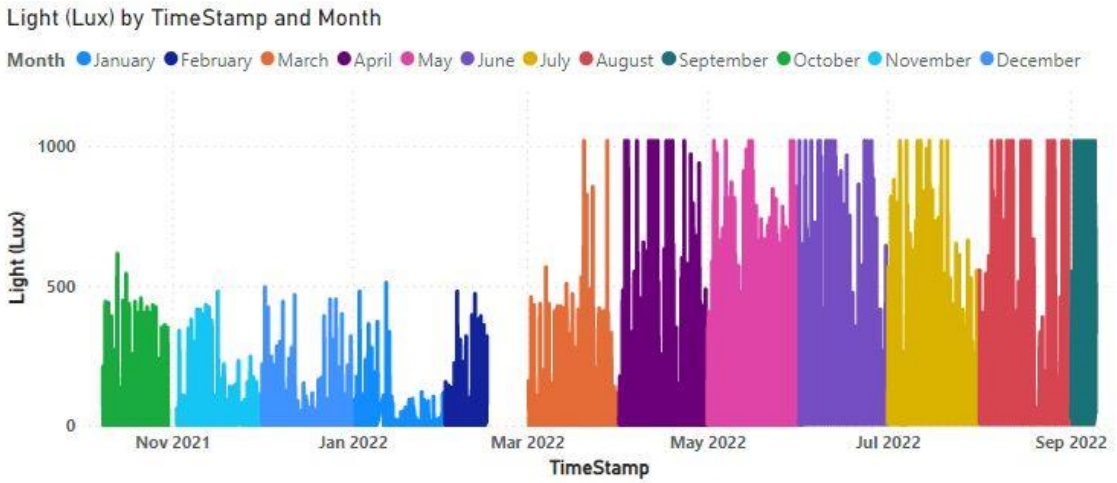


Figure 62: Brightness (Lux) at Room-213

The sensor has no outliers and the dataset is well distributed. Likewise, the sensor has the missing values and the unknown values like other sensors.

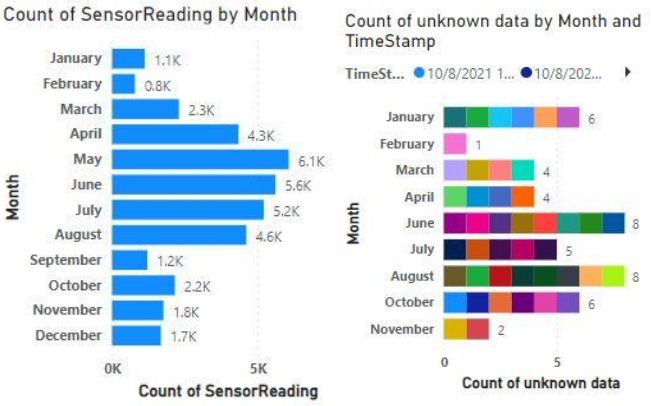


Figure 63: Reading and the unknown data at Room-213

3.4 Comparison of Brightness sensor data

As the brightness sensor reading is directly related to the sunlight the number of recordings of this sensor is huge compare to the other sensor installed in each room. To give a clear explanation, two different rooms are selected among these two rooms, one is mostly exposed to the sunlight and another is totally out of direct sunlight. The data recording in brightness sensor in room 204 and 213 has shown a noticeable difference (figure-64 &65). The room 204 is exposed to sun throughout the year. As a result of that the data recording in the sensor has shown higher lux value also reached the maximum threshold even in the winter time.

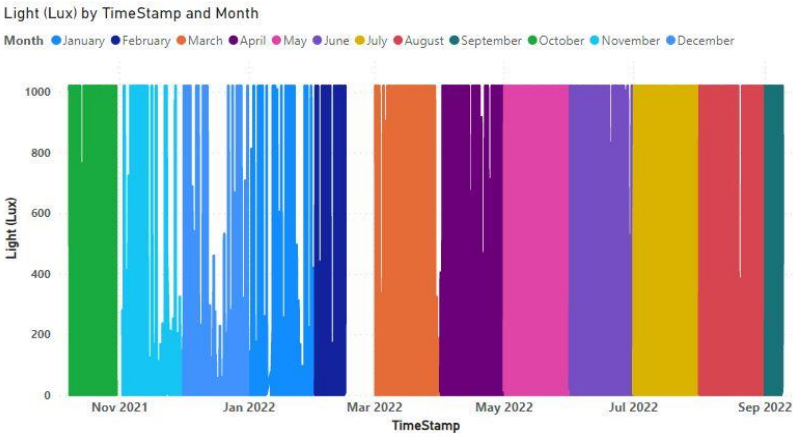


Figure 64: Brightness (room-204)

On the other hand, the brightness sensor placed in the room 213 has shown the difference in data recording (figure-65). In the winter season the reading mostly stayed below 500 lux. Even in the summer season the lux value is not as much as compare to the room 204.

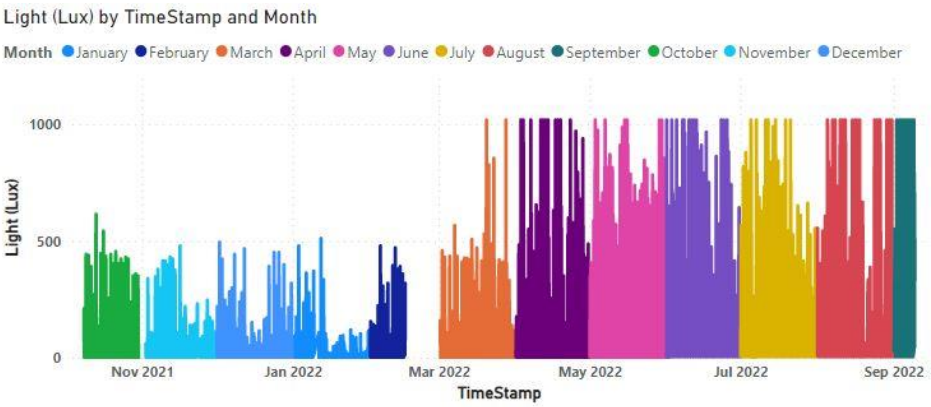


Figure 65: Brightness (room-213)

After anylzing the sun direction in response to the institute building for whole year, it can be obseved that the room 204 (left) exposed to the sun for all the time which produces higher lux values. On the contrary the room 213 (right) has no option to get the direct sunlight (figure-66). (Here the photo taken at 12PM on June 3rd for both room which is a random selection to visualize the situation, other date time can be taken to verify the above mention statement).

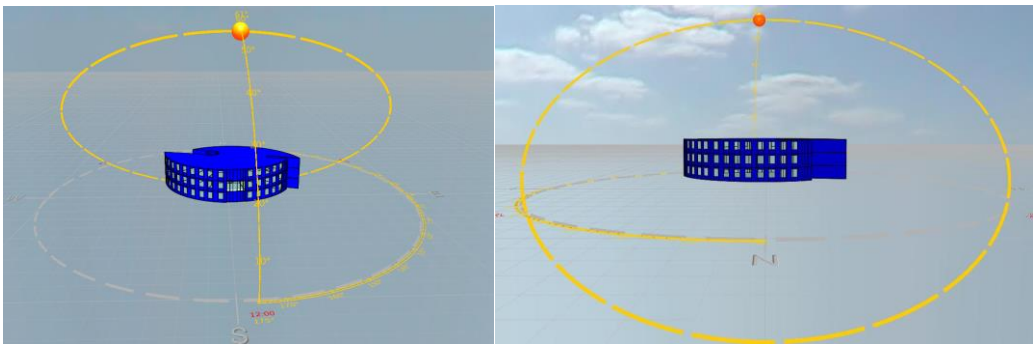


Figure 66: Sun path in response to two different rooms

3.4.1 Placement of brightness sensor

Sensor placement has large impact on produced data, especially in brightness sensor. Here a comparison of sensor data with the placement has explained. In room 201, sensor has placed two different place and data is being observed. The recorded data for placement (left) 280 lux and the reading for second placement (right) recorded 156 lux (figure-67). A small change has given a huge different in data reading. Even the placement of the sensor in some place can also cause the reading as 0 due to the absence of light.

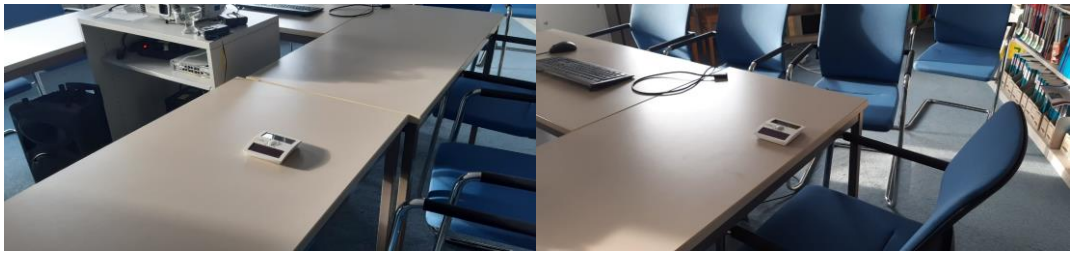


Figure 67: Sensor placement

3.4.2 Correct way to place brightness sensor

The sensor records the value of lux. Lux indicate how much light fall in a particular surface. By placing the sensor in the table light (lumen) is directly falling to the sensor and the sensor is producing the data which is in correct. The sensor should be placed in the ceiling or wall which will record the lux value for the light. Here, the problem with battery and charging of battery of sensor in institute building. The sensor has battery which is charged by the sun-light. Placing the sensor on the ceiling or wall will hinder the charging system. There are two possible solutions for this. (1) Changing the sensor with a sensor which can be connected with electrical wire or fully depend on the battery. (2) Placing the available sensor in 2-3 different place and then fusion of multi-sensor a single reading can be recorded for that specific room.

3.5 Possible reason for sensor data inconsistency

Shutdown of the system: A complete shutdown of the sensor system for certain days for maintenance produce the incomplete or missing data which is considered as Inconsistency in dataset. In the given dataset has missing data which is due to the complete shutdown of the system.

System rebooting: System rebooting also a reason for data inconsistency. When system goes for reboot, the sensor does not record the data and produce “unknown” instead a numerical value.

Battery problems: Battery failure or depletion is another reason for power outages. When a battery is not producing enough power, there are two potential causes: (1) insufficient battery charge, and (2) hardware failure of the battery.

Connection failure: Connection failure or pause also caused the inconsistency in sensor reading.

External attack: Security threat (external attack on sensor server) could be the reason of data anomalies.

Placement of sensor: Placement of sensor plays a very important role of sensor reading. In a room, if sensor placed close to the window or to the heater which can produce some abnormal reading due to closeness of heat source.

4 Solution for data inconsistency

4.1 Point outlier and Unknowns

Point outlier is found in the temperature sensor and humidity sensor in room 213. The unknown reading is found in every sensor.

4.1.1 Deletions

If an outlier value results from a data input error, a data processing error, or a small number of outlier observations, we delete the observation. To get rid of outliers, we can also trim at both ends. The unknowns are available in all 20 sensors and the number of unknown distributions in all months are the same. The following table consists of timestamp and the sensor reading of the temperature sensor in room 201. In sensor reading, some reading are showing as 'unknown' that is considered as outlier.

SensorReference	TimeStamp	SensorReading
sensor.sr04_01_temp	10/8/21 12:14 PM	unknown
sensor.sr04_01_temp	10/8/21 12:29 PM	unknown
sensor.sr04_01_temp	10/8/21 12:45 PM	25.2
sensor.sr04_01_temp	10/8/21 1:17 PM	unknown
sensor.sr04_01_temp	10/8/21 1:19 PM	25.4
sensor.sr04_01_temp	10/8/21 1:19 PM	unknown
sensor.sr04_01_temp	10/8/21 1:35 PM	25.8
sensor.sr04_01_temp	10/8/21 1:47 PM	27
sensor.sr04_01_temp	10/8/21 1:52 PM	27.4
sensor.sr04_01_temp	10/8/21 2:09 PM	28
sensor.sr04_01_temp	10/8/21 2:25 PM	28.2
sensor.sr04_01_temp	10/8/21 2:42 PM	28.4

Table 3: Original Dataset (sample)

Above mentioned unknown reading can be deleted from the dataset as the dataset contain with large number of sensors reading. The following table is showing the dataset after the deletion of outlier.

SensorReference	TimeStamp	SensorReading
sensor.sr04_01_temp	10/8/21 12:45 PM	25.2
sensor.sr04_01_temp	10/8/21 1:19 PM	25.4
sensor.sr04_01_temp	10/8/21 1:35 PM	25.8
sensor.sr04_01_temp	10/8/21 1:47 PM	27
sensor.sr04_01_temp	10/8/21 1:52 PM	27.4
sensor.sr04_01_temp	10/8/21 2:09 PM	28
sensor.sr04_01_temp	10/8/21 2:25 PM	28.2
sensor.sr04_01_temp	10/8/21 2:42 PM	28.4

Table 4: Dataset after deletions

Outlier is observed in the temperature and humidity sensor in room 213. The temperature sensor has given a reading of 51°C which is an outlier (table-5). The previous reading of

51°C was 25.2°C which was just 17 minutes earlier recorded. And after the reading of 51°C it was 26 °C that one also took 17 minutes. So, this sudden increase obviously an outlier.

SensorReference	TimeStamp	SensorReading
sensor.sr04_05_temp	7/5/22 8:53 AM	25.2
sensor.sr04_05_temp	7/5/22 9:10 AM	24.8
sensor.sr04_05_temp	7/5/22 10:00 AM	24.6
sensor.sr04_05_temp	7/5/22 10:16 AM	25.2
sensor.sr04_05_temp	7/5/22 10:33 AM	51
sensor.sr04_05_temp	7/5/22 10:50 AM	26
sensor.sr04_05_temp	7/5/22 11:33 AM	26.4
sensor.sr04_05_temp	7/5/22 11:56 AM	26.6
sensor.sr04_05_temp	7/5/22 12:46 PM	26.4

Table 5: Outlier in temperature sensor

The sensor recorded a single abnormal value and the total reading for this sensor is 5032. As a solution the single outlier data can be deleted from the dataset. The following table is showing the dataset after the deletion of point outlier.

SensorReference	TimeStamp	SensorReading
sensor.sr04_05_temp	7/5/2022 8:53	25.2
sensor.sr04_05_temp	7/5/2022 9:10	24.8
sensor.sr04_05_temp	7/5/2022 10:00	24.6
sensor.sr04_05_temp	7/5/2022 10:16	25.2
sensor.sr04_05_temp	7/5/2022 10:50	26
sensor.sr04_05_temp	7/5/2022 11:33	26.4
sensor.sr04_05_temp	7/5/2022 11:56	26.6
sensor.sr04_05_temp	7/5/2022 12:46	26.4

Table 6: Dataset after outlier deletion

The humidity sensor also recorded a reading of humidity 84% though the outside humidity was 74.8% on that time. In addition to that the reading before 84% was 44% and after the reading it was 43.5% (table-7). So, it is also an outlier in the dataset.

SensorReference	TimeStamp	SensorReading
sensor.sr04_05_hum	10/8/21 8:38 PM	44.5
sensor.sr04_05_hum	10/8/21 10:01 PM	45
sensor.sr04_05_hum	10/8/21 10:18 PM	44.5
sensor.sr04_05_hum	10/9/21 2:28 AM	44
sensor.sr04_05_hum	10/9/21 4:58 AM	84
sensor.sr04_05_hum	10/9/21 5:15 AM	43.5
sensor.sr04_05_hum	10/9/21 7:45 AM	43
sensor.sr04_05_hum	10/9/21 9:08 AM	42.5
sensor.sr04_05_hum	10/9/21 9:25 AM	42

Table 7: Outlier in humidity sensor

The above-mentioned outlier can simply be deleted like the temperature dataset. The following table is showing the dataset after the deletion of point outlier.

SensorReference	TimeStamp	SensorReading
sensor.sr04_05_hum	10/8/21 8:38 PM	44.5
sensor.sr04_05_hum	10/8/21 10:01 PM	45
sensor.sr04_05_hum	10/8/21 10:18 PM	44.5
sensor.sr04_05_hum	10/9/21 2:28 AM	44
sensor.sr04_05_hum	10/9/21 5:15 AM	43.5
sensor.sr04_05_hum	10/9/21 7:45 AM	43
sensor.sr04_05_hum	10/9/21 9:08 AM	42.5
sensor.sr04_05_hum	10/9/21 9:25 AM	42

Table 8: Dataset after outlier deletion

4.1.2 Imputation

We can also impute outliers, just like we do with missing values. The mean, median, and mode imputation methods are available. We should determine whether the outlier is artificial or natural before imputing values. We can use imputing values if it is artificial. We can also anticipate the values of an outlier observation using a statistical model, and then we can impute the observed values to the predicted values.

For point outlier and unknown reading mean imputation is also a solution. In the dataset, by calculating the mean value of the available temperature reading, it is replaced with outliers. The following table showing the that, the temperature reading 51°C is replaced by the mean value (25.65°C) of the dataset.

SensorReference	TimeStamp	SensorReading
sensor.sr04_05_temp	7/5/22 8:53 AM	25.2
sensor.sr04_05_temp	7/5/22 9:10 AM	24.8
sensor.sr04_05_temp	7/5/22 10:00 AM	24.6
sensor.sr04_05_temp	7/5/22 10:16 AM	25.2
sensor.sr04_05_temp	7/5/22 10:33 AM	25.65
sensor.sr04_05_temp	7/5/22 10:50 AM	26
sensor.sr04_05_temp	7/5/22 11:33 AM	26.4
sensor.sr04_05_temp	7/5/22 11:56 AM	26.6
sensor.sr04_05_temp	7/5/22 12:46 PM	26.4

Table 9: Temperature dataset after mean imputation

The method can be applied for all unknowns and also for the outlier in humidity sensor as a solution.

4.2 Missing value

When the missing data is very few, a simple imputation (mean, median, mode) can be use. But in the institute dataset, the data is missing for a long time. As a result, other imputation technique is considered in the study.

Limitations

-There is no previous year dataset to compare and find the relations with current dataset.

-The timestamp is also missing due to complete shutdown of the system.

-Weather data: The relation in between outdoor temperature and indoor temperature is not linear. When the outside is warm compare to indoor, there is a strong relation in between outdoor and indoor temperature but cooler temperature in outdoor has a weaker correlation [55]. The humidity has strong relation with the temperature; therefore, weather data cannot be used in indoor data prediction. The CO2 value completely depend on the occupancy, ventilation etc. and the brightness depend on the orientation of the building, artificial light in the room. So outdoor data cannot be used in imputation of indoor data.

Due to above-mentioned limitations, forecasting is an option for the solution. Here three approximate methods have applied for the data prediction. Three methods (Linear regression, exponential smoothing and ARIMA model) a gave us the approximate value for the missing fifteen days (15th February to 1st March in 2022) and two days (30th October to 2nd November in 2021).

4.2.1 Linear regressing model

Linear regression is an attempt to model the relationship between two variables by fitting a linear equation to observed data, where one variable is considered to be an explanatory variable and the other as a dependent variable[56].

Why linear regression for prediction?

-Linearity: The linear regression model forces the prediction to be a linear combination of features.
-Normality: It is assumed that the target outcome given the features follows a normal distribution.
-Independence: It is assumed that each instance is independent of any other instance.
-Fixed features: The input features are considered "fixed". Fixed means that they are treated as "given constants" and not as statistical variables.

y=m*x+c

Here,

- **y** is the predicted value of the dependent variable **(y)** for any given value of the independent variable **(x).**
- **c** is the intercept, the predicted value of **y** when the **X** is **0.**
- **m** is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).

m (Slope)

m = r*(Sx/Sy)

Here, r is Pearson Correlation coefficient

Sx is Standard deviation of X samples

Sy is standard deviation of Y samples

c (Y-intercept)

$$\mathbf{c} = \bar{y} - m * \bar{x}$$

Here, $\bar{y}$ is the mean of the y samples

$\bar{x}$ is the mean of the x samples

Correlation coefficient (r) is used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

A standard deviation (or s) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Here a sample dataset has been used to illustrate Linear Regression model. The dataset contains with dependent variable (y) and independent variable (x). Then a scatter chart has drowned. In scatter chart linear regressing line also added, through the added line an unknown value of dependent variable can be calculated. For the calculation of the dependent variable with the help of independent variable, a detailed procedure is given in Table 9 and Table 10.

X (Independent variable)	Y (Dependend variable)
1	4
2	12
3	8
4	12
5	16
6	14
7	10
8	15
9	22
10	24

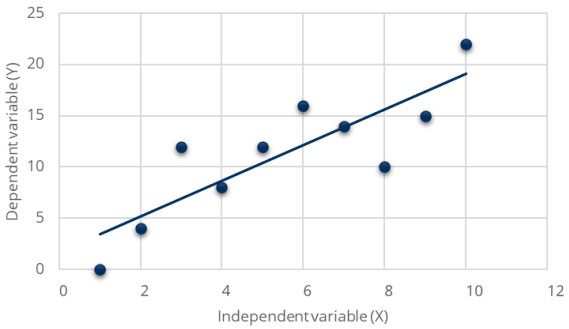


Table 10: Sample dataset and scatter distribution

N	X	Y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	1	4	-4.5	-9.7	43.65	20.25	94.09
2	2	12	-3.5	-1.7	5.95	12.25	2.89
3	3	8	-2.5	-5.7	14.25	6.25	32.49
4	4	12	-1.5	-1.7	2.55	2.25	2.89
5	5	16	-0.5	2.3	-1.15	0.25	5.29
6	6	14	0.5	0.3	0.15	0.25	0.09
7	7	10	1.5	-3.7	-5.55	2.25	13.69
8	8	15	2.5	1.3	3.25	6.25	1.69
9	9	22	3.5	8.3	29.05	12.25	68.89
10	10	24	4.5	10.3	46.35	20.25	106.09
	5.5	13.7			138.5	82.5	328.1

Table 11: Procedure for Linear regression (1)

$r=\frac{(x-\bar{x})*(y-\bar{y})}{\sqrt{((x-\bar{x})^2*(y-\bar{y})^2}}$	$Sy=\sqrt{\frac{\sum(x-\bar{x})^2}{N-1}}$	$Sy=\sqrt{\frac{\sum(y-\bar{y})^2}{N-1}}$	m = r*(Sx/Sy)	c= $\bar{y}$-m* $\bar{x}$	y= m*x+c
0.841820861	6.037843618	3.027650354	1.678787879	4.46666667	6.145455

Table 12: Procedure for Linear regression (2)

In the institute dataset, the data is missing from 15th February to 1st March in 2022. With the help of dataset from 1st February to 14th February from temperature sensor of room 20, a regression line has drowned by extending the regression line with assumed time series (independent variable), the next 15 days sensor reading has approximated (figure-69). The increase and decrease of datapoints in the following figure showing the actual reading from the sensor and the straight line is indicating the prediction for missing data. The same method applied for the 20 sensor and the approximation has done for missing data.

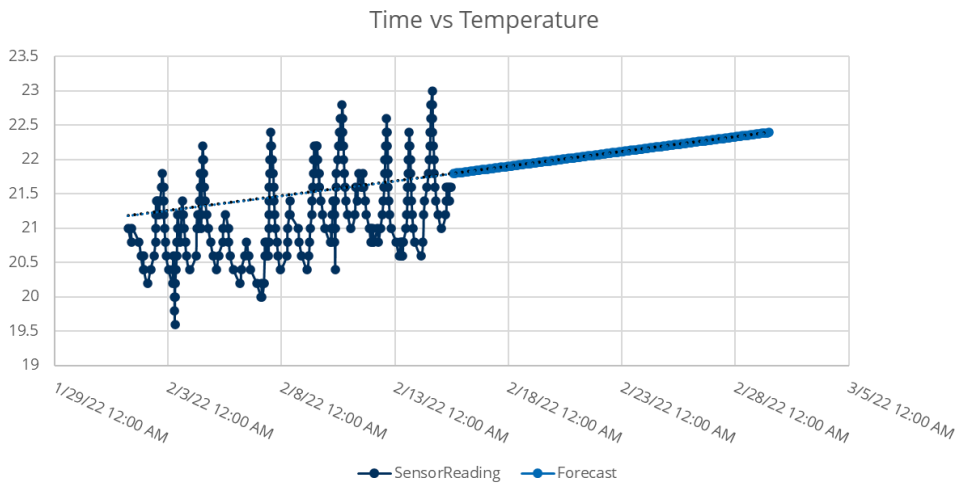


Figure 68: Linear regression for prediction (1)

The second dataset is missing from 30th October to 2nd November in 2021. To predict the missing data, five days data (26th October to 30th October) is taken from the temperature sensor of room 202. In this forecasting, one hour interval is considered in timestamp. By applying the linear regression in the dataset, the missing dataset is predicted (figure-70).

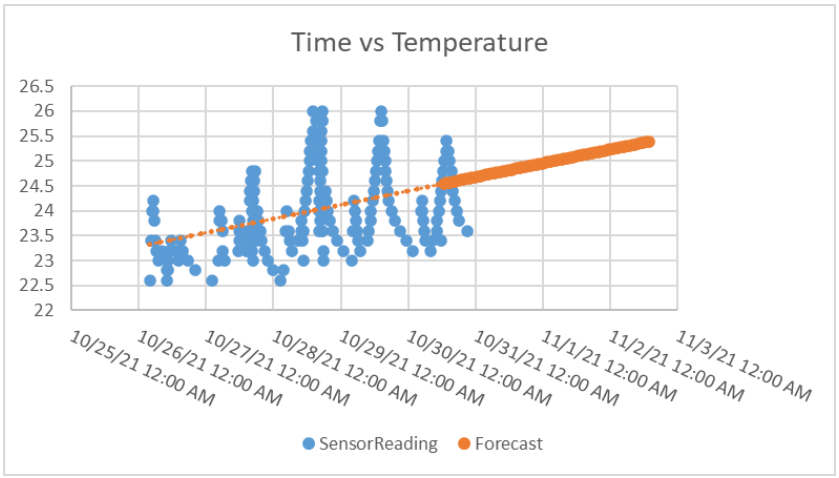


Figure 69: Linear regression for prediction (2)

Above mentioned method has limitation, the regression line showing the upward trend. But in practical distribution, there should be ups and downs with the time. To solve the problem a new approach of forecasting with exponential smoothing has implemented.

4.2.2 Forcasing by Exponential Smoothing

The algorithm used behind the forecasting is exponential smoothing. "Exponential smoothing is a rule of thumb technique for smoothing time series data using the exponential window function. Whereas in the simple moving average the past observations are weighted equally, exponential functions are used to assign exponentially decreasing weights over time"[57].

The simple formula for exponential smoothing[58].

$$F_{t+1} = F_t + \alpha(A_t - F_t)$$

Here, **Ft+1** is the forecast value for the time **t+1**

At is the actual value at time **t**

And **α** is the smoothing constant

In exponential smoothing there are two different algorithm (1) seasonal data (ETS AAA) and (2) non seasonal data (ETS AAN) also introduced in the method. 'The seasonal algorithm (ETS AAA) models the time series using an equation that accounts for additive error, additive trend, and additive seasonality. This algorithm is also popularly known as the Holt-Winters algorithm, after the researchers who described the characteristics of the model. The non-seasonal algorithm (ETS AAN) uses a simpler equation to model the time series, which includes only a term for additive trend and additive error, and does not consider seasonality at all. We assume data values increase or decrease in some way that can be described by a formula, but that the increase or decrease is not cyclical" (Microsoft).

Application in institute dataset:

The data recorded in the institute has not regular interval. The system uses discrete method to record the data. It means when a change is occurred in the environmental components the sensor records the data. As the sensor is powered by the battery and also with the direct connection of electricity, it is necessary to consider the power requirement and consumptions. Therefore, the discrete time series method is taken into account. The regular time interval in the system required more power supply compare to irregular time series as it records more data. In addition to that the regular time series has some issues; like, if the interval of the reading is large, there is a possibility of missing important reading. And if the interval of regular time series is small, the total number of readings will be higher compare to discrete time series. As a result, the discrete method has been used in the institute to record the sensor data.

TimeStamp	SensorReading		TimeStamp	SensorReading
2/1/22 12:37 AM	20.4		2/1/22 12:37 AM	20.4
2/1/22 1:10 AM	20.6			
2/1/22 1:27 AM	20.8		2/1/22 1:27 AM	20.6
			2/1/22 2:17 AM	20.8
2/1/22 2:34 AM	20.6			
2/1/22 3:07 AM	20.4	→	2/1/22 3:08 AM	20.4
2/1/22 3:57 AM	20.6		2/1/22 3:58 AM	20.6
2/1/22 4:14 AM	20.8			
2/1/22 4:47 AM	21		2/1/22 4:48 AM	21
2/1/22 5:04 AM	20.8			
2/1/22 5:20 AM	21			
2/1/22 5:37 AM	21.2		2/1/22 5:39 AM	21.2

Table 13: Original (irregular) dataset (left), regular dataset (right)

To use the exponential smoothing method, it is required the time series with equal interval between two readings. Therefore, the data from institute has changed from the irregular dataset to regular dataset (table-13).

The dataset is about 15 days (15th of February to 1st March in 2022) from the temperature sensor of room 202. The total reading is divided by 15 days. The average time interval is about 50 minutes therefore in the regular time series the interval between two readings is 50 minutes. The following figure is showing the actual data by blue line and the predicted value by red line. In forecasting section, the figure is also showing the lower boundary value and upper boundary value for each datapoints.

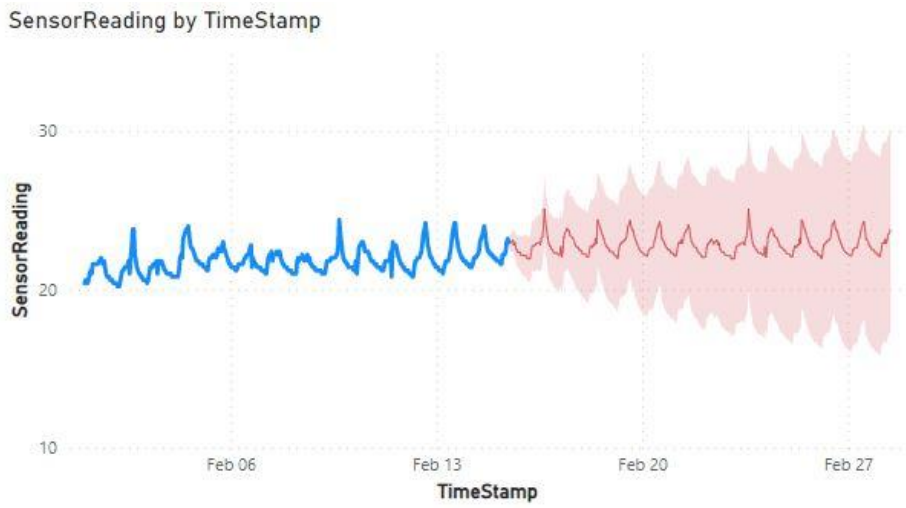


Figure 70: Forecasting missing value (1)

By exponential smoothing the forecasted value is generated from the previous value. Moreover, it also produces the maximum and minimum reading for each forecasted data (table-14).

TimeStamp	Forecast	Higher Boundary	Lower Boundary
2/15/2022 11:54	22.87	23.3	22.43
2/15/2022 12:44	22.99	23.53	22.44
2/15/2022 13:34	23.01	23.65	22.38
2/15/2022 14:24	23.12	23.84	22.4
2/15/2022 15:14	22.78	23.57	22
2/15/2022 16:04	22.98	23.83	22.13
2/15/2022 16:54	22.58	23.49	21.66
2/15/2022 17:44	22.5	23.46	21.53
2/15/2022 18:34	22.33	23.35	21.32
2/15/2022 19:24	22.35	23.42	21.29
2/15/2022 20:14	22.36	23.47	21.25
2/15/2022 21:04	22.16	23.32	21.01
2/15/2022 21:54	22.16	23.36	20.96
2/15/2022 22:44	22.16	23.4	20.92
2/15/2022 23:34	22.12	23.4	20.84

Table 14: Dataset after forecasting (1)

The similar method also applied for the missing dataset on October 30th to November 2nd in 2021. In this case five days temperature data is taken (26th October to 30th October) from temperature sensor in room-202. By converting the irregular time series to regular time series, the forecasting is done (figure-71).

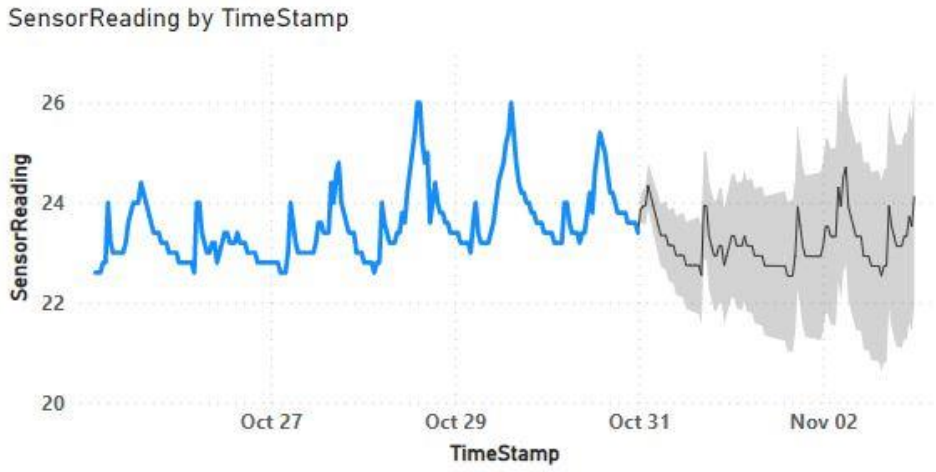


Figure 71: Forecasting missing value (2)

The following table is showing the forecasted value along with the Higher boundary and Lower boundary of each prediction.

TimeStamp	Forecast	Higher Boundary	Lower Boundary
10/31/21 12:25 AM	23.86	24.13	23.59
10/31/21 1:05 AM	23.92	24.27	23.58
10/31/21 1:44 AM	23.94	24.34	23.54
10/31/21 2:24 AM	24.34	24.80	23.88
10/31/21 3:04 AM	24.14	24.64	23.64
10/31/21 3:43 AM	23.94	24.48	23.41
10/31/21 4:23 AM	23.75	24.31	23.18
10/31/21 5:03 AM	23.55	24.15	22.95
10/31/21 5:42 AM	23.35	23.97	22.72
10/31/21 6:22 AM	23.35	24.01	22.69

Table 15: Dataset after forecasting (2)

4.2.3 ARIMA Model

The ARIMA (Auto Regressive Integrated Mean Average) methodology is a statistical method for analyzing and building a forecasting model which best represents a time series by modeling the correlations in the data. ARIMA models only require past data from a time series, they can generalize forecasts and boost prediction accuracy while still maintaining a compact model [59]. The study [60] explained two major terms of ARIMA model in following: (1) The data are differenced in an autoregressive integrated moving average model to make it stationary. A model that demonstrates stationarity demonstrates that the data remain constant across time. The goal of differencing is to eliminate any patterns or seasonal structures that are present because most economic and market data exhibit trends. (2) Seasonality, or when data exhibit recurring, predictable trends over the course of a year, may have a negative impact on the regression model. Many of the calculations throughout the process cannot be performed with great efficiency if a trend develops and stationarity is not obvious.

According to [61] an ARIMA model is characterized by 3 terms: p, d, q. Where **p** is the order of the Auto Regressive term, **q** is the order of the Moving Average term and **d** is the number of differencing required to make the time series stationary.

Steps for the model:

1. Different Python library is called for data reading, and plotting the data for visualization.
2. Then adfuller (ADF) test is being executed for checking the stationarity of the dataset. In the study case the p value for the dataset is .000172 which indicates the dataset is stationary.
3. By using the ARIMA model of python, the p, d, q values are calculated. In this case, the p, d, q values are (3,1,2) considered as the best model for the dataset.
4. The dataset contains 280 sensor readings. Among these last 100 points are taken as the training dataset. Then the dataset is compared with the predicted data. Here the difference between the actual and predicted data is noticeable.
5. Then root mean squared error is calculated for the model. The RMSE value in this case is 0.57.
6. Finally, the time interval is introduced in the model and future prediction is done for the given dataset. Here the model predicted few values from 15th February which has a variation the reading. Then it produced the similar reading for the future.

Python code for the model:

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df = pd.read_csv('abc4.csv', index_col = 'TimeStamp',parse_dates=True)
df =df.dropna()
print('Shape of the data', df.shape)
df.head(10)
```

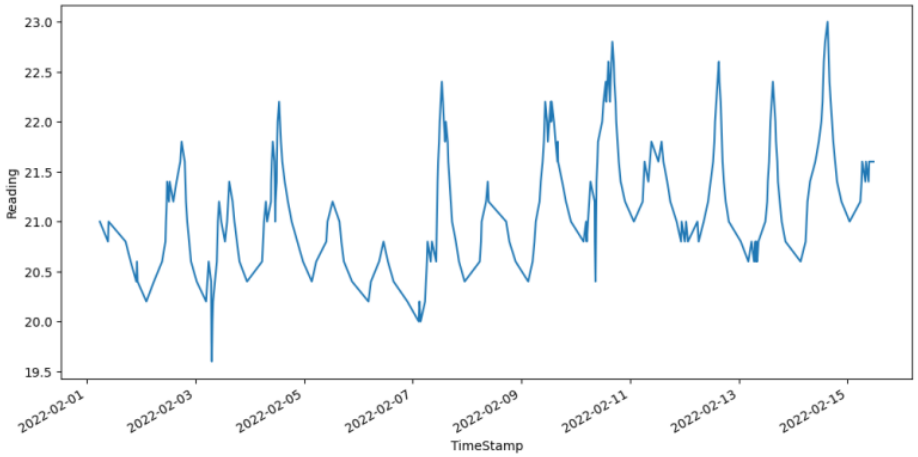
```
Shape of the data (280, 1)
```

Out[2]:

TimeStamp	Reading
2022-02-01 05:50:00	21.0
2022-02-01 09:27:00	20.8
2022-02-01 09:43:00	21.0
2022-02-01 17:13:00	20.8
2022-02-01 19:27:00	20.6
2022-02-01 21:57:00	20.4
2022-02-01 22:13:00	20.6
2022-02-01 22:30:00	20.4
2022-02-02 02:23:00	20.2
2022-02-02 05:43:00	20.4

In [3]:
```
df['Reading'].plot(figsize=(12,6))
plt.xlabel('TimeStamp')
plt.ylabel('Reading')
```

Out[3]:
```
Text(0, 0.5, 'Reading')
```



In [4]:
```
from statsmodels.tsa.stattools import adfuller

def ad_test(dataset):
     dftest = adfuller(dataset, autolag = 'AIC')
     print("1. ADF : ",dftest[0])
     print("2. P-Value : ", dftest[1])
     print("3. Num Of Lags : ", dftest[2])
     print("4. Num Of Observations Used For ADF Regression:", dftest[3])
     print("5. Critical Values :")
     for key, val in dftest[4].items():
            print("\t",key, ": ", val)
```

In [5]:
```
ad_test(df['Reading'])
```

```
1. ADF :  -4.5320007820551576
2. P-Value :  0.00017205280275996005
3. Num Of Lags :  3
4. Num Of Observations Used For ADF Regression: 276
5. Critical Values :
         1% :  -3.4542672521624214
         5% :  -2.87206958769775
         10% :  -2.5723807881747534
```

In [6]:
```
from pmdarima import auto_arima
import warnings
warnings.filterwarnings("ignore")
```

In [7]:
```
stepwise_fit = auto_arima(df['Reading'], trace=True,
                          suppress_warnings=True)
stepwise_fit.summary()
```

```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=-118.384, Time=0.56 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=-70.245, Time=0.13 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=-108.678, Time=0.14 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=-95.496, Time=0.20 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=-72.216, Time=0.10 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=-115.782, Time=0.34 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=-120.349, Time=0.59 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=-110.555, Time=0.17 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=-113.945, Time=0.17 sec
 ARIMA(3,1,1)(0,0,0)[0] intercept   : AIC=-118.395, Time=0.59 sec
 ARIMA(3,1,0)(0,0,0)[0] intercept   : AIC=-118.326, Time=0.27 sec
 ARIMA(3,1,2)(0,0,0)[0] intercept   : AIC=-129.313, Time=1.26 sec
 ARIMA(4,1,2)(0,0,0)[0] intercept   : AIC=inf, Time=1.35 sec
 ARIMA(3,1,3)(0,0,0)[0] intercept   : AIC=-134.726, Time=1.35 sec
 ARIMA(2,1,3)(0,0,0)[0] intercept   : AIC=-116.723, Time=0.78 sec
 ARIMA(4,1,3)(0,0,0)[0] intercept   : AIC=-128.108, Time=1.54 sec
 ARIMA(3,1,4)(0,0,0)[0] intercept   : AIC=-125.601, Time=1.75 sec
 ARIMA(2,1,4)(0,0,0)[0] intercept   : AIC=-115.449, Time=1.53 sec
 ARIMA(4,1,4)(0,0,0)[0] intercept   : AIC=-124.364, Time=1.87 sec
 ARIMA(3,1,3)(0,0,0)[0]             : AIC=-137.764, Time=0.90 sec
 ARIMA(2,1,3)(0,0,0)[0]             : AIC=-118.711, Time=0.44 sec
 ARIMA(3,1,2)(0,0,0)[0]             : AIC=-137.826, Time=0.84 sec
 ARIMA(2,1,2)(0,0,0)[0]             : AIC=-120.372, Time=0.27 sec
 ARIMA(3,1,1)(0,0,0)[0]             : AIC=-120.383, Time=0.27 sec
 ARIMA(4,1,2)(0,0,0)[0]             : AIC=-135.146, Time=0.98 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=-122.337, Time=0.33 sec
 ARIMA(4,1,1)(0,0,0)[0]             : AIC=-118.917, Time=0.37 sec
 ARIMA(4,1,3)(0,0,0)[0]             : AIC=-134.159, Time=1.08 sec

Best model:  ARIMA(3,1,2)(0,0,0)[0]
Total fit time: 20.191 seconds
```

In [8]:
```
from statsmodels.tsa.arima.model import ARIMA
```

In [22]:
```
print(df.shape)
train=df.iloc[:-100]
test=df.iloc[-100:]
print(train.shape,test.shape)
```

```
(280, 1)
(180, 1) (100, 1)
```

In [23]:
```
model=ARIMA(train['Reading'],order=(3,1,2))
model=model.fit()
#model.summary()
```

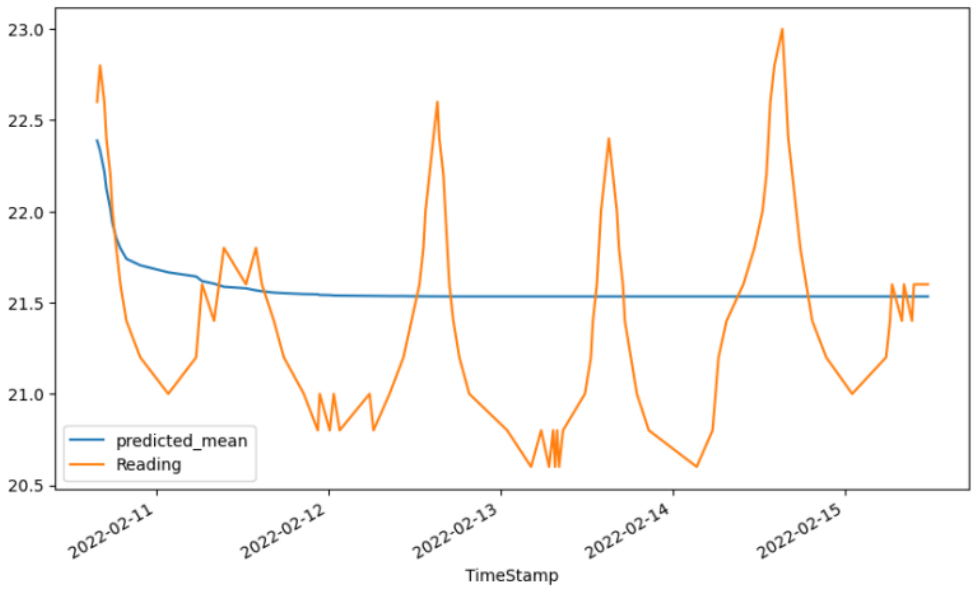
In [24]:
```
start=len(train)
end=len(train)+len(test)-1
pred=model.predict(start=start,end=end,typ='levels')
```

```
pred.index = df.index[start:end+1]
pred.head(10)
```

Out[24]:
```
TimeStamp
2022-02-10 15:46:00    22.388324
2022-02-10 16:11:00    22.333553
2022-02-10 16:47:00    22.217529
2022-02-10 17:04:00    22.124005
2022-02-10 17:37:00    22.013439
2022-02-10 17:54:00    21.935688
2022-02-10 18:27:00    21.852534
2022-02-10 19:01:00    21.797900
2022-02-10 19:51:00    21.740239
2022-02-10 21:47:00    21.704420
Name: predicted_mean, dtype: float64
```

In [25]:
```
pred.plot(legend=True)
test['Reading'].plot(figsize=(10,6),legend=True)
```

Out[25]:
```
<AxesSubplot: xlabel='TimeStamp'>
```



In [26]:
```
test['Reading'].mean()
```

Out[26]:
```
21.573999999999998
```

In [27]:
```
from sklearn.metrics import mean_squared_error
from math import sqrt
test['Reading'].mean()
rmse=sqrt(mean_squared_error(pred,test['Reading']))
print(rmse)
```

```
0.5716680442281084
```

In [28]:
```
model2=ARIMA(df['Reading'],order=(3,1,2))
model2=model2.fit()
```

```
df.tail(5)
```

Out[28]:

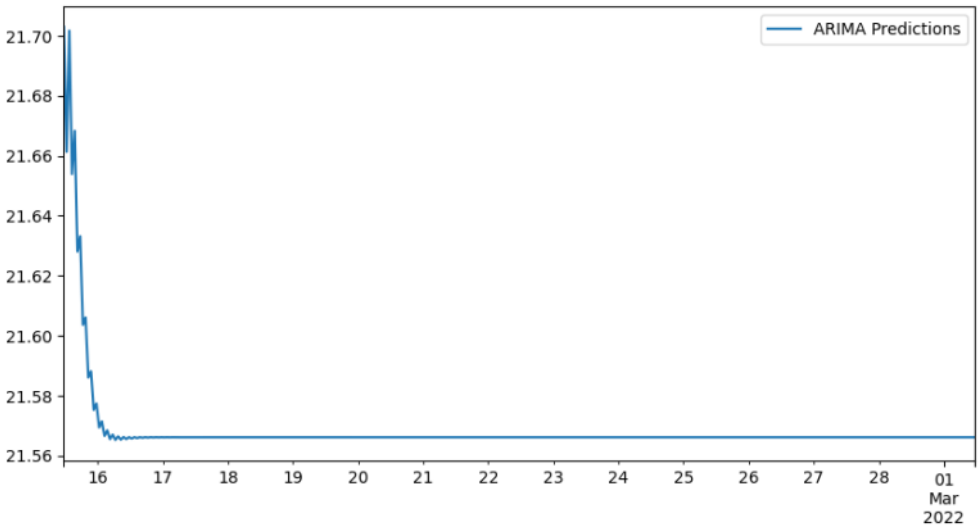
TimeStamp	Reading
2022-02-15 07:54:00	21.4
2022-02-15 08:11:00	21.6
2022-02-15 09:18:00	21.4
2022-02-15 09:34:00	21.6
2022-02-15 11:31:00	21.6

In [31]:
```
index_future_dates = pd.date_range(start='2022-02-15 11:31:00', end='2022-03-01 11:31:
pred=model2.predict(start=len(df),end=len(df)+336,type='levels').rename('ARIMA Predict
pred.index=index_future_dates
pred.head(10)
```

Out[31]:
```
2022-02-15 11:31:00    21.702953
2022-02-15 12:31:00    21.661304
2022-02-15 13:31:00    21.701685
2022-02-15 14:31:00    21.653904
2022-02-15 15:31:00    21.668300
2022-02-15 16:31:00    21.628077
2022-02-15 17:31:00    21.633151
2022-02-15 18:31:00    21.603619
2022-02-15 19:31:00    21.606082
2022-02-15 20:31:00    21.586074
Freq: 60T, Name: ARIMA Predictions, dtype: float64
```

In [33]:
```
pred.plot(figsize=(10,5),legend=True)
```

Out[33]:
```
<AxesSubplot: >
```



5 Conclusion and Future Work

5.1 Conclusion and discussion

The study investigated the dataset from the institute (Nürnberger-Ei) by sensor and room perspective. Each dataset of a sensor has analyzed individually and inconsistency checking is done by visualization with comparison of weather station data. In addition to that, the comparison of data with the standard regulation for temperature, humidity, CO2 and brightness is also done in this study. The research has found some abnormal situation like unknowns and missing data for all twenty different sensors. The outlier is found in the Humidity and Temperature sensor in room 213. Furthermore, some sensor reading reached the upper threshold of a sensor. The CO2 sensor and the brightness sensor of the room 201 and 202 reached the upper threshold 2550 ppm and 1020 lux respectively. The brightness sensor for room 204, 210 and 213 also reached the upper threshold of the sensor reading. According to the humidity sensor data, the indoor humidity for all five different rooms is below the required comfort range. Therefore, the installation of humidifier is suggested to ensure healthy indoor environment. According the "German Committee on Indoor Air Guide Values" the CO2 concentration must be below 2000 ppm in indoor environment but the CO2 sensor reading for the room 201 and 202 crossed the hygiene limit. As a consequence, the existing ventilation system (windows) need to be more functional or additional ventilation can be implemented. In brief, the dataset for the institute has very few unknowns and two noticeable outliers which is solved in this research. One major problem was missing value, which has also been discussed, and a few methods are being used to solve the issue. As a result, the method can be applied for similar data problems in other buildings.

5.2 Future Work

The missing value imputation is done by the method linear regression which has not fulfill our expectations. Then, a forecasting tool has been used to predict the future value. It worked apparently better compare to other two methods, but the procedure was manual. Machine learning approach also taken into account by ARIMA model. This algorithm also worked well but the existing data has lot of variations. Therefore, the algorithm prediction has given the same value after few steps. A better model can also be searched and implemented for the future data prediction. As Power-Bi approach (exponential smoothing) gave the more appropriate solution, this approach can be applied with machine learning with others considerations and make a fully automated system to execute the prediction. For outlier and unknown detection and imputation machine learning or AI can be implemented. Alarm system or the notification system can also be included when data reading shows some abnormal situation.

6 References

[1] S. G. Navada, C. S. Adiga, and S. G. Kini, "A Study on Daylight Integration with Thermal Comfort for Energy Conservation in a General Office", doi: 10.12720/ijoee.1.1.18-22.

[2] D. Wyon, "The effects of indoor air quality on performance and productivity Related papers Pollut ant s emission from building mat erials and t heir influence on indoor air qualit y and peopl… Raimondas Bliūdžius Sick Building Syndrome Sympt oms and Performance in a Field Laborat ory St udy at Different Levels of… The effects of indoor air quality on performance and productivity", Accessed: Nov. 20, 2022. [Online]. Available: www.blackwellpublishing.com/ina

[3] K. Kolcaba, *Comfort theory and practice: a vision for holistic health care and research*. 2003. Accessed: Aug. 30, 2022. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=nduGie_ouQkC&oi=fnd&pg=PR11&dq=Comfort+theory+and+practice:+A+vision+for+holistic+health+care+and+research.&ots=Sa22zLLeIe&sig=3YbxPwjqGCBq9BFsW9g44902rgI

[4] M. Frontczak, P. W.-B. and environment, and undefined 2011, "Literature survey on how different factors influence human comfort in indoor environments," *Elsevier*, Accessed: Aug. 30, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132310003136

[5] K. S.-R. Energy and undefined 1994, "Daylighting design: Enhancing energy efficiency and visual quality," *Elsevier*, Accessed: Aug. 30, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0960148194901163

[6] "WHO AIR QUALITY GUIDELINES GLOBAL UPDATE 2005 MEETING REPORT WHO air quality guidelines global update 2005." [Online]. Available: http://www.euro.who.int/pubrequest.

[7] S. H. Hwang and W. M. Park, "Indoor air concentrations of carbon dioxide (CO2), nitrogen dioxide (NO2), and ozone (O3) in multiple healthcare facilities," *Environ Geochem Health*, vol. 42, no. 5, pp. 1487–1496, May 2020, doi: 10.1007/S10653-019-00441-0/TABLES/4.

[8] L. Pérez-Lombard, J. Ortiz, C. P.-E. and buildings, and undefined 2008, "A review on buildings energy consumption information," *Elsevier*, Accessed: Aug. 31, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778807001016

[9] M. Peña, F. Biscarri, E. Personal, and C. León, "Decision Support System to Classify and Optimize the Energy Efficiency in Smart Buildings: A Data Analytics Approach," *Sensors*, vol. 22, no. 4, Feb. 2022, doi: 10.3390/s22041380.

[10] "Electricity total final consumption by sector, 1971-2019 – Charts – Data & Statistics - IEA." https://www.iea.org/data-and-statistics/charts/electricity-total-final-consumption-by-sector-1971-2019 (accessed Dec. 04, 2022).

[11] “Energieverbrauch nach Energieträgern und Sektoren | Umweltbundesamt.” https://www.umweltbundesamt.de/daten/energie/energieverbrauch-nach-energietraegern-sektoren#entwicklung-des-endenergieverbrauchs-nach-sektoren-und-energietragern (accessed Nov. 21, 2022).

[12] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “A survey on sensor networks,” *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–105, 2002, doi: 10.1109/MCOM.2002.1024422.

[13] D. Jost, “Fierce Electronics,” Jul. 2019. https://www.fierceelectronics.com/sensors/what-a-temperature-sensor (accessed Jun. 13, 2022).

[14] “The relationship between thermal environment and work performance”.

[15] B. W. Olesen, “General rights Indoor environmental input parameters for the design and assessment of energy performance of buildings”.

[16] “Climate and average monthly weather in Dresden (Saxony), Germany.” https://weather-and-climate.com/average-monthly-Rainfall-Temperature-Sunshine,Dresden,Germany (accessed Dec. 04, 2022).

[17] H. Farahani, R. Wagiran, and M. N. Hamidon, “Humidity sensors principle, mechanism, and fabrication technologies: A comprehensive review,” *Sensors (Switzerland)*, vol. 14, no. 5. MDPI AG, pp. 7881–7939, Apr. 30, 2014. doi: 10.3390/s140507881.

[18] “p16-19_Thermal_and_acoustic_comfort_RJ1302”.

[19] “Home Humidity Levels Chart: Manage Your Indoor Humidity & Comfort.” https://www.thecomfortauthority.com/home-humidity-levels-chart/ (accessed Nov. 23, 2022).

[20] J. Boudaden, A. Klumpp, H.-E. Endres, and I. Eisele, “Capacitive CO2 Sensor,” Aug. 2017, p. 472. doi: 10.3390/proceedings1040472.

[21] H. Ruser, “Accurate and reliable CO2 sensors with high long-term stability for demand controlled ventilation IWO-BAY View project SmartPointer View project”, doi: 10.13140/RG.2.2.23650.27849.

[22] “Richtwerte für 2-Propanol in der Innenraumluft : Mitteilung des Ausschusses für Innenraumrichtwerte (AIR),” *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, vol. 64, no. 10, pp. 1318–1327, Oct. 2021, doi: 10.1007/S00103-021-03402-0.

[23] “Light Sensors: Units, Uses, and How They Work.” https://blog.endaq.com/how-light-sensors-work (accessed Dec. 04, 2022).

[24] “What is a Light Sensor?” https://www.apogeeweb.net/electron/basics-of-light-sensor (accessed Dec. 04, 2022).

[25] “Illuminance - Recommended Light Level.” https://www.engineeringtoolbox.com/light-level-rooms-d_708.html (accessed Nov. 29, 2022).

[26] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 2, pp. 159–170, Jun. 2010, doi: 10.1109/SURV.2010.021510.00088.

[27] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 2, pp. 159–170, Jun. 2010, doi: 10.1109/SURV.2010.021510.00088.

[28] L. Erhan *et al.*, "Smart anomaly detection in sensor systems: A multi-perspective review," *Information Fusion*, vol. 67, pp. 64–79, Mar. 2021, doi: 10.1016/J.INFFUS.2020.10.001.

[29] J. Zhang, "ICST Transactions Preprint Advancements of Outlier Detection: A Survey."

[30] Vimal Jyothi Engineering College. Department of Electrical and Electronics Engineering, Institute of Electrical and Electronics Engineers. Kerala Section, and Institute of Electrical and Electronics Engineers, *ICCPCCT-2018 : 2018 International Conference on Control, Power, Communication and Computing Technologies : 23rd-24th March 2018*.

[31] A. Fawzy, H. M. O. Mokhtar, and O. Hegazy, "Outliers detection and classification in wireless sensor networks," *Egyptian Informatics Journal*, vol. 14, no. 2, pp. 157–164, Jul. 2013, doi: 10.1016/J.EIJ.2013.06.001.

[32] V. J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," *Artificial Intelligence Review 2004 22:2*, vol. 22, no. 2, pp. 85–126, Oct. 2004, doi: 10.1023/B:AIRE.0000045502.10941.A9.

[33] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 2, pp. 159–170, Jun. 2010, doi: 10.1109/SURV.2010.021510.00088.

[34] X. Wang, X. L. Wang, Y. Ma, and D. M. Wilkes, "A fast MST-inspired kNN-based outlier detection method," *Inf Syst*, vol. 48, pp. 89–112, Mar. 2015, doi: 10.1016/J.IS.2014.09.002.

[35] N. Chugh, M. Chugh, and A. Agarwal, "Outlier detection in streaming data a research perspective," *Proceedings of 2014 3rd International Conference on Parallel, Distributed and Grid Computing, PDGC 2014*, pp. 429–432, Feb. 2015, doi: 10.1109/PDGC.2014.7030784.

[36] J. Zhang, "ICST Transactions Preprint Advancements of Outlier Detection: A Survey."

[37] Vimal Jyothi Engineering College. Department of Electrical and Electronics Engineering, Institute of Electrical and Electronics Engineers. Kerala Section, and Institute of Electrical and Electronics Engineers, *ICCPCCT-2018 : 2018 International Conference on Control, Power, Communication and Computing Technologies : 23rd-24th March 2018*.

[38] F. Jiang, G. Liu, J. Du, and Y. Sui, "Initialization of K-modes clustering using outlier detection techniques," *Inf Sci (N Y)*, vol. 332, pp. 167–183, Mar. 2016, doi: 10.1016/J.INS.2015.11.005.

[39] S. Di, M. Fabiano, P. Prinetto, and M. Carami, "Design Issues and Challenges of File Systems for Flash Memories," *Flash Memories*, Sep. 2011, doi: 10.5772/22976.

[40] S. Sironi, L. Capelli, S. Vitali, and C. Bax, "Investigation of Electronic Nose Sensor Drift Correction Methods and Their Application to Environmental Samples," *Chem Eng Trans*, vol. 68, 2018, doi: 10.3303/CET1868049.

[41] "(PDF) Faulty sensor detection, identification and reconstruction of indoor air quality measurements in a subway station." https://www.researchgate.net/publication/254012802_Faulty_sensor_detection_identification_and_reconstruction_of_indoor_air_quality_measurements_in_a_subway_station (accessed Nov. 23, 2022).

[42] G. von Arx, M. Dobbertin, and M. Rebetez, "Detecting and correcting sensor drifts in long-term weather data," *Environ Monit Assess*, vol. 185, no. 6, pp. 4483–4489, Jun. 2013, doi: 10.1007/S10661-012-2831-6/FIGURES/4.

[43] S. Munirathinam, "Drift Detection Analytics for IoT Sensors," *Procedia Comput Sci*, vol. 180, pp. 903–912, Jan. 2021, doi: 10.1016/J.PROCS.2021.01.341.

[44] Y. Wang, A. Yang, Z. Li, X. Chen, P. Wang, and H. Yang, "Blind drift calibration of sensor networks using sparse Bayesian learning," *IEEE Sens J*, vol. 16, no. 16, pp. 6249–6260, Aug. 2016, doi: 10.1109/JSEN.2016.2582539.

[45] BIM4EEB, "BIM4EEB ITALIAN BUILDING SENSORS MEASUREMENTS DATASET," Jun. 2022, doi: 10.5281/ZENODO.6783695.

[46] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychol Methods*, vol. 7, no. 2, pp. 147–177, 2002, doi: 10.1037/1082-989X.7.2.147.

[47] M. Caselli, L. Trizio, G. de Gennaro, and P. Ielpo, "A simple feedforward neural network for the PM10 forecasting: Comparison with a radial basis function network and a multivariate linear regression model," *Water Air Soil Pollut*, vol. 201, no. 1–4, pp. 365–377, Jul. 2009, doi: 10.1007/S11270-008-9950-2/TABLES/5.

[48] D. Saini, A. Saxena, and R. C. Bansal, "Electricity price forecasting by linear regression and SVM," *2016 International Conference on Recent Advances and Innovations in Engineering, ICRAIE 2016*, 2016, doi: 10.1109/ICRAIE.2016.7939509.

[49] D. A. Guastella, G. Marcillaud, and C. Valenti, "Edge-Based Missing Data Imputation in Large-Scale Environments," *Information 2021, Vol. 12, Page 195*, vol. 12, no. 5, p. 195, Apr. 2021, doi: 10.3390/INFO12050195.

[50] S. Sunaryo, S. Suhartono, A. J. Endharta, S. Sunaryo, S. Suhartono, and A. J. Endharta, "Double Seasonal Recurrent Neural Networks for Forecasting Short Term Electricity Load Demand in Indonesia," *Recurrent Neural Networks for Temporal Data Processing*, Feb. 2011, doi: 10.5772/15062.

[51] W. Wu, X. Cheng, M. Ding, K. Xing, F. Liu, and P. Deng, "Localized outlying and boundary data detection in sensor networks," *IEEE Trans Knowl Data Eng*, vol. 19, no. 8, pp. 1145–1156, 2007, doi: 10.1109/TKDE.2007.1067.

[52] V. Kotu and B. Deshpande, "Anomaly Detection," in *Data Science*, Elsevier, 2019, pp. 447–465. doi: 10.1016/b978-0-12-814761-0.00013-7.

[53] H. A. R. de Bruin, B. J. J. M. van den Hurk, and L. J. M. Kroon, "On The Temperature-Humidity Correlation And Similarity," *Boundary-Layer Meteorology 1999 93:3*, vol. 93, no. 3, pp. 453–468, 1999, doi: 10.1023/A:1002071607796.

[54] J. L. Nguyen, J. Schwartz, and D. W. Dockery, "The relationship between indoor and outdoor temperature, apparent temperature, relative humidity, and absolute humidity," *Indoor Air*, vol. 24, no. 1, pp. 103–112, Feb. 2014, doi: 10.1111/INA.12052.

[55] J. L. Nguyen, J. Schwartz, and D. W. Dockery, "The relationship between indoor and outdoor temperature, apparent temperature, relative humidity, and absolute humidity," *Indoor Air*, vol. 24, no. 1, p. 103, Feb. 2014, doi: 10.1111/INA.12052.

[56] "Introduction to Linear Regression Analysis - Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining - Google Books." https://books.google.de/books?hl=en&lr=&id=tCIgE-AAAQBAJ&oi=fnd&pg=PR13&dq=linear+regression&ots=lfxgUwd2Jo&sig=qz2oE8M_s7zi_1h1nI3Qebaq4Z0&redir_esc=y#v=onepage&q=linear%20regression&f=false (accessed Nov. 24, 2022).

[57] C. C. Holt, "Forecasting Trends and Seasonal by Exponentially Weighted Averages," *Int J Forecast*, vol. 20, no. 1, pp. 5–10, Jan. 2004, doi: 10.1016/j.ijforecast.2003.09.015.

[58] C. C. Holt, "Forecasting Trends and Seasonal by Exponentially Weighted Averages," *Int J Forecast*, vol. 20, no. 1, pp. 5–10, Jan. 2004, doi: 10.1016/j.ijforecast.2003.09.015.

[59] J. Fattah, L. Ezzine, Z. Aman, H. el Moussami, and A. Lachhab, "Forecasting of demand using ARIMA model," *International Journal of Engineering Business Management*, vol. 10, Oct. 2018, doi: 10.1177/1847979018808673.

[60] "Autoregressive Integrated Moving Average (ARIMA) Definition." https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp (accessed Dec. 04, 2022).

[61] "ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+." https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/ (accessed Nov. 29, 2022).

[62] J. Karlapudi, P. Valluru, and K. Menzel, "An explanatory use case for the implementation of Information Container for linked Document Delivery in Common Data Environments Automatic data transfer and data interlink between BIM and Energy Simulation Models View project sfi-src itobo View project." [Online]. Available: https://www.researchgate.net/publication/353769852