Capstone Project 2

Useful Analytics for Smart Farming

by

Chan Wei Chee
16052755
BSc (Hons) Information Systems (Business Analytics)

Supervisor : Assoc. Prof. Dr Lau Sian Lun
Date: 19 November 2021

Project Title: Useful Analytics for Smart Farming

Date : 19/11/2021
Student : Chan Wei Chee
Supervisor: Associate Professor Dr Lau Sian Lun

Abstract

IoT sensors data can be a useful asset to smart farming and they are used widely in precision agriculture. Sensors are placed across Sunway Future X Farm to collect the climatic data. These data are used to assist the farm team in monitoring and optimizing crop yields by adapting to changes in environmental conditions. The work presented in this paper constitutes a contribution in leveraging data mining techniques to analyze the time series data collected from farm-monitoring sensors between March to July 2021. More specifically, this study demonstrated how the data mining process in terms of data preparation, exploratory analysis and predictive analysis of sensor data could be carried out. Besides, the underlying structure of the time series and appropriate forecasting techniques to forecast the climatic variables by using ARIMA and SARIMA were presented thoroughly in this study. Several ARIMA models were developed and the adequate model was selected according to the lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The result of the best-fitted model for forecasting the climatic variables was presented and evaluated by comparing the prediction values with actual test series data from 29th June to 29th July 2021 using root mean squared error (RMSE). SARIMA model has better performance in forecasting the climatic variables as compared to ARIMA model.

A summary dashboard is designed to visualize the sensor data as well as the information of crop yields. By accessing the climatic forecasts and being able to monitor the sensors data, the farm can leverage insights from these analyses to improve its production while minimizing cost and preserving resources.

Keywords: IoT Sensor, Smart Farming, Precision Agriculture, Time Series Forecasting

TABLE OF CONTENTS

1 Introduction 9
1.1.1 Problem Statement 10
1.1.2 Research Aim 10
1.1.3 Research Objectives 10
2 Literature Review 11
2.1 Smart Farming 11
2.1 Sensor Network Data 11
2.2 Data Mining Techniques 12
2.3 Data Mining Techniques Used In The Field of Agriculture 15
2.4 Time Series Analysis and Forecasting 21
2.4.1 ARIMA Model 21
3 Methodology 23
3.1 Data Collection 23
3.2 Data Preparation 24
3.2.1 Transform and convert timestamp into DateTime 25
3.2.2 Transform “dbl_v” and “long_v” into “values” column 25
3.2.3 Creation of “Device” column 25
3.2.4 Subsetting and filtering valid keys based on conditions 25
3.2.5 Create an individual column for each variable 25
3.2.6 Removal of duplicate data 25
3.2.7 Extract date from DateTime 26
3.2.8 Set DatetimeIndex 26
3.2.9 Derive minimum and maximum value 26
3.2.10 Resampling 26
3.2.11 Deal with missing value 27
3.3 Data Modelling 29
3.3.1 DateTimeIndex and Resampling 29
3.3.2 Decomposition of time series - Trend, Seasonality, Trend and Noise 29
3.3.3 Stationarity 29
3.3.4 Differencing 30
3.3.5 Train/Test validation split 30
3.3.6 Determine the order of AR and MA terms 31
3.3.7 Build and fit the ARIMA/ SARIMA model 33
3.4 Model Evaluation 33

3.5 Data Visualisation 34
3.5.1 Determine the User/audience and their Intended Usage of the Dashboard 34
3.5.2 Selection of Key Data Points 34
3.5.3 Selection of Proper Filters and Drill-downs 35
3.5.4 Selection of Color Design 36
3.5.5 Selection of Charts 37
4 Results and Discussions 43
4.1 Results of Exploratory Data Analysis 43
4.1.1 Average Air Temperature Readings Collected From Indoor and Outdoor Devices 43
4.1.2 Average Air Humidity Readings Collected From Indoor and Outdoor Devices 47
4.1.3 Average Light Intensity Readings Collected From Indoor and Outdoor Devices 50
4.1.4 Correlation Between Air Temperature and Air Humidity 52
4.2 Results Interpretation of Time Series Analysis 52
4.2.1 Results of Trend, Seasonality and Noise 52
4.3 Result Interpretation of ARIMA Model 67
4.4 Results Interpretation of Prediction Dashboard 70
4.4.1 Sensors Dashboard 70
4.4.2 Yield Analysis Dashboard 72
5 Limitation 77
6 Conclusion 78
7 References 79
8 Appendix 84
8.1 Key Identifications of Devices 84
8.2 Screenshots of Dashboard Design 86
8.2.1 Sensors Dashboard 86
8.2.2 Yield Analysis Dashboard 89
8.3 Screenshots of Results of ARIMA model using ACF/PACF plots 91
8.4 Screenshots of Results of ARIMA model using Auto ARIMA 92

LIST OF TABLES

Table 2-1: Summarized tables of common classification algorithms use 13
Table 2-2: Summarized tables of common classification algorithms use 14
Table 2-3: Summarized tables of data mining techniques in agricultural analysis work 18
Table 3-1: Raw Sensor Data Collected 23
Table 3-2: Crop Yield Data 23
Table 3-3: Terminologies of Time Series Analysis 24
Table 4-1: Average Air Temperature Readings collected from Indoor Device 1 and Device 2 43
Table 4-2: Distribution of Air Temperature Readings collected from Outdoor Device 1 & 2 44
Table 4-3: Distribution of Air Humidity Readings collected from Indoor Device1 and Device2 47
Table 4-4: Distribution of Air Humidity Readings collected from Outdoor Device1&2 48
Table 4-5: Distribution of Light Intensity Readings collected from Outdoor Device1&2 50
Table 4-6: Summary Table of the optimal parameters to determine the order of the model. 69
Table 8-1: Key Identification of Devices for Air Temperature 84
Table 8-2: Key Identification of Devices for Air Humidity 84
Table 8-3: Key Identification of Water Temperature 84
Table 8-4: Key Identification of Devices for PH Value 84
Table 8-5: Key Identification of Devices for Water EC Value 84
Table 8-6: Key Identification of Devices for Electric Current 85
Table 8-7: Key Identification of Devices for Light Intensity 85
Table 8-8: Key Identification of Devices for Water Float State 85

TABLE OF FIGURES

Figure 3-1: Results of Resampling 26
Figure 3-2: Density chart of comparison of three imputation techniques 28
Figure 3-3: Visualisation results of the three imputation techniques of air_temp variable 28
Figure 3-4: Visualisation results of the three imputation techniques of water_ temp variable 28
Figure 3-5: Visualisation of splitting of training and testing data 30
Figure 3-6: Process flow for performing manual model orders of AR and MA terms 31
Figure 3-7: Process Flow of Auto Model Orders Selection using AUTO ARIMA 32
Figure 3-8: Filter and Drill-downs of Air Temperature vs. Air Humidity Dashboard 35
Figure 3-9: Filter and Drill-downs of Yield Analysis Dashboard 35
Figure 3-10: Color Design of Dashboard 36
Figure 3-11: Color Design of Dashboard 37
Figure 3-12: Selection of Bar Charts in Sensors Dashboard 38
Figure 3-13: Selection of Gauge Charts in Sensors Dashboard 38
Figure 3-14: Selection of Line Charts in Sensors Dashboard 39
Figure 3-15: Selection of Bar Chart in Yield Analysis Dashboard 39
Figure 3-16: Selection of Tables in Yield Analysis Dashboard 40
Figure 3-17: Selection of Bullet Charts in Yield Analysis Dashboard 41
Figure 3-18: Summary Tables in Yield Analysis Dashboard 41
Figure 3-19: Selection of Tornado Charts in Yield Analysis Dashboard 42
Figure 4-1: Distribution of Air Temperature Readings collected from Indoor Device1 and Device2 43
Figure 4-2: Distribution of Air Temperature Readings collected from Outdoor Device 1&2 44

Figure 4-3: Outliers observed from Outdoor Device 1 and Device 2 45
Figure 4-4: Density Plot of Air Temperature Readings Collected from Indoor and Outdoor Devices 45
Figure 4-5: Overview of Average Air Temperature Readings Collected from Indoor and Outdoor Devices 46
Figure 4-6: Distribution of Air Humidity Readings Collected from Indoor Device1 and Device2 47
Figure 4-7: Distribution of Air Humidity Readings Collected From Outdoor Devices 48
Figure 4-8: Density Plot of Air Humidity Readings Collected from Indoor and Outdoor Devices 49
Figure 4-9: Overview of Average Air Humidity Readings Collected from Indoor and Outdoor Devices 49
Figure 4-10: Average Light Intensity Readings Collected From Outdoor Device 1&2 50
Figure 4-11: Density Plot of Light Intensity Readings Collected from Outdoor Device 1&2 51
Figure 4-12: Correlation Plot Between Air Temperature and Air Humidity 52
Figure 4-13: Time Series Plot A-E 53
Figure 4-14: Decomposition of Time Series Plot A-E 54
Figure 4-15: Rolling mean and Variance of Air Temperature of Indoor Device 1 and Device 2 55
Figure 4-16: Rolling mean and Variance of Air Temperature of Outdoor Device 1&2 55
Figure 4-17: Rolling mean and Variance of Air Humidity of Indoor Device 1 and Device 2 56
Figure 4-18: Rolling mean and Variance of Air Humidity of Outdoor Device 1&2 56
Figure 4-19: Rolling mean and Variance of Light Intensity of Outdoor Device 1 and Device 2 57
Figure 4-20: Auto correlation Plot of Air Temperature in Indoor Device 1 58
Figure 4-21: Auto correlation Plot of Air Humidity in Outdoor Device 1&2 58
Figure 4-22: Results of ADF Test of Air temperature collected from Indoor Device 1 59
Figure 4-23: Results of ADF Test of Air temperature collected from Outdoor Device 1&2 .. 59
Figure 4-24: Results of ADF Test of Air Humidity Collected from Indoor Device 1 60
Figure 4-25: Results of ADF Test of Air Humidity Collected from Outdoor Device 1&2 60
Figure 4-26: Results of ADF Test of Light Intensity Collected from Outdoor Device 1&2 ... 60
Figure 4-27: Auto correlation Plot of Air Temperature and Humidity Before Differencing (Non-Stationary) 61
Figure 4-28: Auto correlation Plot of Air Temperature and Humidity After Differencing (Stationary) 62
Figure 4-29: Partial Auto correlation Plot of Air Temperature and Humidity of Indoor and Outdoor Device Before Differencing (Non-Stationary) 62
Figure 4-30: Partial Auto correlation Plot of Air Temperature and Humidity of Indoor and Outdoor Device After Differencing (Stationary) 63
Figure 4-31: Time Series Plot of Air Temperature Collected From Indoor Device 1 Before and After Differencing 63
Figure 4-32: Time Series Plot of Air Humidity Collected From Outdoor Device 1 and Device 2 Before and After Differencing 64
Figure 4-33: Results of ADF Test of Air Temperature Collected from Indoor Device 1 After Differencing 64

Figure 4-34: Results of ADF Test of Air Humidity Collected from Outdoor Device 1 and Device 2 After Differencing 65
Figure 4-35: Auto Correlation Plot and Partial Correlation of Air Temperature in Outdoor Device 1 and Device 2 (Stationary) 65
Figure 4-36: Auto Correlation Plot and Partial Correlation of Air Humidity in Outdoor Device 1 and Device 2 (Stationary) 66
Figure 4-37: Auto Correlation and Partial Correlation Plot of Light Intensity in Outdoor Device 1 and Device 2 (Stationary) 66
Figure 4-38: Result of ARIMA model of Outdoor Device 1&2 Air Temperature Key 46 generated by determining the parameters order using ACF/PACF plot 67
Figure 4-39: Result of ARIMA model of Outdoor Device 1&2 Air Temperature Key 46 generated by determining the order of the parameters using Auto ARIMA 68
Figure 4-40: Prediction of Air Temperature Key 46 for Next 31 Days 69
Figure 4-41: Prediction of Air Temperature Dashboards 70
Figure 4-42: Prediction of Light Intensity Dashboards 71
Figure 4-43: Yield Analysis Dashboards 72
Figure 4-44: Actual Yield by Types of Vege 72
Figure 4-45: Bar Chart of Actual Yield 73
Figure 4-46: Summary Table of Total Yield For each Vegetable by Month 73
Figure 4-47: Detail Breakdown Table of Yield 74
Figure 4-48: Barchart of Actual Yield by Date of Harvest 75
Figure 4-49: KPI of Actual Yield vs. Expected Yield 75
Figure 4-50: Tornardo Chart of Expected Yield vs Actual Yield 76
Figure 8-1: Dashboard of Air Temperature VS. Air Humidity from Indoor and Outdoor Devices 86
Figure 8-2: Dashboard of Water Temperature Collected From Indoor and Outdoor Devices 87
Figure 8-3: Dashboard of Electric Current Collected From Indoor and Outdoor Devices 88
Figure 8-4: Yield Analysis Dashboard 89
Figure 8-5: Yield KPI Analysis Dashboard 89
Figure 8-6: Prediction Dashboard 90
Figure 8-7: Result of ARIMA of Outdoor Device 1&2 Air Temperature Key 55 generated by determining the parameters order using ACF/PACF plot 91
Figure 8-8 : Result of ARIMA model of Outdoor Device 1&2 Air Temperature Key 55 generated by determining the order of the parameters using Auto ARIMA 92
Figure 8-9: Prediction of Outdoor Device 1&2 Air Temperature Key 55 for Next 31 Days .92
Figure 8-10: Result of ARIMA model of Outdoor Device 1&2 Light Intensity Key 66 generated by determining the order of the parameters using Auto ARIMA 93
Figure 8-11: Prediction of Outdoor Device 1&2 Light Intensity Key 66 for Next 31 Days ... 93
Figure 8-12: Result of ARIMA model of Outdoor Device 1&2 Light Intensity Key 67 generated by determining the order of the parameters using Auto ARIMA 94
Figure 8-13 : Prediction of Outdoor Device 1&2 Light Intensity Key 67 for Next 31 Days .. 94
Figure 8-14: Result of ARIMA model of Outdoor Device 1&2 Light Intensity Key 68 generated by determining the order of the parameters using Auto ARIMA 95
Figure 8-15: Prediction of Outdoor Device 1&2 Light Intensity Key 68 for Next 31 Days ... 95
Figure 8-16 : Result of ARIMA model of Outdoor Device 1&2 Light Intensity Key 69 generated by determining the order of the parameters using Auto ARIMA 96

Figure 8-17: Prediction of Outdoor Device 1&2 Light Intensity of Key 69 for Next 31 Days 96
Figure 8-18: Result of ARIMA model of Outdoor Device 1&2 Air Humidity Key 45 generated by determining the order of the parameters using Auto ARIMA.....97
Figure 8-21:.....97
Figure 8-20: Prediction of Outdoor Device 1&2 Air Humidity Key 45 for Next 31 Days 97
Figure 8-21: Result of ARIMA model of Outdoor Device 1&2 Air Humidity Key 54 generated by determining the order of the parameters using Auto ARIMA.....98
Figure 8-22: Prediction of Outdoor Device 1&2 Air Humidity Key 54 for Next 31 Days 98

1 Introduction

Due to the rapid growth of the population in Malaysia, the demand for food has been increased dramatically. According to Natalie [1], Malaysia's has a high dependency on other countries for food supply with only 8% of agricultural land for agro-food production as well as a lack of research and development in the agriculture industry. Malaysia is food secure on a national level but food insecure on a household level. To address these problems, here come the Sunway Future X Farm [2] which adopts smart farming and precision farming concept to serve local communities.

Smart farming is an emerging concept of connected smart machines and sensors integrated on farms to make farming processes data-driven and data-enabled. The notion of precision farming is being supported by digital technologies includes the Internet of things(IoT) [3], remote sensing technologies [4], geographic information systems(GPS) technologies, robotics, drones and big data applications [5] in the farming practices for tracking, monitoring, automating and analyzing operations while optimizing human labour.

The challenge of farmers includes dealing with in-numerous decision-making and influencing climatic factors every day. A necessary approach is proposed for accomplishing practical and effective solutions for this problem. Data mining is the way of analyzing and discovering hidden patterns in the data from a different perspective and summarizing it into useful information and it is an emerging research field in agriculture crop yield analysis. Sensors are embedded and set up across the fields to collect data on light levels, soil conditions, irrigation, air quality and weather which eased the data acquisition.

With the advent of sensor and GPS technology in agriculture, farmers not only harvest the crops but also harvest large amounts of data. The data collected should be analyzed to provide insights about the farm and make data-based decisions to enhance the productivity of the crop yield. Data mining has a wide range of applications in the field of agriculture. In this study, applications of the data mining techniques in the area of agriculture and its allied areas are studied.

1.1.1 Problem Statement

Sunway Future X farm has set up the sensors across the fields and access to a wealth of data on air humidity, air temperature, light intensity, water pH, water EC value and other climatic variables. The farm management teams are interested in knowing how the sensors data can provide useful insights in improving the farming productivity and other facets of the farming practice. Hence, a thorough analysis is required to extract insights from these massive time-series sensor data.

1.1.2 Research Aim

To our best knowledge, there is little empirical research effort related to sensor data analysis in smart farming. Most of the studies focused only on crop yields analysis. Hence, to bridge the mentioned research gap, the present study aims to provide a thorough study on the applications of data analysis in the sensors data and build a predictive model for forecasting the climatic variables. The insights gathered through the analysis will allow the management team to take necessary actions to improve business operations.

1.1.3 Research Objectives

More specifically, the study intends to achieve the following objectives :

i. To investigate the characteristics of the climatic variables through exploratory data analysis.

ii. To forecast the value of climatic variables for the next 30/31 days through predictive modelling.

iii. To create a summary dashboard for visualizing the sensor data and crop yields information.

The following sections of this document will present materials organized in order and details as stated below:

Section 2: Literature review for related studies on data mining in agriculture.

Section 3: Methodology of the study.

Section 4: Results and discussions.

Section 5: Limitation.

Section 6: Conclusion.

2 Literature Review

2.1 Smart Farming

Smart farming is seen as one way to achieve the goals of resource optimization and sustainable production with controlled costs in the field of agriculture. Such a new form of agriculture has attracted the attention of the research in the agriculture community. Smart farming achieved these goals by leveraging the use of the IoT (Internet of Things) [3] and machine learning. Its goal is to collect data coming from heterogeneous sources to understand, predict and better organize the farming activity. With the wide application of IoT technology, sensors are deployed everywhere to realize the function of real-time data communication and information processing which improve the development of smart agriculture.

2.1 Sensor Network Data

Smart farming is based on the use of different technologies of automation, data capture, data transmission, data processing and decision making. One of the most common data collection tools in this sector is the wireless sensor network [4]. A large number of sensor nodes will form a variety of monitoring networks in the agricultural field to collect data on different environmental parameters such as humidity, temperature, soil moisture, PH value of soil etc. which results in enhancing the quantity and quality of crops. Indeed, meteorological information can warn farmers against infectious diseases and lead to appropriate actions. Farmers can also use the data collected for crop protection and custom fertilization, resulting in higher yields with reduced environmental impact [4].

The study by [3] indicates that the environmental measurements of temperature, humidity and soil moisture are considered the most critical parameters for agriculture and farming by most researchers.

However, manual processing and analysis of these huge data are difficult or impossible for human, so it is essential to leverage analysis tools to transform the raw data into insightful knowledge that elevate the decision-making process. As a result, Data Mining has received a great deal of interest because of its strong ability to extract information from data.

2.2 Data Mining Techniques

Data mining is the process of extracting and discovering patterns and correlations within large datasets and is an interdisciplinary field of machine learning, statistics, database technology, and other disciplines. [6] The 10 most used data mining techniques are discussed in a paper [7] The data mining tasks are twofold and can be classified into two types of categories: descriptive data mining tasks that characterize the general properties of existing data to find interesting patterns describing the data. An example of the most used techniques is Clustering and Association Analysis. [8]

On the other hand, predictive data mining tasks do prediction based on inference on the available data. Regression and Classification and time-series analysis were used widely in many applications [9]

Table 2-1 and Table 2-2 summarized the most common classifications and clustering algorithms that were used [10].

Table 2-1: Summarized tables of common classification algorithms use

DM Techniques	Description	Example Algorithms
Classification [7]	Classification is a supervised learning process that allows the prediction of a class label from a set of training data. It consists of mapping each element of the selected data into one of a predefined class set.	Decision Tree Random Forest Artificial Neural Network (ANN) K-Nearest Neighbours (KNN) Naive Bayes Support Vector Machines (SVM)
Regression [7]	Regression analysis is a statistical method for predicting the relationship between a dependent variable (y) and of one or multiple predictor variables (x).	Linear regression Logistic regression Time series analysis

Table 2-2: Summarized tables of common classification algorithms use

DM techniques	Type of Clustering	Definition	Example algorithms
Clustering [7]		Clustering is an unsupervised machine learning task that finds similarities in the data point and groups similar data points together.	
	Partitioning-Based	Partition based classifies the information into multiple groups based on the characteristics and similarity of the data according to the number of clusters specified.	K-means, Partition around medoids (PAM) and Clustering large applications (CLARA)
	Density-Based	Density-Based Clustering can identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.	DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
	Hierarchical Based	Hierarchical based clustering is an unsupervised clustering algorithm that involves creating clusters that have predominant ordering from top to bottom. This clustering technique is divided into two types: 1. Agglomerative Hierarchical Clustering 2. Divisive Hierarchical Clustering	CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies)
	Fuzzy	Fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster.	C-means

2.3 Data Mining Techniques Used In The Field of Agriculture

In recent years, data mining in the agricultural sector has been carried out extensively as the public has seen the potential of data in improving the productivity and quality of crop yield. In smart agriculture, the data brought back from the surveillance environment and collected by various equipment (drones, sensors, etc.) plays an important role. Datasets from the agricultural domain appear to be significantly more complex than the datasets traditionally used in machine learning [11] .In the agriculture area, predictive data mining techniques such as classification and clustering are mainly used. Different data mining techniques and methodologies were developed for crop yield prediction and to solve a variety of agricultural problems. The application of these techniques in agriculture were discussed below.

A study was undertaken to find the optimal range of best temperature and rainfall to maximize crop production in India. J. Majumdar et al. [12] used modified DBSCAN clustering techniques to cluster the data based on the districts with similar temperature, rainfall and soil type. PAM and CLARA are used to cluster the data based on the districts which are producing maximum crop production and then later multiple linear regression was used to forecast the crop yield. A comparison has been done based on quality metrics like purity, homogeneity, completeness, V Measure, Precision, Recall, etc. to determine the optimal clustering method. The author concluded that DBSCAN gives better results followed by CLARA and PAM.

Similarly, D. Ramesh and B. Vardhan [13] carried out another project with year, rainfall, area of sowing as parameters whereas yield production as the target variable. Four clusters were formed by considering rainfall as the key parameter by adopting k-mean clustering and multiple linear approaches. The findings showed that the multiple linear regression (~98%) has higher accuracy than k-mean (~96%) as compared to the actual average production.

Neural networks have shown to be quite effective in modelling the yield of different crops. In [14] , a Multilayer Perceptron model of Neural Network has been trained for wheat yield prediction by considering sensor input and fertilizers as the parameter. The point of interest would be to investigate how the factor of fertilization determines the yield and additional factors that correlate directly or indirectly with a yield which cannot be discovered using regression and correlation techniques like PCA.

Comparison of regression models of Neural Network has been used by the researchers Georg Ruβ et al. in [15] to show the overview on the capabilities of different regression techniques in the classification of crops and yield prediction: Multi-layered Perceptron (MLP), Radial Basis Function (RBF), Regression Trees, Support Vector Machines. Model performance was evaluated using the root mean squared error (RMSE) and mean absolute error (MAE). A comparison of these four techniques showed that the Support Vector Regression technique produced better yield predictions.

The study in [16] concluded that high precision prediction of agricultural data is realized by neural network model. Data such as soil moisture content, temperature and humidity, light intensity, crop growth status, and weather factors were collected and transmitted to the server through ZigBee and the 3Gnetwork card, and the data are directly imported into the neural network model for processing the data through the Web Service. Finally, by comparing the prediction results with the actual data, it is found that the prediction error of the model designed in this article is within 1%.

M. M. Rahman [17] illustrated the use of an artificial neural network algorithm, Self-Organizing Map (SOM) [18] to depict the influence of daily extreme weather conditions on grapevine phenology, annual crop yield and wine quality. The two sets of data (climate and grapevine yield) are used in this research to see the associations between them. SOM algorithm was used to classify the data associations and the chi-square test was used to verify the SOM results and establish the degree of dependence between the related variable values.

Supervised bi-clustering techniques have been applied to the dataset of wine fermentations by [19]. J. Palanichamy et al. [19] revealed that this technique can simultaneously solve two data mining problems. First, it can select the features that are relevant in the fermentation process. Second, the information that is acquired by finding the bi-clustering of the dataset can be exploited for performing classifications of new fermentations. Therefore, feature selections and supervised classifications can be performed at the same time by using this technique. These studies have recently been published in [20]. The technique can perform good-quality predictions of problematic fermentations.

Following that, J. Palanichamy et al. [19] also used a weighted k-means algorithm to evaluate soil fertility. The comparison was done between the weighted k-mean algorithm and the traditional K-means clustering algorithm, the result showed that the weighted K-means clustering algorithm has better accuracy, operational efficiency, significantly higher than the un-weighted clustering algorithm. Furthermore, a comprehensive evaluation of the changes in soil nutrients after precision fertilization that used algorithm was conducted. The soil fertility status has a significant improvement after years of continuous precision fertilizing. The results show that the improved clustering algorithm is a good method for the comprehensive evaluation of soil fertility.

C. Li and B. Niu [21] determined the relative optimal growth environment curve of seedlings in each fixed cycle by collecting a large number of environmental data, including soil temperature, air temperature, air humidity, soil moisture. The author clusters the data of temperature and soil sensors module every X hours to obtain a relatively optimal environment curve. The k-means algorithm based on the maximum distance method to select the initial cluster centre is proposed to study data mining. The number of iterations is reduced and the clustering efficiency is improved. The crop growth curve is simulated and compared with the improved K-means algorithm and the original k-means algorithm in the experimental analysis. The experimental results show that the improved K-means clustering method has an average reduction of 0.23 s in total time and an average increase of 7.67% in the F metric value.

I. A. Lakhiar et al. [22] comprehensively analyzed the application of intelligent sensor techniques in the aeroponic system to provide insight on how the key parameters of aeroponics correlate with plant growth. The author concluded that temperature, humidity, light intensity, pH and EC of the water nutrient solution, carbon dioxide concentration, nutrient atomization and atomization interval time will be the key parameter during aeroponic plant cultivation. The grower has the responsibility to control and monitor the fluctuations of the above parameters in the desired range for optimal growth.

The artificial neural network was used to identify relationships between plant nitrogen status and sensor data by R. Sui and J. Thomasson [23]. A feed-forward back-propagation ANN was developed to predict nitrogen status in cotton plants based on data from the sensing system. Data including spectral reflectance at four wavebands, plant height, and leaf N concentration

were used. The network was trained with actual leaf nitrogen concentration data that corresponded to sensor spectral data and plant height. Results showed that the spectral information and plant height measured by the sensing system had a significant correlation with the leaf nitrogen concentration of the cotton plants. Trained neural networks were able to predict the nitrogen status of the cotton plants at 90% accuracy when nitrogen status was divided into two categories: deficiency and non-deficiency.

Table 2-3: Summarized tables of data mining techniques in agricultural analysis work

Author	Topic	Parameter used	Target variable	Data mining technique(s)	Results (Accuracy)
[12]	Analysis of agriculture data using data mining techniques: application of big data	Year, Districts, Crop, Area, Season, Production, Average temperature, Average rainfall, PH value, Soil type	MLR - minimum rainfall required, the minimum temperature required	• DBSCAN - District having similar temperature range, soil type and rainfall • CLARA-District with different area size, rainfall and temperature PAM – District with low, moderate, high crop production • MLR – Forecast annual crop yield	DBSCAN gives the better clustering quality followed by CLARA and PAM.
[13]	Data Mining Techniques and Applications to Agricultural Yield Data	The average area of sowing and Yearly rainfall	MLR-average production	• K-mean –Cluster of the district having similar rainfall • MLR-Model relationship between year, area of sowing, average production	K-Mean -96% MLR -98%

[14]	Data Mining with Neural Networks for Wheat Yield Prediction	Nitrogen Fertilizer, Vegetation (REIF value) , Electric Conductivity (EM value),Yield2003, network topology	Yield 2003	• Multi-layered Perceptron (MLP) with back-propagation learning	Not specified
[15]	Data Mining of Agricultural Yield Data: A Comparison of Regression Models	Nitrogen Fertilizer, Vegetation (REIF value) , Electric Conductivity (EM value),Yield2003	Yield 2004 ,Yield 2006	• Multi-layered Perceptron (MLP) • Radial Basis Function (RBF) • Regression Trees • Support Vector Machines	Support Vector Regression technique produced better yield predictions than the other models
[16]	A Study of Big Data Application in Agriculture	soil moisture content, temperature and humidity, light intensity, crop growth status, and weather factor	Crop yield	•Back propagation neural network	The error between the predicted value and the actual value is less than 1%
[17]	Data Mining Techniques for Modelling the Influence of Daily Extreme Weather Conditions on Grapevine, Wine Quality and	Daily weather data, i.e., maximum, minimum and grass minimum temperatures Vintage, Grapes harvested (in Yield tons/hectare, Harvest Date, Brix (dissolved sugar-to-water	The cluster of maximum daily temperature with annual low and high yield year classes	• SOM (Artificial neural network algorithm) - classify the data associations • Chi-square test was used to verify SOM and establish the degree of dependence between the related variable values.	SOM has potential in clustering tasks

	Perennial Crop Yield	mass ratio of a liquid), Acid and pH.			
[19]	A Study of Data Mining Techniques to Agriculture	weight of soil nutrient attributes	The cluster of soil with similar fertility	•K-means – evaluate soil fertility Bi-clustering - feature selection and supervised classification	weighted K-means clustering algorithm has better accuracy than traditional k-mean
[21]	Design of smart agriculture based on big data and the Internet of things	soil temperature, air temperature, air humidity, soil moisture	Clusters of optimum soil temperature and soil humidity	•K-means algorithm based on the maximum distance method	Average reduction of 0.23 s in total time and an average increase of 7.67% in the F metric value
[21]	Monitoring and Control Systems in Agriculture Using Intelligent Sensor Techniques: A Review of the Aeroponic System	temperature, humidity, light intensity, water nutrient solution level, pH and EC value, carbon dioxide concentration	Plant growth	Not specified	Not specified
[22]	Ground-Based Sensing System for Cotton Nitrogen Status Determination	Spectral reflectance at four wavebands, plant height	leaf nitrogen concentration	Feed-forward back-propagation ANN	Predict nitrogen status of the cotton plants at 90% accuracy.

2.4 Time Series Analysis and Forecasting

2.4.1 ARIMA Model

Time series forecasting is widely used in forecasting demand and production. The author in [24] applied time series analysis on agriculture food production while [25] forecast the demand forecasting of finished products in a food manufacturing industry by fitting an ARIMA model to predict the values for the next few years. In the field of agriculture, researches conducted by [26] [27] demonstrated the application of time series forecasting in agricultural products. All of the papers mentioned above that a precise selection of ARIMA model is important in fitting the time series data to forecast future values.

The ARIMA model also known as Auto Regressive Integrated Moving Average developed by Box and Jenkins [28] is the most frequently used time series model and much research has successfully proved its ability in forecasting [29]. An ARIMA model is a class of statistical models that are widely used in forecasting time series data. It consists of 3 parameters ARIMA (p,d,q) which accounts for seasonality, trend, noise as defined below:

Auto regressive components: AR stands for Auto regressive and is denoted by p in which refers to the number of the Auto regressive term (AR order) or lag order observations to be included in the model. Intuitively, this means that how much previous time steps the current value of the variable depends on to predict the future.

Integrated: Integration is the degree of differentiation to be applied and it is denoted by d (differencing order). When d=0 means that it does not require any differencing and is already stationary while d =1 means that the first difference is taken to transform the non-stationary into stationary.

Moving average: MA stands for moving average and is denoted by q to refer to the number of moving average terms (MA order). This means that the number of previous errors (numbers of lags) in the past prediction that is taken into account in order to make a better future prediction. In a simpler term, it corrects future forecasts based on errors made in recent forecasts.

Together, the notation of the ARIMA model is specified as ARIMA (p,d,q)The integrated element allowed ARIMA to handle data with the trend but it does not support time series with a seasonal component.

SARIMA Model

SARIMA stands for Seasonal Auto -Regressive Integrated Moving Average which is the extension of ARIMA that has 3 more hyper parameters (P, D, Q) to explicitly specify the AR, I and MA for the seasonal components of the series as well as an additional parameter (m) for the period of the seasonality [30].

P: Seasonal Auto regressive order.

D: Seasonal difference order.

Q: Seasonal moving average order.

m: The number of time steps for a single seasonal period.

Together , the notation of SARIMA model is specified as : SARIMA(p,d,q)(P,D,Q)m. However, the implementation is called SARIMAX as the “X” addition to the method name means that the implementation also support exogenous variables.

3 Methodology

3.1 Data Collection

The data source for this research project is collected from the sensors data that were set up across the farm. The data collected for each device ranged from March to July 2021. The sensors data being captured included air temperature, air humidity, water temperature, light intensity, the electric current of LED lights and water pump, water pH and EC value. The sensors were placed in both indoor and outdoor. Overall, there are 7 datasets exported which consists of four indoor devices and three outdoor devices. In addition, a dataset that consists of information about crops yield data is utilized in the dashboard for further analysis. The description of raw sensor data and crop yield data are presented in Table 3-1 and Table 3-2.

Table 3-1: Raw Sensor Data Collected

Attributes	Description
entity_id	Unique Device ID
key	Unique key for each device
ts	Time stamp
bool_v	Boolean value
long_v	Character value
dbl_v	Decimal value

To maximize the contextual information that can be attained about the sensor data, the "long_v" and "dbl_v" columns are concatenated into a new column – "values", which will be used as input data in subsequent analysis. The dataset is then pre-processed according to the key identifications table as shown in Table 8-1 to 8-8 in the Appendix.

Table 3-2: Crop Yield Data

Attributes	Description
Types of Vegetables	Type of crops
Date of Harvest	The date that the crops were harvested
Estimated Yield (Plants)	Estimated number of plants successfully germinated from seeds
Actual Harvested (Plants)	Number of plants successfully germinated from seeds
Average weight per plant (g)	Average weight per plant in gram
Expected Yield(kg)	Average weight (g) x Estimated Yield (Plants)
Actual Yield (kg)	Actual weight of the yield harvested
Batch No	Batch number of date of harvest

3.2 Data Preparation

The sensor data collected were subjected to a lot of noise due to issues such as containing missing values due to the sensor is not working or there being variation in reading. To obtain accurate analysis and increase the performance of the time series forecasting for this project, it is critical to perform feature engineering to prepare the proper input dataset that is compatible with the time series analysis algorithm requirements. Therefore, this section will present all the detailed procedures taken to process and transform the raw data into clean data. All the data pre-processing and cleaning is conducted using *Python* as it is proved to be a flexible and efficient tool to work with time-series data [31].

The steps taken and terminologies used while performing time series analysis are presented in the summary Table 3-3. The variables were separated and stored into an individual dataset for easier processing and faster querying.

Table 3-3: Terminologies of Time Series Analysis

Steps	Data Preparation	Description
i.	Convert the timestamp into DateTime	Convert from Unix timestamp into milliseconds.
ii.	Transform “dbl_v” and “long_v” into “values” column	Combine values in ‘long_v’ and ‘dbl_v’ columns.
iii.	Creation of “Device column	To differentiate between different devices.
iv.	Subsetting and filtering valid key based on conditions	Include only valid keys in the dataset for further analysis.
v.	Create an individual column for each variable	Create individual columns for each of the variables according to their key.
vi.	Removal of duplicate data	As time-series analysis cannot contain duplicate time indexes, duplicated data are removed.
vii.	Extract date from DateTime	Extract only date for further analysis.
viii.	Set DateTime index	Set date as the index.
ix.	Derive minimum and maximum value	Derive minimum and maximum values from the original value.
x.	Resample of data	Resample the data into daily frequency by taking the mean value.
xi.	Imputation for missing value	Impute the missing value with interpolation technique.

3.2.1 Transform and convert timestamp into DateTime

In time series analysis, as the name suggests, involves variables that change with time so date-time play a vital role in this analysis. Analysis of time series data required the timestamps of the data in a structure that is amenable to time-based slicing and dicing. In this study, the incoming DateTime is in Unix timestamp and the values are of string type. It is then converted to human-readable dates using Python to_datetime() method from the *pandas* package that interprets the strings and convert them into DateTime objects. Unit = 'ms' is used to calculate the number of milliseconds to the Unix epoch. It is stored in "datetime" column with the data type of datetime64 [32].

3.2.2 Transform "dbl_v" and "long_v" into "values" column

During the data collection phase, the sensors data were recorded according to their data type, for example, "long_v" contains integer data type while "dbl_v" contains double-precision floating-point data type. However, both of the columns represent the sensor reading of each variable. Hence, data transformation is performed by the creation of a new column namely "values" to concatenate the values in both columns. Substitute the null value to 0 took place before the transformation to create new columns by using the operators.

3.2.3 Creation of "Device" column

To differentiate which data values belong to which device, the creation of "Device" column is carried out according to the respective device datasets.

3.2.4 Subsetting and filtering valid keys based on conditions

As the raw dataset consists of millions of data records, filtering is performed to filter only valid keys for the respective datasets (e.g., key 85,86,89,90,91 indicates the valid keys for indoor Device 1 dataset).

3.2.5 Create an individual column for each variable

Next, the creation of an individual column for each variable is carried out by specifying the conditions in a list and using *np.where* function [33] to subset the data based on the condition (e.g., key 86 indicates air temperature, key 93 and 89 indicate water temperature, etc.)

3.2.6 Removal of duplicate data

As we cannot reindex or resample when an index has duplicate values, *df.index.duplicated()* function is used to indicate duplicate index values and removed them.

3.2.7 Extract date from DateTime

Datetime columns without manipulation can be hard and challenging to understand by machine learning algorithms to build an ordinal relationship between the values because the dates can be represented in many formats [34]. For this project, only date is of interest for further processing. Hence, “date” value is extracted from the “datetime” columns.

3.2.8 Set DatetimeIndex

As the time series is the data of observations indexed according to time, the DateTime column is set as an index column of type *DatetimeIndex* for easier time series data manipulation. For instance, we do not always need all the data in the huge dataset, if the date format is in *DatetimeIndex* , an operation such as slice indexes can be carried out easily for thorough analysis [35].

3.2.9 Derive minimum and maximum value

To derive minimum and maximum value from the original data value, the data values are sorted and grouped by “date” , “device” and “key” variable so that maximum and minimum value of the specific date and device can be obtained.

3.2.10 Resampling

Resample is the most important step in preparing the data for time series analysis [36]. It provides the capability to increase or decrease the granularity of the time series data by changing their frequency named upsampling or downsampling. For this project, the dataset is downsampled to a lower frequency from hours and minutes to day by summarizing the higher frequency observations so that the data is available at the same frequency for further processing. As shown in Figure 3-1, the datasets were resampled to a daily frequency. As a result of making the frequency of the dateless granular, it eliminates the noise and enables the discovery of new trends and patterns.

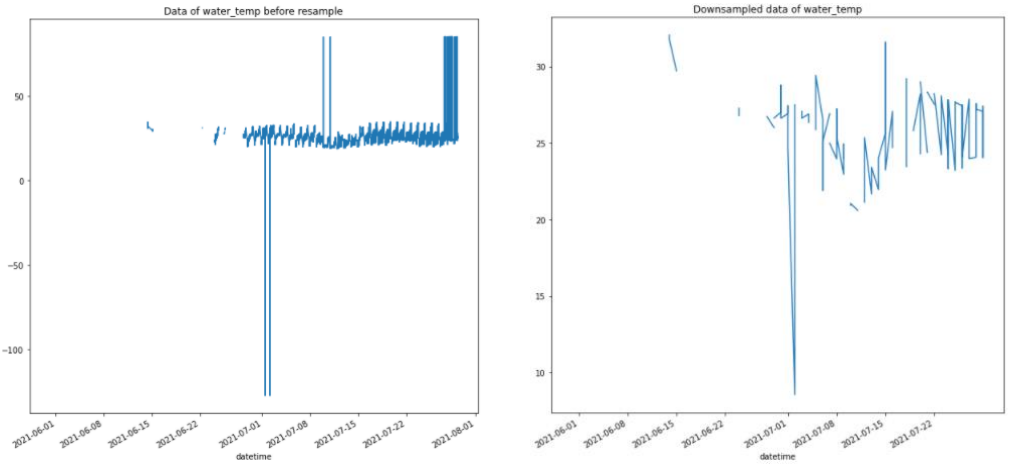


Figure 3-1: Results of Resampling

3.2.11 Deal with missing value

Removing data

Time series data does not always come perfectly clean. Due to hardware failure, the device ran out of power and communication failure, missing value is inevitable to occur during the data collection phase. It can lead to several data quality problems if it is not handled properly [37]. Some of the machine learning algorithms automatically drop the rows with missing values in the model training phase in which reduced the training size, especially in this project where the fraction missing values varies considerably across attributes.

All the variables in the datasets are significant and removing observations with missing values can produce bias in the analysis and model. Having said that, imputation will be the better option in dealing with time-series data [38].

Imputation

There are different strategies used to impute missing data which includes forward fill, backward fill and interpolation [39]. The forward filling approach propagates the last observed non-null value forward to fill the missing gaps while the backward filling approach fills the missing gaps with the next observed non-null data point.

Mean and median imputation is not a good practice to follow for time series data especially if the time series is non-stationary. Forward fill the data with historical values of that time frame will be a more sensible approach [39].

Firstly, the interpolation technique is used for imputation. It is the most commonly used and powerful method to fill missing values in the data while working with time series data because it follows special trends and seasonality. It would be more reasonable to fill missing values with the previous one or two values instead of using mean imputation techniques which caused inconsistency. For instance, the value of today's air temperature would make more sense to be filled with the mean value of the last two days instead of the whole month.

The linear interpolation method considers the distance between any two non-missing points as linearly spaced and it filled the missing value in the same increasing order from the previous data value by connecting dots in a straight line [40].

Figure 3-2 showed the density chart plotted to compare the efficiency of the techniques and choose the imputation that best fits the underlying distribution. A function is created to plot the original distribution against the distribution in which after forward fill and backward fill imputations were performed. As observed in Figure 3-2, the result showed that all three techniques work well on climate data since the differences between nearby data points are small. However, the green line has smoother and most closely resembles the black line as compared to the other two techniques.

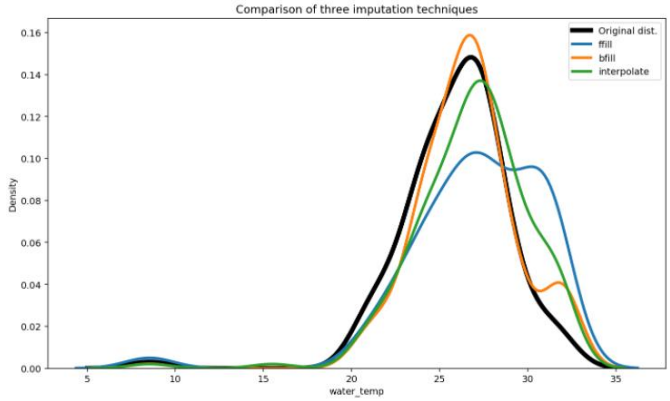


Figure 3-2: Density chart of comparison of three imputation techniques

To choose the best imputation method, visualising the result of imputing the air_temp and water_temp variable using the three methods mentioned above is plotted as shown in Figure 3-3 and Figure 3-4 respectively. It showed a smooth line connecting the missing values in the dataset and the data is essentially fitted from one point to the next by the interpolate technique. The forward fill and backward fill method showed a step-like line because both of these methods will fill the same value for consequent missing values. To conclude, interpolation techniques is chosen as the method to impute the missing value in all the datasets.

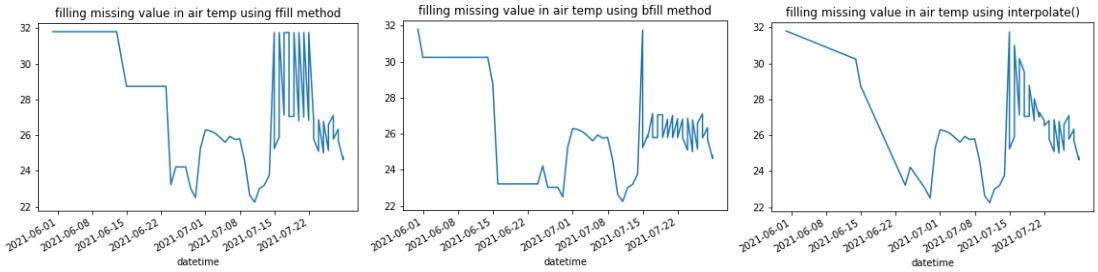


Figure 3-3: Visualisation results of the three imputation techniques of air_temp variable

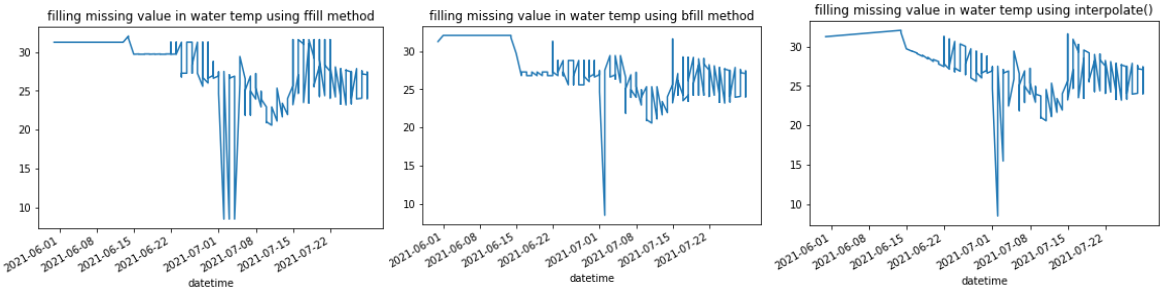


Figure 3-4: Visualisation results of the three imputation techniques of water_ temp variable

After performing the data transformation mentioned above, the datasets were subsetted into an individual dataset for each respective devices. All the cleaned datasets were extracted to the CSV files for uploading to the dashboard and for predictive modelling.

As the dataset for the variables of the indoor device is only available from 31st May 2021 to 29st July 2021, it has insufficient data size to perform time series forecasting. Thus, only the variables from Outdoor Devices 1&2: air temperature, air humidity, light intensity are used for further processing.

3.3 Data Modelling

Important constructs for building Time series forecasting

Besides basic data preparation, this section presented the important constructs and tests needed for building a robust time series forecasting model.

3.3.1 DateTimeIndex and Resampling

Pre-conditions required for the dataset are the object should be in the format of DateTime data type and DateTimeIndex. All the datasets for this project were pre-processed as required and resampled into the frequency of day as shown in Table 3-3.

3.3.2 Decomposition of time series - Trend, Seasonality, Trend and Noise

Decomposition is carried out on the dataset to deconstruct the datasets into seasonality, trend and residual (white noise). These four graph components effectively illustrate how metrics change over time and they will affect our time series analysis if present in excess [41]. The detailed analysis is presented in Section 4.

3.3.3 Stationarity

Stationary is considered an important feature in time series forecasting as a stationary time series is comparatively easy to forecast and is more reliable. The reason being this is that the forecasting models are fundamental works like linear regression where it employ the lags of the series data to form the predictors [42].

Stationary time series refers to the statistical properties of a time series (mean and variance) that are constant over time. The most basic methods for detecting stationarity rely on plotting the data for determining visually whether they present some known property of non-stationary data such as trend, seasonality, and noise. Other visualization methods which included rolling average graphs and Auto correlation (ACF) and partial Auto correlation plots (PACF) were employed as well. The results of these methods will be discussed in Section 4.b However, eyes justification may not be as accurate as of the statistical tests because human judgement and perception tend to introduce biases. Augmented Dickey Fuller tests were employed to validate the stationarity.

3.3.4 Differencing

The concept of differencing stated that time series can be made stationary by subtracting the next period of the time series value by the current time period value. Take as many differences as it takes until the time series is stationary. Making a time series stationary removes all the trend and seasonality components as well as incessant auto correlation between the predictors. Hence, regression algorithms can work better to give a more reliable prediction [43]. This transformation is done on the non-stationary data in the time series as shown in Section 4.

3.3.5 Train/Test validation split

The data is split into training and testing datasets to validate the accuracy of the model. As there is the dependence from one observation to the other, values at the rear of the dataset is used for testing. The data for the last 30 days was assigned as the training while the rest was the training data as shown in Figure 3-5. The blue colour line indicates Training data while the orange colour line indicates Testing data.

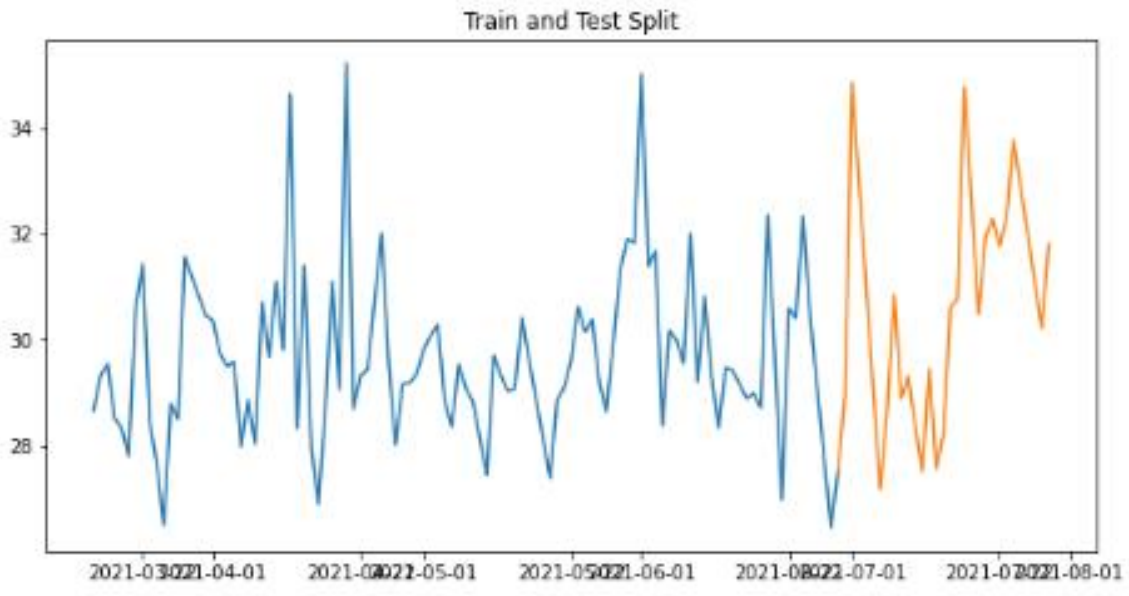


Figure 3-5: Visualisation of splitting of training and testing data

3.3.6 Determine the order of AR and MA terms

The most important part of the ARIMA forecasting will be deciding the order of the AR, I and MA model which are denoted by p,d,q respectively. Two methods were conducted as shown in below:

Method 1: Manual model orders (p,d,q) selection using ACF/PACF plots

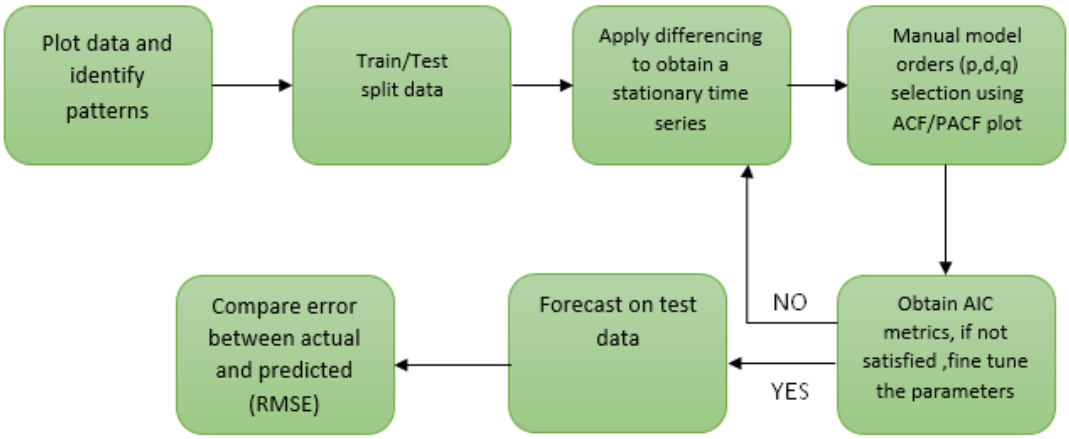


Figure 3-6: Process flow for performing manual model orders of AR and MA terms

A more detailed explanation of the process involved when performing manual selection on the model orders for AR and MA terms is depicted as below:

a. Auto Correlation

Auto correlation is a powerful analysis for modelling time series data with ARIMA models and it involves finding the correlation between a time series and a lagged version of itself. When the correlation is present, it indicates that the past values influence the current value and the impact of lags on forecasting the time series is high.

There are two usages of Auto correlation (ACF) plots in time series analysis [44]. Firstly, it is designed to understand the patterns and properties of the time series. It involves the computing of correlation coefficient to measure the magnitude of association between lagged values of a time series.

In addition, it is also used in the model identification for fitting ARIMA models. It is used to determine the optimal number of **MA(q) terms** which is the order of a moving average model. One of the ways to identify non-stationary time series is by looking for evidence that the ACF plot plummets to zero very fast and the value of r_1 is often large and positive.

b. Partial Auto correlation

Partial Correlation has the same purpose as Auto correlation**.** The only difference is that it only measures the correlation between the variable and its lagged variable after eliminating the correlation from previous lags. In other words, it removes the lags that cause Auto correlation. For example, at lag 3, partial Auto correlation removes the effect lags 1 and 2 have on computing the correlation.

As opposed to ACF plots, PACF plots are useful in identifying the optimum order of an Auto regressive model, AR (p) terms. The number of significant lags for the PACF indicates the order of the AR model. By analyzing the Auto correlation and partial Auto correlation function, an appropriate ARIMA model can be selected for time series prediction. The results of ACP and PACF plots were interpreted in Section 4 to decide the order of parameters in ARIMA.

However, this is a repetitive process to be performed multiple times to get the optimal parameters to determine the order of the model. The purpose of these parameters is to make the model fit the data as well as possible. In this project, a more effective approach involves using the Auto ARIMA modules [45] to automatically select the best order of parameters as shown in the next section. The results are interpreted in Section 4.

Method 2: Auto model orders (p,d,q) selection using Auto ARIMA

Pmdarima library

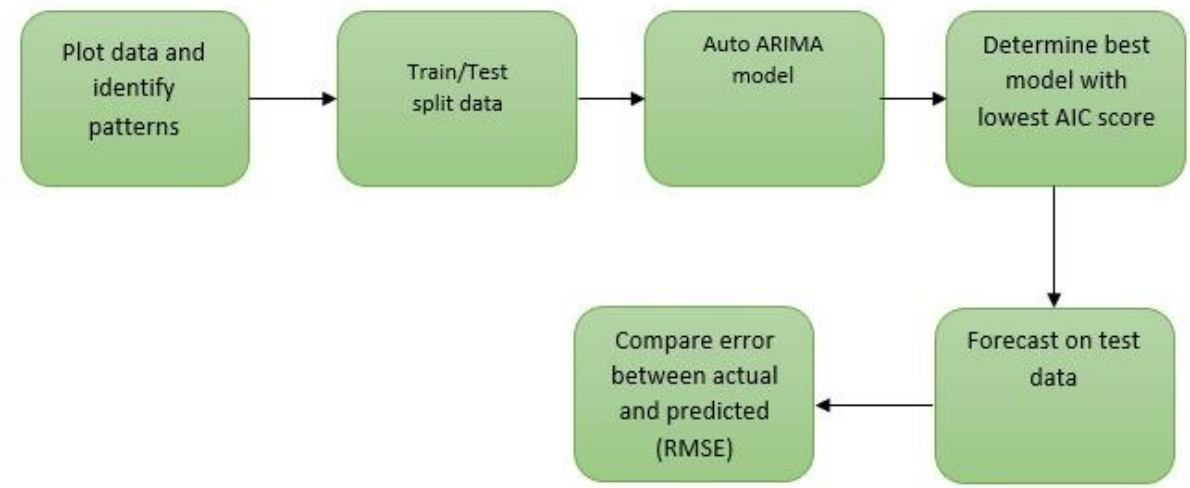


Figure 3-7: Process Flow of Auto Model Orders Selection using AUTO ARIMA

Using the previous method, determining the value for the parameters using the ACP/PACF plots can be time-consuming and less efficient [46].

Pmdarima's Auto ARIMA function is a hyperparameter tuning method that helps to identify the most optimal p,d,q (non-seasonal components) and uppercase P, D, Q (seasonal components) parameters using different combinations and return a fitted ARIMA/SARIMA model. The combinations of parameters p,d,q and P, D, Q values are selected based on lower AIC and BIC values [47].

3.3.7 Build and fit the ARIMA/ SARIMA model

ARIMA model is a widely used statistical method for time series forecasting and used to better understand a single time-dependent variable such as temperature over time. After determining the parameters (p,d,q) (P, D, Q) of the ARIMA/SARIMA model, it is fitted on the training set to predict the test set and see how well it performed.

Model validation is conducted to compare the predicted results to the actual values in the test validation data sample. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) will be used to select the order of parameters for the model. AIC is a measure of the goodness of fit of any statistical and it will penalize complex models in favour of simple ones. Unlike AIC, the BIC penalizes free parameters more strongly. Parameters order with the lowest AIC and BIC score value will be selected [48].

3.4 Model Evaluation

Root Mean Square Error (RMSE) [49]is the squared root of mean squared error, which is a metric that evaluates the quality of a forecasting model or predictor. It takes account into the variance and bias of the data. It is a good measure of how accurately the model predicts future values. The lower the value of RMSE, the higher the accuracy of the model. The RMSE is the same unit as the projected value, which makes it easier to comprehend as compared to other metrics.

Making Future Predictions

Finally, the trained model is used to predict the value for the next 30/31 days from the end date. The result is shown and interpreted in Section 4.

3.5 Data Visualisation

Methodology in Designing Dashboard [50]

3.5.1 Determine the User/audience and their Intended Usage of the Dashboard

The main users of this dashboard will be the person from the farm team. The users are expected to visualize the sensor data and crop yield information through this dashboard so that they can get an insight into the climatic information (temperature, humidity, light intensity) , water properties (pH, EC value) and electric consumption information.

3.5.2 Selection of Key Data Points

There is mainly two dashboard design type used to visualize the sensor and yield data respectively. Our main purpose is to design a summary operational dashboard to give an overview of the performance trends and potential problems of the sensor data. This dashboard displayed critical sensor data information that is time relevant which included: air temperature, air humidity, water temperature, light intensity, pH and EC value of water and electric current consumption for LED lights and water pump in the farm.

In addition, a strategic dashboard is designed to enable users to track the crop yield information and how they are performing against their strategic goals via KPIs. There are a total of six-page of visualization in the dashboard in which represents different key parameters to monitor and analyse by the users.

The first page consists of air temperature and air humidity. By placing these two parameters together in one visualization, it can provide more insight to the user as there is a correlation between them [51].

The second page consists of the dashboard of water temperature, pH and EC. Three of these parameters are related to water properties so they are visualized together.

The third page consists of the dashboard of electric current of LED lights (indoor), water pump and light intensity. This page is focused on the electric current usage by each device and key.

The fourth page is the yield analysis dashboard that summarizes the data related to the yield. It provides the user with an overview of the actual yield harvested from different types of vegetables.

The fifth page is the yield KPI analysis dashboard that shows the key performance indicator (KPI) of Estimated Yield Plants vs Actual Harvested Plants and KPI of Actual Yield (kg) vs Expected Yield (kg). Both of these visualizations allowed users to benchmark the performance of the yield growth and see the bigger picture of the overall performance.

3.5.3 Selection of Proper Filters and Drill-downs

To keep the group particular interests together and information organized, slicers and buttons were used to create a filter as shown in Figure 3-8. Users can filter the dashboard based on date and location (indoor or outdoor). Users can navigate to the previous and next pages using the left and right navigation arrows.

This will help to keep all the information uniform across all the visualization on the same page. Users can drill down the date from year to quarter, month and day for a further breakdown of information.

Slicer is used to filter the date as it enables an interactive interface for quick filtering of the content desired. It also indicated the current filtering state which allows users to be aware of what exactly is currently displayed. As compared to filter, slicer has more options that allowed the user to filter the date through advanced date slicers like a relative date (today, yesterday, last 2 weeks/years, etc.), between, before or after certain dates.

To keep the dashboard consistent, the slicer synchronization [52] allows the same date filter to be applied across the visualization on the other page. The slicer can also use to set up to affect only certain visuals but continue to filter out the others.

As the datasets are displayed on the dashboard based on key and device, a button is used to navigate the user to the specific device indoor or outdoor as shown in the figure. By using the button to subset the data into indoor and outdoor devices, the user will not be overwhelmed by so many visualizations on the same page. The button feature helped in maintaining a clean and simple look.

Figure 3-8: Filter and Drill-downs of Air Temperature vs. Air Humidity Dashboard

While in the crop yield dashboard, the dataset will be filtered by batch number, date of harvest, types of vegetables, year, month and relative date of harvest as shown in Figure 3-9.

Figure 3-9: Filter and Drill-downs of Yield Analysis Dashboard

3.5.4 Selection of Color Design

Use colour to provide context [53]

The colour scheme chosen for this dashboard is green colour as it provides the context of the greenish farm design. Furthermore, the green colour is chosen as it aligned with the farm organization's logo and core concept which helped to indicate that the dashboard is linked to the organization and give the audience a better sense of trust.

Use colour to connect information

In the crop yield analysis dashboard, the colour is used as indicators to connect information from one visual to another. It helps to connect two bits of information while displayed on different elements as shown in Figure 3-10. For example, the colour used in the bar chart is the same as those used on the pie chart which is displaying the same information to provide consistency. It also served as a legend for the pie chart.

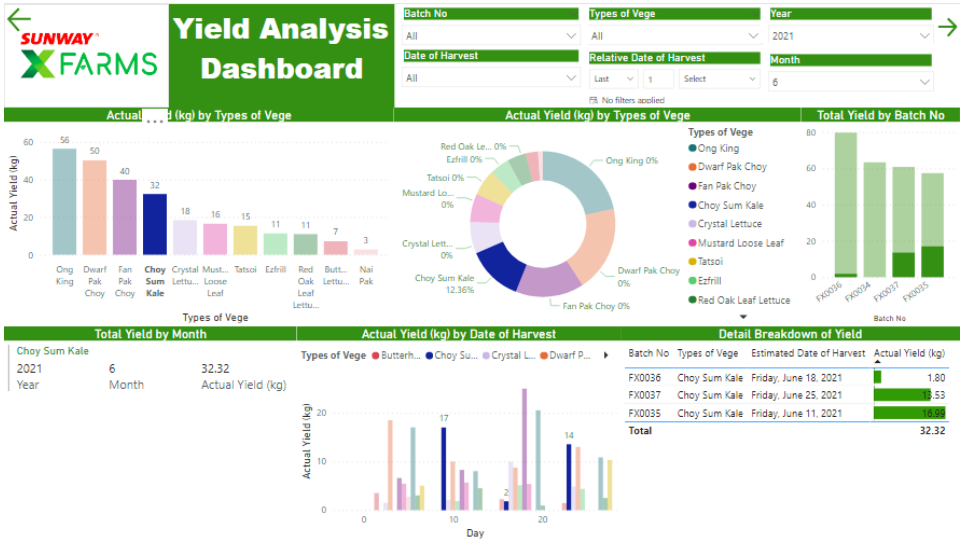


Figure 3-10: Color Design of Dashboard

Use colour to draw attention

The use of two contrasting colours– blue and red enabled users to pay attention to the deviation from the target goals. As shown in Figure 3-11, Users are likely going to look at the red element first because it stands out from the neutral tones. Red is used to indicate how much the actual harvested yield is behind their target yield. Green is not suitable to be used in this visualization as green and red can look the same to a person who has colour blindness [54].

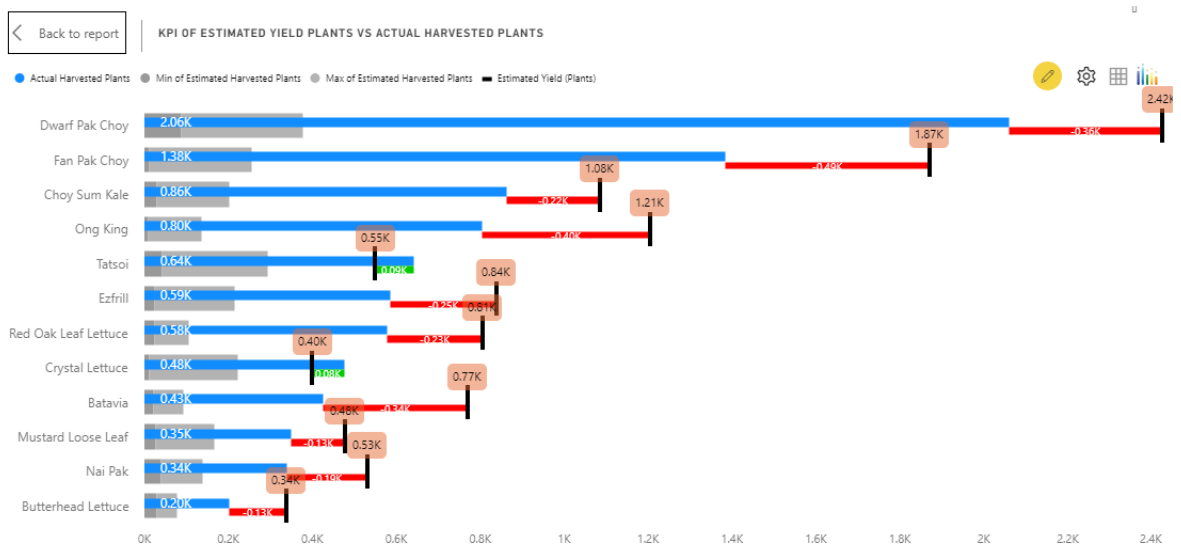


Figure 3-11: Color Design of Dashboard

3.5.5 Selection of Charts

3.5.5.1 Sensor Dashboard

Bar Charts and Line Charts

Lines charts are mainly used for time series data, as the main focus is on trends over time. However, it would be too messy and confused for a line chart to work well if more than two series are being displayed. A combination chart is chosen to visualize the air temperature, water temperature, air humidity, electric current and light intensity data which consists of multiple series represented by a line and column chart in the visualization [55].

As shown in Figure 3-12, the temperatures and humidity values are normally summarized by the average, minimum and maximum values. For instance, users might be interested in normal temperature and humidity for June as well as the highest and lowest value during June. 7-day moving average value provides additional information about the average value for the past week.

While the bar chart displayed average data value, line charts are used to display the minimum, maximum and 7-day moving average data value. They are being coloured separately to represent different values. Some of the charts have two rows representing different keys from the same device.

By combining these two charts in the same visualization, users can compare values in different categories, since the combination gives a clear view of which value is higher or lower over time.

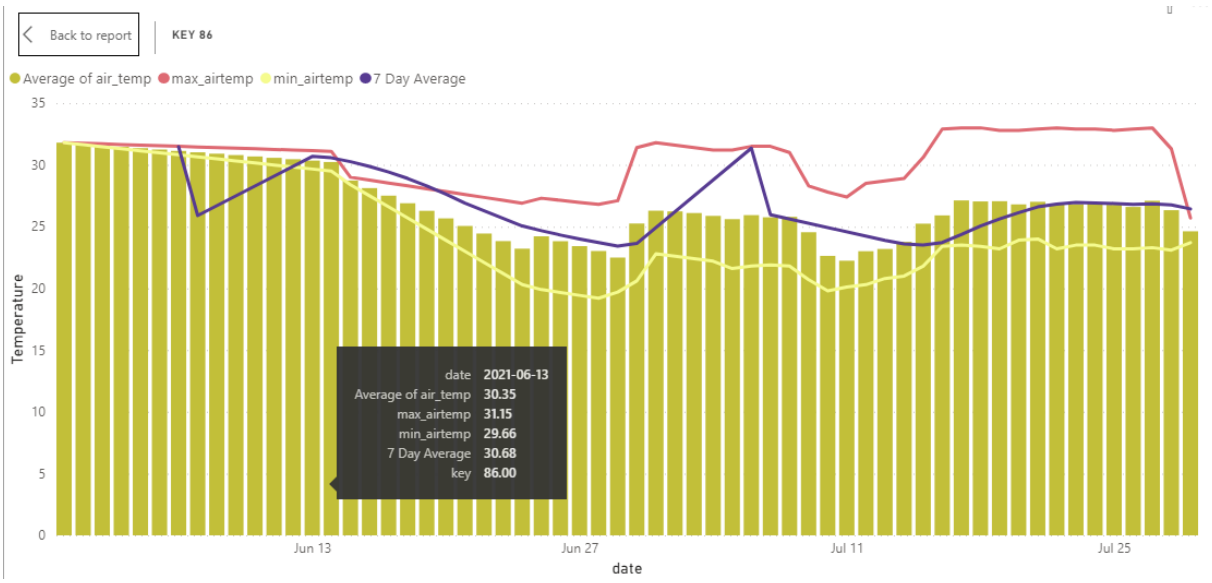


Figure 3-12: Combination charts in Sensors Dashboard

Gauge Chart

To highlight some interesting climatic facts, a gauge chart is used to show the average temperature, minimum and maximum value over the period as shown in Figure 3-13. This can provide a one-time glance at the value so that the user can quickly understand the current status. These gauge charts showed the average value for the date period selected. The user is unable to get this information through the line charts as line charts only show the average value of an individual day.

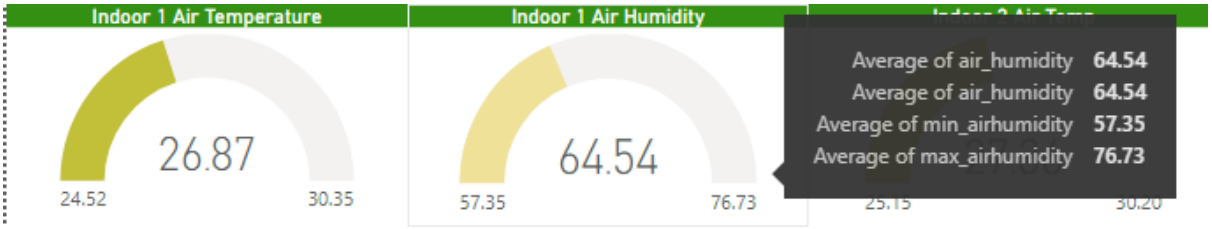


Figure 3-13: Selection of Gauge Charts in Sensors Dashboard

As shown in Figure 3-14, the pH value and electric conductivity (EC) value of water are visualized by line chart and the user can easily spot the trend and track the changes over time. The date is used as the X-axis to give the user an impression of the moving value across the period. Again, the gauge chart enabled the user to know the average value for the selected period.

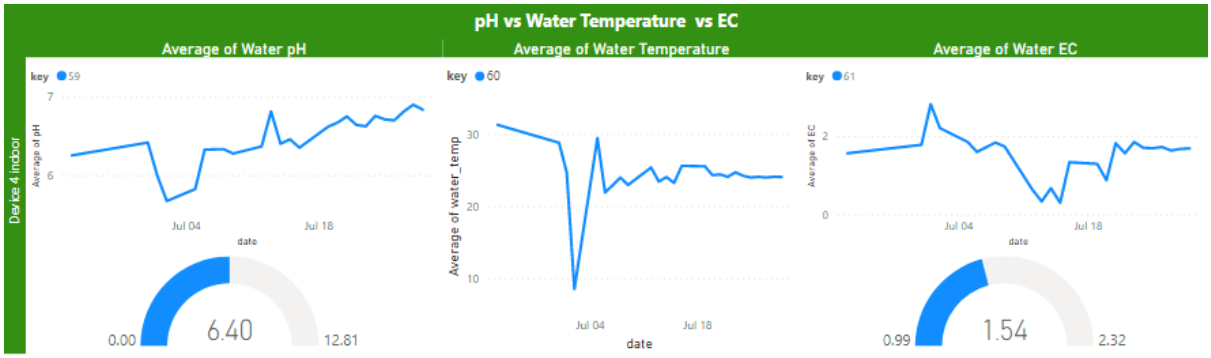


Figure 3-14: Selection of Line Charts in Sensors Dashboard

3.5.5.2 Yield Analysis Dashboard

Bar Graph

Bar graphs are suitable to visualize categorical variables and perform a comparison of values across different subgroups of data. As shown in Figure 3-15, it is used to compare the actual yield between different types of vegetables and total yield by batch number.

Besides, the pie chart is used to represent the percentage and weight of components of the type of vegetables proportional to the overall distribution.

Figure 3-15: Selection of Bar Charts in Yield Analysis Dashboard

Tables

The table is used in the yield analysis dashboard to provide a detailed breakdown of data as shown in Figure 3-16. It allows the user to manipulate the display of data such as sorting and exporting to excel files for reporting purposes. As observed in Figure 3-16, the data bars in the tables enabled the user to understand the context quickly while having the breakdown of details. From the graph, if the user is interested in how much is the actual yield for each type of vegetables harvested on June 25, this table clearly showed the details of it. The largest yield is *Choy Sum Kale* with 13.53 kg while the least yield is *Red Oak Leaf Lettuce* with only 2.50 kg.

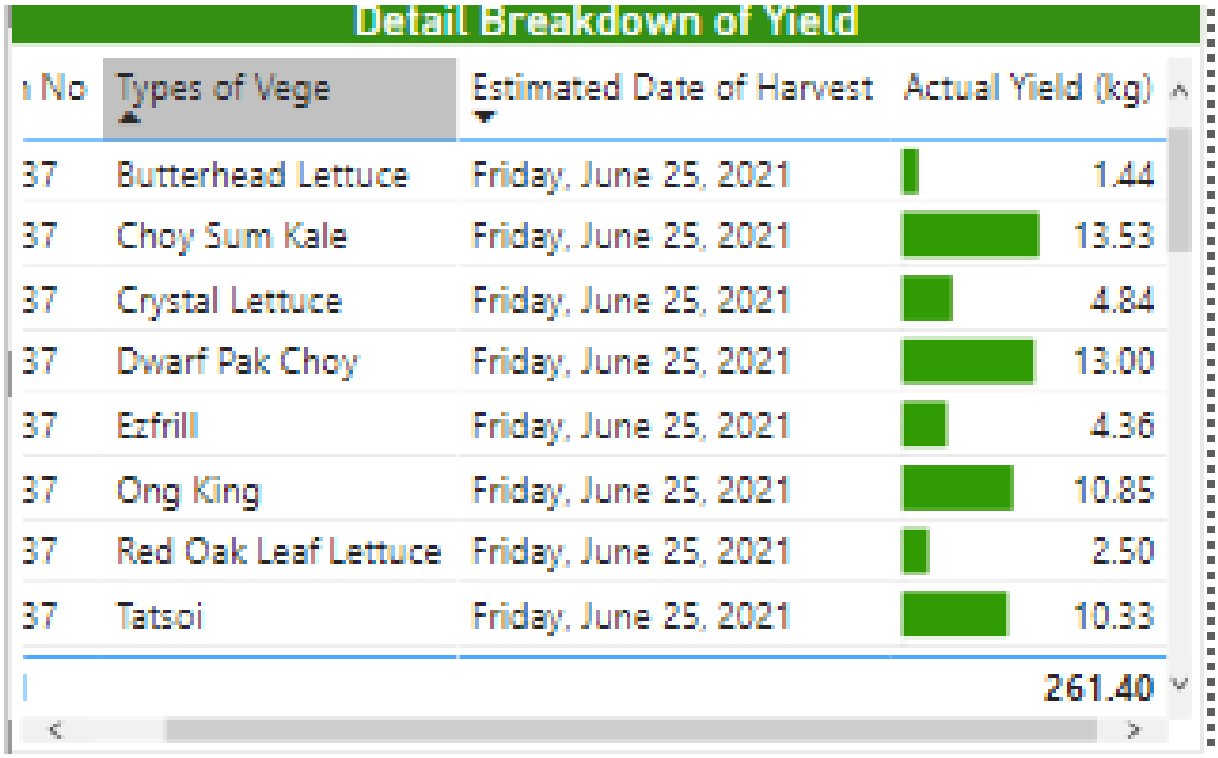


Figure 3-16: Selection of Tables in Yield Analysis Dashboard

Bullet chart

As shown in Figure 3-17, a bullet chart is used to measure the performance of the crop growth by visualizing the estimated yield plants and actual harvested plants. It is a variation of a bar graph but designed to address some of the problems that gauge have [56]. As there are different target values for each type of vegetable, a bar chart is unable to create the desired visual. Besides, it provided a summary table (Figure 3-18) that is clear and precise to understand easily without the need to create another table manually.

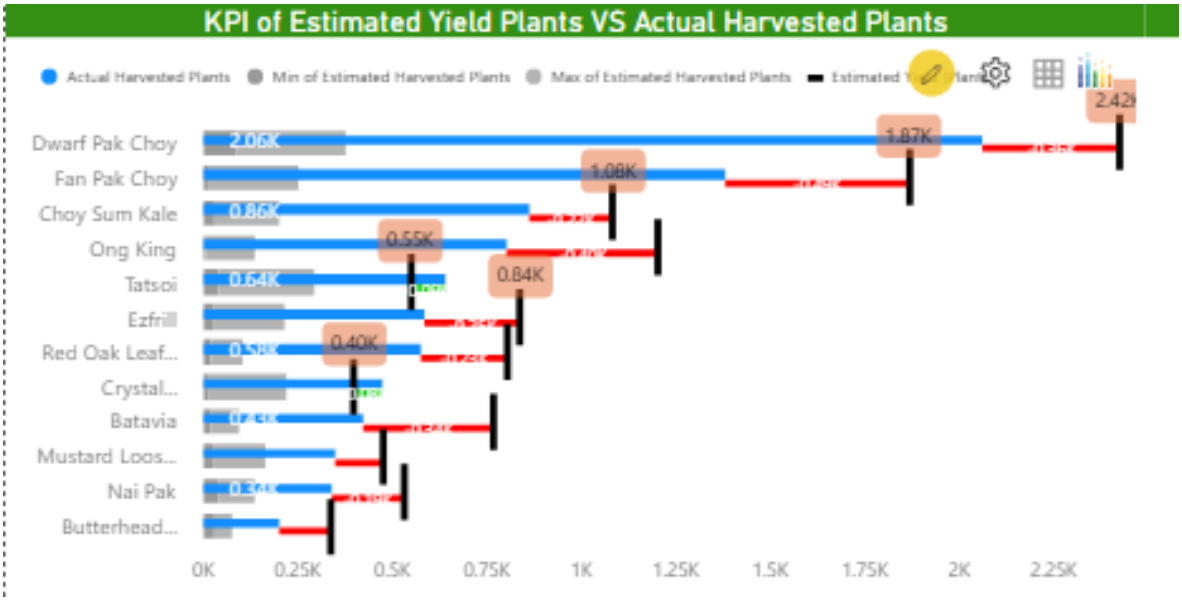


Figure 3-17: Selection of Bullet Charts in Yield Analysis Dashboard

Types of Vege	Actual Harves...	Estimated Yiel...	Min of Estima...	Max of Estima...
Dwarf Pak Choy	3,192.69	3,770.00	88.50	290.18
Fan Pak Choy	1,828.15	2,580.00	11.90	244.08
Red Oak Leaf Lett...	1,492.49	1,515.50	24.36	222.69
Ong King	1,429.35	2,035.00	10.18	153.31
Butterhead Lettuce	886.56	817.00	31.22	195.04
Choy Sum Kale	863.83	1,085.00	31.02	171.78
Tatsoi	843.50	810.00	42.15	251.57
Nai Pak	744.46	786.00	40.88	221.82
Ezfrill	697.03	1,159.00	25.85	189.57
Crystal Lettuce	477.89	400.00	12.05	210.97
Batavia	426.85	770.00	23.76	71.14
Mustard Loose Leaf	351.97	478.00	28.57	139.00

Figure 3-18: Summary Tables in Yield Analysis Dashboard

Next, Tornado charts (Figure 3-19) were used to compare the KPI of estimated yield plants vs actual yield plants. The categories are ordered in such a way that the largest bar appears at the top which allows the user to get the insight quickly.

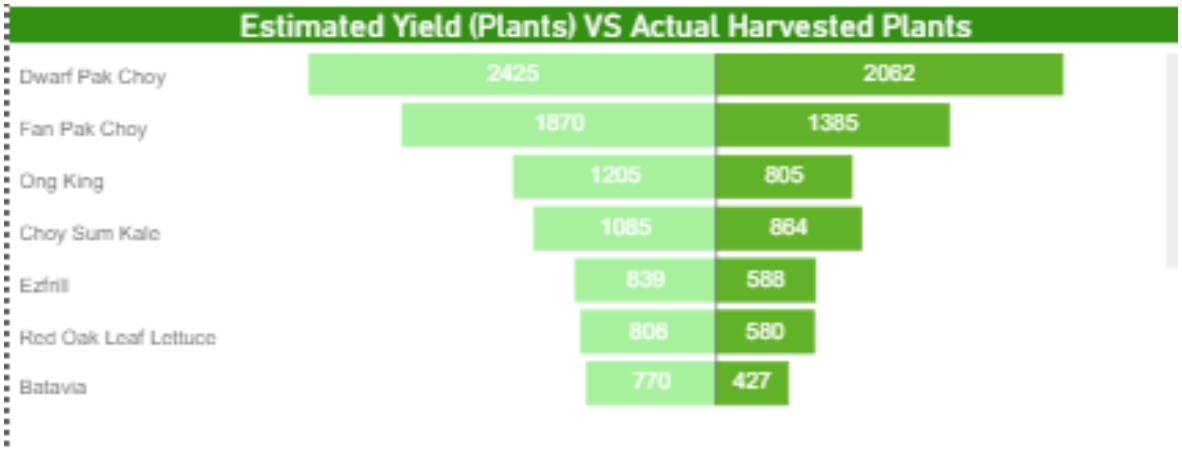


Figure 3-19: Selection of Tornado Charts in Yield Analysis Dashboard

The screenshots of the full dashboard design will be attached in the Appendix. Detail demonstration of the dashboard will be conducted during the presentation.

4 Results and Discussions

4.1 Results of Exploratory Data Analysis

Before proceeding for further analysis, data exploration was carried out by generating their summary statistics which provides quick insight into the characteristics of the data. Box plots and histograms are plotted to visualize the data distribution and display the range of the data for identifying any potential outliers. It shows where the data points are dense and where they are sparse in one dimension. Density plots are plotted to display where the mass of the data is located and it allows for smoother distribution by dampening the effect of noise and outliers [57].

4.1.1 Average Air Temperature Readings Collected From Indoor and Outdoor Devices

To visualize the data distribution and variance of each variable, histogram and boxplot were plotted as shown in Figure 4-1. Visualizing the data distribution helped to determine whether the data points are dense or spread out and variance is low or high. High variance features tend to contribute more to the prediction of the outcome variable.

As shown in Table 4-1, there is a total of 75 data observations in Indoor Device 1 and Indoor Device 2. Both of the devices have an average temperature of 27 °C with a minimum value of 22.3°C and a maximum value of 31.8°C.

Average Air Temperature Readings Collected From Indoor Devices

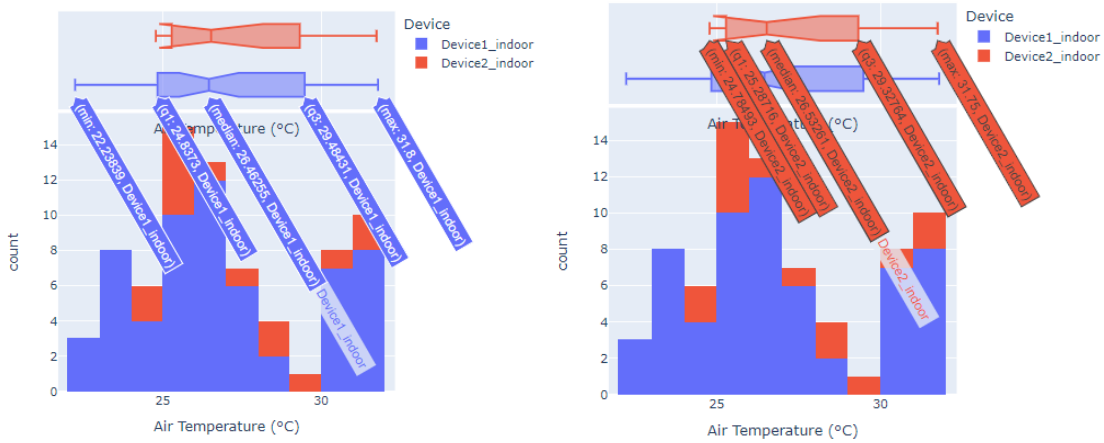


Figure 4-1*: Distribution of Air Temperature Readings collected from Indoor Device1 and Device2*

Table 4-1: *Average Air Temperature Readings collected from Indoor Device 1 and Device 2*

	air_temp	min_airtemp	max_airtemp
count	75.000000	75.000000	75.000000
mean	26.969019	24.650000	30.322667
std	2.717244	3.893514	2.032428
min	22.238388	19.200000	25.700000
25%	25.120508	21.800000	28.616667
50%	26.532608	23.400000	31.200000
75%	29.141304	28.325000	31.606250
max	31.800000	31.800000	33.000000

Average Air Temperature Readings Collected From Outdoor Devices

More data are being collected by the outdoor Device 1&2 because it is the first device that the farm set up to collect the air temperature reading. The distribution of air temperature collected is illustrated in Figure 4-2. As observed in Table 4-2, 250 data observations were captured and the mean air temperature is 29.87 °C , which is much higher than the air temperature in the indoor device.

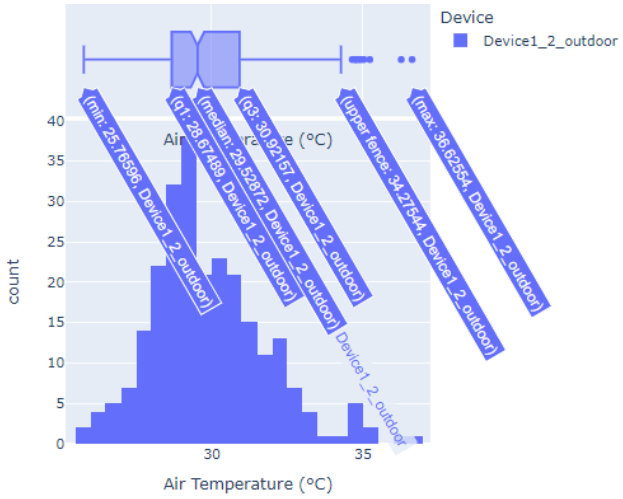


Figure 4-2: Distribution of Air Temperature Readings collected from Outdoor Device 1&2

Table 4-2: Distribution of Air Temperature Readings collected from Outdoor Device 1 & 2

	air_temp	min_airtemp	max_airtemp
count	250.000000	250.000000	250.000000
mean	29.899594	25.112400	38.789800
std	1.918810	2.427289	3.830306
min	25.765957	20.000000	27.000000
25%	28.676586	23.000000	37.000000
50%	29.528716	25.800000	39.000000
75%	30.920879	26.675000	41.950000
max	36.625544	33.000000	45.800000

Furthermore, Figure 4-3 shows a few outliers are discovered in Outdoor Device 1&2. Readings above 34.64 °C are considered as outliers. However, as the aim of this project is to explore and monitor the temperature reading, the observed outlier can be informative about the subject area of our study such as information about the variability inherent in the sensor data. It may cause by unusual conditions and natural variation, for example, power failure or sensor setting drifting off the standard value so the extreme values is a legitimate observation and natural part of the study that should not be removed [58].

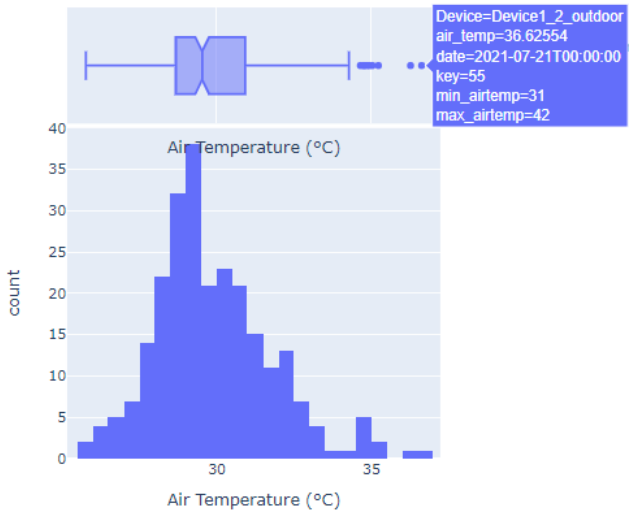


Figure 4-3: Outliers observed from Outdoor Device 1 and Device 2

Moreover, Figure 4-3 shows the distribution of air temperature readings reached its peak with 38 data observations having temperature values ranging between 29 °C to 29.5 °C.

As observed in Figure 4-4, the density plot shows the distribution of air temperature reading in the indoor device is lower than in the outdoor device. There is a peak in the distribution of the indoor devices, the average air temperature readings are range from 26 °C to 28°C. Meanwhile, outdoor devices have average air temperature readings of 30 °C to 32 °C.

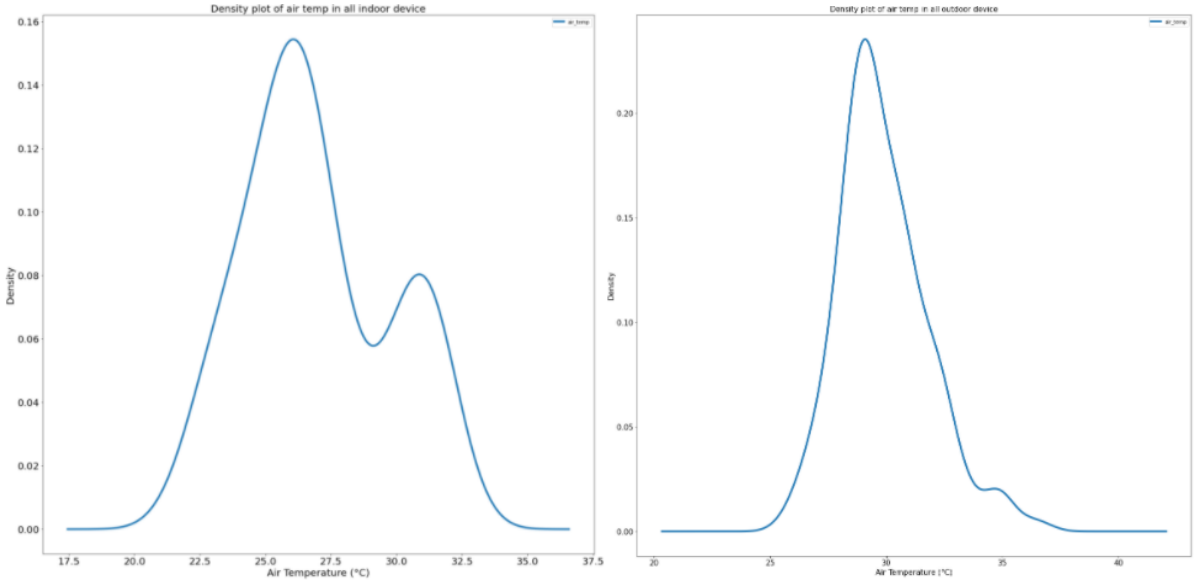


Figure 4-4: Density Plot of Air Temperature Readings Collected from Indoor and Outdoor Devices

Figure 4-5 illustrated the overview of air temperature readings collected from indoor and outdoor devices.

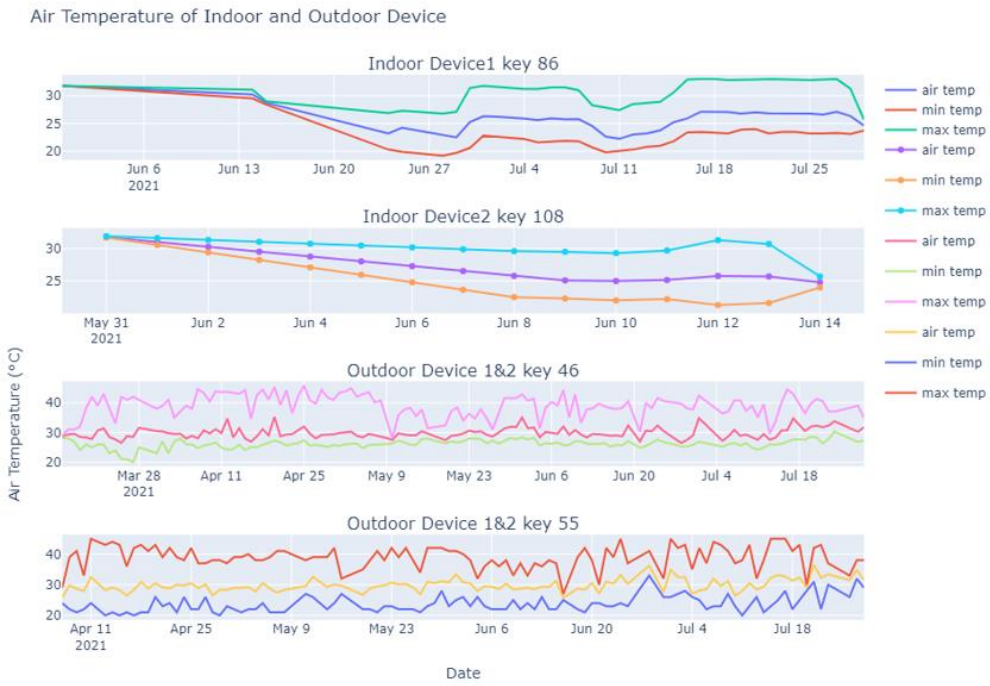


Figure 4-5: Overview of Average Air Temperature Readings Collected from Indoor and Outdoor Devices

4.1.2 Average Air Humidity Readings Collected From Indoor and Outdoor Devices

Average Air Humidity Readings Collected From Indoor Devices

As shown in Table 4-3, there are a total of 75 data observations in Indoor Device 1 and Indoor Device 2 . Both of the devices have average air humidity readings of 63 g/m3 with minimum value of 52 g/m3 and maximum value of 82 g/m3.

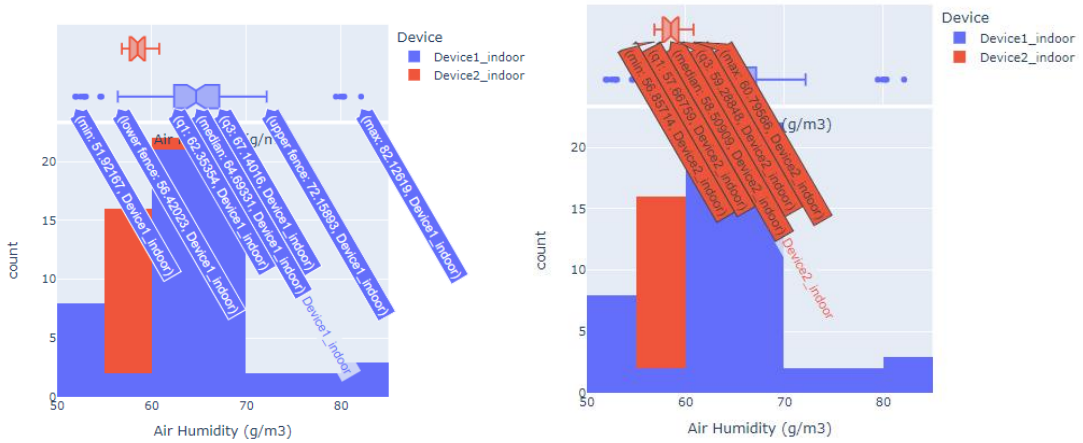


Figure 4-6: Distribution of Air Humidity Readings Collected from Indoor Device1 and Device2

Table 4-3: Distribution of Air Humidity Readings collected from Indoor Device1 and Device2

	air_humidity	min_airhumidity	max_airhumidity
count	75.000000	75.000000	75.000000
mean	63.340699	56.400000	75.340000
std	6.566656	7.118168	10.581032
min	51.921674	44.000000	59.000000
25%	58.703279	51.833333	67.000000
50%	63.361483	55.000000	73.000000
75%	66.228007	62.357143	80.500000
max	82.126193	73.000000	95.000000

Average Air Humidity Readings Collected From Outdoor Devices

As observed in Figure 4-7, a total number of 41 data observations has a humidity level of 40-49.99 g/m3, followed by 38 data observations with humidity level of 70-79.9 g/m3. 36 data observations have humidity levels of 30-39.99 g/m3.

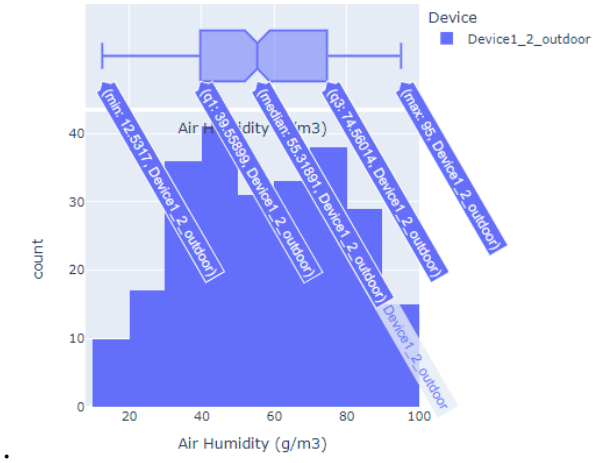


Figure 4-7: Distribution of Air Humidity Readings Collected From Outdoor Devices

As observed from Table 4-4, the outdoor devices have 250 data observations with an average humidity of 57.19 g/m3.

As compared to the indoor device's data, outdoor device data tend to have high variability of air humidity reading. Table 4-4 shows the outdoor devices have minimum value of 12 g/m3 and maximum value of 95 g/m3 and show a high standard deviation of value 21 g/m3.

Table 4-4: Distribution of Air Humidity Readings collected from Outdoor Device1&2

	air_humidity	min_airhumidity	max_airhumidity
count	250.000000	250.000000	250.000000
mean	57.191839	28.910000	72.840000
std	21.616284	20.197134	19.025987
min	12.531700	8.000000	16.000000
25%	39.711981	12.000000	62.000000
50%	55.318905	18.500000	69.000000
75%	74.535604	43.000000	91.187500
max	95.000000	95.000000	95.000000

Moreover, the density plot in Figure 4-8 shows that most of the distribution (43 out of 75 data observations) of air humidity reading in indoor device range between 60 to 70 g/m3. Whereas in outdoor device, it has almost equal distribution of air humidity range from 30 to 50 g/m3 and 60 to 80 g/m3. 41 data observations have the humidity level of 40-49.9 g/m3 , 38 data observations have the humidity level of 70-79.9 g/m3 and 36 data observations have the humidity level of 30-39.9 g/m3.

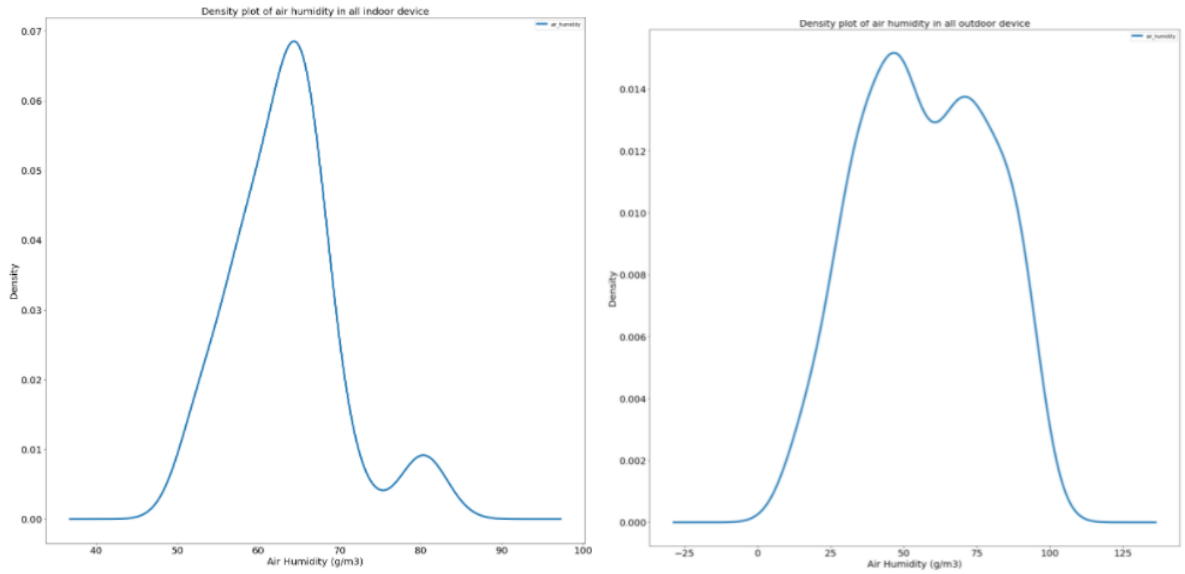


Figure 4-8: Density Plot of Air Humidity Readings Collected from Indoor and Outdoor Devices

Figure 4-9 illustrated the overview of air humidity readings collected from indoor and outdoor devices.

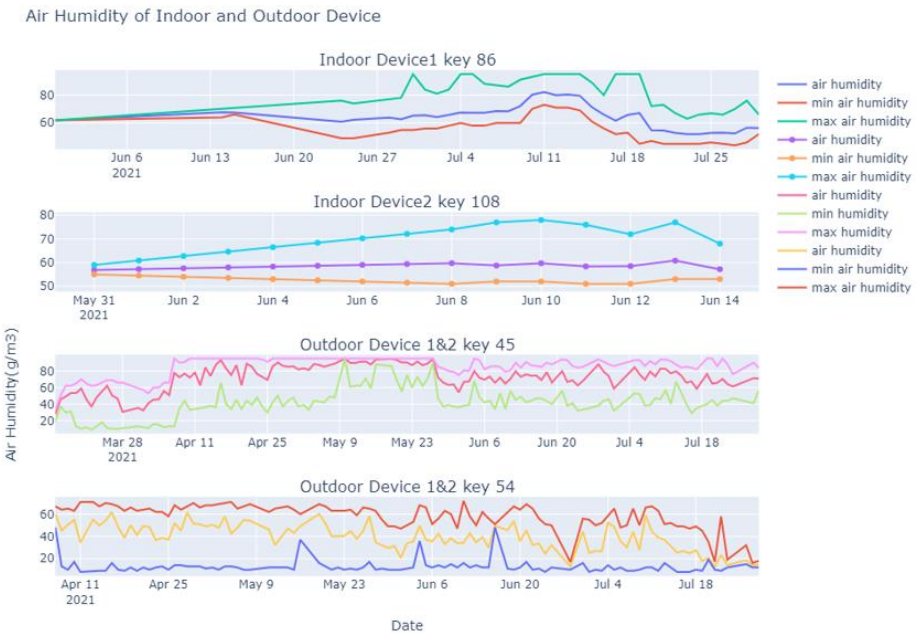


Figure 4-9: Overview of Average Air Humidity Readings Collected from Indoor and Outdoor Devices

4.1.3 Average Light Intensity Readings Collected From Indoor and Outdoor Devices

LUX is the number of lumens that plants receive from a light source in a specific period. Different types of plants require different level of light intensity, so by looking at the light intensity measurement, the farm will be able to locate the plant at a more appropriate location and place additional light if necessary.

Table 4-5 shows that the average light intensity of Outdoor Device 1&2 is 8337 LUX. It has a minimum value of 74 LUX and a maximum value of 34301 LUX. Furthermore, 121 data observations have light lux reading between 0-4990 LUX, followed by 113 data observations having light LUX of 5000 up to 9900 LUX and 139 data observations with light lux value of 10000 LUX to 20000 LUX. Only 4 data points have a value beyond 20000 LUX and 1 data point between 30000-35000 LUX.

Research [59] showed that light intensity of 2000 up to 4000 LUX is ideal for plants that require good and consistent light intensity while 5000 and up LUX is appropriate for plants that require a lot of strong light. Full sun outdoors is about 10000 to 12000 LUX. In this scenario, it meets our expectations as the device is placed outdoor and the light intensity will change throughout the day.

Table 4-5: Distribution of Light Intensity Readings collected from Outdoor Device1&2

	light_LUX	min_light_LUX	max_light_LUX
count	378.000000	378.000000	378.000000
mean	8337.156820	454.884478	35104.276669
std	5264.476214	1029.115860	19336.798611
min	74.153958	1.854839	244.838700
25%	3713.568044	1.854839	20988.192963
50%	8007.488921	1.854839	36007.367143
75%	12088.099700	183.540706	48535.100000
max	34301.057481	5720.323000	87492.740000

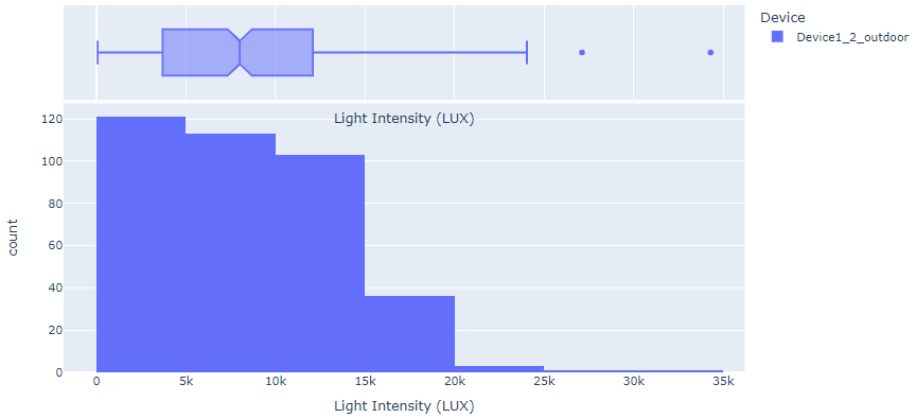


Figure 4-10: Average Light Intensity Readings Collected From Outdoor Device 1&2

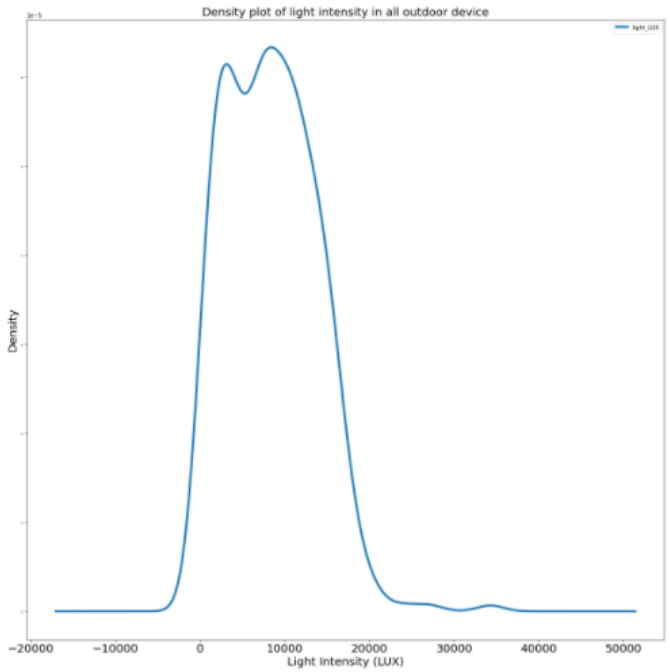
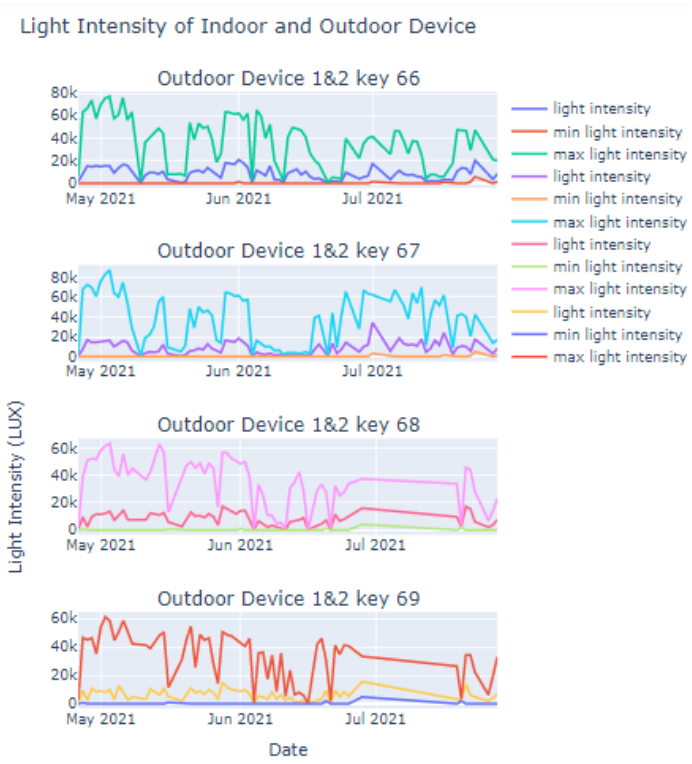


Figure 4-11: Density Plot of Light Intensity Readings Collected from Outdoor Device 1&2

Figure below illustrated the overview of air humidity readings collected from indoor and outdoor devices.



4.1.4 Correlation Between Air Temperature and Air Humidity

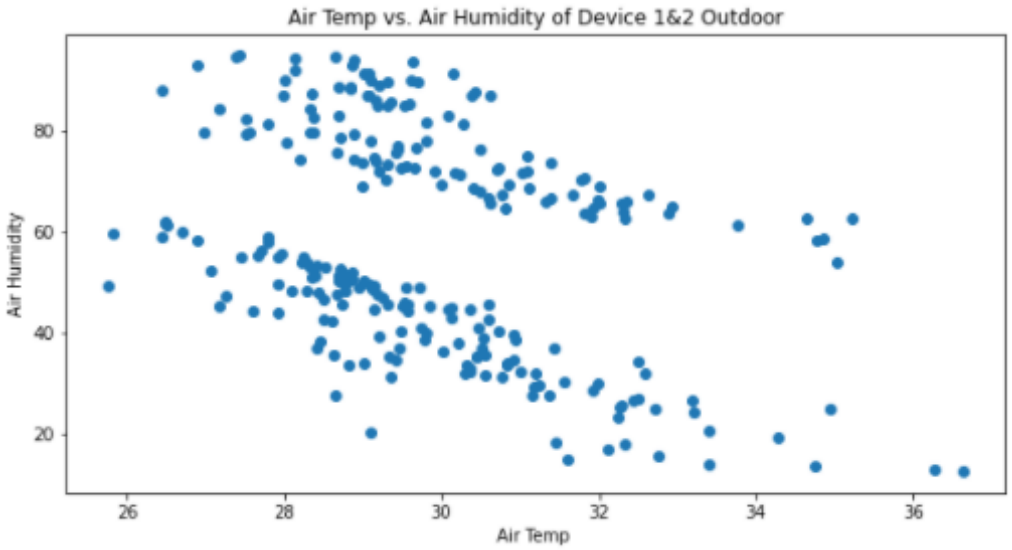


Figure 4-12: Correlation Plot Between Air Temperature and Air Humidity

As observed from the correlation plot shown in Figure 4-12, there is a negative correlation between air temp and air humidity. Air humidity tends to decrease when air temperature increases.

4.2 Results Interpretation of Time Series Analysis

4.2.1 Results of Trend, Seasonality and Noise

The trend shows a general direction of time series data over a period , representing an upward or downward slope. Seasonality refers to whether the data have a cyclical or periodic pattern with respect to timing, magnitude and direction. White noises are the outliers or missing values that are not consistent with the rest of the data and fluctuations in time-series data that is erratic, random and unpredictable.

i. Plot the time series data

The most basic methods for detecting the stationarity of the predictors rely on plotting the data. The visualisation plots help to detect the presence of some known properties of non-stationary data such as trend, seasonality, and noise. The line graph is being plotted for each variable for better visualization of the data. The trends and overview of air temperature, air humidity and light intensity is plotted.

Several heuristics were used to rule out stationarity in the plots in Figure 4-13.The prominent seasonality can be observed in Time Series Plot E while the others time series plot looks like no clear pattern of seasonality. The magnitude of Time Series Plot E changed repeatedly, showing the changes almost similar for different time intervals.

Moreover, noticeable trends and changing levels can be seen in Time Series Plot C and Plot D. This leaves Time Series Plot A (Outdoor Device 1&2 Air Temperature) and B (Outdoor Device 1&2 Light Intensity) as the only stationary series. This justification was further supported by [60] as he mentioned that a stationary time series will show the series to be roughly horizontal (although some cyclic behaviour is possible), with constant variance. In general, a stationary time series will have no predictable patterns in the long term.

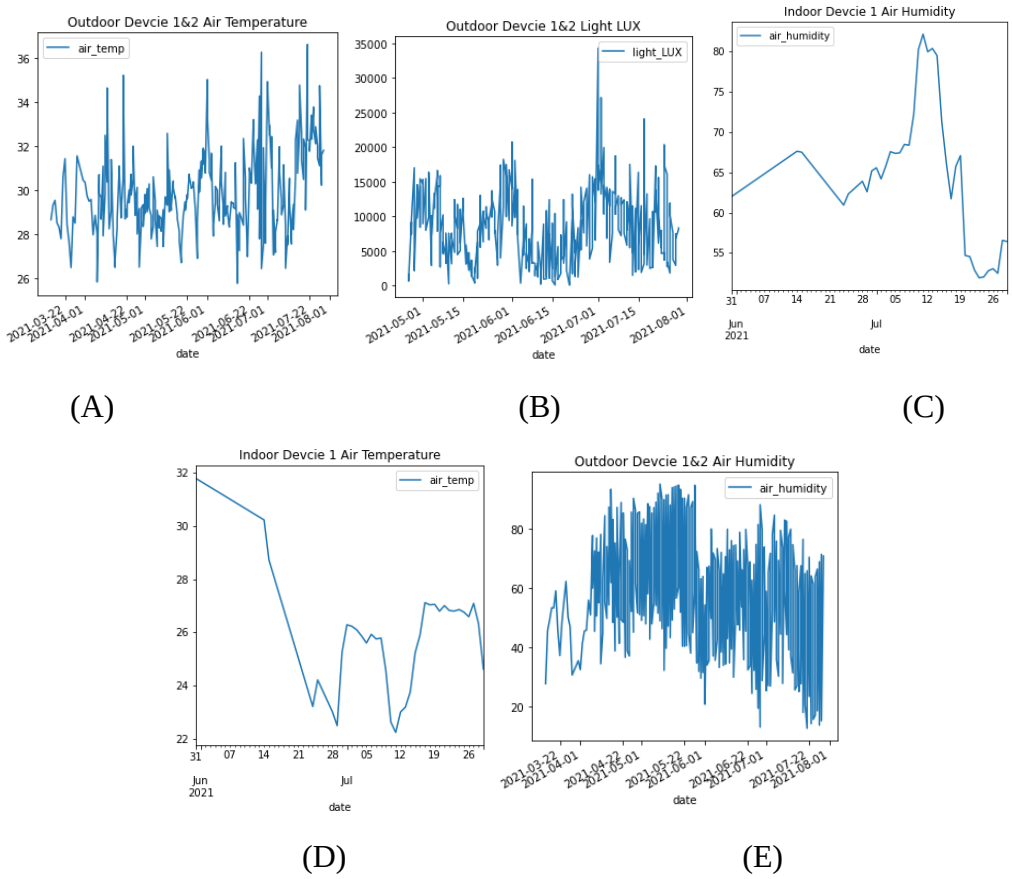
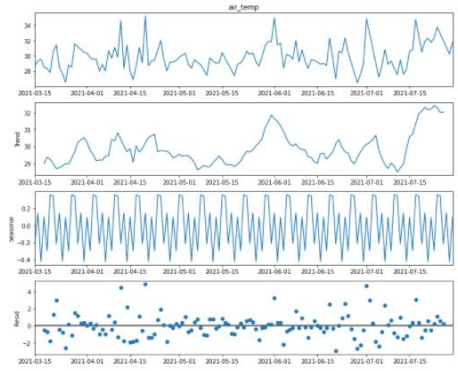
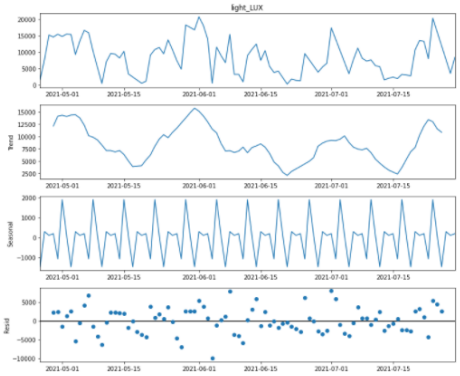


Figure 4-13: Time Series Plot A-E

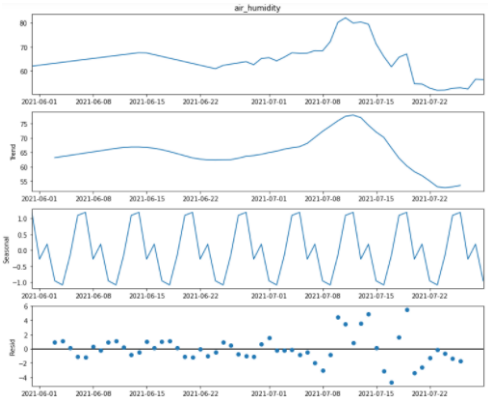
Furthermore, the stationarity of the predictors can be further checked through the decomposition of the Time Series Plot A-E in Figure 4-14. The trends and seasonality exist clearly in Time Series Plot D1 (Indoor Device 1 Air Temperture) and E1(Outdoor 1&2 Air Humidity). D1 has downward trends while E1 has upward trends. Both of them show a clear seasonality pattern in the visualization.



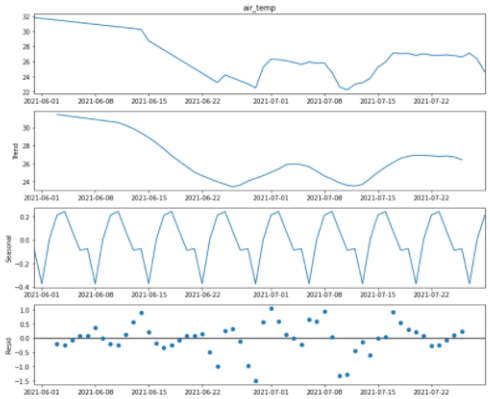
Outdoor Device 1&2 Air Temperature (A1)



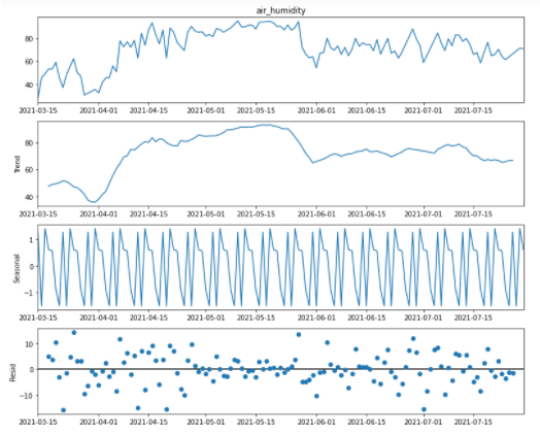
Outdoor 1&2 Light LUX (A2)



Indoor Device 1 Air Humidity(C1)



Indoor Device 1 Air Temperture (D1)



Outdoor 1&2 Air Humidity (E1)

Figure 4-14: Decomposition of Time Series Plot A-E

ii. Rolling means

Another way of determining whether a given time series is stationary is by plotting the rolling average to check the mean and variance of the time series. Besides, a rolling window of 7 is created to obtain the weekly moving average temperature. Figure 4-15 to 4-19 below showed the 7 days rolling mean and standard deviation of the variable.

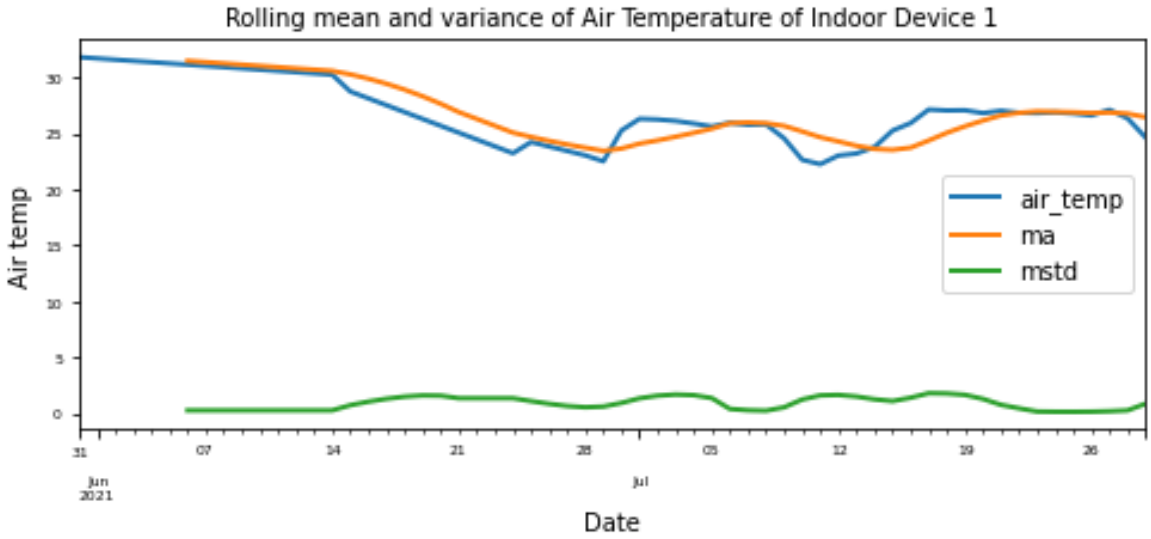


Figure 4-15: Rolling mean and Variance of Air Temperature of Indoor Device 1 and Device 2

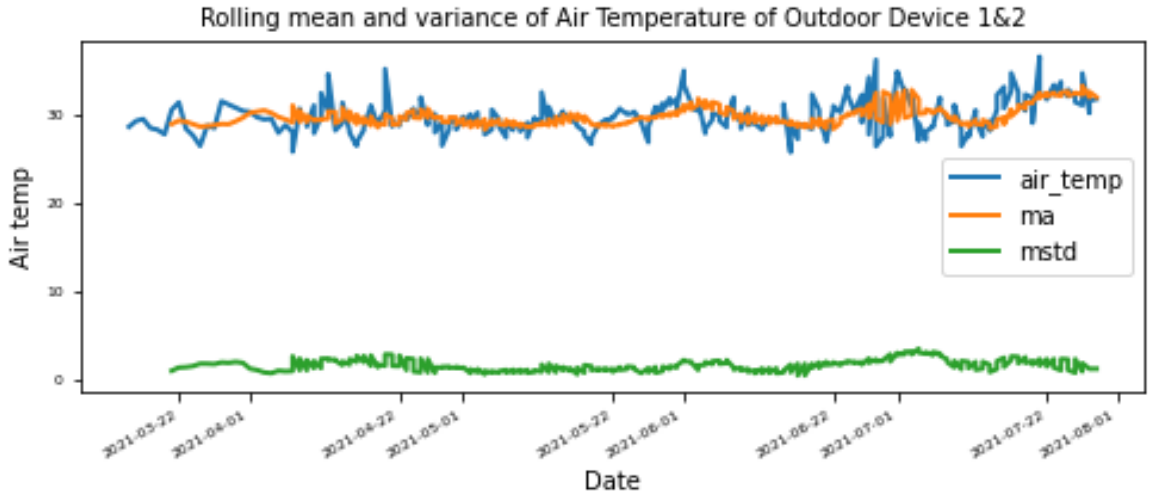


Figure 4-16: Rolling mean and Variance of Air Temperature of Outdoor Device 1&2

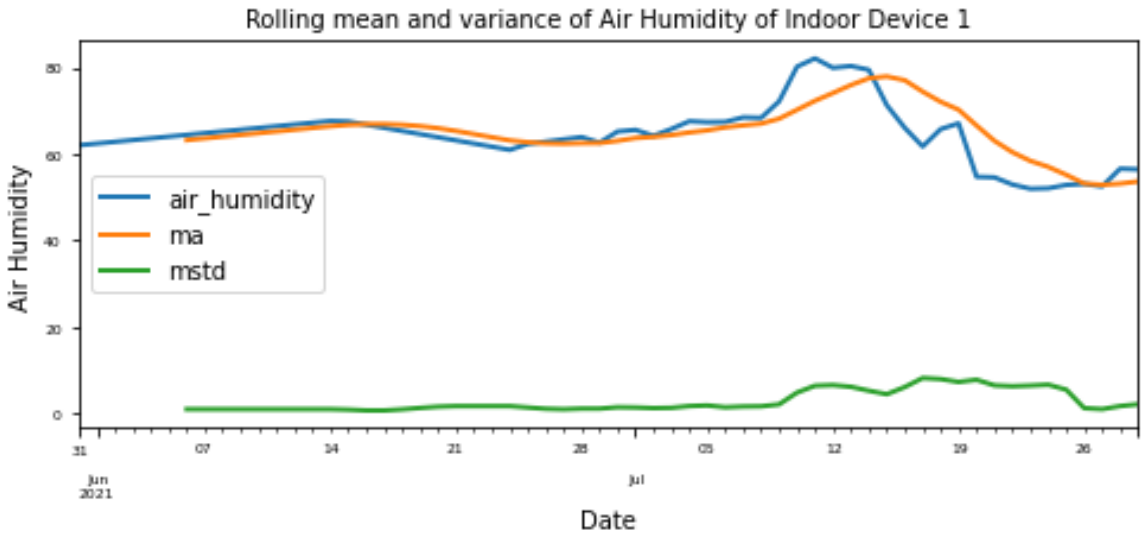


Figure 4-17: Rolling mean and Variance of Air Humidity of Indoor Device 1 and Device 2

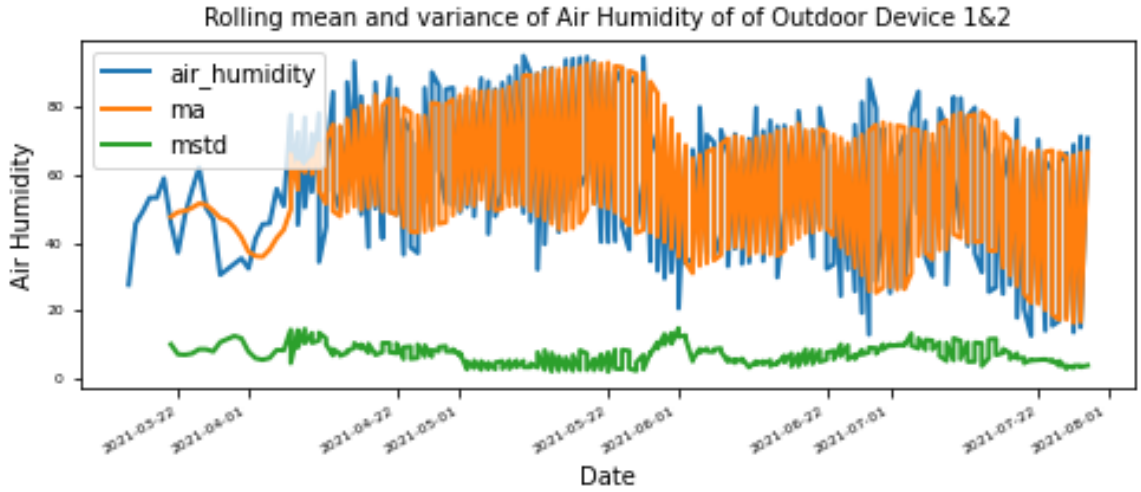


Figure 4-18: Rolling mean and Variance of Air Humidity of Outdoor Device 1&2

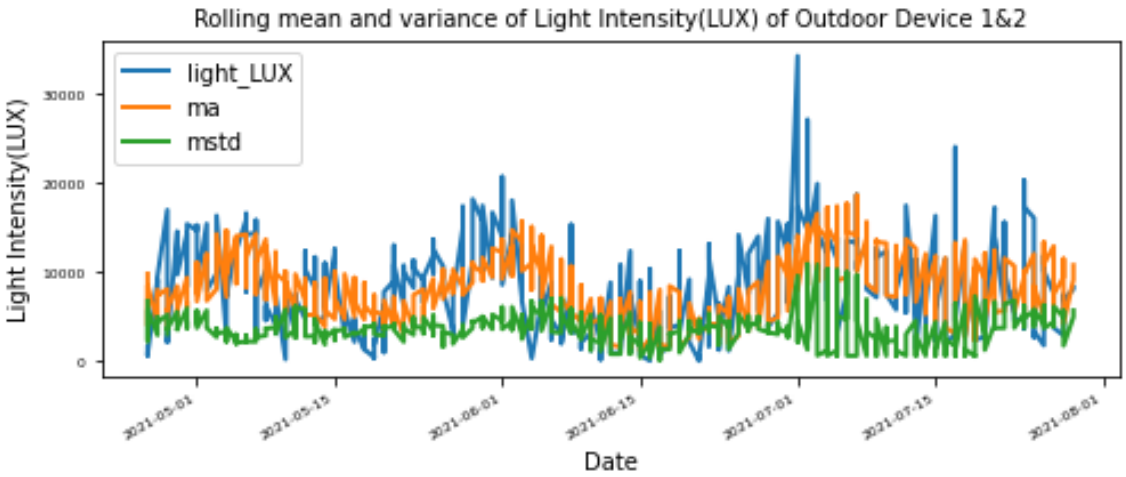


Figure 4-19: Rolling mean and Variance of Light Intensity of Outdoor Device 1 and Device 2

The air temperature of Indoor Device 1 , air humidity of Outdoor Device 1&2 and the light intensity of Outdoor Device 1&2 seems not stationary as the mean and standard deviation do not remain constant over time.

iii. Looking at Auto correlation Function (ACF) plots

Besides looking at the time plot, the Auto Correlation Function Plots (ACF) can be used for identifying non-stationary data.

By looking at the auto correlation plot shown in Figure 4-20 and Figure 4-21 , both the air temperatures of Indoor Device 1 and air humidity of Outdoor Device1&2 showed the traits of the non-stationary time series mentioned by [61] . It can be observed that the time series are very highly correlated, the value of ACF is large and positive and the ACF of non-stationary data decrease gradually.However, trying to determine whether a time series is stationary by just looking at its plot is a dubious venture. A more reliable and effective statistical model is used.

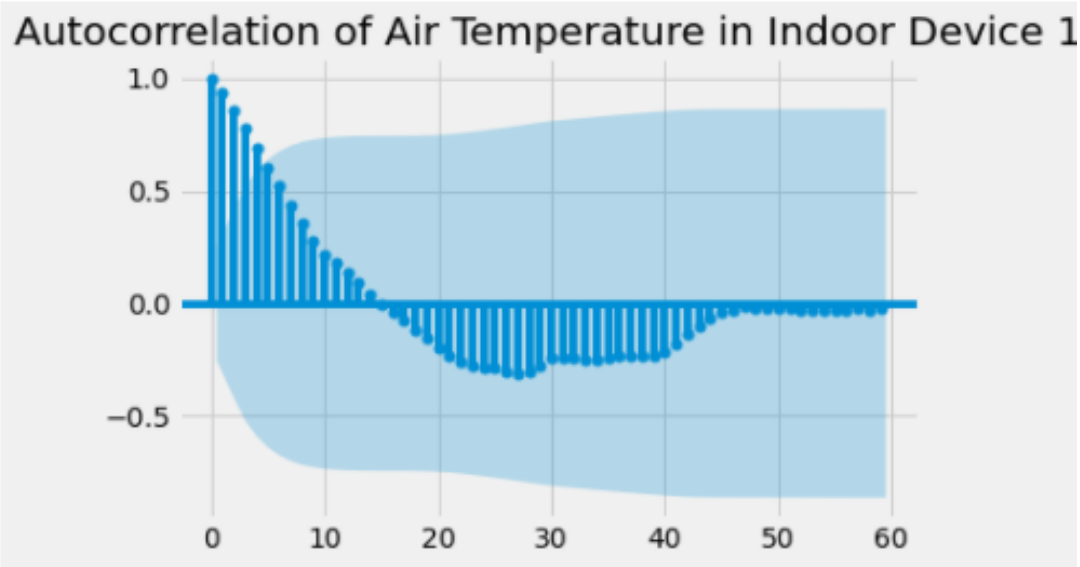


Figure 4-20: Auto correlation Plot of Air Temperature in Indoor Device 1

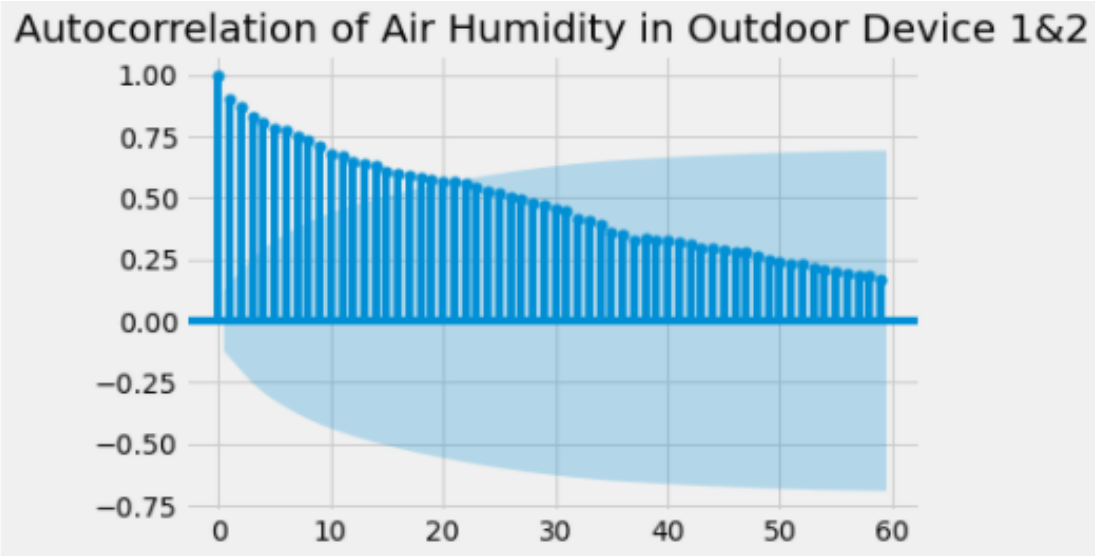


Figure 4-21: Auto correlation Plot of Air Humidity in Outdoor Device 1&2

Augmented Dickey-Fuller Test

The augmented Dickey-Fuller (ADF) test is a common statistical test to test the stationary of a time series [62]. The stationery of a time series is a very important factor to consider before performing ARIMA time series forecasting. The first step in ARIMA forecasting is to determine the number of differences required to make the series stationary as the model is unable to forecast on non-stationary time series data.

We can use the statistical hypothesis unit toot tests to objectively determine whether the series requires differencing. Consequently, small p-values (e.g., > 0.05) suggest that differencing is required

Null Hypothesis (H0): If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time-dependent structure.

Alternate Hypothesis (H1): The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have a time-dependent structure

As observed in Figure 4-22, the result shows that the statistic test value is greater than the critical value and the p-value is also greater than the significant value of 0.05, we fail to reject the null hypothesis and conclude that the time series is non-stationary.

```
ADF Statistic: -1.839307
p-value: 0.361129
n_lags: 1.000000
No of observation: 58.000000
Critical Values:
        1%: -3.548
        5%: -2.913
        10%: -2.594
```

Figure 4-22: Results of ADF Test of Air temperature collected from Indoor Device 1

As observed in Figure 4-23, the result showed that the p-value is lower than 0.05, we reject the null hypothesis and conclude that the time series is stationary.

```
ADF Statistic: -7.151801
p-value: 0.000000
n_lags: 1.000000
No of observation: 248.000000
Critical Values:
        1%: -3.457
        5%: -2.873
        10%: -2.573
```

Figure 4-23: Results of ADF Test of Air temperature collected from Outdoor Device 1&2

As observed in Figure 4-24, the result showed that the p-value is lower than 0.05 , and the critical values at 1%, 5%, 10% confidence intervals are as close as possible to the ADF Statistics , we reject the null hypothesis and conclude that the time series is stationary.

```
ADF Statistic: -3.074541
p-value: 0.028496
n_lags: 11.000000
No of observation: 48.000000
Critical Values:
        1%: -3.575
        5%: -2.924
        10%: -2.600
```

Figure 4-24: Results of ADF Test of Air Humidity Collected from Indoor Device 1

As observed in Figure 4-25, the result showed that the p-value is greater than 0.05 , we fail to reject the null hypothesis and conclude that the time series is non-stationary.

```
ADF Statistic: -0.733460
p-value: 0.837891
n_lags: 5.000000
No of observation: 244.000000
Critical Values:
        1%: -3.457
        5%: -2.873
        10%: -2.573
```

Figure 4-25: Results of ADF Test of Air Humidity Collected from Outdoor Device 1&2

As observed in Figure 4-26, the result showed that the p-value is lower than 0.05, we reject the null hypothesis and conclude that the time series is stationary

```
ADF Statistic: -7.706510
p-value: 0.000000
n_lags: 1.000000
No of observation: 376.000000
Critical Values:
        1%: -3.448
        5%: -2.869
        10%: -2.571
```

Figure 4-26: Results of ADF Test of Light Intensity Collected from Outdoor Device 1&2

To summary, all the time series tested is stationary except the air humidity of Outdoor Device 1&2 and air temperature of Indoor Device 1 which have non-stationary data.

Table 4-6: Summary of ADF Statistical Results from all the Indoor and Outdoor Devices

Time Series	ADF Statistics	p-value	Results
Indoor Device 1 Air Temperature	-1.839307	0.361129	Non-stationary
Outdoor Device 1&2 Air Temperature	-7.151801	0.00	Stationary
Indoor Device 1 Air Humidity	-3.074541	0.028496	Stationary
Outdoor Device 1&2 Air Humidity	-0.733460	0.837891	Non-stationary
Outdoor Device 1&2 Light Intensity	-7.706510	0.00	Stationary

Transform non-stationary series to make it stationary

Differencing

Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality [63]. It is a critical step in ARIMA forecasting to determine the number of differencing required to make the series stationary as the model is unable to forecast on non-stationary time series data.

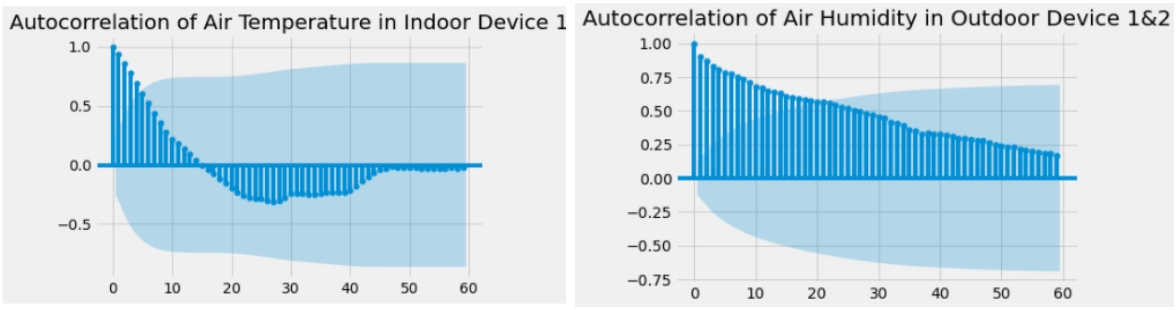


Figure 4-27: Auto Correlation Plot of Air Temperature and Humidity Before Differencing (Non-Stationary)

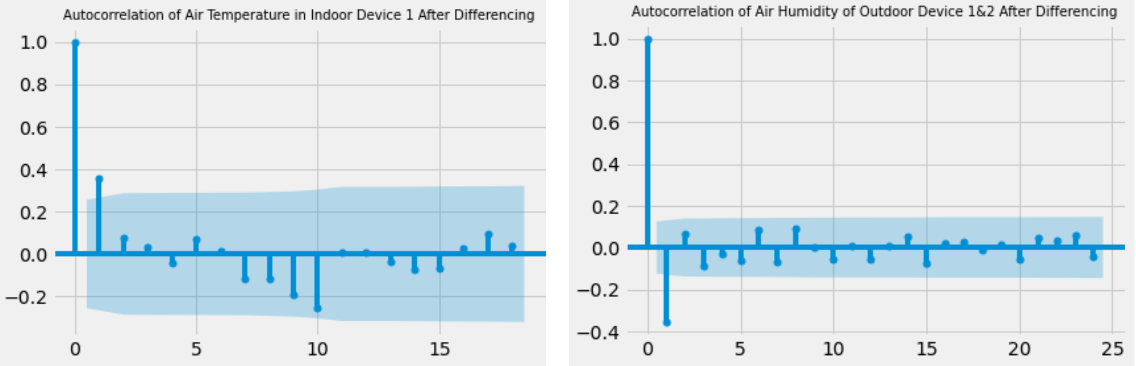


Figure 4-28: Auto Correlation Plot of Air Temperature and Humidity After Differencing (Stationary)

As shown in Figure 4-28 , the air temperature of Device 1 indoor and air humidity of Outdoor Device 1&2 are now stationary. On the graph, indoor device 1 air temperature has significant auto correlation at lag 1 , so **q=1** can be used to determine the MA order. It is also evidenced by observing the value of ACF in both of the data declines to near zero rapidly or non-significant level quickly for a stationary time series.

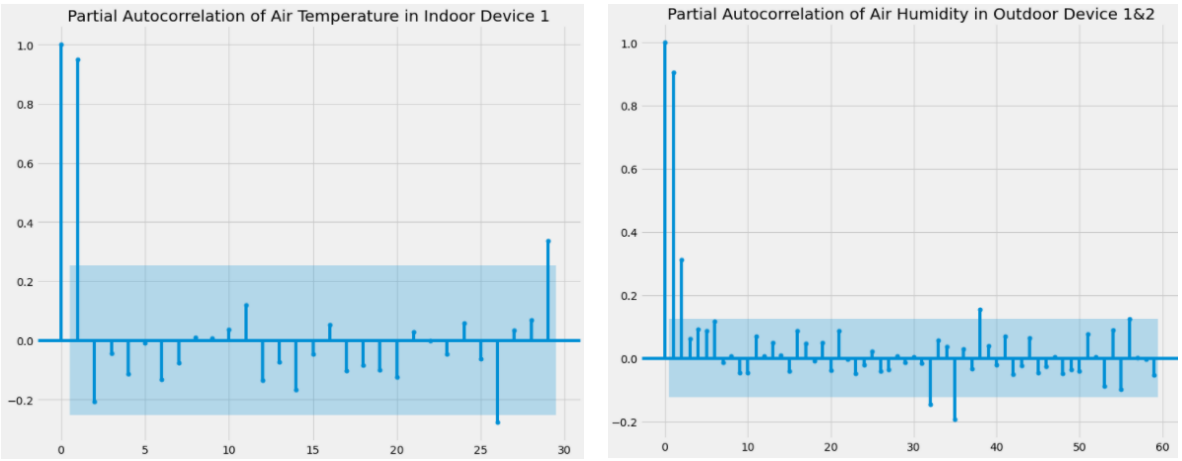


Figure 4-29: Partial Auto Correlation Plot of Air Temperature and Humidity of Indoor and Outdoor Device Before Differencing (Non-Stationary)

As shown in Figure 4-30 , the PACF have a significant lag-1 value, and roughly zeros after that, this PACF suggests that **p=1** will be use in the ARIMA model to refer to the AR order for both of the variable.

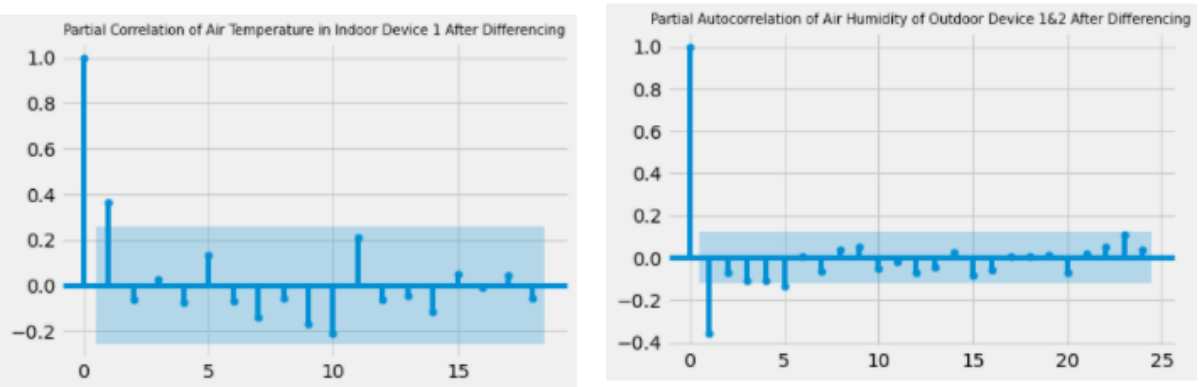


Figure 4-30: Partial Auto correlation Plot of Air Temperature and Humidity of Indoor and Outdoor Device After Differencing (Stationary)

Figure 4-31 and Figure 4-32 showed the time series plot of air temperature in Indoor Device 1 and air humidity of Outdoor Device 1&2 after differencing. The line is much more smoother and parallel to the x-axis.

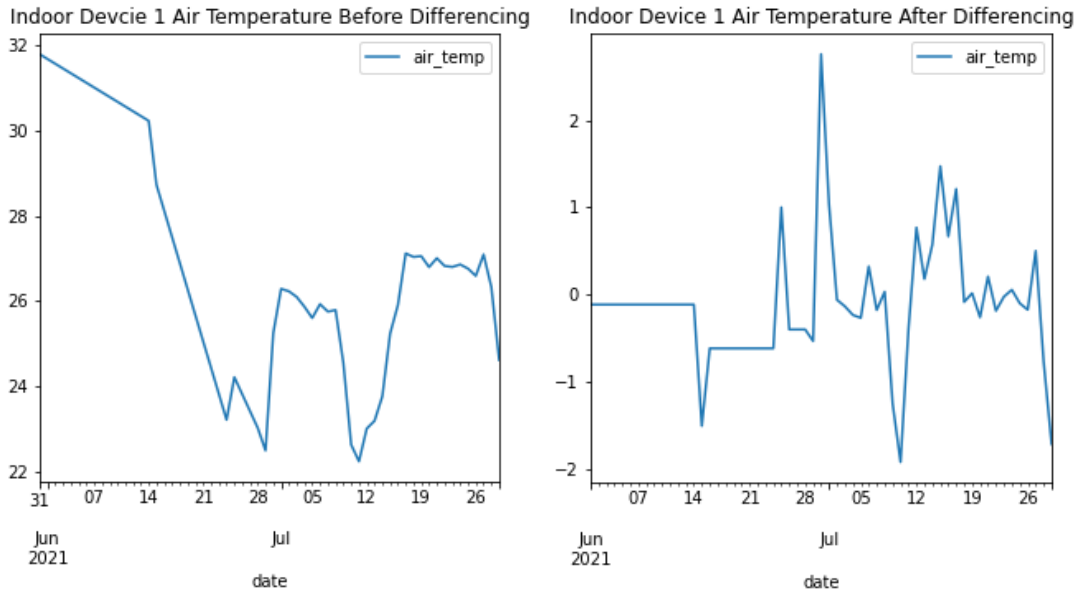


Figure 4-31: Time Series Plot of Air Temperature Collected From Indoor Device 1 Before and After Differencing

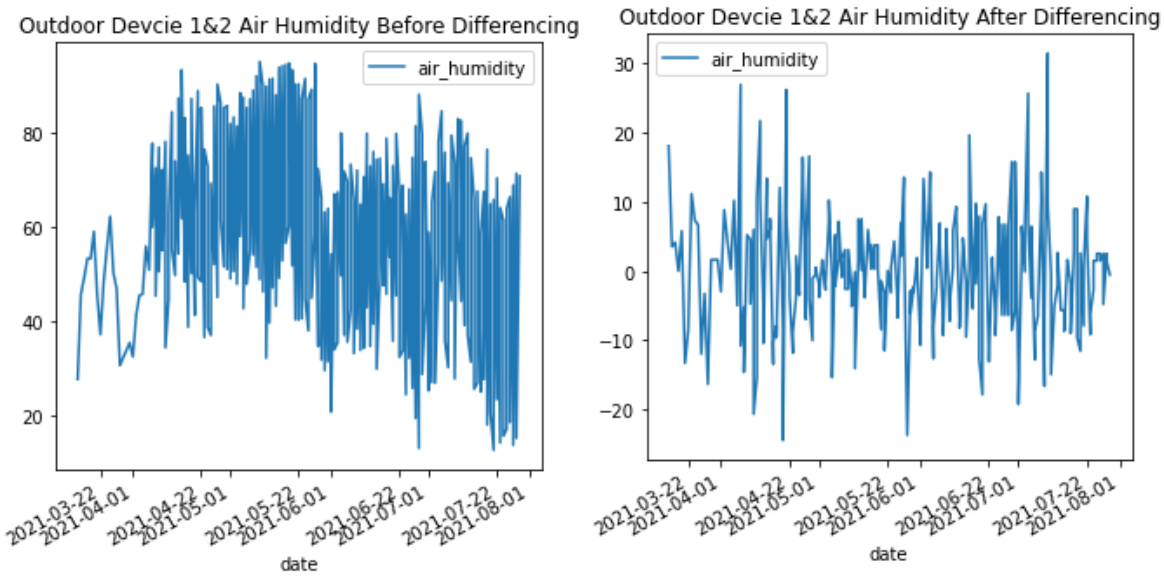


Figure 4-32: Time Series Plot of Air Humidity Collected From Outdoor Device 1 and Device 2 Before and After Differencing

To validate statiscally ,when the Augmented Dickey-Fuller (ADF) test is run on the original distribution of air temperature collected from Indoor Device 1, the p-value is 0.36. As observed in Figure 4-33, the p-value is 0.000084 after differencing is applied, and well within the range after differencing, we would expect for stationary data, suggesting we can reject the null and conclude the differenced data is now stationary.

```
ADF Statistic: -4.701142
p-value: 0.000084
Critical Values:
        1%: -3.548
        5%: -2.913
        10%: -2.594
```

Figure 4-33: Results of ADF Test of Air Temperature Collected from Indoor Device 1 After Differencing

When we run the Augmented Dickey-Fuller (ADF) Test on the original distribution of air humidity of Outdoor Device 1 and Device 2, the p-value is 0.837. As observed in Figure 4-34, the p-value is 0 after differencing, suggesting we can reject the null and conclude that the differenced data is now stationary.

```
ADF Statistic: -10.139884
p-value: 0.000000
Critical Values:
        1%: -3.457
        5%: -2.873
        10%: -2.573
```

Figure 4-34: Results of ADF Test of Air Humidity Collected from Outdoor Device 1 and Device 2 After Differencing

From this point, both of the Indoor Device 1 air temperature and Device Outdoor 1&2 air humidity employs the first-order difference, so **d=1** will be used in the ARIMA model that refers to the number of differencing transformations required by the time series to get stationary.

In summary, we can create and fit an ARIMA model with AR of order 1 (**p=1**), differencing of order 1 (**d=1**) and MA of order 1 (**q=1**) to forecast the air temperature of Indoor Device 1 and air humidity of Outdoor Device 1&2.

As shown in Figure 4-35 , the ACF charts showed that the air temperature in Outdoor Device 1&2 have significant lags up to 2 lags .In this way ,**q=2** will be used for Outdoor Device 1&2 air temperature to denote the MA model. By looking at the PACF , **p=1** can be used to determined the order of AR as it is the lag value where the PACF chart crosses the significance threshold.

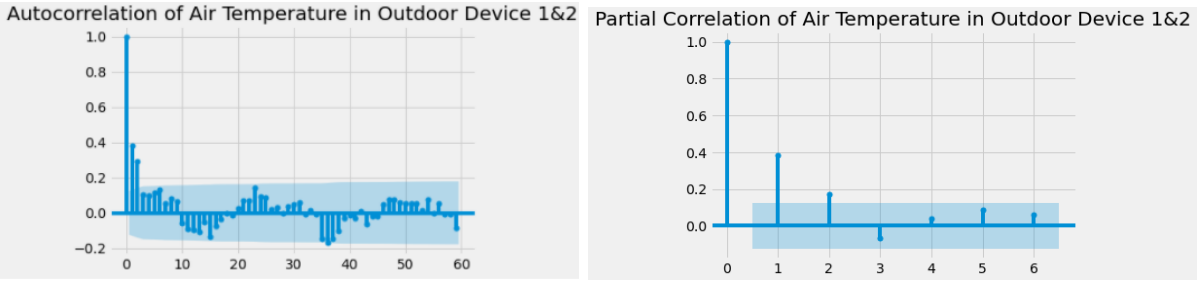


Figure 4-35: Auto Correlation Plot and Partial Correlation of Air Temperature in Outdoor Device 1& 2 (Stationary)

As shown in Figure 4-36 , the ACF charts showed that the air humidity in Indoor Device have significant lags up to 3 lags respectively. In this way , **q=3** will be used for Indoor Device 1 air humidity to denote the MA model. By looking at the PACF , **p=0** can be used to determined the order of AR.

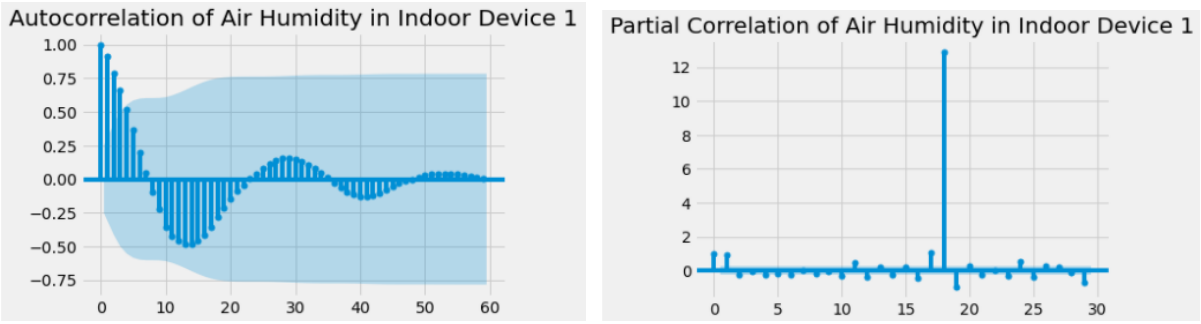


Figure 4-36: Auto Correlation Plot and Partial Correlation of Air Humidity in Indoor Device 1 (Stationary)

As shown in Figure 4-37 , the auto correlations of light intensity in Outdoor Device 1&2 are statistically significant up to 5 lags and the subsequent lags are nearly significant .Consequently , this ACF suggest fitting either a **q=5** or **q=6** order MA model. By looking at the PACF , **p=1 or 2** can be used to determined the order of AR as it is the lag value where the PACF chart crosses the significance threshold.

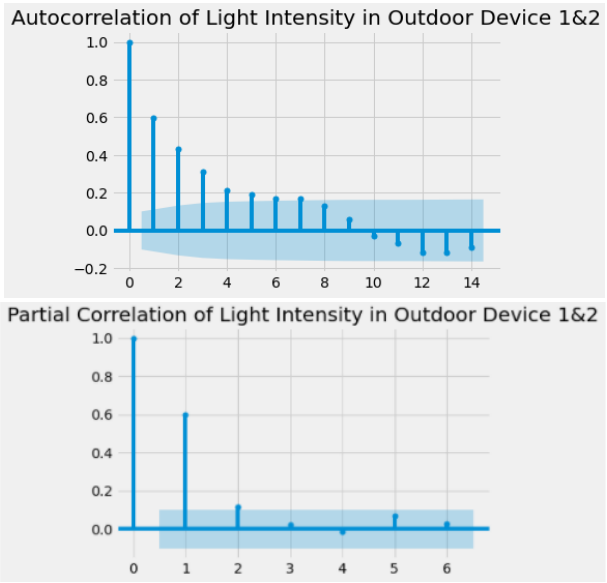


Figure 4-37: Auto Correlation and Partial Correlation Plot of Light Intensity in Outdoor Device 1 and Device 2 (Stationary)

In summary , we can create and fit an ARIMA model by these parameters :

p=0 or 2 , d=0 , q= 3 for Indoor Device 1 air humidity

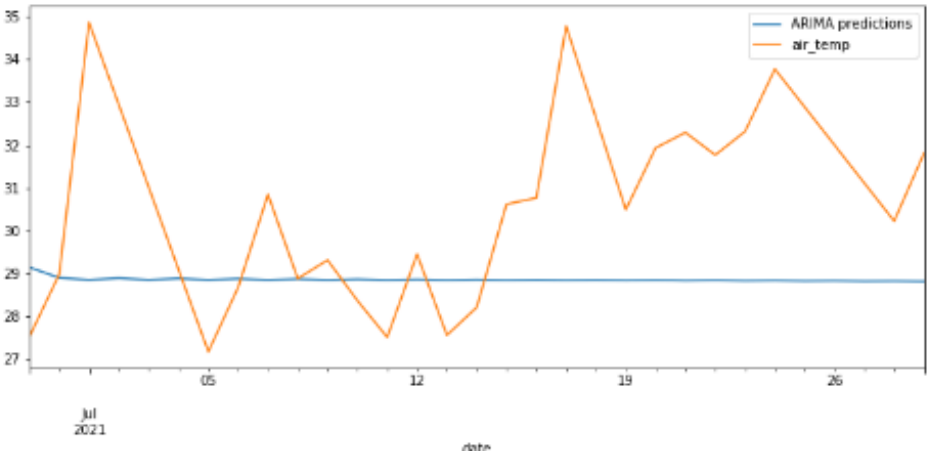
p=1 , d=1 , q=2 for Outdoor Device 1&2 air temperature

p=1 or 2 , d=0 , q=5 or 6 for Outdoor Device 1&2 light intensity

4.3 Result Interpretation of ARIMA Model

As the data collected by the indoor device only took place from 31st May to 29th July, it has insufficient data to perform good predictions. For this project, the focus will be on predicting the air temperature, air humidity and light intensity of the outdoor device.

To compare whether manual or auto method of determining the orders of the parameters is able to yield a better result in predicting the climatic variables, a comparison was done. The result of predicting the air temperature of outdoor device 1&2 (key46) dataset is shown in Figure 4-38.



ARIMA Model Results

Dep. Variable:	D.air_temp	No. Observations:	106
Model:	ARIMA(1, 1, 2)	Log Likelihood	-200.089
Method:	css-mle	S.D. of innovations	1.588
Date:	Tue, 23 Nov 2021	AIC	410.179
Time:	15:31:54	BIC	423.496
Sample:	03-16-2021	HQIC	415.576
	- 06-29-2021		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0019	0.026	-0.072	0.942	-0.053	0.049
ar.L1.D.air_temp	-0.9157	0.149	-6.159	0.000	-1.207	-0.624
ma.L1.D.air_temp	0.0519	0.182	0.286	0.775	-0.304	0.408
ma.L2.D.air_temp	-0.7501	0.173	-4.343	0.000	-1.089	-0.412

Roots

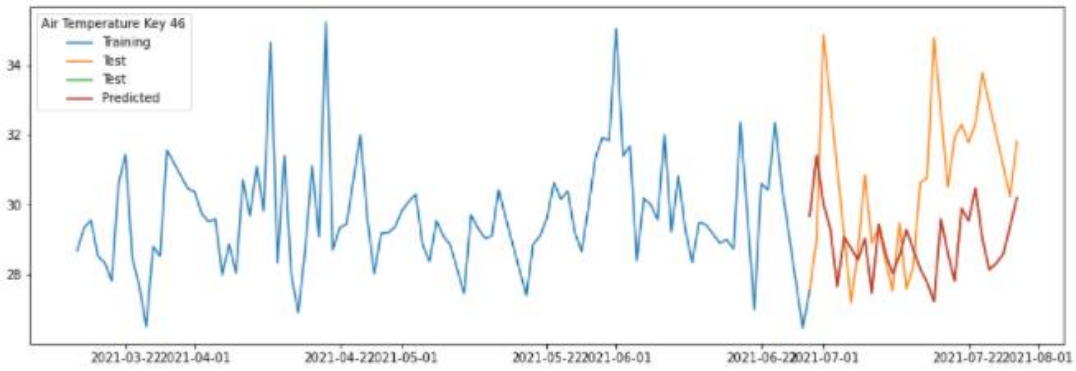
	Real	Imaginary	Modulus	Frequency
AR.1	-1.0920	+0.0000j	1.0920	0.5000
MA.1	-1.1205	+0.0000j	1.1205	0.5000
MA.2	1.1898	+0.0000j	1.1898	0.0000

Figure 4-38: *Result of ARIMA model of Outdoor Device 1&2 Air Temperature Key 46 generated by determining the parameters order using ACF/PACF plot*

From the visualizations in Section 4.2 , the order of ARIMA (1,1,2) is determined and it is used to fit the ARIMA model on the training set to predict the test set. However, the result showed in Figure 4-38 stated that the model did not follow the trend of the test data at all.

By fitting in the model to the auto ARIMA, it will figure out the best order of the ARIMA by selecting the lowest AIC score to judge how good a particular order model is. The model in Figure 4-39 suggests that the best SARIMA model is the order of (0,1,2)x(1,1,0,30) with the minimum AIC score of 323.50 and BIC score of 332.82. Both of the values is lower than the value shown in Figure 4-38.

As seen in the Figure 4-39 , this model did a good job in predicting the air temperature values without compromising with the seasonality effects and exogenous factors. The predicted results closely resembled the actual trend with acceptable lag.



SARIMAX Results

Dep. Variable:	y	No. Observations:	107
Model:	SARIMAX(0, 1, 2)x(1, 1, [], 30)	Log Likelihood	-157.752
Date:	Fri, 19 Nov 2021	AIC	323.503
Time:	13:45:49	BIC	332.826
Sample:	0	HQIC	327.229
	- 107		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.7265	0.121	-5.992	0.000	-0.964	-0.489
ma.L2	-0.1327	0.129	-1.028	0.304	-0.386	0.120
ar.S.L30	-0.5847	0.086	-6.770	0.000	-0.754	-0.415
sigma2	3.1009	0.517	5.997	0.000	2.087	4.114

Ljung-Box (L1) (Q):	0.27	Jarque-Bera (JB):	0.77
Prob(Q):	0.60	Prob(JB):	0.68
Heteroskedasticity (H):	0.84	Skew:	-0.25
Prob(H) (two-sided):	0.66	Kurtosis:	3.03

Figure 4-39: *Result of ARIMA model of Outdoor Device 1&2 Air Temperature Key 46 generated by determining the order of the parameters using Auto ARIMA*

The model has a mean square error of 2.87 which also mean it has a high 97.13 % of accuracy. With this knowledge, this model is used to predict the value for the next 31 days from 28th July to 27th August 2021.

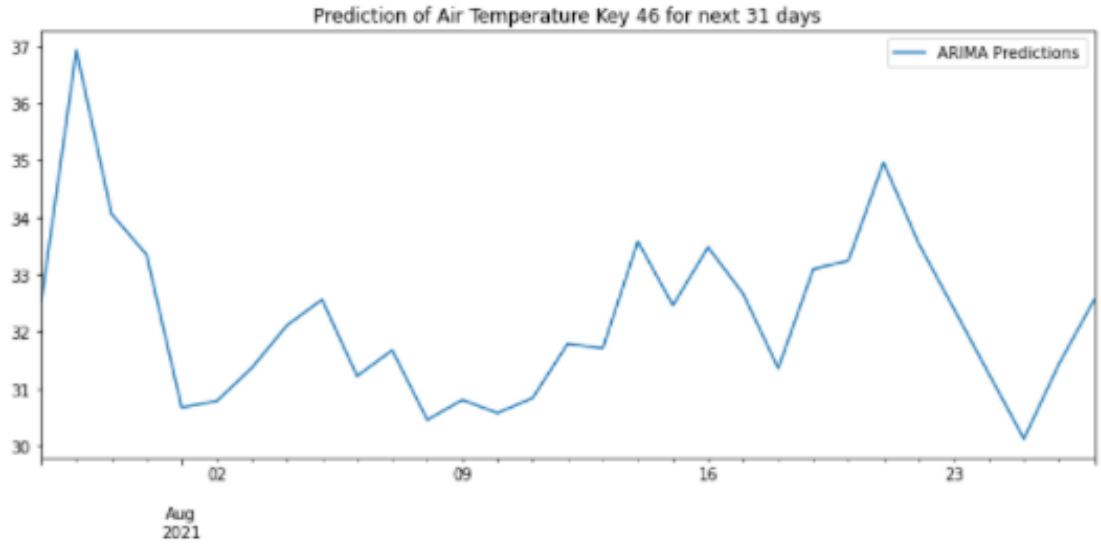


Figure 4-40: Prediction of Air Temperature Key 46 for Next 31 Days

This method is then chosen to determine the optimal order for the ARIMA of air temperature, air humidity and light intensity in Outdoor Device 1&2.The screeshots of the predicted results are attached in Appendix.

Table 4-6 showed the summary of the optimal parameters to determine the order of the model.

Table 4-6: Summary Table of the optimal parameters to determine the order of the model.

Dataset	Order of parameters	AIC	BIC	MSE
Outdoor 1&2 Air Temperature key 46	SARIMAX (0,1,2)x(1,1,0,30)	323.50	332.50	2.88
Outdoor 1&2 Air Temperature key 55	SARIMAX (0,1,1)x(0,1,0,30)	234.10	238.06	3.986
Outdoor 1&2 Air Humidity key 45	SARIMAX (3,1,3)x(1,1,0,30)	586.54	587.08	7.56
Outdoor 1&2 Air Humidity key 54	SARIMAX (3,1,0)x(1,1,0,30)	417.35	427.01	13.96
Outdoor 1&2 Light Intensity key 66	SARIMAX (1,1,0)x(2,1,0,12)	1047.70	1055.43	5443
Outdoor 1&2 Light Intensity key 67	SARIMAX (2,1,1)x(1,1,1,12)	1017.46	1029.05	7730.35
Outdoor 1&2 Light Intensity key 68	SARIMAX (0,1,2)x(2,1,0,12)	1000.08	1009.64	5106.93
Outdoor 1&2 Light Intensity key 69	SARIMAX (0,1,0)x(1,1,1,30)	684.37	688.77	4400.26

4.4 Results Interpretation of Prediction Dashboard

4.4.1 Sensors Dashboard

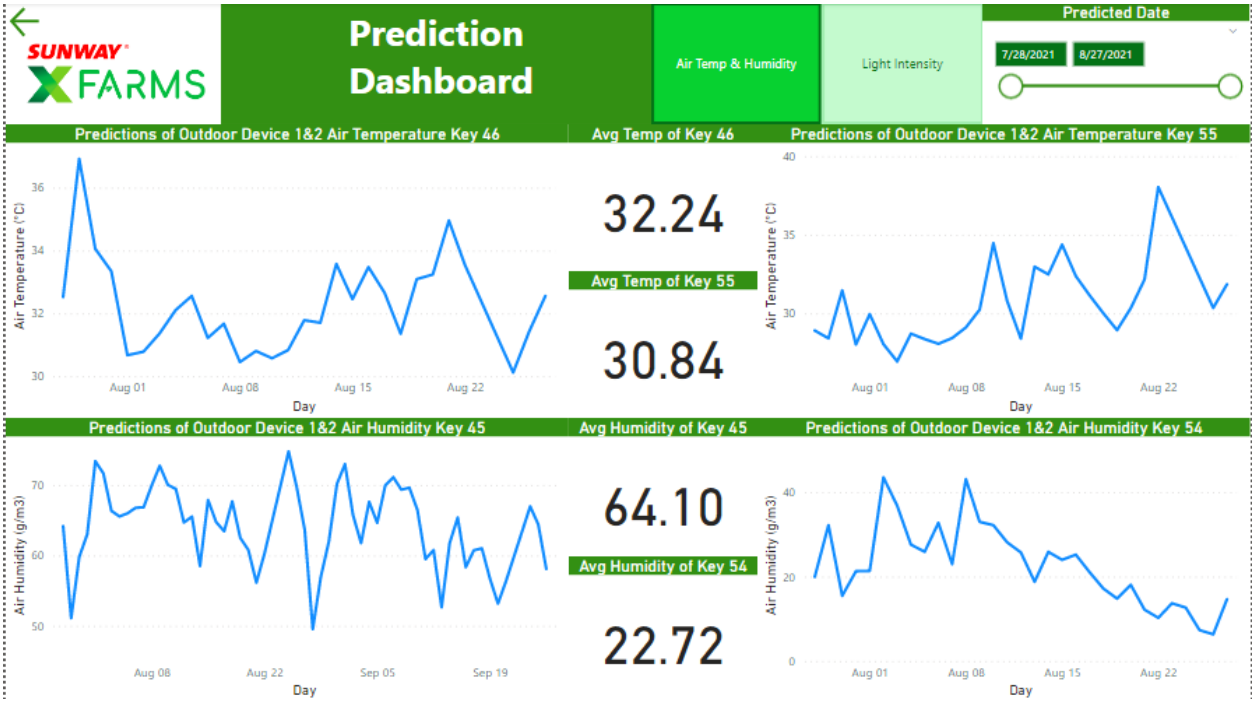


Figure 4-41: Prediction of Air Temperature Dashboards

Figure 4-41 shows the prediction dashboard that displayed the predicted values for air temperature and air humidity of Outdoor Device 1&2 in the next 31 days (28th of July 2021 – 27th of August 2021).

From the dashboard, it showed that the predicted average temperature for the keys 46 and 55 are similar which is around 30 °C. The predicted day with the highest temperature will be on 29th of July with a predicted temperature of 36.92 °C.

The predicted average air humidity of key 45 is a lot more higher than key 54 in the next 30 days with the range from 50 to 70 g/m3 while key 54 only has the maximum humidity level up to 40 g/m3. The air humidity level of key 54 will be decreasing starting from 8th August 2021.

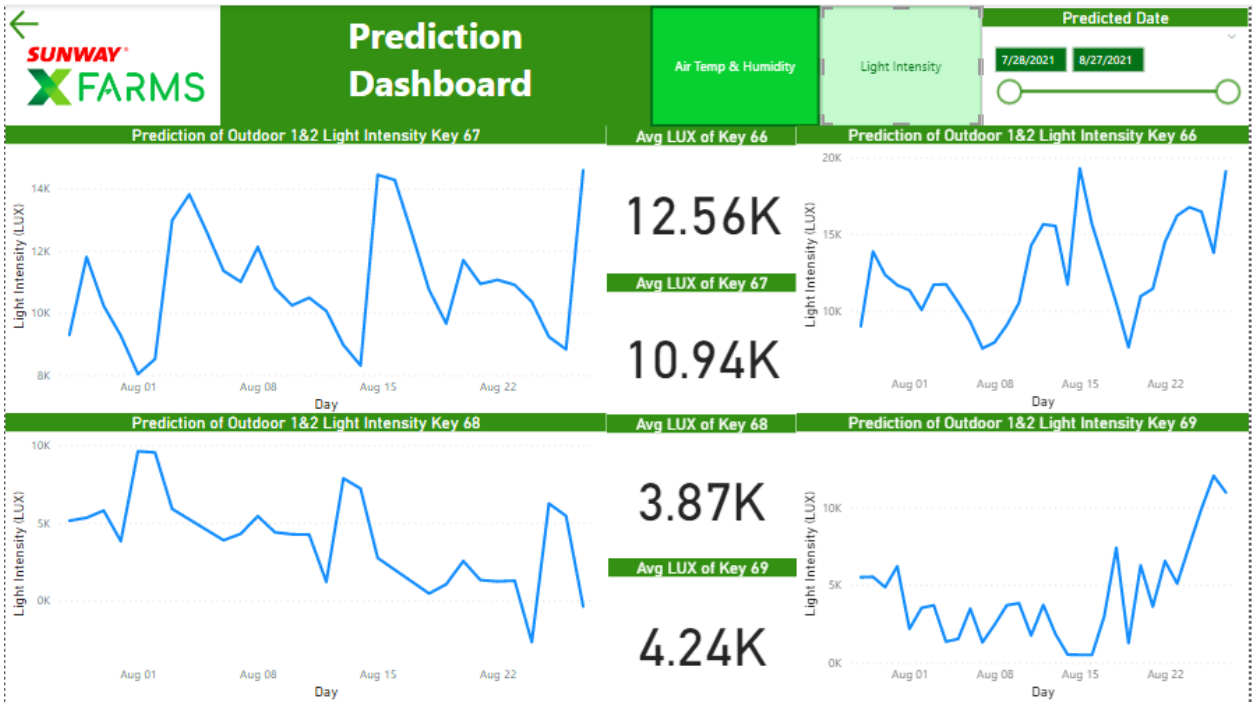


Figure 4-42: Prediction of Light Intensity Dashboards

Figure 4-42 also displayed the prediction value of the light intensity of each key for the next 31 days. (28th of July 2021 – 27th of August 2021) .

Key 66 and key 67 has the highest LUX value than the other keys. The LUX value of Key 68 is predicted to decrease tremendously starting from 01st August 2021 until the 22nd of August 2021.

4.4.2 Yield Analysis Dashboard

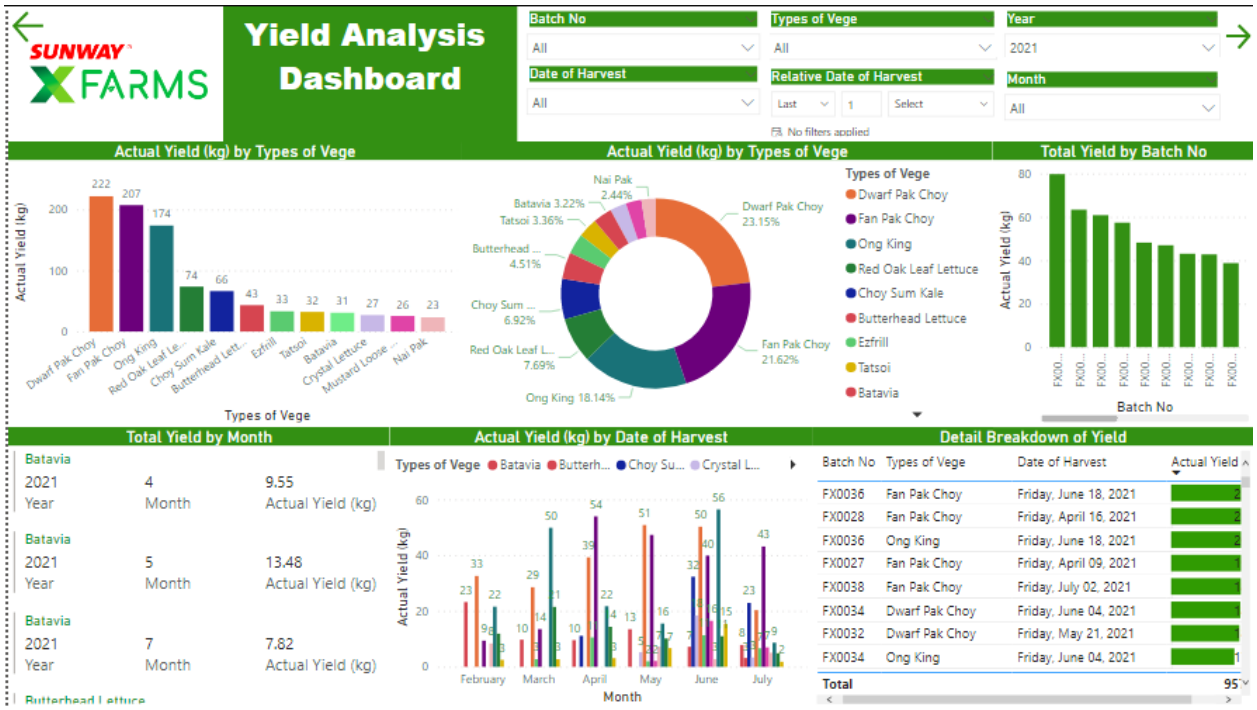


Figure 4-43: Yield Analysis Dashboards

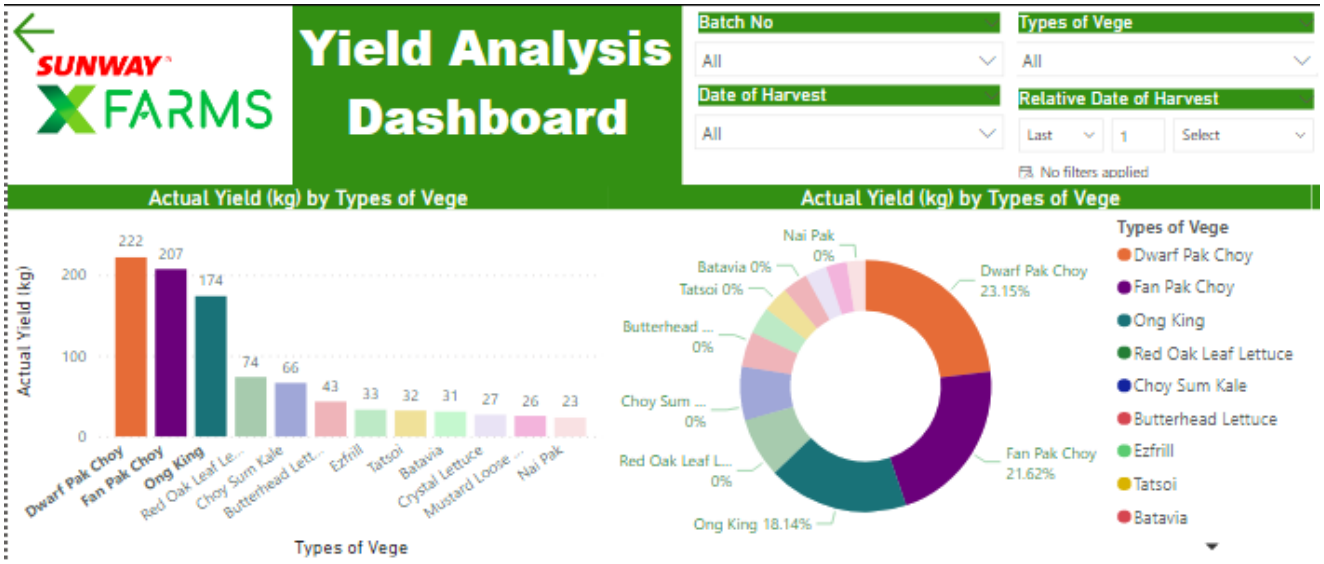


Figure 4-44: Actual Yield by Types of Vege

As shown in Figure 4-44, Dwart Pak Choy and Fan Pak Choy and Ong King were the top 3 crops that were having the largest kilograms among the other type of vegetables. As seen from the pie chart, they made up 63% of the distribution.

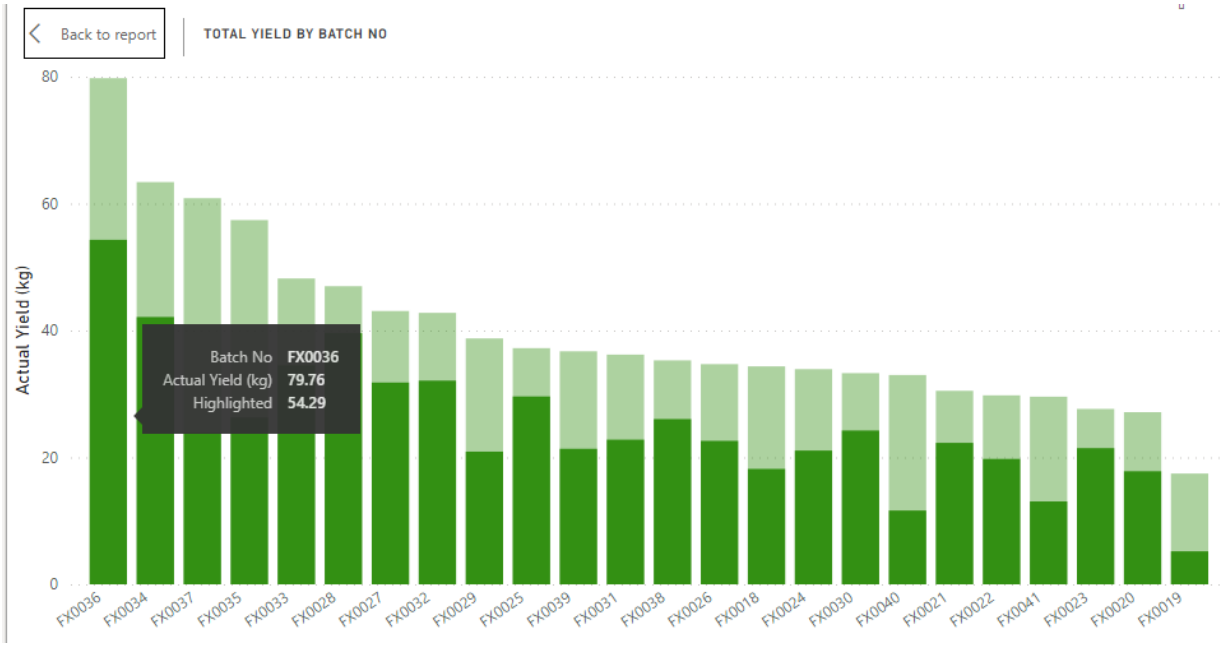
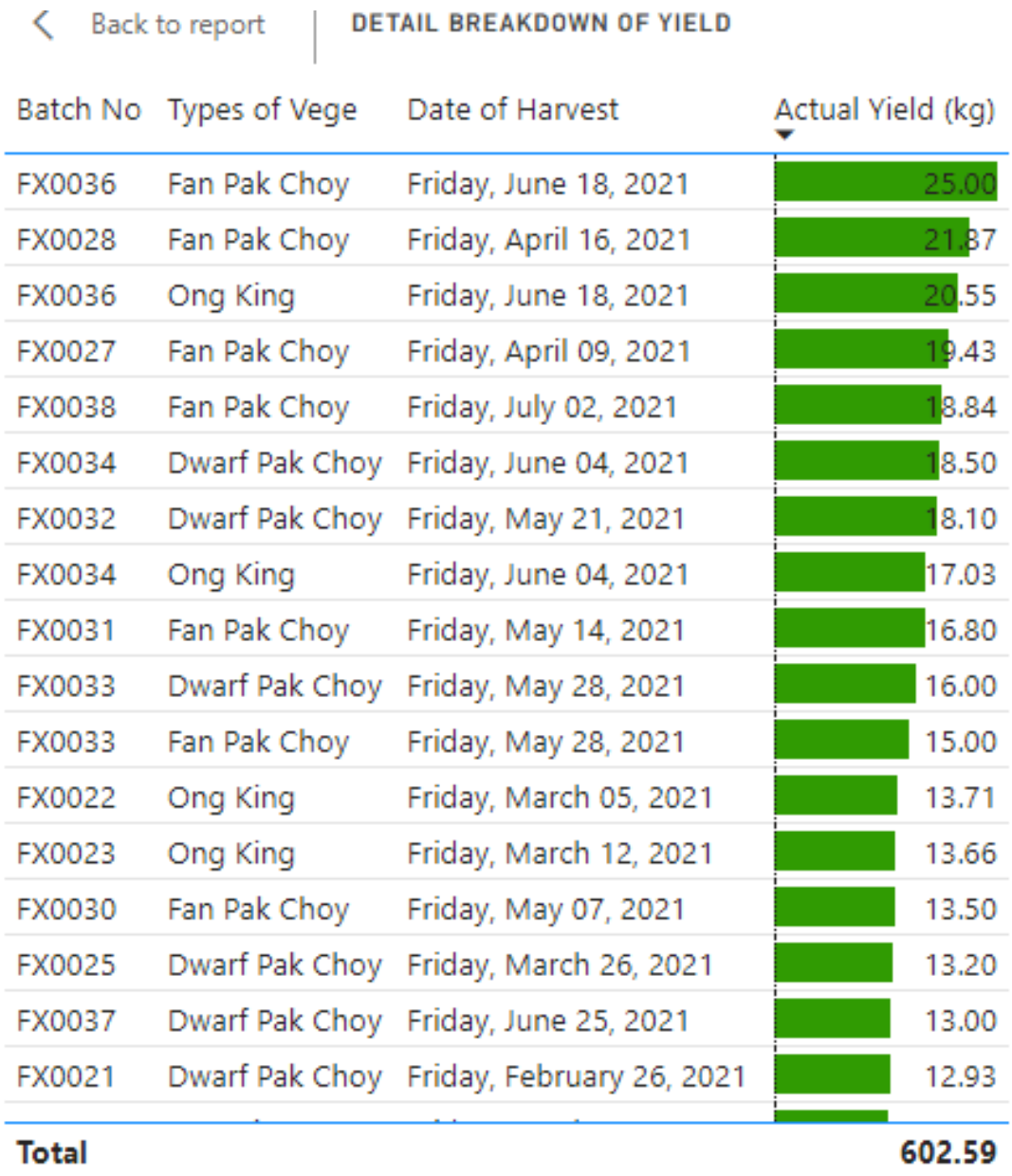


Figure 4-45: Bar Chart of Actual Yield

The batch number of FX0036 was the most harvested batch with a total of 54.29 kg for these three types of vegetables while FX0019 was the least harvested batch with a total weight of 5.2kg.

Figure 4-46: Summary Table of Total Yield for each Type of Vegetables by Month

As shown in Figure 4-46, the total yield for each type of vegetables by month can be tracked in the summary table. It also provides a filter and sort function according to the user need. It is observed that the total harvested yield for Dwarf Pak Choy increased from March and gradually grew to double of its total yield in May and June but it decreased tremendously in July.



Batch No	Types of Vege	Date of Harvest	Actual Yield (kg)
FX0036	Fan Pak Choy	Friday, June 18, 2021	25.00
FX0028	Fan Pak Choy	Friday, April 16, 2021	21.87
FX0036	Ong King	Friday, June 18, 2021	20.55
FX0027	Fan Pak Choy	Friday, April 09, 2021	19.43
FX0038	Fan Pak Choy	Friday, July 02, 2021	18.84
FX0034	Dwarf Pak Choy	Friday, June 04, 2021	18.50
FX0032	Dwarf Pak Choy	Friday, May 21, 2021	18.10
FX0034	Ong King	Friday, June 04, 2021	17.03
FX0031	Fan Pak Choy	Friday, May 14, 2021	16.80
FX0033	Dwarf Pak Choy	Friday, May 28, 2021	16.00
FX0033	Fan Pak Choy	Friday, May 28, 2021	15.00
FX0022	Ong King	Friday, March 05, 2021	13.71
FX0023	Ong King	Friday, March 12, 2021	13.66
FX0030	Fan Pak Choy	Friday, May 07, 2021	13.50
FX0025	Dwarf Pak Choy	Friday, March 26, 2021	13.20
FX0037	Dwarf Pak Choy	Friday, June 25, 2021	13.00
FX0021	Dwarf Pak Choy	Friday, February 26, 2021	12.93
Total			**602.59**

Figure 4-47: Detail Breakdown Table of Yield

From the detail breakdown table shown in Figure 4-47, the user can track the details of the record in terms of batch number, types of vegetables, date of harvest and the actual yield. The Total weight of harvested yield of 602.59 kg.

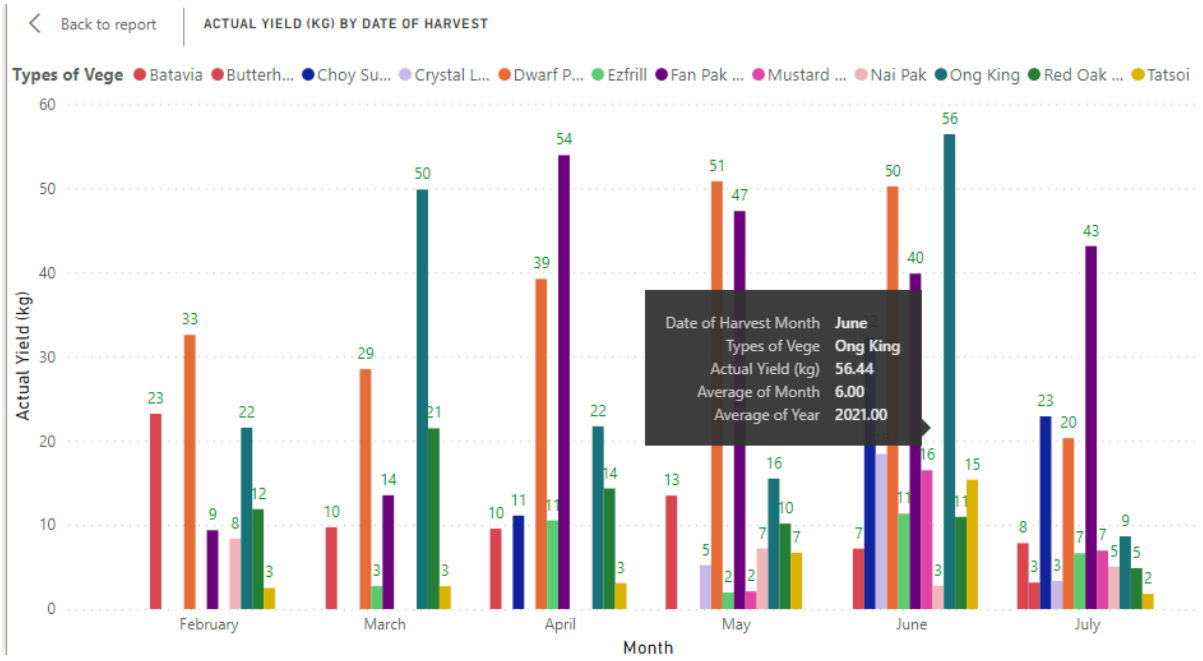


Figure 4-48: Barchart of Actual Yield by Date of Harvest

Users can drill down/up from year to quarter, month and day to visualize the actual yield harvested for each type of vegetable. It showed that *Ong King* has the highest kilogram of actual yield in June as observed in Figure 4-48.

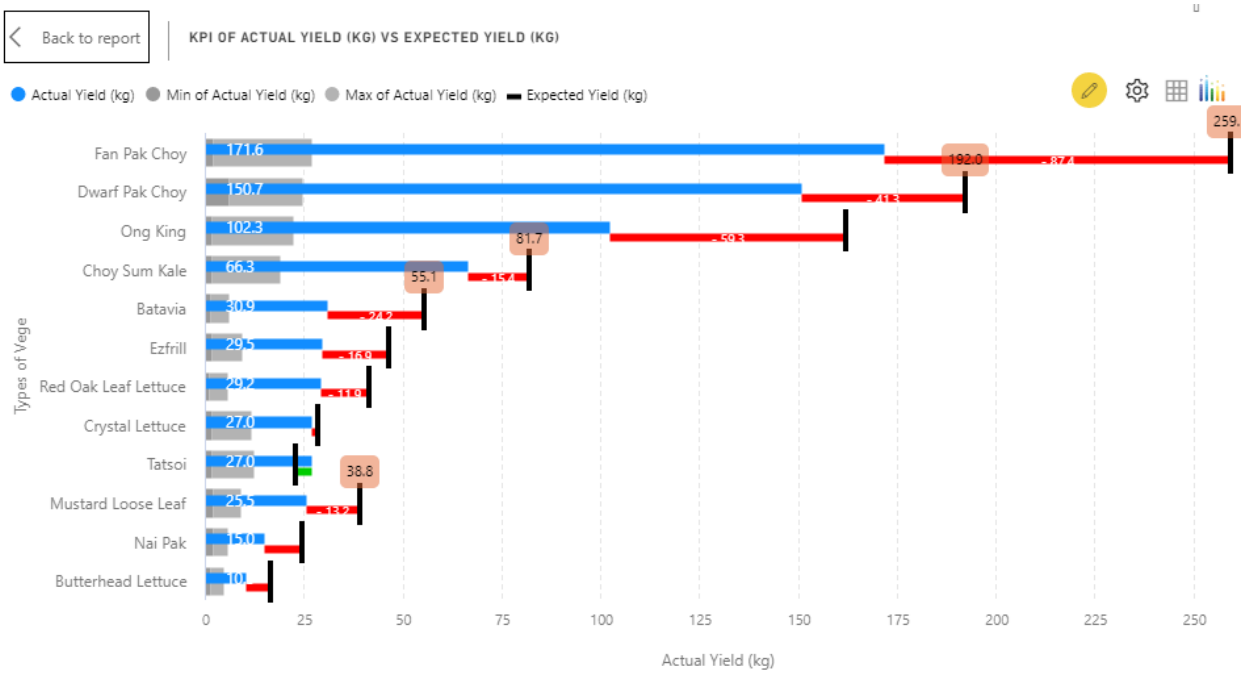


Figure 4-49: KPI of Actual Yield vs. Expected Yield

By looking at the KPI, it showed that *Tatsoi* was the only type of vegetables that achieved the goals with the actual yield of 27 kg as compared to the expected yield of 22.6 kg. On the

other hand, *Fan Pak Choy* showed the largest deviation in terms of achieving the target amount of 259 kg of expected yield. The actual yield was 171.6 kg with a deviation is 87.4kg to its target.

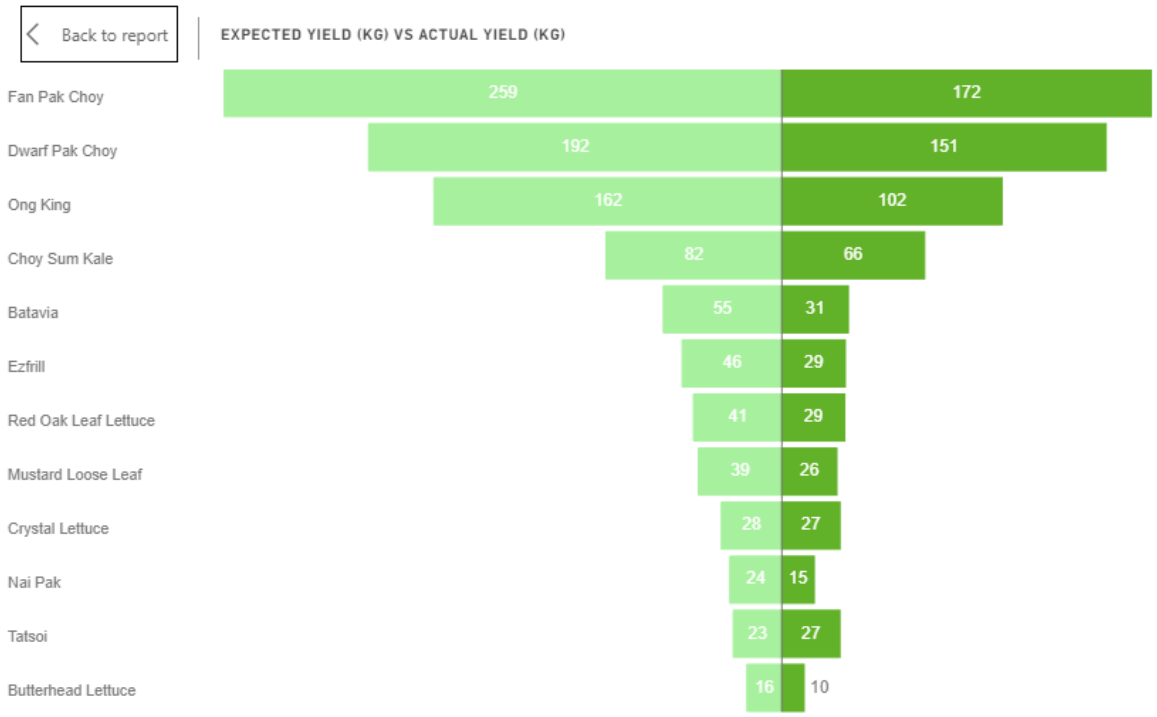


Figure 4-50: Tornardo Chart of Expected Yield vs Actual Yield

The tornado chart provides another way to measure the KPI by looking at and comparing the length of the bars of expected yield and actual yield.

5 Limitation

This study proposed a few limitations due to time constraints in the data collection process. The outdoor device datasets have the data range from March to July while the indoor device datasets have the data range from the end of May to July only. The model building is expected to have better performance with at least one year of data and above. Time series forecasting required more training data in order to build a better prediction model.

Besides, the crops yield dataset is limited to exploratory analysis only as the data provided by the farm team focus more on the total weight of crops harvested rather than sales information. The crops yield data are not synchronized with the sensors data so it is unable to predict the crops yield using the current model.

6 Conclusion

In conclusion, this study aimed to contribute to analysing the sensors data using data mining techniques. This framework presented the whole data mining process from data cleaning to predictive modelling as well as data visualization. Furthermore, manual and auto -selection of ARIMA model are conducted to determine the order of the parameters (p,d,q)(P, D,Q). As demonstrated, auto ARIMA is a more efficient and reliable method to select the order of AR and MA terms as compared to the manual selection. It will select the best optimal parameters orders of ARIMA model with the lowest AIC and BIC values. The forecasted results are then validated by looking at the RMSE between predicted and actual results. The model generated using the Auto_ARIMA() function had better performance as compared to the traditional ARIMA implementation in determining the order of the models using ACF and PACF plots. SARIMA model is more appropriate to forecast the climatic variables as they tend to have seasonal effects such as it is low at night and high in the afternoon. ARIMA model is not capable in dealing with this kind of data.

As future work, other forecasting techniques such as neural networks and support vector machines can be developed to compare the accuracy of the results. The data collection of sensors data can be synchronized with the data collection of crop yields data in order to enable crops yield forecast and sales forecast in the future.

7 References

[1] Natalie, "Should Malaysians be worried about food security? | Covid-19," TheStar, 17 June 2020. [Online]. Available: https://www.thestar.com.my/news/nation/2020/06/17/should-malaysians-be-worried-about-food-security--covid-19. [Accessed 15 May 2021].

[2] "https://sunwayxfarms.com/," Sunway X Farm, [Online]. Available: https://sunwayxfarms.com/. [Accessed 15 May 2021].

[3] R. A. Acharige, M. N. Halgamuge, H. A. H. S. Wirasagoda and A. Syed, "Adoption of the Internet of Things (IoT) in Agriculture and Smart Farming towards Urban Greening: A Review," *International Journal of Advanced Computer Science and Applications,* vol. 10, no. 4, pp. 11-28, 2019.

[4] D. Thakur, Y. Kumar, A. Kumar and P. K. Singh, "Applicability of Wireless Sensor Networks in Precision Agriculture: A Review," *Wireless Personal Communications,* vol. 107, pp. 471-512, 2019.

[5] J. Chen, S. He and X. Li, "A Study of Big Data Application in Agriculture," *Journal of Physics: Conference Series,* vol. 1757, 2021.

[6] "Data Mining," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Data_mining. [Accessed 2 June 2020].

[7] X. W. e. al., "Top 10 algorithms in data mining," *Knowl Inf Syst,* vol. 14, pp. 1-37, 2008.

[8] "Introduction to Data Mining Tasks," WideSkills, [Online]. Available: https://www.wideskills.com/data-mining-tutorial/05-data-mining-tasks. [Accessed 2 June 2021].

[9] "A Survey and Analysis of Various Agricultural Crops," *International Journal of Computer Applications,* vol. 136, no. 11, pp. 25-30, 2016.

[10] H. A. Issad, R. Aoudjit and J. J. Rodrigues, "A comprehensive review of Data Mining techniques in smart agriculture," *Engineering in Agriculture, Environment and Food,* 2019.

[11] D. I. Fabrizio BalducciOrcID and d. Pirlo, "Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement," *Machines,* vol. 6, no. 3, p. 38, 2018.

[12] J. Majumdar, S. Naraseeyappa and S. Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," *Journal of Big Data,* vol. 4, no. 20, 2017.

[13] D. Ramesh and B. Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 2, no. 9, 2013.

[14] G. Ruß, R. Kruse, M. Schneider and P. Wagner, "Data Mining with Neural Networks for Wheat," *Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects.Lecture Notes in Computer Science,* vol. 5077, pp. 47-56, 2008.

[15] G. Ruß, "Data Mining of Agricultural Yield Data: A Comparison of Regression Models," *Advances in Data Mining. Applications and Theoretical Aspects.Lecture Notes in Computer Science,* vol. 5633, pp. 24-37, 2009.

[16] C. Zhang and Z. Liu, "Application of big data technology in agricultural Internet of Things," *International Journal of Distributed Sensor Networks,* vol. 15, no. 10, 2019.

[17] S. Shanmuganathan, P. Sallis and A. Narayanan, "Data Mining Techniques for Modelling the Influence of Daily Extreme Weather Conditions on Grapevine, Wine Quality and Perennial Crop Yield," *2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks,* pp. 90-95, 2020.

[18] "Self-Organizing Map," [Online]. Available: https://en.wikipedia.org/wiki/Self-organizing_map. [Accessed 23 june 2021].

[19] J. Palanichamy, S. Vinothini and B. Periyasamy, "A Study of Data Mining Techniques to Agriculture," *IJRIT International Journal of Research in Information Technology,* vol. 2, no. 4, pp. 306-313, 2014.

[20] A. Urtubia, J. R. Pe´rez-Correa, A. Soto and P. Pszczo´lkowski, "Using data mining techniques to predict industrial wine," *Food Control,* vol. 18, no. 12, pp. 1512-1517, 2007.

[21] C. Li and B. Niu, "Design of smart agriculture based on big data and Internet of things," *International Journal of Distributed Sensor Networks,* vol. 16, no. 5, 2020.

[22] I. A. Lakhiar, G. Jianmin, T. N. Syed, F. A. Chandio, N. A. Buttar and W. A. Qureshi, "Monitoring and Control Systems in Agriculture Using Intelligent Sensor Techniques: A Review of the Aeroponic," *Journal of Sensors,* vol. 1, no. 18, 2018.

[23] R. Sui and J. Thomasson, "Ground-Based Sensing System for Cotton Nitrogen Status Determination," 2016.

[24] P. K. V. SURYA and B. Ravi, "Time Series Data Analysis on Agriculture Food Production," *Conference: Smart Technologies in Data Science and Communication 2017,* December 2017.

[25] J. Fattah, L. Ezzine, Z. Aman and M. M. Harts et métiers, "Forecasting of demand using ARIMA model," *International Journal of Engineering Business Management,* no. 10(2):184797901880867, 2018.

[26] I.-S. Oh and T.-W. Yoo, "Time Series Forecasting of Agricultural Products' Sales Volumes Based on Seasonal Long Short-Term Memory," *Division of Computer Science and Engineering,* 18 November 2020.

[27] P. K. Sharma, "Model, Forecasting Maize Production in India using ARIMA," *Agro Economist - An International Journal,* 2018.

[28] "Box–Jenkins method," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Box%E2%80%93Jenkins_method. [Accessed 11 Novmeber 2021].

[29] J. Fattah, H. E. Moussami, Z. Aman and L. Ezzine, "Forecasting of demand using ARIMA model," *International Journal of Engineering Business Management,* October 2018.

[30] J. Brownlee, "A Gentle Introduction to SARIMA for Time Series Forecasting in Python," 21 August 2019. [Online]. Available: https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/. [Accessed 11 November 2021].

[31] J. Siebert, J. Groß and C. Schroth, "A systematic review of Python packages for time series analysis," 2021.

[32] "Using pandas to_datetime with timestamps," GeeksforGeeks, 23 August 2021. [Online]. Available: https://www.geeksforgeeks.org/using-pandas-to_datetime-with-timestamps/. [Accessed 15 October 2021].

[33] H. d. Harder, "The Ultimate Guide for Column Creation with Pandas DataFrames," towardsdata science, 12 October 2020. [Online]. Available: https://towardsdatascience.com/the-ultimate-guide-for-column-creation-with-pandas-dataframes-83b8c565110e. [Accessed 15 October 2021].

[34] E. Rencberoglu, "Fundamental Techniques of Feature Engineering for Machine Learning," 1 April 2019. [Online]. Available: https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114#8068. [Accessed 15 October 2020].

[35] T. Academy, "Working with Indexes in Time Series," Medium, 25 February 2021 . [Online]. Available: https://levelup.gitconnected.com/working-with-indexes-in-time-series-a2e00d220399. [Accessed 15 October 2021].

[36] B. T., "towards data science," Every Pandas Function You Can (Should) Use to Manipulate Time Series, 10 July 2021. [Online]. Available: https://towardsdatascience.com/every-pandas-function-you-can-should-use-to-manipulate-time-series-711cb0c5c749. [Accessed 15 October 2021].

[37] D. Gong, "How to Address Missing Data," Medium, 7 November 2020. [Online]. Available: https://medium.com/analytics-vidhya/how-to-address-missing-data-531ed964e68. [Accessed 17 October 2021].

[38] A. Swalin, "How to Address Missing Data," towards data science, 31 January 2018. [Online]. Available: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4. [Accessed 17 October 2021].

[39] J. Brownlee, "How To Resample and Interpolate Your Time Series Data With Python," Machine Learning Mastery, 11 February 2020. [Online]. Available: https://machinelearningmastery.com/resample-interpolate-time-series-data-python/. [Accessed 17 October 2021].

[40] R. Agrawal, "Interpolation – Power of Interpolation in Python to fill Missing Values," Analytics Vidhya, 1 June 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/power-of-interpolation-in-python-to-fill-missing-values/. [Accessed 17 October 2021].

[41] B. T., "Advanced Time Series Analysis in Python: Seasonality and Trend Analysis (Decomposition), Autocorrelation," towards data science, 13 July 2021. [Online]. Available: https://towardsdatascience.com/advanced-time-series-analysis-in-python-decomposition-autocorrelation-115aa64f475e. [Accessed 21 October 2021].

[42] D. Mallick, "Interpreting ACF or Auto-correlation plot," Analytics Vidhya, 25 Novenber 2020. [Online]. Available: https://medium.com/analytics-vidhya/interpreting-acf-or-auto-correlation-plot-d12e9051cd14. [Accessed 21 October 2021].

[43] R. Singh, "Interpreting ACF or Auto-correlation plot," Medium, 18 July 2021. [Online]. Available: https://towardsdatascience.com/create-weather-proof-validations-for-your-time-series-forecasting-model-d456a1037c4f. [Accessed 21 October 2021].

[44] C. Maklin, 26 May 2019. [Online]. Available: https://towardsdatascience.com/machine-learning-part-19-time-series-and-autoregressive-integrated-moving-average-model-arima-c1005347b0d7. [Accessed 25 October 2021].

[45] A. Choudhary, S. Kumar, M. Sharma and K. P. Sharma, "A Framework for Data Prediction and Forecasting in WSN with Auto ARIMA," *Wireless Pers Commun,* 2021.

[46] M. Kosaka, "Efficient Time-Series Analysis Using Python's Pmdarima Library," towards data science, 5 January 2021. [Online]. Available: https://towardsdatascience.com/efficient-time-series-using-pythons-pmdarima-library-f6825407b7f0. [Accessed 25 October 2021].

[47] N. Hebbar, "Time Series Forecasting With ARIMA Model in Python for Temperature Prediction," Medium, 18 September 2020. [Online]. Available: https://medium.com/swlh/temperature-forecasting-with-arima-model-in-python-427b2d3bcb53. [Accessed 27 October 2021].

[48] "Difference Between AIC and BIC," DifferenceBetween.net, [Online]. Available: http://www.differencebetween.net/miscellaneous/difference-between-aic-and-bic/. [Accessed 27 October 20201].

[49] J. Moody, "What does RMSE really mean?," towards data science, 6 September 2019. [Online]. Available: https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e. [Accessed 27 October 2021].

[50] "5 Down & Dirty Tips for Clean Data Dashboards," iDashboards, 4 April 2017. [Online]. Available: https://www.idashboards.com/blog/2017/04/04/5-tips-clean-data-dashboards/. [Accessed 28 October 2021].

[51] J. D. Dotson, "How Temperature & Humidity are Related," 23 April 2018. [Online]. Available: https://sciencing.com/temperature-ampamp-humidity-related-7245642.html. [Accessed 28 October 2021].

[52] "Filters vs Slicers – Which Is a Better Choice When Designing Reports in Power BI?," Unity Group, 16 January 2020. [Online]. Available: https://www.unitygroup.com/blog/filters-vs-slicers-which-is-a-better-choice-when-designing-reports-in-power-bi/. [Accessed 16 October 2021].

[53] N. Golabi, "Choosing the right colors for your dashboard," ArcGIS Blog, 30 December 2019. [Online]. Available: https://www.esri.com/arcgis-blog/products/ops-dashboard/decision-support/choosing-the-right-colors-for-your-dashboard/. [Accessed 3 November 2021].

[54] "Color Blindness, Contrast and Data Visualizations," mySidewalk, [Online]. Available: https://dashboards.mysidewalk.com/style-guide-for-dashboards/color. [Accessed 5 November 2021].

[55] "What is a Combination Chart?," [Online]. Available: What is a Combination Chart?. [Accessed 5 November 2021].

[56] B. S. Chadha, "Bullet Chart- Advanced Custom Visuals for Power BI," 18 November 2019. [Online]. Available: https://xviz.com/blogs/bullet-chart-advanced-custom-visuals-for-power-bi/. [Accessed 5 November 2021].

[57] D. Gong, "Feature Selection and EDA in Machine Learning," towards data science, 24 May 2021. [Online]. Available: https://towardsdatascience.com/feature-selection-and-eda-in-python-c6c4eb1058a3. [Accessed 29 October 2021].

[58] J. Frost, "Guidelines for Removing and Handling Outliers in Data," 2021. [Online]. Available: https://statisticsbyjim.com/basics/remove-outliers/. [Accessed 29 October 2021].

[59] "The Best Light Intensity for Plants Indoors," Sunday Gardener, 7 March 2021 . [Online]. Available: https://www.sundaygardener.net/the-best-light-intensity-for-plants-indoors/. [Accessed 29 October 2021].

[60] R. Hyndman and G. Athanasopoulos, "Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia," *2nd edition, OTexts: Melbourne, Australia,* 2018.

[61] J. Frost, "Autocorrelation and Partial Autocorrelation in Time Series Data," [Online]. Available: Autocorrelation and Partial Autocorrelation in Time Series Data. [Accessed 2 November 2021].

[62] Y. Verma, "Complete Guide To Dickey-Fuller Test In Time-Series Analysis," 18 August 2021. [Online]. Available: https://analyticsindiamag.com/complete-guide-to-dickey-fuller-test-in-time-series-analysis/. [Accessed 5 November 2021].

[63] B. T., "How to Remove Non-Stationarity in Time Series Forecasting," towards data science, 18 July 2021. [Online]. Available: https://towardsdatascience.com/how-to-remove-non-stationarity-in-time-series-forecasting-563c05c4bfc7. [Accessed 5 November 2021].

8 Appendix

8.1 Key Identifications of Devices

Table 8-1: Key Identification of Devices for Air Temperature

Air Temperature	
Device	Key
Indoor Device 1	86
Indoor Device 2	108
Outdoor Device 1&2	55,46

Table 8-2: Key Identification of Devices for Air Humidity

Air Humidity	
Device	Key
Indoor Device 1	85
Indoor Device 2	105
Outdoor Device 1&2	54,45

Table 8-3: Key Identification of Water Temperature

Water Temperature	
Device	Key
Indoor Device 1	93,89
Indoor Device 2	103,104
Indoor Device 4	60
Outdoor Device 1&2	58,53
Outdoor Device 3	60

Table 8-4: Key Identification of Devices for PH Value

	pH
Device	Key
Indoor Device 4	59
Outdoor Device 3	59

Table 8-5: Key Identification of Devices for Water EC Value

	Water EC
Device	Key
Indoor Device 4	61
Outdoor Device 3	61

Table 8-6: Key Identification of Devices for Electric Current

Electric current	
Device	Key
Indoor Device 1 (LED Lights)	90,91
Indoor Device 2 (LED Lights)	106,107
Indoor Device 3 (LED Lights)	78
Outdoor Device 4 (Water pump)	78

Table 8-7: Key Identification of Devices for Light Intensity

	Light Intensity (LUX)
Device	Key
Outdoor Device 1&2	66,67,68,69

Table 8-8: Key Identification of Devices for Water Float State

Water float state (0=start refilling water ,1=stop refilling water)	
Device	Key
Indoor Device 3	77
Outdoor Device 4	77

8.2 Screenshots of Dashboard Design

8.2.1 Sensors Dashboard

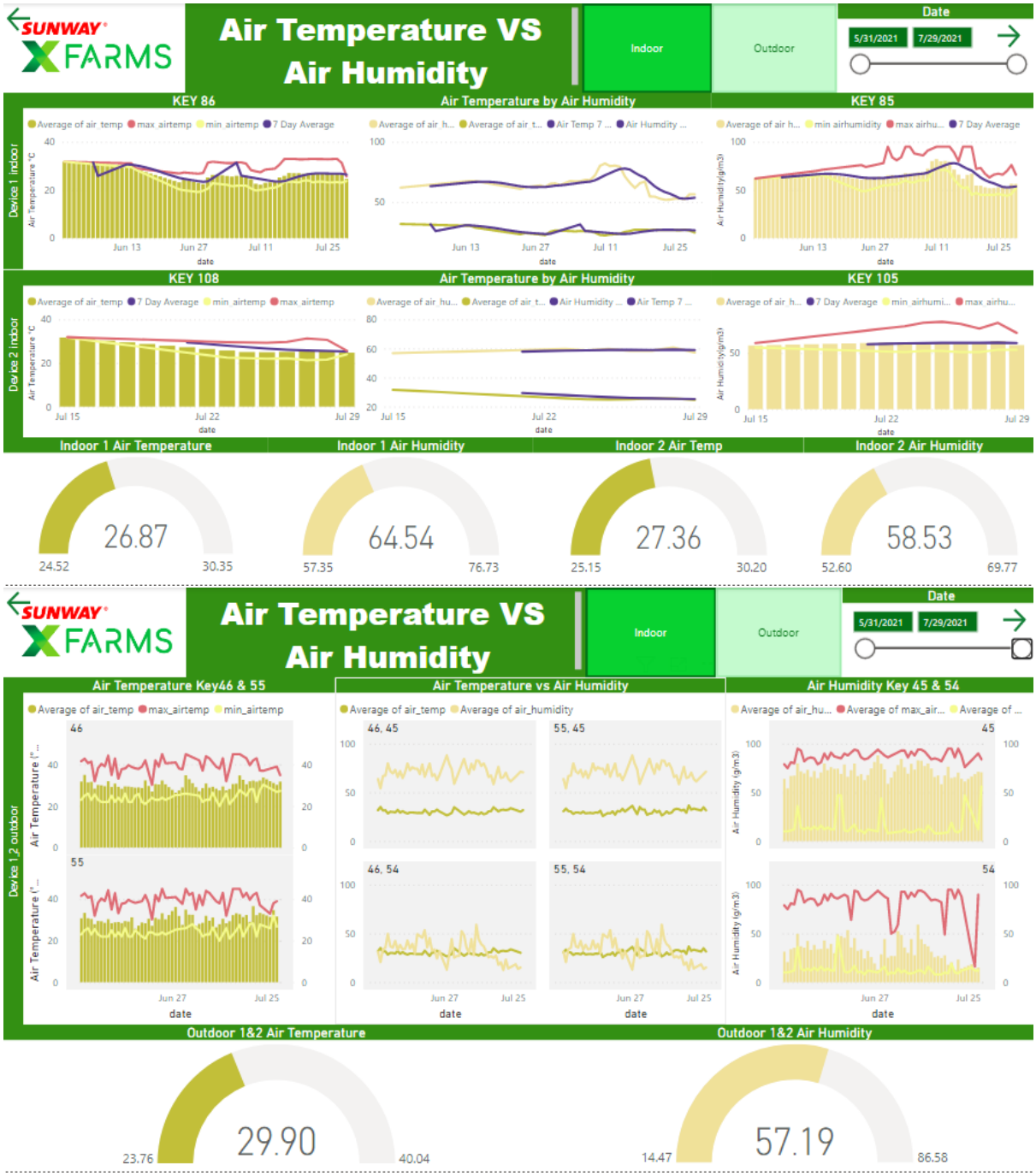


Figure 8-1: Dashboard of Air Temperature VS. Air Humidity from Indoor and Outdoor Devices

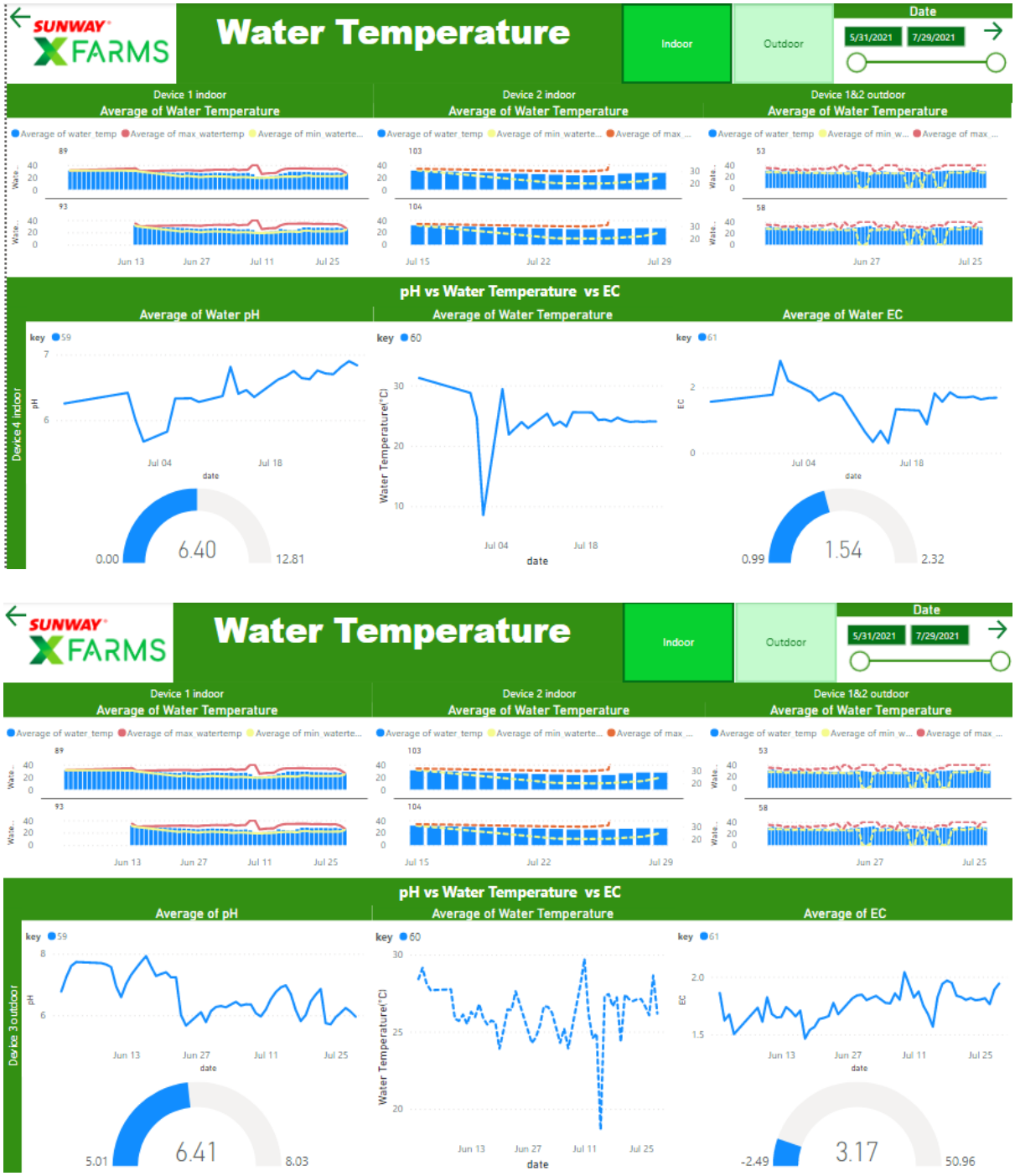


Figure 8-2: Dashboard of Water Temperature Collected From Indoor and Outdoor Devices

Figure 8-3: *Dashboard of Electric Current Collected From Indoor and Outdoor Devices*

8.2.2 Yield Analysis Dashboard

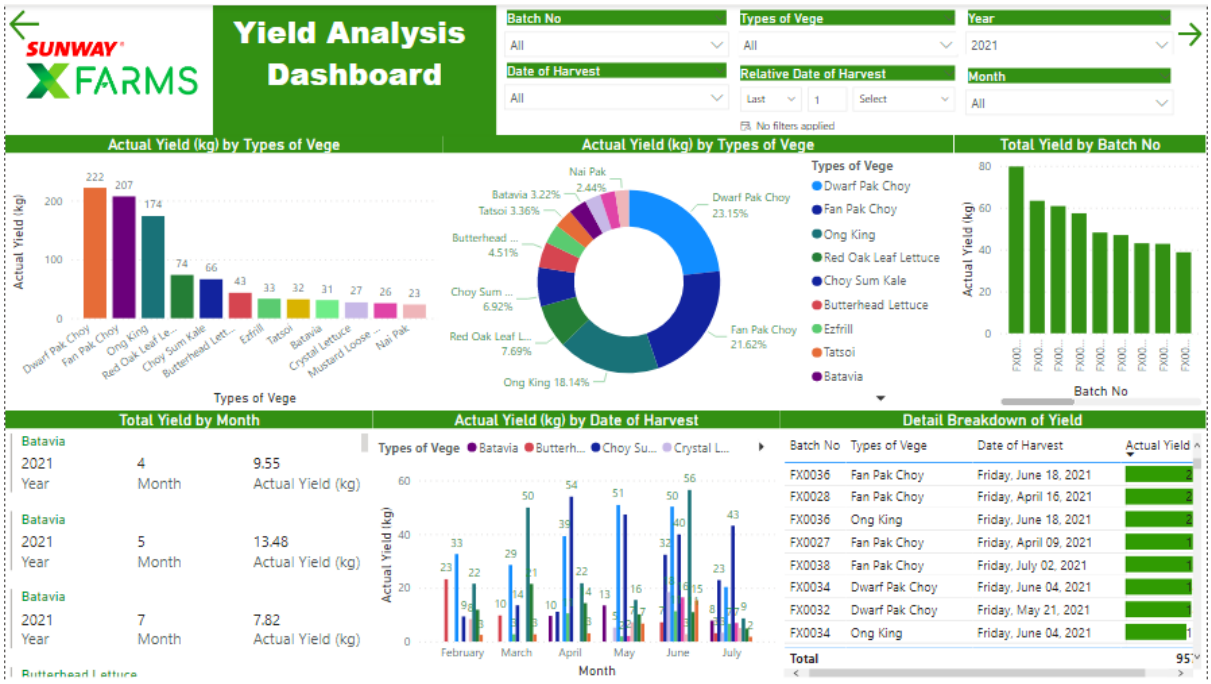


Figure 8-4: Yield Analysis Dashboard

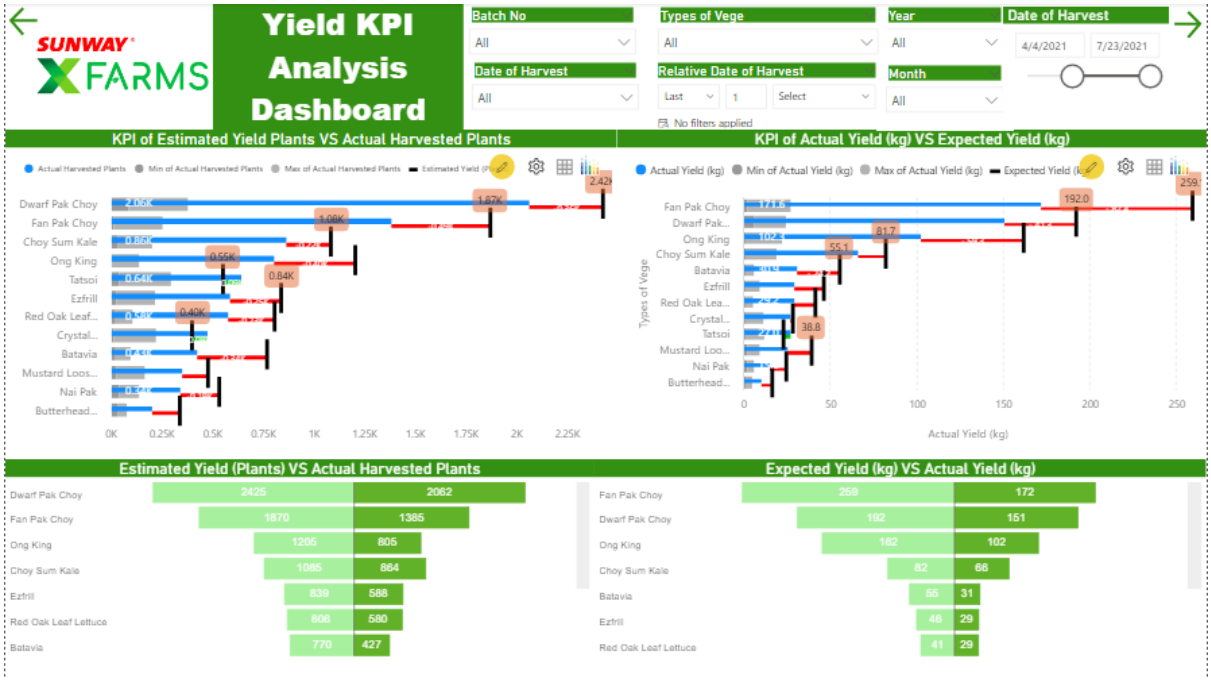


Figure 8-5: Yield KPI Analysis Dashboard

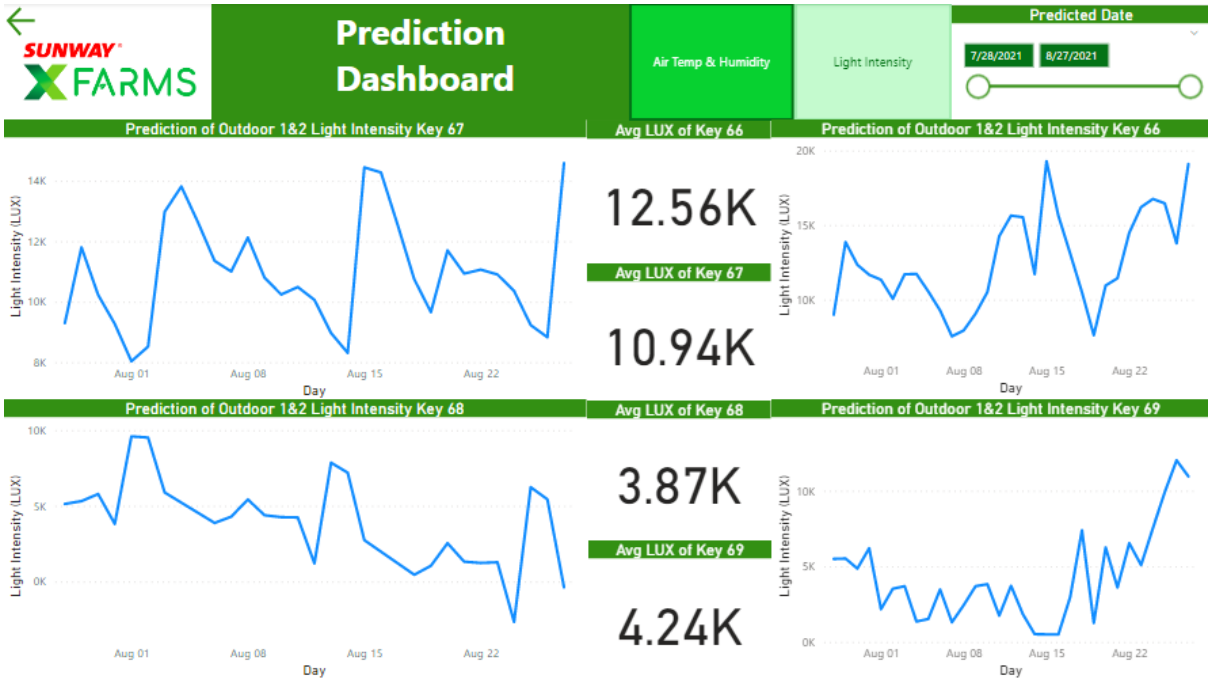
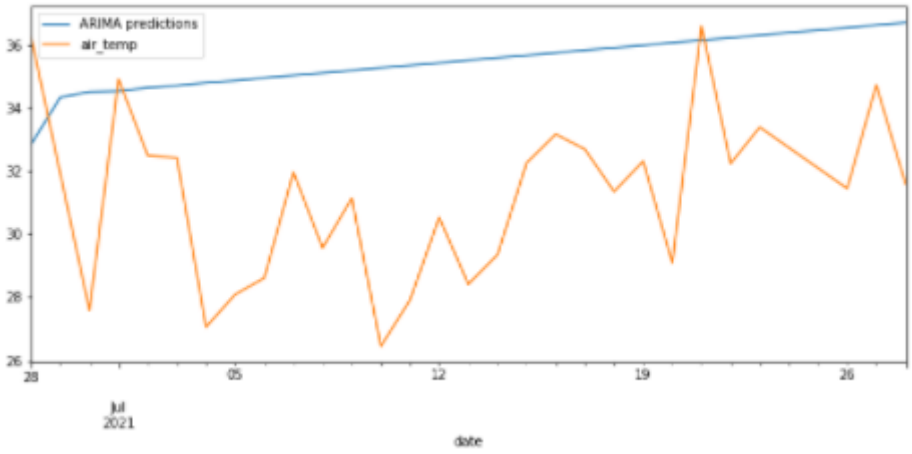


Figure 8-6: Prediction Dashboard

8.3 Screenshots of Results of ARIMA model using ACF/PACF plots



ARIMA Model Results

Dep. Variable:	D.air_temp	**No. Observations:**	82
Model:	ARIMA(1, 1, 2)	**Log Likelihood**	-157.299
Method:	css-mle	**S.D. of innovations**	1.644
Date:	Tue, 23 Nov 2021	**AIC**	324.598
Time:	16:02:49	**BIC**	336.632
Sample:	04-08-2021	**HQIC**	329.430
	- 06-28-2021		

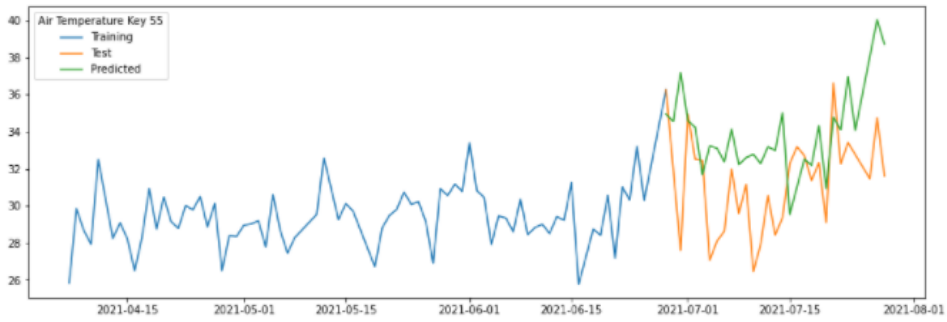
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0798	0.080	0.994	0.320	-0.078	0.237
ar.L1.D.air_temp	-0.6067	0.988	-0.614	0.539	-2.544	1.330
ma.L1.D.air_temp	0.0166	0.994	0.017	0.987	-1.932	1.965
ma.L2.D.air_temp	-0.3284	0.608	-0.540	0.589	-1.520	0.863

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.6482	+0.0000j	1.6482	0.5000
MA.1	-1.7200	+0.0000j	1.7200	0.5000
MA.2	1.7704	+0.0000j	1.7704	0.0000

Figure 8-7: *Result of ARIMA of Outdoor Device 1&2 Air Temperature Key 55 generated by determining the parameters order using ACF/PACF plot*

8.4 Screenshots of Results of ARIMA model using Auto ARIMA



SARIMAX Results

Dep. Variable:	air_temp	No. Observations:	113
Model:	SARIMAX(0, 1, 1)x(0, 1, [], 30)	Log Likelihood	-194.053
Date:	Tue, 23 Nov 2021	AIC	392.106
Time:	18:15:18	BIC	398.919
Sample:	04-07-2021	HQIC	394.038
	- 07-28-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.9649	0.040	-23.966	0.000	-1.044	-0.886
sigma2	6.4401	0.866	7.436	0.000	4.743	8.137

Ljung-Box (L1) (Q):	1.28	Jarque-Bera (JB):	4.59
Prob(Q):	0.26	Prob(JB):	0.10
Heteroskedasticity (H):	2.02	Skew:	0.45
Prob(H) (two-sided):	0.07	Kurtosis:	3.73

Figure 8-8 : *Result of ARIMA model of Outdoor Device 1&2 Air Temperature Key 55 generated by determining the order of the parameters using Auto ARIMA*

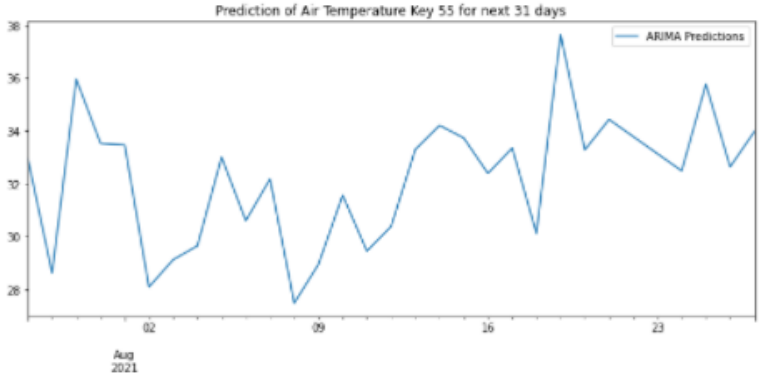
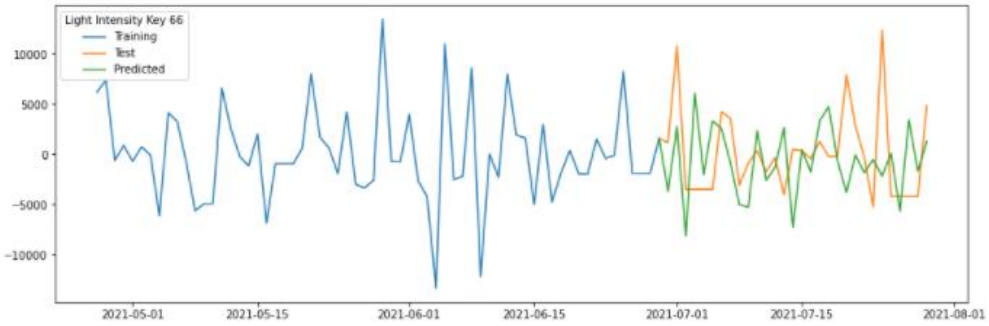


Figure 8-9: *Prediction of* Outdoor Device 1&2 *Air Temperature Key 55 for Next 31 Days*



SARIMAX Results

Dep. Variable:	y	No. Observations:	64
Model:	SARIMAX(1, 1, 0)x(2, 1, 0, 12)	Log Likelihood	-519.854
Date:	Tue, 23 Nov 2021	AIC	1047.707
Time:	18:42:36	BIC	1055.434
Sample:	0	HQIC	1050.660
	- 64		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5259	0.133	-3.961	0.000	-0.786	-0.266
ar.S.L12	-0.9869	0.200	-4.932	0.000	-1.379	-0.595
ar.S.L24	-0.4726	0.201	-2.357	0.018	-0.866	-0.080
sigma2	3.954e+07	1.45e-09	2.72e+16	0.000	3.95e+07	3.95e+07

Ljung-Box (L1) (Q):	1.83	Jarque-Bera (JB):	5.26
Prob(Q):	0.18	Prob(JB):	0.07
Heteroskedasticity (H):	1.10	Skew:	0.73
Prob(H) (two-sided):	0.84	Kurtosis:	3.59

Figure 8-10: *Result of ARIMA model of Outdoor Device 1&2 Light Intensity Key 66 generated by determining the order of the parameters using Auto ARIMA*

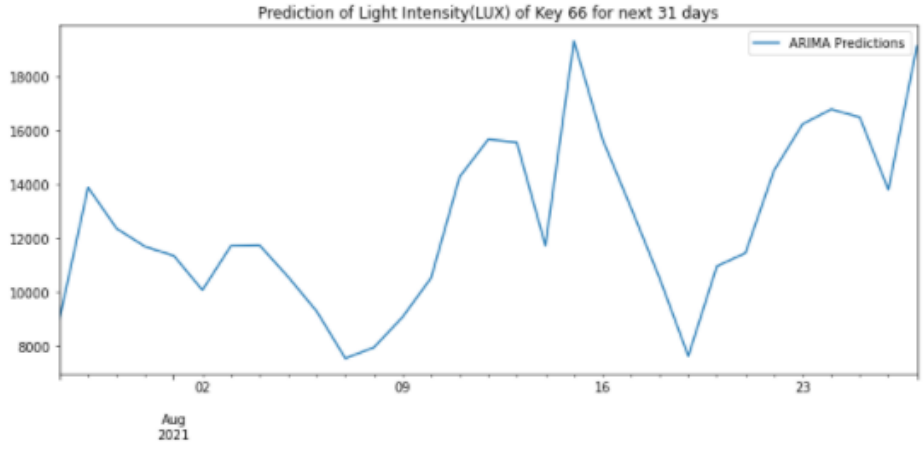
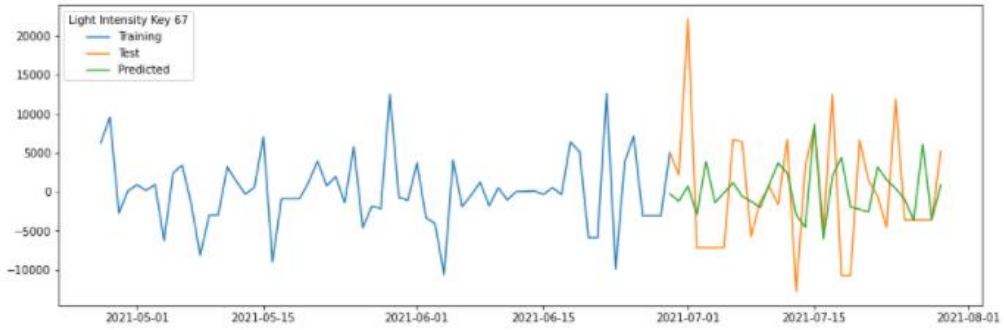


Figure 8-11: *Prediction of Outdoor Device 1&2 Light Intensity Key 66 for Next 31 Days*



SARIMAX Results

Dep. Variable:	light_LUX	**No. Observations:**	95
Model:	SARIMAX(2, 1, 1)x(1, 1, 1, 12)	**Log Likelihood**	-838.368
Date:	Tue, 23 Nov 2021	**AIC**	1688.736
Time:	19:02:21	**BIC**	1703.177
Sample:	04-26-2021	**HQIC**	1694.534
	- 07-29-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4906	0.397	1.235	0.217	-0.288	1.269
ar.L2	0.0293	0.241	0.121	0.903	-0.444	0.503
ma.L1	-0.8870	0.309	-2.805	0.005	-1.473	-0.261
ar.S.L12	-0.2248	0.252	-0.893	0.372	-0.718	0.269
ma.S.L12	-0.6931	0.238	-2.913	0.004	-1.160	-0.227
sigma2	6.84e+07	6.23e-09	1.1e+16	0.000	6.84e+07	6.84e+07

Ljung-Box (L1) (Q):	0.08	**Jarque-Bera (JB):**	39.31
Prob(Q):	0.78	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.97	**Skew:**	0.84
Prob(H) (two-sided):	0.08	**Kurtosis:**	5.95

Figure 8-12: *Result of ARIMA model of Outdoor Device 1&2 Light Intensity Key 67 generated by determining the order of the parameters using Auto ARIMA*

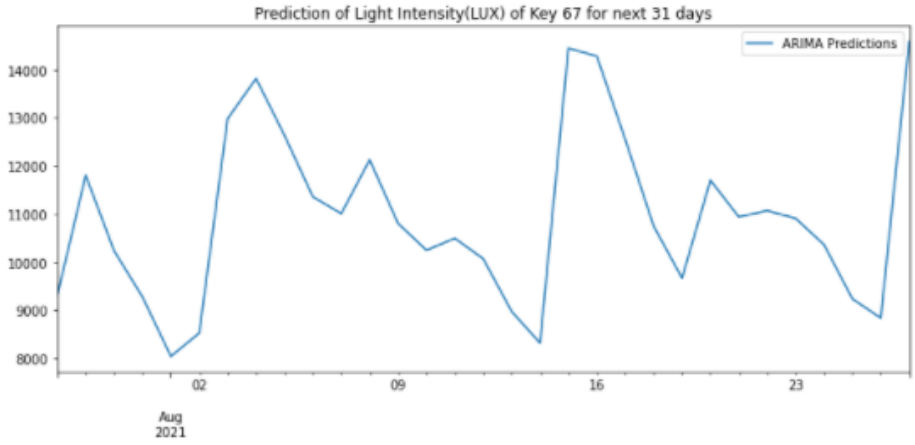
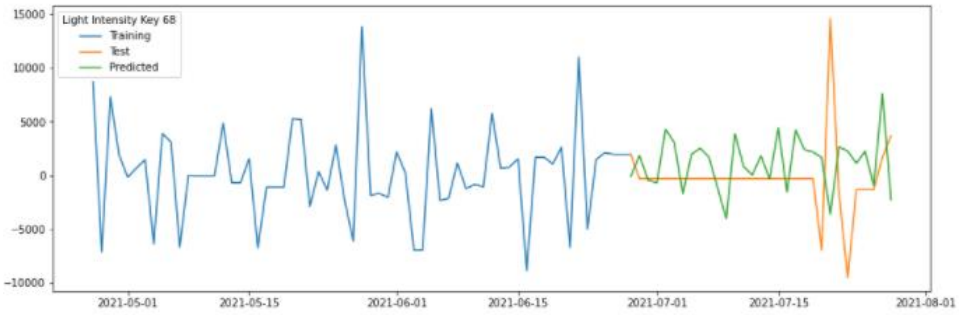


Figure 8-13 : *Prediction of Outdoor Device 1&2 Light Intensity Key 67 for Next 31 Days*



SARIMAX Results

Dep. Variable:	light_LUX	No. Observations:	94
Model:	SARIMAX(0, 1, 2)x(2, 1, [], 12)	Log Likelihood	-795.366
Date:	Tue, 23 Nov 2021	AIC	1600.733
Time:	19:14:56	BIC	1612.705
Sample:	04-26-2021	HQIC	1605.536
	- 07-28-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.3679	0.147	-2.506	0.012	-0.656	-0.080
ma.L2	-0.0808	0.210	-0.385	0.701	-0.493	0.331
ar.S.L12	-0.6395	0.208	-3.074	0.002	-1.047	-0.232
ar.S.L24	-0.3194	0.164	-1.952	0.051	-0.640	0.001
sigma2	2.75e+07	3.19e-09	8.62e+15	0.000	2.75e+07	2.75e+07

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	22.68
Prob(Q):	0.91	Prob(JB):	0.00
Heteroskedasticity (H):	0.59	Skew:	0.73
Prob(H) (two-sided):	0.18	Kurtosis:	5.15

Figure 8-14: *Result of ARIMA model of Outdoor Device 1&2 Light Intensity Key 68 generated by determining the order of the parameters using Auto ARIMA*

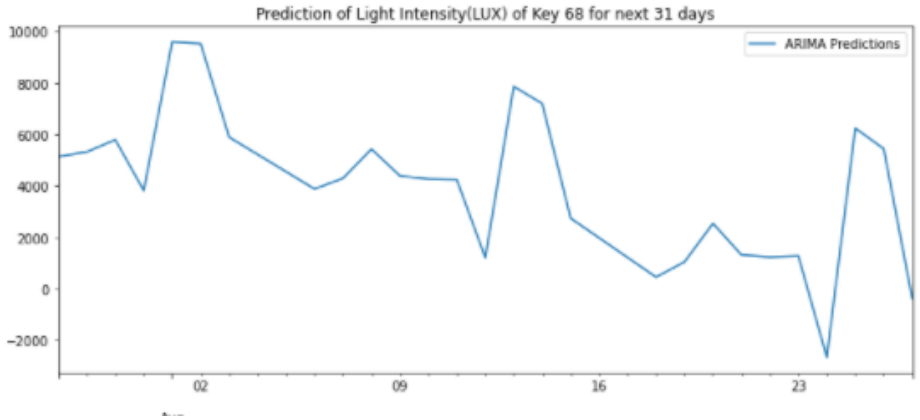
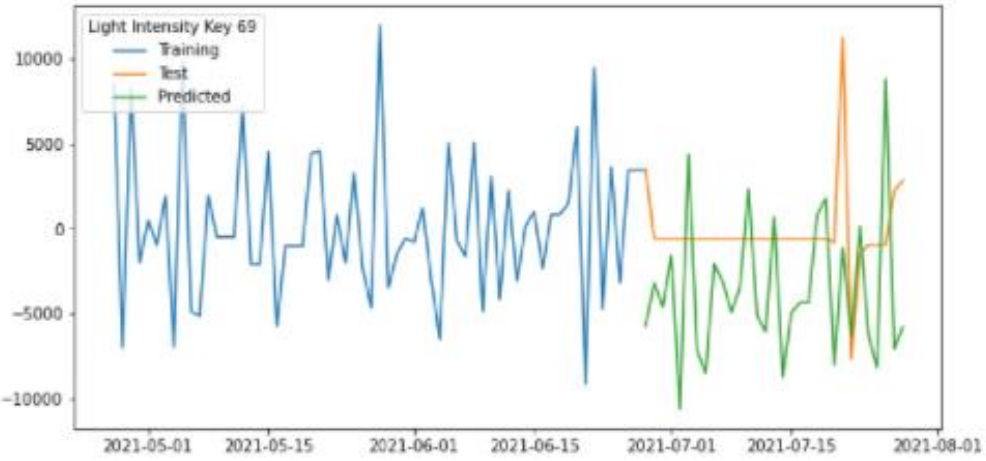


Figure 8-15*: Prediction of Outdoor Device 1&2 Light Intensity Key 68 for Next 31 Days*



SARIMAX Results

Dep. Variable:	light_LUX	**No. Observations:**	94
Model:	SARIMAX(0, 1, 0)x(1, 1, [1], 30)	**Log Likelihood**	-616.898
Date:	Tue, 23 Nov 2021	**AIC**	1239.797
Time:	19:20:48	**BIC**	1246.226
Sample:	04-26-2021	**HQIC**	1242.326
	- 07-28-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.S.L30	-0.3959	0.108	-3.655	0.000	-0.608	-0.184
ma.S.L30	-0.9965	0.158	-6.290	0.000	-1.307	-0.686
sigma2	1.226e+07	1.3e-08	9.45e+14	0.000	1.23e+07	1.23e+07

Ljung-Box (L1) (Q):	22.37	**Jarque-Bera (JB):**	16.53
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.70	**Skew:**	0.81
Prob(H) (two-sided):	0.42	**Kurtosis:**	4.92

Figure 8-16 : *Result of ARIMA model of Outdoor Device 1&2 Light Intensity Key 69 generated by determining the order of the parameters using Auto ARIMA*

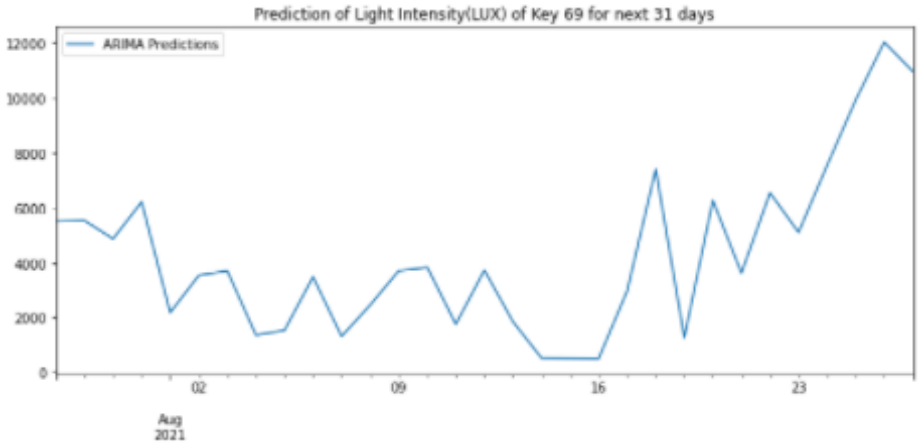
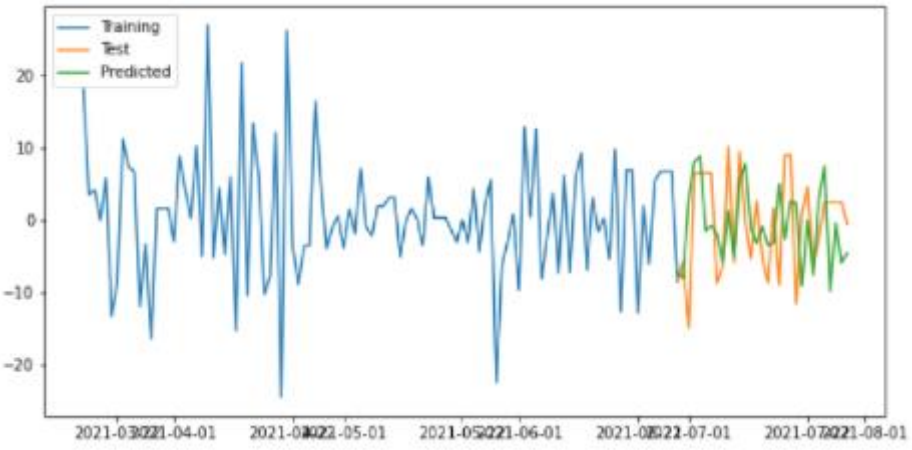


Figure 8-17*: Prediction of Outdoor Device 1&2 Light Intensity of Key 69 for Next 31 Days*



Dep. Variable:	air_humidity	No. Observations:	137
Model:	SARIMAX(3, 1, 3)x(1, 1, [], 30)	Log Likelihood	-382.567
Date:	Tue, 23 Nov 2021	AIC	781.134
Time:	20:24:53	BIC	802.441
Sample:	03-15-2021	HQIC	789.770
	- 07-29-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1119	0.282	0.397	0.691	-0.441	0.664
ar.L2	-0.4162	0.281	-1.481	0.139	-0.967	0.135
ar.L3	0.0247	0.299	0.083	0.934	-0.562	0.611
ma.L1	-0.4768	0.276	-1.728	0.084	-1.018	0.064

Figure 8-18*: Result of ARIMA model of Outdoor Device 1&2 Air Humidity Key 45 generated by determining the order of the parameters using Auto ARIMA*

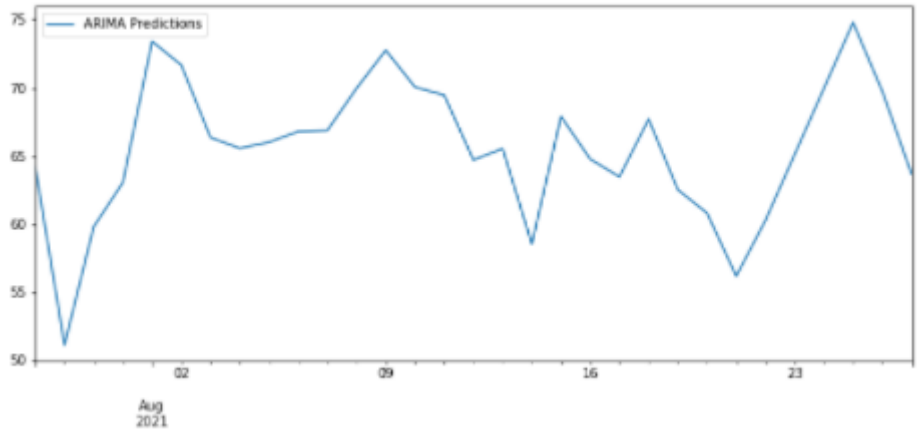


Figure 8-19:

Figure 8-20: *Prediction of* Outdoor Device 1&2 Air Humidity *Key 45 for Next 31 Days*

SARIMAX Results

Dep. Variable:	air_humidity	**No. Observations:**	113
Model:	SARIMAX(3, 1, 0)x(1, 1, 0, 30)	**Log Likelihood**	-315.902
Date:	Tue, 23 Nov 2021	**AIC**	641.805
Time:	20:29:53	**BIC**	653.838
Sample:	04-07-2021	**HQIC**	646.636
	- 07-28-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5480	0.118	-4.638	0.000	-0.780	-0.316
ar.L2	-0.1882	0.154	-1.225	0.221	-0.489	0.113
ar.L3	-0.1001	0.142	-0.704	0.482	-0.379	0.179
ar.S.L30	-0.3144	0.165	-1.909	0.056	-0.637	0.008
sigma2	124.6055	23.905	5.213	0.000	77.753	171.458

Ljung-Box (L1) (Q):	0.03	**Jarque-Bera (JB):**	1.13
Prob(Q):	0.86	**Prob(JB):**	0.57
Heteroskedasticity (H):	0.97	**Skew:**	0.26
Prob(H) (two-sided):	0.93	**Kurtosis:**	2.74

Figure 8-21: *Result of ARIMA model of Outdoor Device 1&2 Air Humidity Key 54 generated by determining the order of the parameters using Auto ARIMA*

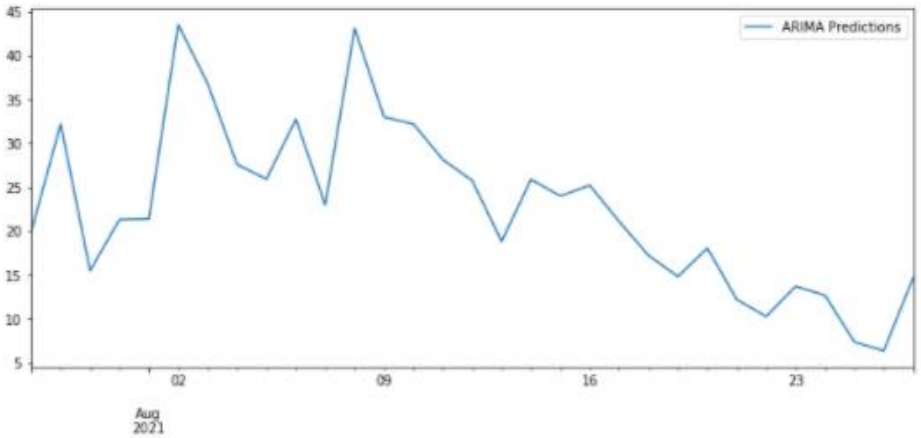


Figure 8-22*: Prediction of Outdoor Device 1&2 Air Humidity Key 54 for Next 31 Days*