|  |  |
| --- | --- |
| **RMIT University**  **School of Science and Technology (SST) EEET2485 Research Methods for Engineers** |  |

**Data Analysis Report**

Project: Room Occupancy Estimation Data Set

**RMIT University Vietnam Assignment Cover Page**

|  |  |
| --- | --- |
| **Course Code** | * EEET2485 |
| **Course Name** | * Research Methods for Engineers |
| **Location/ Campus** | * RMIT Vietnam Saigon South |
| **Assignment** | * Group Project |
| **Student names** | * Nguyen Dang Huynh Chau (s3777214) * Truong Phu Khang (s3814172) * Le Anh Quan (s3777351) |
| **Lecturer** | * Dr. Son Vu Truong Dao |
| **Group** | * 10 |
| **Assignment due date** | * May 28th, 2022 |
| **Date of Submission** | * May 28th, 2022 |
| **Page Count** | * 12 |

Table of Contents

[Table of Contents 3](#_Toc104655156)

[I. Executive Summary: 4](#_Toc104655157)

[II. Introduction: 4](#_Toc104655158)

[**1.** **Dataset Context:** 4](#_Toc104655159)

[**2.** **Research Question:** 5](#_Toc104655160)

[**A.** **Analyze correlations between each parameter: (Spearman’s rho)** 5](#_Toc104655161)

[**B.** **Analyze room occupancy condition: (Factor analysis)** 5](#_Toc104655162)

[**C.** **Analyze room occupancy in different conditions: (plot, t-test)** 5](#_Toc104655163)

[III. Data analysis: 5](#_Toc104655164)

[**1.** **Outliners:** 5](#_Toc104655165)

[**2.** **Analyze correlations between parameters:** 6](#_Toc104655166)

[**A.** **Methodology:** 6](#_Toc104655167)

[**B.** **Correlation Testing:** 6](#_Toc104655168)

[**C.** **Multicorrelation:** 6](#_Toc104655169)

[**3.** **Analyze Room Occupation causation:** 6](#_Toc104655170)

[**A.** **Methodology:** 7](#_Toc104655171)

[**B.** **Bartlett's test:** 7](#_Toc104655172)

[**C.** **Kaiser-Meyer-Olkin (KMO) Test:** 7](#_Toc104655173)

[**D.** **Choosing the number of factors:** 7](#_Toc104655174)

[**E.** **Highest Correlation between features and “Room\_Occupancy\_Count”:** 8](#_Toc104655175)

[**F.** **Decision Tree Model:** 9](#_Toc104655176)

[**4.** **Analyze Room Occupation in different conditions:** 9](#_Toc104655177)

[**A.** **Methodology:** 9](#_Toc104655178)

[**B.** **Relationship plotting:** 9](#_Toc104655179)

[**C.** **Mann-Whitney Testing:** 11](#_Toc104655180)

[IV. Conclusion: 12](#_Toc104655181)

[**1.** **Outlier:** 13](#_Toc104655182)

[**2.** **Correlation Between Parameter:** 14](#_Toc104655183)

# **Executive Summary:**

The world population has shown a considerable upward trend recently, so that the buying and selling an occupation has received great academic attention. In this project, we have conducted research on a provided dataset which is about the “Room Occupancy Estimation” by Adarsh Pal Singh and Dr. Sachin Chaudhari [1]. We examined machine behavior pattern in numerous aspects using approaches and analytical tools provided in Pandas, including factor collaborations, data probability distributions, and room occupancy causes. Based on the defined aspects, we developed a list of research questions to investigate the nature of our dataset based on the mentioned features. We devised our trials using research criteria and came up with the following conclusions:

1. **Finding 1:** The Temperature sensors have a correlation with each other.
2. **Finding 2:** The Light sensors have a correlation with each other.
3. **Finding 3:** The Sound sensors have a correlation with each other.
4. **Finding 4:** The CO2 sensors have a correlation with each other.
5. **Finding 5:** The PIR sensors have a correlation with each other.
6. **Finding 6:** S1\_Light and S5\_CO2 are the most important factors in the dataset
7. **Finding 7:** The S5\_CO2 in the rooms having 2 people is lower than rooms having 1 person since the light and temperature in the 1-person rooms are higher.

# **Introduction:**

## **Dataset Context:**

In this project, our group used a dataset called Room Occupancy Estimation. It is a set of collection of numbers from 4 different types of sensors, each type of sensor has 4 dividual sensors placed on a fixed position in the room. The main idea of this setup is to avoid the room is too big for one sensor to cover or, in the room there are other small room, areas that one sensor cannot cover. The dataset contains over 10,000 data and 18 different attributes

1. **Time**: acknowledge the time in a day that specific data is being collected.
2. **Date**: acknowledge date that specific data is being collected.
3. **S1\_Temp**: temperature sensor nodes 1 in the room.
4. **S2\_Temp**: temperature sensor nodes 2 in the room.
5. **S3\_Temp**: temperature sensor nodes 3 in the room.
6. **S4\_Temp**: temperature sensor nodes 4 in the room.
7. **S1\_Light**: brightness sensor nodes 1 in the room.
8. **S2\_Light**: brightness sensor nodes 2 in the room.
9. **S3\_Light**: brightness sensor nodes 3 in the room.
10. **S4\_Light**: brightness sensor nodes 4 in the room.
11. **S1\_Sound**: sound sensor nodes 1 in the room.
12. **S2\_Sound**: sound sensor nodes 2 in the room.
13. **S3\_Sound**: sound sensor nodes 3 in the room.
14. **S4\_Sound**: sound sensor nodes 4 in the room.
15. **S5-CO2**: CO2 sensor nodes 5 in the room.
16. **S5\_CO2\_Slope**: CO2 slope sensor nodes 1 in the room.
17. **S6\_PIR**: digital passive infrared (PIR) sensor nodes 6 in the room.
18. **S7\_PIR**: digital passive infrared (PIR) sensor nodes 7 in the room.

## **Research Question:**

### **Analyze correlations between each parameter: (Spearman’s rho)**

* **Correlation between Temperature Feature:**
  + Correlation between S1\_Temp and S2\_Temp.
  + Correlation between S1\_Temp and S3\_Temp.
  + Correlation between S1\_Temp and S4\_Temp.
* **Correlation between Light Feature:**
  + Correlation between S1\_Light and S2\_Light.
  + Correlation between S1\_Light and S2\_Light.
  + Correlation between S1\_Light and S4\_Light
* **Correlation between Sound Feature:**
  + Correlation between S1\_Sound and S2\_Sound.
  + Correlation between S1\_Sound and S3\_Sound.
  + Correlation between S1\_Sound and S4\_Sound.
* **Correlation between CO2 Feature:**
* Correlation between S5\_CO2 and S5\_CO2\_Slope.
* **Correlation between PIR Feature:**
* Correlation between S6\_PIR and S7\_PIR.

### **Analyze room occupancy condition: (Factor analysis)**

* + What is the main factor in the growth of room occupancy count?
  + Which factors often occur together when there is at least one person in the room?

1. **Analyze room occupancy in different conditions: (plot, t-test)**
   * How the main features influence the “Room\_Occupancy\_Count”?

Section A of our research explains the link between the parameters and how changing one might influence the others. Section B aids in identifying the primary cause of room occupancy to better forecast how and why room is not empty at a specific time. Section C examines the state of each room occupancy, which serves as an indicator of why a certain failure occurred and aids in predicting in a specific case in a room would be occupied.

# **Data analysis:**

## **Outliners:**

Since most of the dataset is numerical values, the outlier occurrence is extremely high, detecting and dealing with outliers is essential for reducing the noise of the data. Firstly, the box plots were applied to detect the outliers. However, conducting an outlier test on data is the next step in determining whether it is an outlier. It is called IQR-distance from Median. The Z-Score, Quantile Filter, and IQR-distance from Median are three common methods for dealing with outliers. The Z-Score is used to calculate the mean and standard deviation. As a result, the evaluation of outliers may be skewed. The Quantile Filter is a superior strategy; however, it can remove some of the parts that aren't outliers, resulting in a significant quantity of good data being lost while the outliers remain. After dealing with outliers, it is recognized that there are 2 different scenarios that may happen during collecting outliers in the dataset which are the outliers are equal or less than 6 percent and the outliers is over 6 percent. For those outliers that have a percentage lower or equal to 6 percent, we will decide to drop all the outliers that met the condition that we set since the percentage is so small to have any consequence on the dataset. Nevertheless, if the outliers value is greater than 6 percent, further analysis process must be applied. Our data has temperature, light, sound, CO2, and PIR features of the rooms, so if those features are in the extreme states the room should have at least one person, otherwise, it is just the noise data. Since then, unless all the outliers have at least one person in the room, all the noise data having no person are all dropped.

Before removing the outliers, the data includes 10,129 rows and 19 columns. There are still 8,935 rows and 19 columns after removing with the outliers. Since then, there are approximately 12 percent of the outliers in our dataset.

## **Analyze correlations between parameters:**

### **Methodology:**

To find the most appropriate approach to find the correlation between columns, the data distribution plot must be plotted. We can express a big quantity of data and its frequency using a histogram. However, the density plot is a smoothed and continuous version of the histogram calculated from the data. According to the density plot, there is no normal distribution in the dataset. Since then, Spearman’s rho is the most appropriate method to test the correlation. [2]

### **Correlation Testing:**

**RQ1: What is the correlation between S1\_Temp and S2\_Temp?**

**H0:** S2\_Temp increases when S1\_Temp increases from 0 to 26.5.

**H0:** S2\_Temp is not correlated with S1\_Temp**.**

At the Significant Level = 0.05

Text

Description automatically generated with medium confidence

**Observation:**

At Spearman’s rho correlation score = 0.97 and df = 8935, the positive correlation between S1\_Temp and S2\_Temp is proven with p\_value = 0.0 < significant values = 0.05. The positive correlation score 0.97 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S2\_Temp increases when S1\_Temp increases from 0 to 26.5”.

RQ2 to RQ11 will be analyzed in the appendix, since they are all the same methods and results.

### **Multicorrelation:**

In order to avoid the multicorrelation problem, only the “S1\_Light”, “S1\_Temp”, “S1\_Sound”, “S6\_PIR”, “S5\_CO2” features were kept.

## **Analyze Room Occupation causation:**

### **Methodology:**

Our data has a lot of factors, so we desire to reduce its dimension. Since then, the factor analysis is applied to determine the number of factors, and which one effects the room occupancy the most. The Spearman’s rho is also applied to ascertain the factors having the highest correlation with the room occupancy count. Finally, a Decision Tree Model is also adopted to confirm the result once again.

**RQ12: What are the main factors?**

First, the data is queried by the group by “Date” and “Time” column in order to illustrate the most influencing factors. As the result, the S1\_Light and S5\_CO2 has a great influence on the room occupancy. (The result of the query is shown in the appendix). Since then, the hypothesis is made

**Hypothesis: S1\_Light and S5\_CO2 are the main factors**

### **Bartlett's test:**

Before applying the factor analysis, Bartlett's test of sphericity is applied to determine whether the observed variables intercorrelate at all. You should not use a factor analysis if the test results were statistically insignificant.

Graphical user interface, text, application

Description automatically generated

**Observation**:

The p-value in this Bartlett's test is 0. The analysis confirmed that the observed correlation matrix is not an identity matrix.

### **Kaiser-Meyer-Olkin (KMO) Test:**

Additionally, we also have to apply KMO to determine if data is suitable for factor analysis. It assesses the suitability of each observed variable as well as the entire model. The fraction of variance among all observable variables is estimated by KMO. A smaller fraction is better for factor analysis. KMO values vary from 0 to 1. A KMO value of less than 0.6 is deemed insufficient.

Graphical user interface, text, application

Description automatically generated

**Observation:**

Our data has a KMO of 0.82, which is great. This number implies that your intended factor analysis can proceed.

### **Choosing the number of factors:**

The Kaiser criterion and the scree plot may be used to determine the number of factors. Eigenvalues are used in both

Chart, line chart

Description automatically generated

**Observation:**

Only for 2-factors are eigenvalues greater than one, as seen above. This means we only have to choose 2 factors.

### **Highest Correlation between features and “Room\_Occupancy\_Count”:**

Graphical user interface, application

Description automatically generated

**Observation:**

“S1\_Light” and “S5\_CO2” has the highest correlation with the “Room\_Occupancy\_Count”. But we still desire to have a model to test these results. Since the target variable is Room\_Occupancy\_Count has more than 2 classes so we cannot use the Logistic Regression. Since then, the Decision Tree Model is applied to build for testing the important features.

### **Decision Tree Model:**

After building our decision tree model, the important features are extracted.

Chart

Description automatically generated

**Observation:**

The two most important features in the dataset are S1\_Light and S5\_CO2 so that we are right.

## **Analyze Room Occupation in different conditions:**

### **Methodology:**

Firstly, the relationship between the main factors which are S1\_Light and S5\_CO2 were plotted. The initial hypothesizes are the higher the S1\_Light and S5\_CO2 are the more people in the room. The Mann-Whitney testing is applied for testing our hypothesizes.

### **Relationship plotting:**

**RQ13: How the S1\_Light of each Room Occupancy Type?**

Chart, bar chart

Description automatically generated

**Observation:**

* Surprisingly, the S1\_Light in the rooms having 3 people is lower than the room having 1 and 2 people. However, it is not our concern since the people in the 3-people room may desire to consume the light more economically.

**RQ13: How the S5\_CO2 of each Room Occupancy Type?**

Chart, box and whisker chart

Description automatically generated

**Observation:**

* Surprisingly, the S5\_CO2 in the rooms having 2 people are lower than the room having 1 people. Since then, we will have a statistical test on this.

### **Mann-Whitney Testing:**

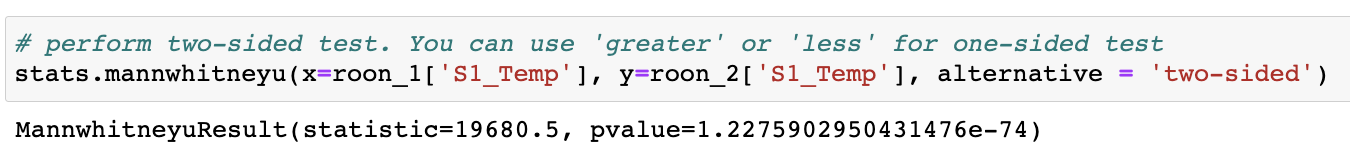
The CO2 in the rooms having 1 people is higher than 2 people, the reason maybe because the 1-person rooms have higher temperature, light than the rooms having 2 people.

**RQ13: How high the S1\_Temp in the room having 1 person compare to the 2-people room?**

**H0:** S1\_Temp in 1-person rooms is lower than 2-people rooms.

**H0:** S1\_Temp in 1-person rooms is equal or higher than 2-people rooms.

At the Significant Level = 0.05



**Observation:**

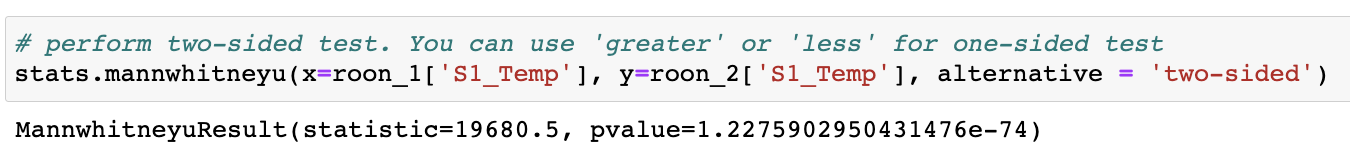
* The column S1\_Temp in the rooms having 1 person is lower:
  + Room\_1\_person has lower mean than Room\_2\_person
  + P-values is lower than the Significant level so we cannot reject the null hypothesis.

**RQ13: How high the S1\_Light in the room having 1 person compare to the 2-people room?**

**H0:** S1\_Light in 1-person rooms is lower than 2-people rooms .

**H0:** S1\_Light in 1-person rooms is equal or lower than 2-people rooms .

At the Significant Level = 0.05



**Observation:**

* The column S1\_Light in the rooms having 1 person is lower:
  + Room\_1\_person has lower mean than Room\_2\_person
  + P-values is lower than the Significant level so we cannot reject the null hypothesis.

# **Conclusion:**

In this project, we exhibited several data analysis approaches on a synthetic predictive maintenance dataset in this research. Many questions about the dataset, as well as the hypotheses we made while working on the data exploration process, were answered. The parameters' correlations were investigated and tested with the Spearman’s rho. The main factors causing the number of room occupancy are also discovered by factor analysis and tested with a decision tree model. Finally, the relationship of the room occupancy and the main factos is also explored by plots and tested with the Man-Whitney testing method. Thanks to the project, we learnt how to use various research methodologies and statistical analysis to evaluate our grasp of the subject and test hypotheses throughout the project. Furthermore, designing tests and determining outcomes has broadened our knowledge and has become a necessary ability for our young engineering team.

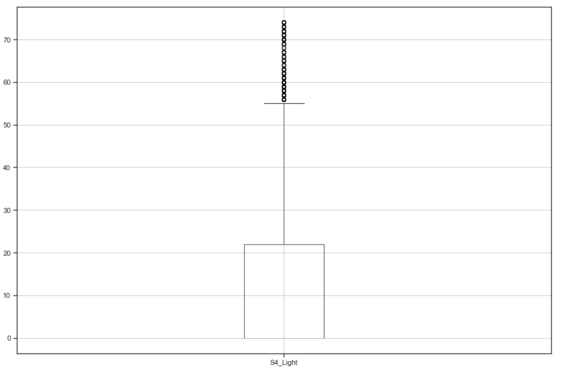
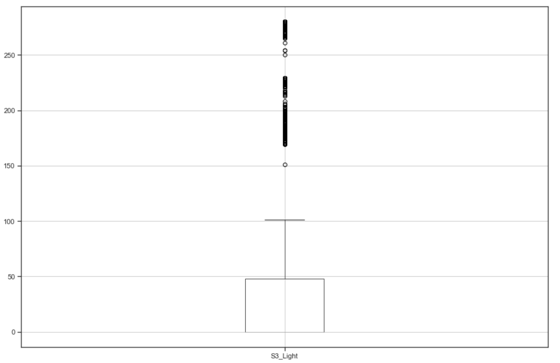
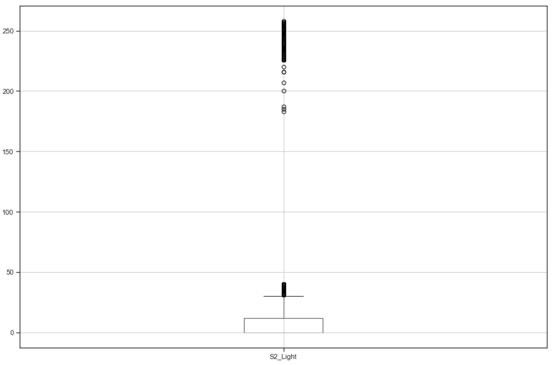
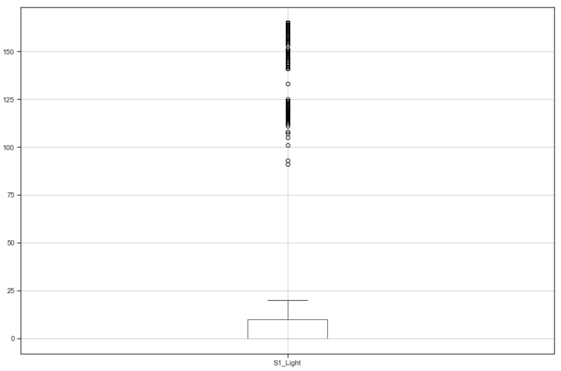
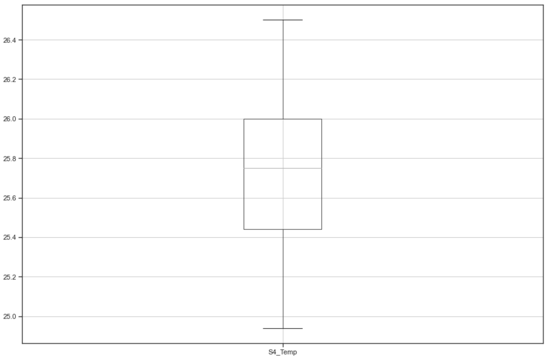
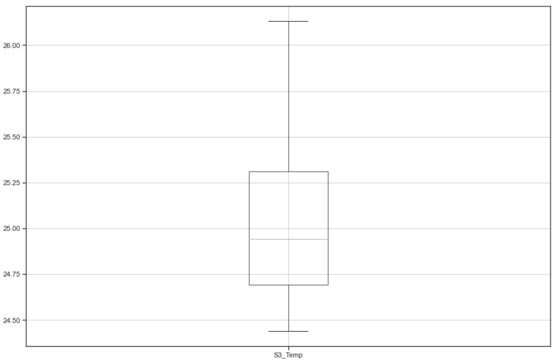
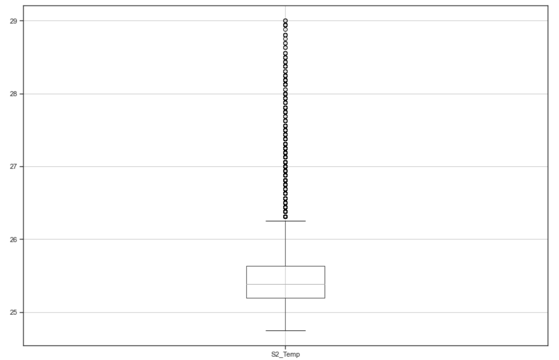
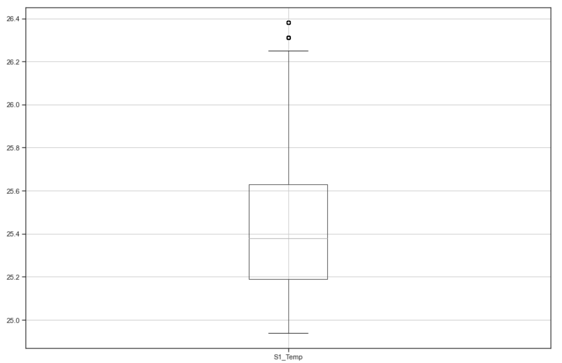
1. **Reference:**

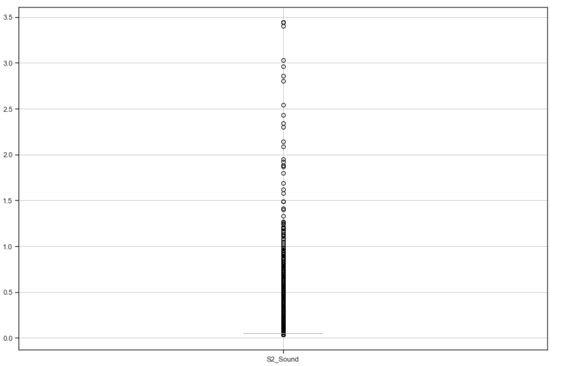
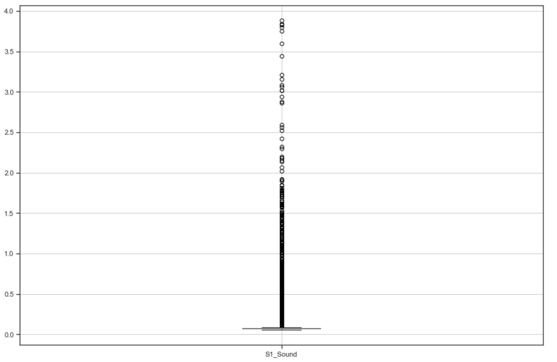
[1] Adarsh Pal Singh, Vivek Jain, Sachin Chaudhari, Frank Alexander Kraemer, Stefan Werner and Vishal Garg, â€œMachine Learning-Based Occupancy Estimation Using Multivariate Sensor Nodes,â€ in 2018 IEEE Globecom Workshops (GC Wkshps), 2018.

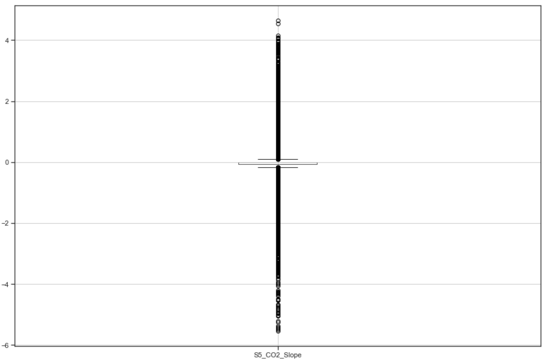
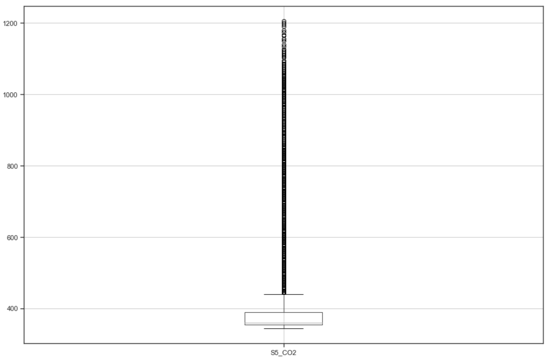
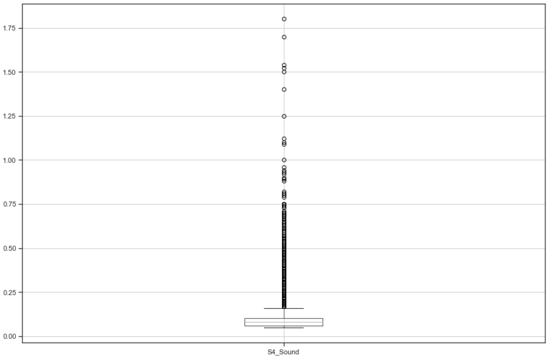
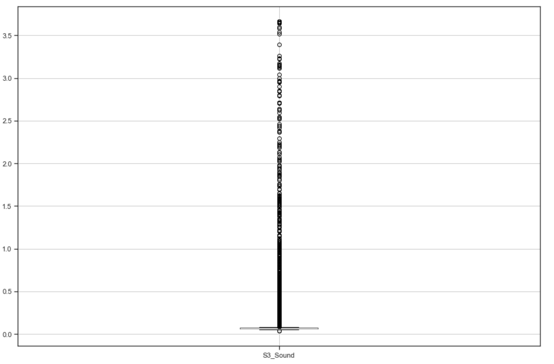
[2] P. Bhandari, “Correlation coefficient: Types, formulas & examples,” *Scribbr*, 19-May-2022. [Online]. Available: https://www.scribbr.com/statistics/correlation-coefficient/. [Accessed: 27-May-2022].

1. **Appendix:**

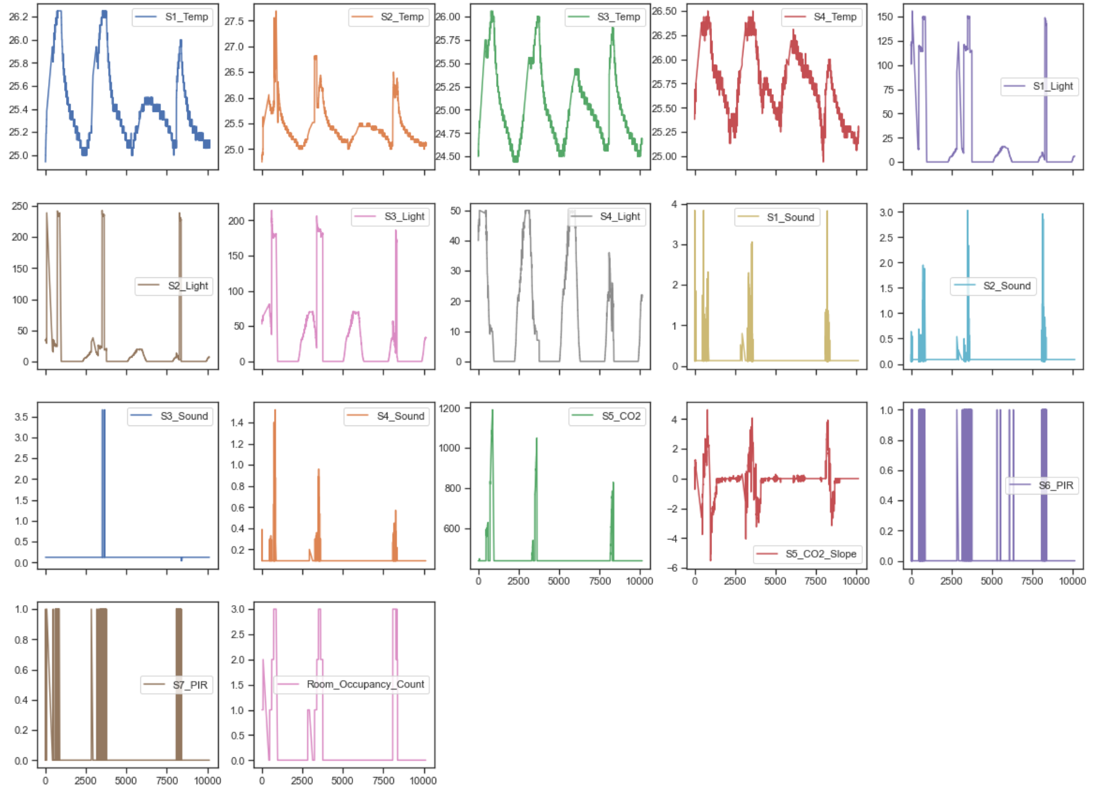
## **Outlier:**







## **Correlation Between Parameter:**



**RQ1: What is the correlation between S1\_Temp and S2\_Temp?**

**H0:** S2\_Temp increases when S1\_Temp increases from 0 to 26.5.

**H0:** S2\_Temp is not correlated with S1\_Temp**.**

At the Significant Level = 0.05

Text

Description automatically generated with medium confidence

**Observation:**

At Spearman’s rho correlation score = 0.97 and df = 8935, the positive correlation between S1\_Temp and S2\_Temp is proven with p\_value = 0.0 < significant values = 0.05. The positive correlation score 0.97 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S2\_Temp increases when S1\_Temp increases from 0 to 26.5”.

**RQ2: What is the correlation between S1\_Temp and S3\_Temp?**

**H0:** S3\_Temp increases when S1\_Temp increases from 0 to 26.5.

**H0:** S3\_Temp is not correlated with S1\_Temp**.**

At the Significant Level = 0.05

Text

Description automatically generated with medium confidence

**Observation:**

At Spearman’s rho correlation score = 0.96 and df = 8935, the positive correlation between S1\_Temp and S3\_Temp is proven with p\_value = 0.0 < significant values = 0.05. The positive correlation score 0.96 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S3\_Temp increases when S1\_Temp increases from 0 to 26.5”.

**RQ3: What is the correlation between S1\_Temp and S4\_Temp?**

**H0:** S4\_Temp increases when S1\_Temp increases from 0 to 26.5.

**H0:** S4\_Temp is not correlated with S1\_Temp**.**

At the Significant Level = 0.05

Text

Description automatically generated with medium confidence

**Observation:**

At Spearman’s rho correlation score = 0.93 and df = 8935, the positive correlation between S1\_Temp and S4\_Temp is proven with p\_value = 0.0 < significant values = 0.05. The positive correlation score 0.93 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S4\_Temp increases when S1\_Temp increases from 0 to 26.5”.

**RQ4: What is the correlation between S1\_Light and S2\_Light?**

**H0:** S2\_Light increases when S1\_Light increases from 0 to 156.

**H0:** S2\_Light is not correlated with S1\_Light**.**

At the Significant Level = 0.05

Graphical user interface, text

Description automatically generated

**Observation:**

At Spearman’s rho correlation score = 0.99 and df = 8935, the positive correlation between S1\_Light and S2\_Light is proven with p\_value = 0.0 < significant values = 0.05. The positive correlation score 0.99 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S2\_Light increases when S1\_Light increases from 0 to 156”.

**RQ5: What is the correlation between S1\_Light and S3\_Light?**

**H0:** S3\_Light increases when S1\_Light increases from 0 to 156.

**H0:** S3\_Light is not correlated with S1\_Light**.**

At the Significant Level = 0.05

Text

Description automatically generated with medium confidence

**Observation:**

At Spearman’s rho correlation score = 0.96 and df = 8935, the positive correlation between S1\_Light and S3\_Light is proven with p\_value = 0.0 < significant values = 0.05. The positive correlation score 0.96 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S3\_Light increases when S1\_Light increases from 0 to 156”.

**RQ6: What is the correlation between S1\_Light and S4\_Light?**

**H0:** S4\_Light increases when S1\_Light increases from 0 to 156.

**H0:** S4\_Light is not correlated with S1\_Light**.**

At the Significant Level = 0.05

Text

Description automatically generated with low confidence

**Observation:**

At Spearman’s rho correlation score = 0.93 and df = 8935, the positive correlation between S1\_Light and S4\_Light is proven with p\_value = 0.0 < significant values = 0.05. The positive correlation score 0.93 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S4\_Light increases when S1\_Light increases from 0 to 156”.

**RQ7: What is the correlation between S1\_Sound and S2\_Sound?**

**H0:** S2\_Sound increases when S1\_Sound increases from 0 to 3.84.

**H0:** S2\_Sound is not correlated with S1\_Sound**.**

At the Significant Level = 0.05

Graphical user interface, text

Description automatically generated with medium confidence

**Observation:**

At Spearman’s rho correlation score = 0.46 and df = 8935, the positive correlation between S1\_Sound and S2\_Sound is proven with p\_value = 0.0 < significant values = 0.05. The positive correlation score 0.46 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S2\_Sound increases when S1\_Sound increases from 0 to 3.84”.

**RQ8: What is the correlation between S1\_Sound and S3\_Sound?**

**H0:** S3\_Sound increases when S1\_Sound increases from 0 to 3.84.

**H0:** S3\_Sound is not correlated with S1\_Sound**.**

At the Significant Level = 0.05

Text

Description automatically generated with low confidence

**Observation:**

At Spearman’s rho correlation score = 0.03 and df = 8935, the positive correlation between S1\_Sound and S3\_Sound is proven with p\_value = 0.0008 < significant values = 0.05. The positive correlation score 0.03 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S3\_Sound increases when S1\_Sound increases from 0 to 3.84”.

**RQ9: What is the correlation between S1\_Sound and S4\_Sound?**

**H0:** S4\_Sound increases when S1\_Sound increases from 0 to 3.84.

**H0:** S4\_Sound is not correlated with S1\_Sound**.**

At the Significant Level = 0.05

Graphical user interface, text

Description automatically generated

**Observation:**

At Spearman’s rho correlation score = 0.46 and df = 8935, the positive correlation between S1\_Sound and S4\_Sound is proven with p\_value = 0.0 < significant values = 0.05. The positive correlation score 0.46 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S4\_Sound increases when S1\_Sound increases from 0 to 3.84”.

**RQ10: What is the correlation between S5\_CO2 and S5\_CO2\_Slope?**

**H0:** S5\_CO2\_Slope increases when S5\_CO2 increases from 439.693341 to 1190.

**H0:** S5\_CO2\_Slope is not correlated with S5\_CO2**.**

At the Significant Level = 0.05

A picture containing text

Description automatically generated

**Observation:**

At Spearman’s rho correlation score = 0.36 and df = 8935, the positive correlation between S5\_CO2 and S5\_CO2\_Slope is proven with p\_value = 8.69e-276 < significant values = 0.05. The positive correlation score 0.36 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S5\_CO2\_Slope increases when S5\_CO2 increases from 439.693341 to 1190”.

**RQ11: What is the correlation between S6\_PIR and S7\_PIR?**

**H0:** S7\_PIR increases when S6\_PIR increases from 0 to 1.

**H0:** S7\_PIR is not correlated with S6\_PIR**.**

At the Significant Level = 0.05

Text

Description automatically generated

**Observation:**

At Spearman’s rho correlation score = 0.56 and df = 8935, the positive correlation between S7\_PIR and S6\_PIR is proven with p\_value = 0.0 < significant values = 0.05. The positive correlation score 0.56 indicates that the two columns are positively correlated. Since then, the null hypothesis is accepted that “S7\_PIR increases when S6\_PIR increases from 0 to 1”.