
# EB5101 Foundation Business Analytics:
# Data Preparation Assignment

**Lecturer:** Mr. Prakash Chandra Sukhwal



| Student ID | Name | E-mail |
|---|---|---|
| A0178551X | Choo Ming Hui Raymond | e0267862@u.nus.edu |
| A0178431A | Huang Qingyi | e0267742@u.nus.edu |
| A0178415Y | Jiang Zhiyuan | e0267726@u.nus.edu |
| A0178365R | Wang Jingli | e0267676@u.nus.edu |
| A0178500J | Yang Chia Lieh | e0267811@u.nus.edu |

# Table of Contents

# 1 Introduction

With the rapid development of the Internet of Things (IoT) technology, sensors are increasingly being implemented to measure and monitor environmental, clinical and other related parameters to provide insights on our daily lives. Based on two datasets obtained from four sensors embedded in a commercial building, six environmental parameters such as temperature, humidity, $CO_2$, Volatile Organic Compound (VOC), light and noise of a closed room were recorded. This report explores these datasets, and aims to uncover findings related to this commercial building.

# 2 Data Preparation

Prior to using the dataset for analysis, a series of data preparation and sanity check of the data reported for each variable was carried out to identify abnormal records, remove duplicated data and inspect missing values based on domain knowledge.

## 2.1 Abnormal Data Exploration

Since these four sensors are designed to measure environmental data, preliminary abnormal data exploration was carried out based on the variables summary and domain knowledge of international logical indoor index. These findings are as shown in Table 1.

**Table 1: Preliminary Summary of Variables**

| | Temperature/℃ | Noise/dB | Light/ Lux | $CO_2$/ppm | VOC/ppm | Humidity/% |
|---|---|---|---|---|---|---|
| **Min** | 22.85 | 50.7 | 1.917 | 424 | 311 | 46.4 |
| **1st Qu.** | 24.29 | 51.74 | 6.583 | 433 | 317.5 | 60.65 |
| **Median** | 24.7 | 51.89 | 8.042 | 438.1 | 321.2 | 62.17 |
| **Mean** | 24.73 | 52.07 | 147.211 | 438.7 | 321.7 | 62.67 |
| **3rd Qu.** | 25.08 | 52.15 | 377.708 | 442.2 | 324.3 | 64.45 |
| **Max** | 26.1 | 66.81 | 749 | 470.9 | 345.3 | 70.4 |
| **Logical Values** | <50 | 50-60 | 320-500 | 350-1000 | <500 | <100 |

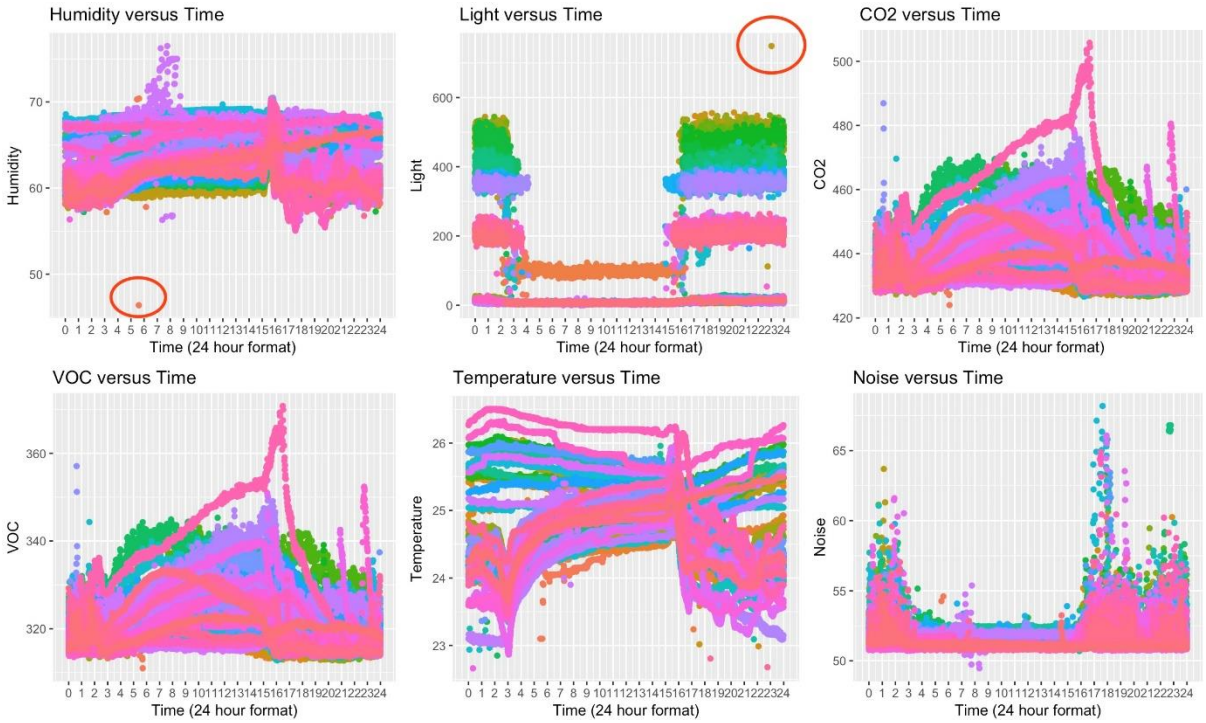


**Figure 1: Scatter Plot of Six Environment Parameters versus Time by Day**

As shown in Figure 1, several abnormal values were found:

1. There was an occurrence of extreme low humidity at 05:36 on 2017-03-02, where humidity as recorded by sensor SS0050 was 46.4% whereas other occurrences preceding and following that occurrence were more than 50%.
2. There was an occurrence of extreme luminance at 23:02 on 2017-03-10, where luminance as recorded by sensor SS0036 was 749 Lux, whereas other occurrences preceding and following that occurrence were less than 600 Lux.

It is likely that either these sensors have malfunctioned and read an abnormal value or something unusual might have occurred at that instance. As these occurrences only occurred for a minute, these records were removed to avoid influencing the analysis. Further to this, no other abnormal records could be definitely identified at this stage.

## 2.2 Duplicated Data Cleaning

After the preliminary abnormal data exploration, checks for duplicated data were carried out to ensure that no repeated data points were posted. The variables *"Date_time"* and *"unitid"* were first chosen to undergo this check to see if there were repeated data posted by the same sensor at the same instance. Since none were found, the same was done with the other six environmental variables to identify instances of delayed transmission. As no such instance was observed, no duplicated data were required to be removed.

## 2.3 Missing Data Exploration

In the final stage of data preparation, the dataset was checked for missing data in two stages. First, it was found that all the existing 43085 rows of observations in March and 41720 rows in April contains no missing value (N/A or blank). However, this is only part of the missing data check, as there may be instances where an entire occurrence was not sent out by the sensor.

Based on the summary of all the records by day in both datasets corresponding to the normal dates distributions in March and April[1], it was observed that:

1. There was a maximum of 1440 records in a day with all sensors combined. This corresponds to 1440 minutes in a day, which means that there were 1440 records in a day, if there is no network or sensor issues.
2. For days where there is data collected from the sensors, there are a total of 155 data points missing in these two months, with 115 points from March and 40 points from April. The plot of missing data quantity by day is given in Figure 2, with the detailed timestamps of missing data included in Table 2:

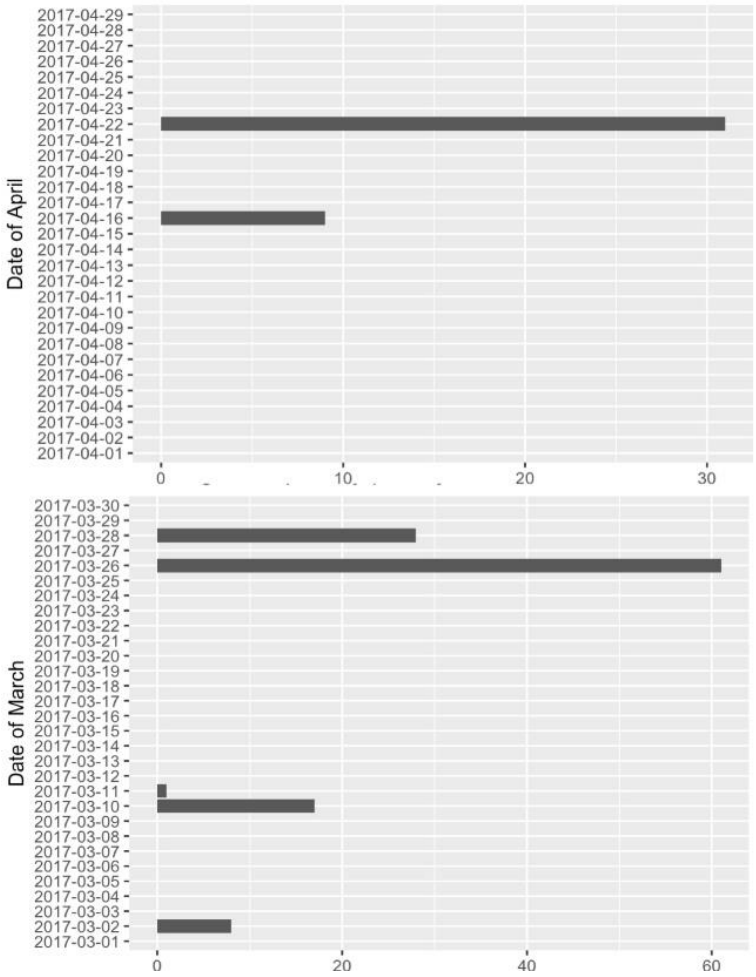


**Figure 2: Plot of Missing Data Quantity by Day**

[1] Posted Data Points Statistics of Sensors by Day is included in Appendix A

**Table 2: Statistics of Missing Dates and Time Points**

| | Missing Dates | Missing Time Points | Count |
|---|---|---|---|
| **1** | 2017-03-02 | "05:26" "05:27" "05:38" "05:43" "05:44" "06:06" "06:07" "06:08" | 8 |
| **2** | 2017-03-10 | "22:45" "22:46" "22:47" "22:48" "22:49" "22:50" "22:51" "22:52" "22:53" "22:54""22:55" "22:56" "22:57" "22:58" "22:59" "23:00" "23:01" | 17 |
| **3** | 2017-03-11 | "00:00" | 1 |
| **4** | 2017-03-26 | "10:00" "10:01" "10:02" "10:03" "10:04" "10:05" "10:06" "10:07" "10:08" "10:09" "10:10" "10:11" "10:12" "10:13" "10:14" "10:15" "10:16" "10:17" "10:18" "10:19" "10:10" "10:11" "10:12" "10:13" "10:14" "10:15" "10:16" "10:17" "10:18" "10:19" "10:20" "10:21" "10:22" "10:23" "10:24" "10:25" "10:26" "10:27" "10:28" "10:29" "10:30" "10:31" "10:32" "10:33" "10:34" "10:35" "10:36" "10:37" "10:38" "10:39" "10:40" "10:41" "10:42" "10:43" "10:44" "10:45" "10:46" "10:47" "10:48" "10:49" "10:50" "10:51" "10:52" "10:53" "10:54" "10:55" "10:56" "10:57" "10:58" "10:59" "15:52" | 61 |
| **5** | 2017-03-28 | "18:28" "22:38" "22:39" "22:40" "22:41" "22:42" "22:43" "22:44" "22:45" "22:46" "22:47" "22:48" "22:49" "22:50" "22:51" "22:52" "22:53" "22:54" "22:55" "22:56" "22:57" "22:58" "22:59" "23:00" "23:01" "23:02" "23:03" "23:04" | 28 |
| **6** | 2017-04-16 | "07:37" "07:52" "07:57" "08:01" "08:05" "08:07" "08:13" "08:28" "08:30" | 9 |
| **7** | 2017-04-22 | "10:01" "10:02" "10:03" "10:04" "10:05" "10:06" "10:07" "10:08" "10:09" "10:10" "10:11" "10:12" "10:13" "10:14" "10:15" "10:16" "10:17" "10:18" "10:19" "10:20" "10:21" "10:22" "10:23" "10:24" "10:25" "10:26" "10:27" "10:28" "10:29" "10:30" "10:31" | 31 |
| **Total** | | | **155** |

3. There are two days where there were no data collected for the entire day (31st March and 30th April). It's likely that these sensors are dormant or in maintenance mode at every end of the month.
4. There are instances where there is bulk data missing. For example, 17 minutes, 1 hour, 27 minutes, and 31 minutes of continuous data for 10th March, 26th March, 28th March and 22nd April respectively were missing.
5. Comparing with the daily posted records by each sensor, there is no visible duty-standby cycle trend between these four sensors. These sensors appear to upload data at random.
6. With reference to Table A in the Appendix, the March and April statistics of the monthly data recorded per sensor are not consistent. In particular, it was noted that the data collected in April is not evenly distributed across all sensors as compared to those in March.

With the above observations in mind, it is nearly impossible to impute values for the missing data as these are missing completely at random. Since these missing data do not form a major part of the dataset (3.5% of expected data from March and April), these are ignored in further analytics.

# 3 Related Findings

After data preparation, further data visualization and analysis need to be carried out to explore the latent features of this commercial building with four sensors.

## 3.1 Correlations of Variables

To better understand the structural correlations between variables, a correlation matrix was generated. In Figure 3, it was found that *"VOC"* has the highest positive correlation with *"$CO_2$"*, while the positive correlation of *"Humidity" and "Temperature"* is also obvious. This leads us to believe that there is a central Heating, Ventilation and Air-Conditioning (HVAC) system installed in this commercial building, which helps to mix, circulate and disperse air from/to all of the rooms.

In addition, *"Humidity"* and *"Temperature"* both have relatively high negative correlations with *"Light"*. This means that when there is light, the humidity and temperature of the building is low. This further reinforces the belief that there is a central HVAC system operating during working hours.

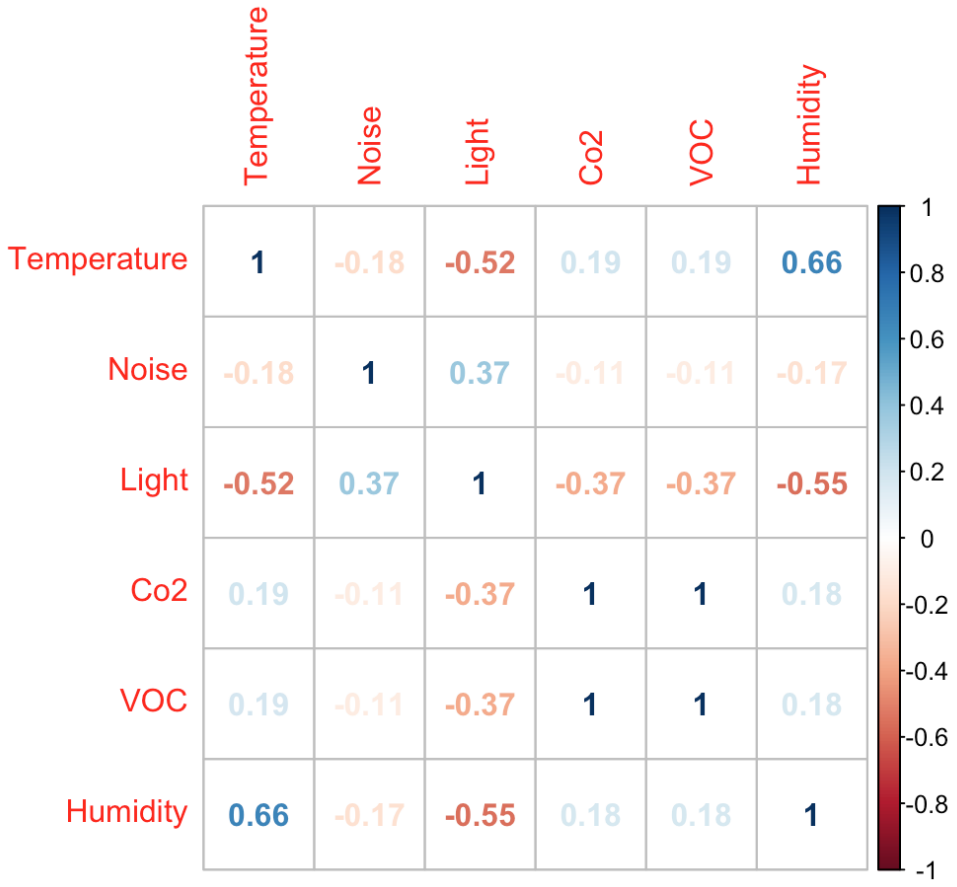


**Figure 3: Plot of Correlation of Variables**

## 3.2 Heating, Ventilation and Air-Conditioning (HVAC) System Operating Hours

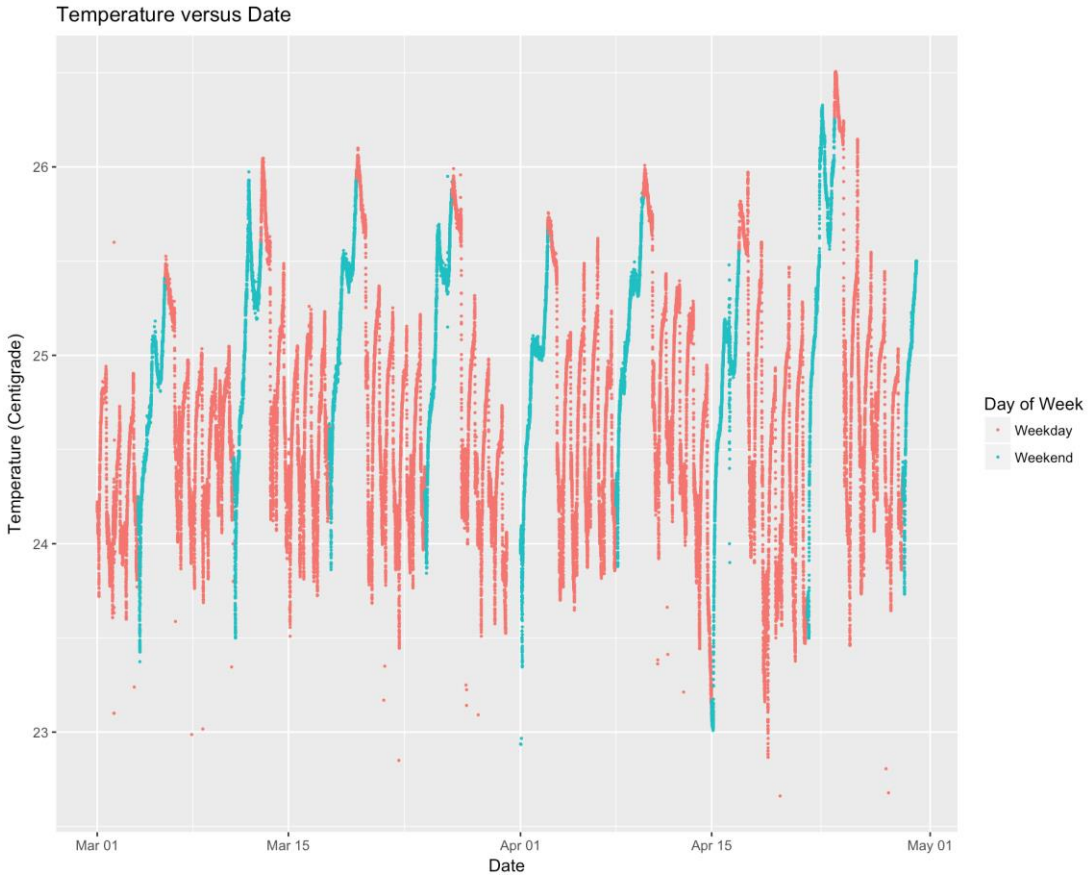


**Figure 4: Temperature versus Date (March & April)**

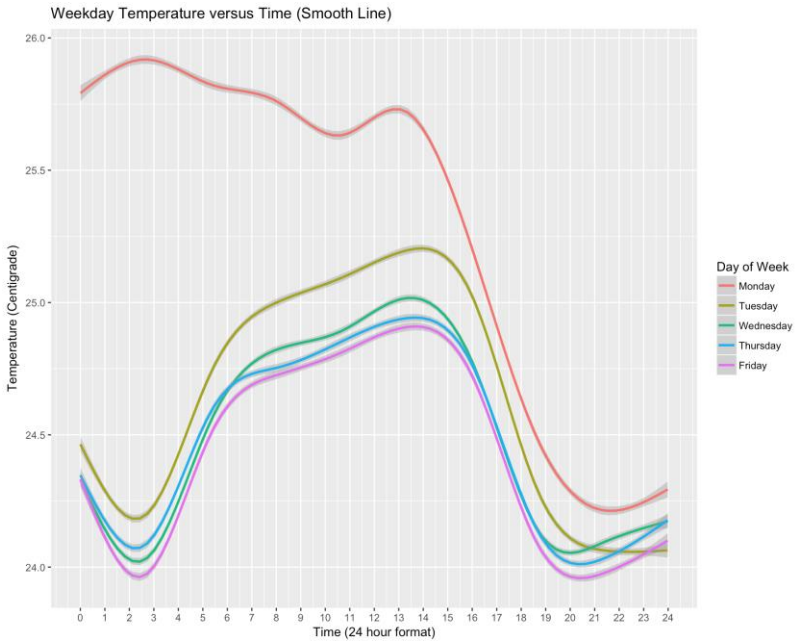


**Figure 5: Weekday Temperature versus Time (Smooth Line)**

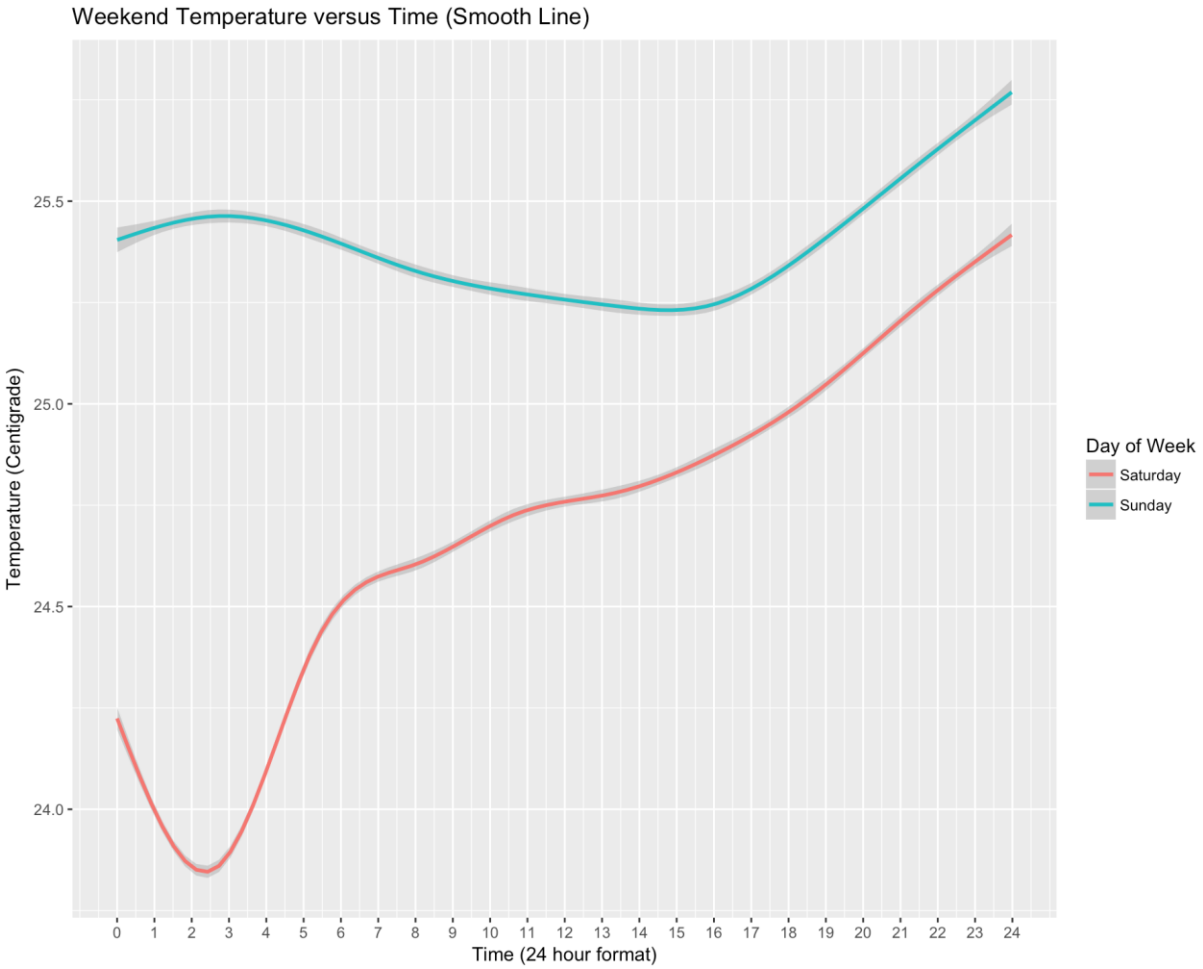


**Figure 6: Weekend Temperature versus Time (Smooth Line)**

From Figure 4, 5 & 6, it's noted that the temperature profile moves in a regulated manner, which is consistent with the expectation that the building's air-conditioning is controlled by a central HVAC system. Based on point of inflections of the temperature profile on weekdays, the general observation is that the building's HVAC system operates on a 5-day work week basis, from Monday afternoon to the early hours of Saturday morning. From Figure 5 & 6, the operating hours of the HVAC system is narrowed down to "Mondays to Fridays, 2.30pm to 3am".

## 3.3 Operating Hours of the Building

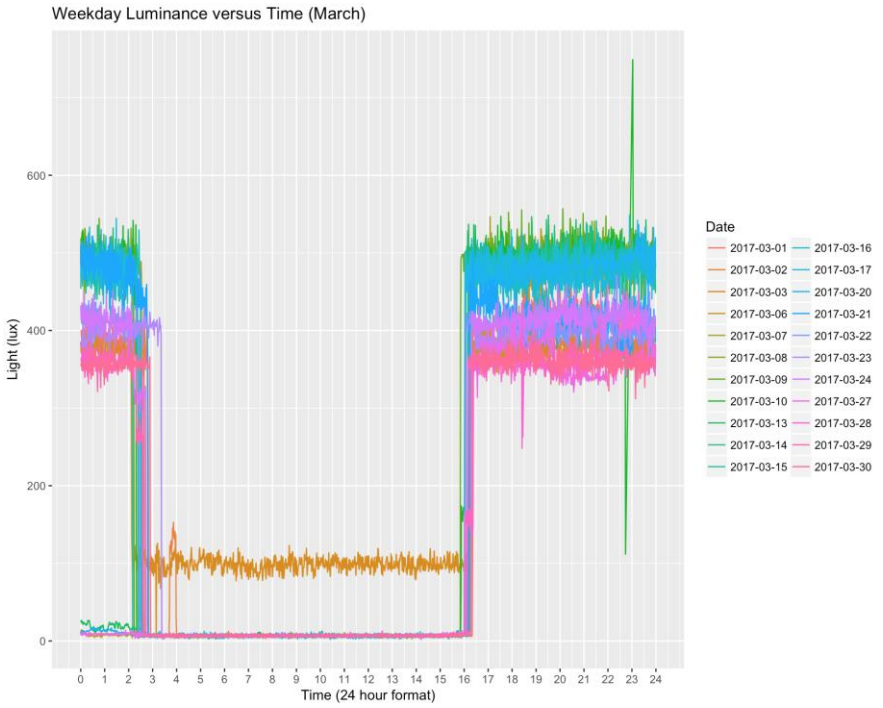


**Figure 7: Weekday Luminance versus Time (March)**

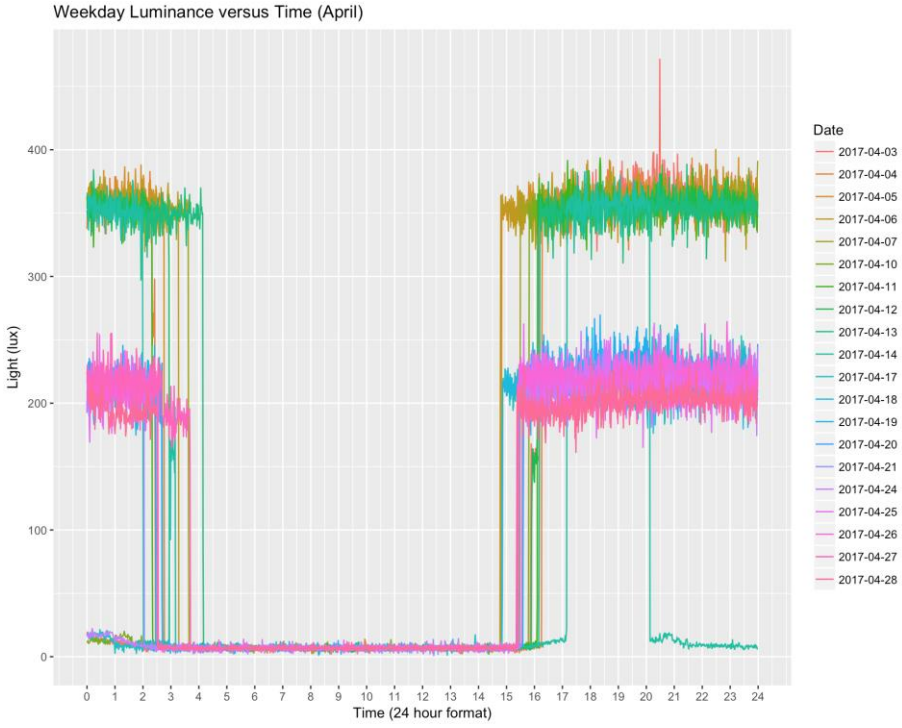


**Figure 8: Weekday Luminance versus Time (April)**

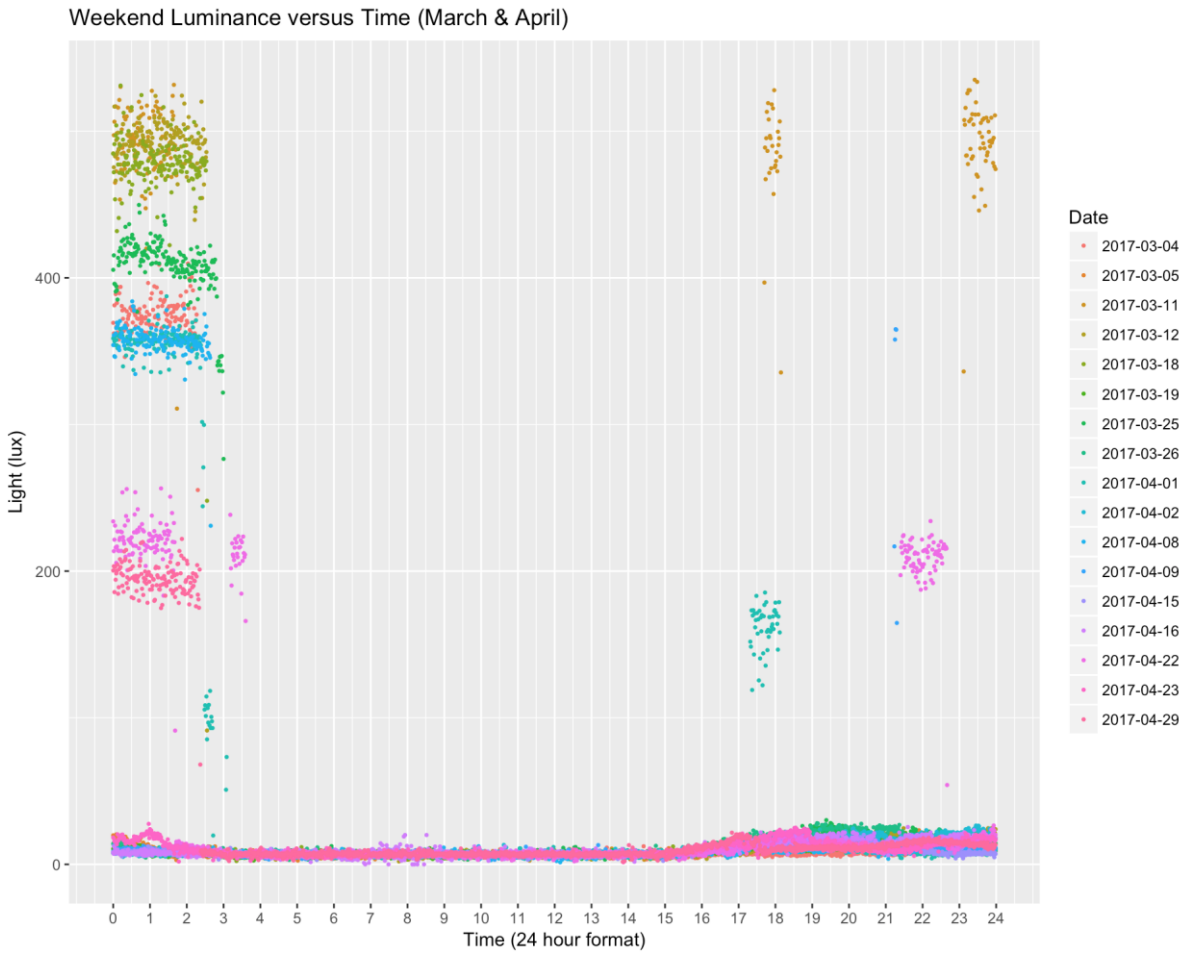


**Figure 9: Weekend Luminance versus Time (March & April)**

Considering that the luminance of a typical office lighting is between 320 and 500 Lux[1], the initial observation from Figures 7, 8 and 9 reinforces our observation that the building operates on a 5-day work week from around 3 to 4pm on Monday afternoon to the early hours of Saturday, around 3am.

[1] The Luminance values and the respective lighting conditions were referenced from "Lux" published by Wikipedia on https://en.wikipedia.org/wiki/Lux accessed on 10 March 2018.

## 3.4 Abnormal Lighting Conditions on 3rd March

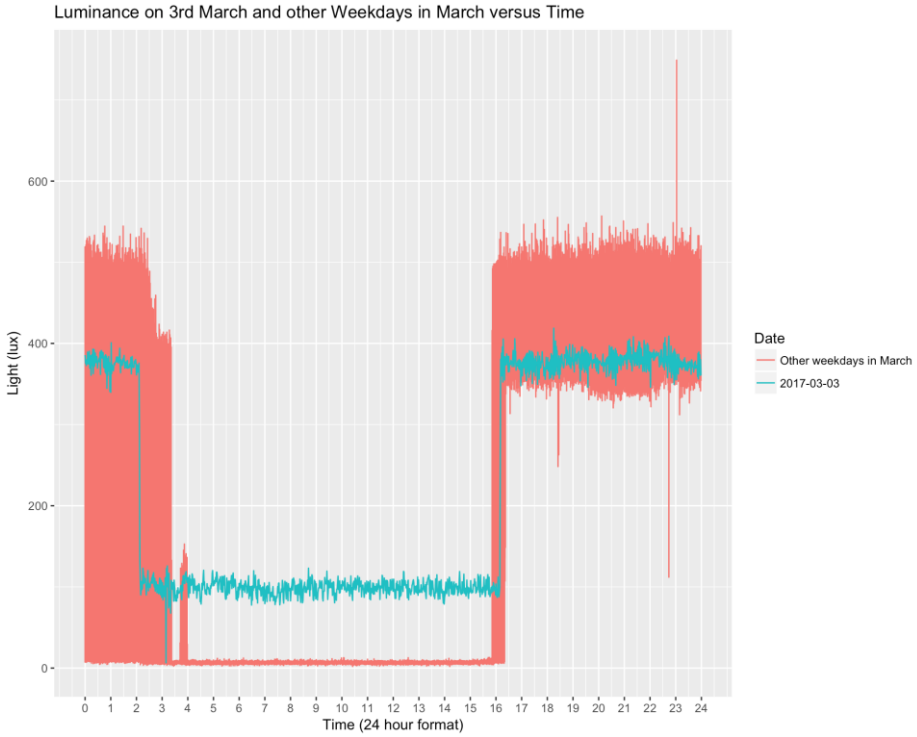


**Figure 10: Lights left on 3rd March**

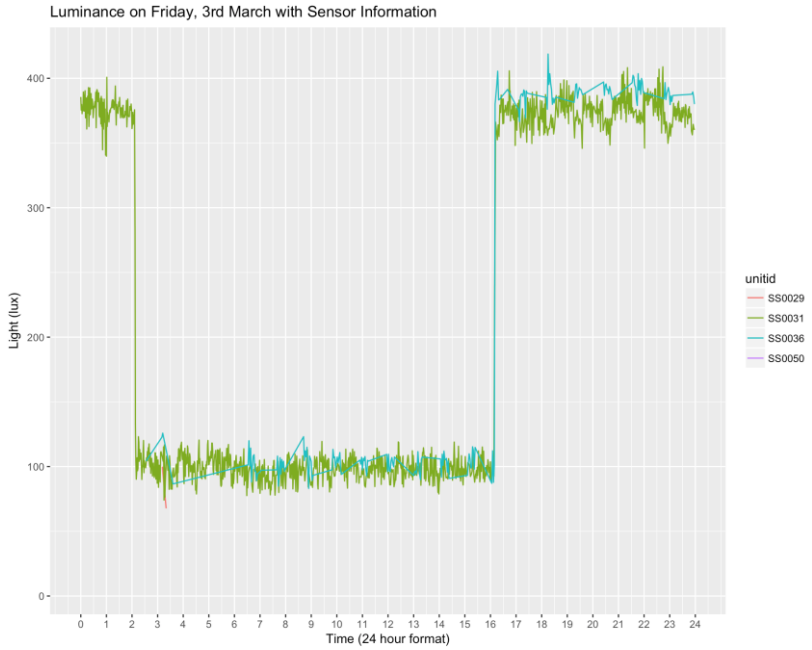


**Figure 11: Luminance on Friday, 3rd March with Sensor Information**

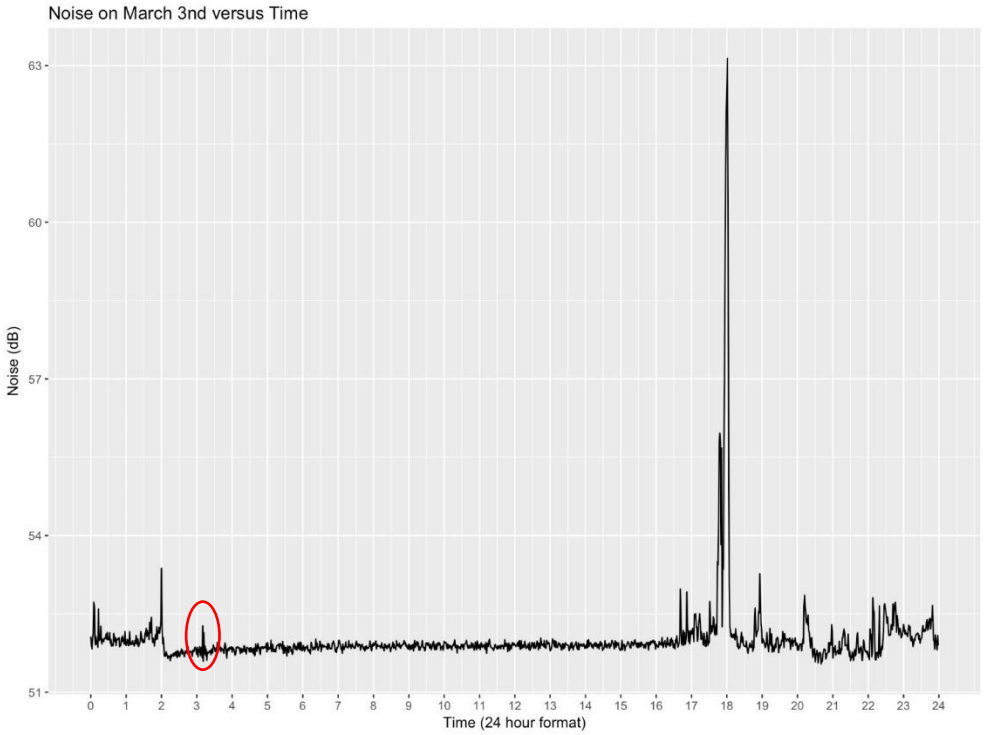


**Figure 12: Noise data on Friday, 3rd March**

Based on Figures 7 & 10, it's noted that on Friday, 3rd March, the luminance of the building was abnormal from 3am to 4pm when compared to other weekdays. After reviewing the sensors responsible for that day's data (Figure 12), it was concluded that the sensors were healthy as they report similar values, with the lights dimmed at approximately 2am till 4pm. However, from 2am onwards, a spike in noise was noted at about 3.15am. Since a luminance of 100 lux corresponds to a very dark overcast sky, it's likely that some building occupants were working late beyond 2am that night, and left at some point either at or after 3.15am, forgetting to turn off the lights on their way out.

## 3.5 Varying Light Intensities from March to April

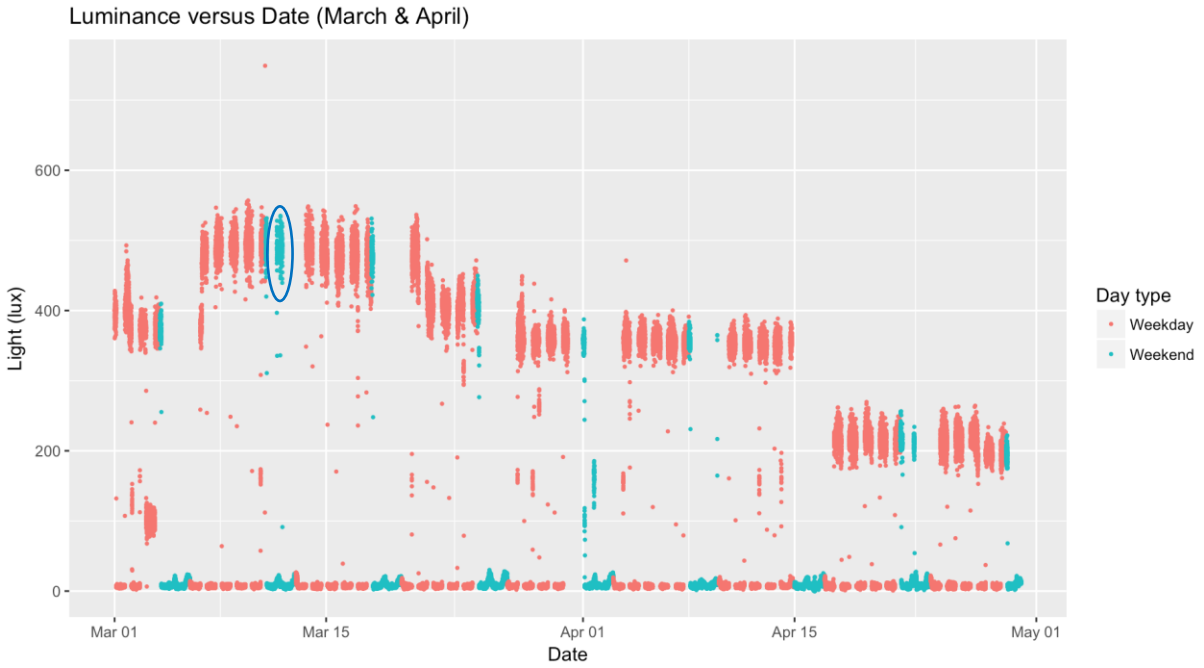


**Figure 13: Luminance versus Date in March & April**

From Figure 13, it's observed that during these two months, there are three main levels of luminance, approximately 400 lux, 500 lux and 200 lux in chronological order. From 7th March till 21st March, it's observed that the brightness of the building increased from the 400 lux to 500 lux range, with an additional day (Saturday, 11th March) having the same lighting conditions as the previous days when there shouldn't have been any light observed as per other Saturdays (see 11th March data from Figure 9). From Wednesday, 22nd March till 14th April, the luminance reverts to the 400 Lux range, before dimming to the 200 Lux range after 18th April. This leads us to believe that there is a likelihood the building occupants might be involved in a single project, with the project peak being from 7th March till 21st March, and with the project subsequently ramping down after 18th April, resulting in more unoccupied desks and less lights required to illuminate these desks.

## 3.6 Abnormal Lighting Conditions on 14th April


**Figure 14: Luminance on Friday, 14th April**

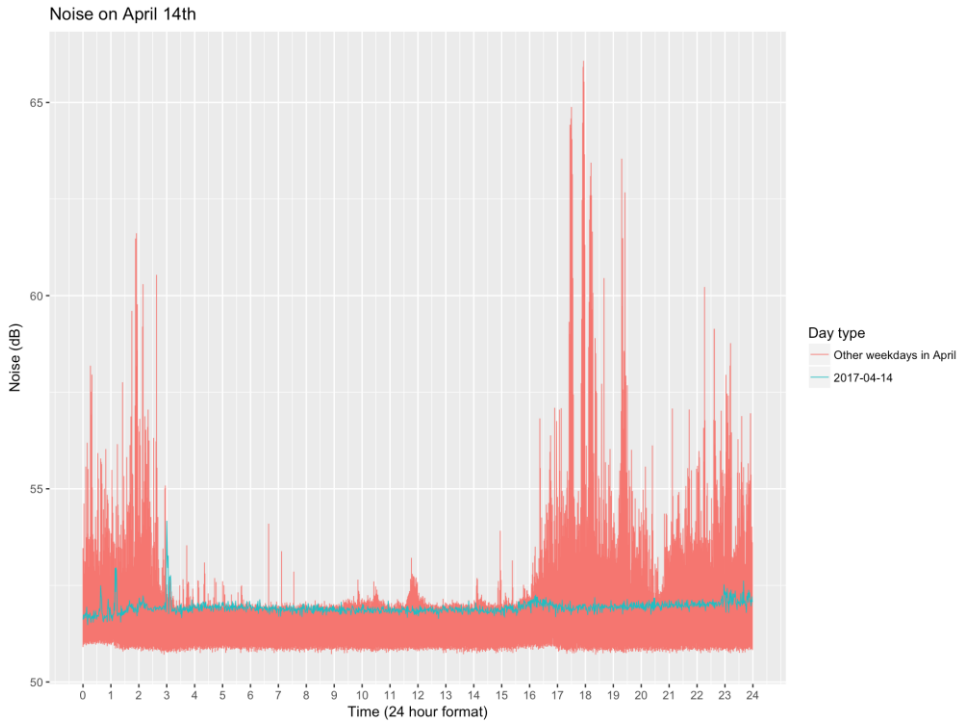


**Figure 15: Noise conditions on Friday, 14th April**

From Figure 14, its noted that on Friday, 14th April, the building occupants appear to have left for the day earlier than usual at 8pm, instead of working till the early hours of Saturday morning. Considering our previous assumption in section 3.5 that the project is ramping down on 18th

April, it's likely that the occupants might have achieved a major milestone at this point, thus leaving earlier, and also requiring less manpower / resources after 18th April.

## 3.7 Drop in ambient Noise level on 16th April

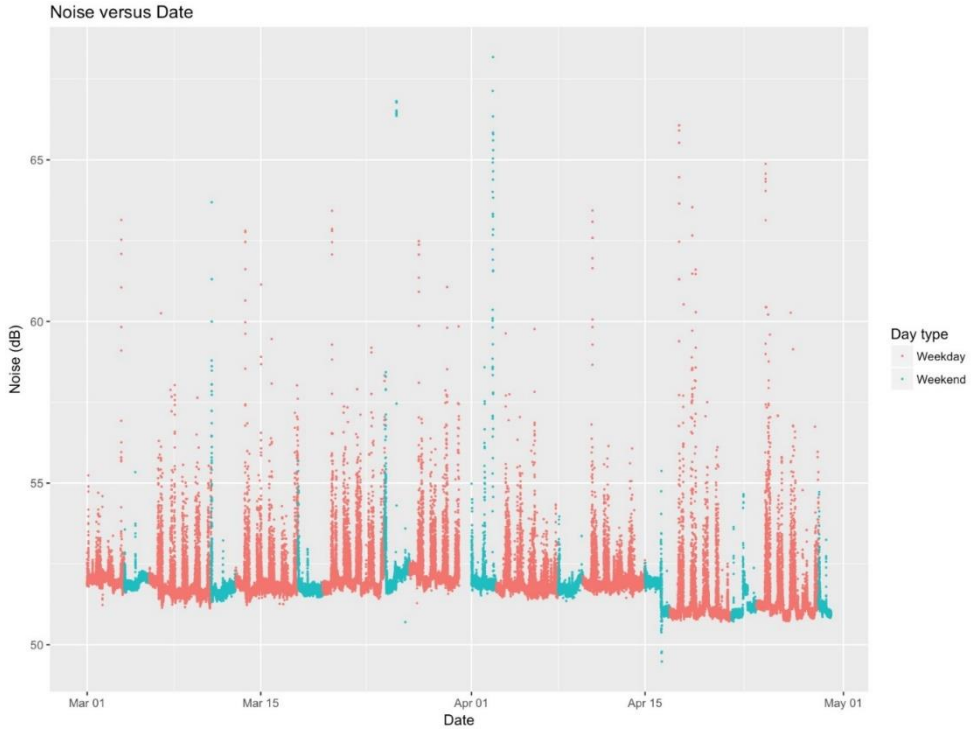


**Figure 16: Noise versus Date (March & April)**

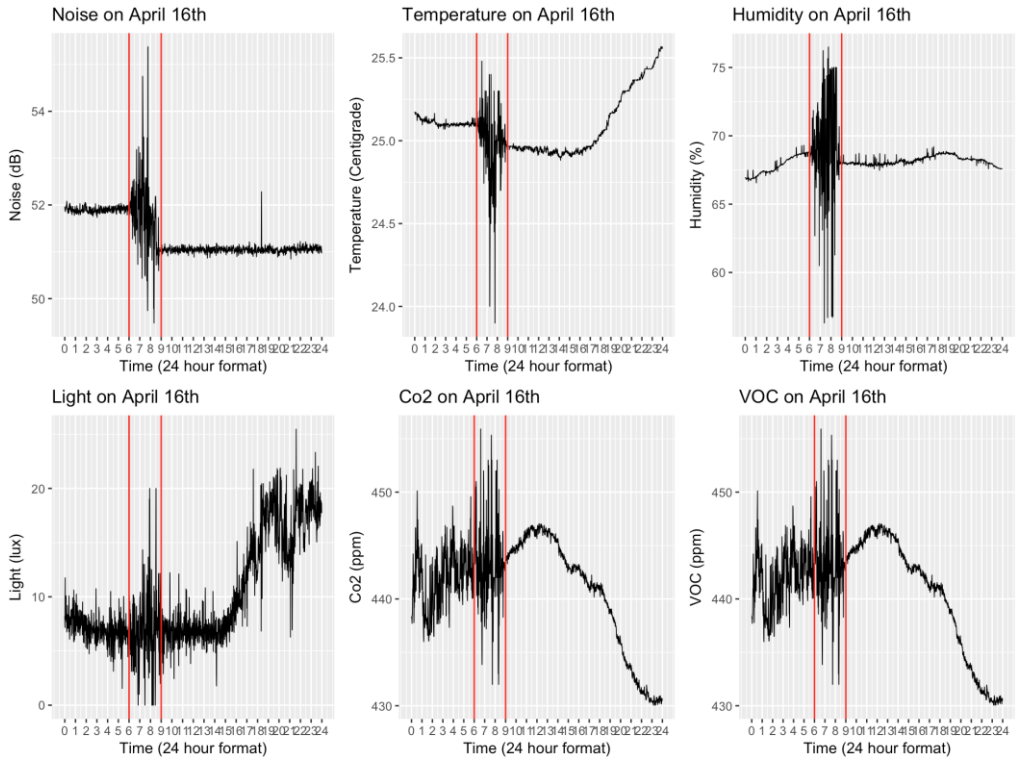


**Figure 17: Noise, Temperature, Humidity, Light, $CO_2$, VOC on Sunday, 16th April**

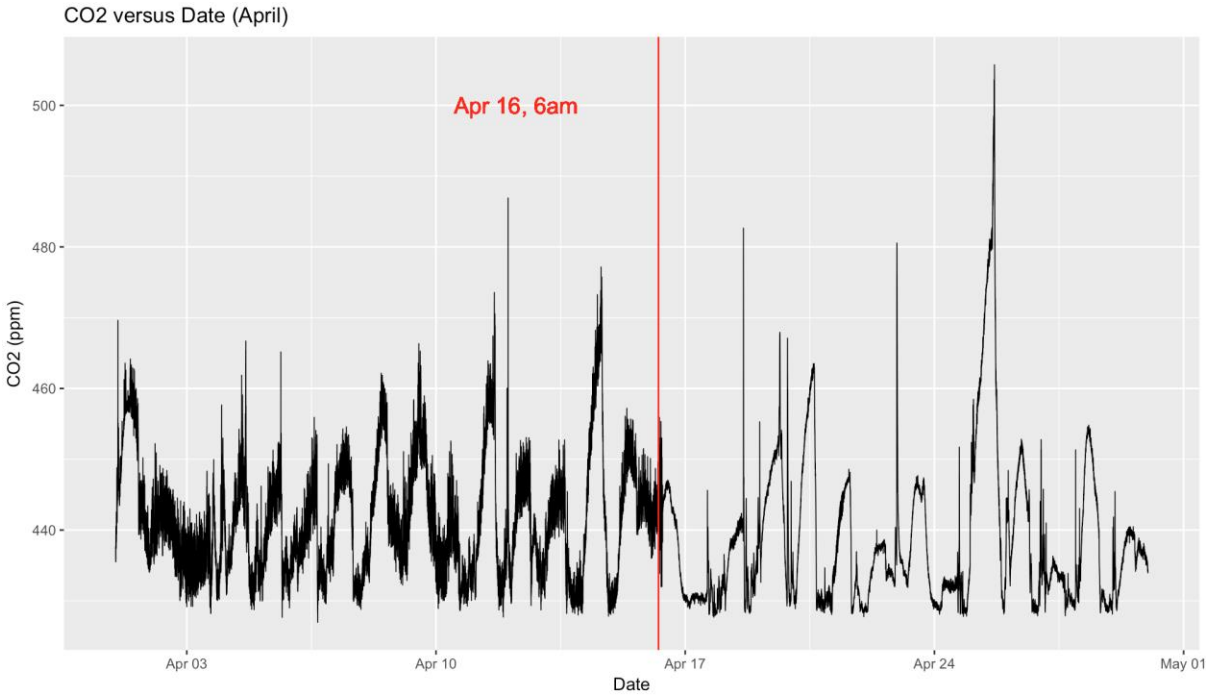


**Figure 18: $CO_2$ versus Date (April)**

From Figure 16, it's observed that there is a permanent decrease in ambient noise level from 16th April onwards. In addition, as seen in Figure 17 that from 6am to 9am, there's some disturbance to Temperature, Humidity and Noise parameters. Consequently, the external disturbances acting on $CO_2$ and VOC have disappeared from 9am onwards. It's likely that during this 3-hour window, some maintenance work was carried out on the building (such as the HVAC system and other equipment in the building), with test runs carried out, resulting in temperature, humidity and noise fluctuations during this 3-hour window, and better performance of these equipment (e.g. less ambient noise from running equipment, improved seals on windows etc.) after 9am.

## 3.8 Location of Sensors

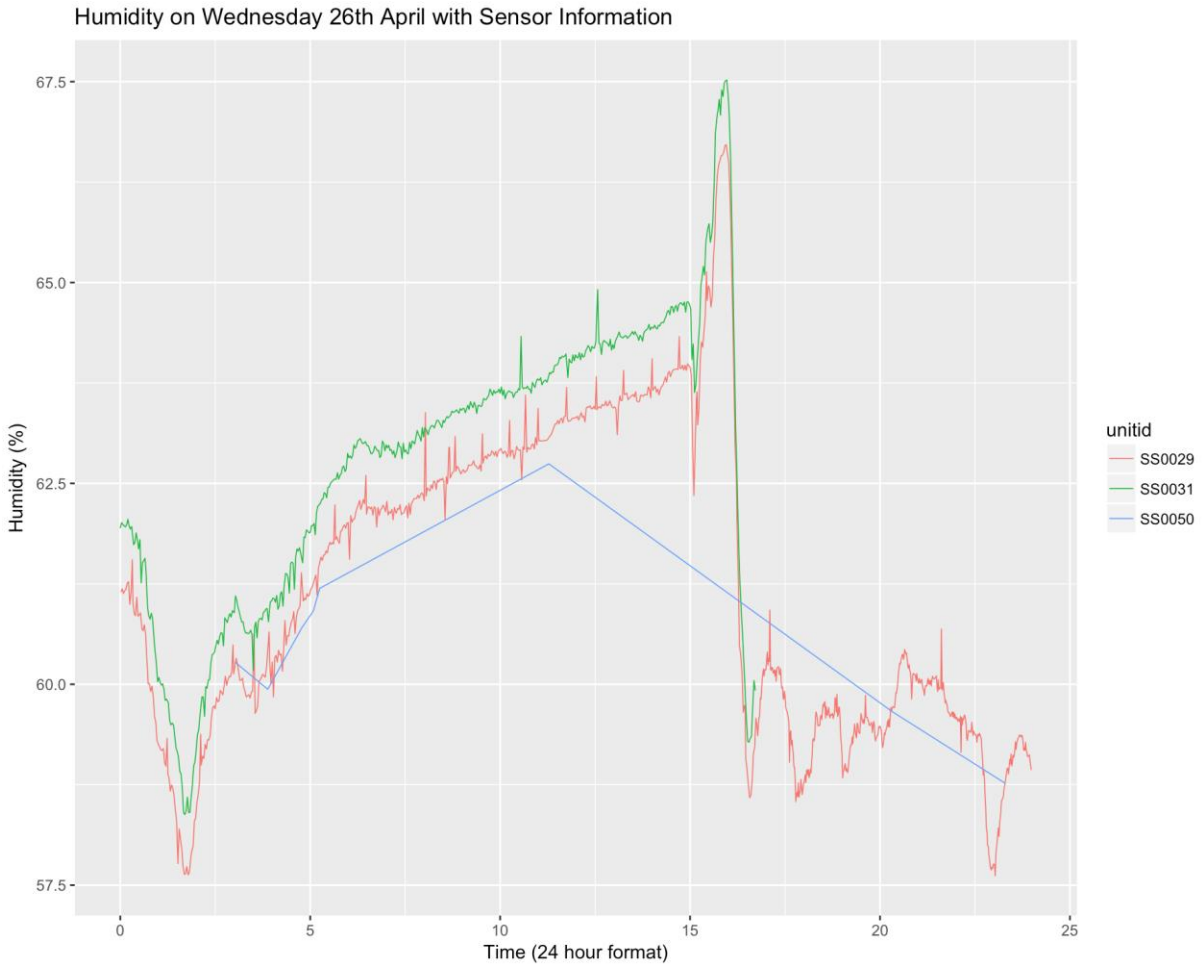


**Figure 19: Humidity versus Time on Wednesday, 26th April**

To verify whether sensors are located at the same place, sensor data where at least two sensors operating on the same day were used. If there are a gap between the measured values, it's likely that these sensors are in a different location. Based on the Figure 19, sensors SS0029 and SS0031 can be considered to be installed at different places, due to the obvious "gap" of humidity obtained from these two sensors during the same period.

## 3.9 Humidity

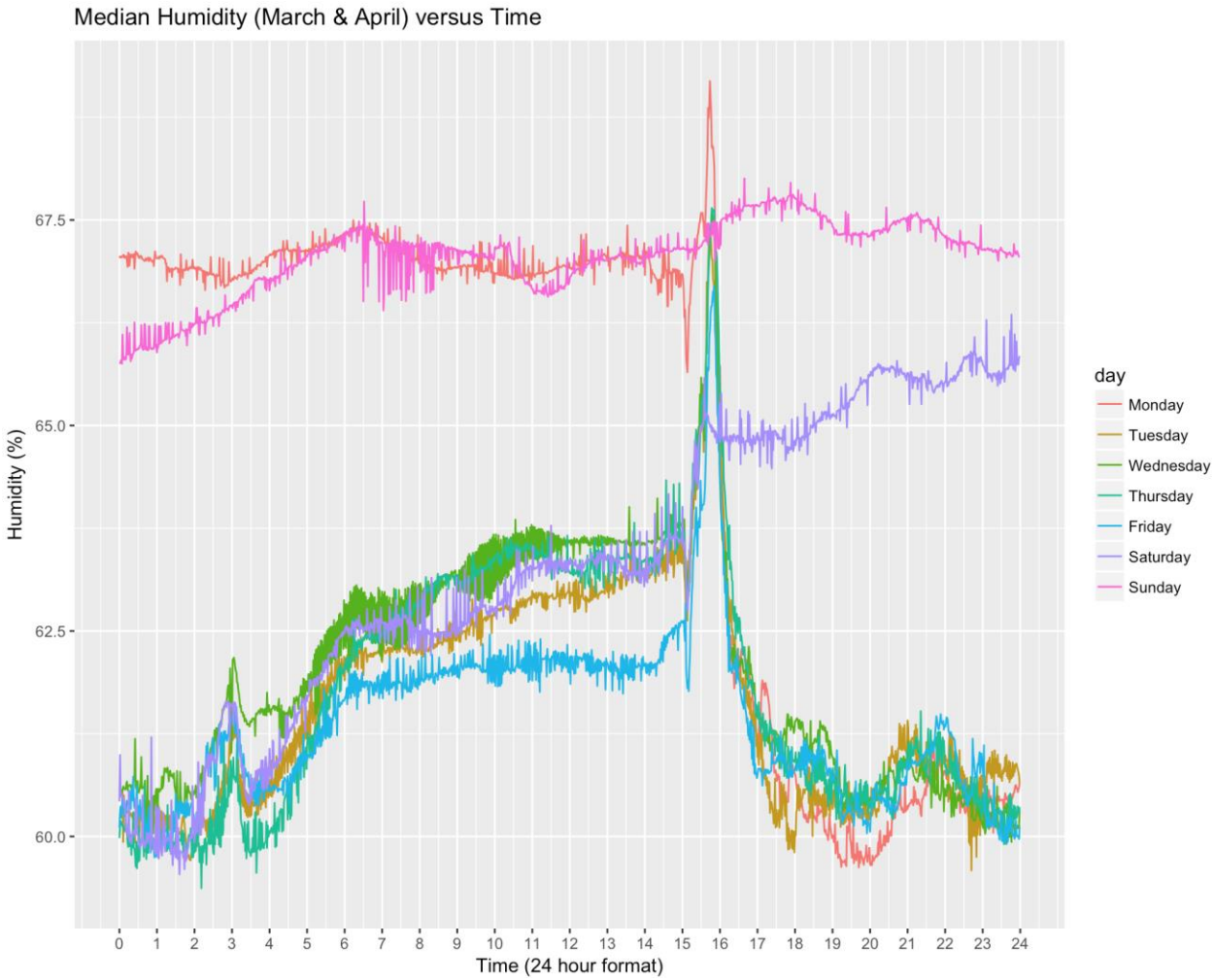


**Figure 20: Median Humidity (March & April) versus Time**

From Figure 20, it's noted that on weekdays, humidity will start rising at 3pm and peak at about 4pm before coming back down to the usual levels. Based on the previous finding in section 3.2, this phenomenon is perfectly matches the start operation time of the HVAC system. This is likely due to the design of the HVAC system, where a fresh air intake occurs during cold start of the system, resulting in an increase of humidity during this phase. It's also seen that humidity increases when the HVAC system is not operating, with humidity reaching equilibrium on Sunday and Monday (before 3pm), as it is not being controlled by the HVAC system.

## 3.10 Abnormal Carbon Dioxide and VOC on 25th April

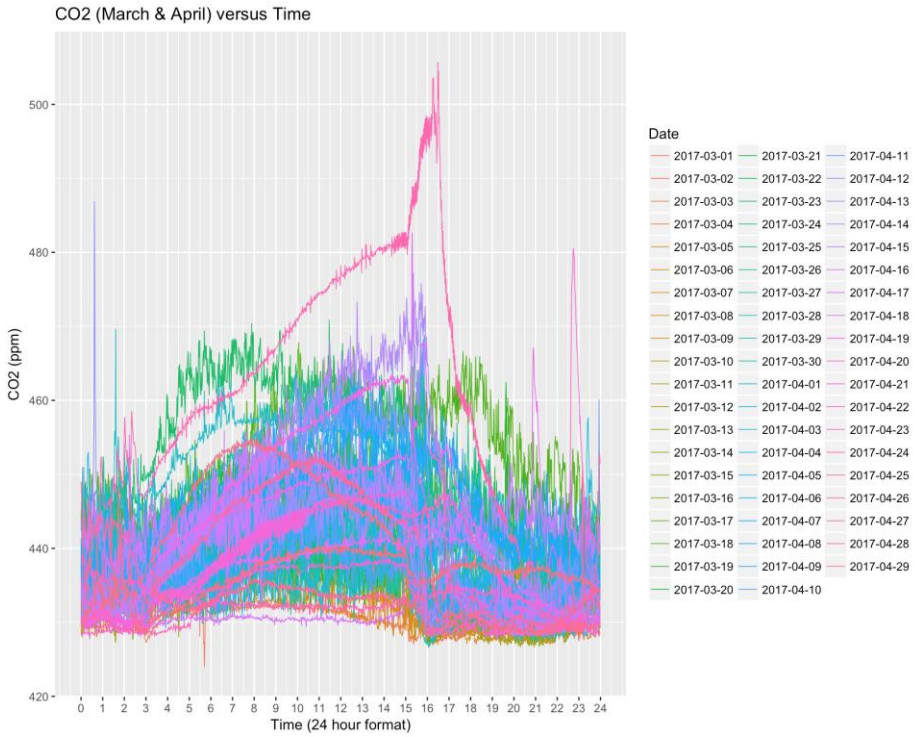


**Figure 21: $CO_2$ (March & April) versus Time**

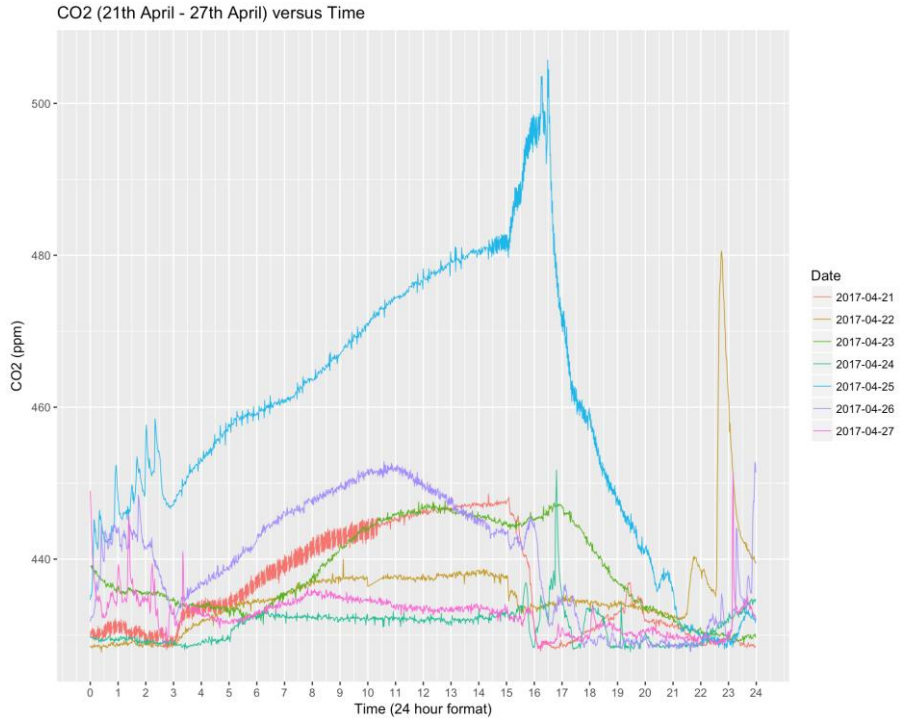


**Figure 22: $CO_2$ ($21^{th}$ April – $27^{th}$ April) versus Time**

Based on the correlation finding between $CO_2$ and VOC in section 3.1, it is noted that these two parameters change in the same way. Based on this finding, the line chart of weekday $CO_2$ concentration over a 24-hour period was plotted based on data collected over March and April

(Figure 21). From Figure 22, an abnormally high $CO_2$ concentration was observed on 25th April, Tuesday, from 2am to 3pm. To establish the reason behind this phenomenon, the profile of other parameters was used for comparison. However, no unusual change of other parameters was observed. As such, it is difficult to explain this abnormal change of $CO_2$ level due to a lack of understanding of the building characteristics.

## 3.11 Evening Breaks

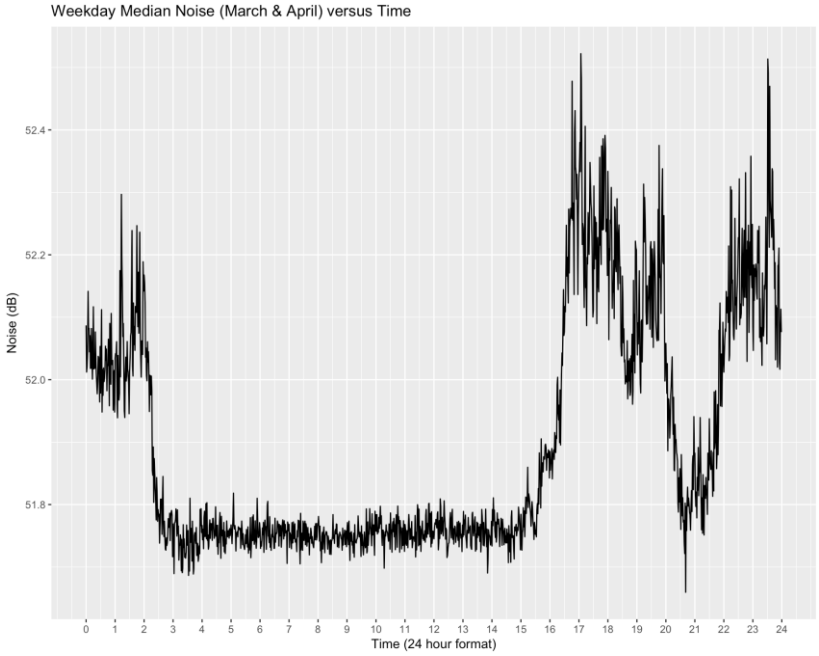


**Figure 23: Weekday Median Noise (March & April) versus Time**

With reference to Figure 23, a significant drop was found at approximately 8pm on weekdays. Based on the working hours assumed in section 3.3, such a drop can be considered as working breaks in the evening, approximately four hours after operation start time at around 3 to 4pm. During evening breaks, employees were likely to go out for dinner, leading to a decrease in noise.

# 4 Conclusion

In this report, two datasets from four IoT sensors embedded in a commercial building were analysed to derive characteristics of its operations. During the data preparation stage, no duplicated values or incomplete recorded observations were found. However, it was observed that around 3.5% of expected data to be collected during this period were missing completely at random. No data was imputed as a result, and a total of 11 findings such as the operating hours of the HVAC system in the building, and the working hours of the building occupants were found. Based on our observations, it's likely that the sensors located in this building are either located underground or located in a room without any windows, and is close to a location where there are equipment running 24-7 (e.g. server room) due to the consistent and uninterrupted ambient noise recorded over these 2 months.

# Appendix A

All the posted data points statistics of four sensors by day in both datasets of March and April corresponding to the normal dates distributions is listed in Table A below:

**Table A: Posted Data Points Statistics of Sensors by Day**

| NO. | Date | Weekday | SS0029 | SS0031 | SS0036 | SS0050 | Sum |
|---|---|---|---|---|---|---|---|
| **March** | | | | | | | |
| **1** | 2017-03-01 | Wednesday | 0 | 0 | 1438 | 2 | 1440 |
| **2** | **2017-03-02** | **Thursday** | **4** | **1088** | **335** | **5** | **1432** |
| **3** | 2017-03-03 | Friday | 2 | 1247 | 190 | 1 | 1440 |
| **4** | 2017-03-04 | Saturday | 8 | 173 | 1259 | 0 | 1440 |
| **5** | 2017-03-05 | Sunday | 9 | 0 | 1431 | 0 | 1440 |
| **6** | 2017-03-06 | Monday | 8 | 0 | 1432 | 0 | 1440 |
| **7** | 2017-03-07 | Tuesday | 7 | 0 | 1433 | 0 | 1440 |
| **8** | 2017-03-08 | Wednesday | 8 | 0 | 1432 | 0 | 1440 |
| **9** | 2017-03-09 | Thursday | 3 | 1025 | 412 | 0 | 1440 |
| **10** | **2017-03-10** | **Friday** | **22** | **1387** | **1** | **13** | **1423** |
| **11** | **2017-03-11** | **Saturday** | **1** | **1048** | **0** | **390** | **1439** |
| **12** | 2017-03-12 | Sunday | 0 | 1 | 5 | 1434 | 1440 |
| **13** | 2017-03-13 | Monday | 0 | 0 | 10 | 1430 | 1440 |
| **14** | 2017-03-14 | Tuesday | 0 | 0 | 3 | 1437 | 1440 |
| **15** | 2017-03-15 | Wednesday | 1244 | 131 | 15 | 50 | 1440 |
| **16** | 2017-03-16 | Thursday | 1209 | 194 | 0 | 37 | 1440 |
| **17** | 2017-03-17 | Friday | 0 | 833 | 39 | 568 | 1440 |
| **18** | 2017-03-18 | Saturday | 0 | 1 | 96 | 1343 | 1440 |
| **19** | 2017-03-19 | Sunday | 0 | 1 | 3 | 1436 | 1440 |
| **20** | 2017-03-20 | Monday | 16 | 26 | 12 | 1386 | 1440 |
| **21** | 2017-03-21 | Tuesday | 619 | 815 | 1 | 5 | 1440 |
| **22** | 2017-03-22 | Wednesday | 1402 | 31 | 0 | 7 | 1440 |
| **23** | 2017-03-23 | Thursday | 1436 | 0 | 0 | 4 | 1440 |
| **24** | 2017-03-24 | Friday | 1393 | 2 | 36 | 9 | 1440 |
| **25** | 2017-03-25 | Saturday | 1425 | 0 | 15 | 0 | 1440 |
| **26** | **2017-03-26** | **Sunday** | **884** | **478** | **17** | **0** | **1379** |
| **27** | 2017-03-27 | Monday | 7 | 881 | 552 | 0 | 1440 |
| **28** | 2017-03-28 | Tuesday | 53 | 8 | 390 | 961 | 1412 |
| **29** | 2017-03-29 | Wednesday | 1290 | 140 | 3 | 7 | 1440 |
| **30** | 2017-03-30 | Thursday | 0 | 1355 | 69 | 16 | 1440 |
| **31** | **2017-03-31** | **Friday** | **0** | **0** | **0** | **0** | **0** |
| **March Total** | | | **11050** | **10865** | **10629** | **10541** | **43085** |

| NO. | Date | Weekday | SS0029 | SS0031 | SS0036 | SS0050 | Sum |
|---|---|---|---|---|---|---|---|
| **April** | | | | | | | |
| **1** | 2017-04-01 | Saturday | 2 | 1 | 944 | 493 | 1440 |
| **2** | 2017-04-02 | Sunday | 0 | 0 | 1429 | 11 | 1440 |
| **3** | 2017-04-03 | Monday | 0 | 202 | 1222 | 16 | 1440 |
| **4** | 2017-04-04 | Tuesday | 0 | 1435 | 5 | 0 | 1440 |
| **5** | 2017-04-05 | Wednesday | 0 | 1417 | 23 | 0 | 1440 |
| **6** | 2017-04-06 | Thursday | 0 | 1302 | 137 | 1 | 1440 |
| **7** | 2017-04-07 | Friday | 0 | 53 | 1377 | 10 | 1440 |
| **8** | 2017-04-08 | Saturday | 618 | 0 | 629 | 193 | 1440 |
| **9** | 2017-04-09 | Sunday | 1438 | 0 | 0 | 2 | 1440 |
| **10** | 2017-04-10 | Monday | 1433 | 0 | 0 | 7 | 1440 |
| **11** | 2017-04-11 | Tuesday | 1440 | 0 | 0 | 0 | 1440 |
| **12** | 2017-04-12 | Wednesday | 1421 | 0 | 0 | 19 | 1440 |
| **13** | 2017-04-13 | Thursday | 856 | 15 | 13 | 556 | 1440 |
| **14** | 2017-04-14 | Friday | 521 | 1 | 32 | 886 | 1440 |
| **15** | 2017-04-15 | Saturday | 486 | 1 | 784 | 169 | 1440 |
| **16** | **2017-04-16** | **Sunday** | **23** | **971** | **400** | **37** | **1431** |
| **17** | 2017-04-17 | Monday | 1 | 1436 | 0 | 3 | 1440 |
| **18** | 2017-04-18 | Tuesday | 0 | 1342 | 0 | 98 | 1440 |
| **19** | 2017-04-19 | Wednesday | 39 | 1293 | 0 | 108 | 1440 |
| **20** | 2017-04-20 | Thursday | 2 | 1432 | 0 | 6 | 1440 |
| **21** | 2017-04-21 | Friday | 2 | 1053 | 0 | 385 | 1440 |
| **22** | **2017-04-22** | **Saturday** | **3** | **1406** | **0** | **0** | **1409** |
| **23** | 2017-04-23 | Sunday | 700 | 740 | 0 | 0 | 1440 |
| **24** | 2017-04-24 | Monday | 1399 | 41 | 0 | 0 | 1440 |
| **25** | 2017-04-25 | Tuesday | 1350 | 90 | 0 | 0 | 1440 |
| **26** | 2017-04-26 | Wednesday | 943 | 489 | 0 | 8 | 1440 |
| **27** | 2017-04-27 | Thursday | 1435 | 0 | 0 | 5 | 1440 |
| **28** | 2017-04-28 | Friday | 1432 | 3 | 0 | 5 | 1440 |
| **29** | 2017-04-29 | Saturday | 1436 | 0 | 0 | 4 | 1440 |
| **30** | **2017-04-29** | **Sunday** | **0** | **0** | **0** | **0** | **0** |
| **April Total** | | | **16980** | **14723** | **6995** | **3022** | **41720** |
| **Total** | | | **28030** | **25588** | **17624** | **13563** | **84805** |