



EB5101 - Foundations of Business Analytics

Assignment:

IOT Sensor Dataset – Data Preparation using R

Lecturer: Prakash C Sukhwai

Student ID	Name	E-mail
A0178495J	Ankita Avadhani	e0267806@u.nus.edu
A0178493M	Bethapudi Keerthana	e0267804@u.nus.edu
A0178314A	Sabrish Gopalakrishnan	e0267625@u.nus.edu
A0178507W	Sanchit Mittal	sanchitmittal@u.nus.edu
A0056418A	Tan Zhi Wen David	e0267355@u.nus.edu

Submission Date: 11th March, 2018

Contents

Objective	3
Data Preparation.....	3
Analysis and Insights	4
Correlation Study	4
Exploratory Study.....	5
Sensor Level Study	11
Time-based Study	13
Hourly Plots.....	13
Monthly Plots.....	15
Conclusion.....	17
APPENDIX- Anomaly and Outlier Study – Additional Insights	18
Humidity and Temperature	18
CO ₂ and VOC	19
Light.....	20
Noise	23

Objective

Data from IOT sensors in commercial building measuring 6 environmental parameters was analyzed. The objective is to prepare the dataset to discover patterns and insights.

Data Preparation

The given data set provides observations about 6 environmental parameters: temperature, noise, light, CO₂, VOC and humidity. Four different sensors were used to record the observations. It is not clear as to whether the sensors are in the same room or location. This needs to be established as part of the analysis. It was observed that data obtained from these sensors had certain issues. The issues and their treatment are discussed below.

Issue 1: Sometimes the sensors may malfunction and read abnormal values

Treatment: The sensor-malfunctioning can be interpreted as outliers being present in the data. An initial analysis shows that 12% of the observations could be categorized as outliers. It was however decided to not exclude them since these observations can sometimes be recordings of anomalous behavior of indoor systems instead of sensor malfunction. Data collected from these systems is intended to monitor and act on such anomalies. Excluding data points blindly using the trend line, without this consideration, defeats the purpose of IOT monitoring. A separate section on outliers and anomalies is included in the end as appendix to examine the possible reasons for these observations.

Issue 2: Sometimes due to network issues, same data points would be posted more than once.

Treatment: It was observed that all observations had unique timestamps. This shows that there were no duplicates in the time-stamp. It is however possible that an observation is not recorded at exact intervals in seconds. It was decided that these readings would be not removed as a moving average analysis using the rolled-up data would even out such anomalies.

Issue 3: Sometimes sensors get disconnected with the network and data will not be recorded for that period.

Treatment: A look out for missing values showed that there were no NAs or blanks in the data. However, when the timestamps were rolled-up to a minute level, it was observed that readings were missing during certain missing minute-time intervals. The missing timestamps were then identified and evaluated on the type of “missingness”. As this is a time series data, the missingness can be classified as ‘missing conditionally at random’, conditionally dependent on the timestamp and the previous non-NA values. Imputation of numeric variables done using linear interpolation produced best results. Then to impute the sensor ID, which is a categorical variable, K-Nearest neighbors (KNN) technique was used, which examines entire dataset and complete row slices (i.e. all columns of the row) and imputes missing categories comparing the measurements of Temperature, Light, Humidity, etc. with the nearest non-NA N neighbors having a similar pattern.

Although, March has 31 days, the original dataset provided contains data only from 1st of March to 30th of March. Therefore, values for 31st of March were not imputed, assuming the possibility of the facility being shut on that day.

Analysis and Insights

Correlation Study

A study was conducted to understand the correlation between the variables.

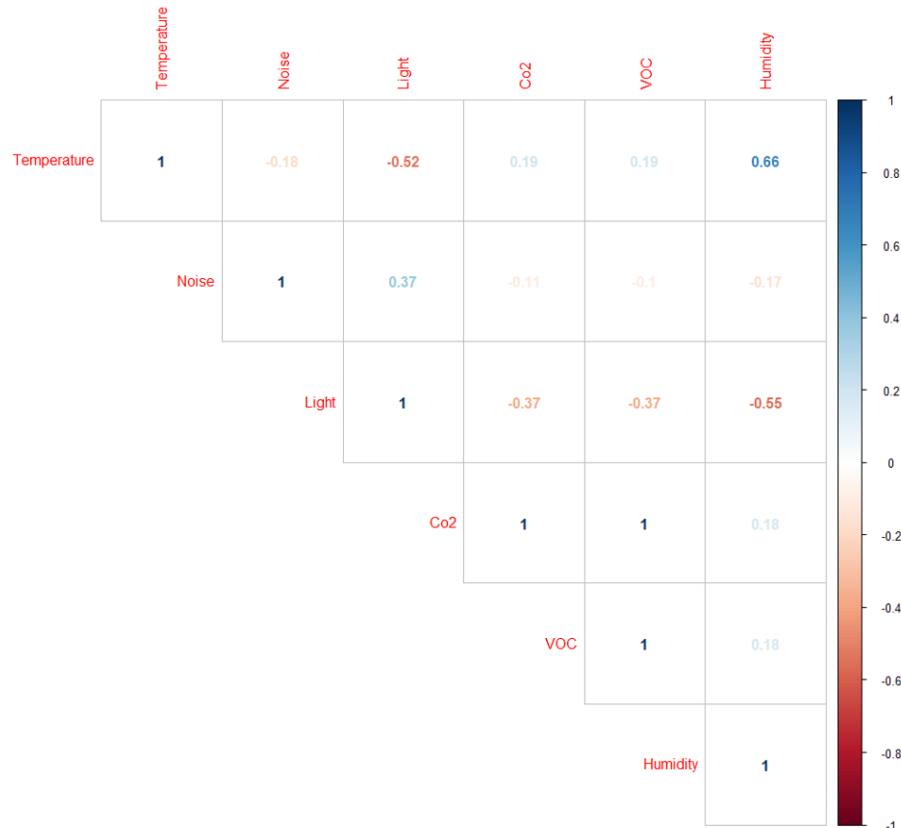


Figure 1: Correlation matrix of all variables

Insight1: The correlation plot as seen above reveals that the CO₂ and VOC (Volatile Organic Compound) exhibit perfect positive correlation ($r=1$). Assuming human respiration to be the major source of CO₂ in the closed room, we could infer that as the number of people in a room increases, CO₂ increases and VOCs increase correspondingly with the usage of equipment that release them.

Insight2: Temperature and humidity were two other variables that showed significant positive correlation ($r=0.66$). This is explained by the general phenomenon of warm air holding more moisture than cold air.

Exploratory Study

An exploratory study was conducted to understand the nature and behaviour of each variable. This study was conducted on the entire data (includes data of all four sensors).

Temperature:

Minimum	Maximum	No. of outliers (as seen in the box plot)
23.01667	26.4444	161

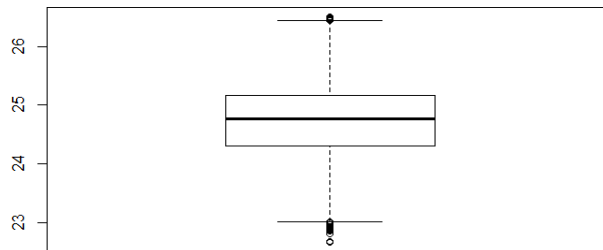


Figure 2: Temperature Boxplot

Insight 3: Since the range is only 3°C, it can be inferred that the environment has been maintained at a relatively constant temperature.

Noise:

Minimum	Maximum	No. of outliers (as seen in the box plot)
49.48	68.18	11094

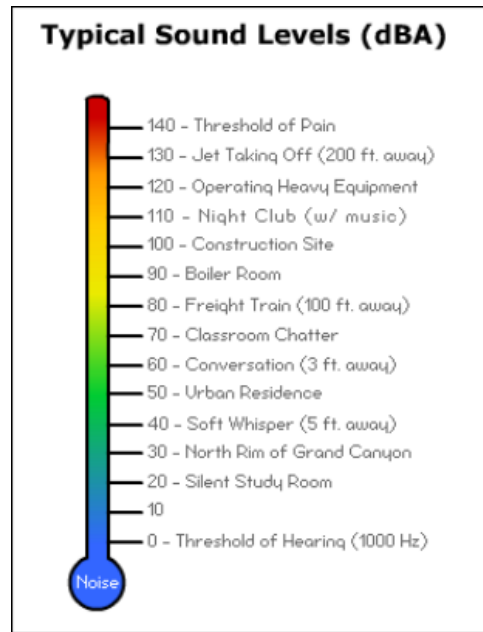


Figure 3: Typical sound levels(dB)

Source: <https://www.osha.gov/SLTC/noisehearingconservation#loud>

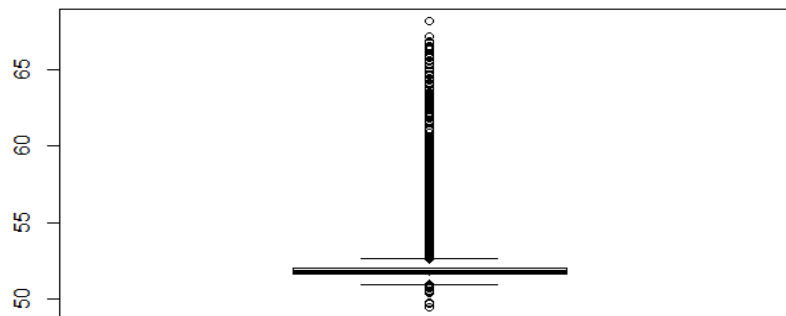


Figure 4: Noise Boxplot

Insight 4:

The readings on Noise were measured against the standard on noise levels at workplace from OSHA's website. It was observed that the environment is subject to basic conversation level noise which is around 50-60 decibels. From the box plot above, it can be further inferred that there is always a low-decibel background noise which may indicate that an equipment or a machine could be operating throughout the day.

Light:





	5,000	Overcast sky
	500	Well-lit office
	300	Minimum for easy reading
	50	Passageway/outside working area
	15	Good main road lighting
	10	Sunset
	5	Typical side road lighting
	2	Minimum security risk lighting
	1	Twilight
	0.3	Clear full moon
	0.1	Typical moonlight/cloudy sky
	0.001	Typical starlight

Figure 5: Lighting levels (Lux)

Source: https://www.use-ip.co.uk/datasheets/lux_light_level_chart.pdf

Minimum	Maximum	No. of outliers (as seen in the box plot)
~ 0.00	749.00	6

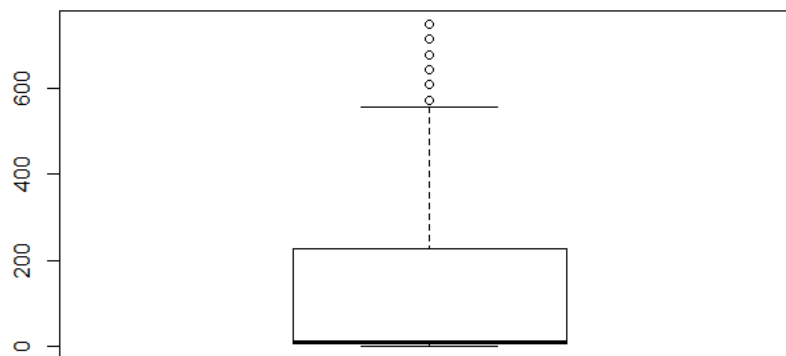


Figure 6: Light Boxplot

Insight 5: As seen in the boxplot above, as the median is very close to zero, it can be inferred that around 50% of the time the lights were turned off. Comparing the other 50% of the readings ranging between 0 to 500 lux (excluding outliers) with the chart above, it can be said that the environment is conducive for reading and working. This may also indicate a cyclic pattern of switching of lights which is indicative of an office or working space.

CO₂:

Minimum	Maximum	No. of outliers (as seen in the box plot)
424.0	505.7	2318

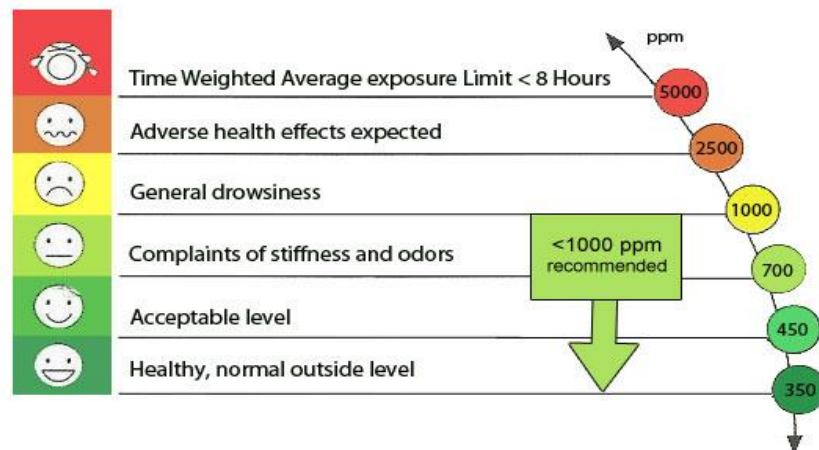


Figure 7: CO₂ Levels chart

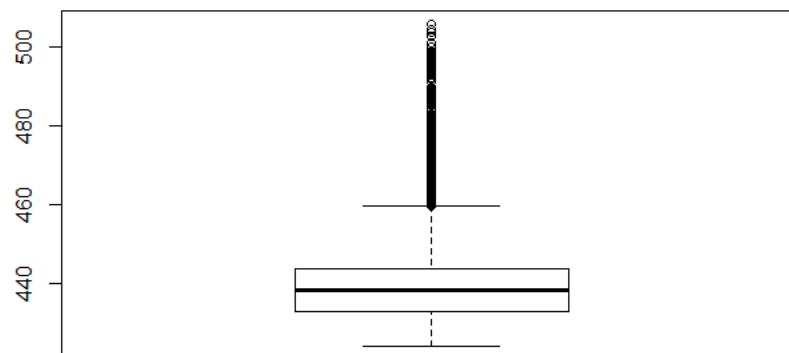


Figure 8: CO₂ Boxplot

Insight 6: The observed values seem to lie within acceptable limits when compared to CO₂ levels in the chart shown above. However, the points indicated as outliers in the box plot represent spikes in CO₂ levels which will be analysed further based on the timestamp.

VOC

Minimum	Maximum	No. of outliers (as seen in the box plot)
311.0	370.8	2318

Threshold Limit Values for several hazardous air pollutants [11].

Hazardous Air Pollutant	Threshold Limit Value (ppm)	
	8-Hour Time Weighted Average	15-Minute Short-Term Exposure Limit
Benzene	0.5	2.5
Xylenes	100	150
Trichloroethylene	50	100

Figure 9: Threshold limit for VOCs

(Source: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3909362/table/t8-sensors-05-00004/>)

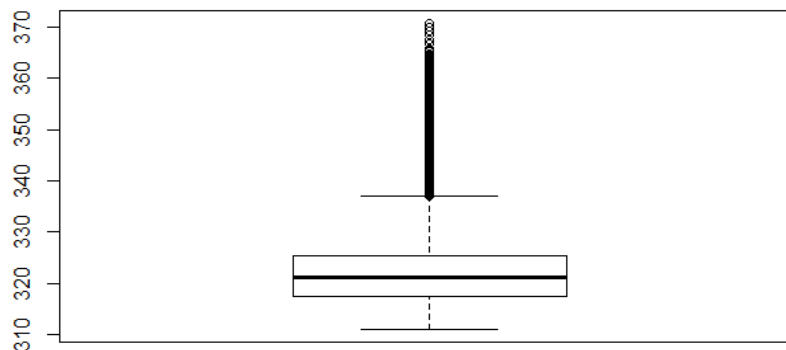


Figure 10: VOC Boxplot

Insight 7: The VOC readings show numerous spikes(outliers) which confirms a persistent emitter of VOC. As the most common sources of VOCs are certain chemicals and electronics, the readings could be attributed to the presence or usage of some object that is chemical or electronic in nature.

Humidity

Minimum	Maximum	No. of outliers (as seen in the box plot)
55.05556	72.4333	35

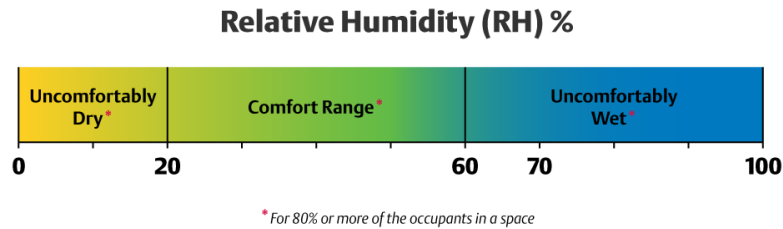


Figure 11: Relative Humidity Levels (source: <http://www.ac-heatingconnect.com/wp-content/uploads/ACHC-Home-Humidity-Relative-Humidity.png>)

Insight 8: Comparing the results of the boxplot (Figure 12) with the relative humidity levels (Figure 11) shows that the environment where the sensors are placed, is usually very humid with more than half of the observations lying above the 60% humidity level.

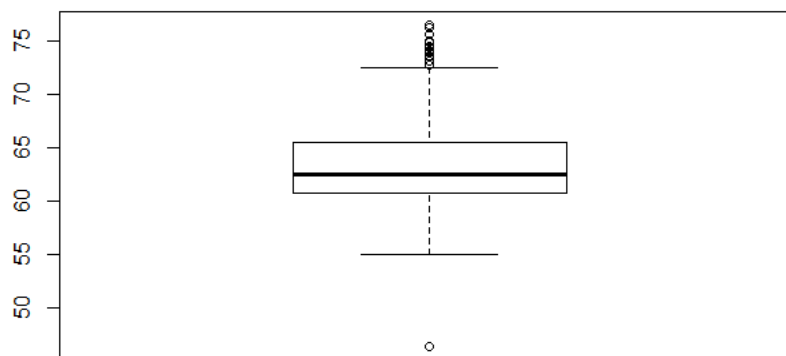


Figure 12: Humidity Boxplot

The values above 60% could indicate that this facility might be in a place with a tropical and humid climate like Singapore. While the other half of the readings, which are lower than 65%, could be due to the air-conditioned environment of the facility during its operating hours.

Based on the above exploratory study, it is likely that this facility could be housing one of these:

- Data Centre
- Chemical Lab
- Warehouse

Further sensor level and time- based analysis needs to be done in order to establish other facts.

Sensor Level Study

Temperature, noise and light measurements were compared from Sensors SS0029, SS0031, SS0036 and SS0050 to build a hypothesis of whether the sensors were in a close proximity or not. If they were in the same enclosure, Temperature, Noise and Light readings should be similar in magnitude. To evaluate this, visual inspections using boxplots (medians) and one-way ANOVA test were conducted on these 3 variables:

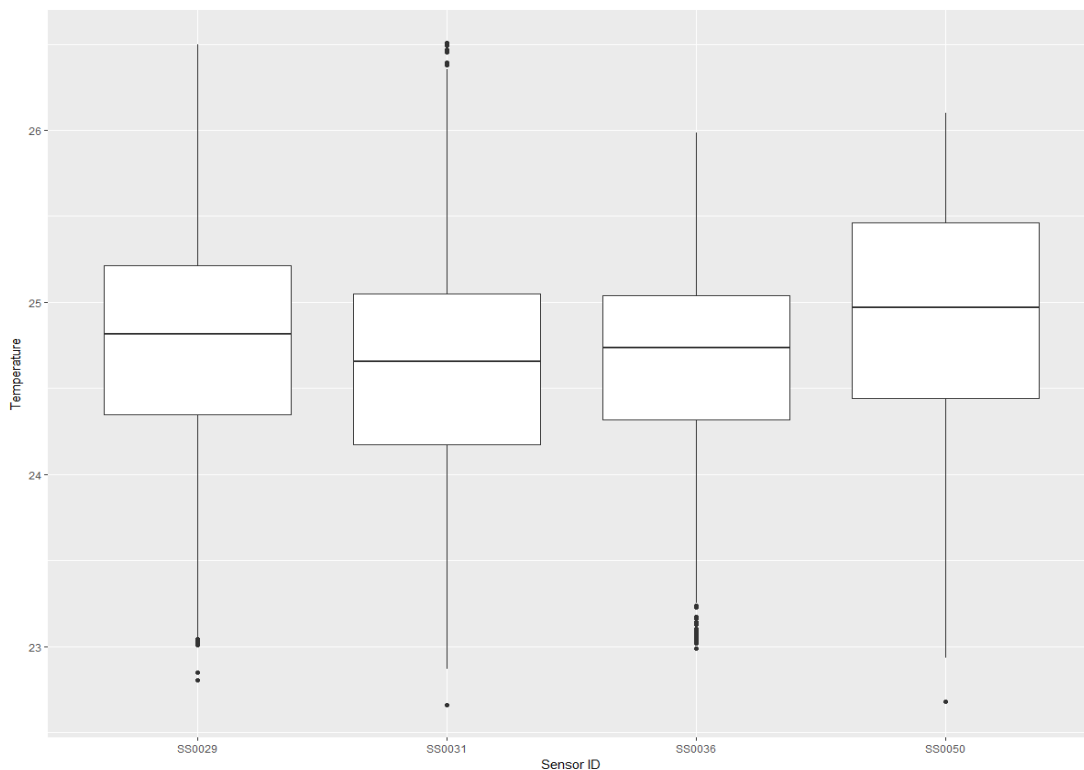


Figure 13: Temperature

Insight 9: Visual Inspection shows minor differences in the medians; the highest median temperature being recorded by sensor SS0050. An ANOVA test produced the following results:

```
> summary(aov(data_impute_3$Temperature~data_impute_3$unitid))
              Df Sum Sq Mean Sq F value Pr(>F)
data_impute_3$unitid    3    777   258.87   718.8 <2e-16 ***
Residuals             84956   30597     0.36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic tells us that there is statistically significant difference in the levels of recordings for

each sensor since null hypothesis does not hold. This indicates a high probability of the sensors being located far apart from each other.

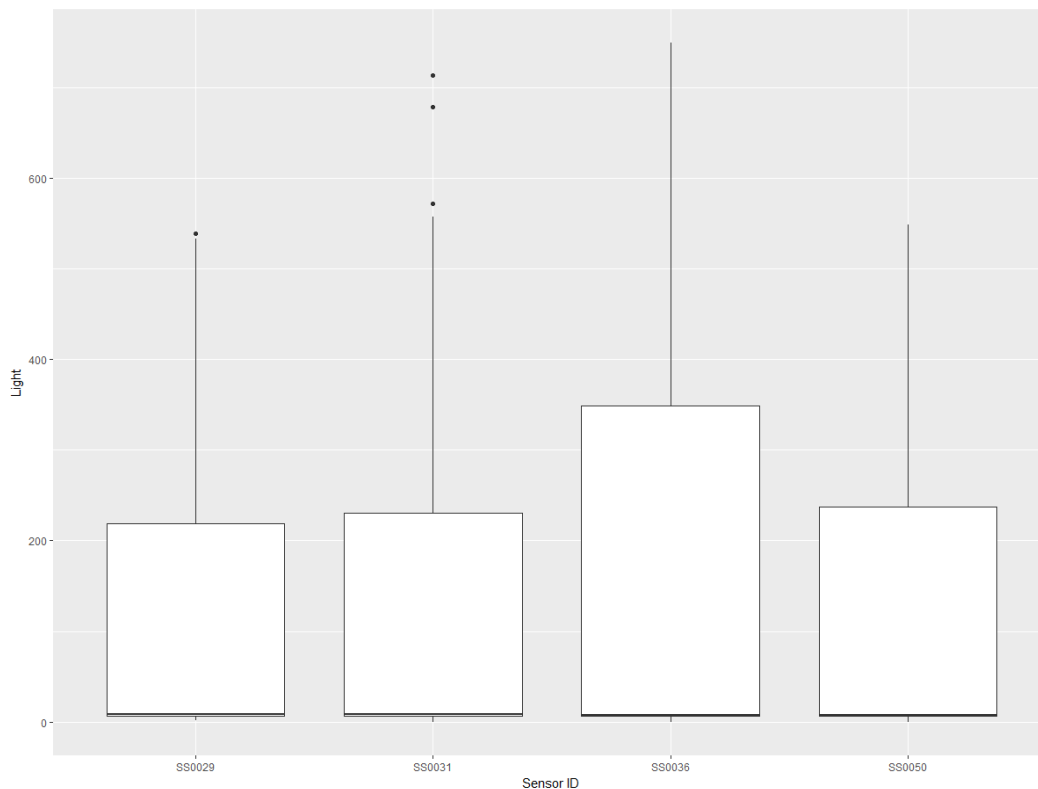


Figure 14: Light

Insight 10: The lower median of 8LUX for sensors SS0036, SS0050, SS0029 and SS0031 indicates that the lights in the facility are turned off 50% of the time. As it is slightly above 0, it indicates the presence of a constant dull/dim source of light and this could mean that these sensors might be placed in or near a corridor or lobby or some equipment emitting a dull light.

Furthermore, SS0036 seems to be located at or near a well-lit workspace. The results from the ANOVA test are shown below:

```
> summary(aov(data_impute_3$Light~data_impute_3$unitid))
              Df    Sum Sq Mean Sq F value    Pr(>F)    
data_impute_3$unitid    3  3.917e+06 1305705   43.51 <2e-16 ***
Residuals              84956  2.550e+09    30011
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

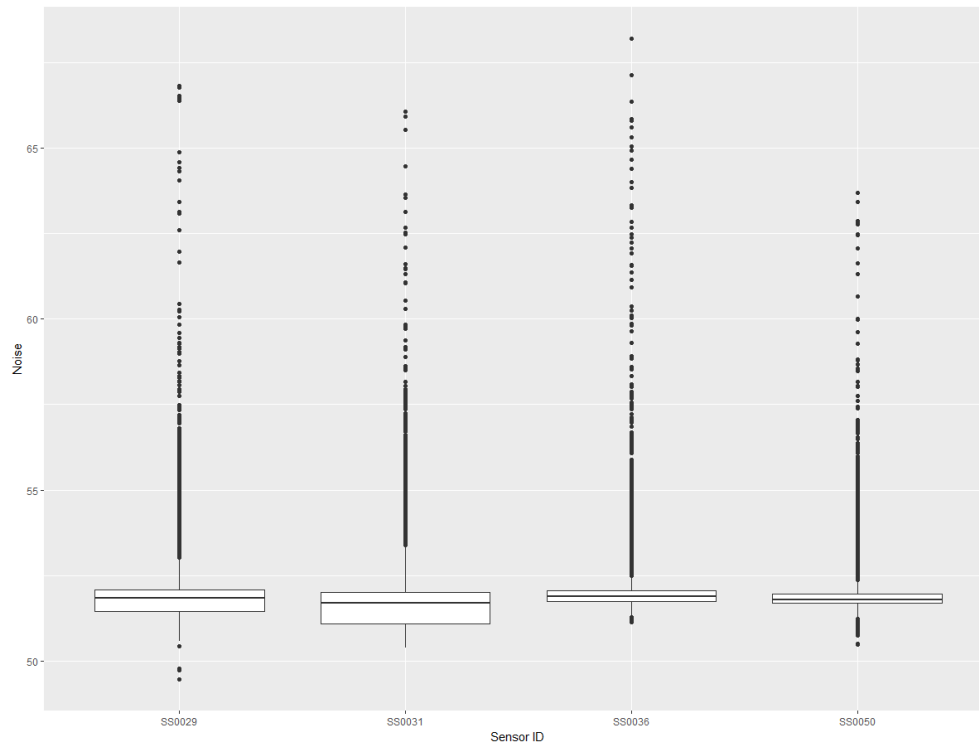


Figure 15: Noise

Insight 11: Sensors SS0036 and SS0050 have lesser spread than sensors SS0031 and SS0029. As the sound level drop in SS0036 and SS0050 is lesser as compared to other sensors, it can be inferred that these sensors may be located near a constant source of sound such as an electronic equipment that is in operation throughout.

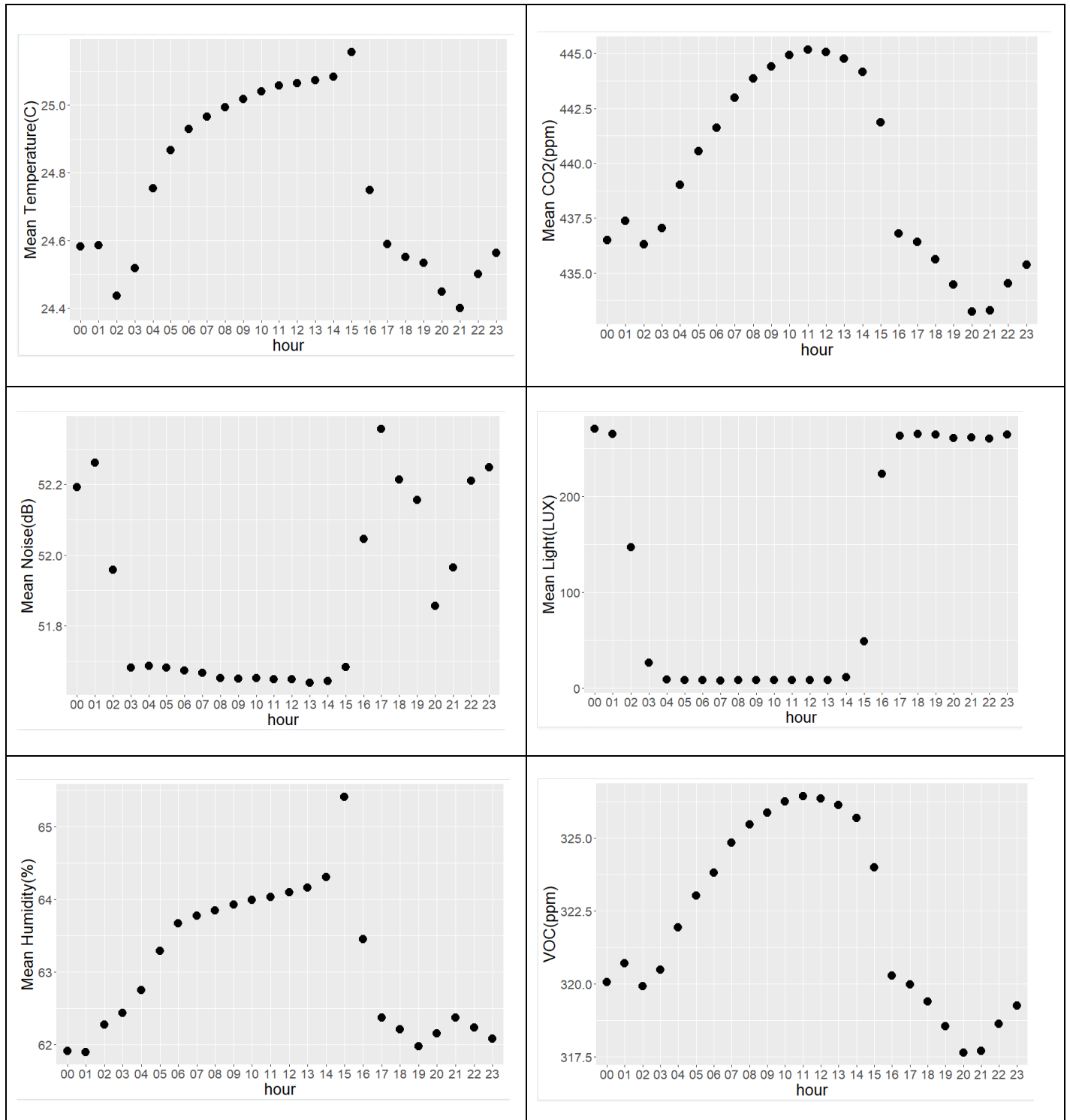
```
> summary(aov(data_impute_3$Noise~data_impute_3$unitid))
              Df Sum Sq Mean Sq F value Pr(>F)
data_impute_3$unitid    3    905   301.60   492.9 <2e-16 ***
Residuals              84956   51988     0.61
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test results of Noise also point to the same observation that the sensors may be apart from each other.

Time-based Study

Hourly Plots

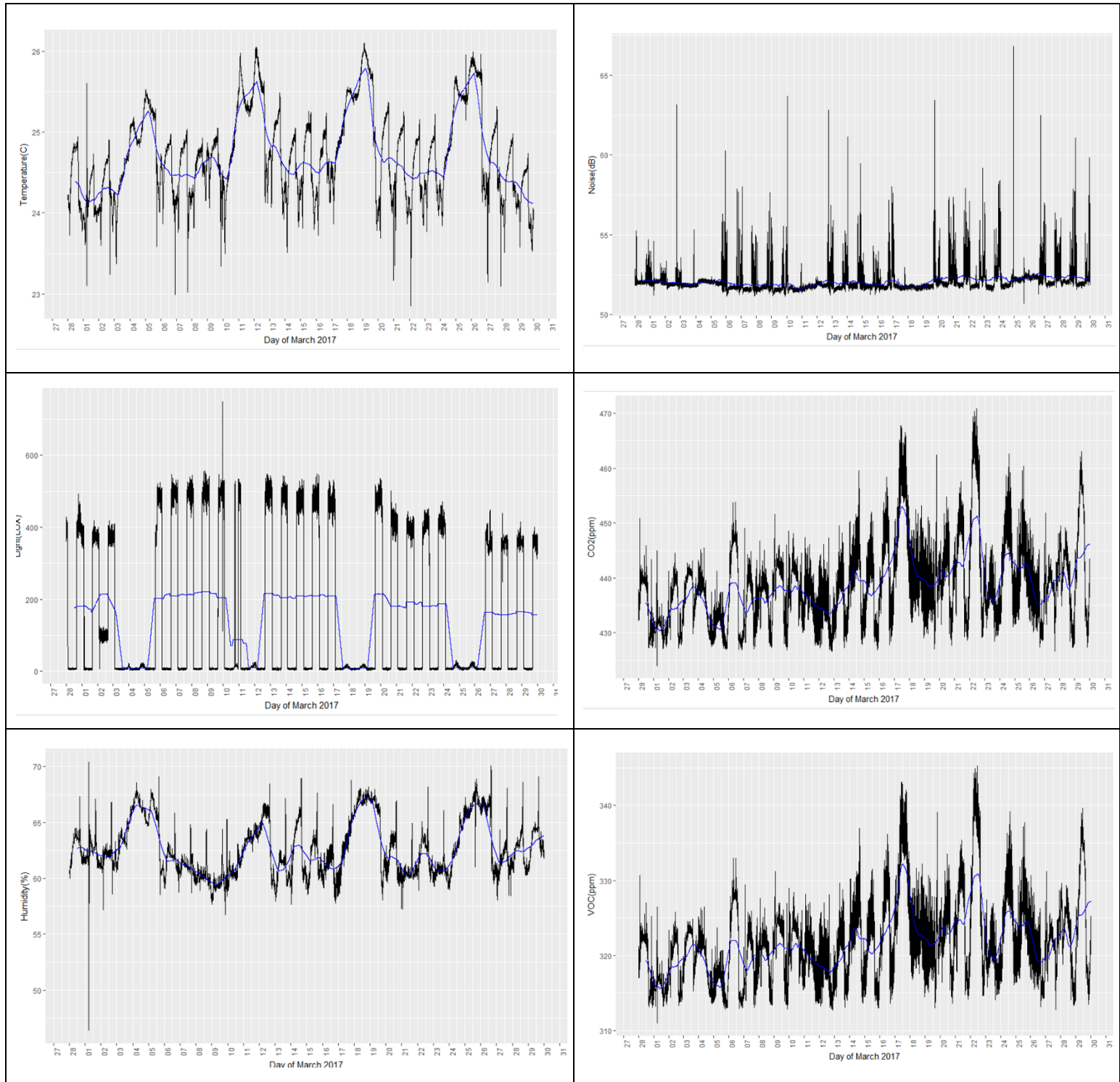
The time series data was analysed to find patterns and draw inferences on time-based measures such as operating hours of the facility and time of maximum occupancy on a given day. A **new feature called hour** was created from the date time stamp over the entire dataset of March and April to examine the patterns of the environment. Plotted below are hourly averages of all 6 variables: Temperature, Co2, Noise, Light, VOC and Humidity.



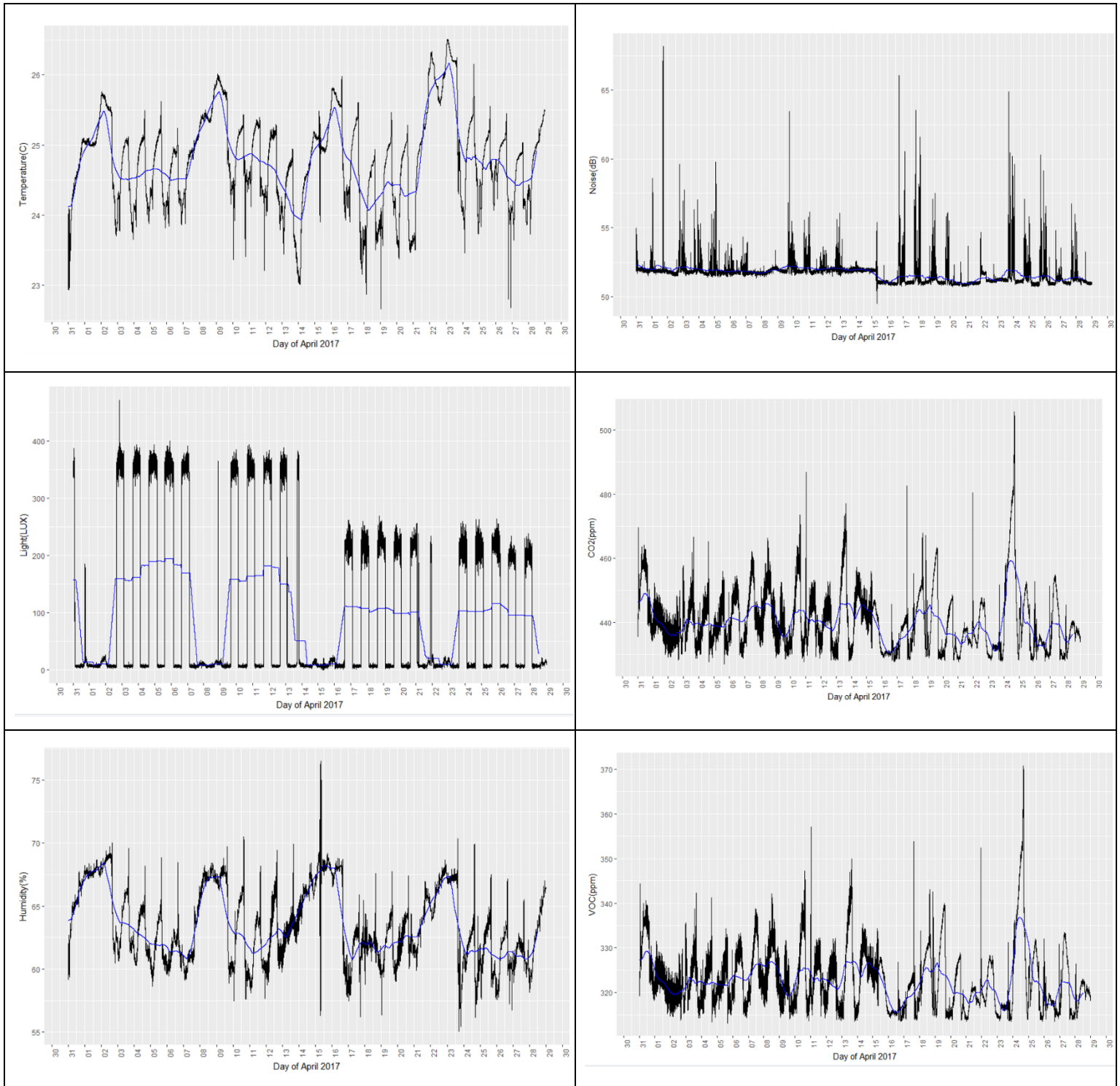
Insight 12: The hourly plots revealed a pattern suggesting that the facility under study could be operating on a shift basis. Variables such as Temperature, Humidity, CO₂ and VOC were increasing during the period 3AM to 3PM while as Noise and Light were relatively low and constant. It was also noticed that while Noise and Light increased between 3PM and 3AM, all other parameters decreased.

Monthly Plots

The time-series plots for the month of March are shown below. 6 new features which are **Moving daily average of each of the 6 variables** were created by rolling up to $N = 1440$ minutes = 1 Day. This was done for ease of interpretation and for better clarity in trends/patterns.



The time-series plots for the month of April are shown below. The blue line is the moving average of the data.



Insight 13: The monthly plot for March revealed the following patterns:

- On an average, temperature and humidity were lower during weekdays and increased during weekends. As most work places do not operate during weekends, this pattern could be due to the air-conditioning system being switched off during weekends allowing outside air to affect the indoor environment. The work week pattern of this facility is further substantiated by the Light trends which also indicated that the facility might not be operating on Saturdays and Sundays.
- An anomalous activity was observed on March 3rd wherein the lights were kept ON during the general non-working hours of 2am – 4pm. However, no spikes in CO₂ emissions or Noise levels (which are indicators of human activity) were observed. This indicates that the light may not have been turned OFF as per normal schedule.
- An anomaly, again, was noticed during the weekend of 11th March 2017 as lights were turned on around 6pm, turned off around 9pm and were turned on again by 11pm. They continued to stay on until the next morning 2am. It is possible that one or more of the employees came back to the workplace to finish some pending work.
- On 23rd of March, a sudden spike was observed in the CO₂ and VOC levels, which could be due to a large group of people gathering at the workplace.
- Also, the VOC and CO₂ levels were relatively higher in the second half of March (after 15th) and this could mean that more people were working in the office towards the end of the financial year.

Insight 14: The monthly plot for April revealed the following patterns:

- A similar pattern of variables during weekdays and weekends was noticed in April like in March. On a certain weekend, the facility seems to have remained closed for longer than 2 days. It was noticed that this was a long weekend when Good Friday fell on 14th April and the subsequent 2 days were Saturday and Sunday. Work seems to have resumed on the 17th, which was a Monday.
- On 25th of April, a spike was again noticed in the CO₂ levels which could imply a large gathering in the office.
- The average background noise levels dropped after 16th April and a difference was noticed in the noise levels during the 1st and 2nd halves of the month. This could be attributed to various reasons: such as fewer equipment operating, or lesser number of people around. It is therefore difficult to attribute this to one specific cause.

Conclusion

Based on the insights drawn at different stages, the facility under study closely fits with the characteristics of a data center. Although this cannot be established with utmost certainty, the high levels of VOCs in the environment point towards usage of equipment such as computers, printers, photocopiers and other electronic equipment that are known to emit VOCs. The irregular work shifts, noise and light patterns also point towards the maintenance and down-time of the servers in the data center.

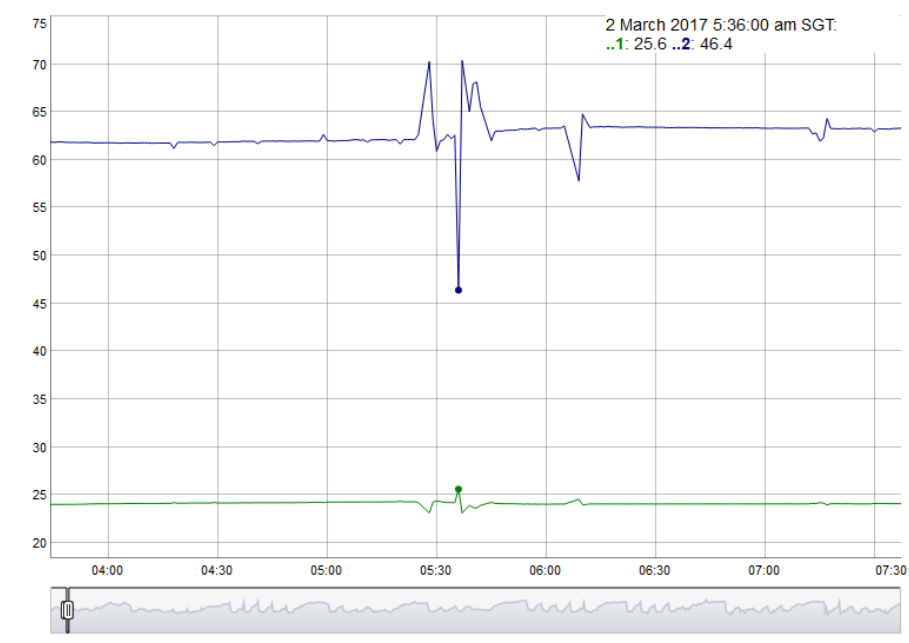
It is also noted that the humidity and VOCs levels in this facility sometimes fall in to the unacceptable ranges as this could cause serious health concerns in the personnel working in this environment. It is therefore important that proper ventilation systems are installed in this facility to improve the indoor air quality.

APPENDIX- Anomaly and Outlier Study – Additional Insights

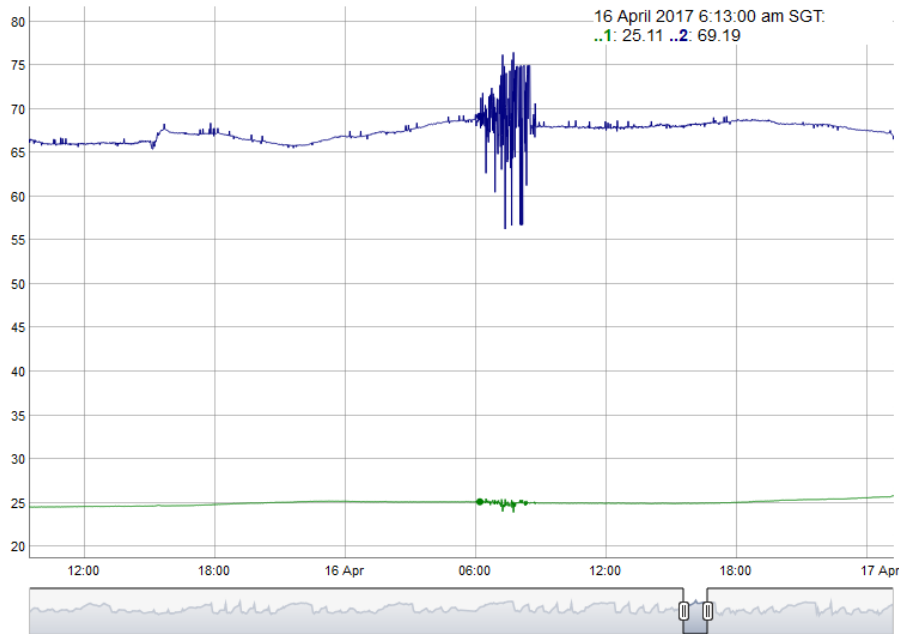
This study was done using package dygraphs and xts Time-Series Data Frames.

Humidity and Temperature

There is an abnormally low value of humidity on 2nd March 05:36 am of 46.4 with a minor surge of temperature. This phenomenon started at 5:35 am and receded by 5:37 am. The humidity stabilized by 5:45 am. This is in fact, the global minima for humidity

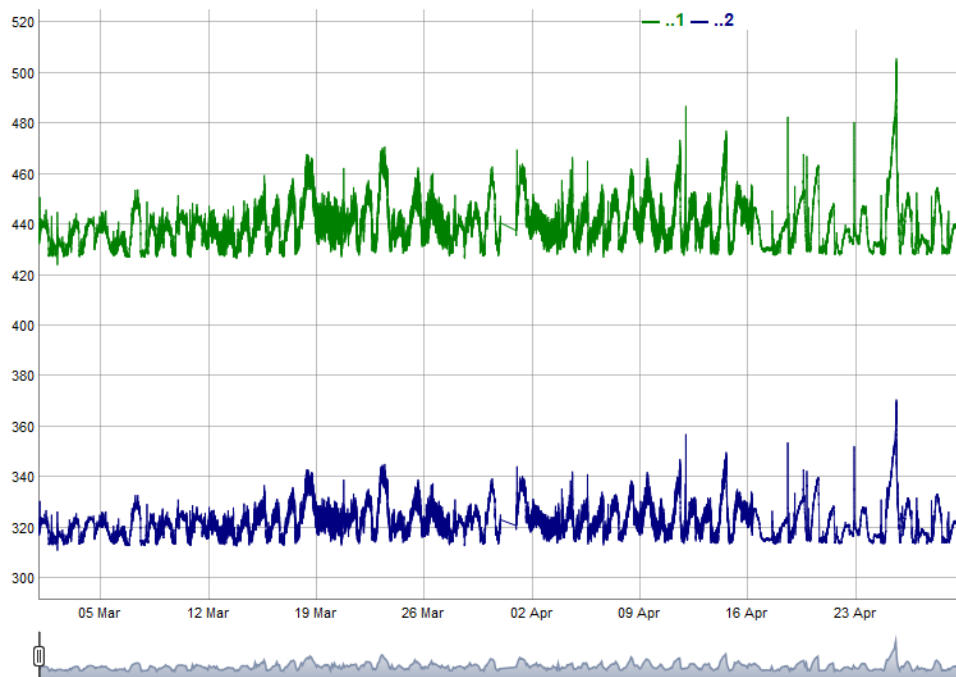


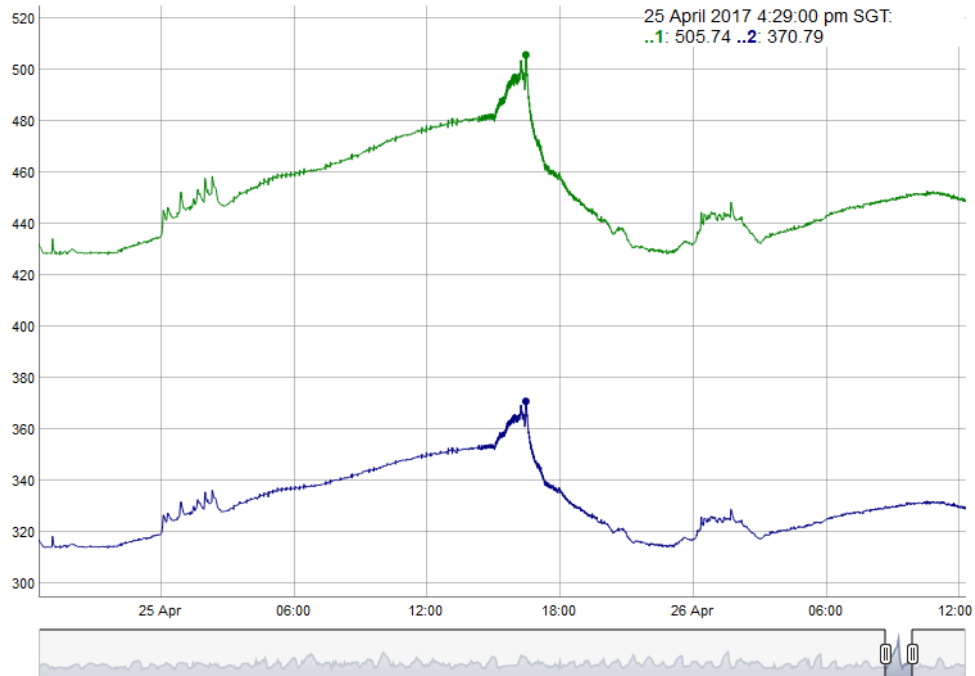
There is an erratic pattern of humidity fluctuations observed on 16th April from 6:13 am all the way up to 8:48 am. This is accompanied by small fluctuations in temperature also. Humidity seems to have adhered thereafter to cyclicity.



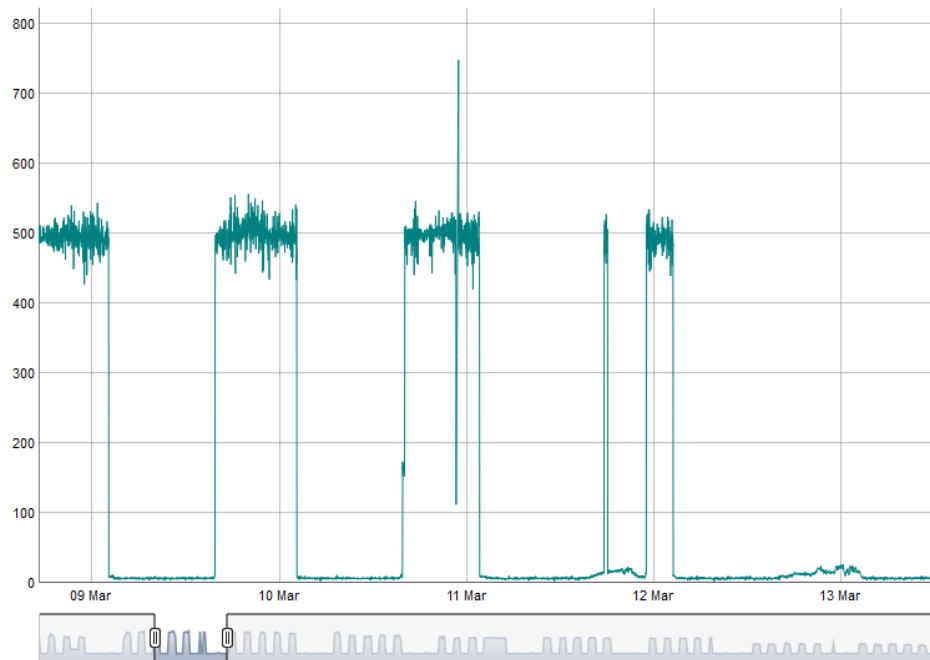
CO₂ and VOC

A comparison over two months reveals that the pattern gradually spreads out over the two months and hence finding outliers here would eliminate important information such as the spike in both CO₂ and VOC on 25th April at 4:29 pm which are the global maxima and indicative of a flash rise in emissions detected by system due to some event near the sensor SS0029. The maxima reached together along with the fact that their sample correlation is 1 justifies the observation's non-exclusion as an outlier observation

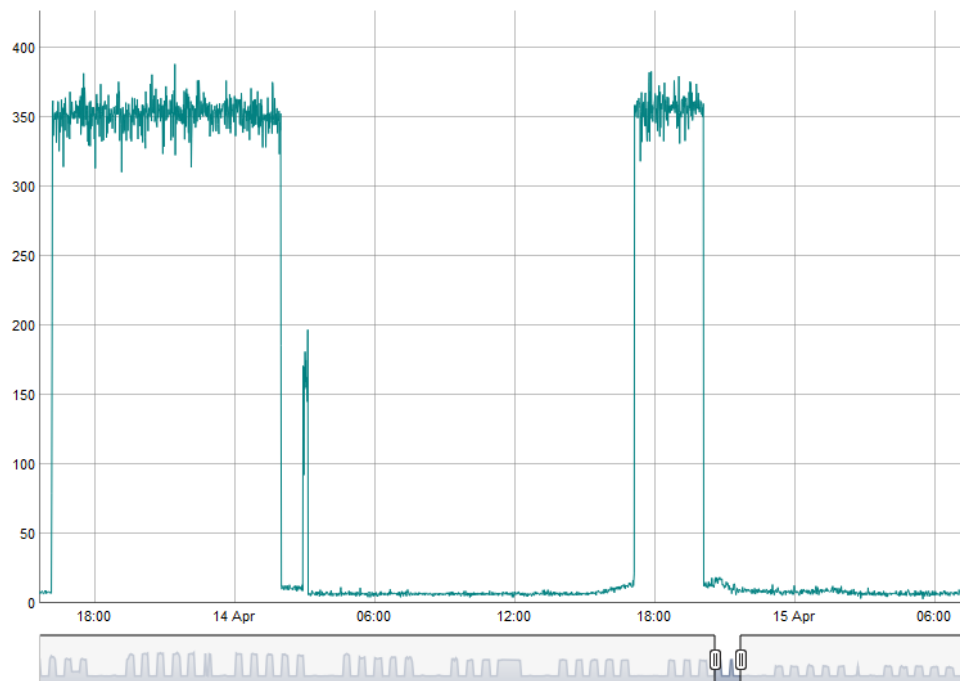
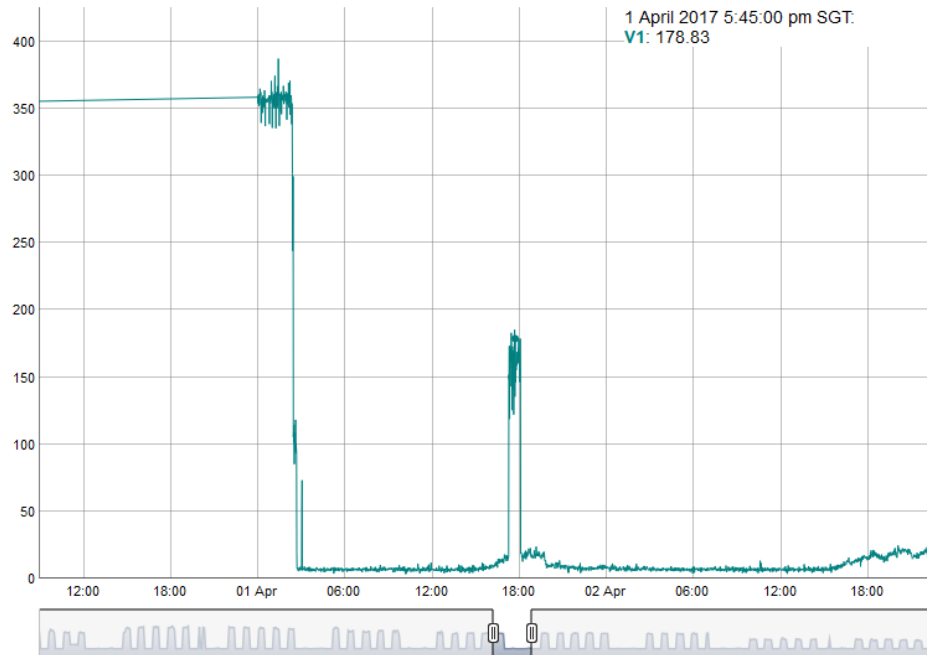


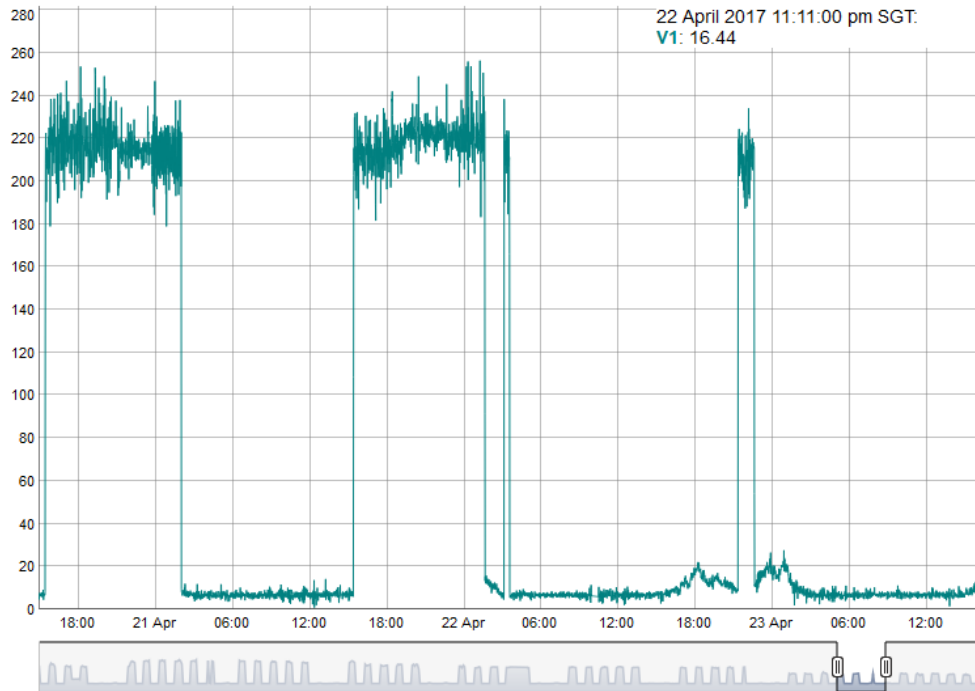


Light

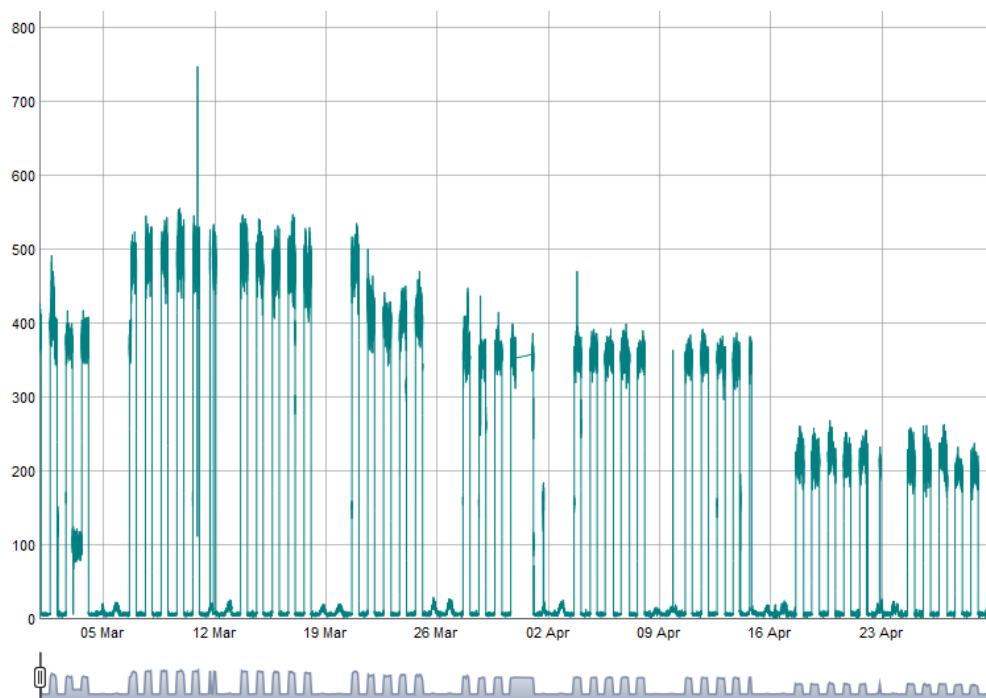


There is a sudden and random dimming at 10:44 pm followed by spiked brightening of light source up to 11:02 pm on 10th March. This phenomenon may be indicative of issues with neutral wiring of the light source and may warrant inspection.

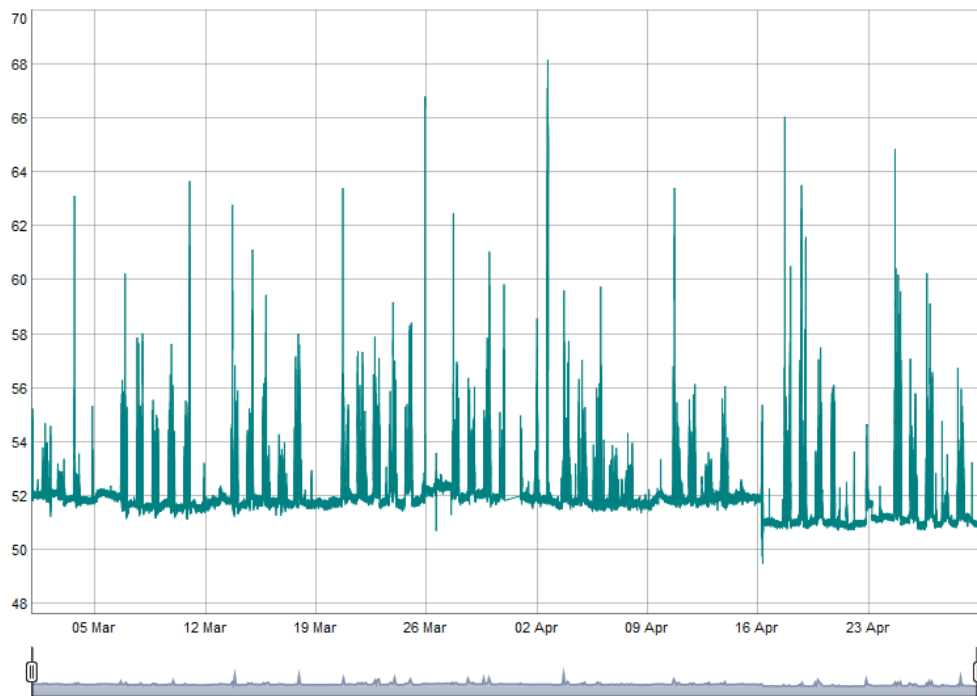




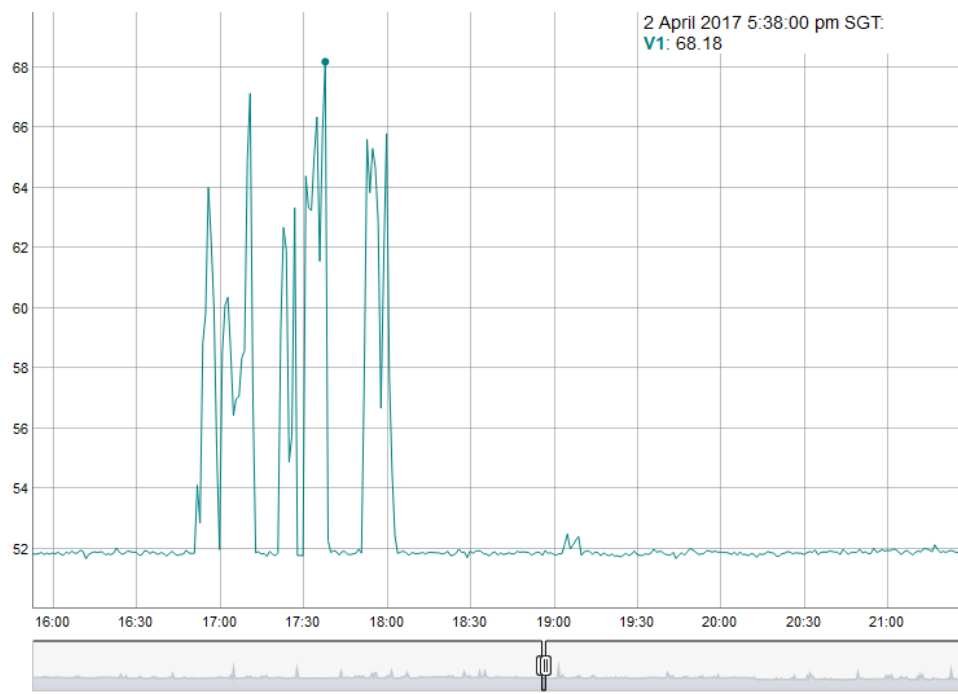
There was an anomalous usage of light on 11 & 12 March Sunday & Monday: at 5:41 to 6:10 pm on 11th and 11:05pm to 02:35 am on 11-12th. Similarly, on 1st April 5:17pm to 6:10 pm, 14th April (Good Friday) 5:04pm to 8:10 pm and 22nd April 9:23pm to 10:46 pm.



Noise



As discussed earlier there is a minute but sharp drop in background noise from 16th April onwards. This may be a deliberate calibration effort by the administration team to improve quality of noise recorded over background noise.



Global maxima is reached on 2nd April at 5:38 pm. The pattern may indicate some short event or announcement/broadcast in workspace.