3/11/2018

DATA PREPARATION ASSIGNMENT

An analysis of IOT sensors data

Prepared by:
TEAM DATA INSIGHTERS
Sreekar Bethu
Mohammed Ismail Khan
Meghna Vinay Amin
Balacoumarane Vetrivel

Table of Contents

1. Introduction:

IOT sensors have been designed to measure 6 environmental parameters such as temperature, humidity, CO2, VOC (Volatile Organic Compound), Light and Noise in a closed room. Data has been collected from sensors for two months (March'17 & April'17).

Few issues while collecting data are as follows:

a) Sometime the sensors malfunctioned and read abnormal values.
b) IOT sensors post data to database using Wi-Fi network. Sometimes due to network issues, same data points are posted consecutively.
c) Sometimes sensors get disconnected from the network and data does not get recorded for that period.

1.1 Objective:

1. To analyze the given issues in the data, impute missing, duplicate & abnormal data where necessary.
2. To analyze the prepared data and provide findings/inferences based on the understanding of the recordings

2. Data preparation:

Two data files were handed for analysis, one from 1st to 30thMarch 2017 and the other from 1st to 29th April 2017.

a) Two files were merged into a single file for further processing.
b) In merged file, 'date_time' column is renamed to 'time_1'.
c) Data type of 'date_time' is changed to POSIXct and time zone is forced to 'Asia/Kuala_Lumpur'. This data frame (Data_1_2) is copied to new data frame (Data_merge) for further analysis.
d) 'date_time' value of the next recording of each record is stored as a new variable feature created as 'time_2', to obtain the time interval between two consecutive recordings.
e) For each observation, interval between the two times 'time_1' and 'time_2' is calculated and stored in terms of seconds, minute, hours and days in the variables 'seconds', 'minute', 'hours' and 'days' respectively.
f) As there are no recordings on 31st of March, the time difference between last recording of 30th March and first recording of 1st April is 1 day, and hence the last observation 30th march is ignored.
g) Date is fetched from 'time_1' variable and placed into a new 'only_date' variable, to get the number of records in each day.
h) Hour & minute values of each recording is fetched from 'time_1' variable and placed in 'time_hours' & 'time_minutes' variable respectively.

2.1 Feature Details:

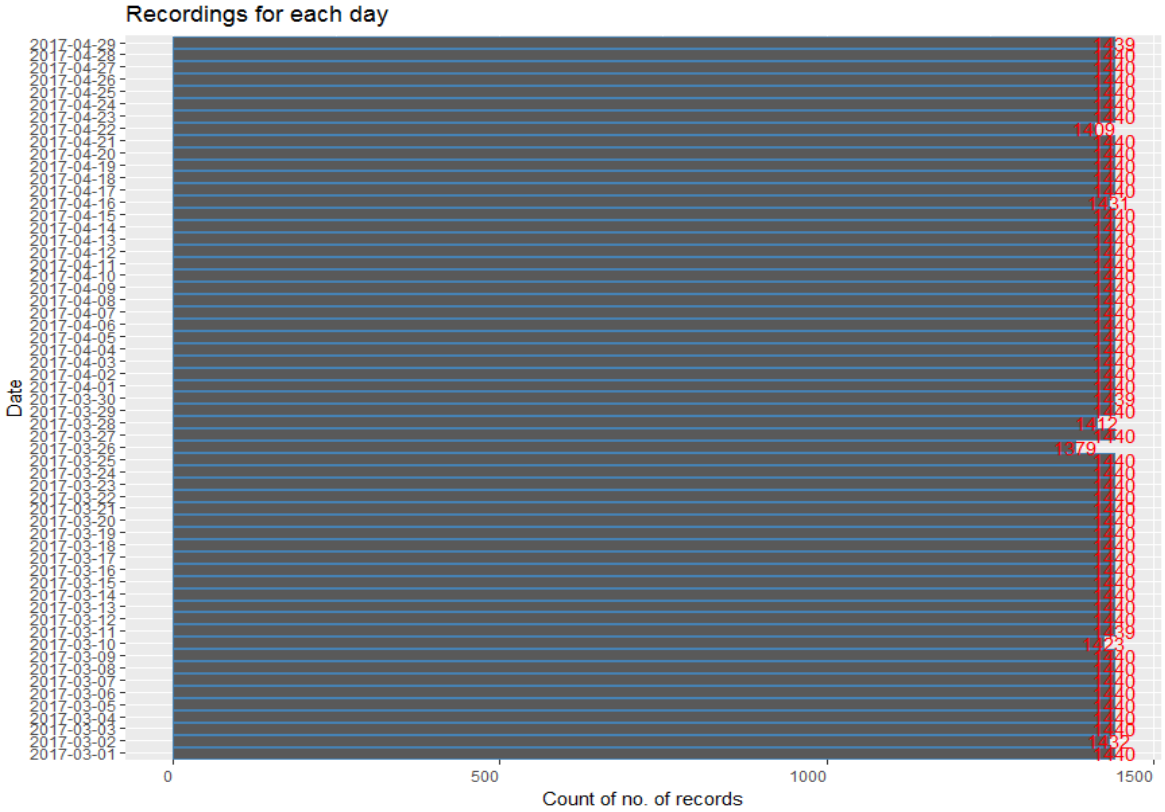
For any time-series dataset, creating new features with given variables is imperative to a certain extent and helpful in analysis and understanding. In this case, the below features are created:

Feature Name	How is it generated	Feature Use
time_2	Data_1_2 = Data_1_2 %>% mutate(time_2 = c(Data_1_2$time_1[-1], NA)) %>% na.omit()	'time_2' feature has been created to get the time difference between two consecutive recordings.
seconds **minutes** **dhours** **days**	inter_Data_1_2 <- interval(Data_1_2$time_1,Data_1_2$time_2) Data_1_2$seconds <- inter_Data_1_2/dseconds() Data_1_2$minutes <- inter_Data_1_2/dminutes() Data_1_2$dhours <- inter_Data_1_2/dhours() Data_1_2$days <- inter_Data_1_2/ddays()	'seconds', 'minutes', 'dhours' & days' gives interval between two consecutive recordings in number of seconds, minutes, hourss & days respectively.
only_date	Data_1_2$only_date = as.Date(format(Data_1_2$time_1, format="%Y-%m-%d"))	To get only the date vale from 'time_1' column
weekday	Data_1_2$weekday <- weekdays(Data_1_2$only_date)	To get day of week for the particular date
time_hours	Data_1_2$time_hours <- substr(Data_1_2$time_1, 12, 13) Data_1_2$time_hours <- as.integer(Data_1_2$time_hours)	To get hour of recording from 'time_1'
time_minutes	Data_1_2$time_minutes <- substr(Data_1_2$time_1, 15, 16) Data_1_2$time_minutes <- as.integer(Data_1_2$time_minutes)	To get minute of recording from 'time_1'

2.2 Data Imputation:

2.2.1 Sensor disconnected from Network (Missing values):

- Data is recorded for every minute in a day from March 1st to March 30th and from April 1st to April 29th.
- Total record count for each day should be 1440 (24hrs x 60mins). Below plot shows the number of observations recorded on each day:



- As per the above plot, below are the days that are having recordings less than 1440:

Date	No. of recordings
02-03-2017	1432
10-03-2017	1423
11-03-2017	1439
26-03-2017	1379
28-03-2017	1412
30-03-2017	1439
16-04-2017	1431
22-04-2017	1409
29-04-2017	1439

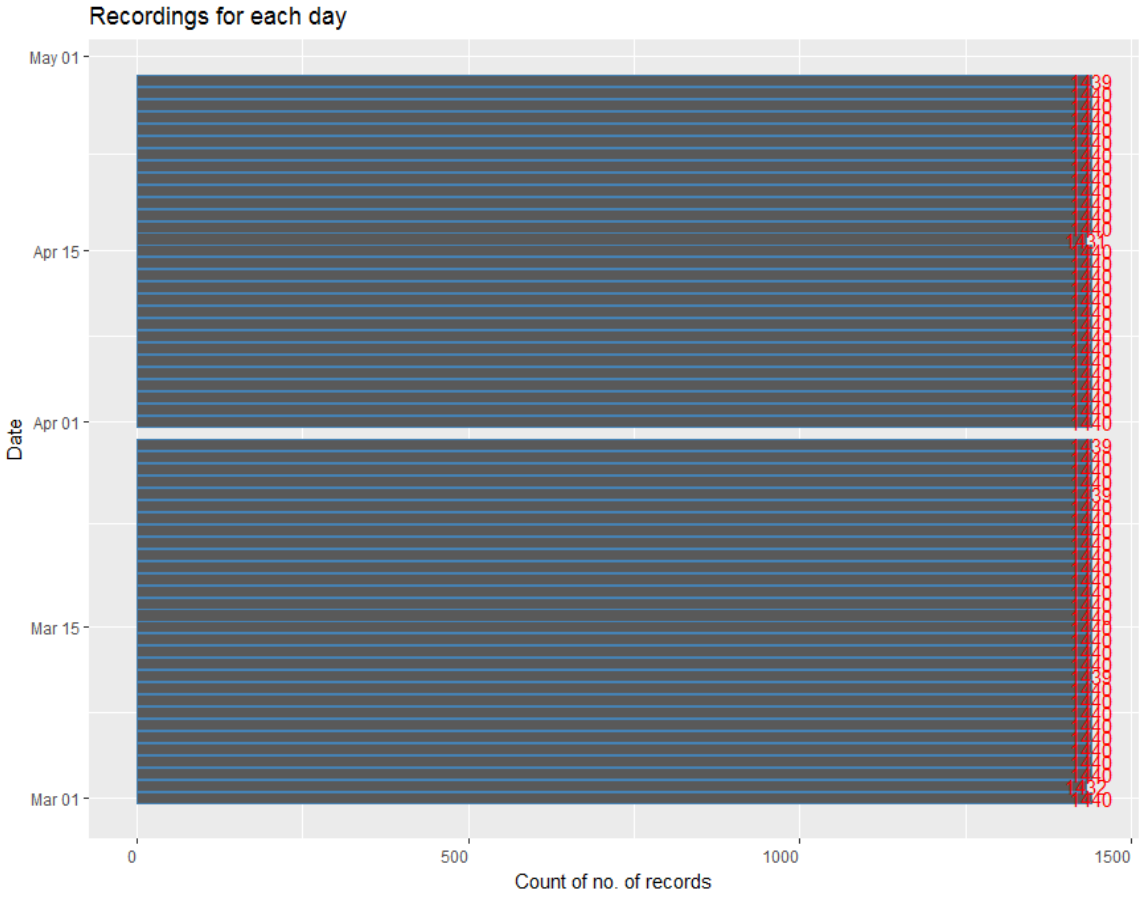
- Recordings of environmental parameters are made for every minute, and hence it's assumed that one or two missing recordings in an hour or so will not show major deviations in the environmental parameters.
- In any given day in the data file, if no records are found continuously for more than 15 minutes then it can be assumed that there was a power cut and sensor was disconnected from the network and so the sensor could not record data during this time.
- Considering the threshold of missing values as 15, from the time interval in 'minutes' variable, records with value greater than 15 are obtained and imputation is performed as:
  - 'time_1' value is fetched from the current observation with value greater than 15 in 'minutes' variable.
  - 24hours is subtracted from the 'time_1' of current observation to get the previous day date. Value of 'minutes' variable of current observation is added to the obtained previous day date to get the interval from the previous day.

- 24hours is added to the 'time_1' of current observation to get the next day date. Value of 'minutes' variable of current observation is added to the obtained next day date to get the interval from the next day.
- Mean values of all attributes of each record is calculated from the previous day and next day to get the similar values for the imputed records in current day.

- 139 new observations are imputed as per the above process and are added to the data frame (Data_merge) as mentioned in third step of data preparation.
- All the data preparation steps are repeated to get the same features for the newly imputed observations.
- Number of observations of each day is again calculated as it is found that, for two days recordings are more than 1440:

Date	No. of recordings
02-03-2017	1432
10-03-2017	1442
11-03-2017	1439
26-03-2017	1439
30-03-2017	1439
16-04-2017	1431
22-04-2017	1441
29-04-2017	1439
02-03-2017	1432

- This is due to imputing new record when there is a recording already for that hour & minute, in such cases imputed recordings are deleted by comparing the hours and minutes variables and final count is as below.



2.2.2 Network issue causing Duplicate entries:

- As values of environmental parameters do not change much every minute, it is assumed that if the same point for each variable is repeated consecutively beyond the threshold number then there is an issue with the network, thus imputation is performed as below:
  - Consider duplicate imputation for temperature, threshold is assumed as 15 for the variable. If the same temperature point is repeated more than 15 times consecutively then the number of repetitions is taken as count.
  - Assume count (no. of times a same point is repeated) of a temperature point is 51, the quotient of 'count/2' is calculated which is 25. In the set of 51 consecutive records with same temperature point. 25th record will be taken as base, and then the value of recordings from 24th to 1st will be decremented with 0.01*1, 0.01* 2, 0.01*3 and so on till 0.01*24. In the same way, value of recording from 26th to 51st will be incremented with 0.01*1, 0.01* 2, 0.01*3 and so on till 0.01*25.
  - Same process was applied for every variable with different threshold & Increment/Decrement levels.
  - Below is the list of threshold and details assumed for each variable.

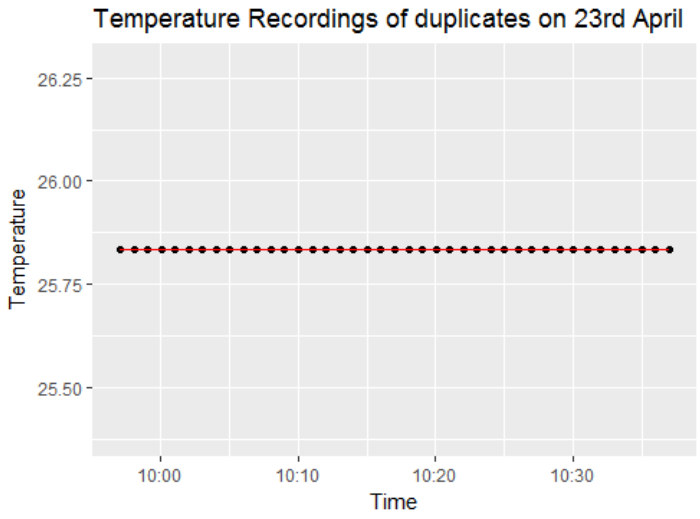
Variable	Threshold	Increment/ Decrement level	Imputation	Comments
Temperature	15	0.01	Done	Few temperature points were repeated consecutively beyond threshold
Noise	3	0.1	Not needed	No Noise point repetitions were there consecutively beyond threshold
Light	5	0.5	Not needed	No Light point repetitions were there consecutively beyond threshold
Co2	5	0.25	Not needed	No Co2 point repetitions were there consecutively beyond threshold
VOC	5	0.25	Not needed	No VOC point repetitions were there consecutively beyond threshold
Humidity	7	0.1	Done	Few Humidity points were repeated consecutively beyond threshold

Few duplicate recordings in temperature are as follows:

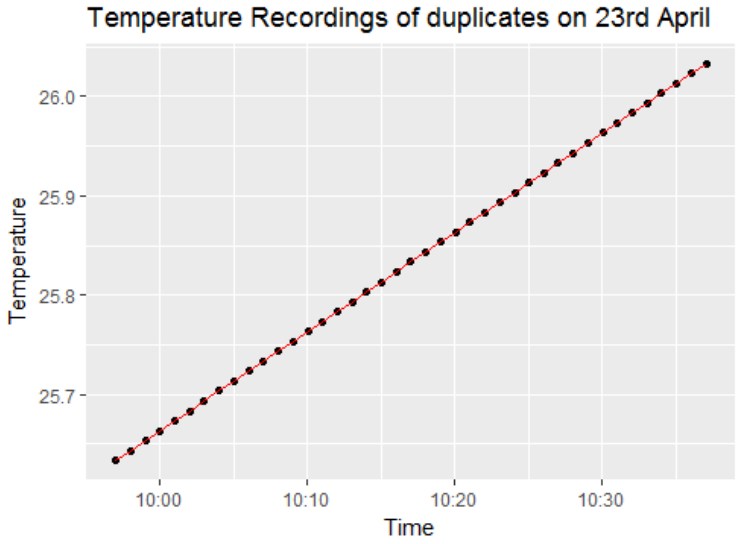
	time_1	unitid	Temperature	Noise	Light	Co2	VOC	Humidity	Count
75502	2017-04-23 10:37:02	SS0031	25.83333	51.17861	6.7222	445.2778	326.6667	66.40000	41
82576	2017-04-28 08:31:09	SS0029	24.80000	50.94287	6.4444	453.1111	332.3333	62.13889	40
69808	2017-04-19 11:44:03	SS0031	24.70000	51.08107	6.5000	449.7778	329.6667	61.98889	39
66562	2017-04-17 05:38:01	SS0031	25.73333	50.99828	5.2778	430.2222	315.4444	68.10556	38
6912	2017-03-05 19:19:03	SS0036	25.07500	52.05639	21.0000	428.9167	314.5417	65.41667	36
82534	2017-04-28 07:49:09	SS0029	24.80000	50.97556	6.9444	454.6111	333.1667	62.12222	34
84858	2017-04-29 22:33:08	SS0029	25.43333	50.92930	15.2778	436.0556	319.6667	66.39444	34
73936	2017-04-22 08:32:03	SS0031	24.90000	50.95931	6.1111	437.0000	320.3333	63.23889	33
41216	2017-03-29 15:03:01	SS0029	24.85000	52.02645	6.9583	439.3333	322.3333	64.92083	31
74233	2017-04-22 13:28:02	SS0031	25.10000	50.99514	8.3333	437.8889	320.7222	64.28333	29
84889	2017-04-29 23:04:08	SS0029	25.43333	50.94293	14.8889	436.2222	319.6667	66.42778	29

Value of temperature 25.83333 is repeated 41 times consecutively from '2017-04-23 09:56:09'to '2017-04-23 10:37:02'.

Before imputation, plot is shown below:



After imputation, below is the plot. Values between the two time-interval ranges from 25.63333 to 26.03333.



2.2.3 Abnormal values in the Sensor recordings:

- Abnormal values are identified by considering three consecutive records as group. It is assumed that a threshold value for each variable and imputation is performed as below:
  - Consider 'Noise' variable, its assumed that threshold of Noise as 2. Now, three consecutive recordings of noise are i^{th} row, $i+1^{st}$ row and $i+2^{nd}$ row.
  - If absolute difference between i & i+1 is greater than 2 (threshold) and absolute difference between i+1 & i+2 is greater than 2 (threshold) and absolute difference between i & i+2 is less than 2 (threshold), then i+1 recording of Noise is considered as abnormal.
  - This abnormal i+1 recording is imputed with mean value of i and i+2 recording for a variable.
  - Once this iteration is done then i+1, i+2 & i+3 recordings of variable will be considered for next iteration.

Below is list of variables and their assumed threshold value.

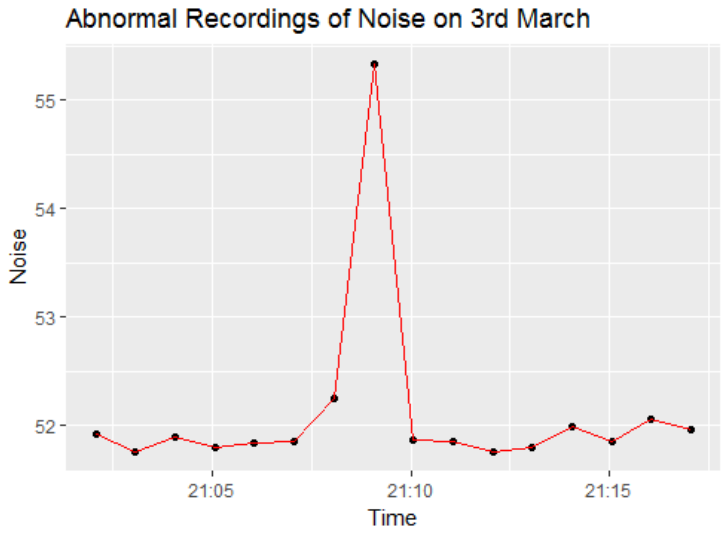
Variable	Threshold	Imputation
Temperature	1	Done
Noise	2	Done
Light	250	Done
Co2	5	Done
VOC	5	Done
Humidity	2	Done

Few abnormal values obtained in Noise are as follows:

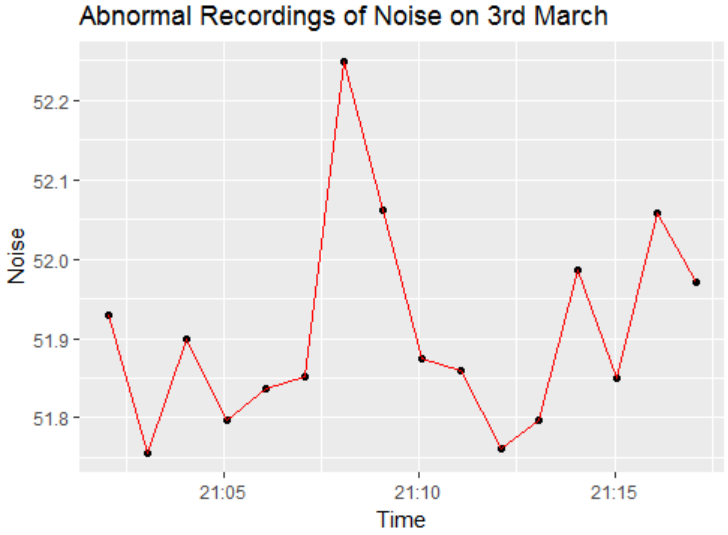
	time_1	unitid	Temperature	Noise	Light	Co2	VOC	Humidity
1776	2017-03-02 05:37:01	SS0036	23.10000	54.59880	11.0000	430.0000	315.0000	70.40000
5582	2017-03-04 21:09:04	SS0036	24.67917	55.33893	8.2083	437.5833	320.9583	65.56667
8311	2017-03-06 18:38:02	SS0036	24.46250	55.39947	438.0417	428.0000	313.7500	60.59167
8315	2017-03-06 18:42:02	SS0036	24.47500	56.30286	427.2917	429.4167	314.8750	60.59583

From the above, it is inferred that 5582 recording at '2017-03-24 21:09:04' is considered as abnormal when compared to before & next recordings.

Below is the plot of graph **before imputing the abnormality**. Recording of noise is 55.33893.



Below is the plot of graph **after imputing the abnormality**. The values of Noise '2017-03-24 21:09:04' is 52.06168.



3. Findings and Inferences from the prepared data

For better analysis/understanding of the data, an assumption is made and based on the assumption, findings have been presented.

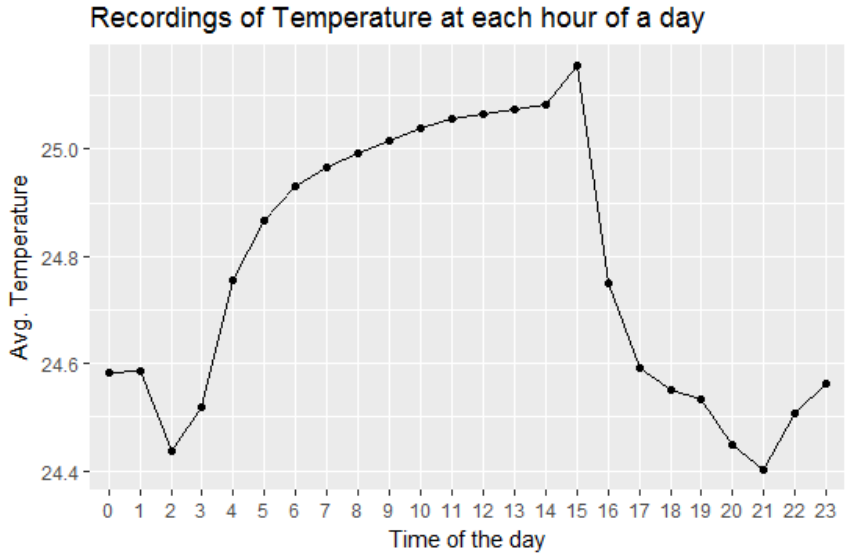
Assumption:

The IOT sensors are assumed to be present in an air-conditioned Café that is located near the seashore. The sensors are placed at different locations inside the Café.

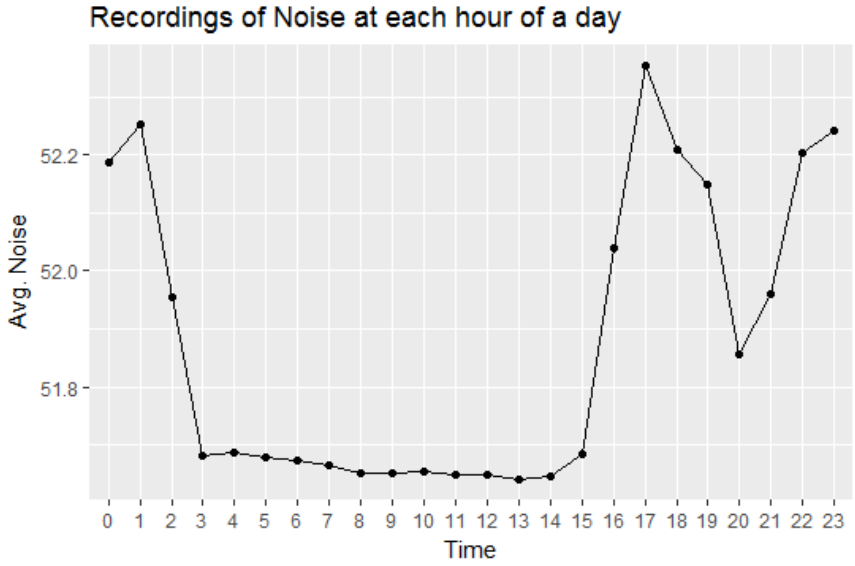
Based on this assumption, the below graphs have been plotted as such:

3.1 Comparing Average level of variable at each hour of the day:

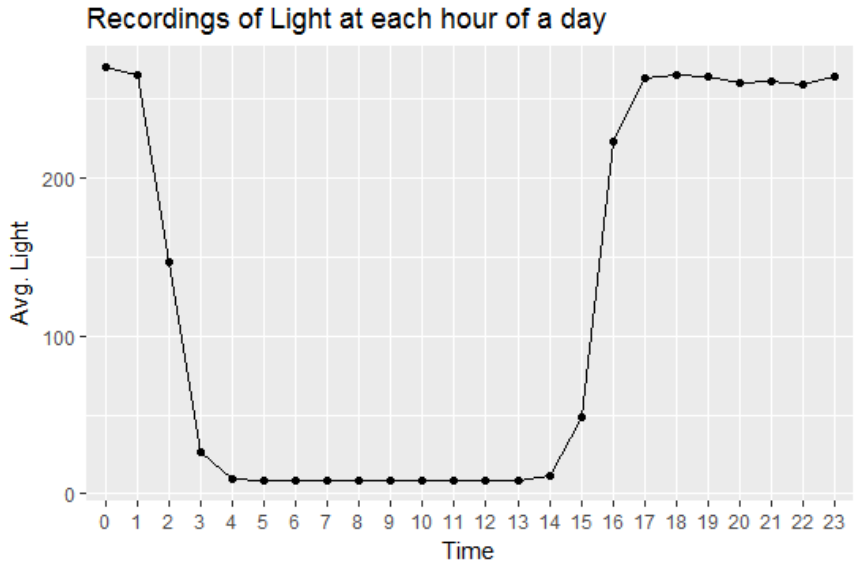
a) Temperature:



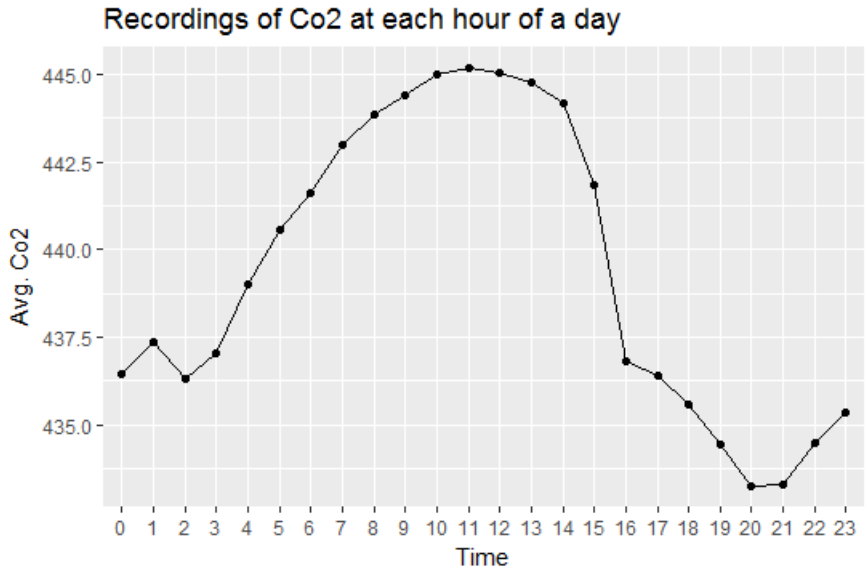
b) Noise:



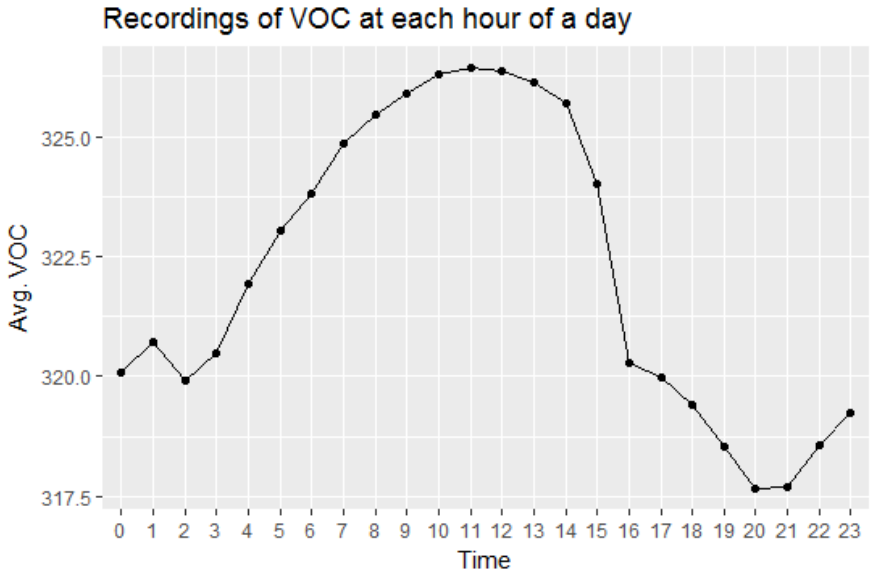
c) Light:



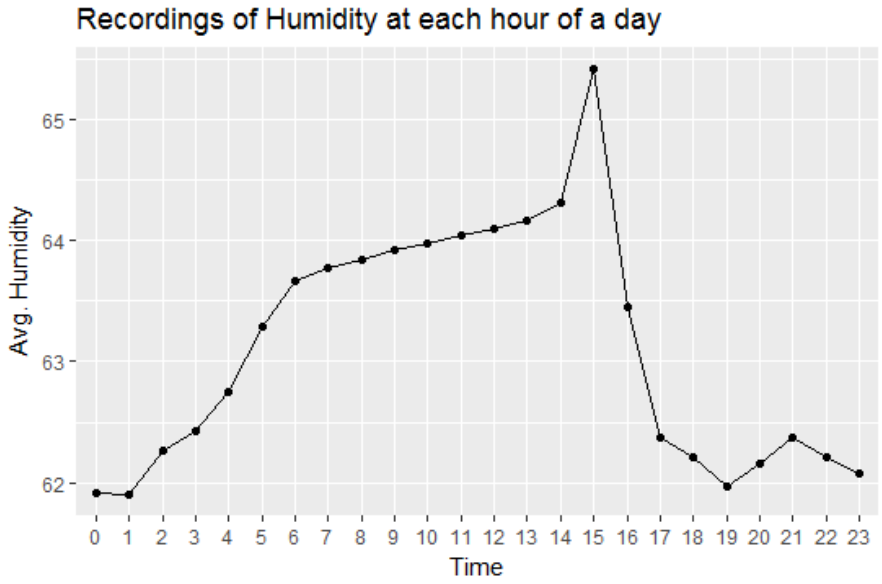
d) Co2:



e) VOC:



f) Humidity:



Findings from above plots:

1) Temperature & Humidity start to increase from 2AM and are in high level till 4PM. There is a sudden spike in the level at 3PM.
2) Level of Noise starts to decrease from 2AM and is very less till 3PM and starts to rise from 4PM. Noise level starts to increase again from 8PM.
3) Level of Light starts to decrease from 2AM and is very less till 3PM. It increases from 4PM and the level is constant for rest of the hours.
4) Level of Co2 & VOC follow the same pattern; their level starts to increase from 2AM and is very high till 2PM and starts to decrease from 3PM.

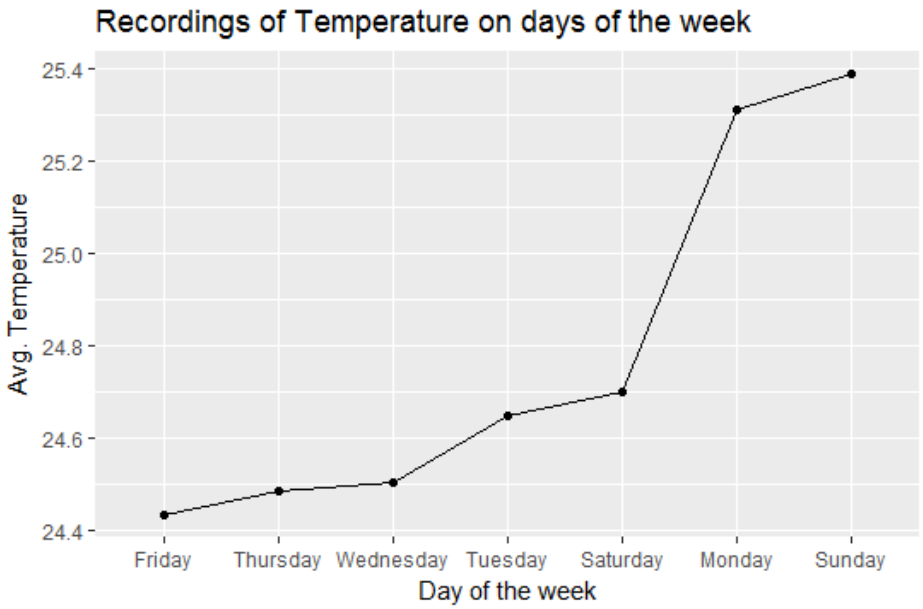
Inferences from above findings:

- From the above findings, its is assumed that the Café opens in the evening around 3PM to 4PM and runs till around 1AM.
- As the Café remains closed from 2AM to 3PM with the doors and most of the windows shut, there would be no free ventilation. Air conditioning would also be switched off and hence the level of Temperature & Humidity remains high. As the Café opens around 3PM, staff would come to the Café and open all doors and windows for ventilation. Since the Café is near the seashore/beach, hot breeze from the nearby sea flowing through the Café at 3PM in the afternoon would be a reason for causing a sudden rise. After some time, as the Café starts operations, Air conditioning in the room brings the temperature/humidity down.
- Level of light is also very less from 2AM to 3PM indicating that Café is closed in that time.
- High level of Co2 indicates that the Café is closed and level of ventilation required is high. High level of VOC indicates that organic compounds are emitted from household products present in the closed Café around 2AM to 3PM. Sudden decrease in Co2 & VOC level at 3PM indicates that openings of doors & windows reduced the level of ventilation needed.
- An increase in the Noise level from 8PM would suggest that the Café starts playing music as a sort of entertainment. It may also suggest that TV's in the Café are turned on as most sport events are telecasted in night.

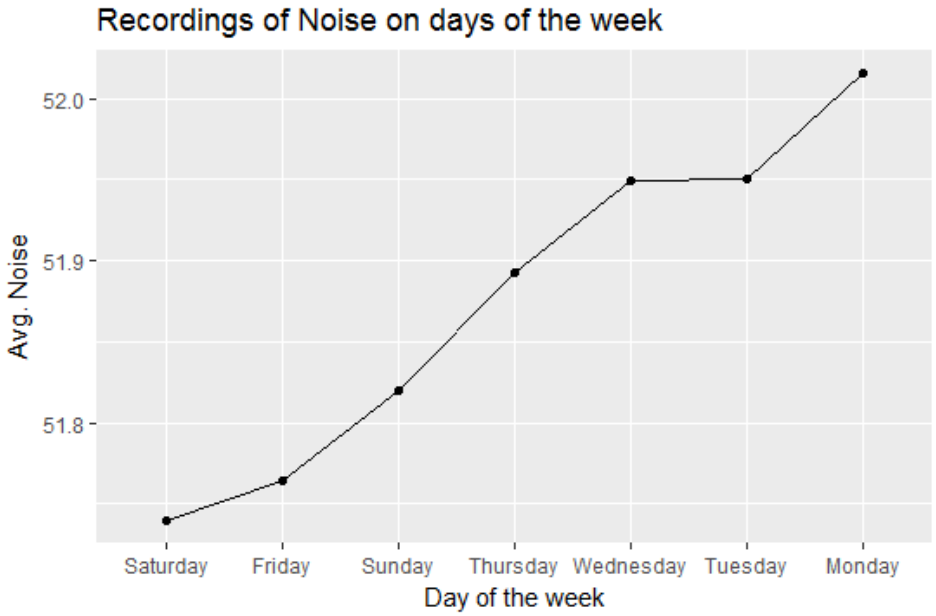
- Very slight increase in CO2 and VOC after 8PM may be attributed to the reason that the Café allows smoking in its lounges.

3.2 Comparing Average level of attributes on each day:

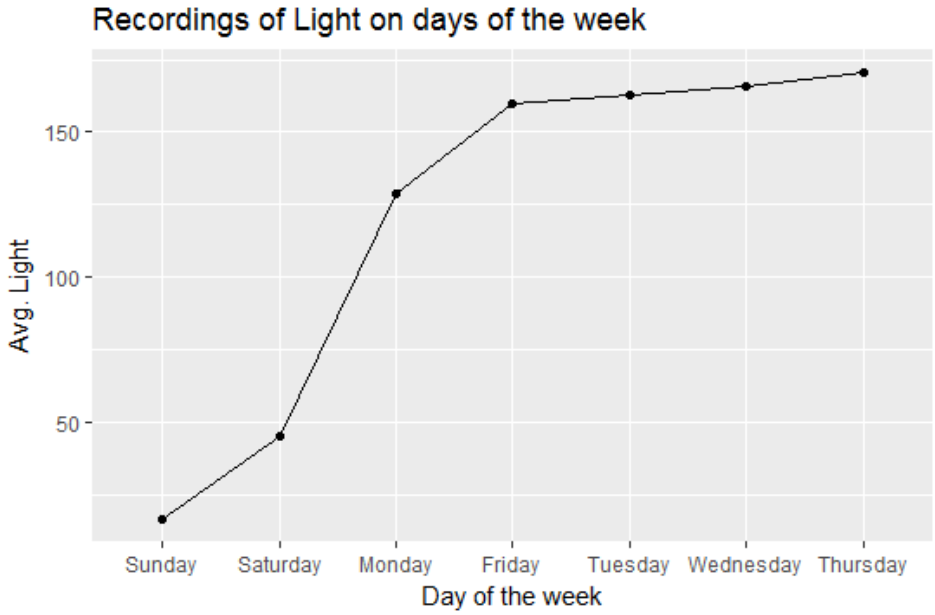
a) Temperature:



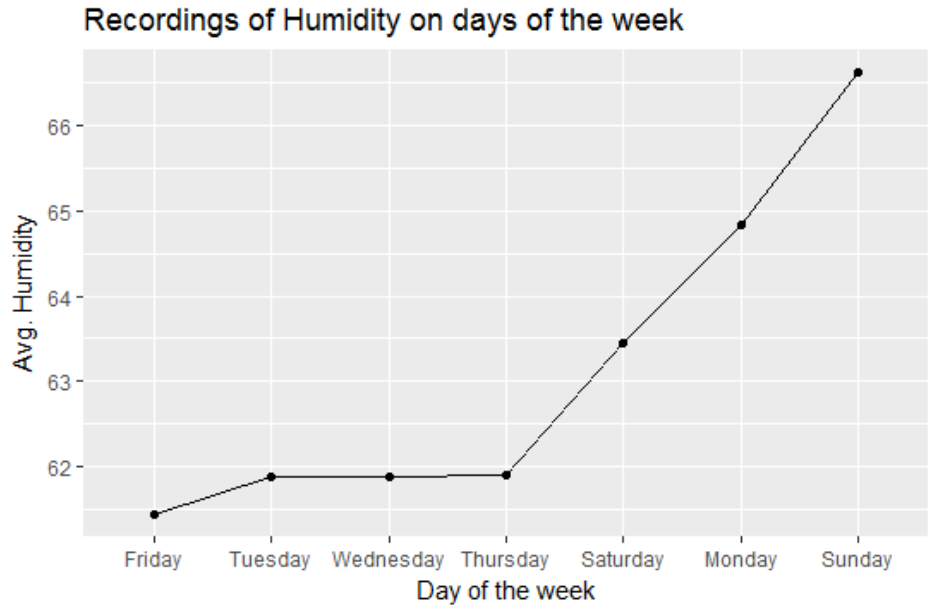
b) Noise:



c) Light:



d) Humidity:



Findings from the above plots:

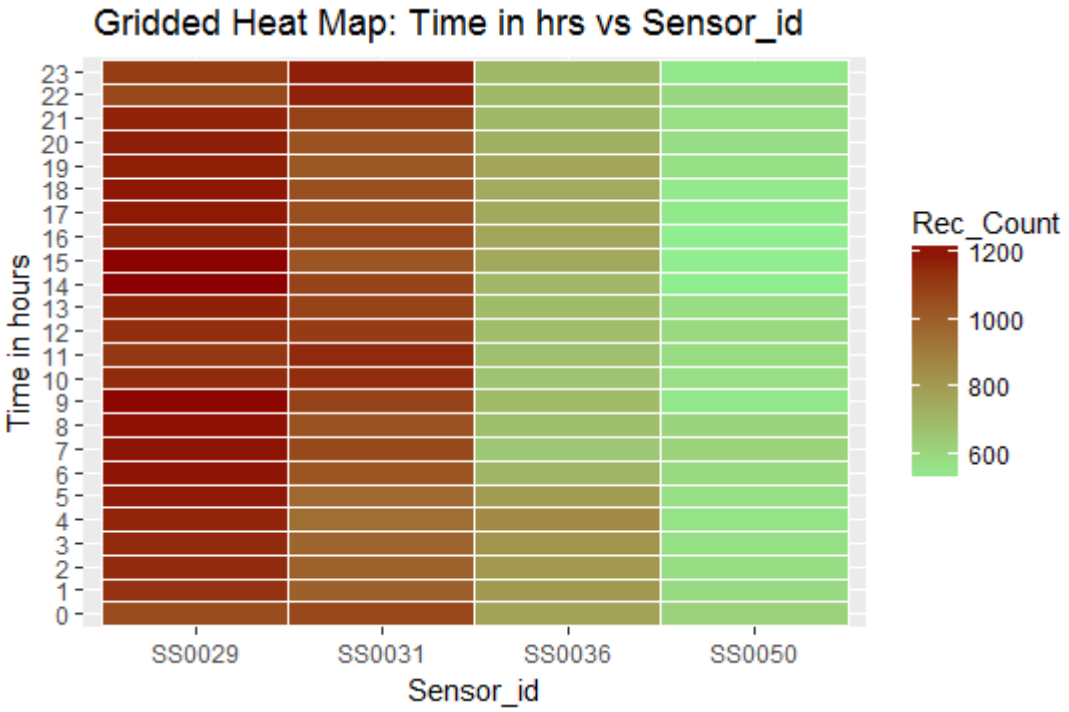
5. Temperature and Humidity is high on Sundays, moderately high on Mondays and lower when compared to other days on Saturday.
6. Light is the least on Sunday, slightly higher on Saturday and moderately high on Monday.
7. Noise is the least on Saturday, moderately high on Sunday and highest on Monday

Inferences from above findings:

- From the above findings, it can be inferred that the Café is closed on Sunday and as the Temperature is high it can be assumed that the Air Conditioner is turned off. Noise is moderately low because there are no people but only the machines inside the Café that are running. Lowest level of light indicates that the Café is closed.
- Saturday is a special day in the Café; the low levels of light and noise indicate that the ambience of the Café is maintained with dimmed lights and silent serving. Only certain people with prior booking would be allowed to enjoy the special ambience and food.
- Light and Noise level on Tuesday and Wednesday suggest that Café activity is highest on these days.
- Thorough cleaning takes place in the Café on Monday. The temperature is higher indicating that the cleaning takes place without switching on the AC. The amount of noise produced from the cleaning equipment can be seen by the high level of noise on Monday.

3.3 Comparing hourly sensor recordings to identify most and least active sensor

a) Heat map of the plot

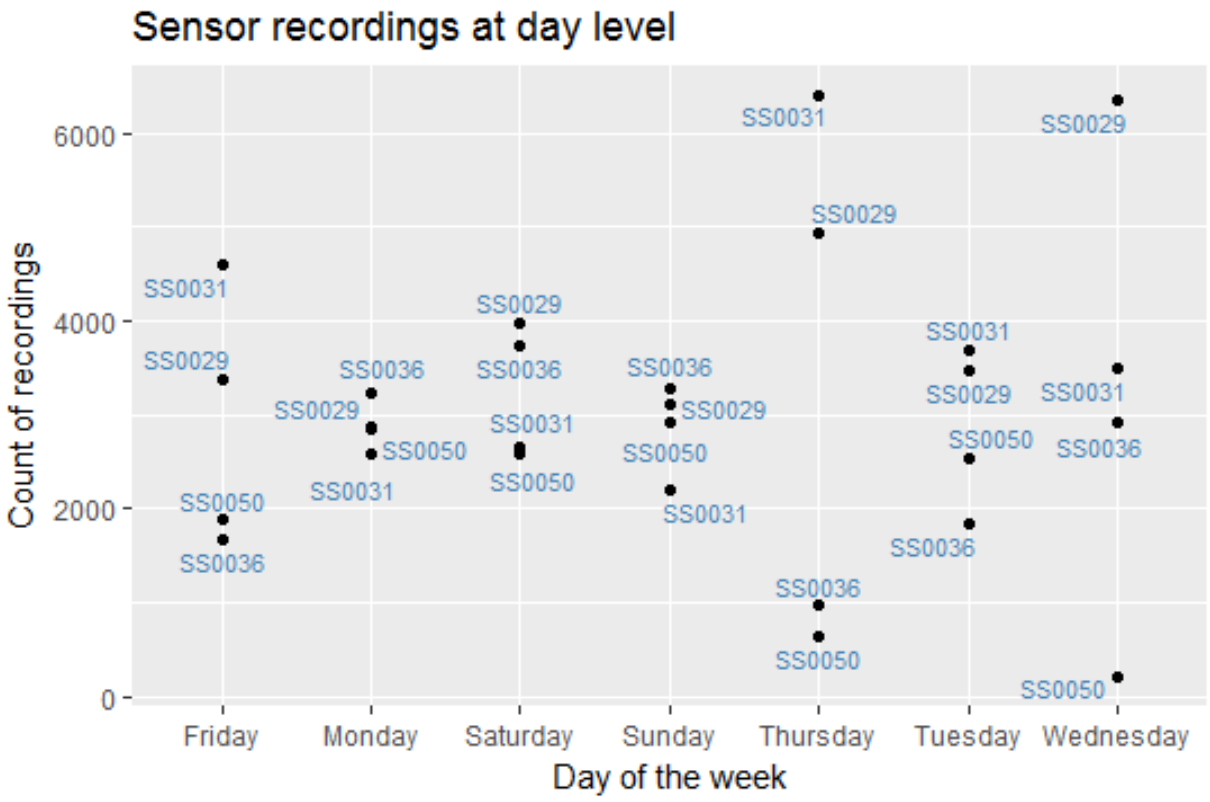


Findings from above plot:

8) Most active at every hour of day is SS0029, indicating that sensor is placed at a busy place of the Café, for example, the entrance or kitchen.
9) Least active sensor is SS0050, indicating that it is placed in some corner of the room where there is minimal activity.

3.4 Comparing sensor recordings:

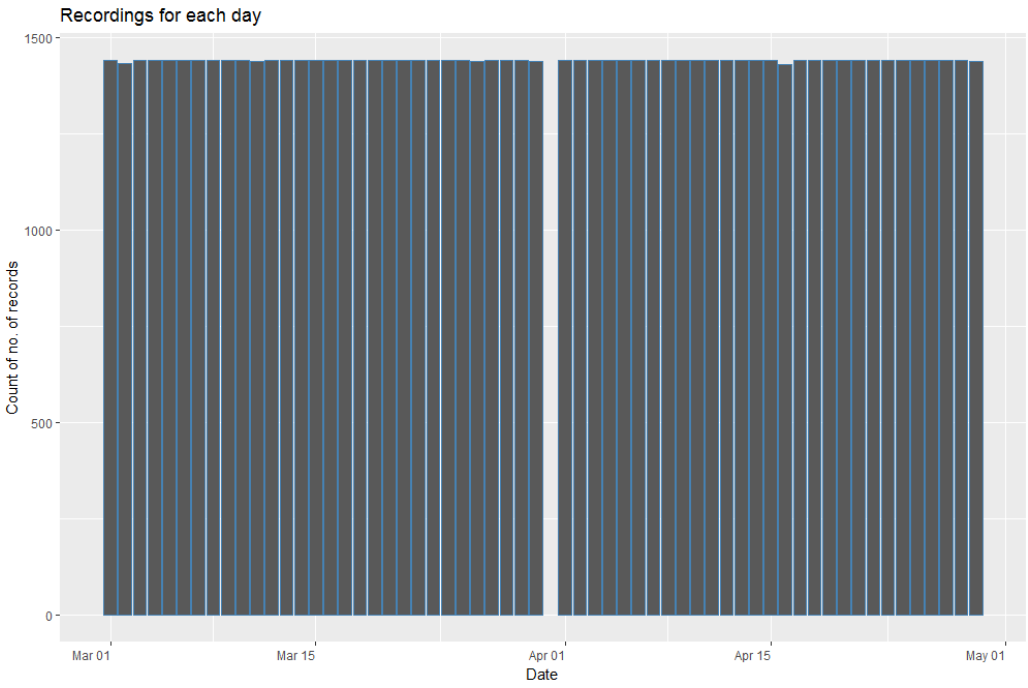
a) Sensor recordings at day level:



Findings from above plot:

10) Sensor SS0050 records lesser number of recordings almost every day when compared to other sensors, this supports the claim made in #9. Also, on Wednesday, the sensor barely does any recordings indicating it could be faulty one which couldn't take much load on every day of the week.

b) Sensor recordings at month end:



Findings from above plot:

11) There are no recordings on March 31th and April 30th. From this, it can be assumed that, on every day of month sensors will be taken for servicing and hence recordings were not made.

4. Conclusion:

The inconsistencies and issues in the dataset have been eliminated through analysis and imputation. The prepared data was used to obtain findings and thus inferences were drawn.