

DẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC BÁCH KHOA  
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH



## CẤU TRÚC RỜI RẠC CHO KHMT (CO1007)

Thống kê khảo sát kết quả  
ứng dụng IoT trong nông nghiệp môn Cấu trúc rời rạc

GVHD: Huỳnh Tường Nguyên

Nguyễn Ngọc Lẽ

SV thực hiện: Nguyễn Xuân Nam – 2012516

Nguyễn Minh Phát – 2011795

Thới Duy Phát – 2120049



## Mục lục

1 Động cơ nghiên cứu	2
2 Mục tiêu	2
3 Mô tả dữ liệu	2
4 Nhiệm vụ (Mã đề của nhóm là 4263)	2



## 1 Động cơ nghiên cứu

The Internet of things (IoT) mô tả các thiết bị (cảm biến - sensors) kết nối đến internet để thu thập dữ liệu và trao đổi dữ liệu. IoT sẽ cách mạng hóa cách chúng ta thực hiện mọi thứ từ giao thông vận tải đến truyền thông. Ngành nông nghiệp cũng được hưởng lợi từ việc ứng dụng khả năng công nghệ vào đó. Các cảm biến sẽ hỗ trợ người nông dân quản lý tốt hơn cây trồng, nguồn nước, nhiệt độ, độ ẩm,...cùng với dự báo thời tiết sẽ làm gia tăng năng suất lương thực và tiết kiệm chi phí.

Dữ liệu có được từ các cảm biến có thể xử lý trước với một vài thống kê cơ bản trước khi nó được truyền đi để khai thác dữ liệu thông minh sâu hơn.

## 2 Mục tiêu

Trong bài tập lớn này, các sinh viên sẽ bắt đầu với các bài toán thống kê đơn giản từ những dữ liệu được cung cấp. Qua đó, các em sẽ tìm ra những con số thú vị, có ý nghĩa đối với các dữ liệu thực tế từ ứng dụng IoT trong nông nghiệp. Những kết quả mà các em tìm ra sẽ là bước khởi đầu cho việc khai phá nguồn dữ liệu của hệ thống sau này, nhằm đạt tới mục tiêu nâng cao kỹ năng lập trình, kỹ năng giải quyết vấn đề cho người học, kỹ năng làm việc nhóm cũng như hướng tới mục tiêu cao hơn là đam mê trong làm việc, học tập và nghiên cứu.

## 3 Mô tả dữ liệu

Dữ liệu được đo từ các thiết bị cảm biến gồm các thuộc tính chính “**date**, **deviceid**, **devicetype**, **data**, **Temp**, **Humi**” được lưu trong file **csv**. Số thiết bị có trong mỗi csv file có thể khác nhau.

1. **date**: chứa thông tin thời gian mà thiết bị thực hiện đọc dữ liệu môi trường với định dạng Year-Month-Day: Hour:Minute:Second.Microsecond (2021-03-01 11:56:40.629095+00)
2. **deviceid**: định danh cho thiết bị ("SS0825910591458270")
3. **devicetype** cung cấp thông tin về công dụng của thiết bị ("TempAndHumi": đo độ nhiệt độ và độ ẩm đất)
4. **data**: danh sách các cặp khóa, giá trị của nhiệt độ và độ ẩm "Humi": 59.8, "Temp": 30
5. **Temp**: chứa giá trị là nhiệt độ đất
6. **Humi**: chứa giá trị là độ ẩm đất

## 4 Nhiệm vụ (Mã đề của nhóm là 4263)

Nhóm sẽ thao tác với các file số liệu:

[3-01\\_2021.csv](#)  
[4-03\\_2021.csv](#)  
[5-04\\_2021.csv](#)

Nhóm sẽ thực hiện bài tập các phần  $\{i, \dots, x_{iii}\}$  và trả lời các câu hỏi 2, 3, 4, 6 trong phần [ix](#)



i) Nhóm câu hỏi liên quan đến tổng quát dữ liệu

Dùng tập dữ liệu để trả lời các câu hỏi và trình bày theo định dạng.  
Đọc dữ liệu file bằng *read*

```
data <- read.csv("3-01_2021.csv")
data <- read.csv("4-03_2021.csv")
data <- read.csv("5-04_2021.csv")
```

1) Số lượng thiết bị cảm biến và định danh của mỗi thiết bị.

Sử dụng hàm *table* cho cột *deviceid*

```
table(data$deviceid)
```

Ví dụ kết quả thu được:

```
> table(data$deviceid)
"SS0866974771458415" "SS0866977441458415"
 10150                 3110
```

Bảng 1: Số lượng thiết bị cảm biến và định danh của mỗi thiết bị

	Số lượng	Định danh	thiết bị
File 3	2	SS0866974771458415	dv1
		SS0866977441458415	dv2
File 4	3	SS0866974771458415	dv1
		SS0866977441458415	dv2
		SS0466973631458415	dv3
File 5	3	SS0866974771458415	dv1
		SS0866977441458415	dv2
		SS0466973631458415	dv3

2) Số lượng giá trị cảm biến thu thập trên từng thiết bị cảm biến

Nhóm các dữ liệu theo deviceid và coi loại giá trị cảm biến thu thập dựa trên cột data.

```
data_d1 = data[data$deviceid == "\"SS0866974771458415\"",]
data_d2 = data[data$deviceid == "\"SS0866977441458415\"",]
data_d3 = data[data$deviceid == "\"SS0466973631458415\"",]
data_d1$data
data_d2$data
data_d3$data
```

Bảng 2: Số lượng giá trị cảm biến thu thập trên từng thiết bị cảm biến

Sensors	Số lượng	Giá trị
SS0866974771458415	2	Temp Humi
SS0866977441458415	2	Temp Humi
SS0466973631458415	3	Temp pH EC



- 3) Số lượng dữ liệu được trong từng thiết bị và tổng số.

Phân loại dữ liệu thu được dựa trên *deviceid*

```
data_d1 = data[data$deviceid == "\"SS0866974771458415\"",]  
data_d2 = data[data$deviceid == "\"SS0866977441458415\"",]  
data_d3 = data[data$deviceid == "\"SS0466973631458415\"",]
```

Bảng 3: Số lượng dữ liệu được trong từng thiết bị và tổng số

	Devices	Số lượng	Tổng
File 3	dv1	10150	13260
	dv2	3110	
	dv3		
File 4	dv1	13256	19821
	dv2	6178	
	dv3	387	
File 5	dv1	7390	11266
	dv2	3129	
	dv3	747	
			44347

- 4) Cho biết tần số của các ngày trong tập dữ liệu?

Tạo một cột mới tên date\_new chỉ chứa 10 kí tự đầu trong cột date và dùng hàm table cho cột date\_new

```
data_d = data  
date_new = substr(data_d$date, 1, 10)  
date_new = as.Date(date_new, format = "%Y-%m-%d")  
data_d = cbind(data_d, date_new)  
table(date_new)
```

Kết quả thu được:

*File3* (3-01\_2021.csv):

2021-01-04	2021-01-05	2021-01-06	2021-01-07	2021-01-08	2021-01-09	2021-01-10	2021-01-11
158	601	459	476	304	286	534	641
2021-01-12	2021-01-13	2021-01-14	2021-01-15	2021-01-16	2021-01-17	2021-01-18	2021-01-19
633	377	473	473	599	755	686	619
2021-01-20	2021-01-21	2021-01-22	2021-01-23	2021-01-24	2021-01-25	2021-01-26	2021-01-27
621	591	601	471	382	390	423	475
2021-01-28	2021-01-29	2021-01-30	2021-01-31				
444	267	218	303				

*File4* (4-03\_2021.csv):

2021-03-01	2021-03-02	2021-03-03	2021-03-04	2021-03-05	2021-03-06	2021-03-07	2021-03-08
670	900	858	437	476	475	357	451
2021-03-09	2021-03-10	2021-03-11	2021-03-12	2021-03-13	2021-03-14	2021-03-16	2021-03-17
473	686	908	904	918	647	379	651
2021-03-18	2021-03-19	2021-03-20	2021-03-21	2021-03-22	2021-03-23	2021-03-24	2021-03-25
478	715	1102	847	874	925	745	688
2021-03-26	2021-03-27	2021-03-28	2021-03-29	2021-03-30	2021-03-31		
584	540	518	585	552	478		



**File5** (5-04\_2021.csv):

2021-04-01	2021-04-02	2021-04-03	2021-04-04	2021-04-05	2021-04-06	2021-04-07	2021-04-08
573	594	334	352	485	68	143	144
2021-04-09	2021-04-10	2021-04-11	2021-04-12	2021-04-13	2021-04-14	2021-04-15	2021-04-16
169	234	917	948	954	947	335	204
2021-04-17	2021-04-18	2021-04-19	2021-04-20	2021-04-21	2021-04-22	2021-04-23	
821	552	235	479	720	840	218	

- 5) Từ ngày 1 đến 31, hãy cho biết tần số thu thập dữ liệu trên từng thiết bị cảm biến.

Phân loại dữ liệu dựa trên deviceid và thực hiện tương tự câu 4.

```
data <- read.csv("3-01_2021.csv")
data <- read.csv("4-03_2021.csv")
data <- read.csv("5-04_2021.csv")
data_d = data
data_d = data[data$deviceid == "\"SS0866974771458415\"", ]#1
data_d = data[data$deviceid == "\"SS0866977441458415\"", ]#2
data_d = data[data$deviceid == "\"SS0466973631458415\"", ]#3
date_new = substr(data_d$date,1, 10)
date_new = as.Date(date_new,format = " %Y-%m-%d")
table(date_new)
```

**File3** (3-01\_2021.csv):

**dv1** (SS0866974771458415):

2021-01-04	2021-01-05	2021-01-06	2021-01-07	2021-01-08	2021-01-09	2021-01-10
131	474	408	476	182	174	253
2021-01-11	2021-01-12	2021-01-13	2021-01-14	2021-01-15	2021-01-16	2021-01-17
355	393	377	473	473	463	432
2021-01-18	2021-01-19	2021-01-20	2021-01-21	2021-01-22	2021-01-23	2021-01-24
469	426	363	412	434	463	373
2021-01-25	2021-01-26	2021-01-27	2021-01-28	2021-01-29	2021-01-30	2021-01-31
339	342	421	350	267	204	223

**dv2** (SS0866977441458415):

2021-01-04	2021-01-05	2021-01-06	2021-01-08	2021-01-09	2021-01-10	2021-01-11
27	127	51	122	112	281	286
2021-01-12	2021-01-16	2021-01-17	2021-01-18	2021-01-19	2021-01-20	2021-01-21
240	136	323	217	193	258	179
2021-01-22	2021-01-23	2021-01-24	2021-01-25	2021-01-26	2021-01-27	2021-01-28
167	8	9	51	81	54	94
2021-01-30	2021-01-31					
14	80					

**File4** (4-03\_2021.csv):

**dv1** (SS0866974771458415):

2021-03-01	2021-03-02	2021-03-03	2021-03-04	2021-03-05	2021-03-06	2021-03-07
371	464	435	437	476	475	357
2021-03-08	2021-03-09	2021-03-10	2021-03-11	2021-03-12	2021-03-13	2021-03-14
451	473	472	477	477	477	420
2021-03-16	2021-03-17	2021-03-18	2021-03-19	2021-03-20	2021-03-21	2021-03-22
307	478	341	334	476	453	476
2021-03-23	2021-03-24	2021-03-25	2021-03-26	2021-03-27	2021-03-28	2021-03-29
473	437	469	463	425	428	479
2021-03-30	2021-03-31					
477	478					



*dv2* (SS0866977441458415):

2021-03-01	2021-03-02	2021-03-03	2021-03-10	2021-03-11	2021-03-12	2021-03-13
299	436	423	214	431	427	441
2021-03-14	2021-03-16	2021-03-17	2021-03-18	2021-03-19	2021-03-20	2021-03-21
227	72	173	137	210	410	394
2021-03-22	2021-03-23	2021-03-24	2021-03-25	2021-03-26	2021-03-27	2021-03-28
398	452	308	219	121	115	90
2021-03-29	2021-03-30					
106	75					

*dv3* (SS0466973631458415):

2021-03-19	2021-03-20
171	216

*File5* (5-04\_2021.csv):

*dv1* (SS0866974771458415):

2021-04-01	2021-04-02	2021-04-03	2021-04-04	2021-04-05	2021-04-06	2021-04-07
442	466	334	352	386	65	125
2021-04-08	2021-04-09	2021-04-10	2021-04-11	2021-04-12	2021-04-13	2021-04-14
137	166	167	458	476	477	476
2021-04-15	2021-04-16	2021-04-17	2021-04-18	2021-04-19	2021-04-20	2021-04-21
167	136	410	276	229	479	477
2021-04-22	2021-04-23					
476	213					

*dv2* (SS0866977441458415):

2021-04-01	2021-04-02	2021-04-05	2021-04-06	2021-04-07	2021-04-08	2021-04-09
131	128	32	3	18	7	3
2021-04-10	2021-04-11	2021-04-12	2021-04-13	2021-04-14	2021-04-15	2021-04-17
67	459	472	477	471	168	411
2021-04-18	2021-04-19					
276	6					

*dv3* (SS0466973631458415):

2021-04-05	2021-04-16	2021-04-21	2021-04-22	2021-04-23
67	68	243	364	5

- 6) Cho biết ngày và có số lần thu thập dữ liệu thấp nhất của ngày đó trên từng thiết bị cảm biến. Phân loại dữ liệu dựa trên deviceid. Sử dụng hàm min() và which.min() cho table(date\_new) để tìm ra giá trị nhỏ nhất và ngày chứa giá trị đó

```
data_d = data[data$deviceid == "\"SS0866974771458415\"", ]#1
data_d = data[data$deviceid == "\"SS0866977441458415\"", ]#2
data_d = data[data$deviceid == "\"SS0466973631458415\"", ]#3
date_new = substr(data_d$date,1, 10)
date_new = as.Date(date_new,format = " %Y-%m-%d")
table(date_new)
#-----cau6-----#
which.min(table(date_new))
min(table(date_new))
```



Bảng 4: Ngày và số lần thu thập dữ liệu thấp nhất trong ngày

	Device	Date	Số lần
<b>File 3</b>	dv1	1/4/2021	131
	dv2	1/23/2021	8
<b>File 4</b>	dv1	3/16/2021	307
	dv2	3/16/2021	72
	dv3	3/19/2021	171
<b>File 5</b>	dv1	4/6/2021	65
	dv2	4/6/2021	3
	dv3	4/23/2021	5

- 7) Cho biết ngày và có số lần thu thập dữ liệu cao nhất của ngày đó trên từng thiết bị cảm biến.  
Tương tự như câu 6 nhưng thay thế hàm min() bằng hàm max() và which.max() cho table(date\_new)

```
which.max(table(date_new))  
max(table(date_new))
```

Bảng 5: Ngày và số lần thu thập dữ liệu cao nhất trong ngày

	Device	Date	Số lần
<b>File 3</b>	dv1	1/7/2021	476
	dv2	1/17/2021	323
<b>File 4</b>	dv1	3/29/2021	479
	dv2	3/23/2021	452
	dv3	3/20/2021	216
<b>File 5</b>	dv1	4/20/2021	479
	dv2	4/13/2021	477
	dv3	4/22/2021	364



- 8) Cho biết số ngày dài nhất liên tiếp mà dữ liệu không thu thập được trên từng thiết bị cảm biến.

Bảng 6: Số ngày dài nhất liên tiếp mà dữ liệu không thu thập được

	Device	Day	Date
<b>File 3</b>	dv1	0	0
	dv2	3	13 -> 15
<b>File 4</b>	dv1	0	0
	dv2	6	3 -> 9
	dv3	18	1 -> 18
<b>File 5</b>	dv1	8	24 -> 31
	dv2	12	20 -> 31
	dv3	10	6 -> 15

- 9) Cho biết số ngày ngắn nhất liên tiếp mà dữ liệu không thu thập được trên từng thiết bị cảm biến.

Bảng 7: Số ngày ngắn nhất liên tiếp mà dữ liệu không thu thập được

	Device	Day	Date
<b>File 3</b>	dv1	0	0
	dv2	1	7
<b>File 4</b>	dv1	0	0
	dv2	1	15
	dv3	11	21 - 31
<b>File 5</b>	dv1	0	0
	dv2	2	3 -> 4
	dv3	4	17 -> 20

- 10) Cho biết số lượng mẫu thu được trong các khoảng thời gian theo ngày trên từng thiết bị: 1-7, 8-14, 15-21, 22-28.

Phân loại dữ liệu dựa trên deviceid, sử dụng hàm filter trong thư viện dplyr để gộp

```
library(dplyr)
data <- read.csv("3-01_2021.csv")
data <- read.csv("4-03_2021.csv")
data <- read.csv("5-04_2021.csv")
data_d = data[data$deviceid == "\\"SS0866974771458415\\",]
data_d = data[data$deviceid == "\\"SS0866977441458415\\",]
data_d = data[data$deviceid == "\\"SS0466973631458415\\",]
count_devices=data.frame(table(data_d$deviceid))
l=data.frame(data_d%>%select(date,deviceid,Temp,Humi))
l$date=strptime(l$date,"%Y-%m-%d %H:%M:%S")
tuan1=1%>%filter(as.integer(format(as.Date(date),"%d"))>=1 & as.integer(format(as.Date(date),"%d"))<=7)
tuan2=1%>%filter(as.integer(format(as.Date(date),"%d"))>=8 & as.integer(format(as.Date(date),"%d"))<=14)
tuan3=1%>%filter(as.integer(format(as.Date(date),"%d"))>=15 & as.integer(format(as.Date(date),"%d"))<=21)
tuan4=1%>%filter(as.integer(format(as.Date(date),"%d"))>=22 & as.integer(format(as.Date(date),"%d"))<=28)
```



Bảng 8: Số lượng mẫu thu được trong File 3 - 3-01\_2021.csv

Khoảng thời gian	dv1	dv2	dv3
1 -> 7	1489	205	0
8 -> 14	2207	1041	0
15 -> 21	3038	1306	0
22 -> 28	2722	464	0

Bảng 9: Số lượng mẫu thu được trong File 4 - 4-03\_2021.csv

Khoảng thời gian	dv1	dv2	dv3
1 -> 7	3015	1158	0
8 -> 14	3247	1740	0
15 -> 21	2389	1396	387
22 -> 28	3171	1703	0

Bảng 10: Số lượng mẫu thu được trong File 5 - 5-04\_2021.csv

Khoảng thời gian	dv1	dv2	dv3
1 -> 7	2170	312	67
8 -> 14	2357	1956	0
15 -> 21	2174	861	311
22 -> 28	689	0	369

- 11) Có ngày nào mà tất cả các thiết bị số lượng thu thập dữ liệu là bằng nhau không? Hãy cho biết các ngày đó.



- 12) Với một ngày k biết được, hãy cho biết thiết bị nào có lượng dữ liệu thu thập được thấp nhất. Thay ngày "2021-01-05" bằng một ngày bất kỳ, code sẽ tạo một dataframe lưu các giá trị thu được trong ngày đó và phân loại dựa trên *deviceid*. Cuối cùng, so sánh số lượng các dataframe thu được của mỗi thiết bị sẽ biết được thiết bị nào thu được thấp/nhiều nhất trong từng file.

```
setwd("E:/DOCUMENT/second year/sem1/Cau truc roi rac/BTL/BTL1")
data <- read.csv("3-01_2021.csv")
data <- read.csv("4-03_2021.csv")
data <- read.csv("5-04_2021.csv")
data_d = data
date_new = substr(data_d$date,1, 10)
date_new = as.Date(date_new,format = "%Y-%m-%d")
data_d = cbind(data_d, date_new)
data_d = data_d[data_d$date_new" == "2021-01-05",]
data_d1 = data_d[data_d$deviceid == "\"SS0866974771458415\"",]#1
data_d2 = data_d[data_d$deviceid == "\"SS0866977441458415\"",]#2
data_d3 = data_d[data_d$deviceid == "\"SS0466973631458415\"",]#3
datelength <- c(length(data_d1$deviceid), length(data_d2$deviceid), length(data_d3$deviceid))
min(datelength)
which.min(datelength)
```

- 13) Với một ngày k biết được, hãy cho biết thiết bị nào có lượng dữ liệu thu thập được nhiều nhất. Tương tự câu 12, thay thế bằng hàm max và which.max

```
setwd("E:/DOCUMENT/second year/sem1/Cau truc roi rac/BTL/BTL1")
data <- read.csv("3-01_2021.csv")
data <- read.csv("4-03_2021.csv")
data <- read.csv("5-04_2021.csv")
data_d = data
date_new = substr(data_d$date,1, 10)
date_new = as.Date(date_new,format = "%Y-%m-%d")
data_d = cbind(data_d, date_new)
data_d = data_d[data_d$date_new" == "2021-01-05",]
data_d1 = data_d[data_d$deviceid == "\"SS0866974771458415\"",]#1
data_d2 = data_d[data_d$deviceid == "\"SS0866977441458415\"",]#2
data_d3 = data_d[data_d$deviceid == "\"SS0466973631458415\"",]#3
datelength <- c(length(data_d1$deviceid), length(data_d2$deviceid), length(data_d3$deviceid))
max(datelength)
which.max(datelength)
```



ii) Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

Trên từng thiết bị cần tính số liệu thống kê lần lượt cho 2 thể loại nhiệt độ và độ ẩm như sau:

```
data <- read.csv("3-01_2021.csv")
data <- read.csv("4-03_2021.csv")
data <- read.csv("5-04_2021.csv")
#chạy 1 trong 3 dòng trên#
data_d1 = data[data$deviceid == "\"SS0866974771458415\"",]
data_d2 = data[data$deviceid == "\"SS0866977441458415\"",]
data_d3 = data[data$deviceid == "\"SS0466973631458415\"",]
```

Chọn số liệu của một trong ba file và phân loại dựa trên *deviceid* của thiết bị

Đáp án các câu {1,...,5} được tổng hợp trong bảng số liệu ở câu 6

Hàm *summary* trả các giá trị: min, max, tứ phân vị, giá trị trung bình. Nên có thể sử dụng hàm để trả lời các câu hỏi 1,2,3

Sử dụng hàm:

```
summary(data_d1)
summary(data_d2)
summary(data_d3)
```

Kết quả thu được(SS0866974771458415 - file 3 – 01\_2021.csv):

Temp	Humi
Min. : 8.70	Min. :22.80
1st Qu.:16.00	1st Qu.:56.90
Median :19.50	Median :76.90
Mean :21.36	Mean :69.19
3rd Qu.:26.00	3rd Qu.:80.90
Max. :42.00	Max. :98.10
NA's :3737	

1) Tính giá trị nhỏ nhất, lớn nhất

Sử dụng hàm *min, max* hoặc hàm *summary*

```
#vd câu 1#
max(data_d1[["Temp"]], na.rm = TRUE)
min(data_d1[["Temp"]], na.rm = TRUE)
max(data_d1[["Humi"]], na.rm = TRUE)
min(data_d1[["Humi"]], na.rm = TRUE)
```

2) Tính tứ phân vị thứ nhất(Q1), thứ hai(Q2), thứ ba(Q3)

Sử dụng hàm *quantile* hoặc hàm *summary*

```
#vd câu 2#
quantile(data_d1$Temp, na.rm = TRUE)
quantile(data_d1$Humi, na.rm = TRUE)
```

3) Tính giá trị trung bình (Avg)

Sử dụng hàm *mean*

```
#vd câu 3#
mean(data_d1$Temp, na.rm = TRUE)
mean(data_d1$Humi, na.rm = TRUE)
```



- 4) Tính giá trị độ lệch chuẩn (Std)

Sử dụng hàm *sd*

```
#vd câu 4#
sd4 = data_d1$Temp
sd4 = data_d2$Temp
sd4 = data_d3$Temp
sd4 = data_d1$Humi
sd4 = data_d2$Humi
sd4 = data_d3$Humi
#chạy 1 trong 6 dòng trên#
standd = sd(sd4, na.rm = TRUE)
standd
```

- 5) Dếm xem có bao nhiêu outliers, một quan sát mà giá trị của nó nằm trong khoảng sau:

$$IQR = Q3 - Q1$$

$$\text{outliers} < Q1 - 1.5 * IQR \text{ hoặc outliers} > Q3 + 1.5 * IQR$$

Tạo ra các biến lưu trữ các giá trị outliers và đếm tổng số lượng phần tử trong đó

```
#vd câu 5#
qnt_dur = data_d1$Temp
qnt_dur = data_d1$Humi
qnt_dur = data_d2$Temp
qnt_dur = data_d2$Humi
qnt_dur = data_d3$Temp
qnt_dur = data_d3$Humi
#chọn 1 trong 6 dòng trên#
qnt = quantile(qnt_dur, na.rm = TRUE)
qnt
outl_low = qnt["25%"] - 1.5*(qnt["75%"]-qnt["25%"])
outl_hig = qnt["75%"] + 1.5*(qnt["75%"]-qnt["25%"])
outliers_low = qnt_dur[qnt_dur<outl_low & is.na(qnt_dur) == 0]
outliers_high= qnt_dur[qnt_dur>outl_hig & is.na(qnt_dur) == 0]
outliers = length(outliers_high) + length(outliers_low)
outliers
```

- 6) Lập bảng mô tả số liệu thống kê các thiết bị cho từng thẻ loại:

Temp[Humi]:

Devices	Min	Q1	Q2	Q3	Max	Avg	Std	Outlier
dvi	?	?	?	?	?	?	?	?

Bảng 11: Bảng số liệu Humi

		Min	Q1	Q2	Q3	Max	Avg	Std	Outlier
SS0866974771458415	File 3	22.8	56.9	76.9	80.9	98.1	69.19	17.32832	0
	File 4	16.1	45.9	75.9	80.5	95.6	64.66	19.81573	0
	File 5	23	51.9	76	80.9	95.3	68.03	18.36201	0
SS0866977441458415	File 3	32.9	60	80.9	83	96.4	72.19	15.12838	0
	File 4	17.7	51.8	81	86.3	98.1	71.2	20.95864	0
	File 5	25.9	59	84	90	97.4	75.2	19.53726	0
SS0466973631458415	File 3	NA	NA	NA	NA	NA	NA	NA	NA
	File 4	NA	NA	NA	NA	NA	NA	NA	NA
	File 5	NA	NA	NA	NA	NA	NA	NA	NA

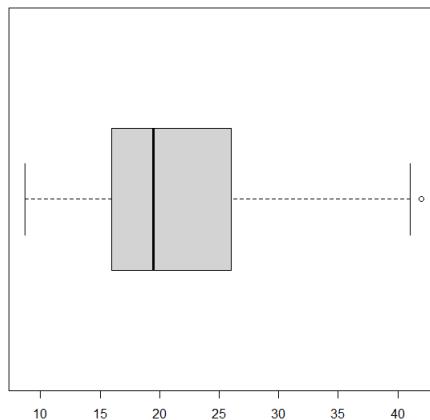
Bảng 12: Bảng số liệu Temp

		Min	Q1	Q2	Q3	Max	Avg	Std	Outlier
SS0866974771458415	File 3	8.7	16	19.5	26	42	21.36	7.11075	2
	File 4	9.3	17.2	21.1	30.8	48	23.9	8.658778	0
	File 5	15.4	20.5	22.6	31.4	47.5	26.01	7.415446	0
SS0866977441458415	File 3	9.9	14.2	17	22.9	35.9	18.89	5.818591	0
	File 4	8.6	16.6	18.6	27.98	44.4	22.01	7.834721	0
	File 5	14.2	19.1	20.9	28.7	44.9	24.12	7.336232	14
SS0466973631458415	File 3	NA	NA	NA	NA	NA	NA	NA	NA
	File 4	0	27.91	28.44	28.81	29.07	28.12	2.441374	3
	File 5	0	18.04	19.84	22	28.33	19.46	5.534976	46

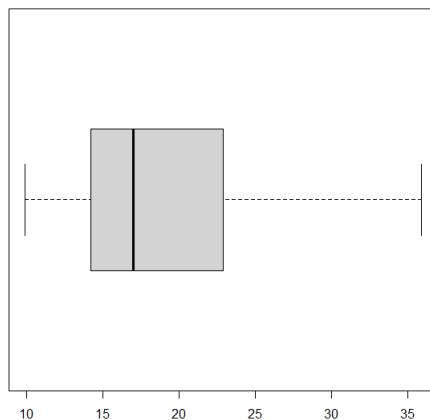
- 7) Vẽ biểu đồ boxplot hay còn được gọi là box-and-whisker cho nhiệt độ

Dùng hàm boxplot và chọn số liệu cột Temp để vẽ biểu đồ

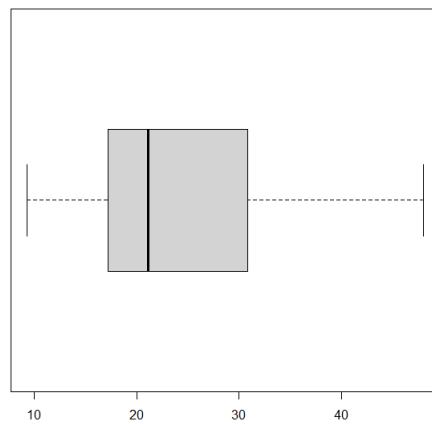
```
boxplot(data_d1$Temp, horizontal = TRUE)
boxplot(data_d2$Temp, horizontal = TRUE)
boxplot(data_d3$Temp, horizontal = TRUE)
```



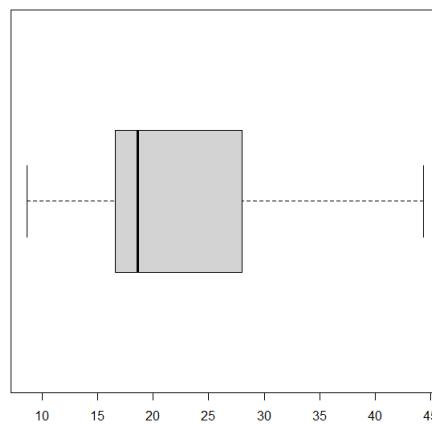
Biểu đồ boxplot cho nhiệt độ File 3 - thiết bị 1



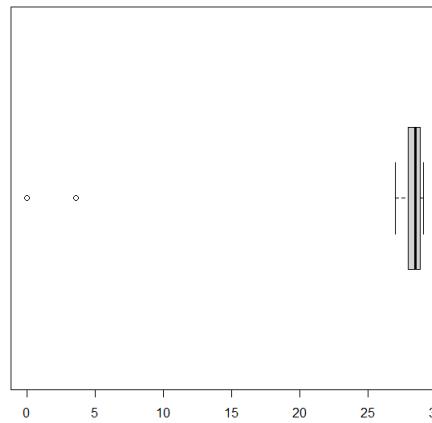
Biểu đồ boxplot cho nhiệt độ File 3 - thiết bị 2



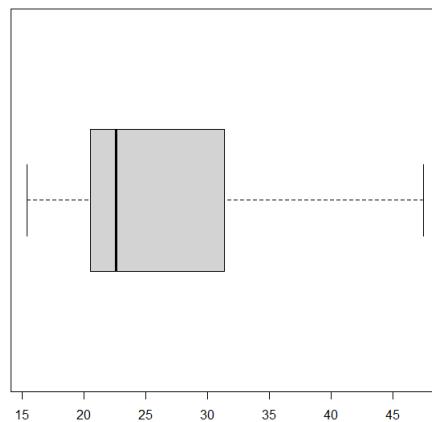
Biểu đồ boxplot cho nhiệt độ File 4 - thiết bị 1



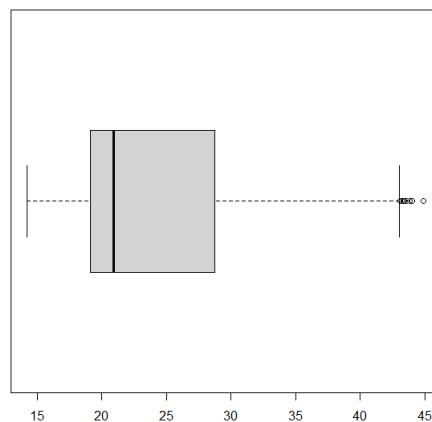
Biểu đồ boxplot cho nhiệt độ File 4 - thiết bị 2



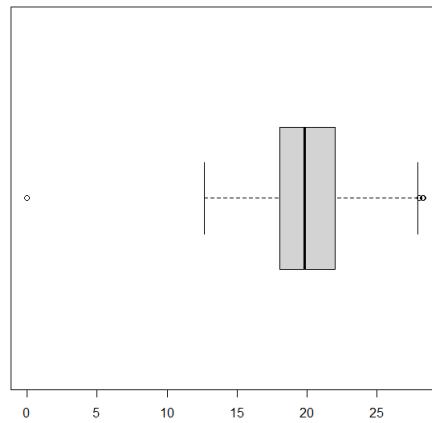
Biểu đồ boxplot cho nhiệt độ File 4 - thiết bị 3



Biểu đồ boxplot cho nhiệt độ File 5 - thiết bị 1



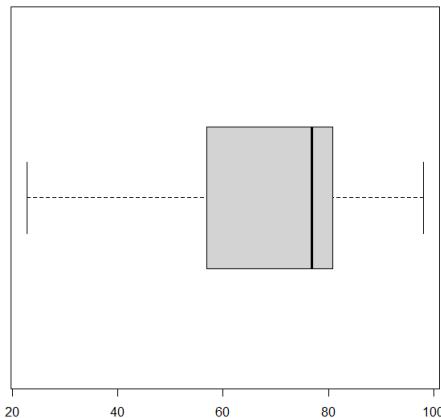
Biểu đồ boxplot cho nhiệt độ File 5 - thiết bị 2



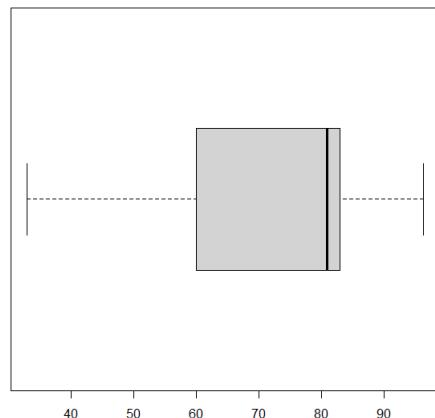
Biểu đồ boxplot cho nhiệt độ File 4 - thiết bị 5

- 8) Vẽ biểu đồ boxplot hay còn được gọi là box-and-whisker cho độ ẩm  
Dùng hàm boxplot và chọn số liệu cột Humi để vẽ biểu đồ

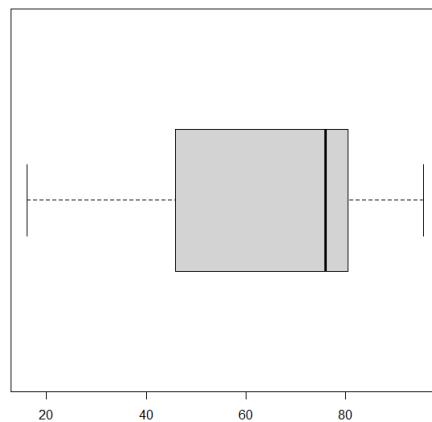
```
boxplot(data_d1$Humi, horizontal = TRUE)  
boxplot(data_d2$Humi, horizontal = TRUE)  
boxplot(data_d3$Humi, horizontal = TRUE)
```



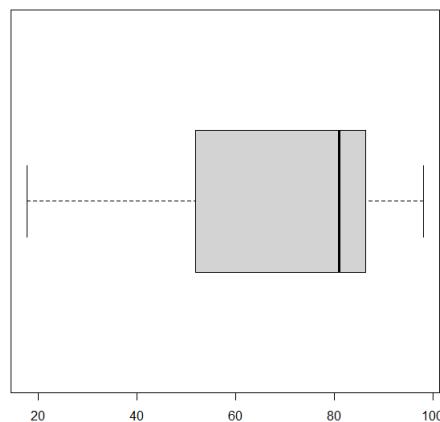
Biểu đồ boxplot cho độ ẩm File 3 - thiết bị 1



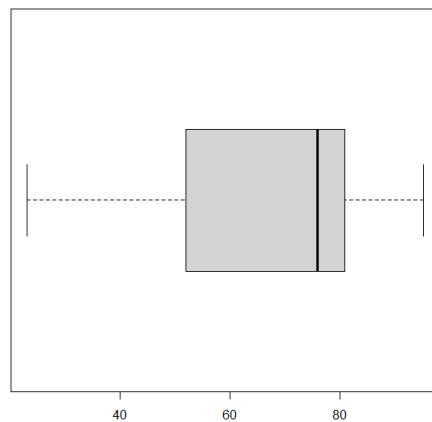
Biểu đồ boxplot cho độ ẩm File 3 - thiết bị 2



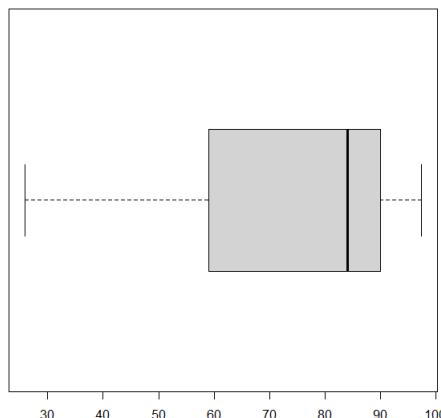
Biểu đồ boxplot cho độ ẩm File 4 - thiết bị 1



Biểu đồ boxplot cho độ ẩm File 4 - thiết bị 2



Biểu đồ boxplot cho độ ẩm File 5 - thiết bị 1



Biểu đồ boxplot cho độ ẩm File 5 - thiết bị 2

iii) Nhóm câu hỏi liên quan đến sai lệch

Trên từng thiết bị cần xác định độ sai số dữ liệu theo nhiệt độ (Temp) và theo độ ẩm (Humi) theo 2 cách sau:

- 1) Tính giá trị phần lẻ nhỏ nhất của các giá trị thu thập được

Trước tiên ta khởi tạo khán biến alo1, alo2 để lọc dữ liệu theo thiết bị

Dùng lệnh dropna để loại bỏ các giá trị NA trong bảng dữ liệu

Sau đó dùng câu lệnh min(giatri - as.integer(giatri)) để tính phần lẻ nhỏ nhất

```
library(dplyr)
library(ggplot2)
a=read.csv("Documents/5-04_2021.csv")
#khởi tạo khán biến alo1, alo2 để lọc dữ liệu theo thiết bị
sp1 <- filter(a, deviceid=='SS0866974771458415')
alo=sp1%>%select(date,deviceid,Temp,Humi)
sp2 <- filter(a, deviceid=='SS0866977441458415')
alo2=sp2%>%select(date,deviceid,Temp,Humi)
##y1
#min(giatri - as.integer(giatri)) để tính phần lẻ nhỏ nhất
beta_1= min(al0$Temp - as.integer(al0$Temp))
beta_2= min(al02$Temp - as.integer(al02$Temp))
#drop_na(Humi) để loại bỏ giá trị NA
beta_3= min((al0 %>% drop_na(Humi))$Humi - as.integer((al0 %>% drop_na(Humi))$Humi))
beta_4= min((al02 %>% drop_na(Humi))$Humi - as.integer((al02 %>% drop_na(Humi))$Humi))
```

- 2) Đo khoảng cách chênh lệch nhỏ nhất thu được nếu có thể sắp xếp các giá trị thu được theo nhiệt độ (Temp) và theo độ ẩm (Humi) từ nhỏ đến lớn

Đầu tiên, dùng lệnh order để sắp xếp dữ liệu từ thấp đến cao, dùng lệnh dropna để loại bỏ các giá trị NA trong bảng dữ liệu

Tiếp đến sử dụng lệnh diff để tính hiệu 2 phần tử liên tiếp

Cuối cùng, dùng lệnh min để tìm chênh lệch nhỏ nhất

```
small_1=min(diff(al0[order(al0$Temp),]$Temp))
small_2=min(diff(al02[order(al02$Temp),]$Temp))
small_3=min(diff((al0 %>% drop_na(Humi))[order((al0 %>% drop_na(Humi)),]$Humi]))
small_4=min(diff((al02 %>% drop_na(Humi))[order((al02 %>% drop_na(Humi)),]$Humi)))
```



3) Thiết hiện bảng số liệu như sau:

File 3:

Sai lệch theo số lẻ nhỏ nhất:

Devices	Temp	Humi
SS0866974771458415	0	0
SS0866977441458415	0	0

Sai lệch theo chênh lệch nhỏ nhất:

Devices	Temp	Humi
SS0866974771458415	0	0
SS0866977441458415	0	0

File 4:

Sai lệch theo số lẻ nhỏ nhất:

Devices	Temp	Humi
SS0866974771458415	0	0
SS0866977441458415	0	0
SS0466973631458415	0	

Sai lệch theo chênh lệch nhỏ nhất:

Devices	Temp	Humi
SS0866974771458415	0	0
SS0866977441458415	0	0

File 5:

Sai lệch theo số lẻ nhỏ nhất:

Devices	Temp	Humi
SS0866974771458415	0	0
SS0866977441458415	0	0
SS0466973631458415	0	

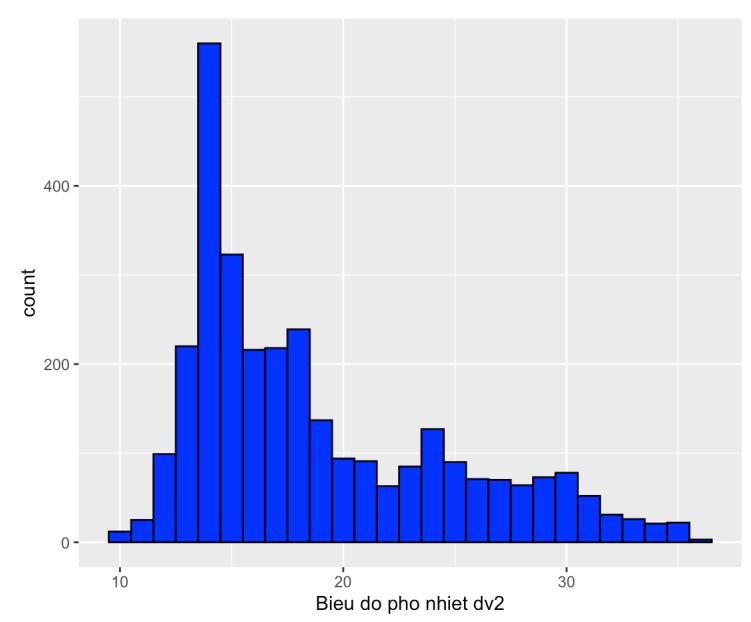
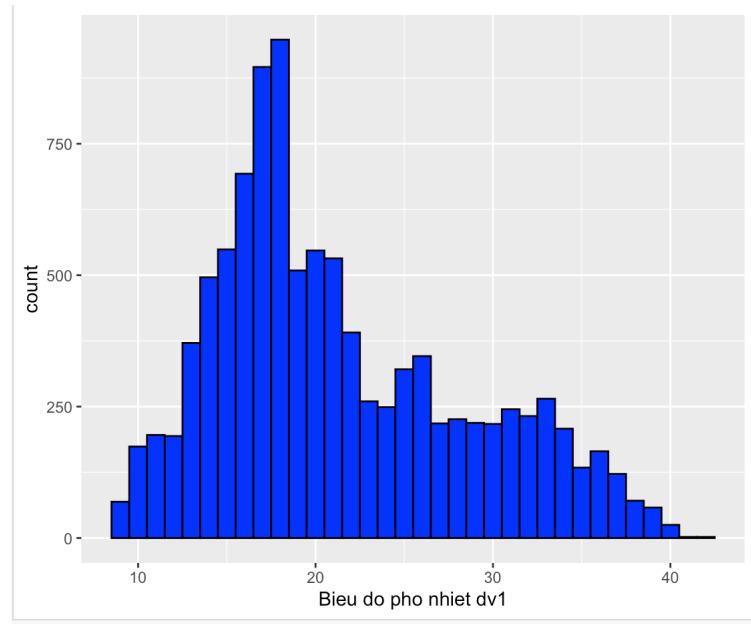
Sai lệch theo chênh lệch nhỏ nhất:

Devices	Temp	Humi
SS0866974771458415	0	0
SS0866977441458415	0	0

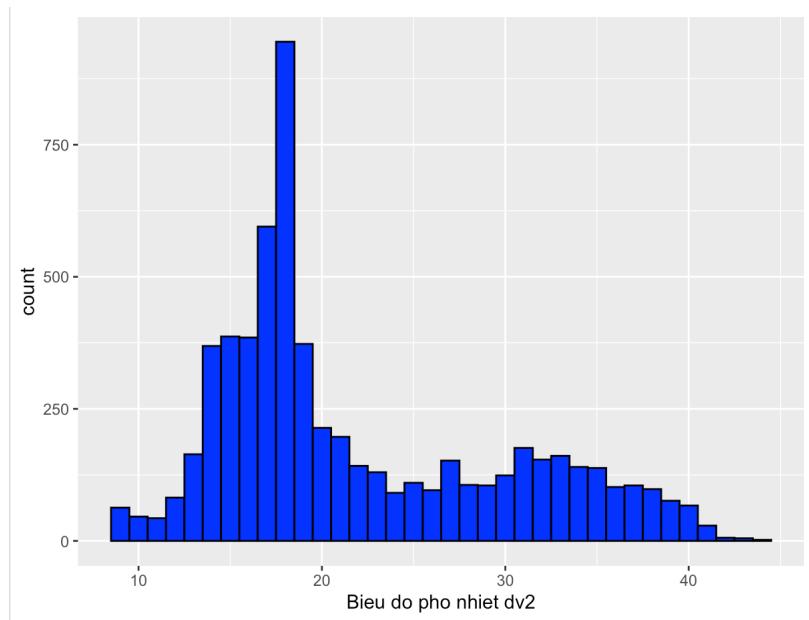
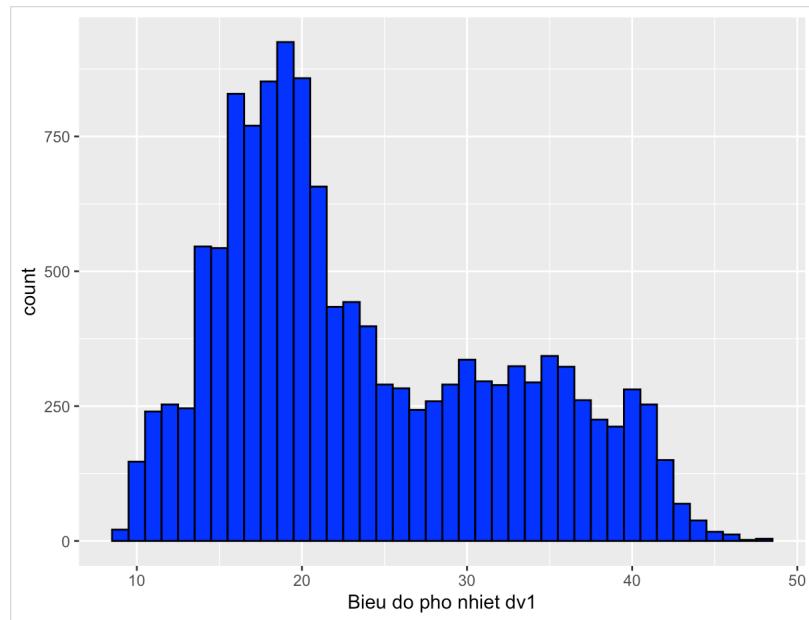
4) Vẽ biểu đồ phô nhiệt độ theo từng thiết bị cảm biến

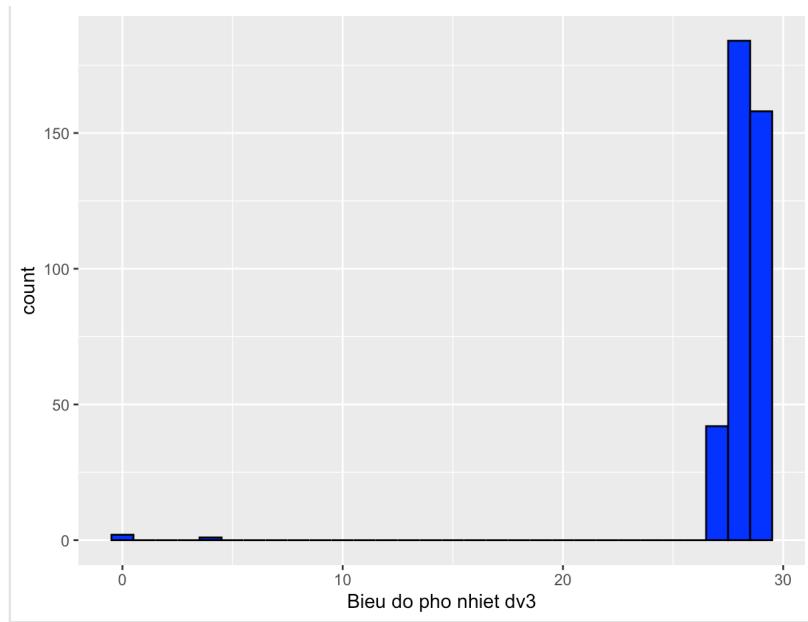
```
data=a%>%select(deviceid,Temp,Humi)
#loc du lieu theo thiet bi roi ve do thi
Bang_nhiet1=data%>%filter(deviceid=="SS0866974771458415")%>%ggplot(aes(Temp))
+ geom_histogram(col="black",binwidth = 1,fill="blue") + xlab("Bieu do pho nhiet dv1")
Bang_nhiet2=data%>%filter(deviceid=="SS0866977441458415")%>%ggplot(aes(Temp))
+ geom_histogram(col="black",binwidth = 1,fill="blue") + xlab("Bieu do pho nhiet dv2")
```

File 3

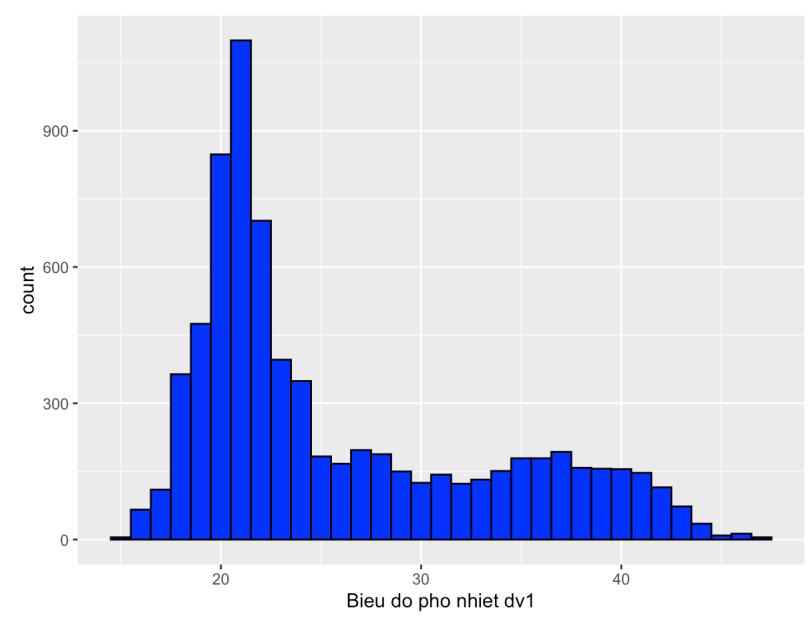


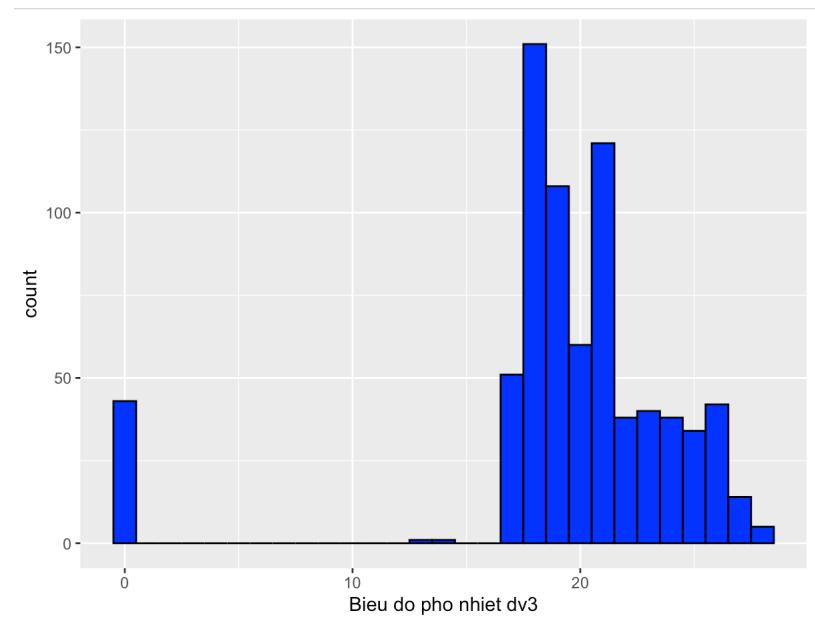
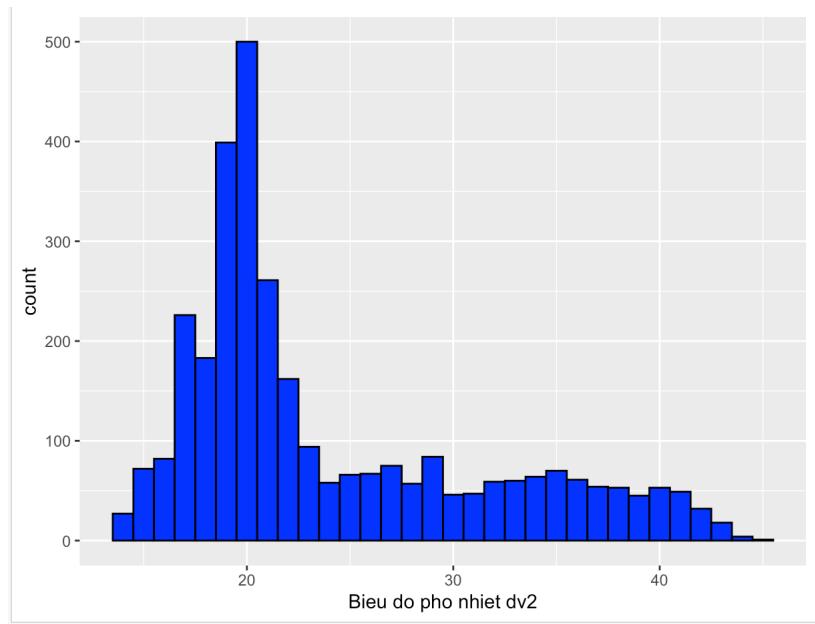
File 4





File 5

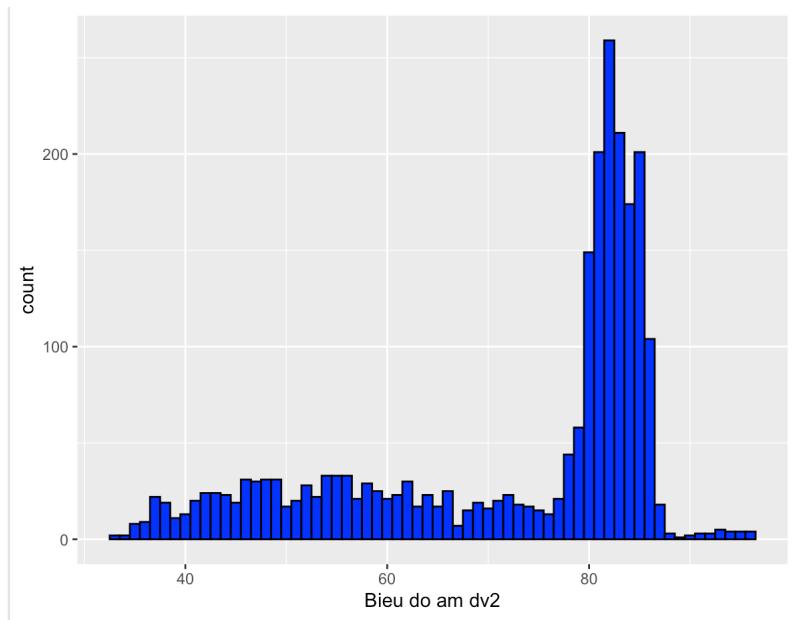
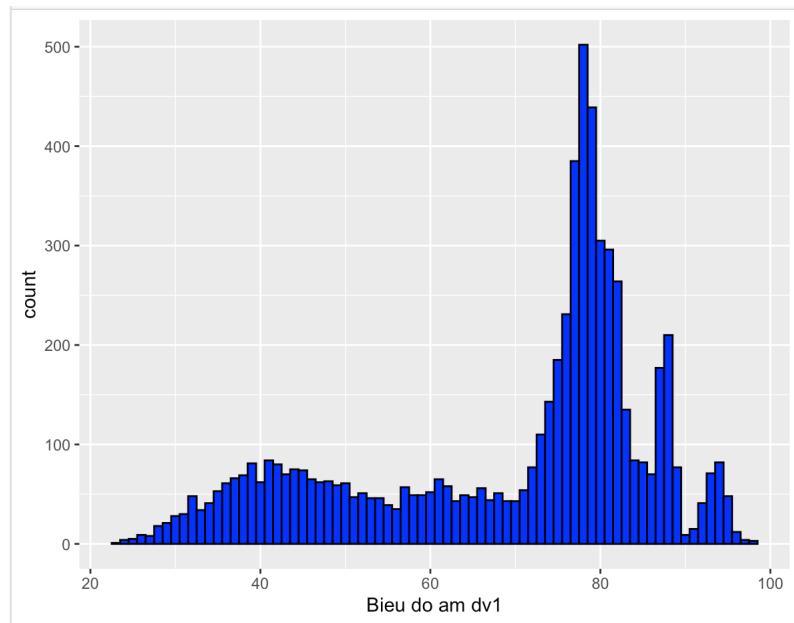




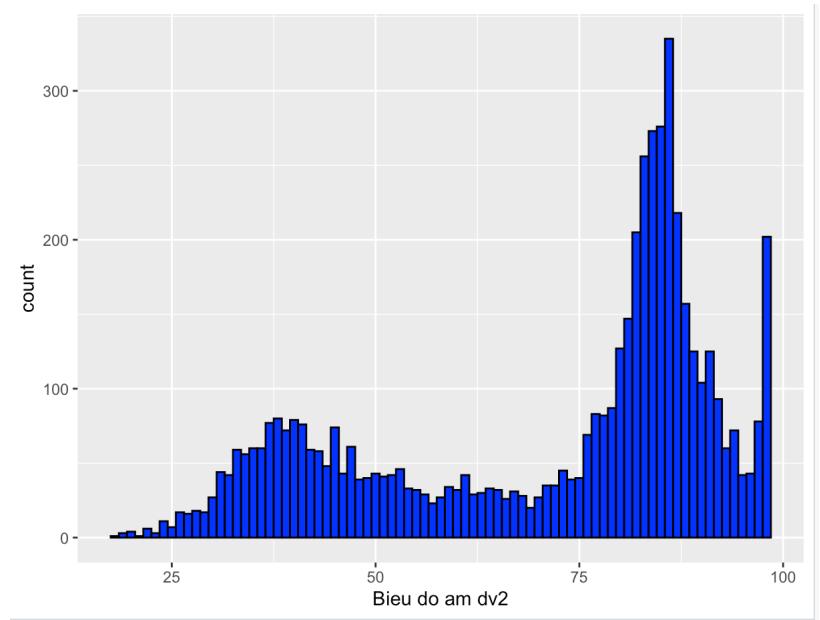
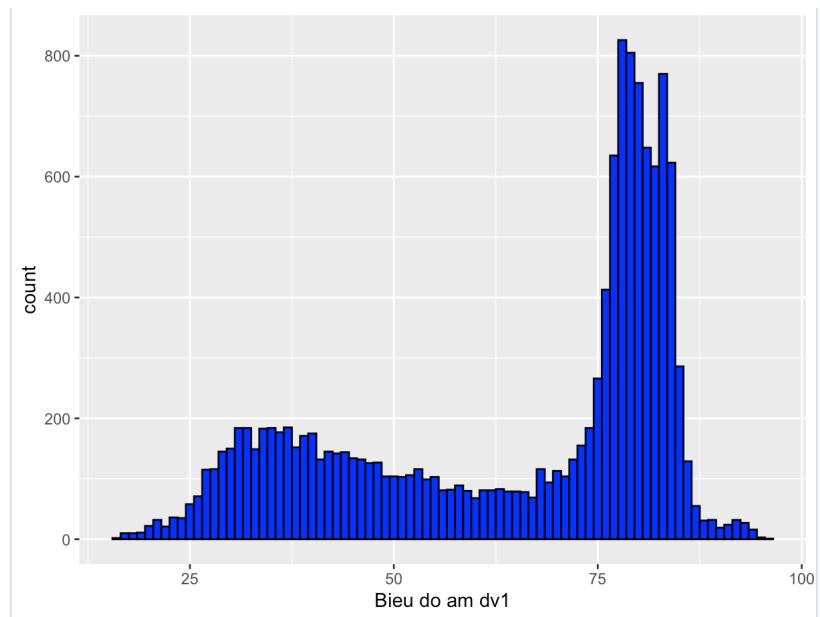
5) Vẽ biểu đồ phô độ ẩm theo từng thiết bị cảm biến

```
#loc du lieu theo thiet bi roi ve do thi
Bang_am_1=data%>%filter(deviceid=='SS0866974771458415')%>%ggplot(aes(Humi))+  
  geom_histogram(col="black",binwidth = 1,fill="blue") + xlab("Bieu do am dv1")
Bang_am_2=data%>%filter(deviceid=='SS0866977441458415')%>%ggplot(aes(Humi))+  
  geom_histogram(col="black",binwidth = 1,fill="blue") + xlab("Bieu do am dv2")
```

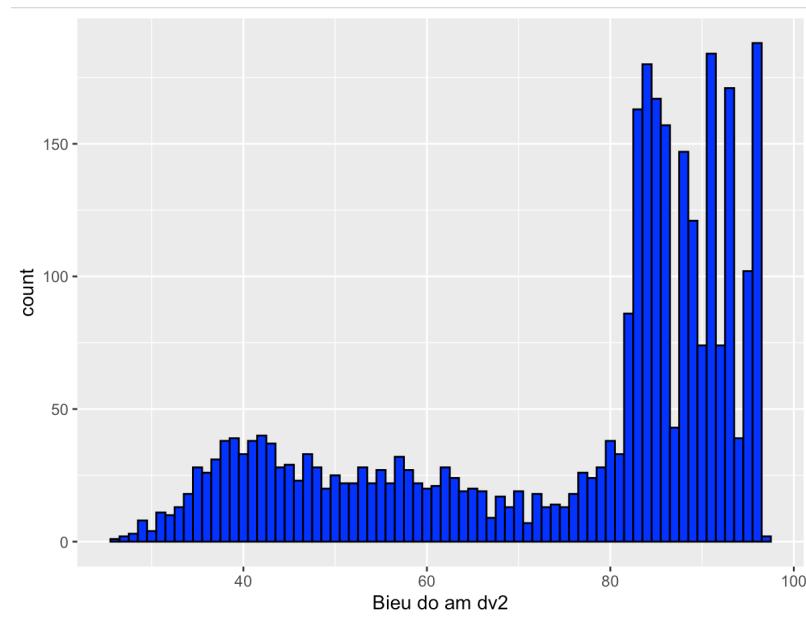
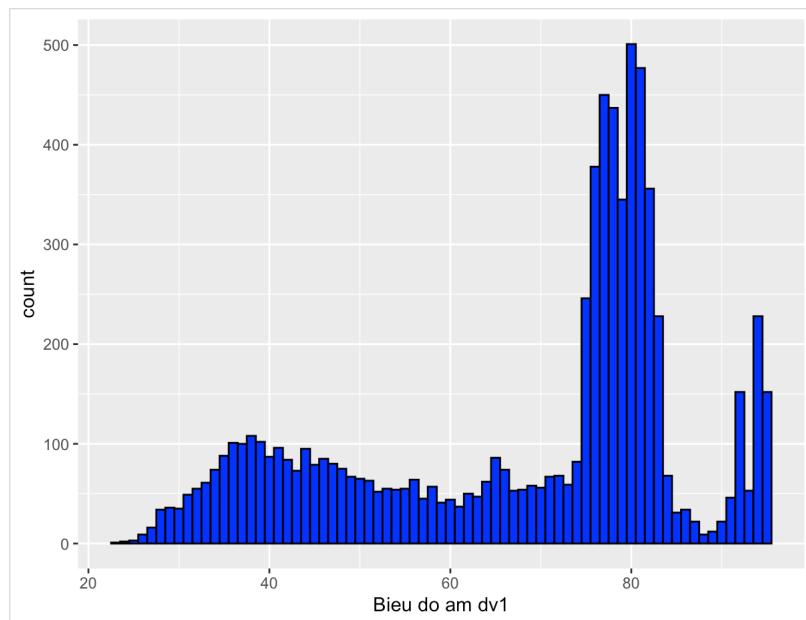
File 3



File 4



File 5







iv) Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là ngày

Trên từng thiết bị hãy vẽ biểu đồ thể hiện trực Ox là thời gian, trục Oy là nhiệt độ/độ ẩm. Hãy dùng 4 ký số của mã để vẽ 4 ngày tương ứng theo ký số đó. Nếu ký số là 0 thì lấy ngày là 10.

1 Biểu đồ thu thập nhiệt độ theo thời gian là ngày của từng thiết bị

FILE 3:

Thực hiện vẽ đồ thị vào các ngày 2,3,4,6 (nếu có)

Ta thực hiện các bước như sau:

- Tạo một Data-frame chứa các dữ liệu của ngày mà ta muốn vẽ biểu đồ bằng cách dùng chức năng Filter trên thanh công cụ R, ta xác định được nhóm dữ liệu theo từng ngày.
- Tạo một data Frame mới với điều kiện là phân loại theo từng thiết bị.
- Dùng hàm ggplot trong gói ggplot-2 để vẽ biểu đồ với trục Ox là thời gian trong 1 ngày và trục Oy thể hiện giá trị Temp theo ngày đó.

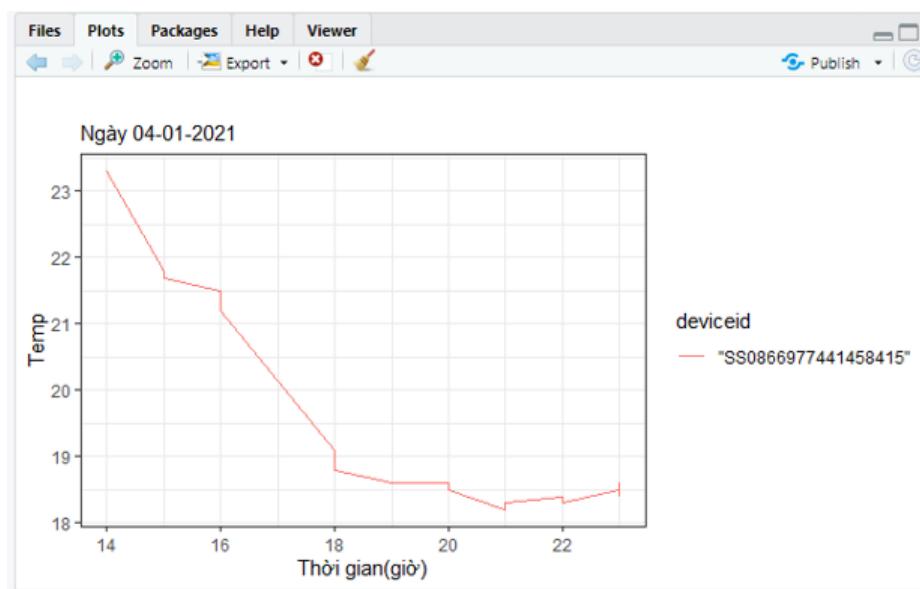
```
setwd('C:/Users/DELL/Desktop/R')
FILE_3 <- read.csv("3-01_2021.csv")
install.packages("lubridate")
library(lubridate)

attach(FILE_3)

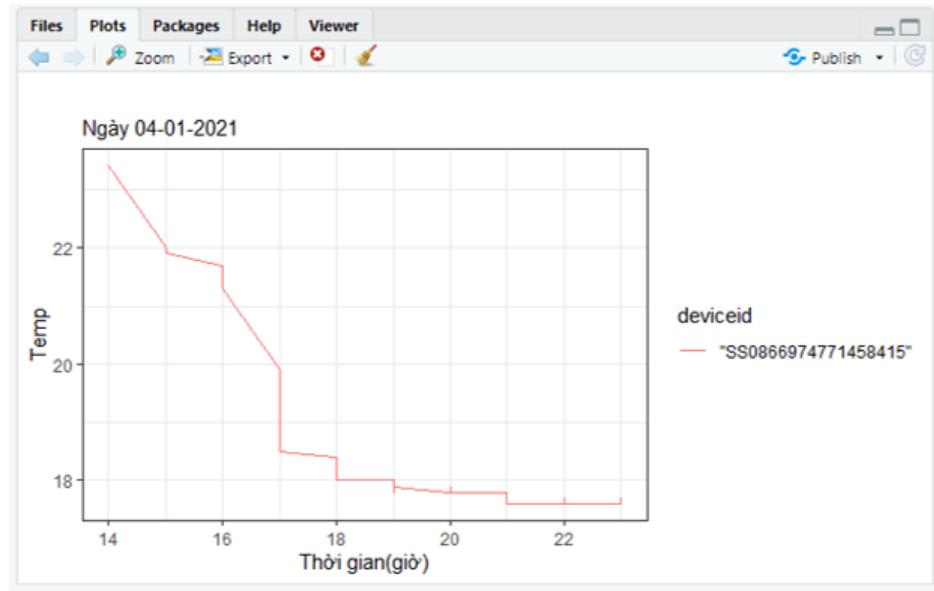
Date_4 <- FILE_3[671:1570,]

Date_44 <- Date_4[Date_4$deviceid=="\"SS0866977441458415\"",]
attach(Date_44)
Date_4f33 = ggplot(data=Date_44, aes(x=hour(date), y=Temp))
Date_4f33 + geom_line(aes(col=deviceid)) + labs(title = "", subtitle =
"Ngày 04-01-2021",x = "Thời gian(giờ)",y = "Temp") + theme_bw()
```

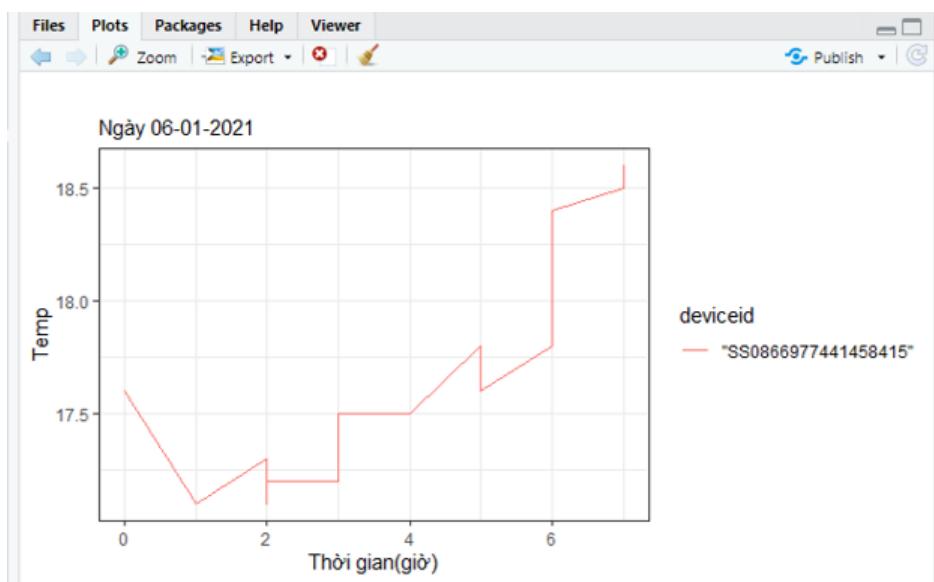
- Cuối cùng ta được biểu đồ như bên dưới:

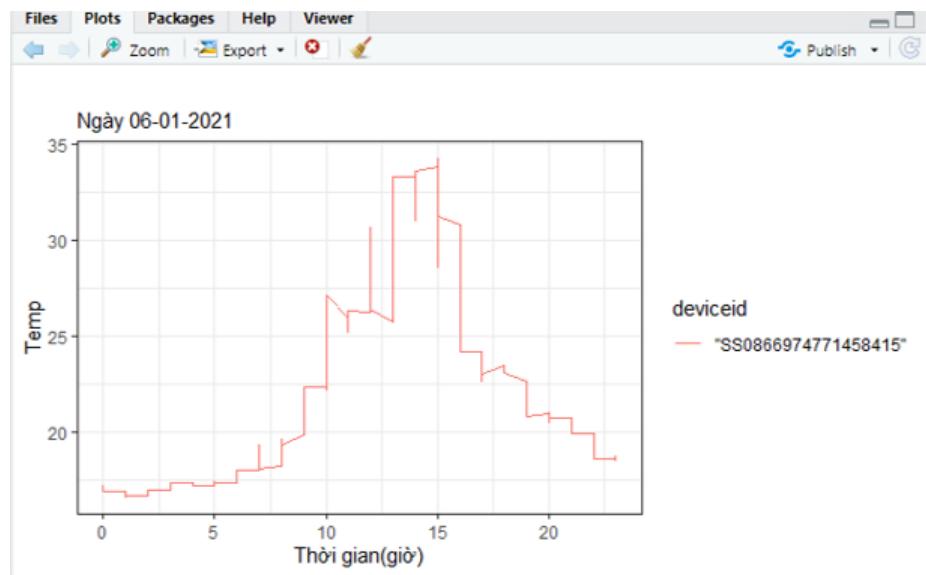


- Loại thiết bị thứ 2:



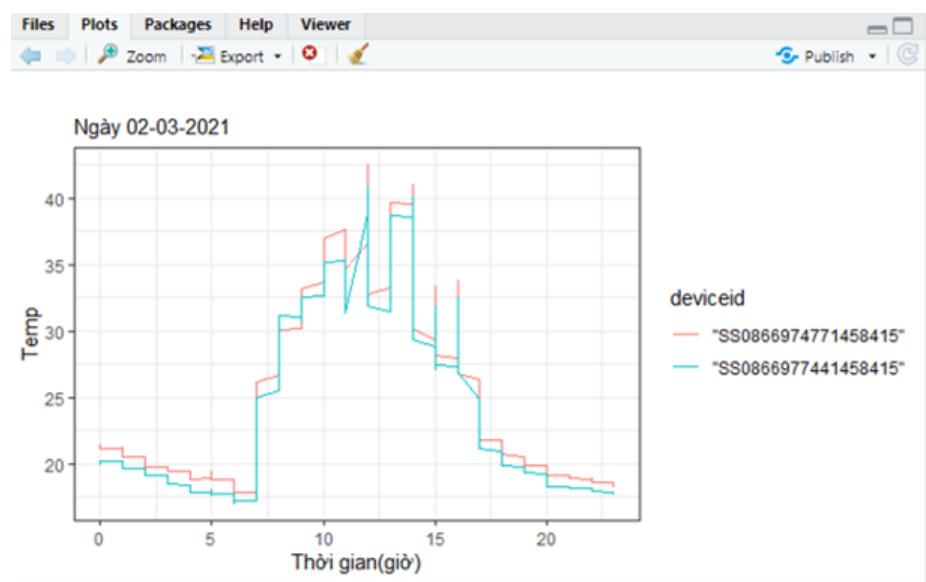
- Tương tự, ta vẽ được biểu đồ tương quan của nhiệt độ và thời gian trong ngày 06/01/2021 theo từng thiết bị như sau:

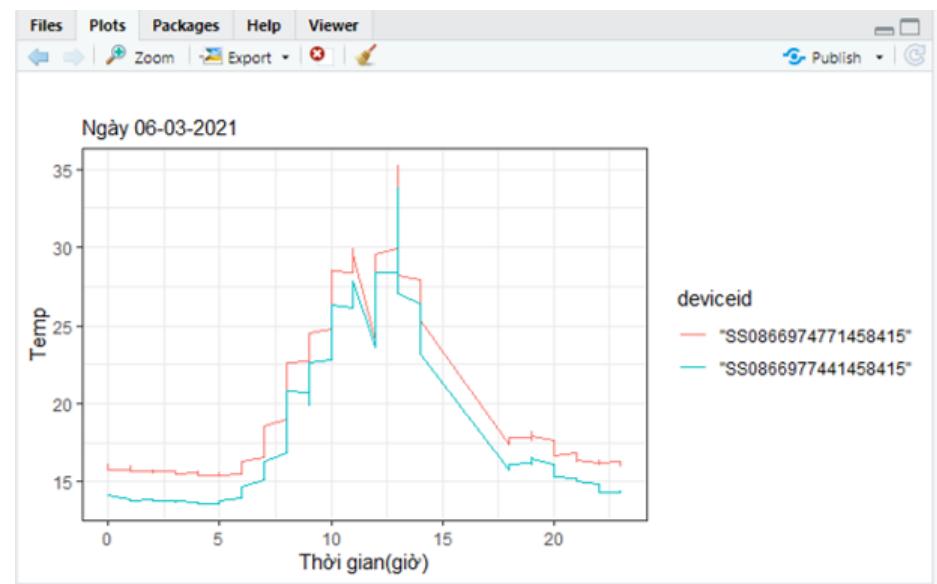
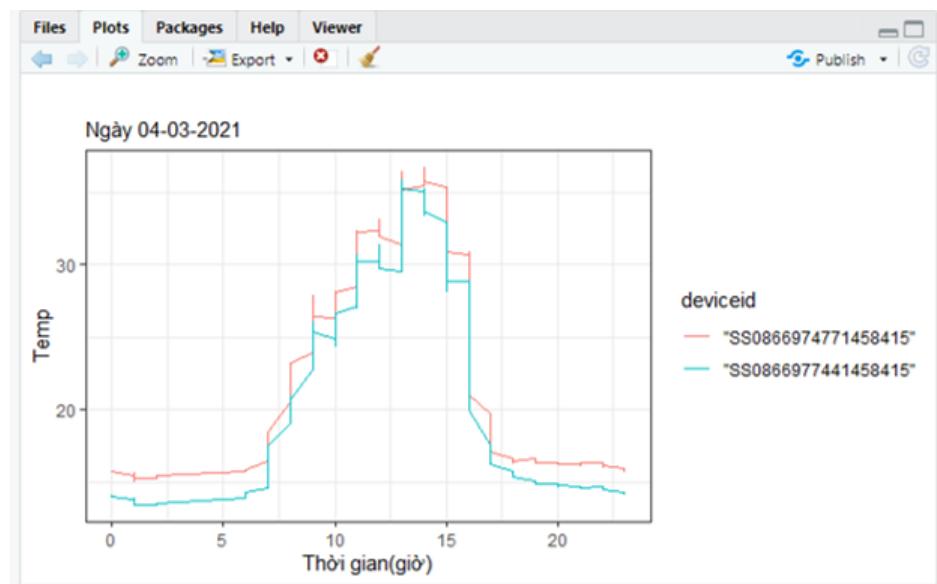
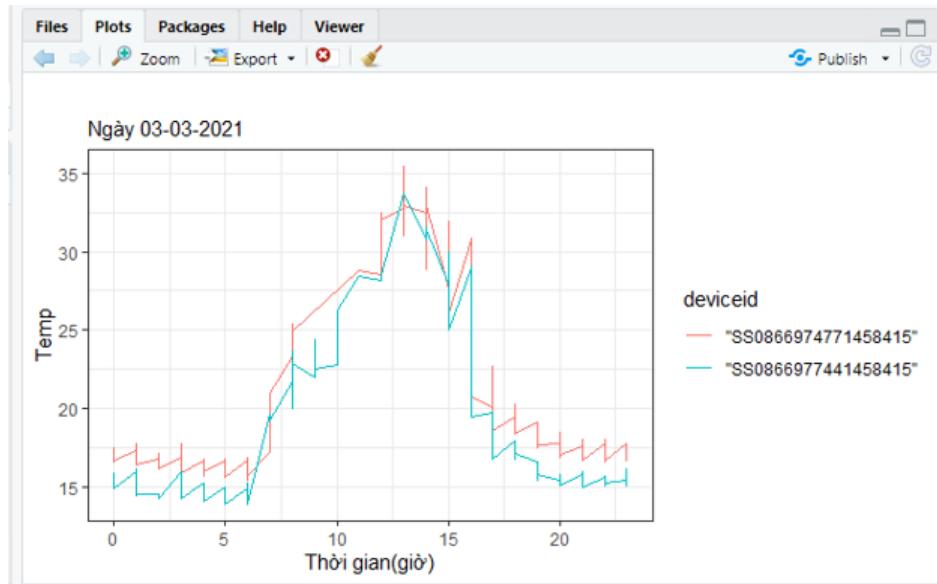




FILE 4:

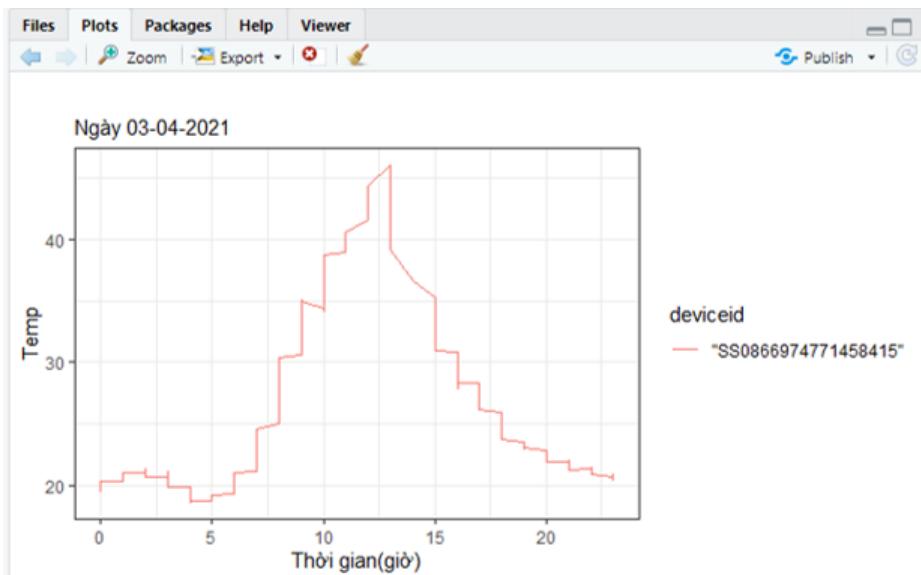
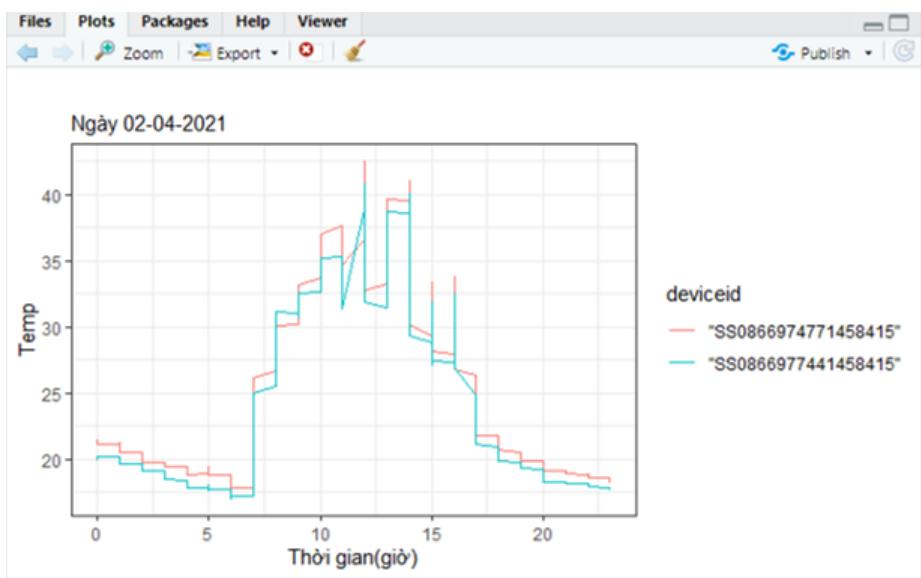
Thực hiện tương tự như file 3, ta cũng thu được biểu đồ tương quan của nhiệt độ và thời gian trong các ngày 2,3,4 và 6/03/2021, chi tiết thể hiện như bên dưới:

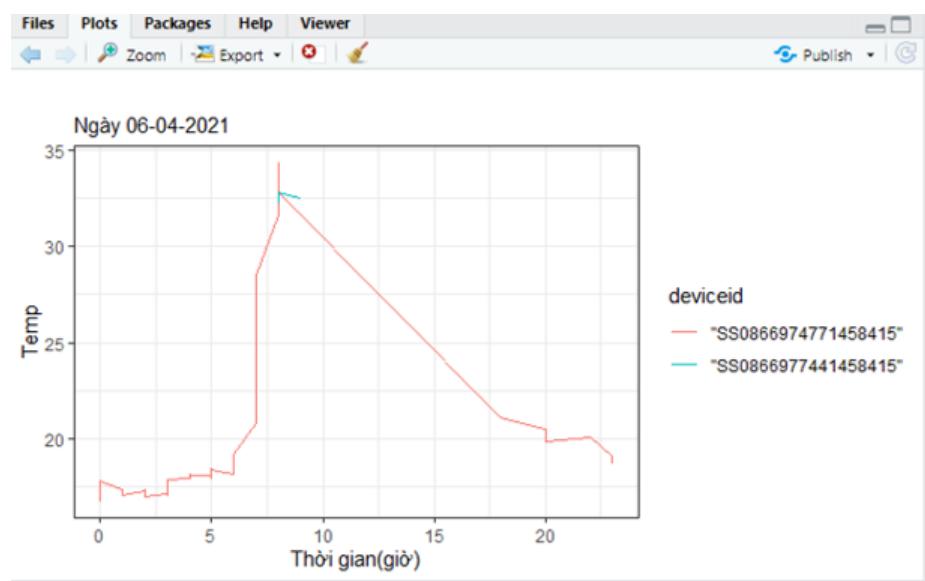
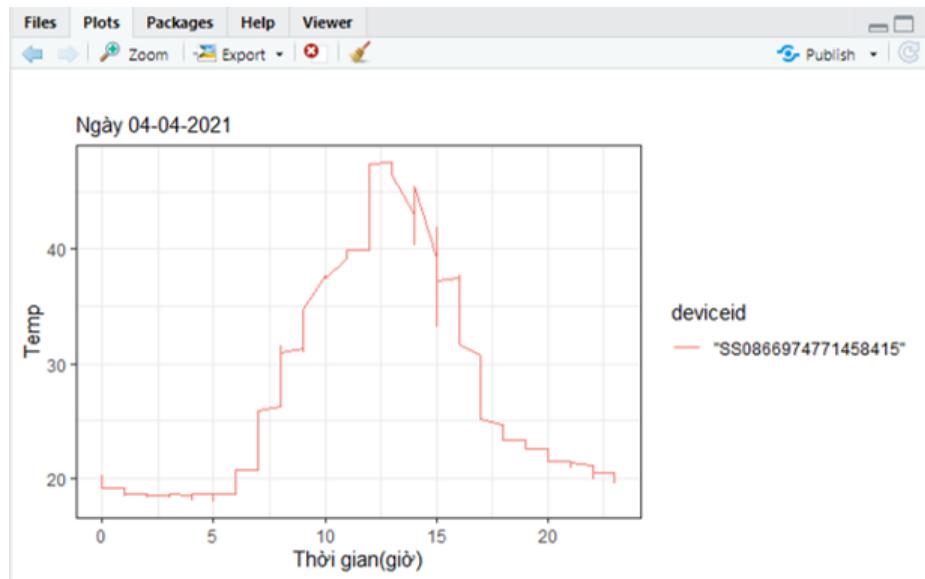




FILE 5:

Thực hiện tương tự như file 3 và file 4, ta vẽ được biểu đồ tương quan của nhiệt độ và thời gian theo từng thiết bị như sau:





- 2 Biểu đồ thu thập độ ẩm theo thời gian là ngày của từng thiết bị  
FILE 3:

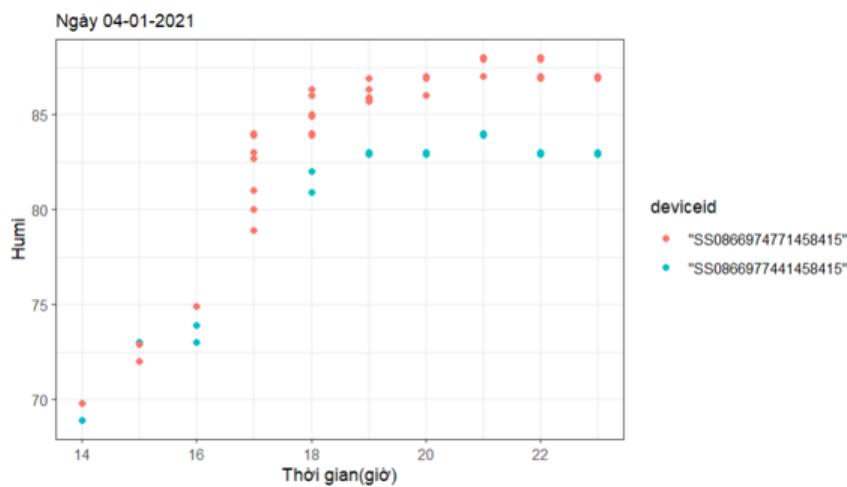
Thực hiện vẽ đồ thị vào các ngày 2,3,4,6 (nếu có)

Ta thực hiện các bước như sau:

- Tạo một Data-frame chứa các dữ liệu của ngày mà ta muốn vẽ biểu đồ bằng cách dùng chức năng Filter trên thanh công cụ R, ta xác định được nhóm dữ liệu theo từng ngày.
- Dùng hàm ggplot trong gói ggplot-2 để vẽ biểu đồ với trục Ox là thời gian trong 1 ngày và trục Oy thể hiện giá trị Humi/Temp theo ngày đó.

```
+ Date_4f3<-file_3[1:158,]
+ attach(Date_4f3)
+ Date_4f33 = ggplot(data=Date_4f3, aes(x=hour(date), y=Humi))
+ Date_4f33 + geom_point(aes(col=deviceid)) +
  labs(title = "", subtitle = "Ngày 04-01-2021",
  x = "Thời gian(giờ)", y = "Humi") + theme_bw()
```

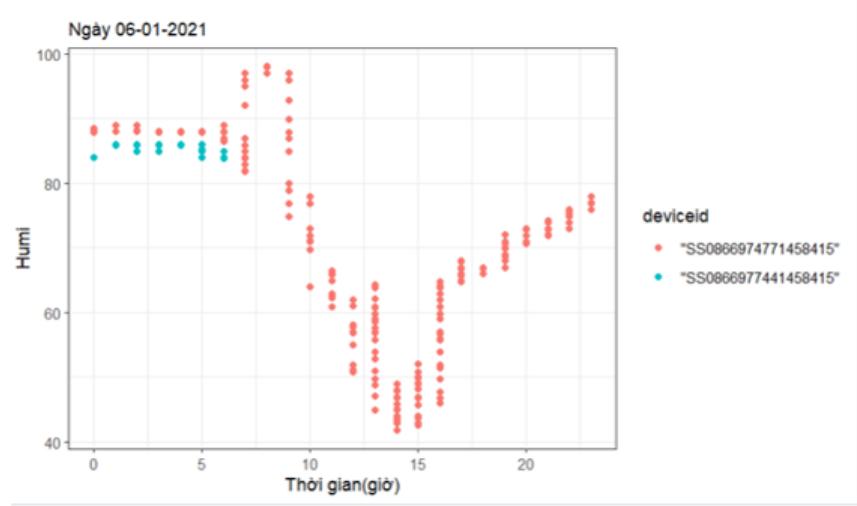
- Cuối cùng ta được biểu đồ như bên dưới:



- Tương tự, biểu đồ tương quan của Temp/Humi theo các ngày còn lại như sau:

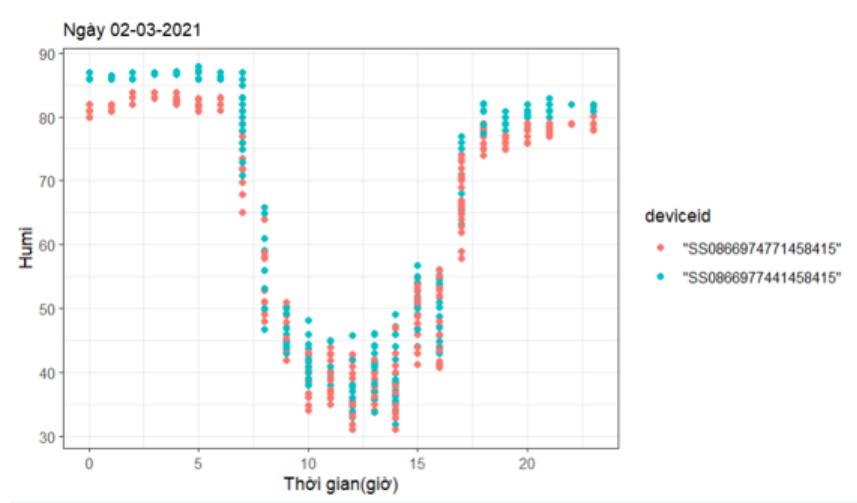
```
Date_6f3<-file_3[760:1218,]
Date_6f33 = ggplot(data=Date_6f3,aes(x=hour(date), y=Humi))
Date_6f33 + geom_point(aes(col=deviceid)) +
labs(title = "", subtitle = "Ngày 06-01-2021",x ="Thời gian(giờ)",y = "Humi")
+ theme_bw()
```

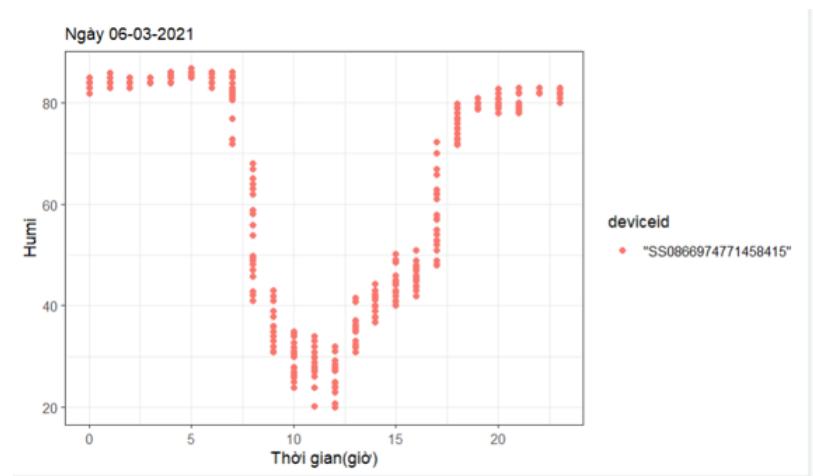
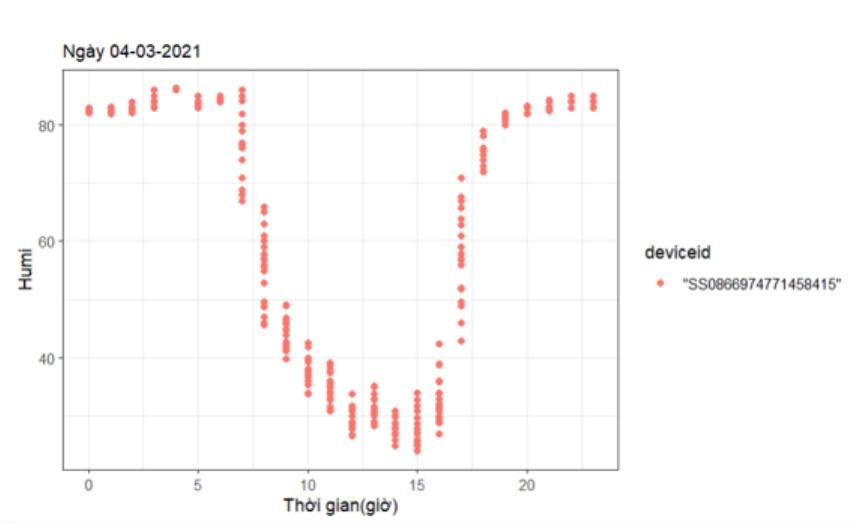
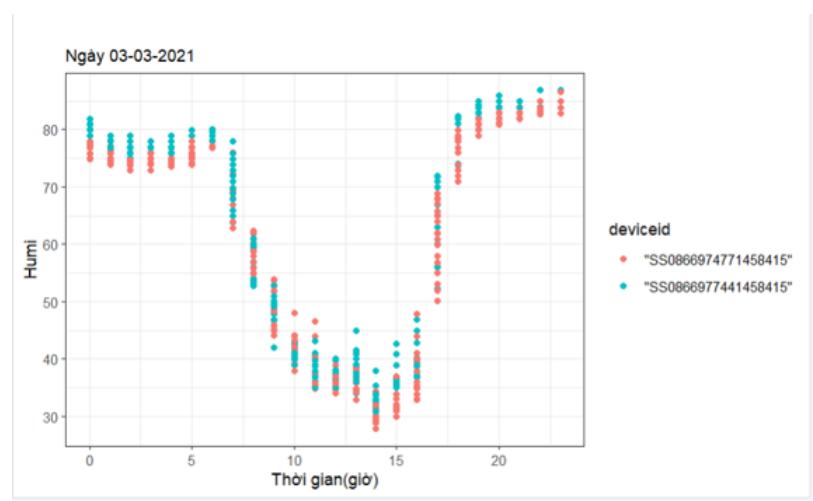
-Cuối cùng ta được biểu đồ tương quan giữa giá trị Humi và thời gian trong ngày 06-01-2021 như bên dưới:



#### FILE 4

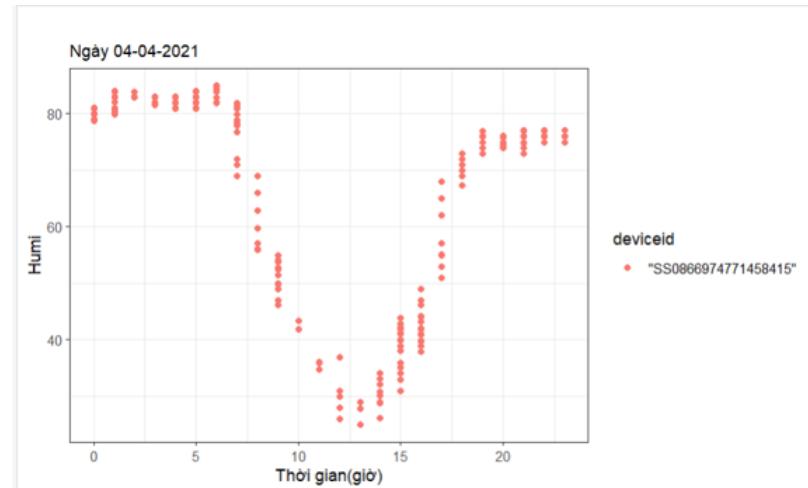
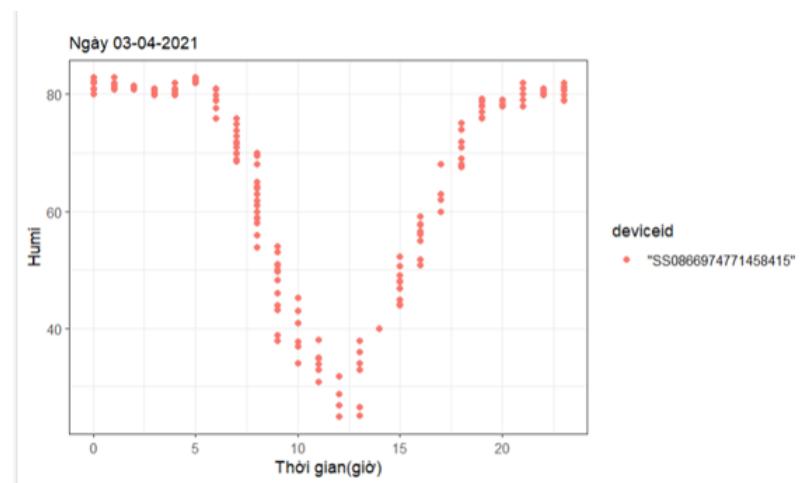
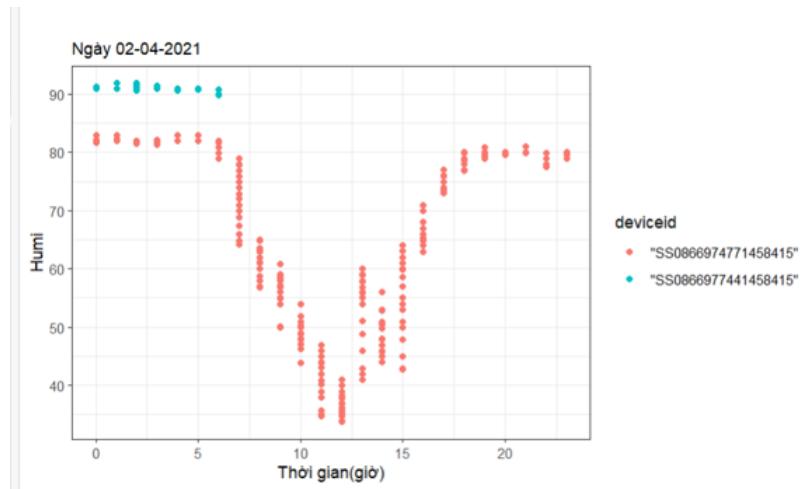
Thực hiện hoàn toàn tương tự như đối với File 3, ta lần lượt vẽ biểu đồ thể hiện tương quan của giá trị Humi với thời gian(giờ) của lần lượt các ngày 2,3,4,6/03.  
Kết quả của 4 ngày lần lượt được trình theo thứ tự hình bên dưới .

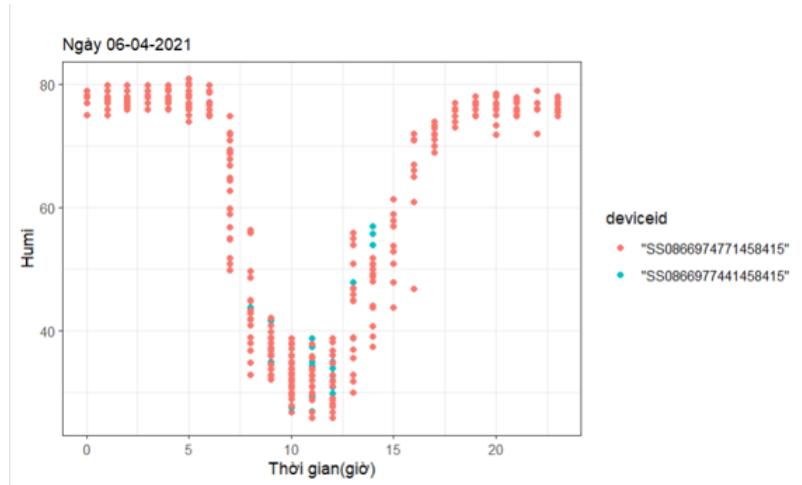




FILE 5

Thực hiện tương tự như đối với 2 file trên, ta thu được kết quả là biểu đồ thể hiện tương quan của Humi với thời gian(giờ) của lần lượt các ngày 2,3,4,6/04





3 Biểu đồ thu thập gồm cả hai nhiệt độ và độ ẩm theo thời gian là ngày của từng thiết bị

FILE 3:

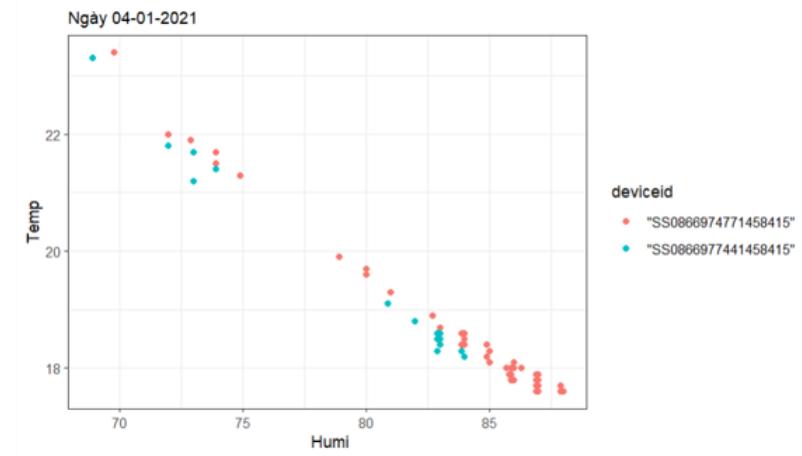
Thực hiện vẽ đồ thị vào các ngày 4,6 (File 3 không có dữ liệu ngày 2 và 3)

Ta thực hiện các bước như sau:

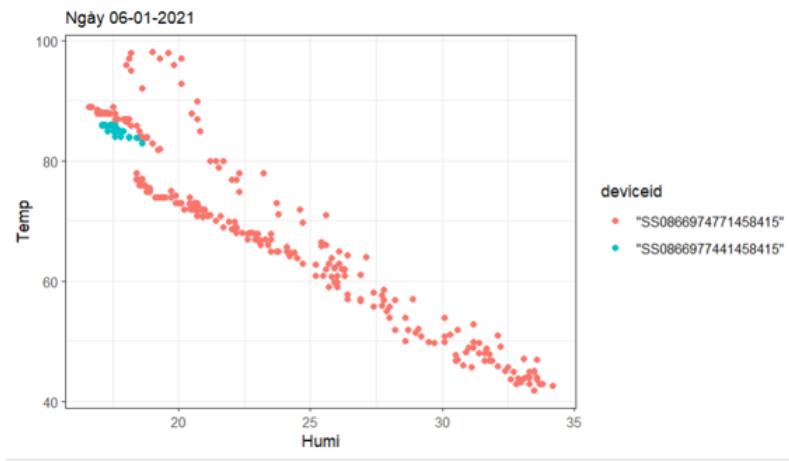
- Tạo một Data-frame chứa các dữ liệu của ngày mà ta muốn vẽ biểu đồ bằng cách dùng chức năng Filter trên thanh công cụ R, ta biết được nhóm dữ liệu theo ngày.
- Dùng hàm ggplot trong gói package ggplot2 để vẽ biểu đồ với trục Ox là Humi, Oy là Temp theo ngày đó.

```
Date_4f333 = ggplot(data=Date_4f3, aes(x=Humi, y=Temp))
Date_4f333 + geom_point(aes(col=deviceid))
labs(title = "", subtitle = "Ngày 04-01-2021",x =" Humi",y = "Temp") + theme_bw()
```

Kết quả thu được biểu đồ như sau:

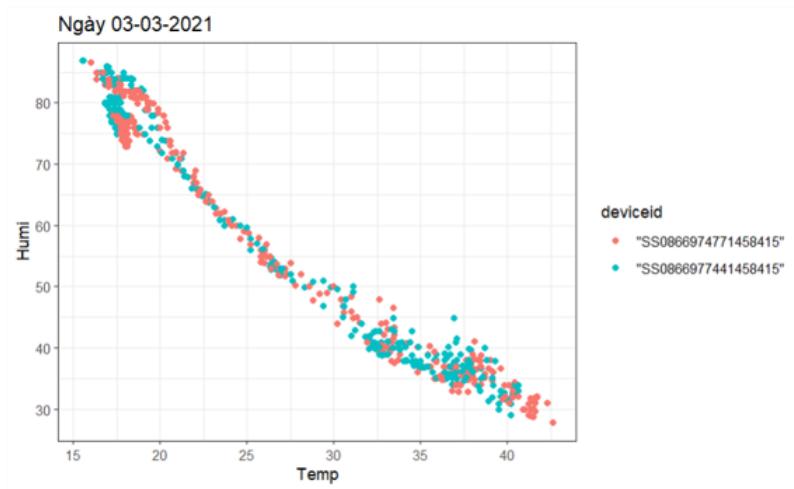
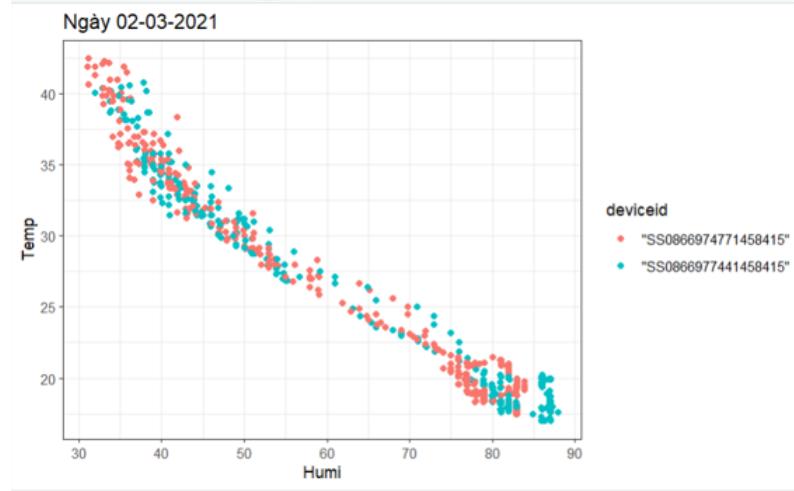


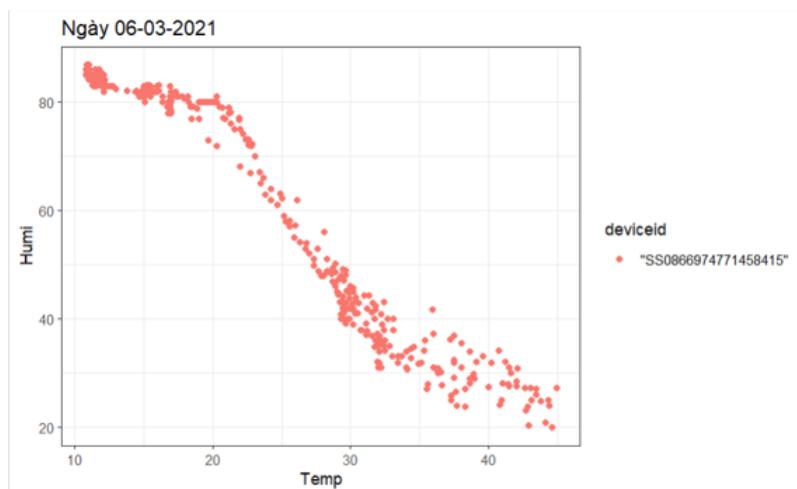
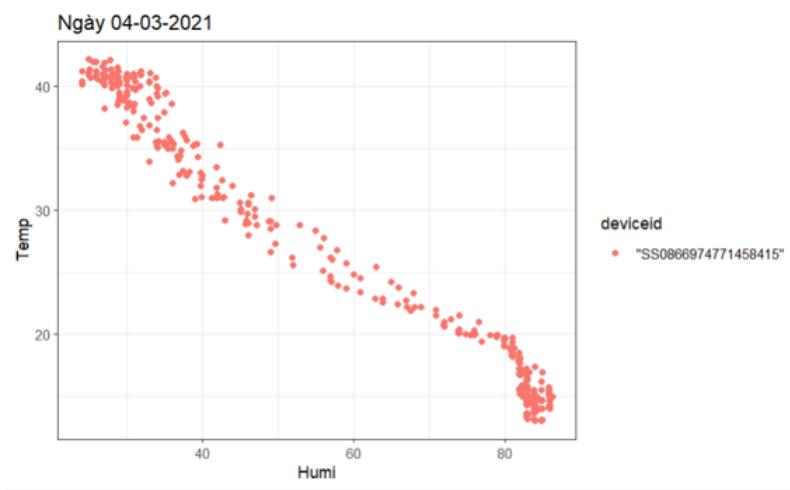
Tương tự, ta cũng vẽ được biểu đồ thể hiện tương quan trong ngày 06-01-2021



FILE 4

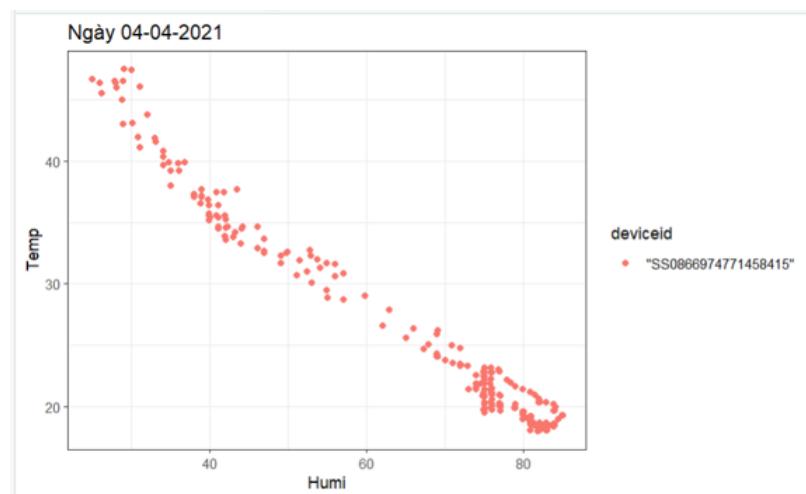
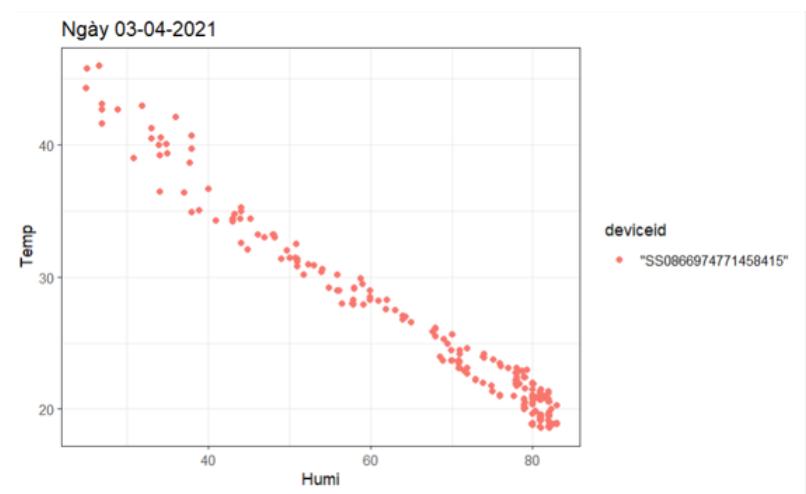
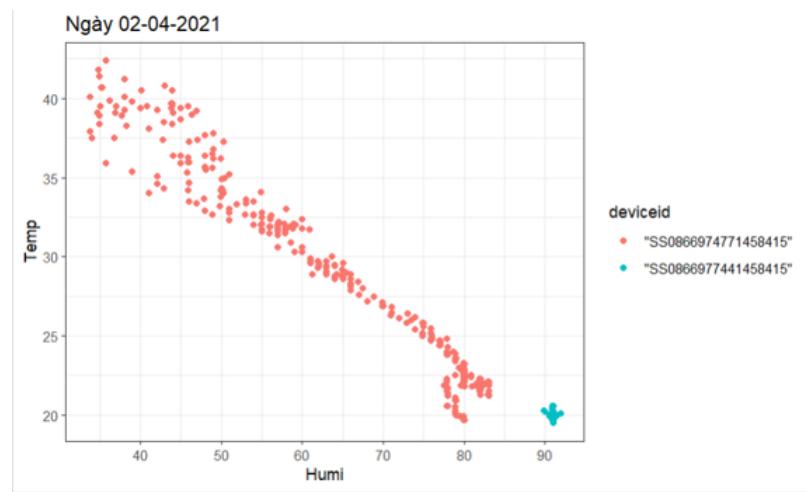
Thực hiện tương tự như đối với file 3, ta cũng vẽ được biểu đồ thể hiện tương quan của nhiệt độ và độ ẩm theo thời gian

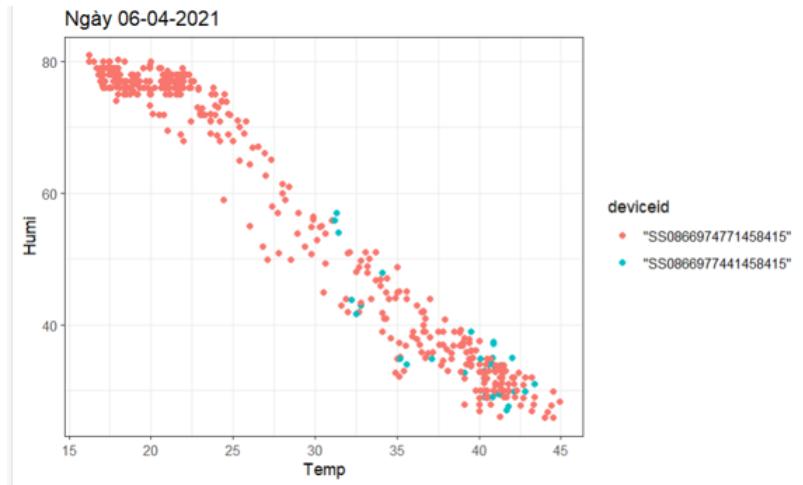




FILE 5

Thực hiện tương tự như đối với file 3,4 ta cũng vẽ được biểu đồ thể hiện tương quan của nhiệt độ và độ ẩm theo thời gian.





4 Vẽ biểu đồ tần số từng ngày của tập dữ liệu

FILE-3:

Thực hiện vẽ đồ thị vào các ngày 4,6 (Tập dữ liệu tháng 1 bắt đầu từ ngày 04-01-2021)

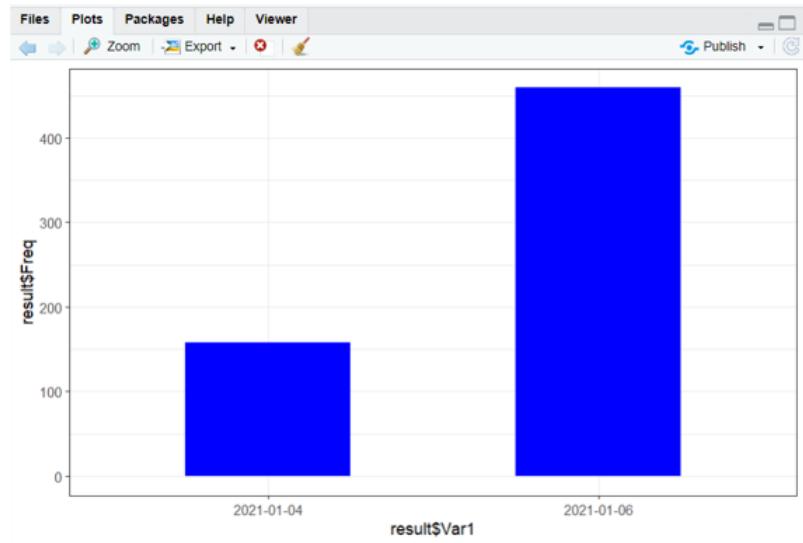
Ta thực hiện các bước như sau:

- Dùng hàm as.Date để chiết xuất số dữ liệu theo ngày từ bảng số liệu.
- Tạo data frame mới với điều kiện là ngày mà ta cần thống kê tần số.
- Dùng hàm vẽ đồ thị Barplot để thể hiện tần số của tập dữ liệu theo ngày trong tháng 01-2021.

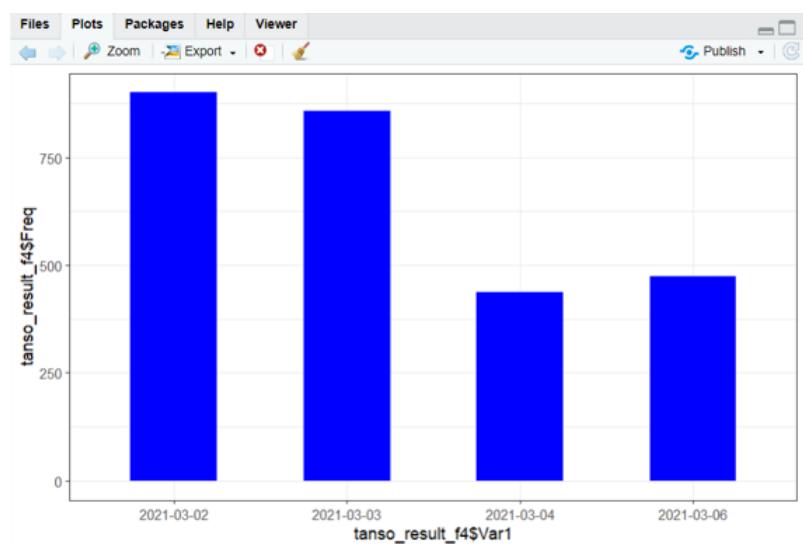
```
attach(file_3)
tansof333<- table(as.Date(date))
tansof3333<-data.frame(tansof333)
tansof3333
filter_data4<-tansof3333[tansof3333$Var1 == "2021-01-04",]
filter_data4
filter_data6<-tansof3333[tansof3333$Var1=="2021-01-06",]
filter_data6
tanso_result <- merge(filter_data4,filter_data6, all=TRUE)
tanso_result

ggplot(data=tanso_result, aes(y=result$Freq, x=result$Var1)) +
  geom_bar(stat="identity", width=0.5, fill="blue") + theme_bw()
```

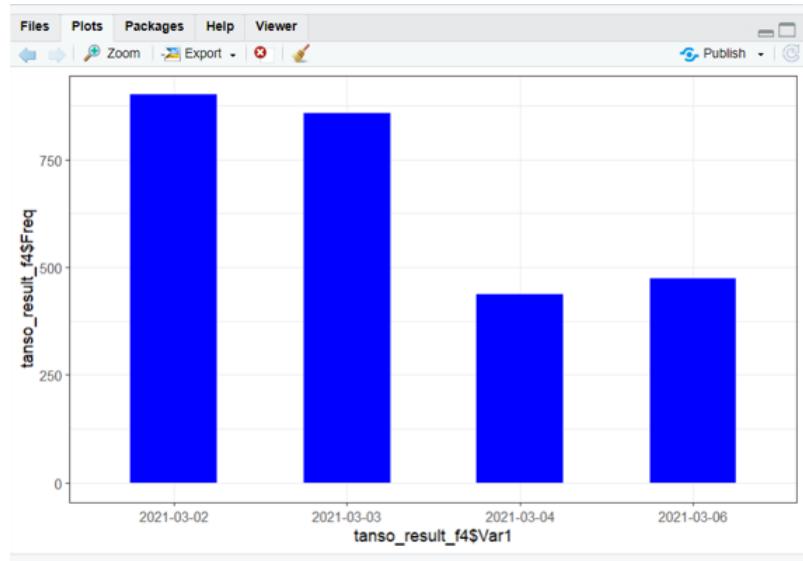
Ta thu được biểu đồ thể hiện tần số trong tháng 01-2021 như sau:



FILE 4: Thực hiện tương tự như đối với File 3, ta vẽ được biểu đồ thể hiện tần số xuất hiện của dữ liệu theo ngày trong tháng 04-2021.



FILE-5: Thực hiện tương tự như đối với File 3 và file 4, ta vẽ được biểu đồ thể hiện tần số xuất hiện của dữ liệu theo ngày trong tháng 05-2021



5 Vẽ biểu đồ tần số tích lũy ngày của tập dữ liệu  
FILE 3:

Để vẽ biểu đồ tần số tích lũy ngày của tập dữ liệu, ta thực hiện các bước sau:

- Dùng hàm Table kết hợp với as.Date để tìm ra tần số tích lũy của từng ngày. Gán kết quả cho x1.
- Tạo một ma trận bất kỳ x2 có độ dài trùng với số ngày của tập dữ liệu.
- Dùng vòng lặp For để thực hiện gán các giá trị của bảng tần số tìm được ở bước 1 cho ma trận ở bước 2.
- Tiếp tục dùng lặp For để cộng dồn giá trị tần số tích lũy qua từng ngày của tập dữ liệu.
- Chuyển đổi các ma trận sang kiểu dữ liệu Data Frame sau đó gộp lại bằng hàm cbind(), rồi gán kết quả cho biến Tanso-Tichluy.
- Dùng hàm ggplot để vẽ đồ thị đường thể hiện tần số tích lũy.

Chi tiết thể hiện như bên dưới.

```
attach(FILE_3)
x1 <- table(as.Date(date))
x1<-data.frame(x1)
x2<-rep(1,28)

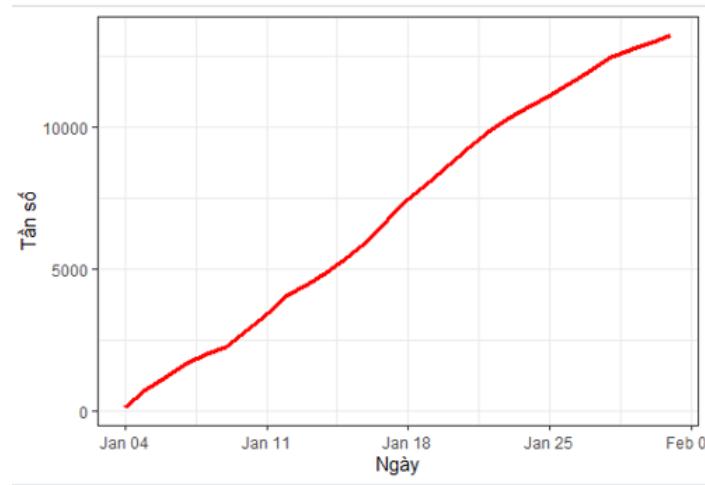
for(i in 1:28){
    x2[i] <- x1$Freq[i]
}
for(i in 1:28){
    x2[i+1] = x2[i+1] + x2[i]
}

x3<-data.frame(x1$Var1)
x2<-data.frame(x2)
x2<-x2[-c(29),]
Tanso_tichluy<- cbind(x3,x2)

Bieu_do_tan_so<-ggplot(data=Tanso_tichluy, aes(x=as.Date(Tanso_tichluy$x1.Var1),
, y = x2))

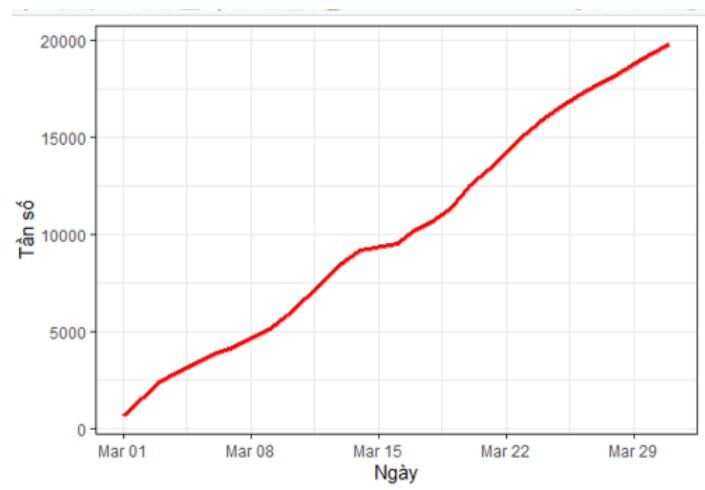
Bieu_do_tan_so + geom_line(col="red", size=1.2) + labs(x="Ngày", y="Tần số")
+ theme_bw()
```

Cuối cùng, ta được đồ thị thể hiện tần số tương đối tích lũy theo ngày của tập dữ liệu như sau:



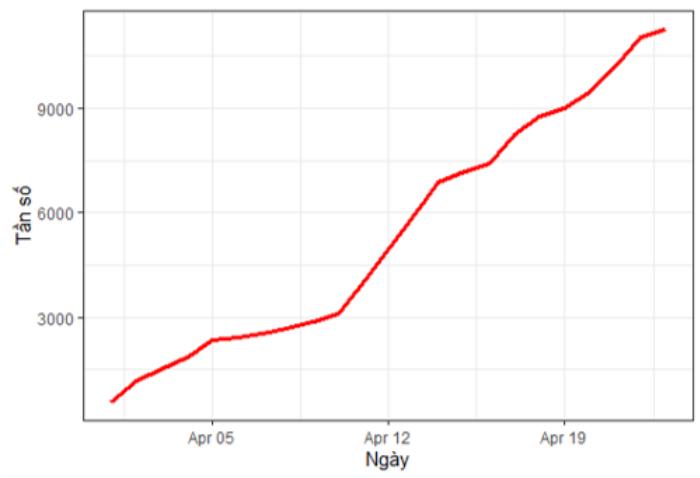
FILE 4:

Thực hiện hoàn toàn tương tự ở file 3, ta cũng vẽ được biểu đồ tần số tích lũy cho tập dữ liệu.



FILE 5:

Biểu đồ tần số tích lũy tương đối theo ngày của File 5 là:



## 6 Vẽ phổ giờ mà outliers xuất hiện trên từng thiết bị

Thực hiện các bước như sau:

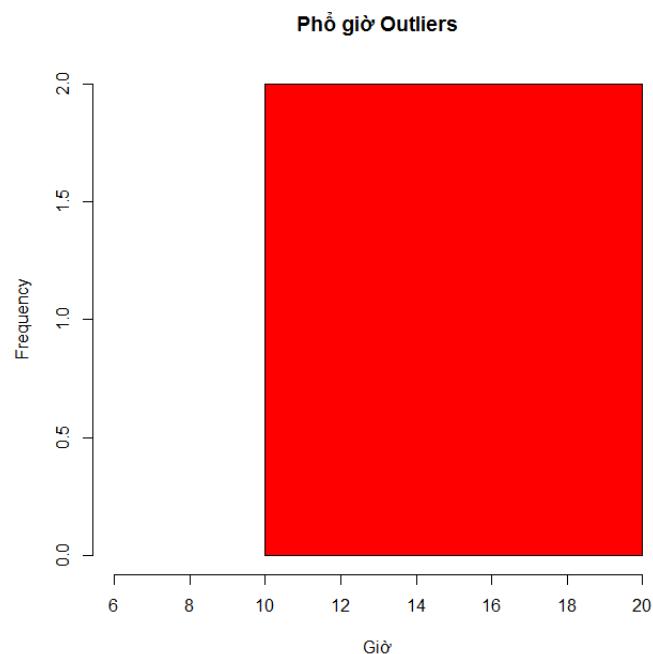
- Cài đặt gói Lubridate để lọc dữ liệu theo giờ, qua hàm hour.
- Phân loại và thu thập dữ liệu của mỗi thiết bị dựa trên deviceid
- Dùng hàm boxplot.stats() để tìm ra các giá trị outliers trong tập dữ liệu.
- Dùng hàm filter() để lọc các mốc thời gian có các giá trị outliers tương ứng đã tìm được.
- Quy đổi dữ liệu thời gian vừa tìm được ra giờ.
- Lặp 1 véc-tơ lặp số 1 với số lần lặp là số mốc thời gian (giờ) mà ta vừa tìm được.
- Gộp 2 véc-tơ ở bước 4 và 5 thành 1 véc tơ, sau đó dùng hàm vẽ đồ thị HITS() để vẽ biểu đồ thanh thể hiện sự phân bố của phổ giờ theo số lần xuất hiện của giá trị outliers tại thời điểm đó.

```
data <- read.csv("3-01_2021.csv")
data <- read.csv("4-03_2021.csv")
data <- read.csv("5-04_2021.csv")
data_d = data
data_d = data[data$deviceid == "\\"SS0866974771458415\\",]#1
data_d = data[data$deviceid == "\\"SS0866977441458415\\",]#2
data_d = data[data$deviceid == "\\"SS0466973631458415\\",]#3

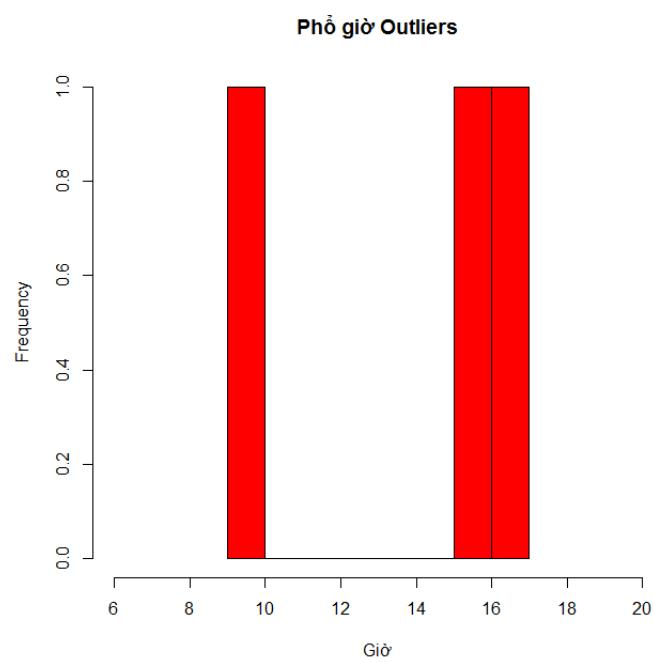
library("dplyr", lib.loc="C:/Program Files/R/R-4.1.1/library")
library("lubridate", lib.loc="C:/Program Files/R/R-4.1.1/library")
hala<-boxplot.stats(data_d$Temp)$out
check<-data.frame(hala)

result_time<-data_d%>%filter(data_d$Temp %in% check$hala)
new<-hour(result_time$date)
hala<-data.frame(new)

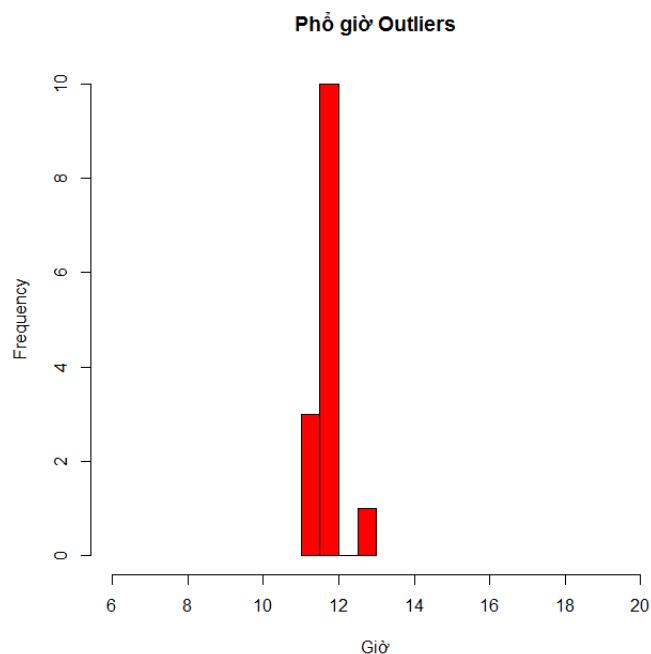
vamos<-data.frame(hala)
hist(vamos$new, xlim=c(6,20),breaks=6,col="red",
     main="Phổ giờ Outliers",xlab="Giờ")
```



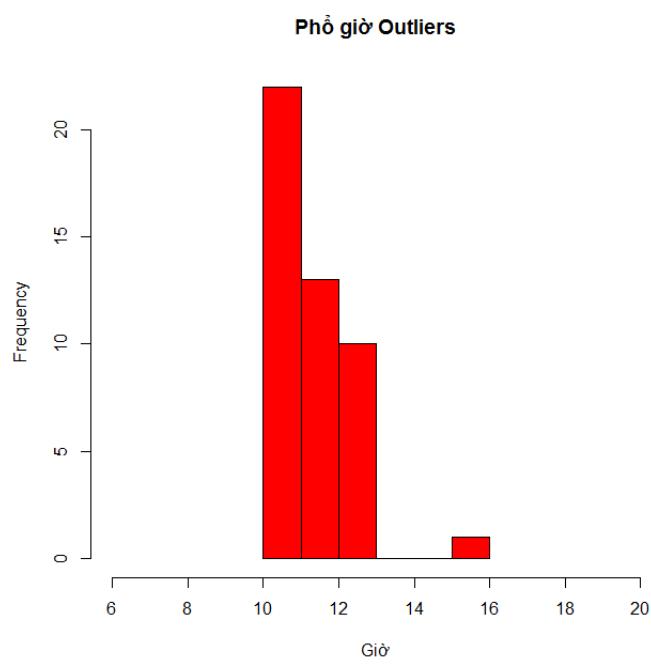
outlier thiết bị 1 file 3



outlier thiết bị 3 file 4



outlier thiết bị 2 file 5



outlier thiết bị 3 file 5

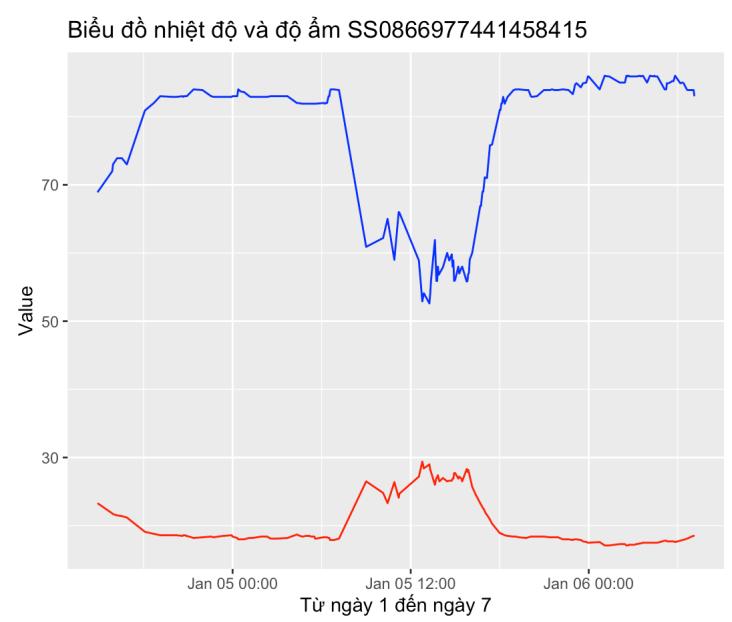
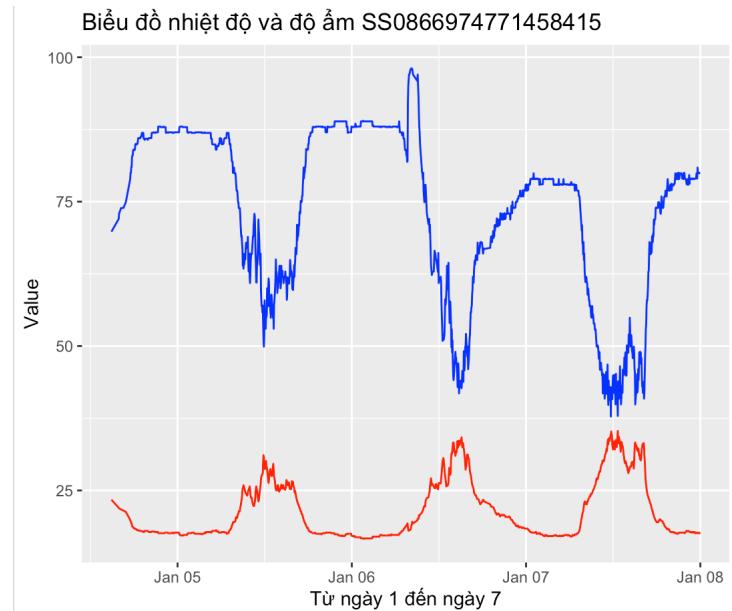


v) Nhóm câu hỏi liên quan đến trực quan dữ liệu theo khoảng thời gian: (từ ngày - đến ngày)

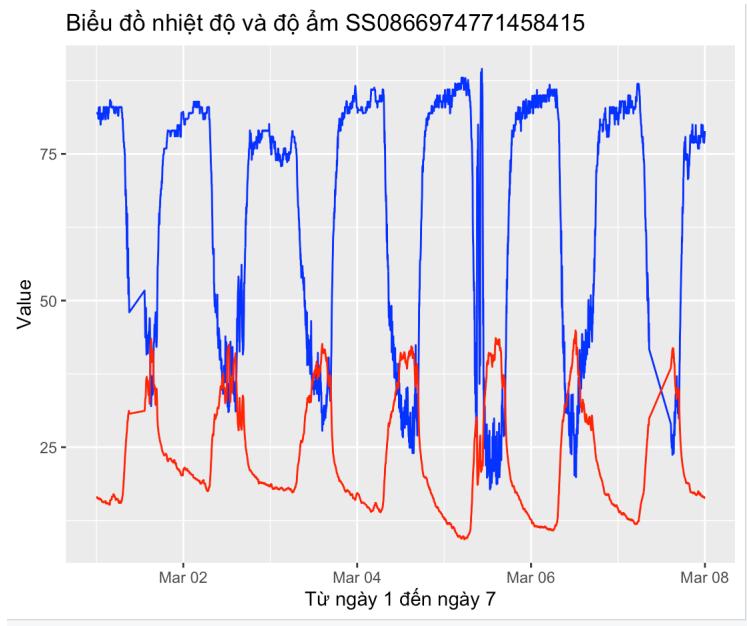
```
library(dplyr)
library(ggplot2)
#Tạo các biến tuan1, tuan2, tuan3, tuan4 để lọc dữ liệu theo ngày
cau5=read.csv("Documents/5-04_2021.csv")
count_devices=data.frame(table(cau5$deviceid))
l=data.frame(cau5%>%select(date,deviceid,Temp,Humi))
l$date=strptime(l$date,"%Y-%m-%d %H:%M:%S")
tuan1=l%>%filter(as.integer(format(as.Date(date), "%d"))>=1 &
as.integer(format(as.Date(date), "%d"))<=7)
tuan2=l%>%filter(as.integer(format(as.Date(date), "%d"))>=8 &
as.integer(format(as.Date(date), "%d"))<=14)
tuan3=l%>%filter(as.integer(format(as.Date(date), "%d"))>=15 &
as.integer(format(as.Date(date), "%d"))<=21)
tuan4=l%>%filter(as.integer(format(as.Date(date), "%d"))>=22 &
as.integer(format(as.Date(date), "%d"))<=28)
#sau do ve bieu do
tuan1_dv1=tuan1%>%filter(deviceid=='SS0866974771458415')%>%
ggplot(aes(as.POSIXct(date)))+geom_line(aes(y=Humi),colour="blue")+
geom_line(aes(y=Temp),colour="red")+
xlab("Từ ngày 1 đến ngày 7")+ylab("Value")+ggtitle("Biểu đồ nhiệt độ và độ ẩm SS0866974771458415")
tuan2_dv1=tuan2%>%filter(deviceid=='SS0866974771458415')%>%
ggplot(aes(as.POSIXct(date)))+geom_line(aes(y=Humi),colour="blue")+
geom_line(aes(y=Temp),colour="red")+
xlab("Từ ngày 8 đến ngày 14")+ylab("Value")+ggtitle("Biểu đồ nhiệt độ và độ ẩm SS0866974771458415")
tuan3_dv1=tuan3%>%filter(deviceid=='SS0866974771458415')%>%
ggplot(aes(as.POSIXct(date)))+geom_line(aes(y=Humi),colour="blue")+
geom_line(aes(y=Temp),colour="red")+
xlab("Từ ngày 15 đến ngày 21")+ylab("Value")+ggtitle("Biểu đồ nhiệt độ và độ ẩm SS0866974771458415")
tuan4_dv1=tuan4%>%filter(deviceid=='SS0866974771458415')%>%
ggplot(aes(as.POSIXct(date)))+geom_line(aes(y=Humi),colour="blue")+
geom_line(aes(y=Temp),colour="red")+
xlab("Từ ngày 22 đến ngày 28")+ylab("Value")+ggtitle("Biểu đồ nhiệt độ và độ ẩm SS0866974771458415")
```

- 1) Biểu đồ thu thập gồm nhiệt độ và độ ẩm theo thời gian từ ngày 1 đến ngày 7 của từng thiết bị  
File 3  
Thiết bị 1

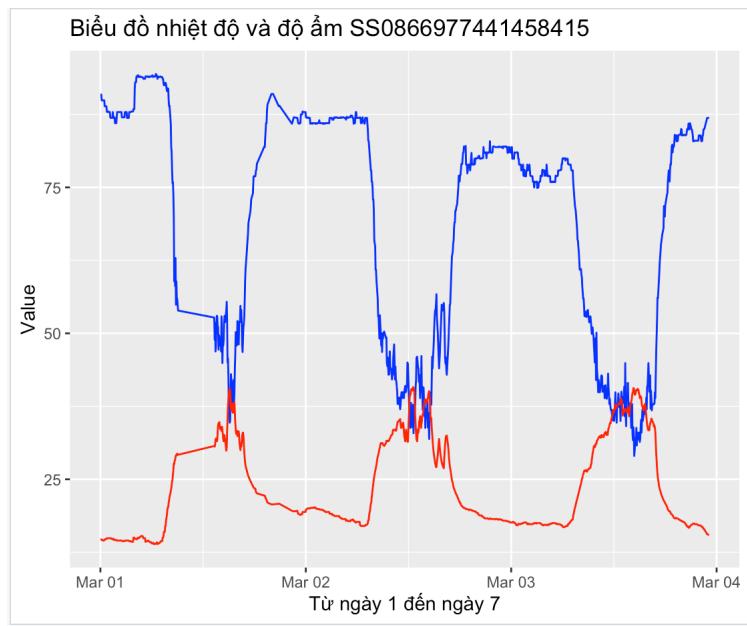
Thiết bị 2



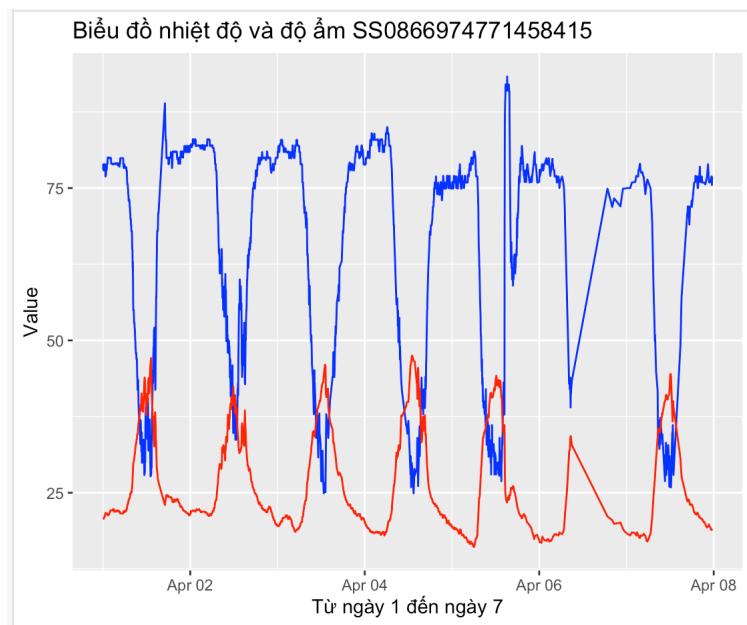
File 4  
Thiet bi 1



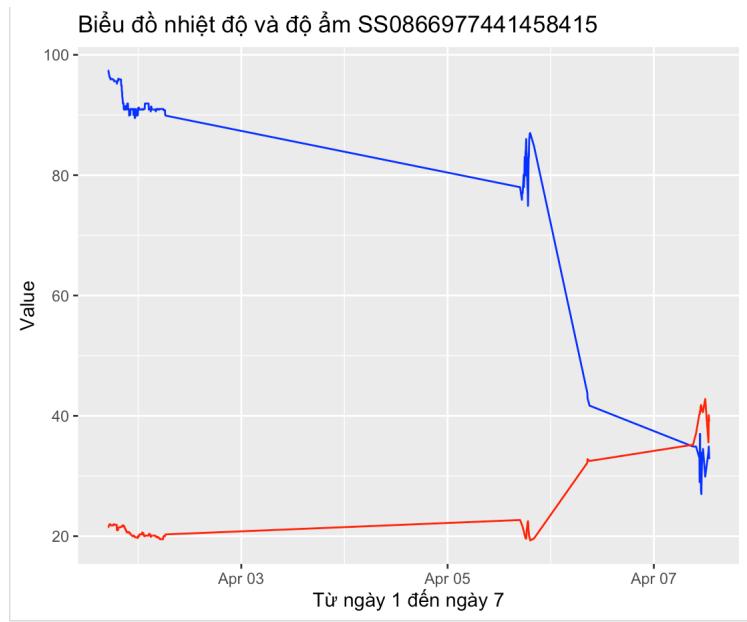
Thiet bi 2



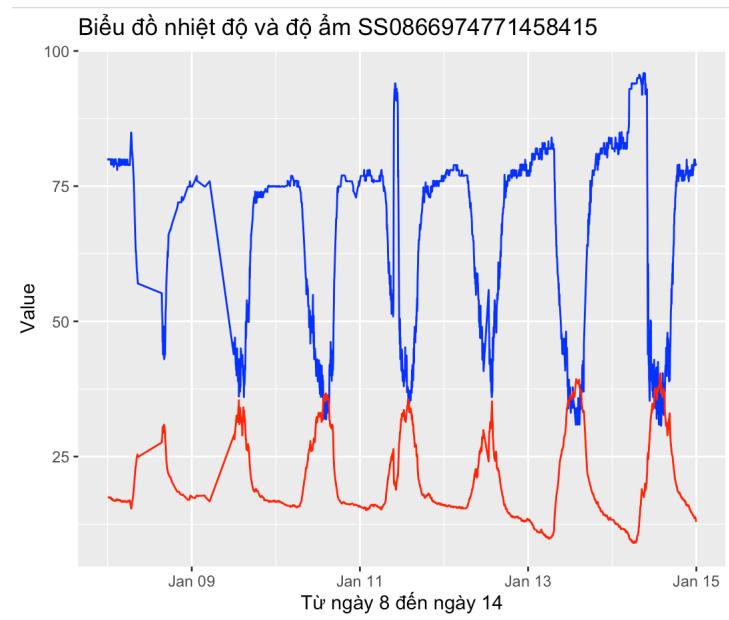
File 5  
Thiet bi 1



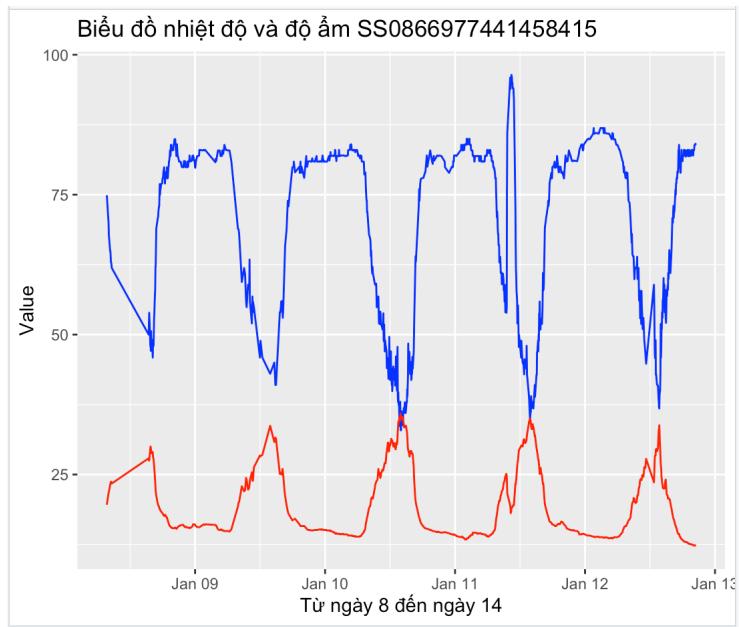
Thiet bi 2



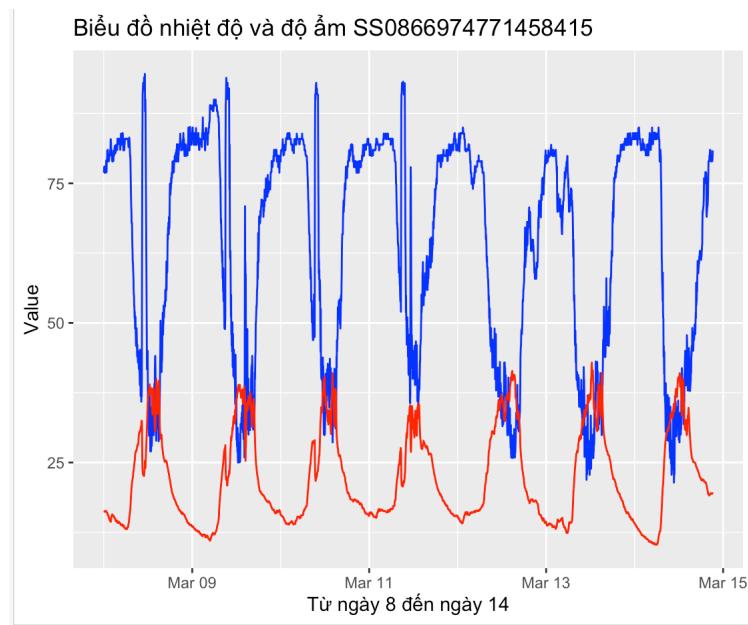
- 2) Biểu đồ thu thập nhiệt độ và độ ẩm theo thời gian từ ngày 8 đến ngày 14 của từng thiết bị  
File 3  
Thiết bị 1



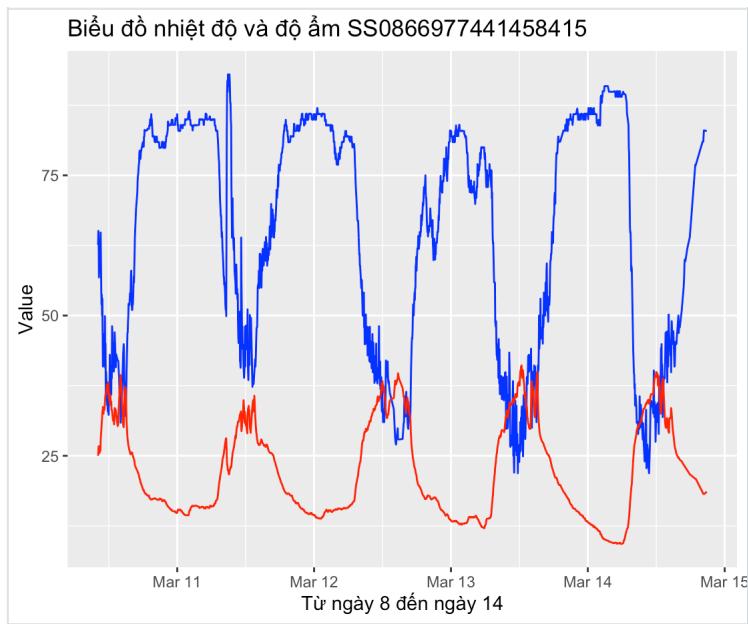
Thiết bị 2



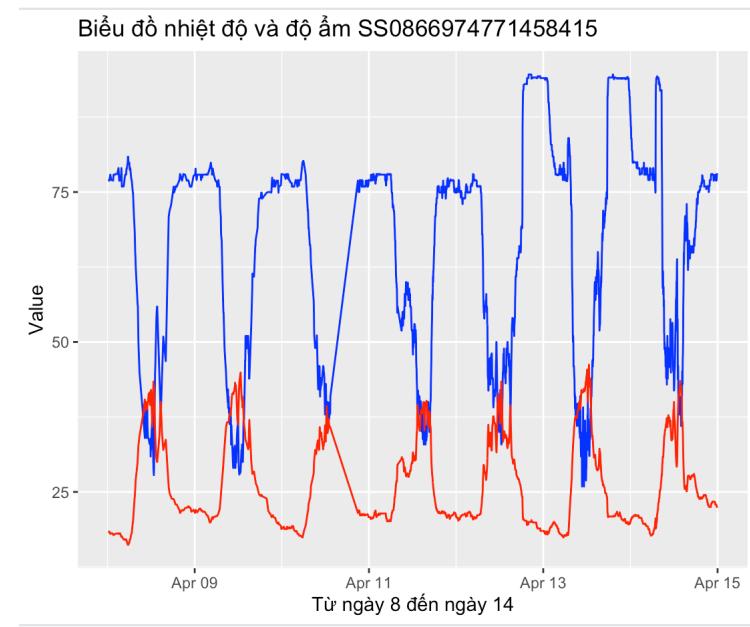
File 4  
Thiet bi 1



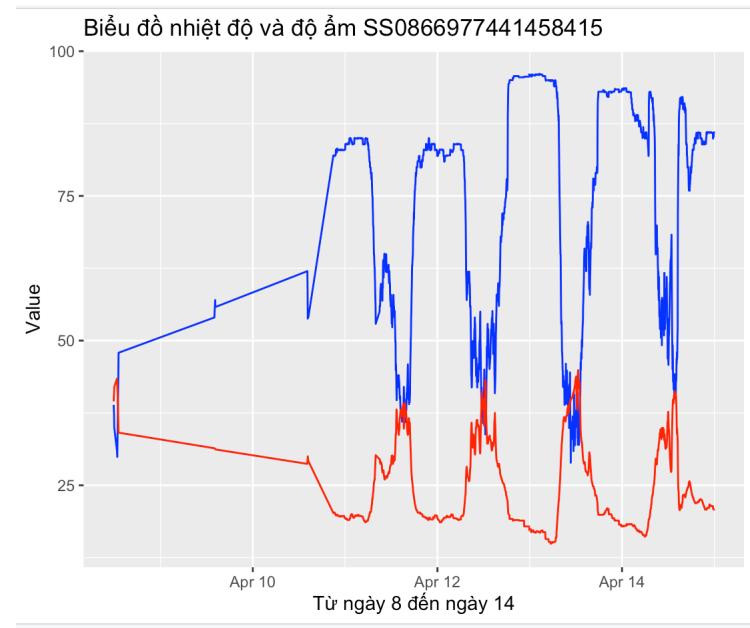
Thiet bi 2



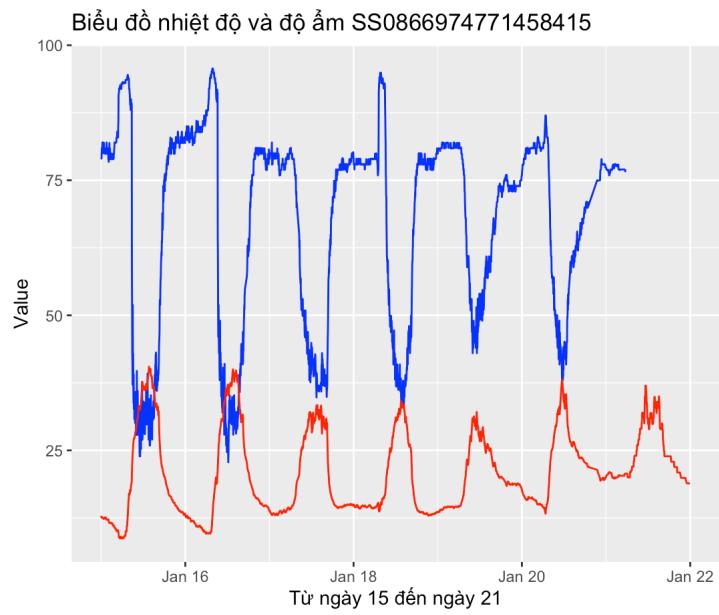
File 4  
Thiet bi 1



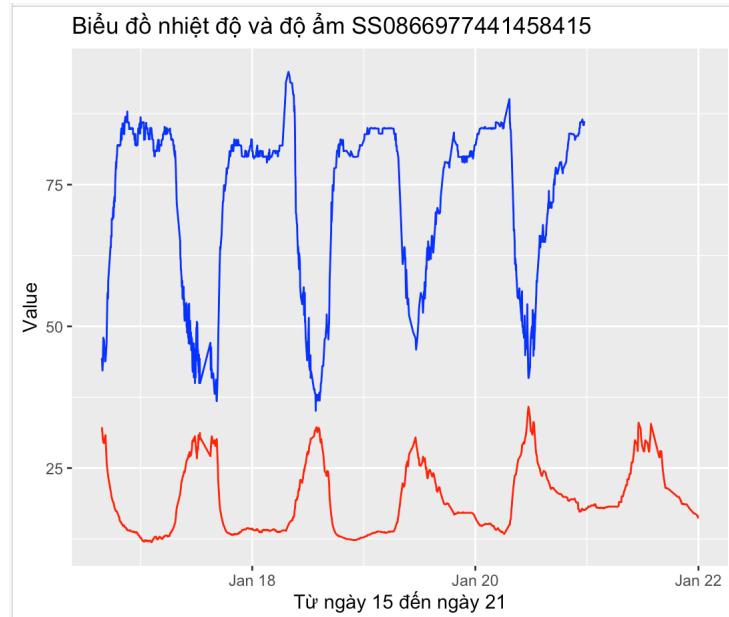
Thiet bi 2



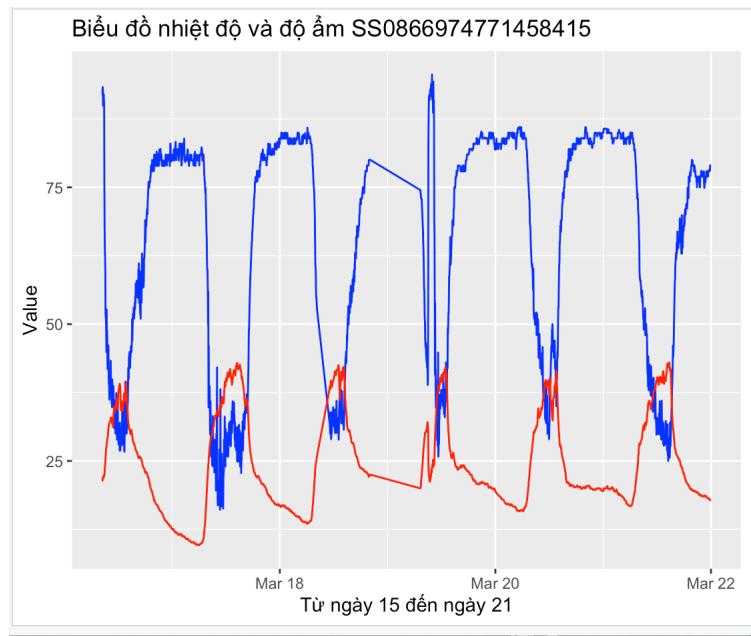
- 3) Biểu đồ thu thập gồm nhiệt độ và độ ẩm theo thời gian từ ngày 15 đến ngày 21 của từng thiết bị File 3
- Thiết bị 1



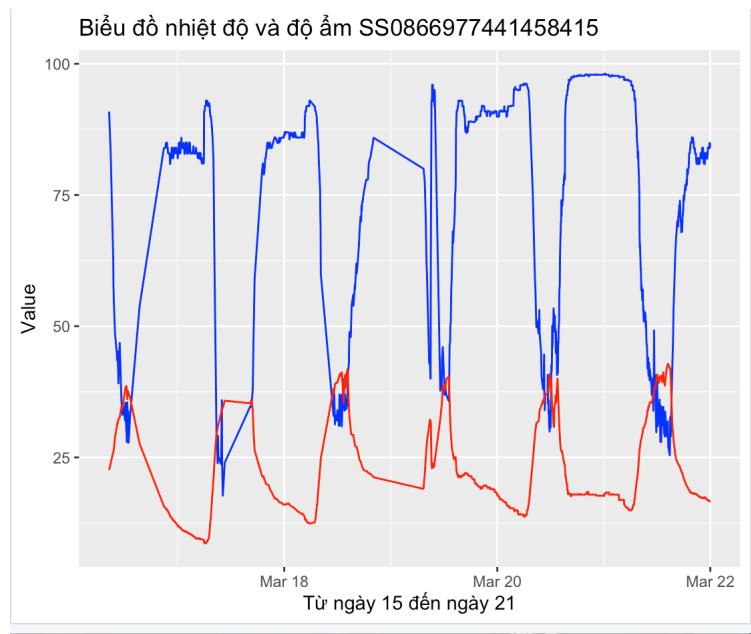
Thiết bị 2



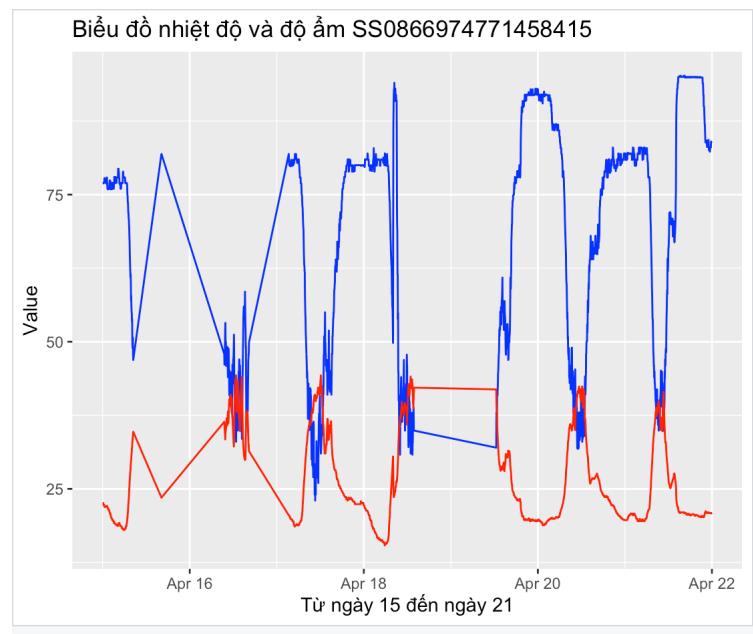
File 4  
Thiet bi 1



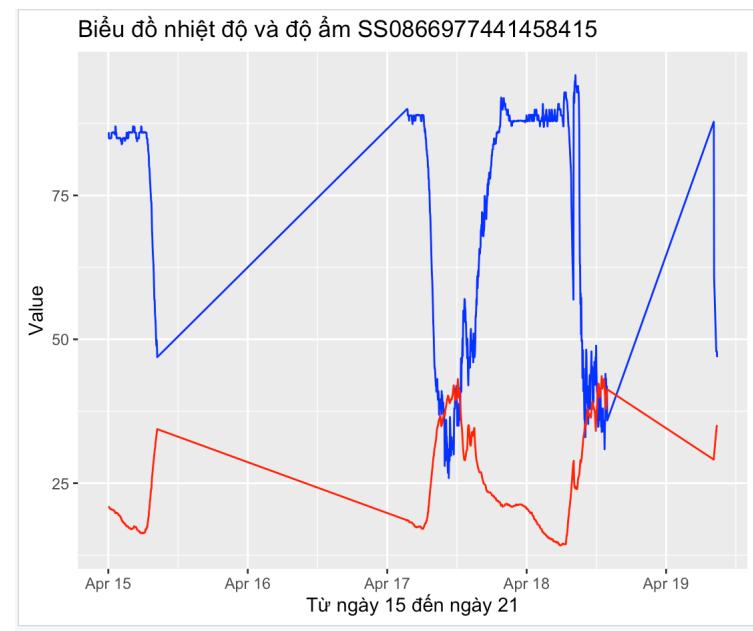
Thiet bi 2



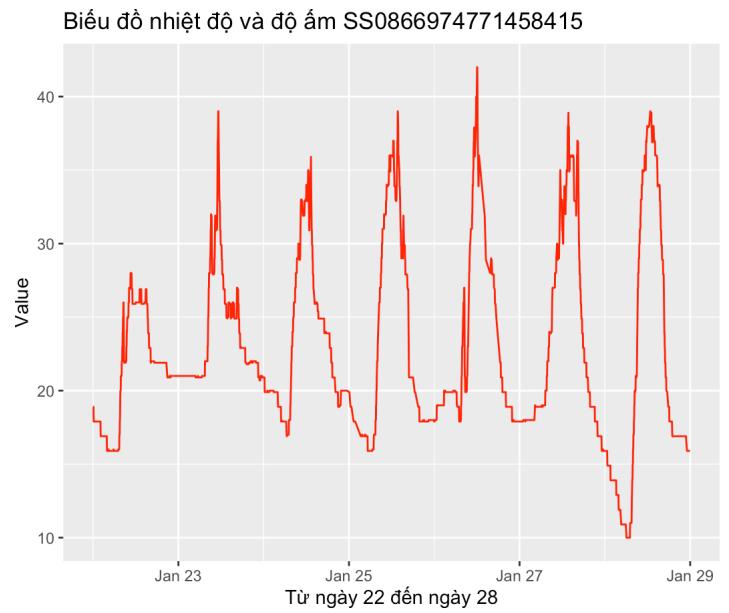
File 5  
Thiet bi 1



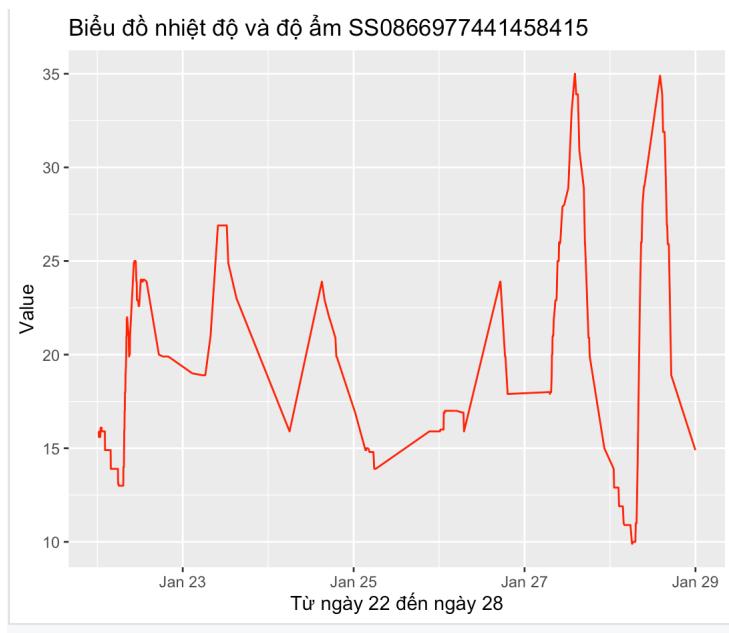
Thiet bi 2



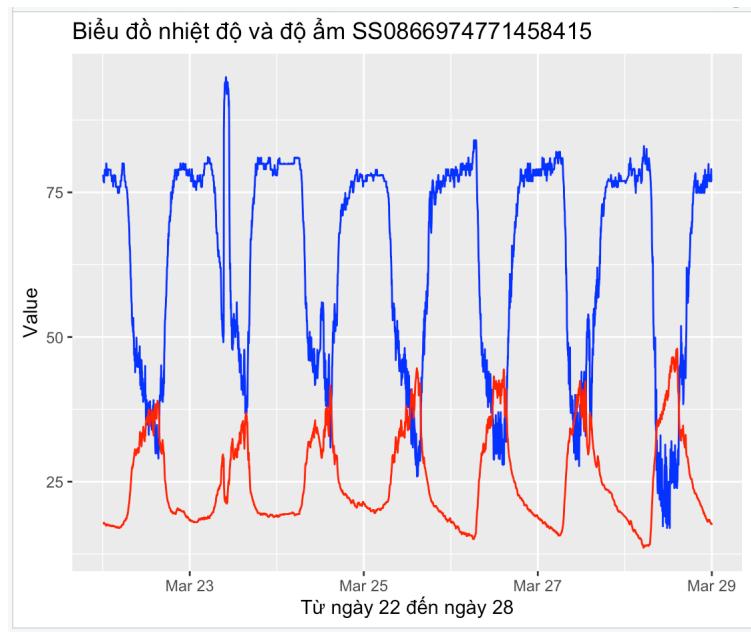
- 4) Biểu đồ thu thập gồm nhiệt độ và độ ẩm theo thời gian từ ngày 22 đến ngày 28 của từng thiết bị File 3
- Thiết bị 1



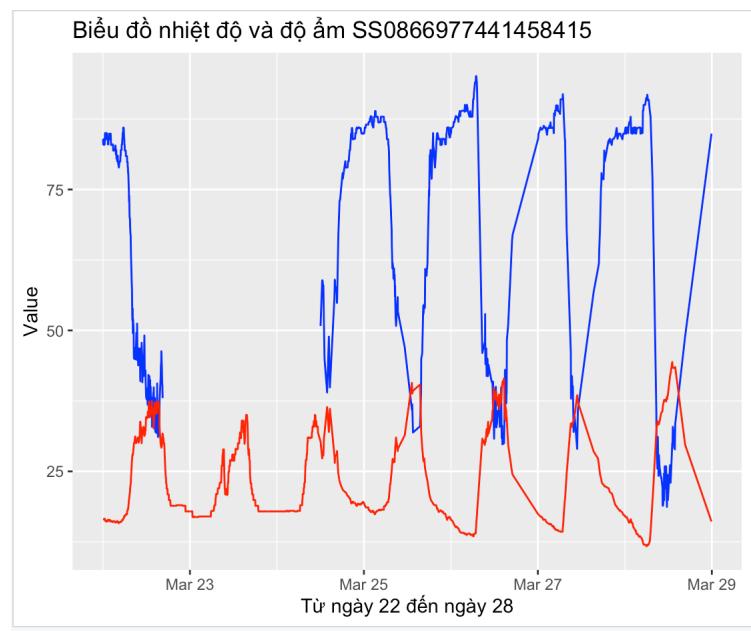
Thiết bị 2



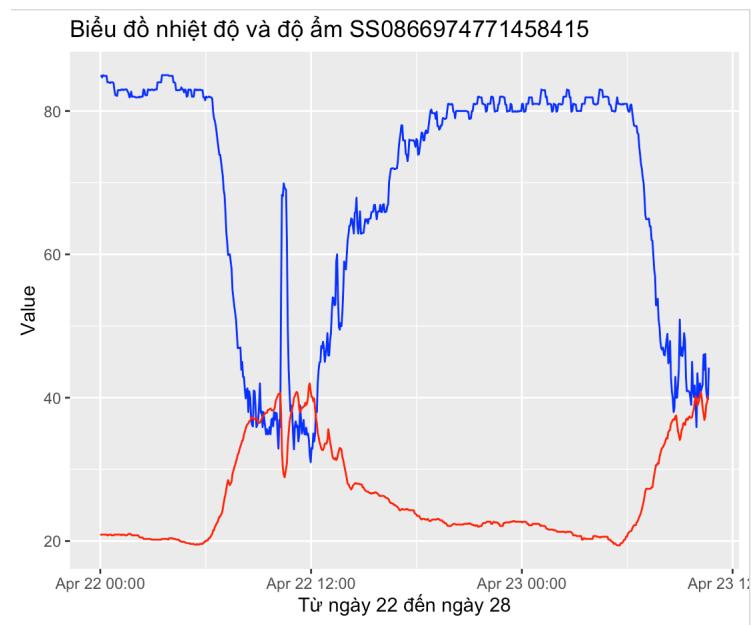
File 4  
Thiet bi 1



Thiet bi 2



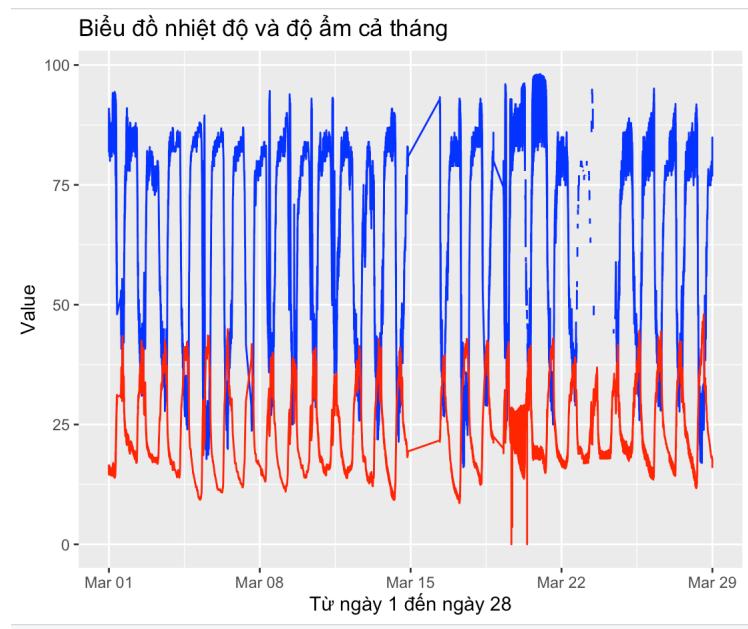
File 5  
Thiết bị 1



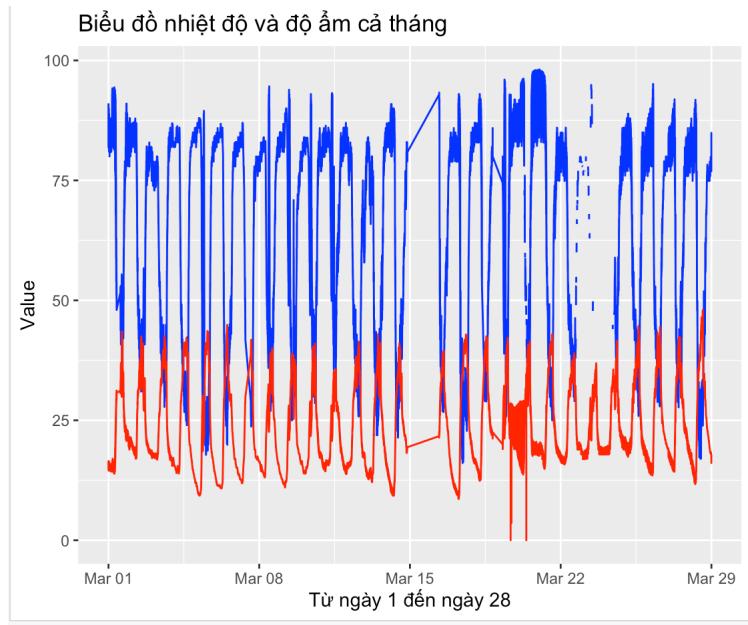
- 5) Biểu đồ thu thập gồm nhiệt độ và độ ẩm theo tháng (cả tập dữ liệu)

```
thang=1%>%filter(as.integer(format(as.Date(date), "%d"))>=1 &
as.integer(format(as.Date(date), "%d"))<=28)
thang_dv1=thang%>%filter(is.na(Humi))%>%ggplot(aes(as.POSIXct(date)))+
geom_line(aes(y=Humi), colour="blue")+geom_line(aes(y=Temp), colour="red")+
xlab("Từ ngày 1 đến ngày 28")+ylab("Value")+ggtitle("Biểu đồ nhiệt độ và độ ẩm cả tháng")
```

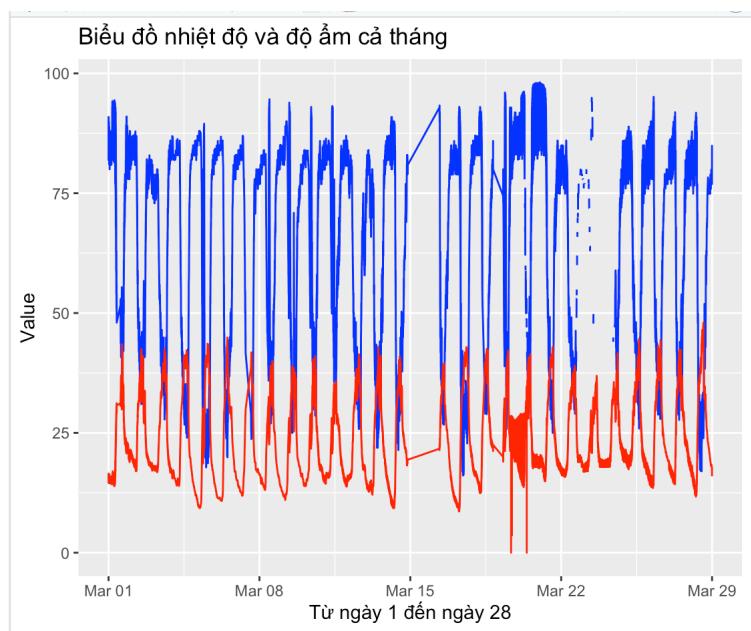
File 3



File 4



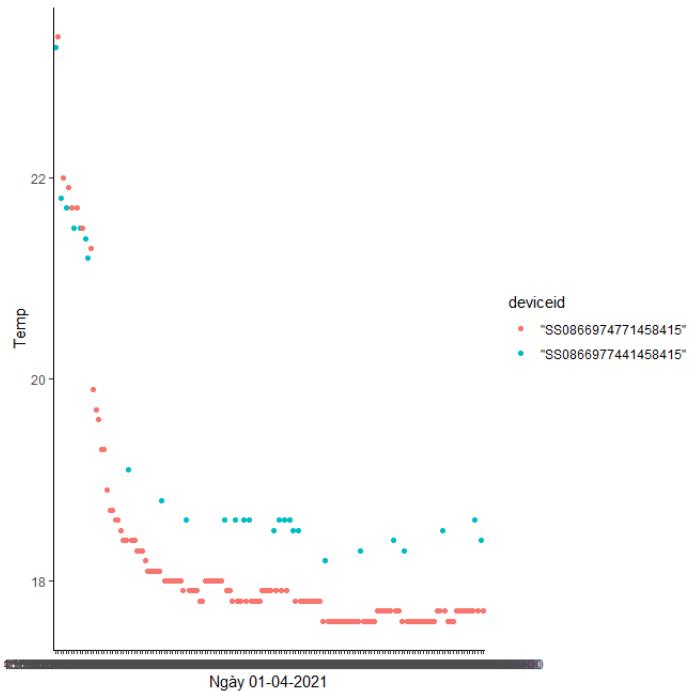
File 5

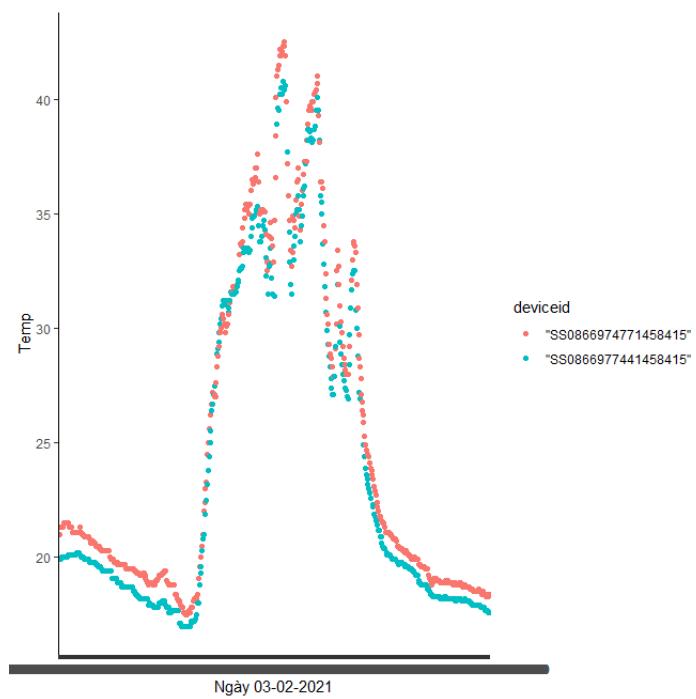
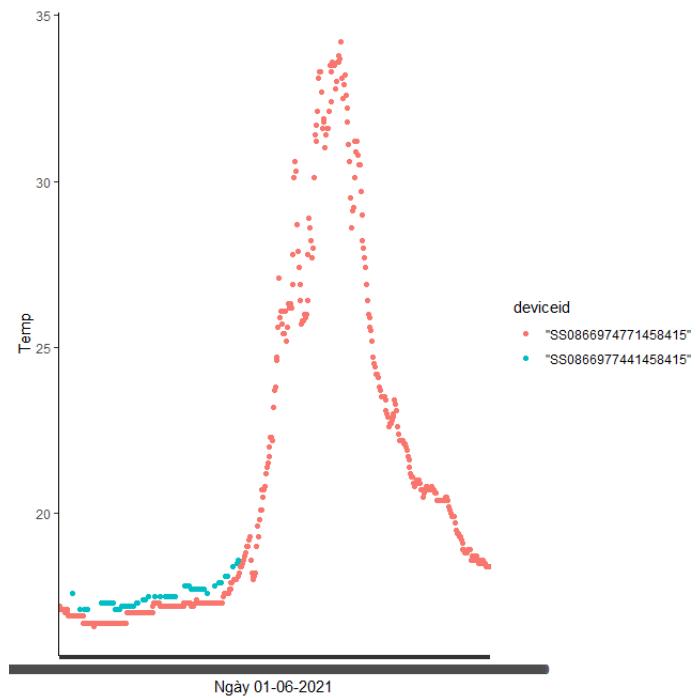


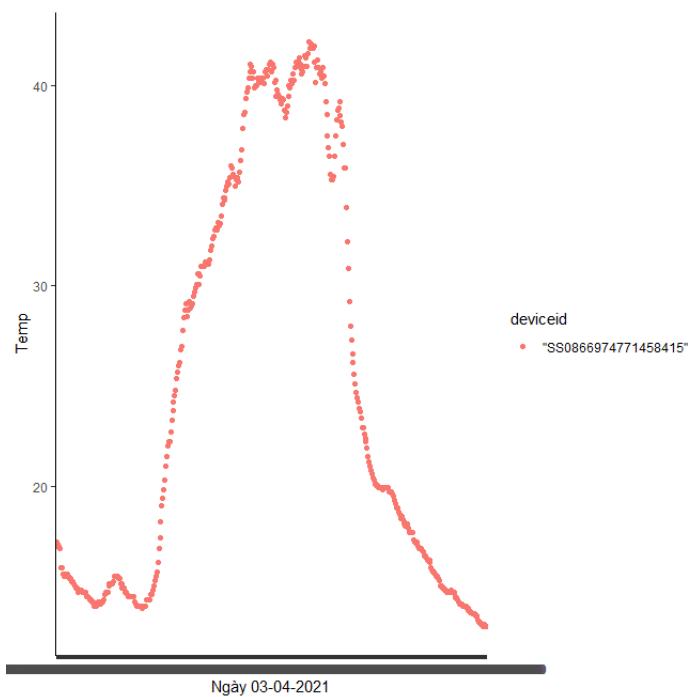
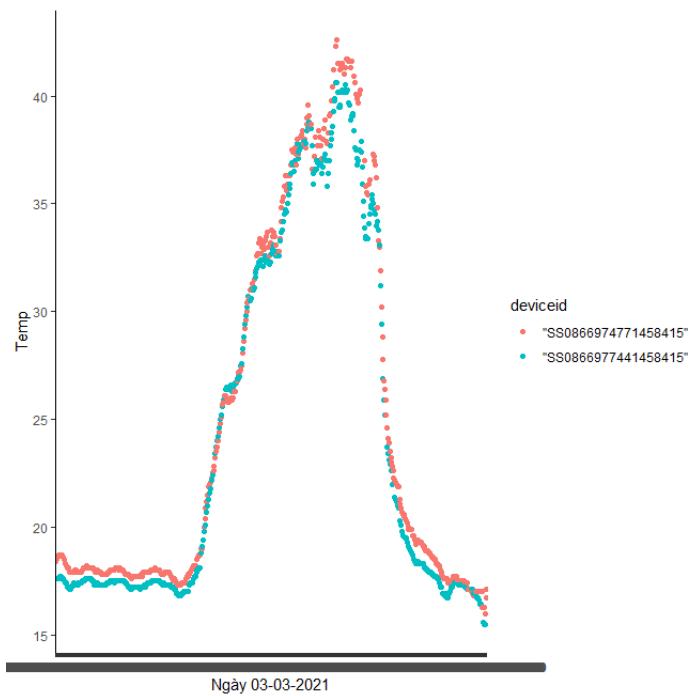
- vi) Nhóm câu hỏi liên quan đến tất cả thiết bị theo thời gian là ngày Hãy dùng 4 ký số của mã đè để vẽ 4 ngày tương ứng theo ký số đó, nếu ký số là 0 thì lấy ngày là 10. Trục Ox là thời gian, Oy là Temp/Humi, số đường thể hiện trên đồ thị là số thiết bị đo.

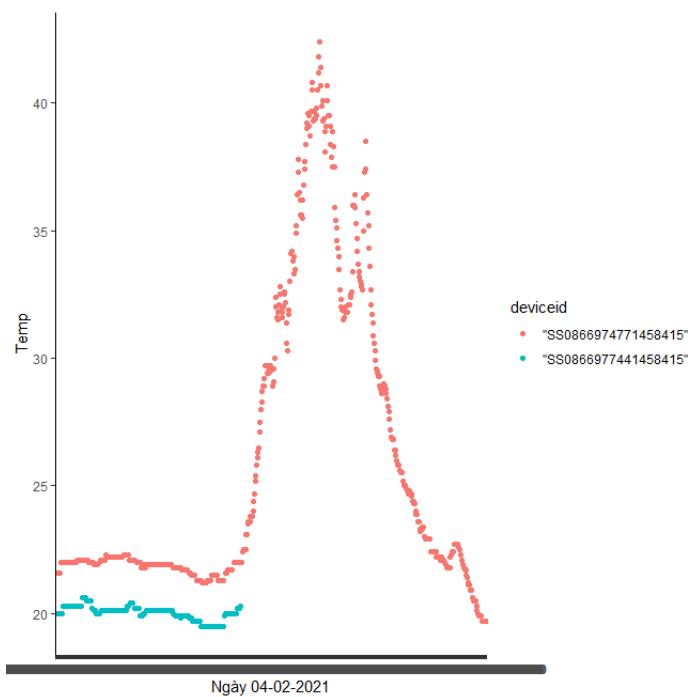
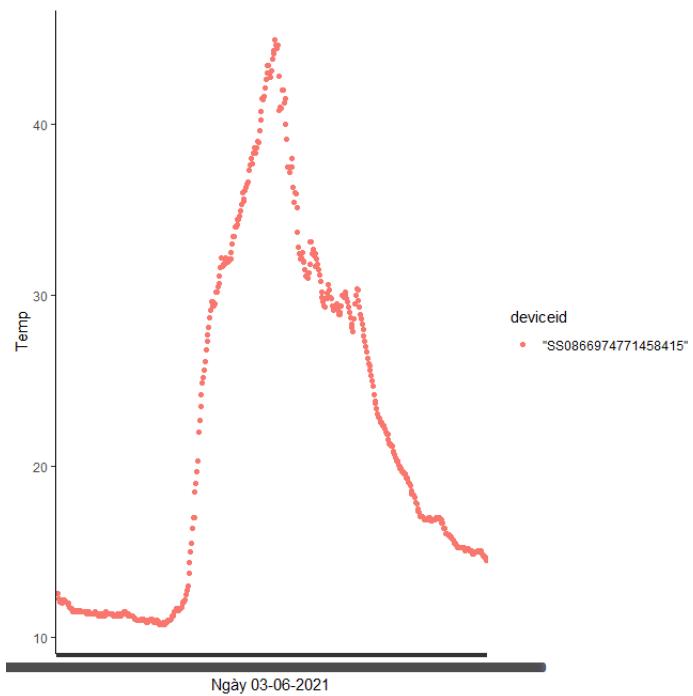
- 1 Biểu đồ thu thập nhiệt độ theo thời gian là ngày của tất cả thiết bị  
Lọc lấy dữ liệu bằng cách thêm cột date\_new chỉ chứa ngày trong cột date vào bảng số liệu và vẽ biểu đồ bằng các hàm trong thư viện ggplot2 cho cột temp

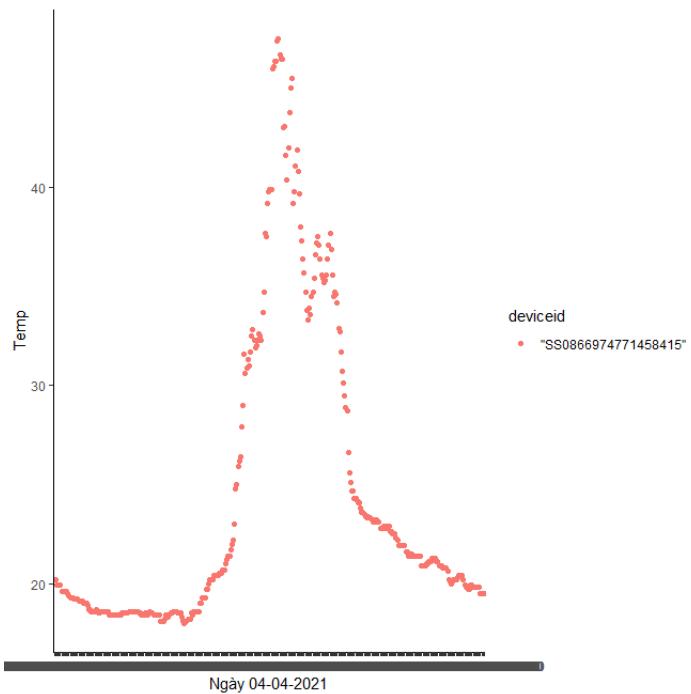
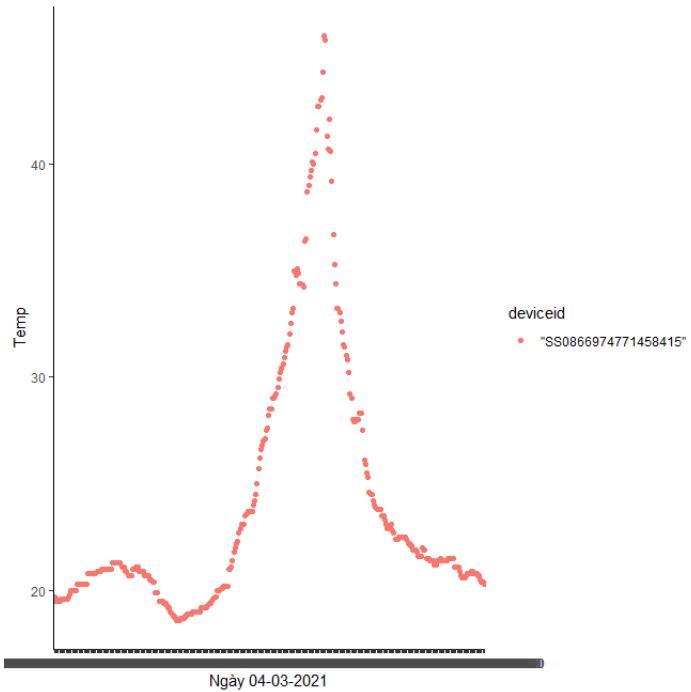
```
date_new = substr(data_d$date,1, 10)
clock= substr(data_d$date,12,30)
date_new = as.Date(date_new,format = " %Y-%m-%d")
class(date_new)
data_d = cbind(data_d, date_new)
data_d = cbind(data_d, clock)
#-----
library("ggplot2", lib.loc="C:/Program Files/R/R-4.1.1/library")
data_graph = subset(data_d, data_d$date_new == "2021-01-04")
attach(data_graph)
p1 = ggplot(data = data_graph, aes(x = clock, y = Humi))
p1 + geom_point(aes(col = deviceid)) + xlab("Ngày 01-04-2021")+theme_classic()
```

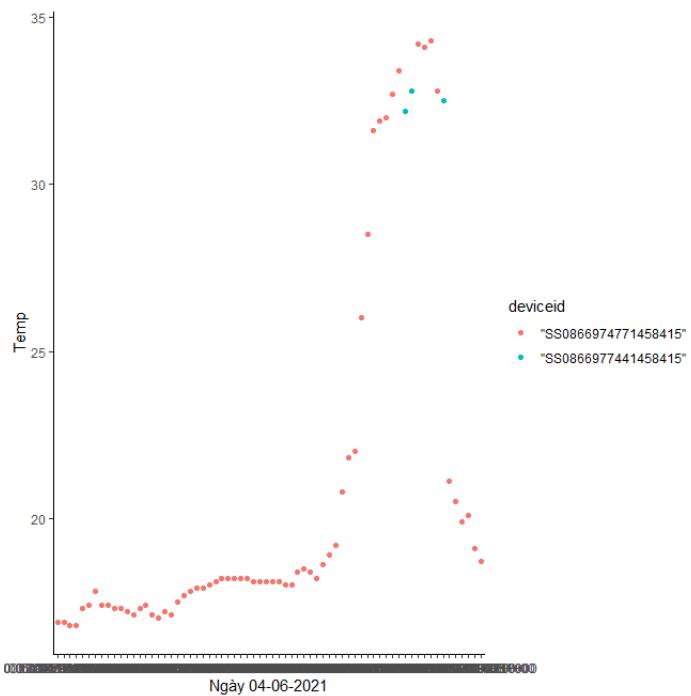






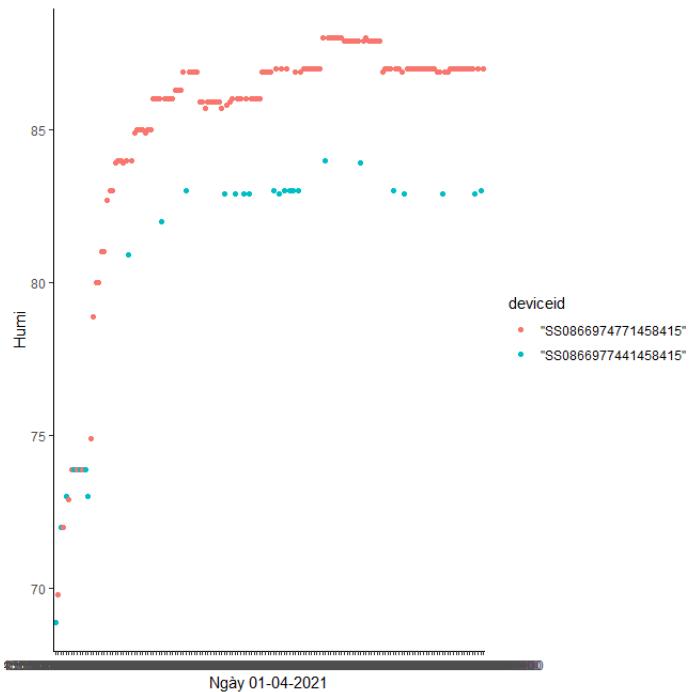


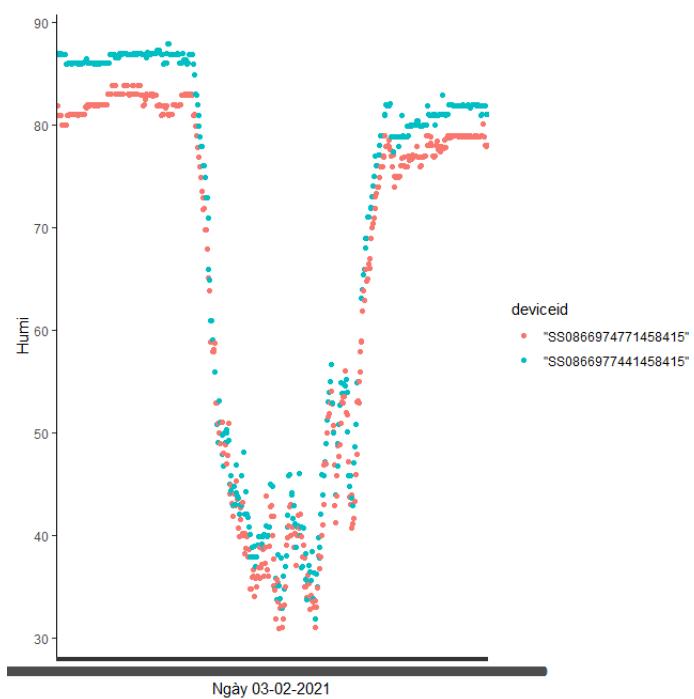
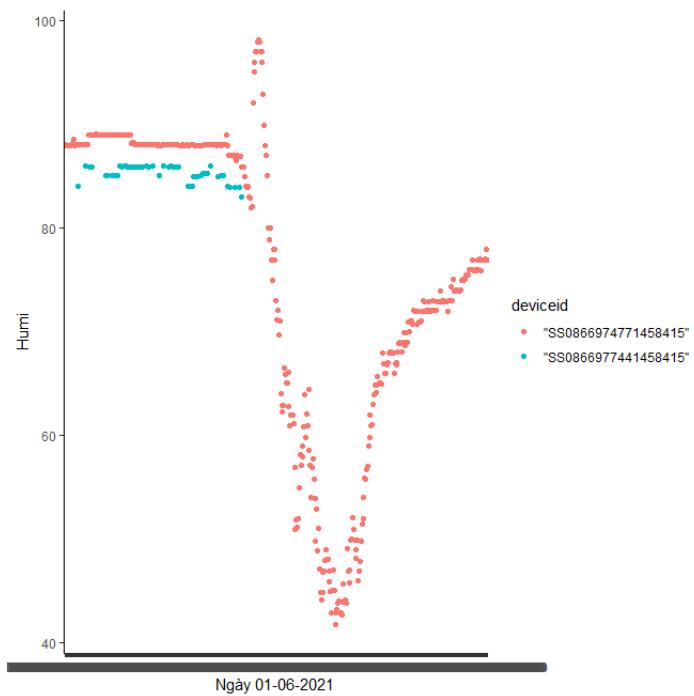


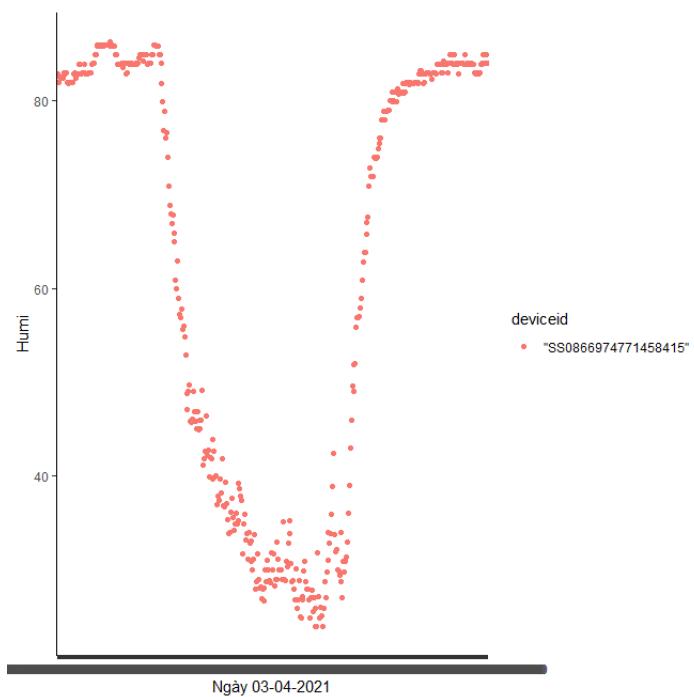
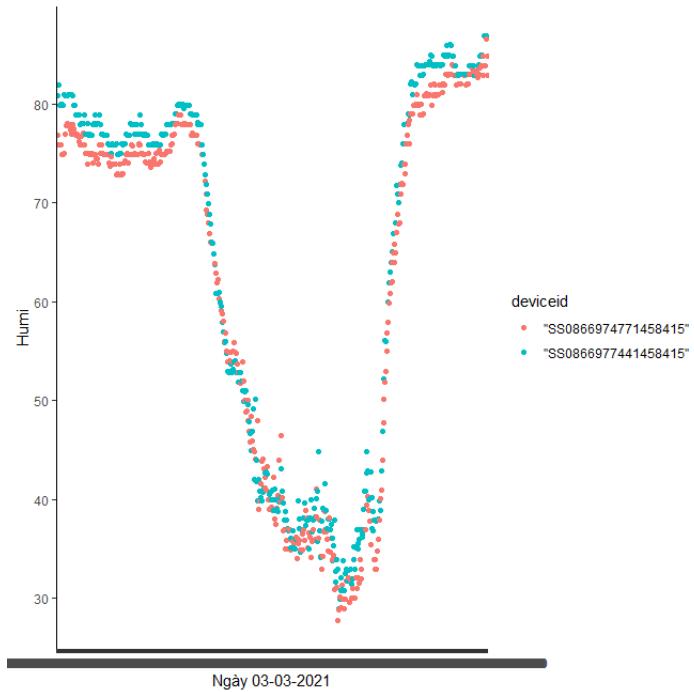


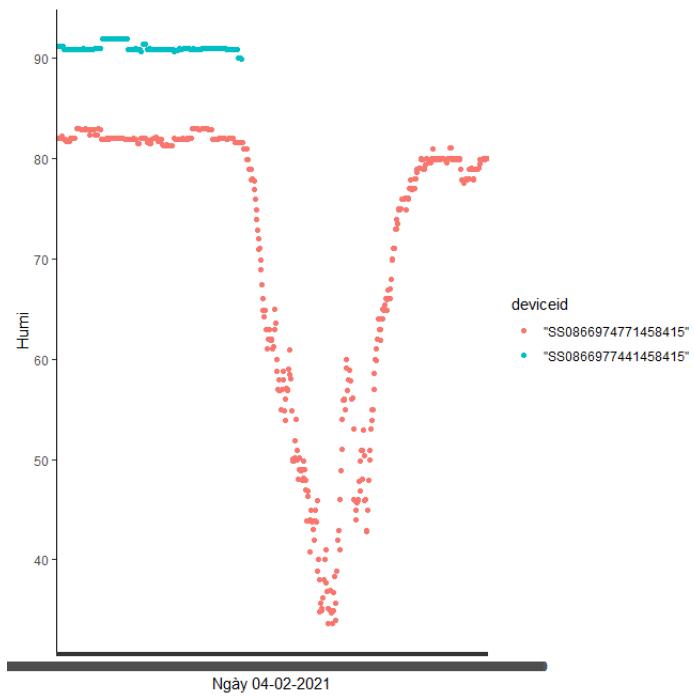
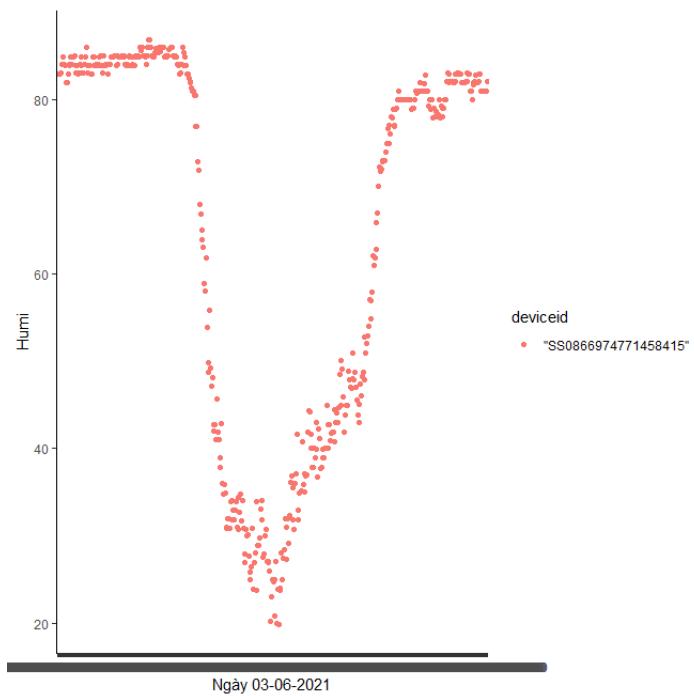
2 Biểu đồ thu thập độ ẩm theo thời gian là ngày của tất cả thiết bị Lọc lấy dữ liệu bằng cách thêm cột date\_new chỉ chứa ngày trong cột date vào bảng số liệu và vẽ biểu đồ bằng các hàm trong thư viện ggplot2 cho cột humi

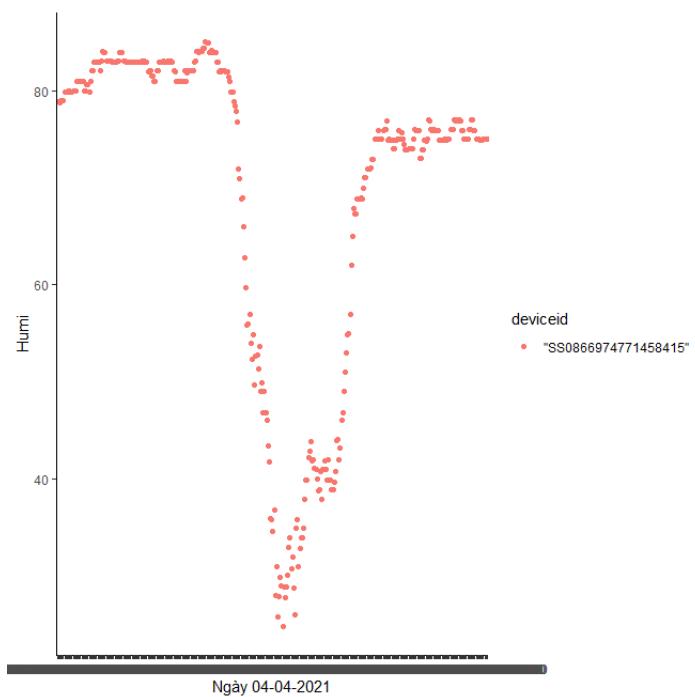
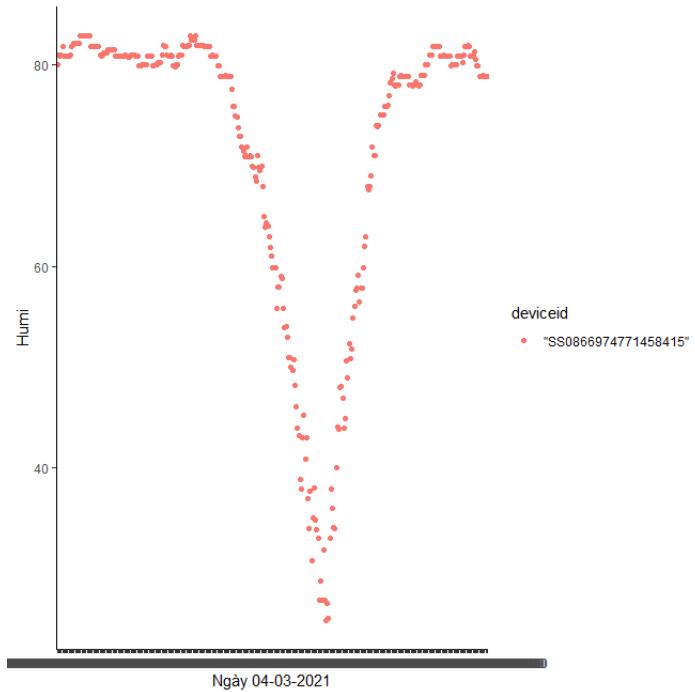
```
library("ggplot2", lib.loc="C:/Program Files/R/R-4.1.1/library")
data_graph = subset(data_d, data_d$date_new == "2021-01-04")
attach(data_graph)
p1 = ggplot(data = data_graph, aes(x = clock, y = Humi))
p1 + geom_point(aes(col = deviceid)) + xlab("Ngày 01-04-2021") + theme_classic()
```

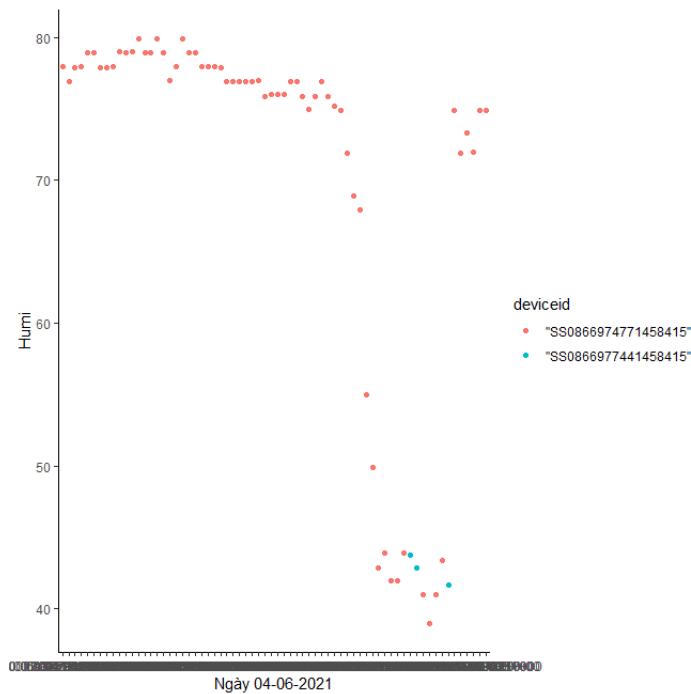












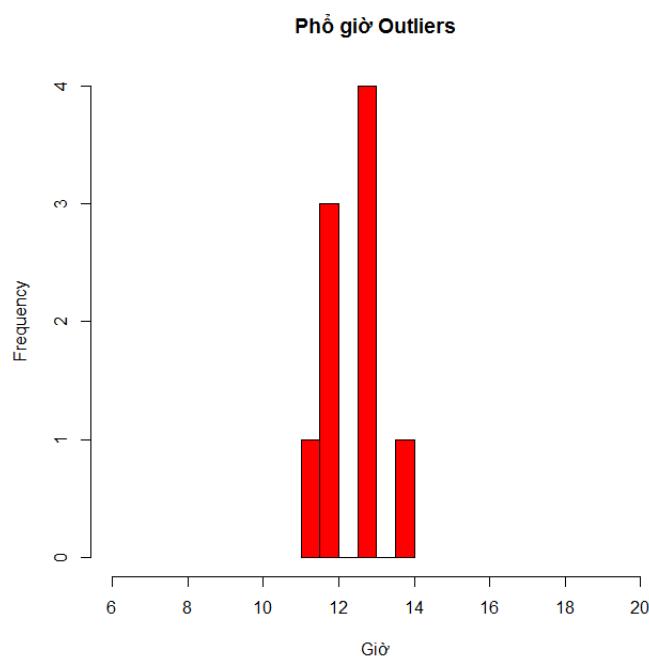
### 3 Vẽ phô giờ mà outliers xuất hiện trên tất cả các thiết bị cảm biến

Thực hiện các bước như sau:

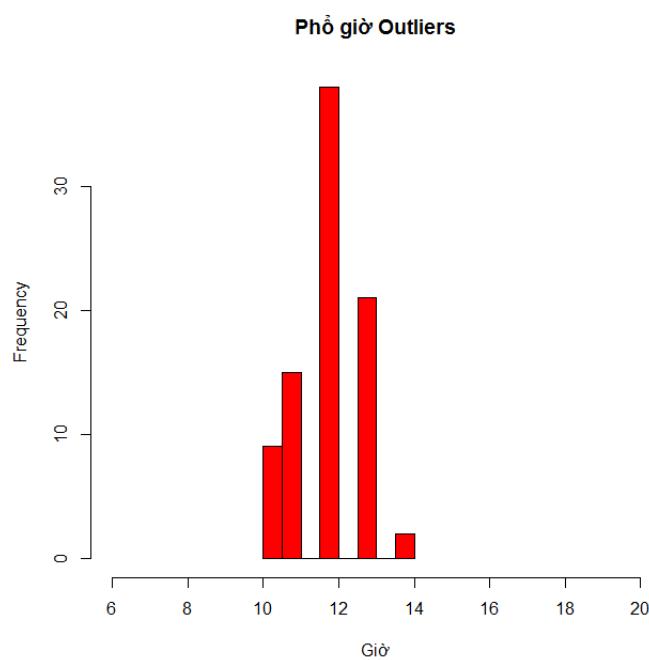
- Cài đặt gói Lubridate để lọc dữ liệu theo giờ, qua hàm hour.
- Dùng hàm boxplot.stats() để tìm ra các giá trị outliers trong tập dữ liệu.
- Dùng hàm filter() để lọc các mốc thời gian có các giá trị outliers tương ứng đã tìm được.
- Quy đổi dữ liệu thời gian vừa tìm được ra giờ.
- Lập 1 véc-tơ lặp số 1 với số lần lặp là số mốc thời gian (giờ) mà ta vừa tìm được.
- Gộp 2 véc-tơ ở bước 4 và 5 thành 1 véc-tơ, sau đó dùng hàm vẽ đồ thị HITS() để vẽ biểu đồ thanh thể hiện sự phân bố của phô giờ theo số lần xuất hiện của giá trị outliers tại thời điểm đó.

```
data <- read.csv("3-01_2021.csv")
data <- read.csv("4-03_2021.csv")
data <- read.csv("5-04_2021.csv")
data_d = data
library("dplyr", lib.loc="C:/Program Files/R/R-4.1.1/library")
library("lubridate", lib.loc="C:/Program Files/R/R-4.1.1/library")
hala<-boxplot.stats(data_d$Temp)$out
check<-data.frame(hala)
result_time<-data_d%>%filter(data_d$Temp %in% check$hala)
new<-hour(result_time$date)
hala<-data.frame(new)
vamos<-data.frame(hala)
hist(vamos$new, xlim=c(6,20),breaks=6,col="red",
      main="Phô giờ Outliers",xlab="Giờ")
```

\*Không tìm thấy outliers trong File 4



Phô giờ outliers tất cả thiết bị file 3

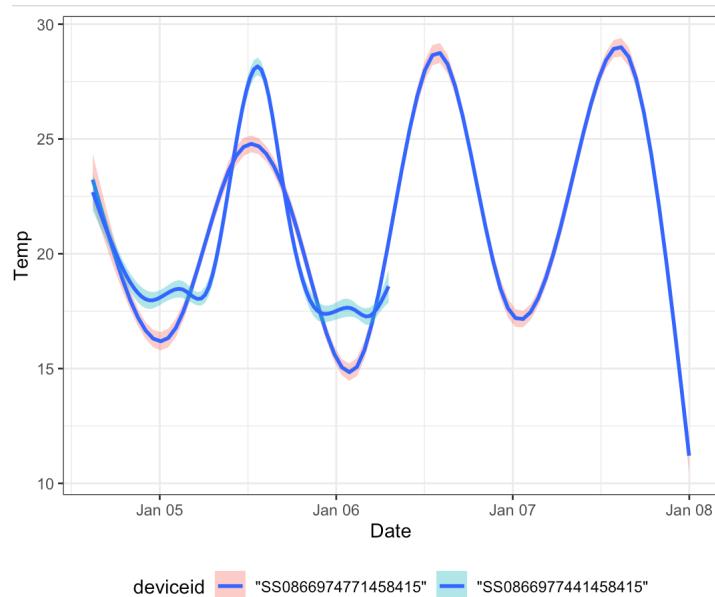


Phô giờ outliers tất cả thiết bị File 5

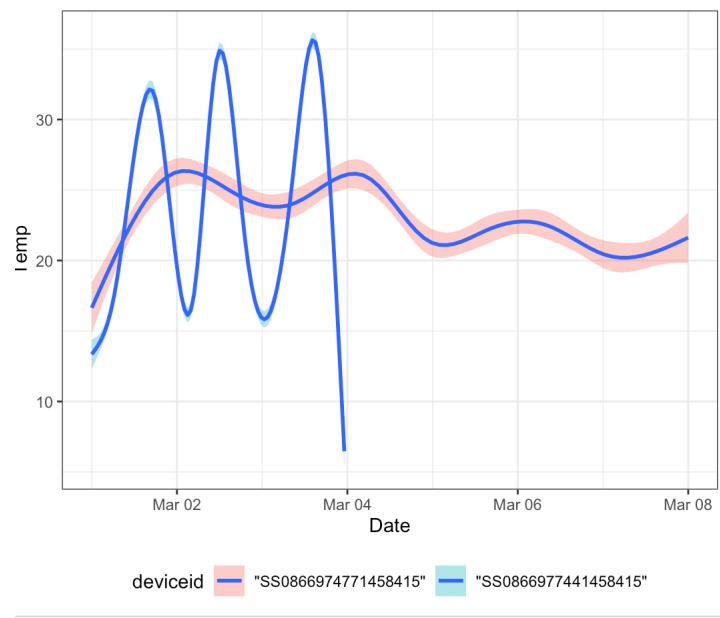
- vii) Nhóm câu hỏi liên quan đến tất cả thiết bị theo khoảng thời gian Trục Ox là thời gian, Oy là Temp/Humi, số đường thể hiện trên đồ thị là số thiết bị đo.

```
#loc du lieu theo ngay
cau7 <- X5_04_2021 %>% select(date,deviceid,Temp,Humi) %>%
filter(as.integer(format(X5_04_2021$date, "%d")) >= 1
& as.integer(format(X5_04_2021$date, "%d")) <=7)
date <- cau7$date
deviceid <- cau7$deviceid
temp <- cau7$Temp
humi <- cau7$Humi
#ve bieu do
a = ggplot(cau7, aes(x=date, y=temp, fill=deviceid))
a + geom_smooth() + xlab("Date") + ylab("Temp") + theme_bw() +
theme(legend.position = "bottom")
```

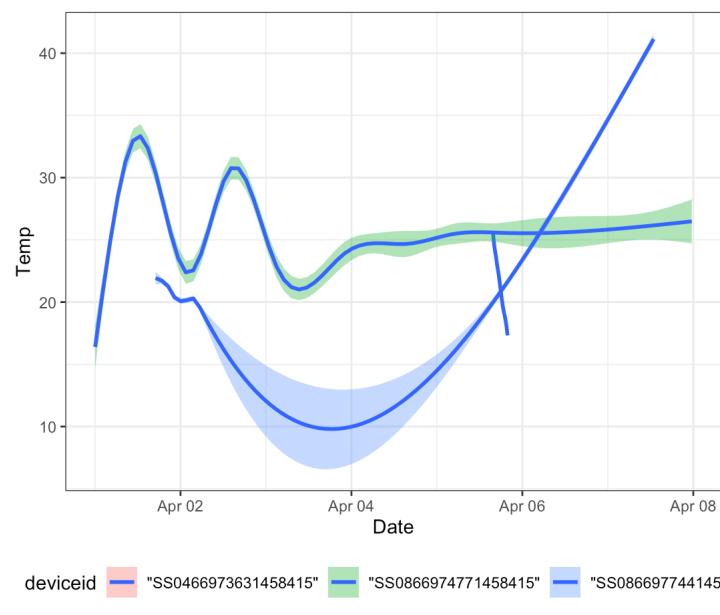
- 1) Biểu đồ thu thập nhiệt độ/ độ ẩm theo thời gian từ ngày 1 đến ngày 7 của tất cả thiết bị.  
File 3



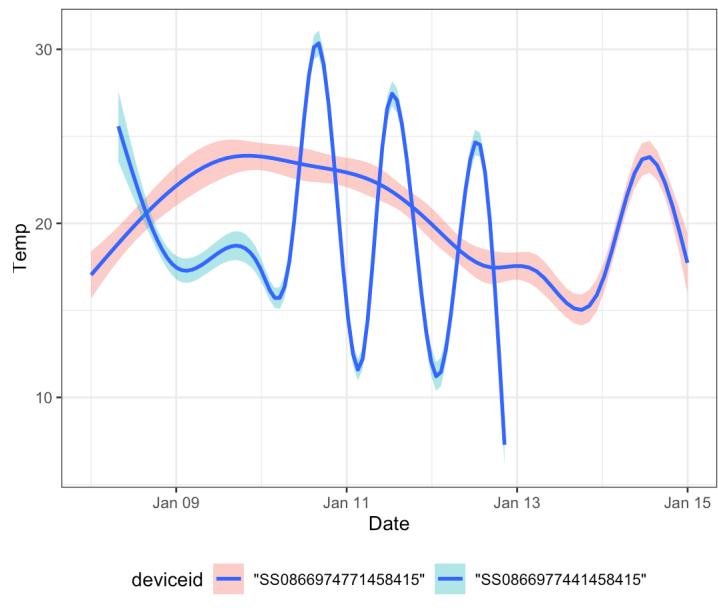
File 4



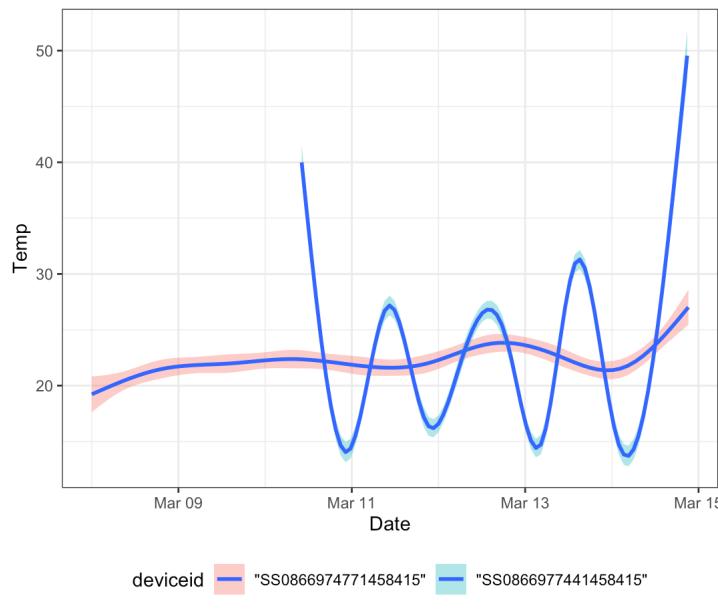
File 5



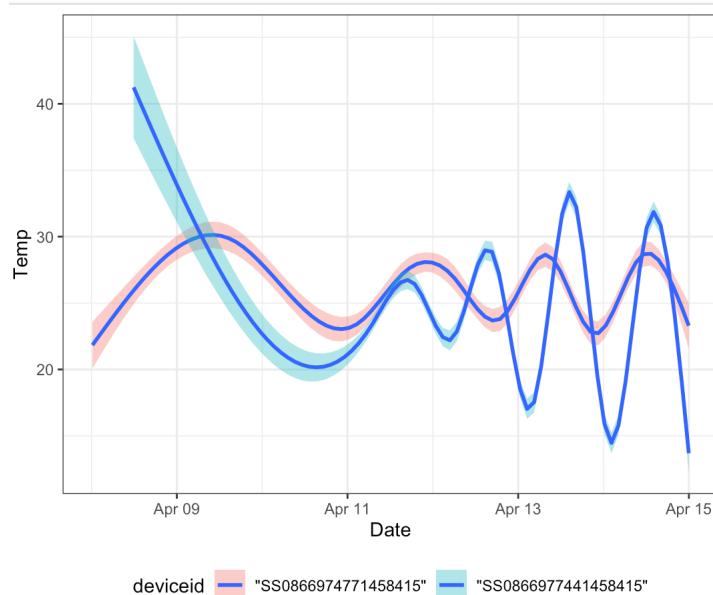
- 2) Biểu đồ thu thập nhiệt độ/ độ ẩm theo thời gian từ ngày 8 đến ngày 14 của tất cả thiết bị.  
File 3



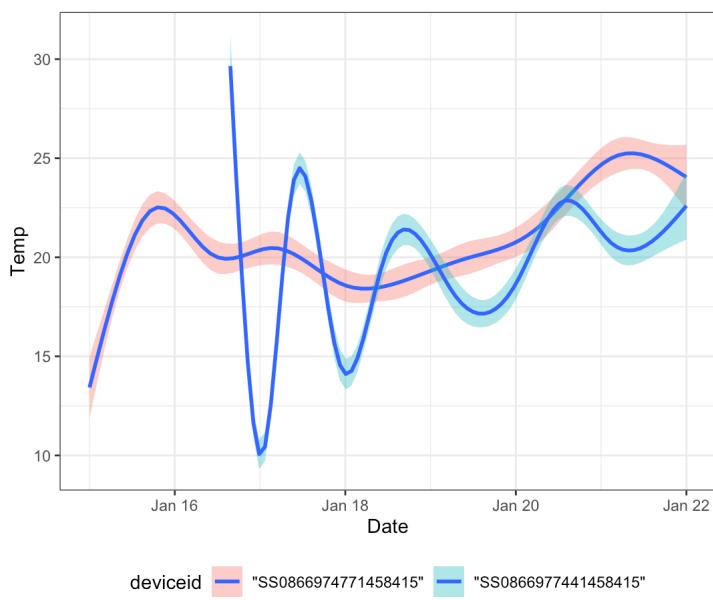
File 4



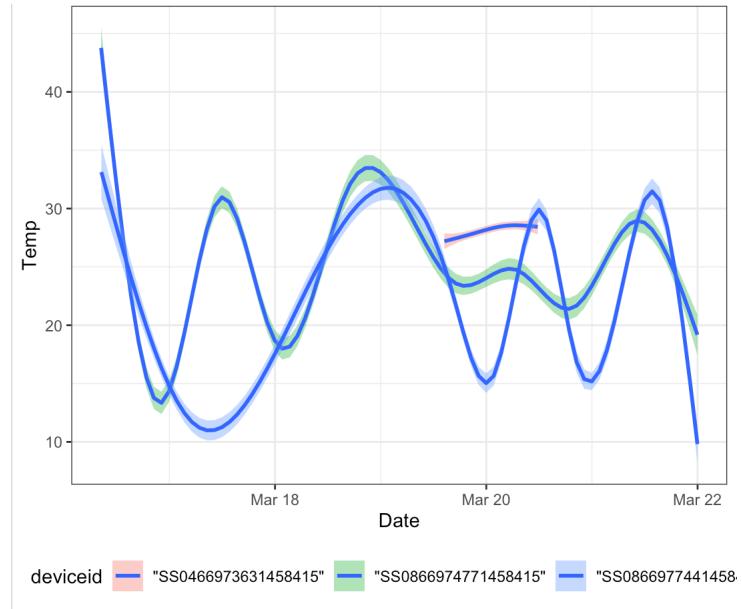
File 5



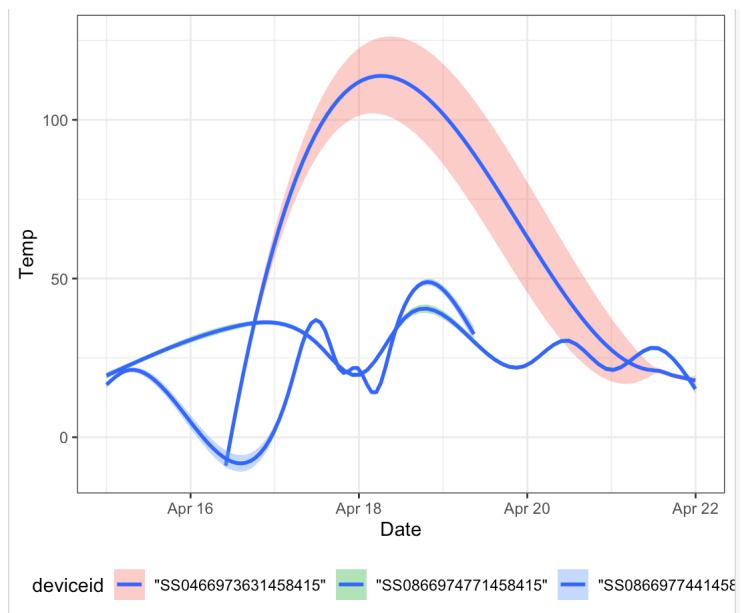
3) Biểu đồ thu thập nhiệt độ/ độ ẩm theo thời gian từ ngày 15 đến ngày 21 của tất cả thiết bị.  
File 3



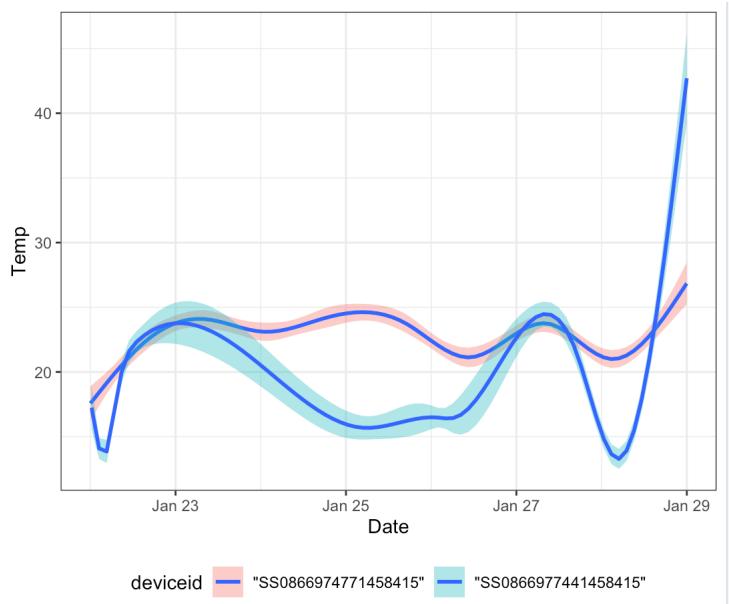
File 4



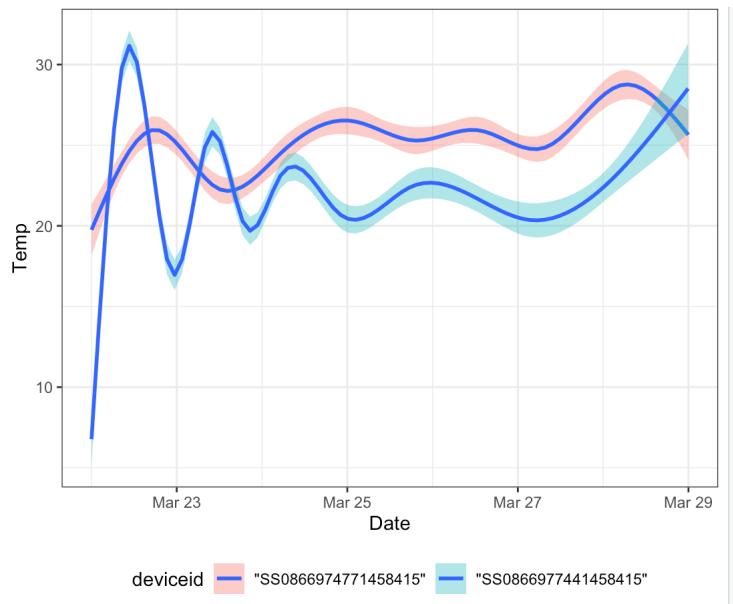
File 5



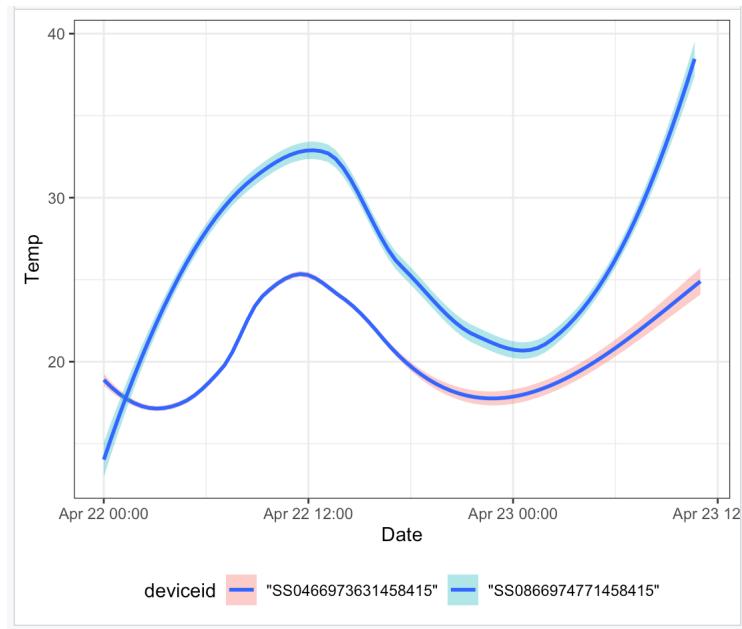
- 4) Biểu đồ thu thập nhiệt độ/ độ ẩm theo thời gian từ ngày 22 đến ngày 28 của tất cả thiết bị.  
File 3



File 4



File 5



viii) Nhóm câu hỏi liên quan đến sự tương quan giữa nhiệt độ và độ ẩm

Trên từng thiết bị hãy vẽ biểu đồ thể hiện trục Ox là nhiệt độ, trục Oy là độ ẩm.

- 1) Xét tương quan trong một ngày, hãy lấy 4 ngày theo 4 ký số mã để thể hiện. Nếu ký số là 0 thì lấy ngày là 10.

FILE 3:

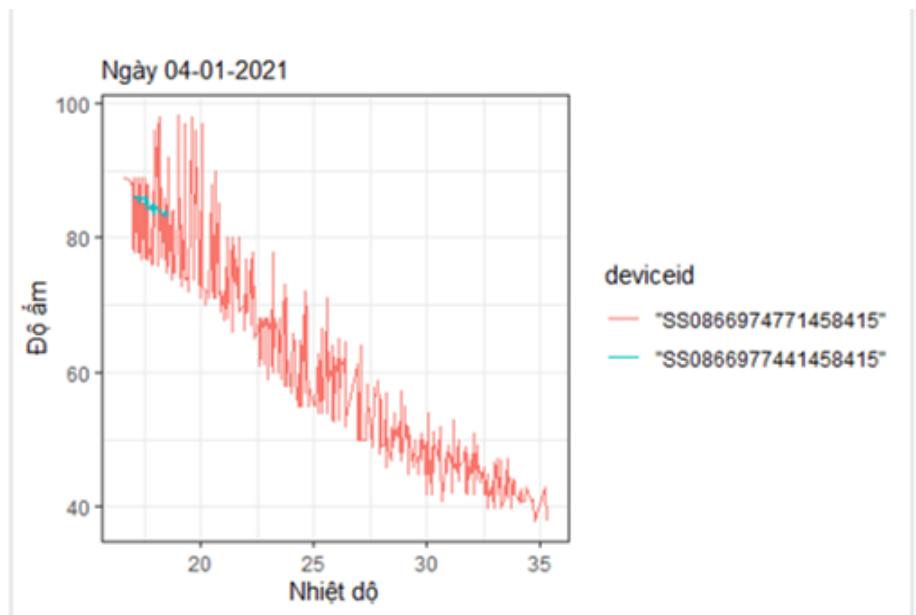
Để vẽ biểu đồ tương quan của nhiệt độ và độ ẩm trong 1 ngày, ta thực hiện các bước sau:

- Dùng công cụ Filter lọc dữ liệu theo ngày.
- Dùng hàm ggplot để vẽ biểu đồ đường thể hiện sự tương quan, với trục Ox là nhiệt độ, Oy là độ ẩm.

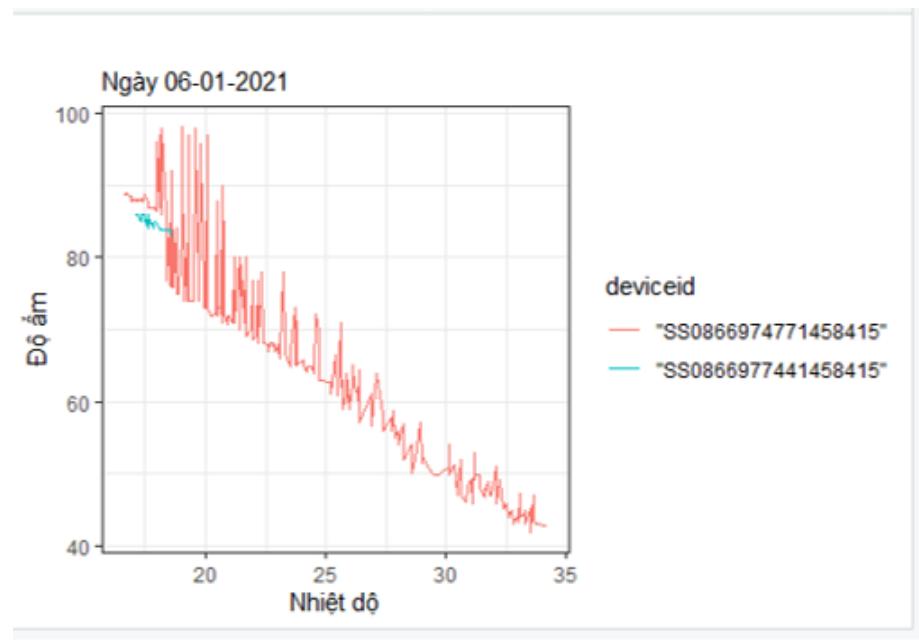
Cụ thể như sau:

```
Date_4 <- FILE_3[671:1570,]  
Date_4f33 = ggplot(data=Date_4, aes(x=Temp, y=Humi))  
Date_4f33 + geom_line(aes(col=deviceid)) + labs(title = "",  
subtitle = "Ngày 04-01-2021",x = "Nhiệt độ",y = "Độ ẩm") + theme_bw()
```

Kết quả thu được biểu đồ như sau:

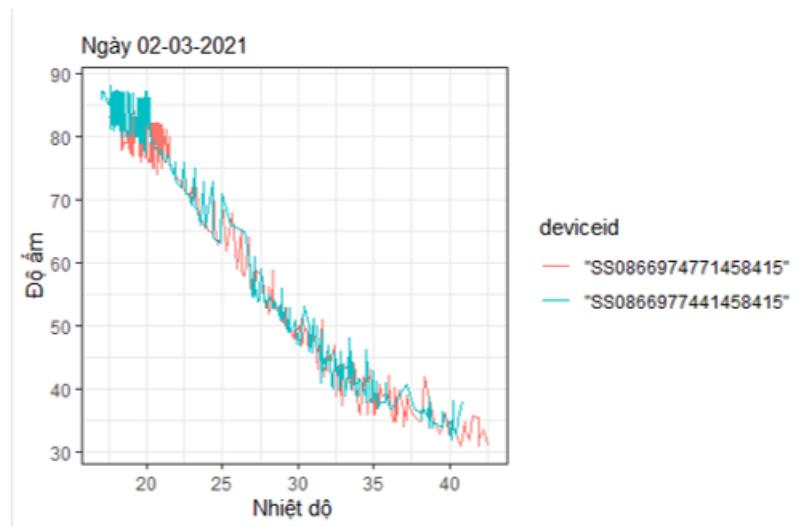


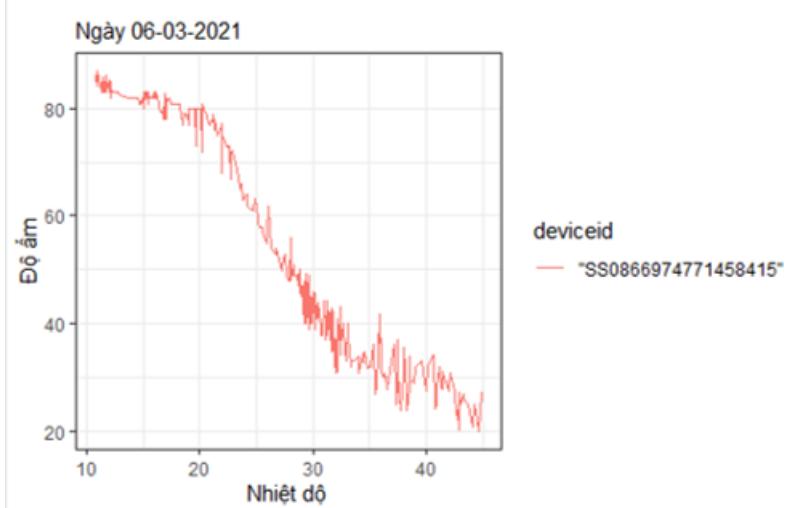
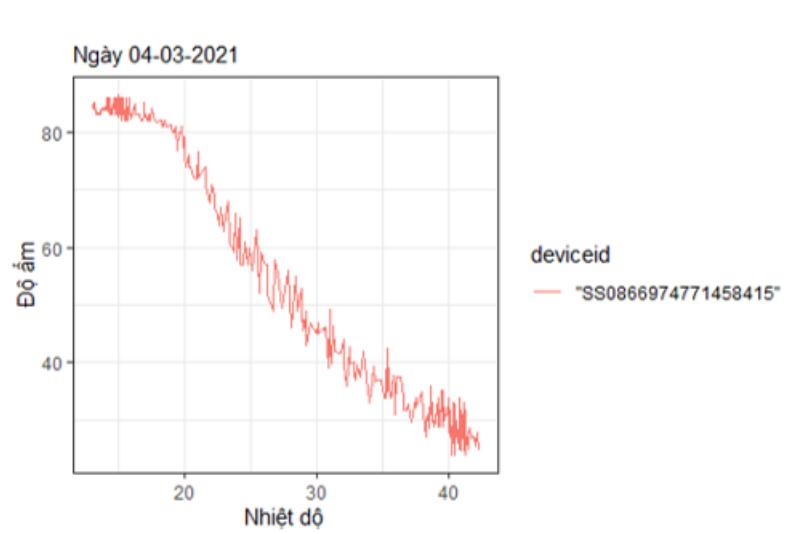
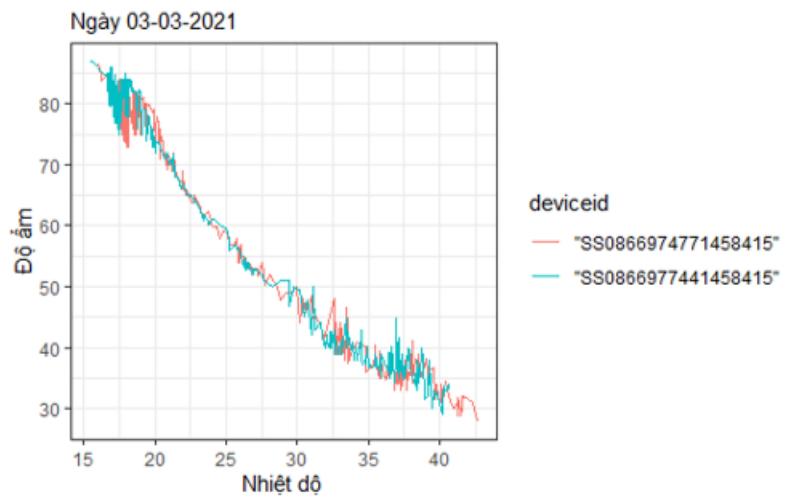
Thực hiện tương tự cho ngày 06-01-2021, ta thu được biểu đồ tương quan như sau:



FILE 4:

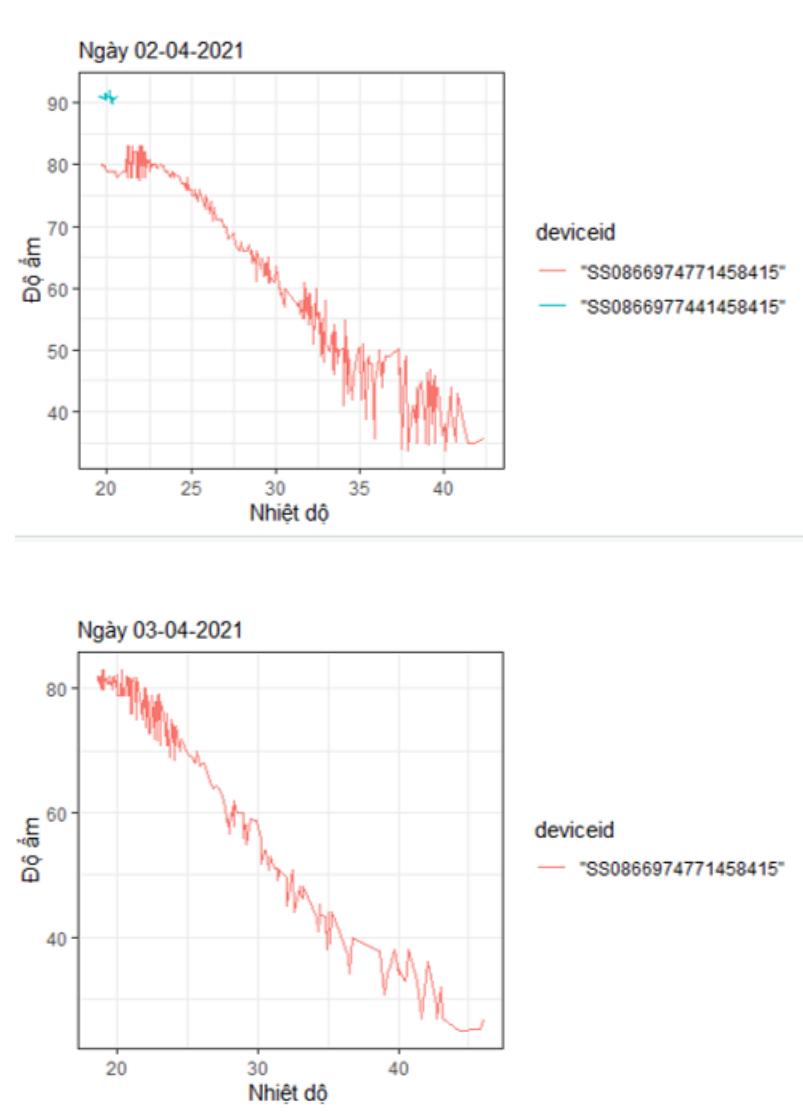
Thực hiện tương tự nhu file 3, ta lần lượt thu được đồ thị biểu diễn tương quan của nhiệt độ và độ ẩm trong các ngày 2,3,4,6/03/2021.

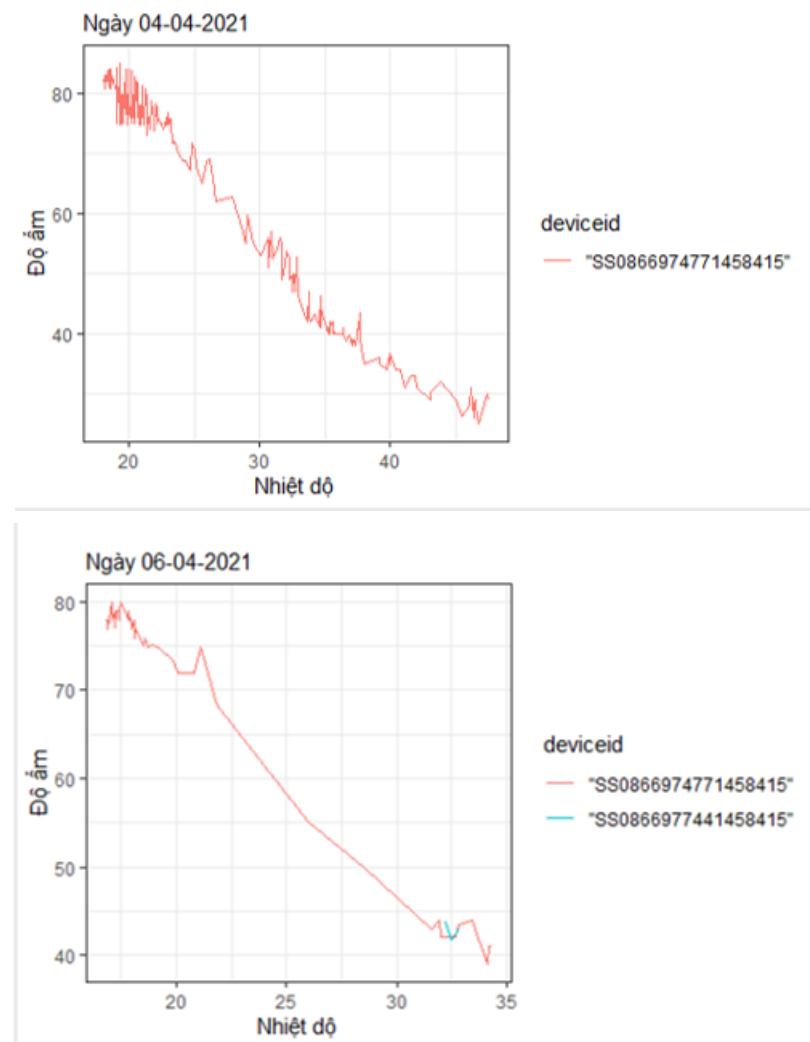




FILE 5:

Biểu đồ tương quan của nhiệt độ và độ ẩm của các ngày 2,3,4,6/04/2021 như sau:





2) Xét tương quan từ ngày 1 đến ngày 7

FILE 3:

Thực hiện các bước sau:

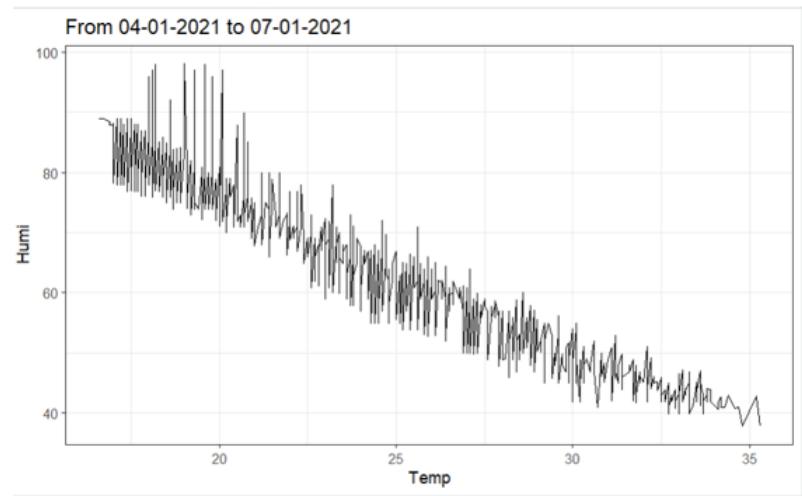
- Dùng công cụ filter của R để lọc dữ liệu và sau đó gom dữ liệu từ ngày 1 đến ngày 7 lại thành 1 data frame.

- Dùng các hàm vẽ đồ thị trong ggplot2 để vẽ biểu đồ thể hiện tương quan của nhiệt độ và độ ẩm.

\*Ghi chú: Dữ liệu trong File 3 bắt đầu từ ngày 04-01-2021.

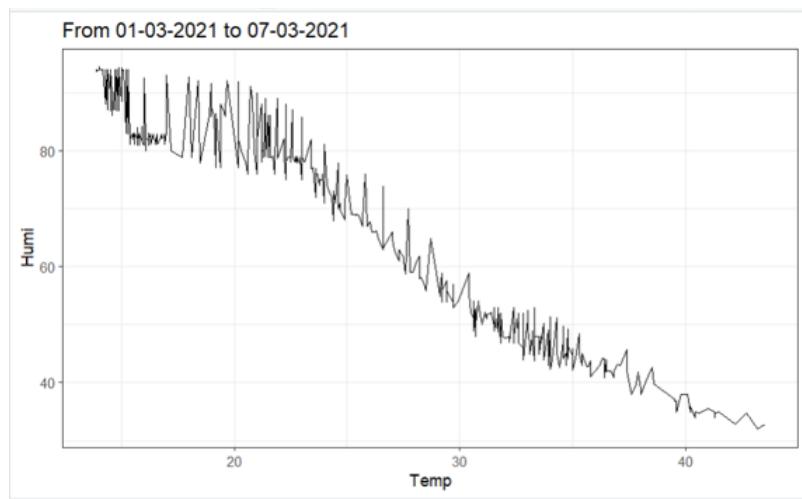
```
Tuongquan8_14f3<-file_3[4943:9286,]
Tuongquan_grapf3.3<-ggplot(Tuongquan8_14f3, aes(x=Temp, y=Humi))
Tuongquan_grapf3.3 + geom_line() + ggtitle("From 04-01-2021 to 07-01-2021")
+ theme_bw()
```

- Ta được kết quả như sau:



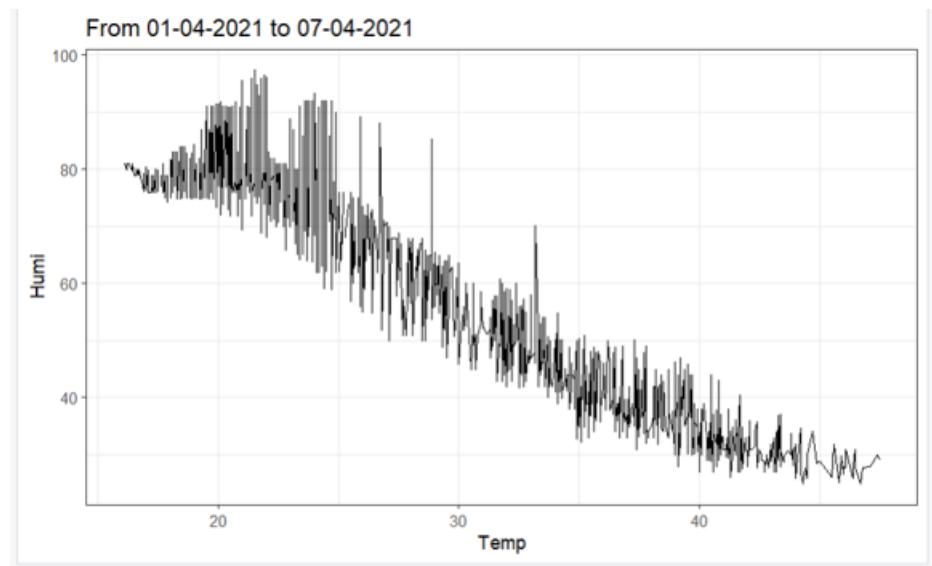
FILE 4:

Thực hiện tương tự như file 3, ta vẽ được đồ thị tương quan ở file 4 như sau:

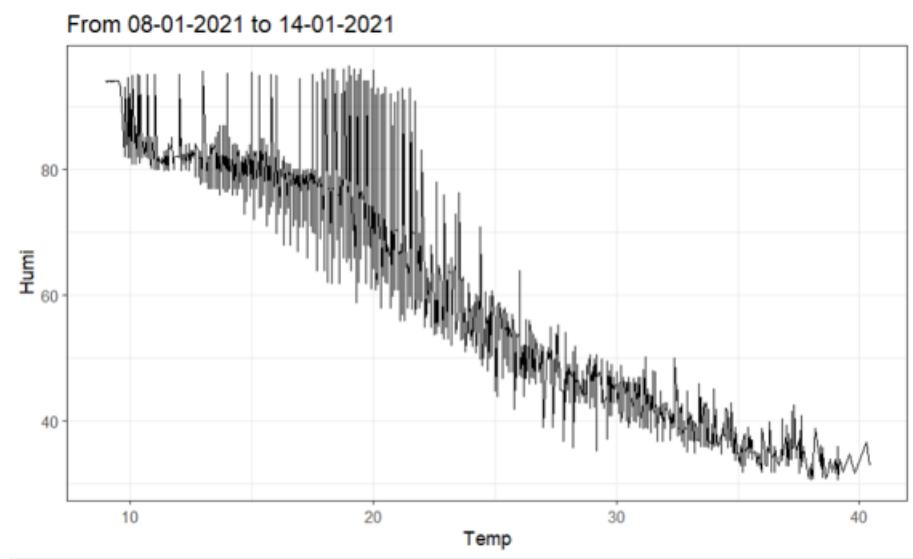


FILE 5:

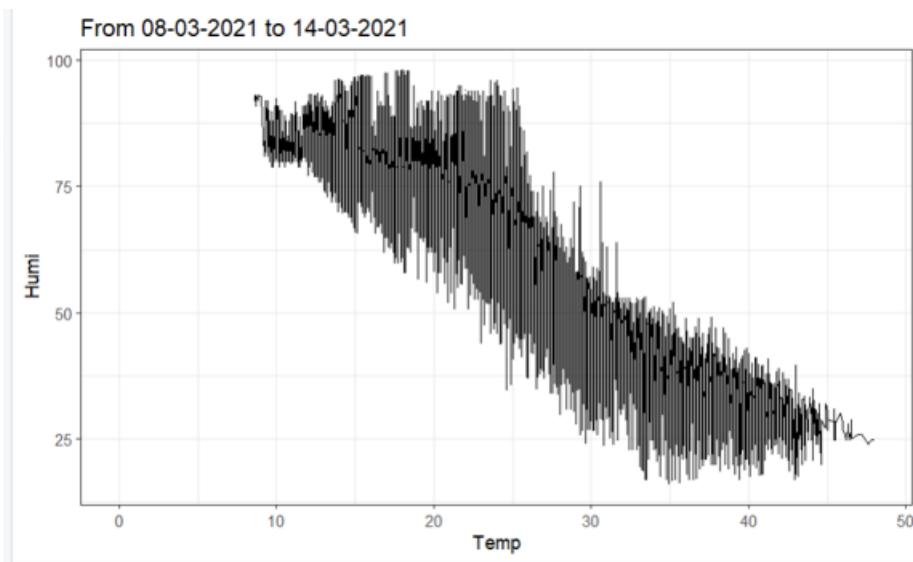
Thực hiện tương tự đối với các số liệu File 5, ta vẽ được đồ thị tương quan từ ngày 01/04-07/04/2021 như sau:



- 3) Xét tương quan từ ngày 8 đến ngày 14  
FILE 3: Thực hiện trong tháng 01-2021.

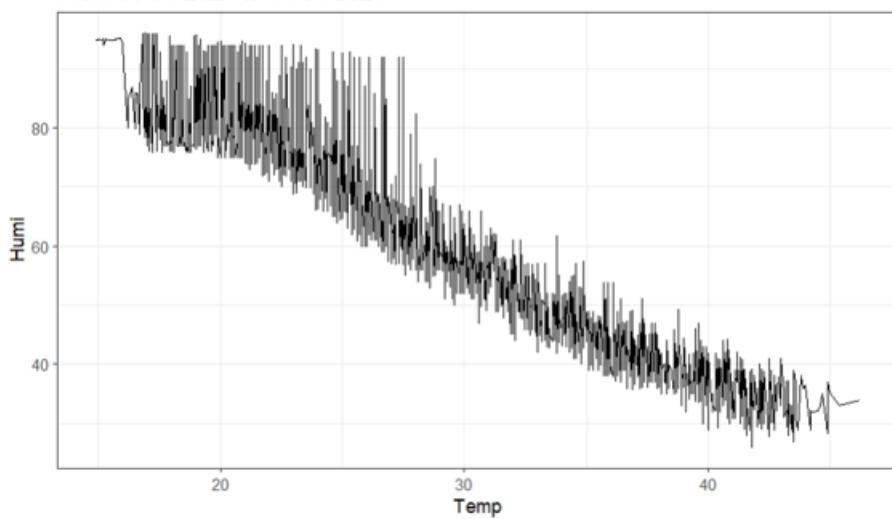


FILE 4: Thực hiện trong tháng 03-2021.



FILE 5: Thực hiện trong tháng 04-2021.

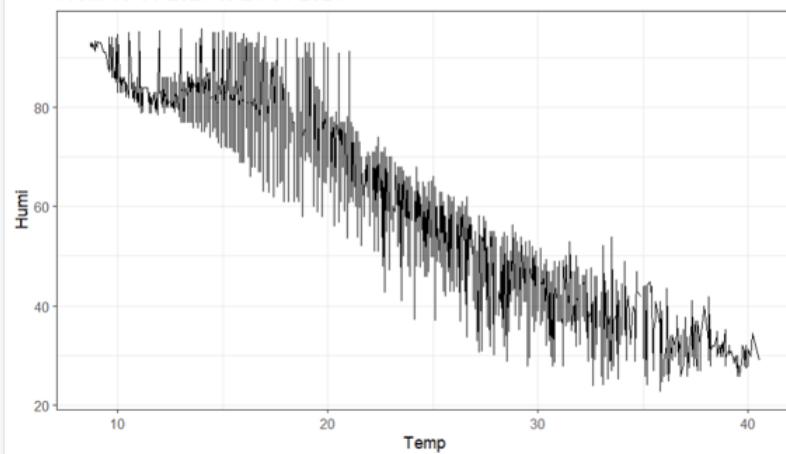
From 08-04-2021 to 14-04-2021



4) Xét tương quan từ ngày 15 đến ngày 21

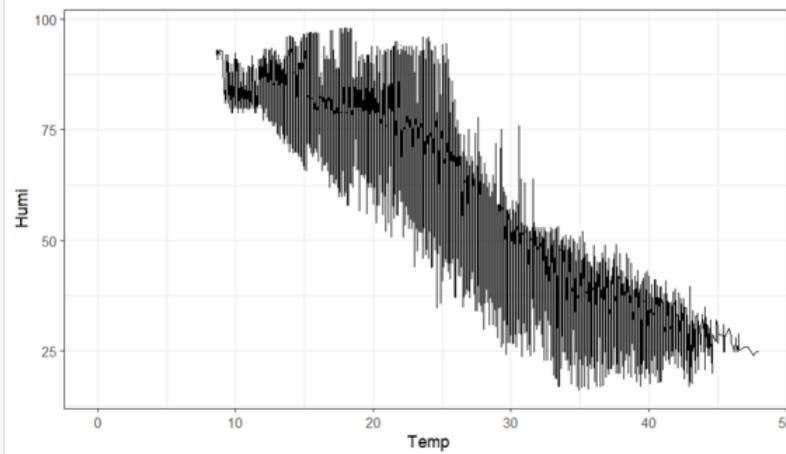
FILE 3: Thực hiện trong tháng 01-2021

From 15-01-2021 to 21-01-2021

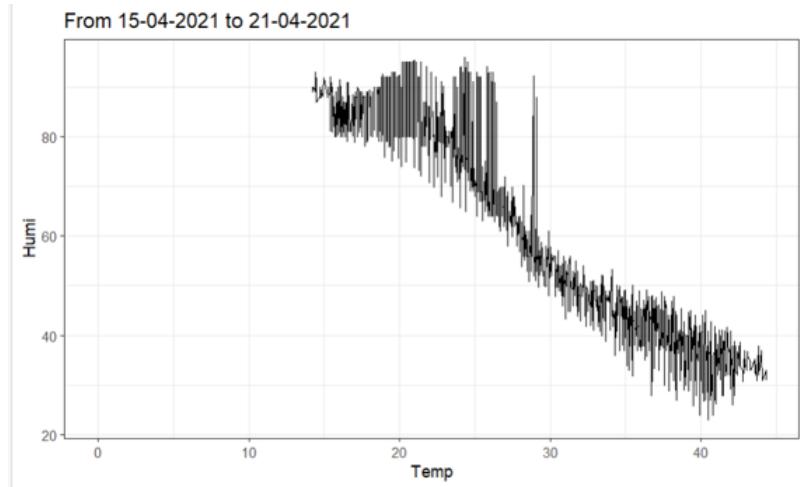


FILE 4: Thực hiện trong tháng 03-2021

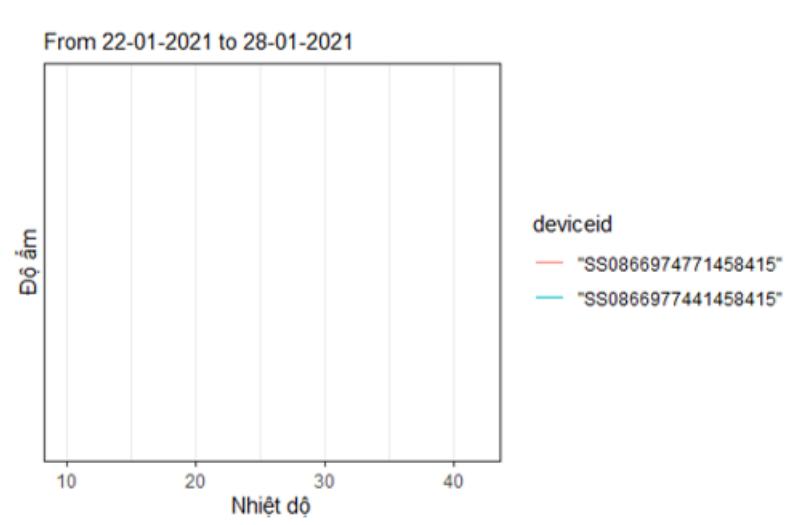
From 15-03-2021 to 21-03-2021



FILE 5: Thực hiện trong tháng 04-2021

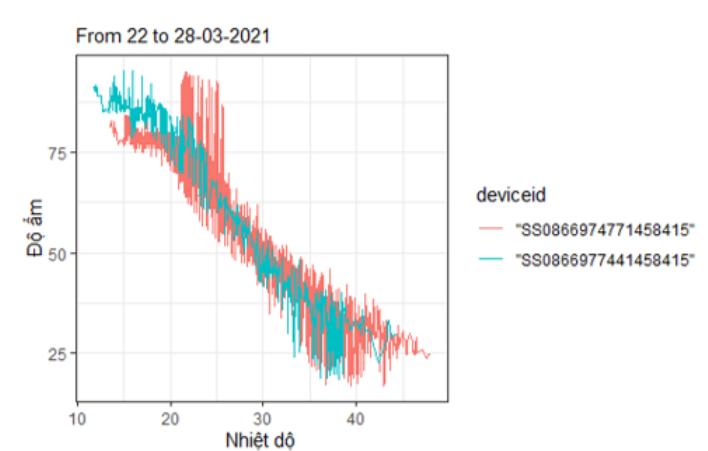


- 5) Xét tương quan từ ngày 22 đến ngày 28  
FILE 3: Thực hiện trong tháng 01-2021.

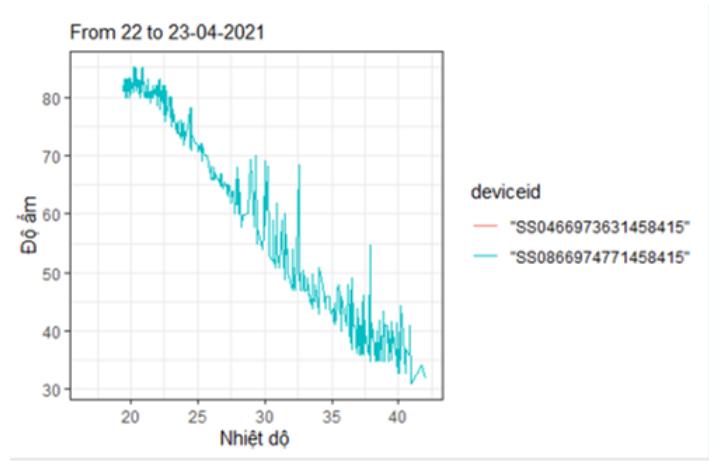


Ghi chú: Cột giá trị của Humi nhận giá trị N/A nên đồ thị không được thành lập ở tập dữ liệu này.

FILE 4: Thực hiện trong tháng 03-2021.



FILE 5: Thực hiện trong tháng 04-2021.





ix) Nhóm câu hỏi riêng

- 1) Cho biết đơn vị thời gian lấy mẫu của thiết bị cảm biến
- 2) Số lượng mẫu thu thập có ổn định không

FILE 3:

Để thực hiện đánh giá. Ta thực hiện các bước như sau:

- Gom dữ liệu theo ngày.
- Lập bảng thống kê tần suất hay số lượng dữ liệu thu thập theo từng ngày.
- Cuối cùng là gộp 2 data frame trên lại với nhau.
- Sử dụng hàm oav để phân tích phương sai Anova của tập data mới tạo thành theo 2 trường dữ liệu là ngày và số lượng dữ liệu theo ngày. Từ đó đi đến kết luận sự ổn định của tập dữ liệu.

```
Table_day_f3<-table(as.Date(date))
Sort_day_f3<-data.frame(cbind(Table_day_f3))
Sort_day_f3

date_f3<- c(4:31)
date_f3

Data_check<-data.frame(date_f3, Sort_day_f3$Table_day_f3)
Data_check
```

Ta được kết quả là giá trị của 2 cột, ngày trong tháng và tần số hay số lượng dữ liệu.



```
> Data_check<-data.frame(date_f3, Sort_day_f3$Table_day_f3)
> Data_check
  date_f3 Sort_day_f3.Table_day_f3
1         4                  158
2         5                  601
3         6                  459
4         7                  476
5         8                  304
6         9                  286
7        10                  534
8        11                  641
9        12                  633
10       13                  377
11       14                  473
12       15                  473
13       16                  599
14       17                  755
15       18                  686
16       19                  619
17       20                  621
18       21                  591
19       22                  601
20       23                  471
21       24                  382
22       25                  390
23       26                  423
24       27                  475
25       28                  444
26       29                  267
27       30                  218
28       31                  303
```

Tiếp tục dùng hàm aov để phân tích Anova cho dữ liệu trên như sau:

```
Result<-aov(Table_day_f3 ~ date_f3)
summary(Result)
```

Kết quả thu được như sau:

```
R 4.1.1 - C:/Users/Admin/Desktop/THỰC HÀNH R/ ↵
> summary(Result)
  Df Sum Sq Mean Sq F value Pr(>F)
date_f3     1 10853   10853    0.467   0.5
Residuals  26 604419   23247
> |
```

\* Ta thấy giá trị  $Pr = 0.5 > 0.05$ , ta đi đến kết luận: Không có sự khác biệt có ý nghĩa thống kê giữa các nhóm hay trong trường hợp này là giữa các ngày trong tập dữ liệu. Vậy đi đến kết luận là số lượng mẫu thu được ổn định.

FILE 4:

Thực hiện tương tự như file 3, ta thu được kết quả như sau:

```
R 4.1.1 - C:/Users/Admin/Desktop/THỰC HÀNH R/ ↵
> summary(Result_f4)
  Df Sum Sq Mean Sq F value Pr(>F)
date_f4     1      154      154    0.004  0.951
Residuals  28 1133760    40491
> |
```

\* Giá trị  $Pr = 0.951 > 0.05$ . Vậy dữ liệu thu được ở mẫu dữ liệu này cũng ổn định theo ngày.

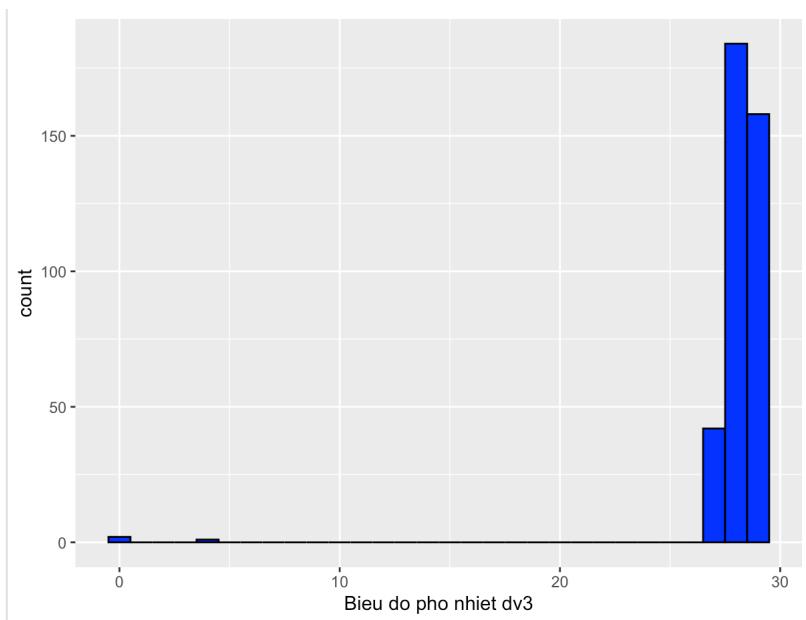
FILE 5:

Thực hiện tương tự như file 3 và file 4, ta thu được kết quả như sau:

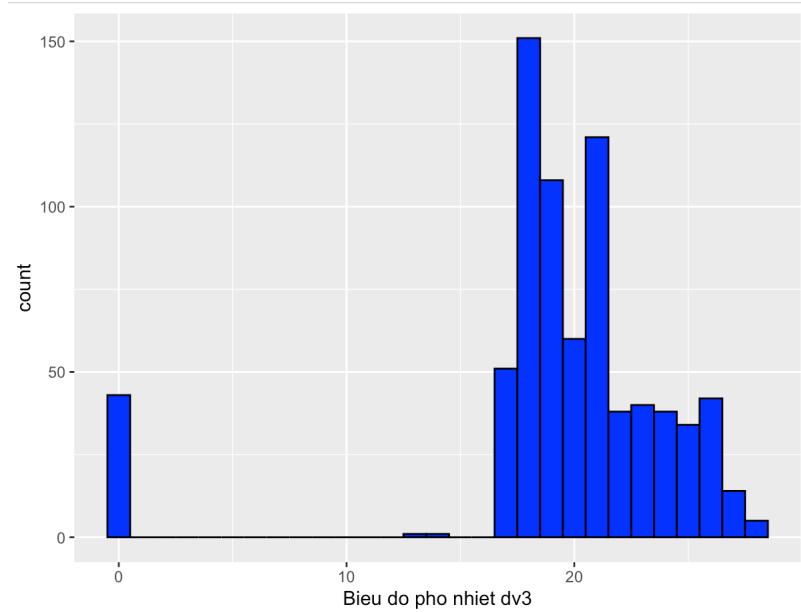
```
R 4.1.1 · C:/Users/Admin/Desktop/THỰC HÀNH R/ ↵
> summary(Result_f5)
   Df Sum Sq Mean Sq F value Pr(>F)
date_f5     1  94708   94708    1.045  0.318
Residuals  21 1903182   90628
>
```

\* Giá trị  $Pr = 0.318 > 0.05$ . Vậy dữ liệu thu được ở mẫu dữ liệu này ổn định theo ngày.

- 3) Có khoảng thời gian dài dữ liệu không thu thập không? Cho biết khoảng thời gian đó là bao nhiêu?  
File 4



File 5



Nhìn vào biểu đồ, ta thấy các thiết bị 1,2 luôn thu thập dữ liệu xuyên suốt 1 tháng, chỉ có thiết bị 3 khoảng thời gian đầu tháng không thu thập dữ liệu.

- 4) Nếu có khoảng thời gian dài dữ liệu không thu thập thì có thường xuyên lặp lại không?  
Nhìn vào biểu đồ ta thấy với thiết bị 3 ở file 4 thì thời gian dữ liệu không thu thập gần như lặp lại cả tháng trừ những ngày cuối. Còn ở file 5 thì khoảng thời gian ấy lặp lại từ đầu tới giữa tháng.
- 5) Các khung giờ mà thu được các giá trị trong nhóm Q1, Q2, Q3 cho từng thiết bị
- 6) Các mốc thời điểm max, min, mean, median có trùng nhau không?

#### FILE 3:

Ta thực hiện các bước như sau:

- Dùng hàm Summary thống kê ra các giá trị Min, Max, Medium, Mean của Humi, Temp
- Sau khi có được các dữ liệu trên ta dùng công cụ Filter của R lọc dữ liệu ra với điều kiện lọc là các giá trị min, max, mean, medium của humi/temp.

```
- attach(file_3)
- summary(Humi)
```

```
> summary(Humi)
Min. 1st Qu. Median     Mean 3rd Qu.      Max.    NA's
22.8      57.9    77.2    70.0    81.9    98.1    4479
>
```

```
- Min_Humi_f3 <- file_3[file_3$Humi==22.8,2]
- head(Min_Humi_f3)
```

Ta có được thời điểm tương ứng với giá trị Humi(min) bên dưới:

```
[1] "2021-01-16 12:15:45.987595+00" NA
[3] NA                               NA
[5] NA                               NA
>
```

Thực hiện tương tự, ta cũng tìm được các mốc thời gian tương ứng với các giá trị max, median và kết quả gần đúng cho giá trị mean.



- Thời điểm tương ứng với Humi(max):

```
> Max_Humi_f3 <- file_3[file_3$Humi==98.1,2]
> head(Max_Humi_f3)
[1] "2021-01-06 08:13:57.521602+00" NA
[3] NA                               NA
[5] NA                               NA
> |
```

- Thời điểm tương ứng với Humi(median):

```
> #Thời điểm tương ứng với Median(Humi) = 77.2
> Median_Humi_f3 <- file_3[file_3$Humi==77.2,2]
> head(Median_Humi_f3)
[1] "2021-01-12 00:33:19.296246+00" "2021-01-12 00:39:21.077327+00"
[3] "2021-01-12 00:42:22.150528+00" "2021-01-12 01:18:32.72751+00"
[5] "2021-01-17 07:13:08.116817+00" NA
```



- Thời điểm tương ứng với Humi(mean):

```
> #Thời điểm tương ứng với Mean(Humi) = 70.0
> Mean_Humi_f3 <- file_3[file_3$Humi==70.0,2]
> head(Mean_Humi_f3)
[1] "2021-01-05 10:41:03.059382+00" "2021-01-05 11:11:05.928964+00"
[3] "2021-01-06 19:43:08.48363+00"   "2021-01-07 07:39:58.765215+00"
[5] "2021-01-12 07:46:51.908345+00"   "2021-01-12 16:13:41.57879+00"
> |
```

\* Ta nhận thấy các thời điểm Min, max, median, mean của Humi không trùng nhau.

Thực hiện tương tự cho các giá trị của cột Temp, ta cũng thu được kết quả như sau:

```
> -----
> summary(Temp)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  8.70 15.60 18.80 20.78 25.40 42.00
> head(Min_Temp_f3)
[1] "2021-01-15 05:22:25.037176+00" "2021-01-15 05:46:29.77315+00"
[3] "2021-01-15 06:07:33.246804+00"
> head(Max_Temp_f3)
[1] "2021-01-26 12:02:47.025584+00" "2021-01-26 12:05:47.827859+00"
> head(Median_Temp_f3)
[1] "2021-01-04 18:41:19.961542+00" "2021-01-05 17:45:22.034593+00"
[3] "2021-01-05 18:06:23.874895+00" "2021-01-05 18:09:24.348306+00"
[5] "2021-01-06 07:25:26.094895+00" "2021-01-06 07:28:26.715896+00"
> head(Mean_Temp_f3)
[1] "2021-01-05 17:09:16.060694+00" "2021-01-05 17:12:23.313063+00"
[3] "2021-01-06 09:23:48.791238+00" "2021-01-06 19:58:12.433082+00"
[5] "2021-01-06 20:37:21.472981+00" "2021-01-06 20:40:22.145894+00"
> |
```

\* Ta nhận thấy các thời điểm Min, max, median, mean của Temp cũng không trùng nhau.

FILE 4:

Thực hiện tương tự như File, ta cũng thu được kết quả Min, max, median, mean của Humi/Temp như bên dưới:

- Humi:

```
R 4.1.1 - C:/Users/Admin/Desktop/THỰC HÀNH R/ 
> head(Min_Humi_f4)
[1] "2021-03-17 10:40:29.471525+00" NA
[3] NA                               NA
[5] NA                               NA
> head(Max_Humi_f4)
[1] NA NA NA NA NA NA
> head(Median_Humi_f4)
[1] "2021-03-01 17:54:07.592066+00" "2021-03-01 17:57:28.898233+00"
[3] "2021-03-01 18:00:29.076897+00" "2021-03-01 18:03:30.030314+00"
[5] "2021-03-01 18:06:30.398472+00" "2021-03-02 07:27:25.242371+00"
> head(Mean_Humi_f4)
[1] "2021-03-02 17:29:27.551021+00" "2021-03-19 10:26:14.280307+00"
[3] NA                               NA
[5] NA                               NA
> |
```

\* Ta nhận thấy các thời điểm min, max, median, mean của Humi trong file 4 này cũng không trùng nhau.



- Temp:

```
R 4.1.1 - C:/Users/Admin/Desktop/THỰC HÀNH R/ ◊
> summary(Temp)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  0.00 17.00 20.60 23.39 29.90 48.00
> head(Min_Temp_f4)
[1] "2021-03-19 16:11:44.420558+00" "2021-03-20 09:41:05.49627+00"
> head(Max_Temp_f4)
[1] "2021-03-28 14:27:19.503172+00"
> head(Median_Temp_f4)
[1] "2021-03-01 07:40:33.997404+00" "2021-03-01 23:10:51.619374+00"
[3] "2021-03-01 23:13:51.772935+00" "2021-03-02 01:41:20.600227+00"
[5] "2021-03-02 01:50:22.615939+00" "2021-03-02 01:53:23.005152+00"
> head(Mean_Temp_f4)
[1] "2021-03-01 08:16:29.279843+00" "2021-03-01 18:00:29.076897+00"
[3] "2021-03-01 18:03:30.030314+00" "2021-03-01 18:06:30.398472+00"
[5] "2021-03-02 17:19:25.676121+00" "2021-03-02 17:38:29.032862+00"
> |
```

\* Ta nhận thấy các thời điểm min, max, median, mean của Temp trong file 4 này không trùng nhau.

FILE 5:

- Temp:

```
R 4.1.1 - C:/Users/Admin/Desktop/THỰC HÀNH R/ ◊
> summary(Temp)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  0.00 19.80 21.90 25.05 29.50 47.50
> head(Min_Temp_f5)
[1] "2021-04-16 10:00:05.709939+00" "2021-04-16 10:06:06.802943+00"
[3] "2021-04-16 10:09:10.467633+00" "2021-04-16 10:12:16.120087+00"
[5] "2021-04-16 10:24:35.380194+00" "2021-04-16 10:43:35.4235+00"
> head(Max_Temp_f5)
[1] "2021-04-04 13:04:30.508598+00"
> head(Median_Temp_f5)
[1] "2021-04-01 01:56:14.46329+00" "2021-04-01 03:02:22.927818+00"
[3] "2021-04-01 03:05:23.339842+00" "2021-04-01 03:08:23.722822+00"
[5] "2021-04-01 03:11:24.908052+00" "2021-04-01 03:14:24.40741+00"
> head(Mean_Temp_f5)
[1] "2021-04-01 01:56:14.46329+00" "2021-04-01 03:02:22.927818+00"
[3] "2021-04-01 03:05:23.339842+00" "2021-04-01 03:08:23.722822+00"
[5] "2021-04-01 03:11:24.908052+00" "2021-04-01 03:14:24.40741+00"
> |
```

\* Ta nhận thấy các thời điểm median và mean của giá trị Temp trong trường hợp này trùng nhau, còn các giá trị min, max không trùng nhau.

- Humi:



```
R 4.1.1 - C:/Users/Admin/Desktop/THỰC HÀNH R/ ↵
> summary(Humi)
   Min. 1st Qu. Median    Mean 3rd Qu.    Max.    NA's
23.00   54.00   77.00   70.16   83.00   97.40    747
> head(Min_Humi_f5)
[1] NA NA NA NA NA NA
> head(Max_Humi_f5)
[1] "2021-04-01 17:00:22.399762+00" "2021-04-01 17:03:25.451493+00"
[3] NA NA NA NA NA NA
[5] NA NA NA NA NA NA
> head(Median_Humi_f5)
[1] "2021-04-01 00:44:03.006903+00" "2021-04-01 00:47:03.374573+00"
[3] "2021-04-02 17:59:28.895791+00" "2021-04-02 18:14:33.10279+00"
[5] "2021-04-02 18:17:33.947351+00" "2021-04-02 18:20:34.62644+00"
> head(Mean_Humi_f5)
[1] "2021-04-05 14:29:13.423476+00" NA NA NA NA NA NA
[3] NA NA NA NA NA NA NA
[5] NA NA NA NA NA NA NA
> |
```

- 7) Các mốc thời điểm Q1, Q2, Q3 có trùng nhau không?
- 8) Các mốc thời điểm các outlier có trùng nhau không?
- 9) Cho nhận xét của các bạn theo các tập mẫu mà nhóm đã phân tích
- 10) Hãy mô tả mối quan hệ tuyến tính giữa nhiệt độ và độ ẩm bằng cách đo độ kết hợp của mối quan hệ dùng correlation r (correlation coefficient) và hướng kết hợp.