

COS3302 MIDTERM

Basic Statistical Analysis and

Visualization

with

Occupancy Detection Dataset

จัดทำโดย

นายศุภณัฐ แซ่เตีย

ID: 6505000270



Agenda

- ที่มาของชุดข้อมูล
- ลักษณะของชุดข้อมูล
- คุณสมบัติของแต่ละ Feature
- ความสัมพันธ์ของแต่ละ Feature
- Data Preprocess

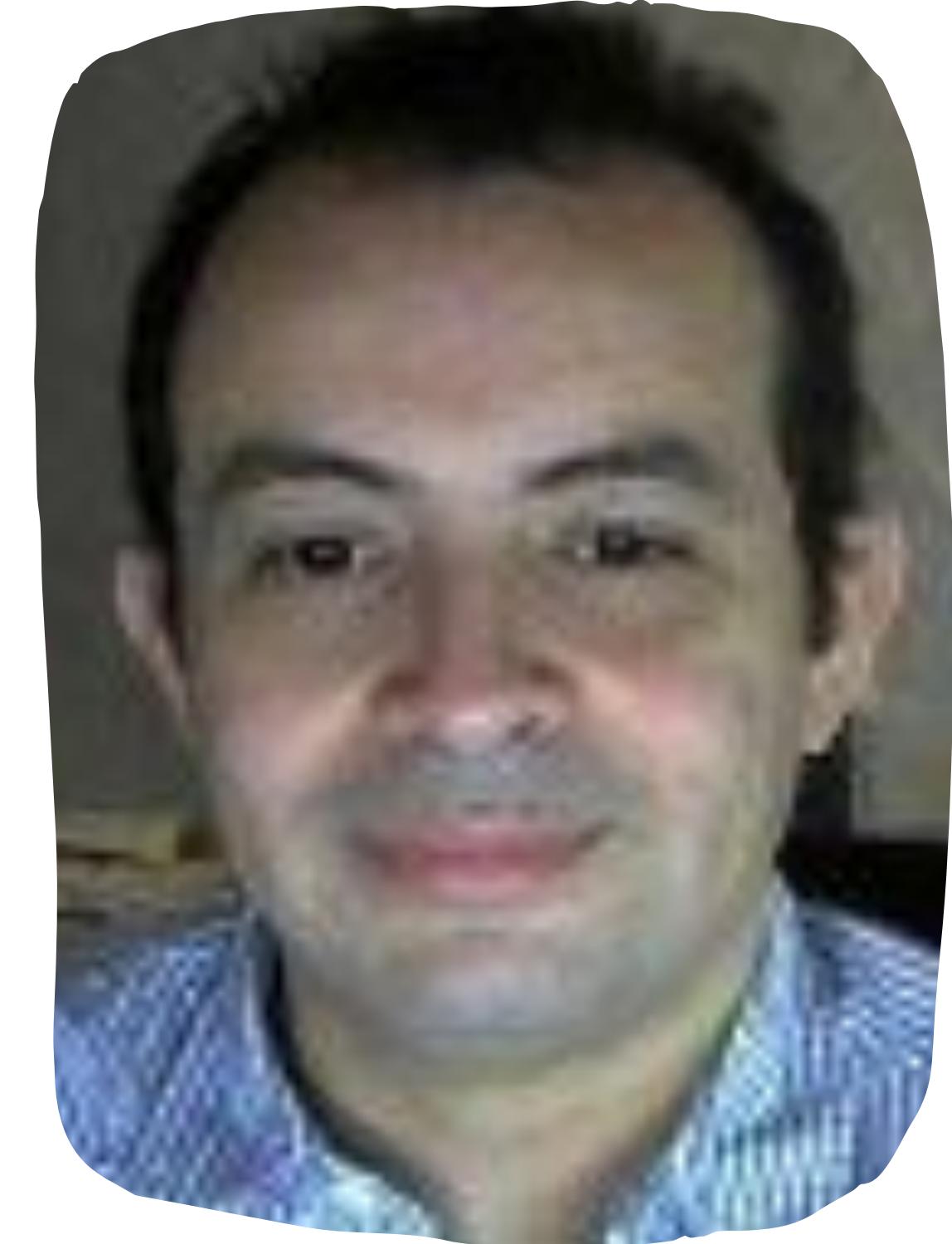
ที่มาของชุดข้อมูล



ที่มาของชุดข้อมูล

Occupancy Detection Dataset เป็นชุดข้อมูลที่มาจากการทดลองในงานวิจัยเรื่อง “*Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models*” โดยคุณ Luis Miguel Candanedo Ibarra ในปี ค.ศ. 2015

คุณ Luis M. Candanedo

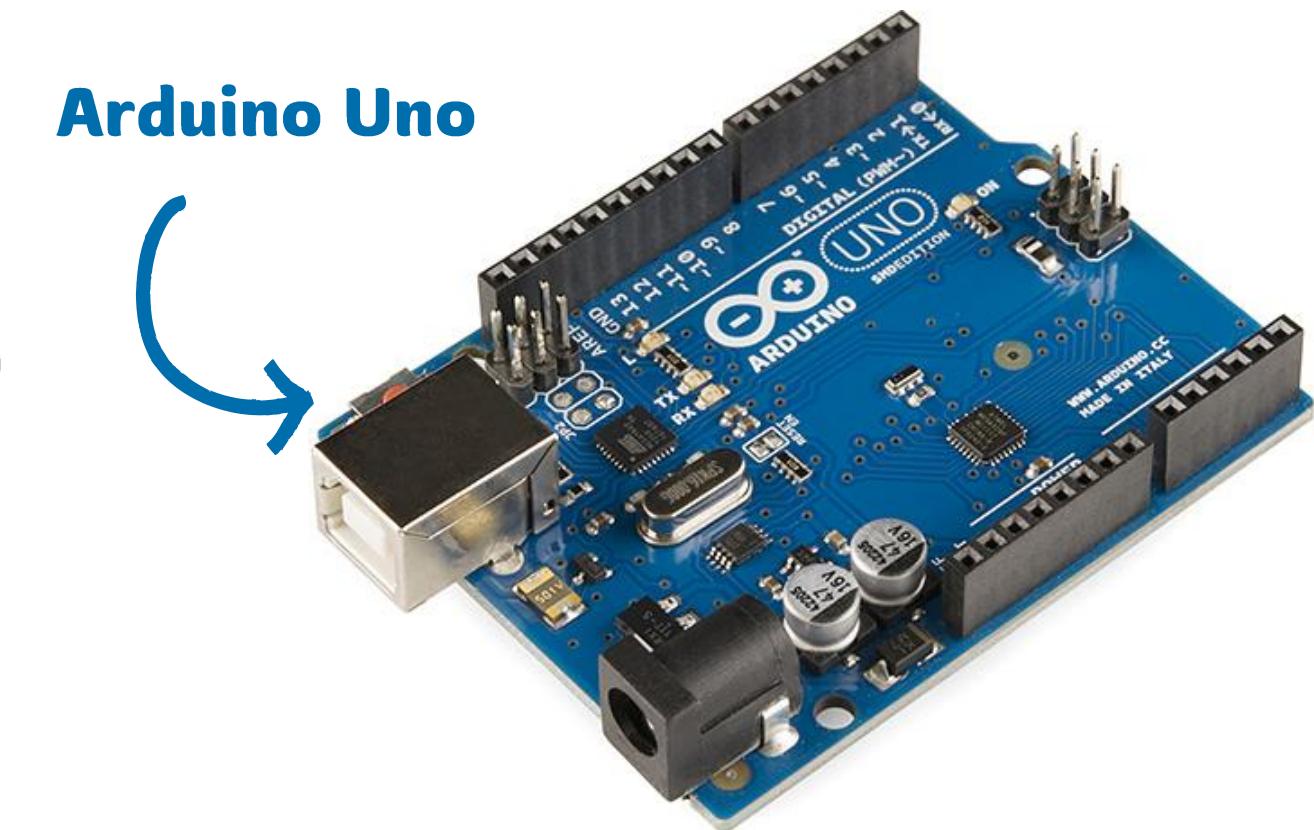


ที่มาของชุดข้อมูล

- เป็นการทดลองเก็บค่าอุณหภูมิ ความชื้น แสง และระดับน้ำบนไดอ็อกไซด์ กายในห้องสำนักงานขนาด 5.85 ม. × 3.50 ม. × 3.53 ม. (กว้าง × ลึก × สูง) โดยใช้บอร์ด Arduino และ Raspberry Pi ร่วมกับเซ็นเซอร์ตรวจวัดค่าต่าง ๆ



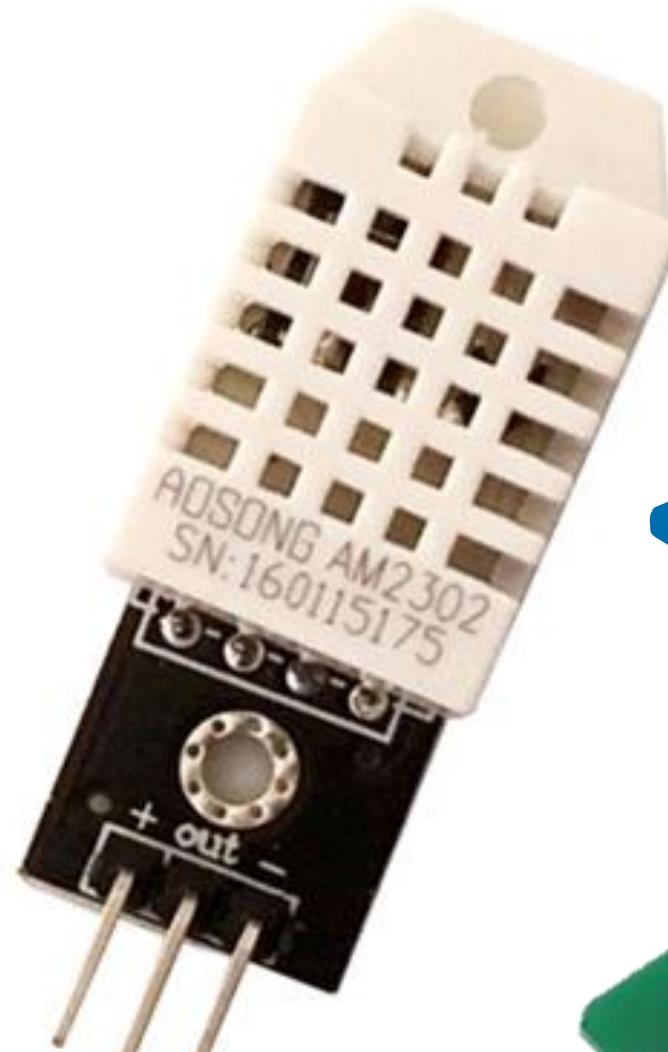
ภาพจาก: <https://pcnautic.com/en/product/raspberry-pi-1-model-b-1>



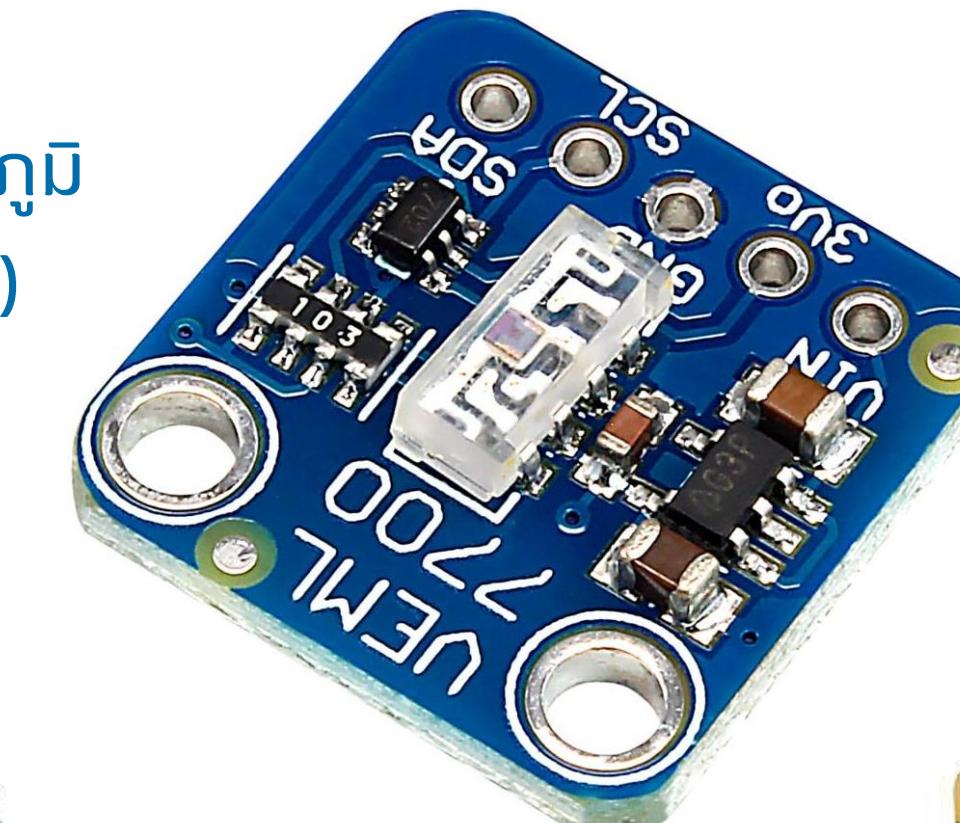
ภาพจาก: https://en.wikipedia.org/wiki/File:Arduino_Uno_-_R3.jpg

- การทดลองจะใช้กล้องในการถ่ายภาพเพื่อบันทึกคนหรือไม่ และใช้เซ็นเซอร์ต่าง ๆ ในการเก็บค่าภายในห้อง โดยการเก็บค่าจะทำในทุก 1 นาที

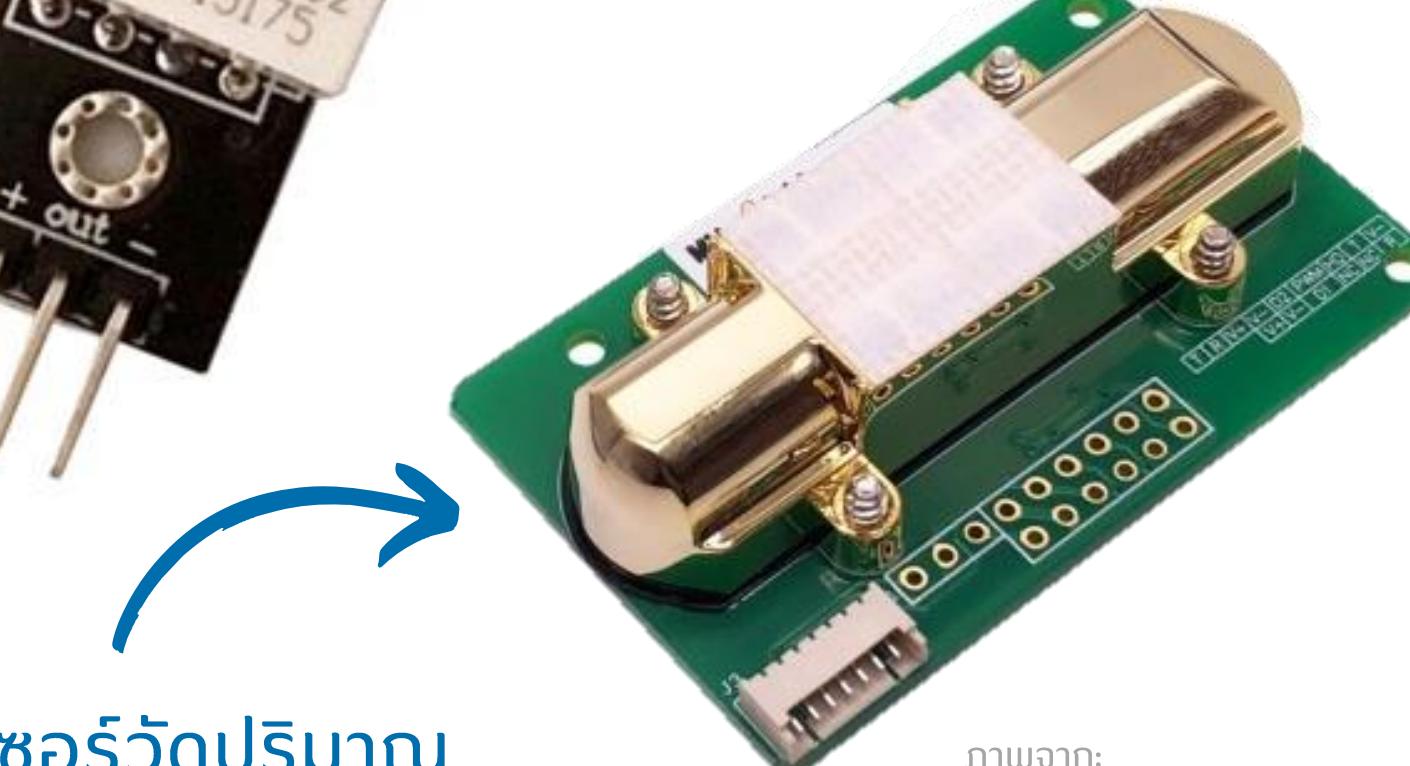
ที่มาของชุดข้อมูล



เซ็นเซอร์ตรวจวัดอุณหภูมิ
และความชื้น (DHT22)

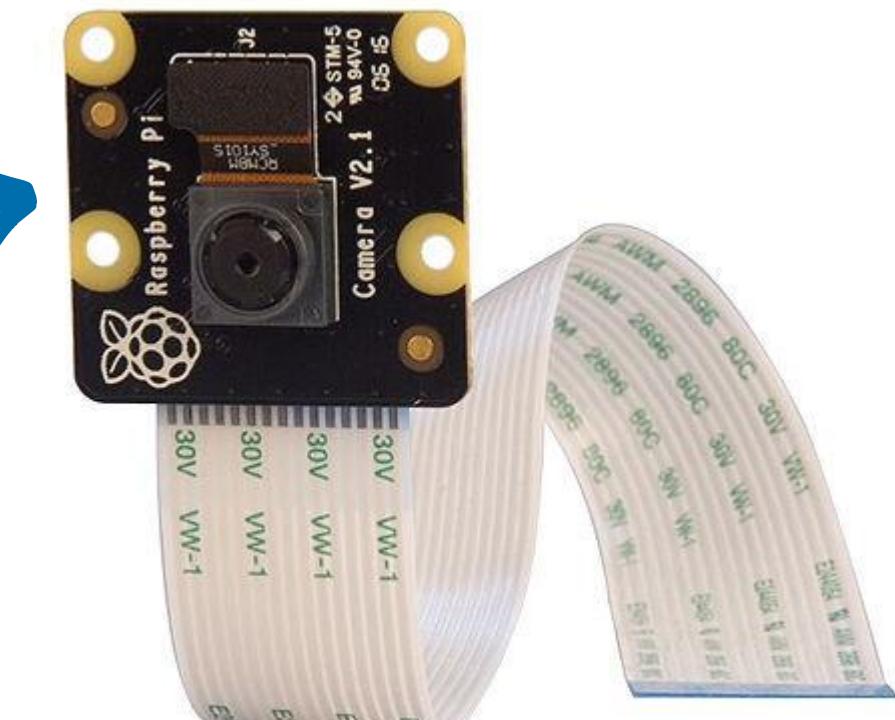


เซ็นเซอร์ตรวจวัดแสงสว่าง



เซ็นเซอร์วัดปริมาณ
คาร์บอนไดออกไซด์

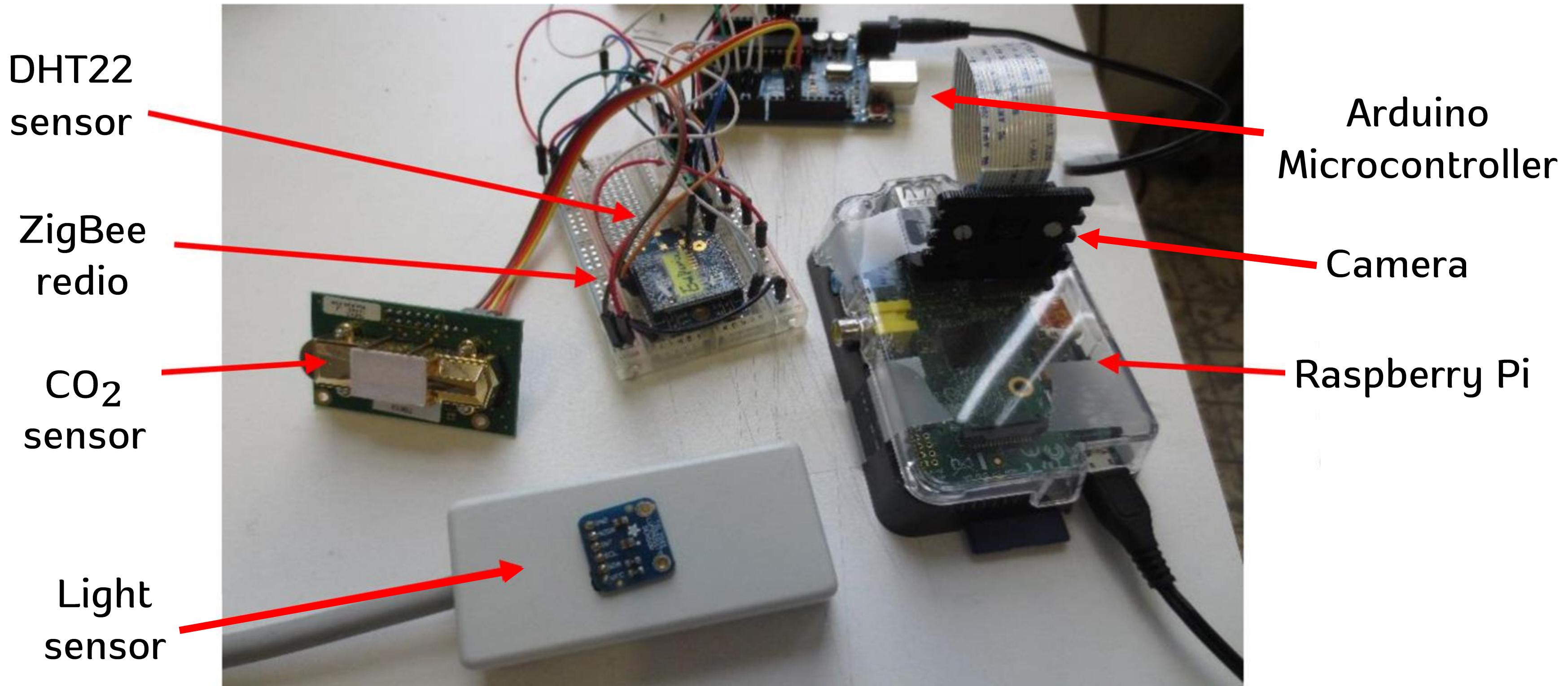
กล้อง



ภาพจาก:

- <https://components101.com/sensors/dht22-pinout-specs-datasheet>
- <https://electropeak.com/veml7700-ambient-light-sensor-module>
- <https://www.tinytronics.nl/en/sensors/air/gas/winsen-mh-z14a-co2-sensor-with-cable>
- <https://www.thaieasyelec.com/product/raspberry-pi-camera-v2-8mp/11000833173000465>

ที่มาของชุดข้อมูล



តាកេណៈខែងចុះមូល



ลักษณะของชุดข้อมูล

- ชุดข้อมูล Occupancy Detection ประกอบด้วย 20,560 ตัวอย่าง
- แบ่งออกเป็น 3 ชุด ได้แก่
 - data training มี 8,143 ตัวอย่าง
 - data test 1 มี 2,665 ตัวอย่าง
 - data test 2 มี 9,752 ตัวอย่าง
- ชุดข้อมูลประกอบด้วย 8 Feature ได้แก่
 - ID
 - Date
 - Temperature
 - Humidity
 - Light
 - CO₂
 - Humidity Ratio
 - Occupancy

គុណសមប័តីទូទៅនៃ Feature

ID

- เป็น Identifier Features ประเภทเชิงคุณภาพ (Categorical)
ชนิดไม่มีลำดับ (Nominal)
- กำหนดให้ชี้ไปยังตัวอย่างใด ๆ ในชุดข้อมูล
- ข้อมูลจะเป็นตัวเลขจำนวนเต็มที่ไม่ซ้ำกัน
- เป็น Feature ที่ไม่มีผลต่อการคำนایของโมเดล (Irrelevant Features)

id
1
2
3
4
5

ตัวอย่างข้อมูล



Date

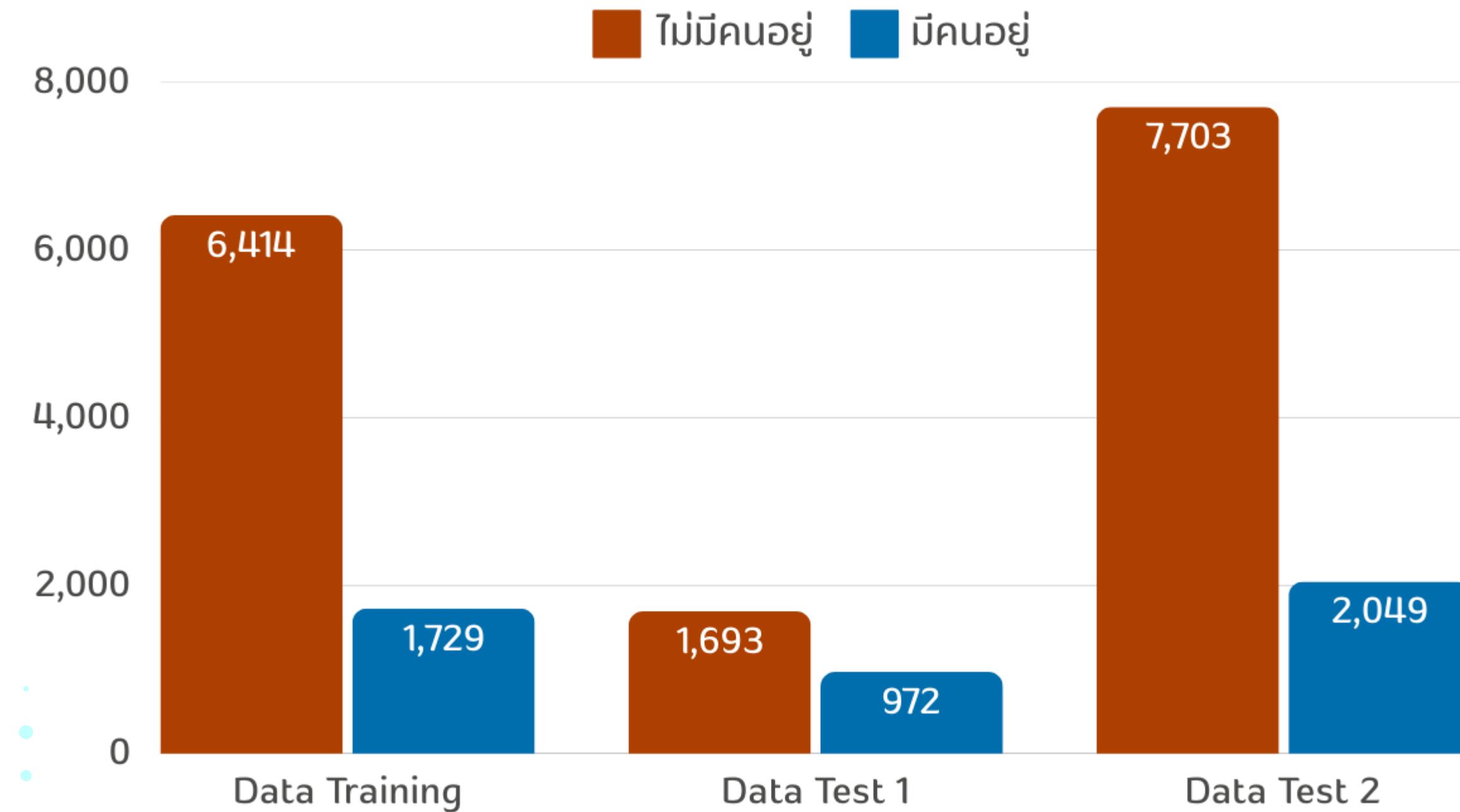
- เป็น Features ประเภท Time Feature
- กำหนดที่ระบุวันที่และเวลาที่ทำการบันทึกข้อมูล
ตัวอย่างนี้ ๆ
- โดยข้อมูลจะถูกໃນลักษณะ YYYY-MM-DD
hh:mm:ss

ตัวอย่างข้อมูล

date
2015-02-11 14:48:00
2015-02-11 14:49:00
2015-02-11 14:50:00
2015-02-11 14:51:00
2015-02-11 14:51:59

Occupancy

- เป็น Target Feature ประเภทเชิงคุณภาพ (Categorical) ในกลุ่ม Binary
- กำหนดให้เก็บค่าว่ามีคนอยู่ในห้องจริงหรือไม่ โดย 0 แทนไม่มีคน และ 1 แทนมีคน



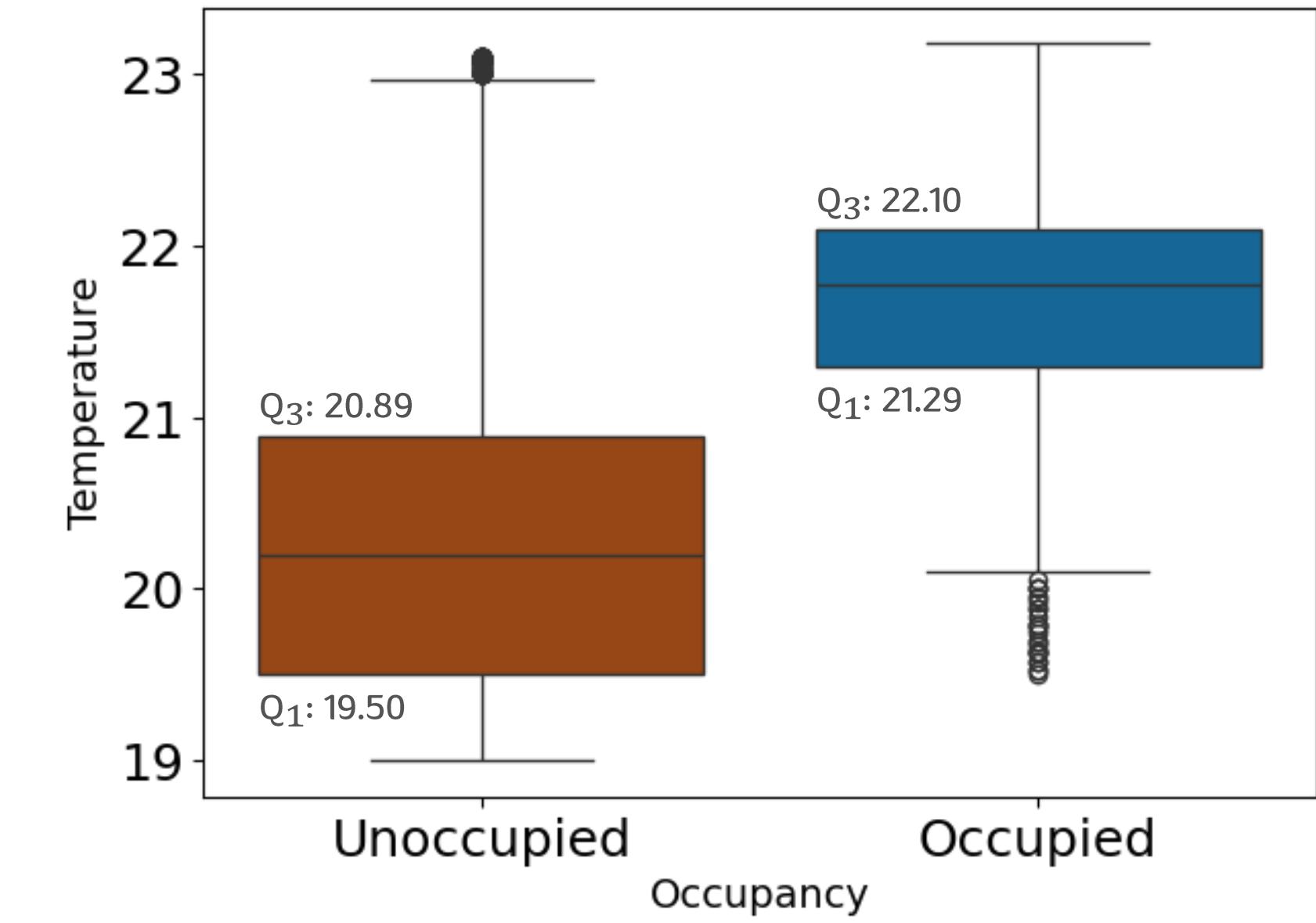
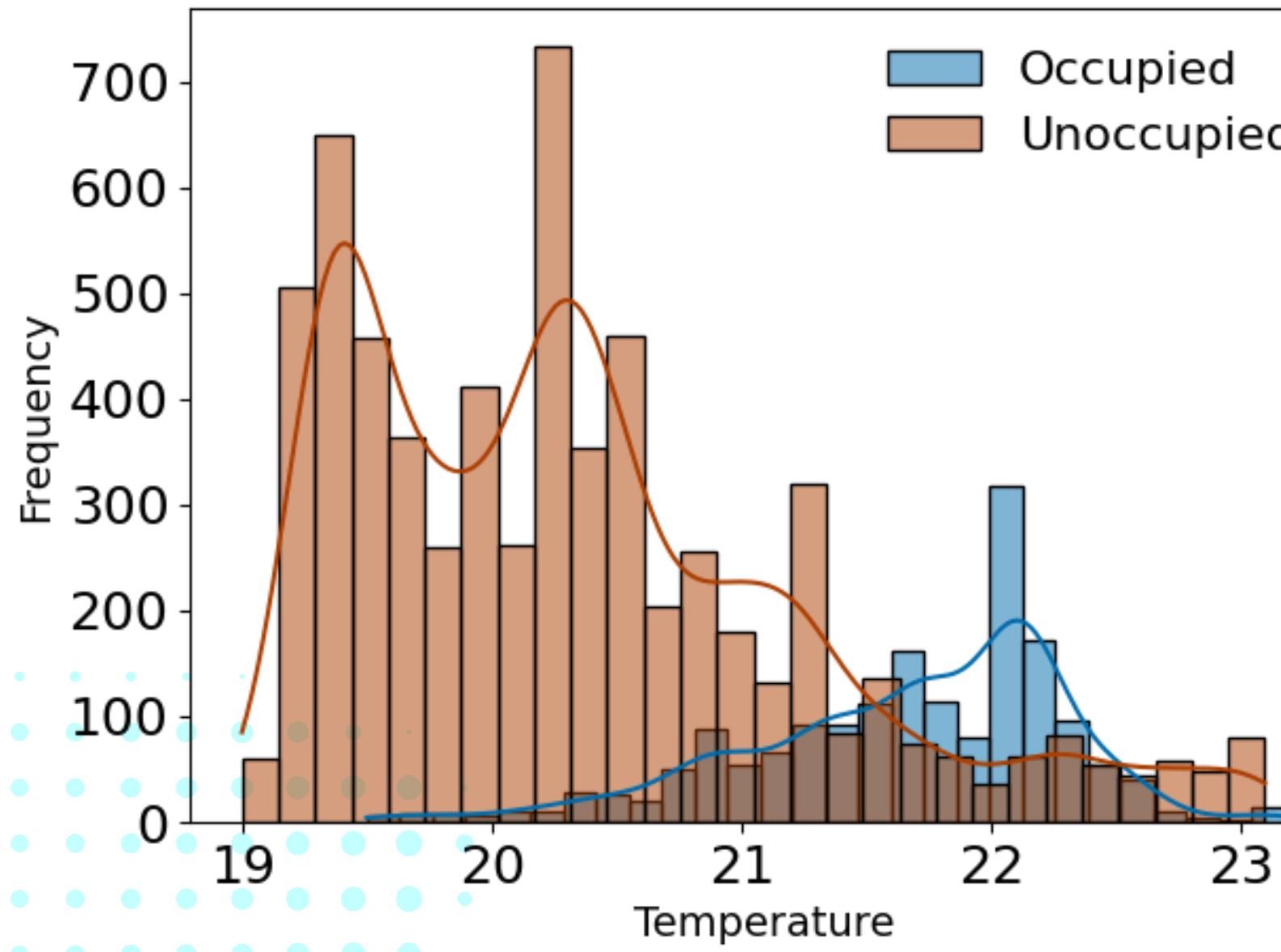
Temperature

- เป็น Feature ประเภทเชิงปริมาณ (Numerical) ชนิดอันตรภาค (Interval)
- กำหนดให้ค่าอุณหภูมิภายในห้อง ในหน่วยองศาเซลเซียส ($^{\circ}\text{C}$)

	Training	Test 1	Test 2		Training	Test 1	Test 2
Mean	20.62	21.43	21.00	Q₁	19.70	20.65	20.29
Median	20.39	20.89	20.79	Q₂	20.39	20.89	20.79
S.D.	1.02	1.03	1.02	Q₃	21.39	22.36	21.53
Min	19.00	20.20	19.50	Max	23.18	24.41	24.39

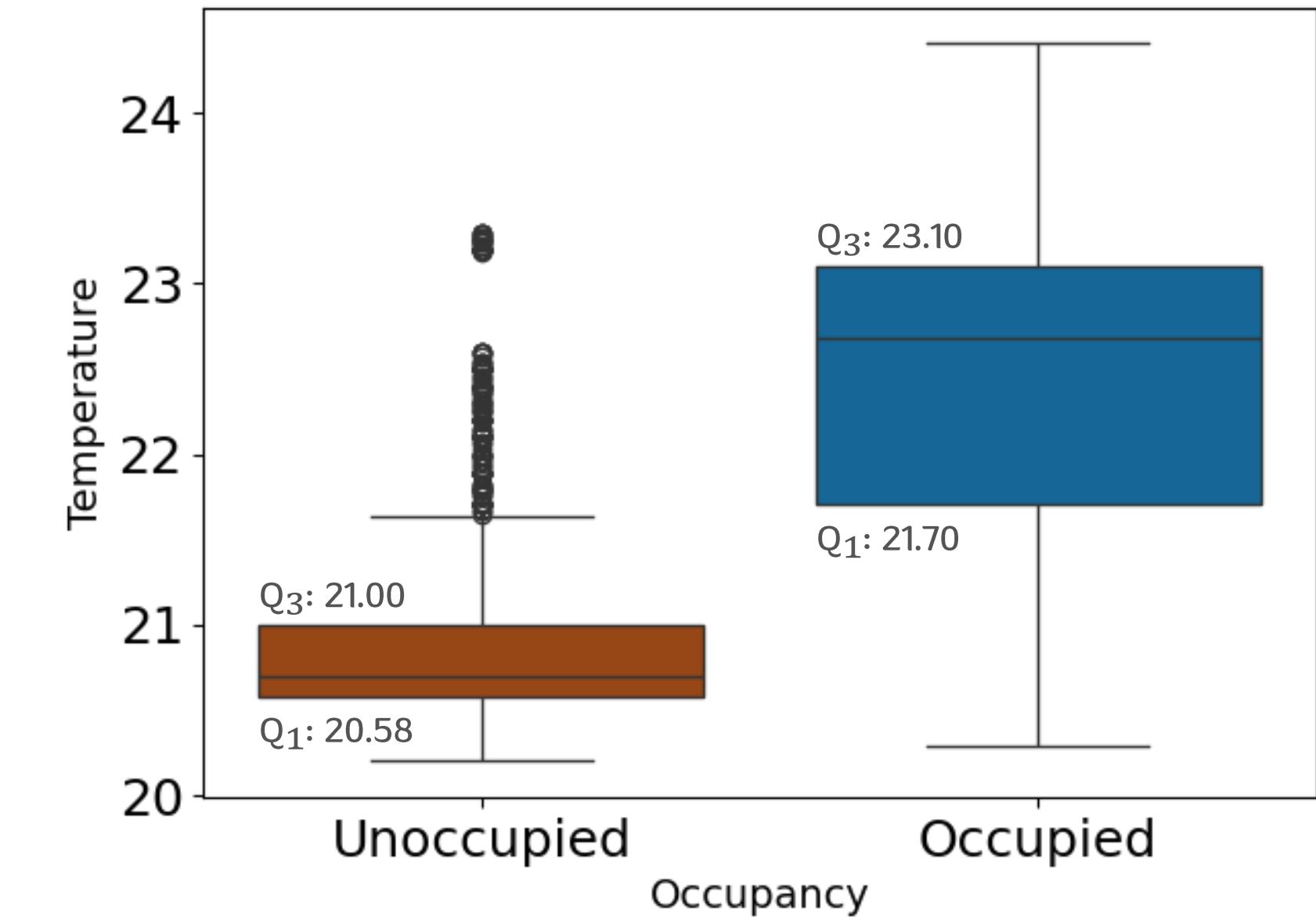
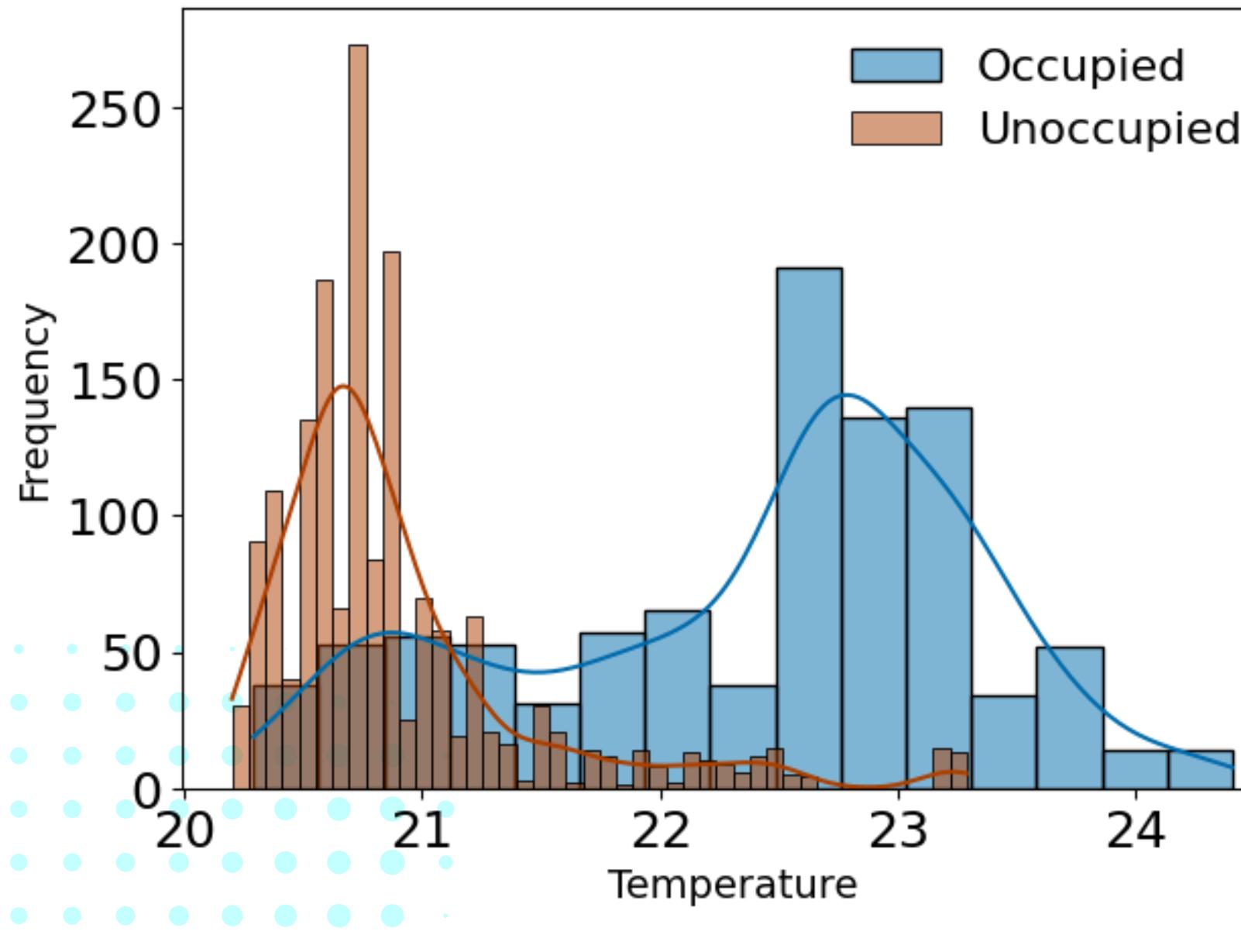
Temperature

- ข้อมูลในชุด Data Training มีลักษณะเป้าไปทางขวาเล็กน้อย (Right-skewed)
- อุณหภูมิในช่วง $19.50 - 20.89^{\circ}\text{C}$ จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- อุณหภูมิในช่วง $21.29 - 22.10^{\circ}\text{C}$ จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



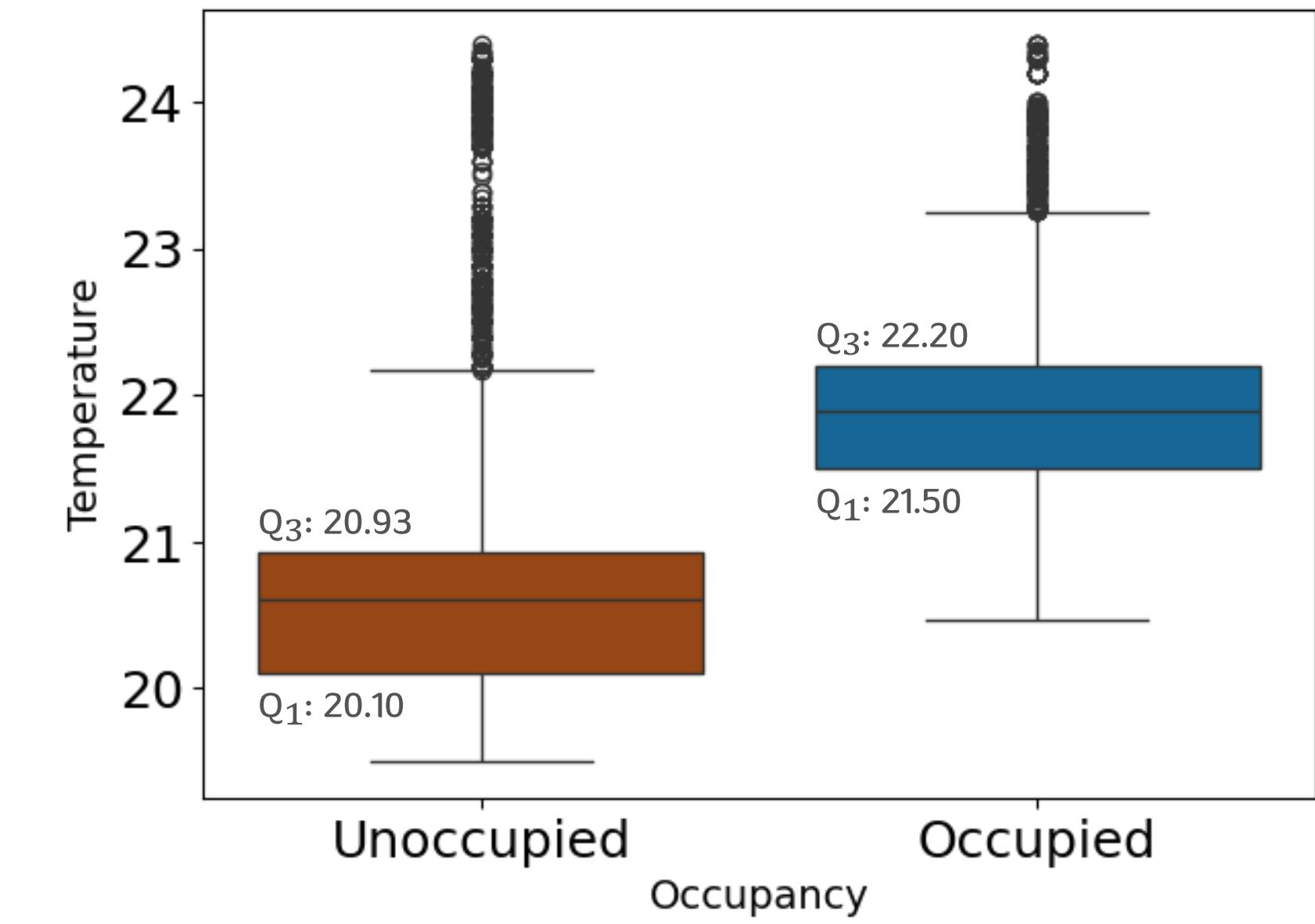
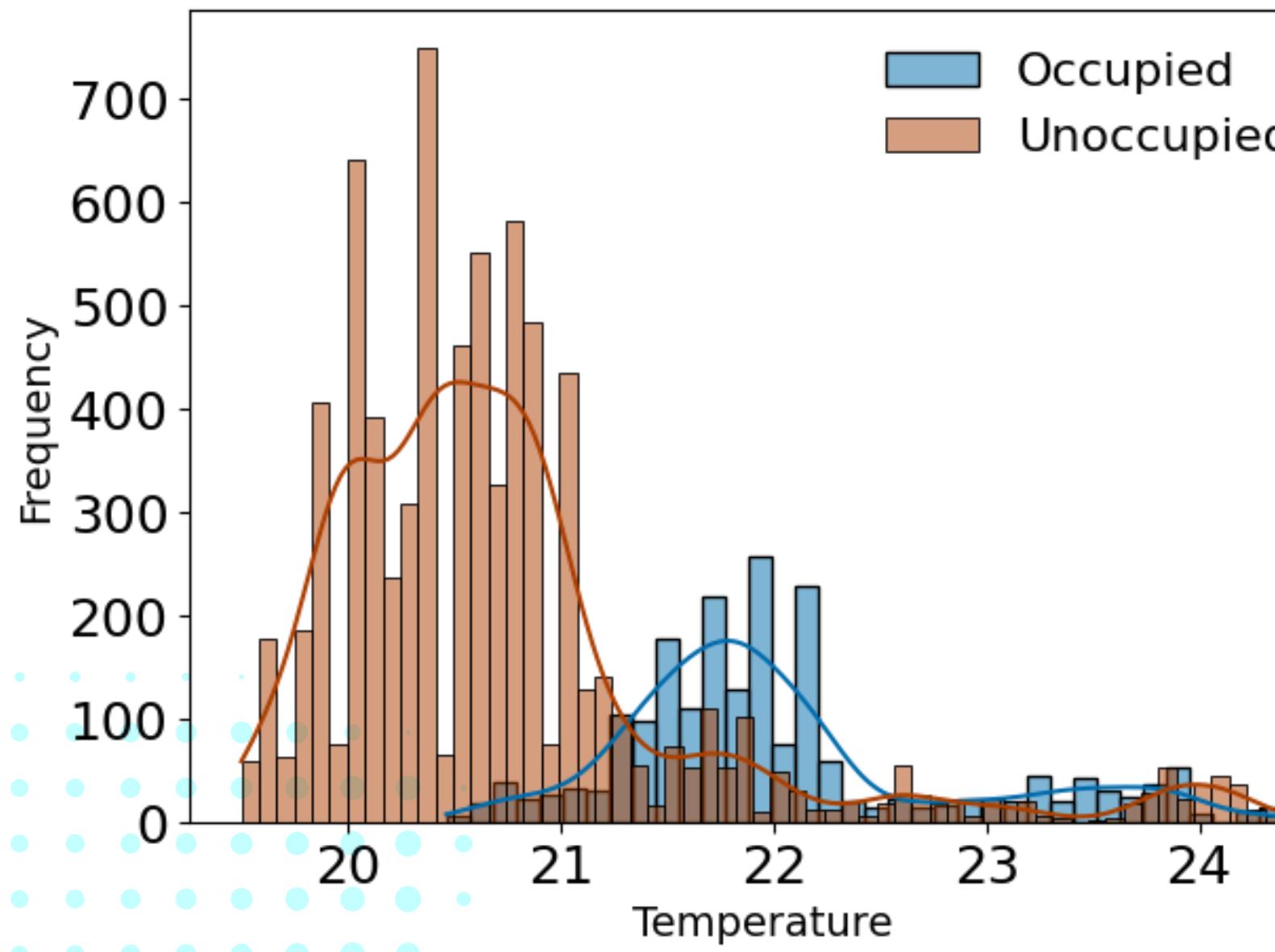
Temperature

- ข้อมูลในชุด Data Test 1 มีลักษณะเบ้าไปทางขวาในระดับปานกลาง (Right-skewed)
- อุณหภูมิในช่วง $20.58 - 21.00\text{ }^{\circ}\text{C}$ จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- อุณหภูมิในช่วง $21.70 - 23.10\text{ }^{\circ}\text{C}$ จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



Temperature

- ข้อมูลในชุด Data Test 2 มีลักษณะเบ้าไปทางขวาในระดับสูง (Right-skewed)
- อุณหภูมิในช่วง $20.10 - 20.93^{\circ}\text{C}$ จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- อุณหภูมิในช่วง $21.50 - 22.20^{\circ}\text{C}$ จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



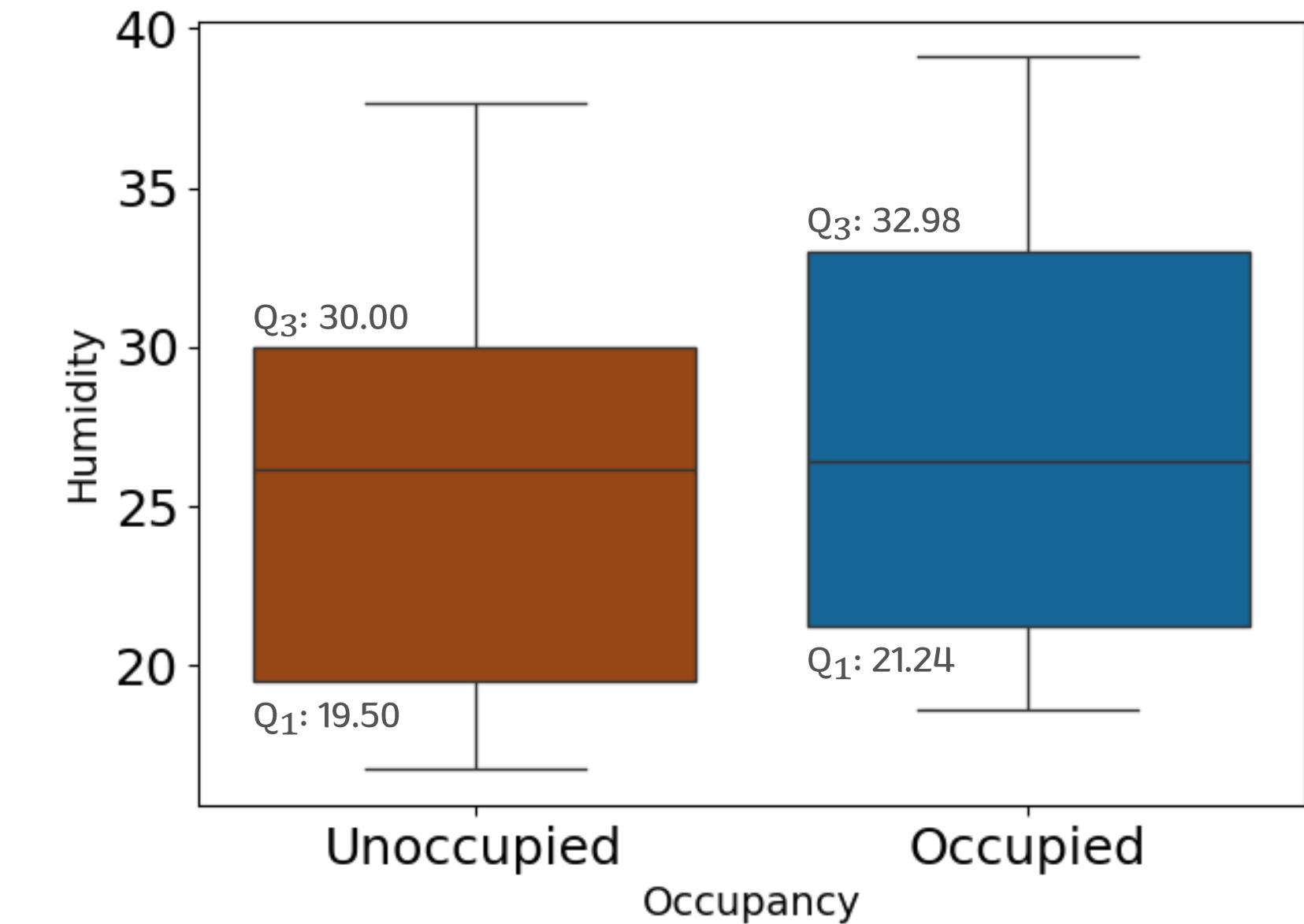
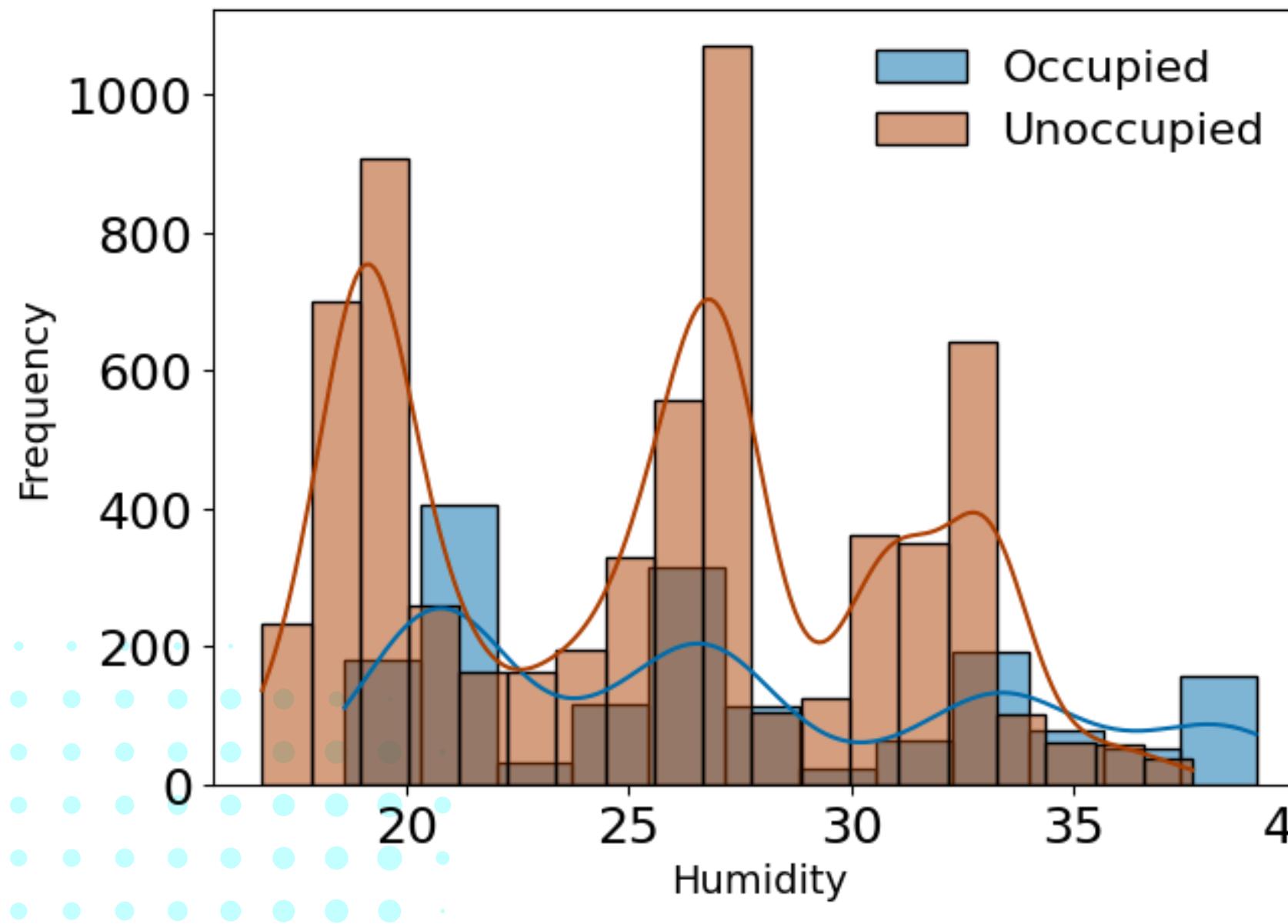
Humidity

- เป็น Feature ประเภทเชิงปริมาณ (Numerical) ชนิดอัตราส่วน (Ratio)
- กำหนดให้ค่าร้อยละความชื้นสัมพัทธ์ภายในห้อง

	Training	Test 1	Test 2		Training	Test 1	Test 2
Mean	25.73	25.35	29.89	Q₁	20.20	23.26	26.64
Median	26.22	25.00	30.20	Q₂	26.22	25.00	30.20
S.D.	5.53	2.44	3.95	Q₃	30.53	26.86	32.70
Min	16.75	22.10	21.87	Max	39.12	31.47	39.50

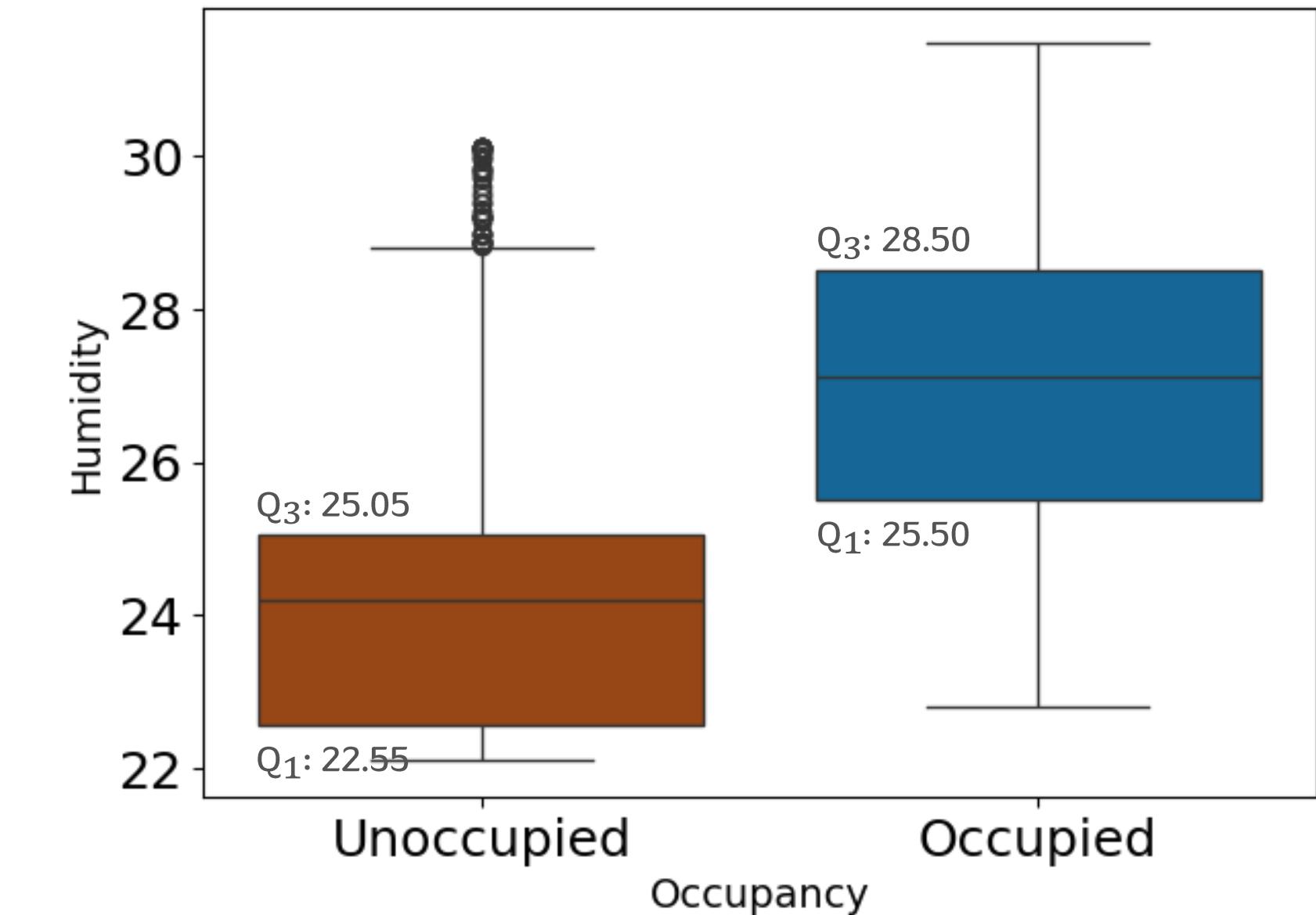
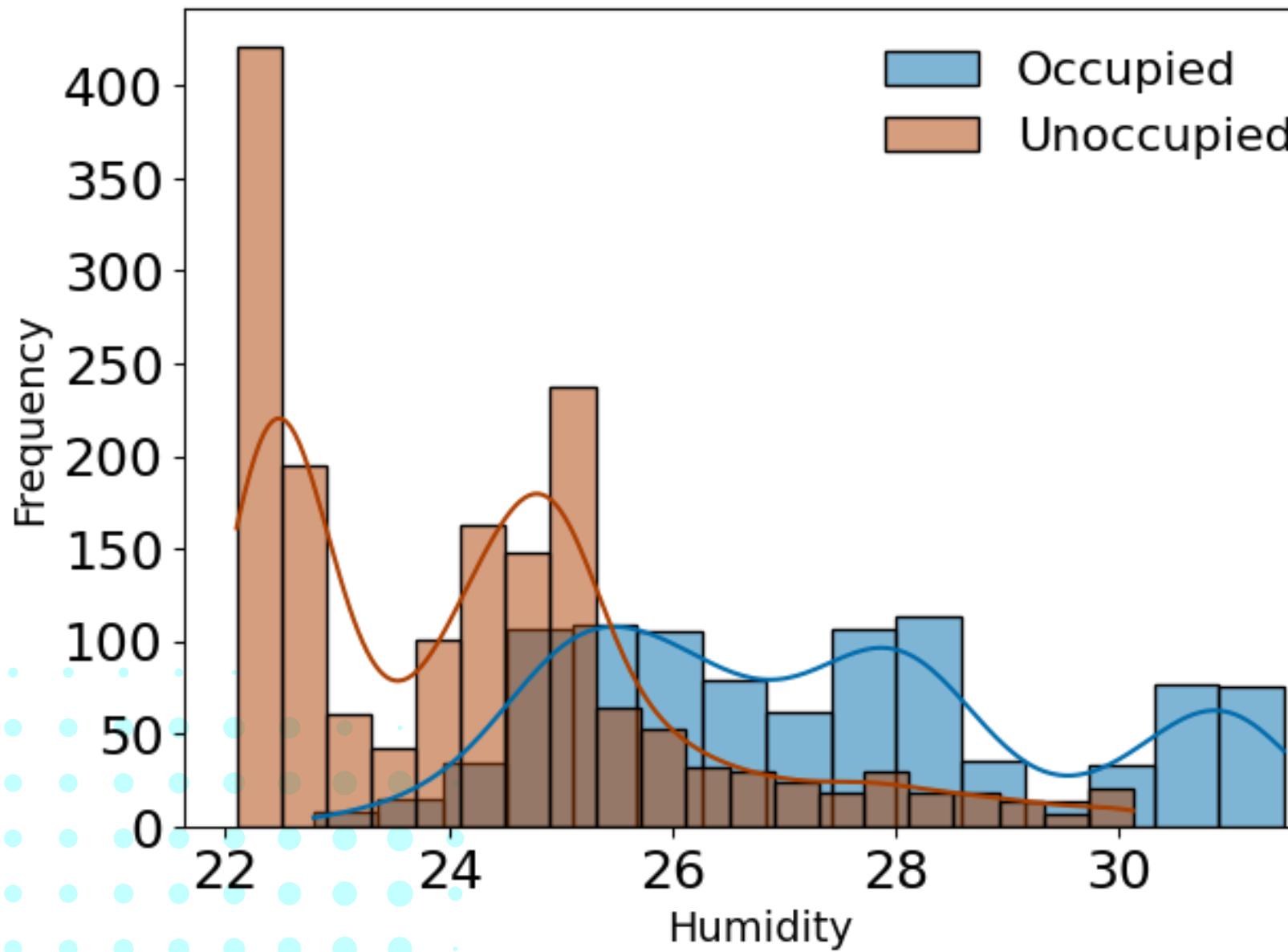
Humidity

- ข้อมูลในชุด Data Training มีลักษณะเป้าไปทางขวาเล็กน้อย (Right-skewed)
- ความชื้นสัมพัทธ์ในช่วง 19.50% - 30.00% จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- ความชื้นสัมพัทธ์ในช่วง 21.24% - 32.98% จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



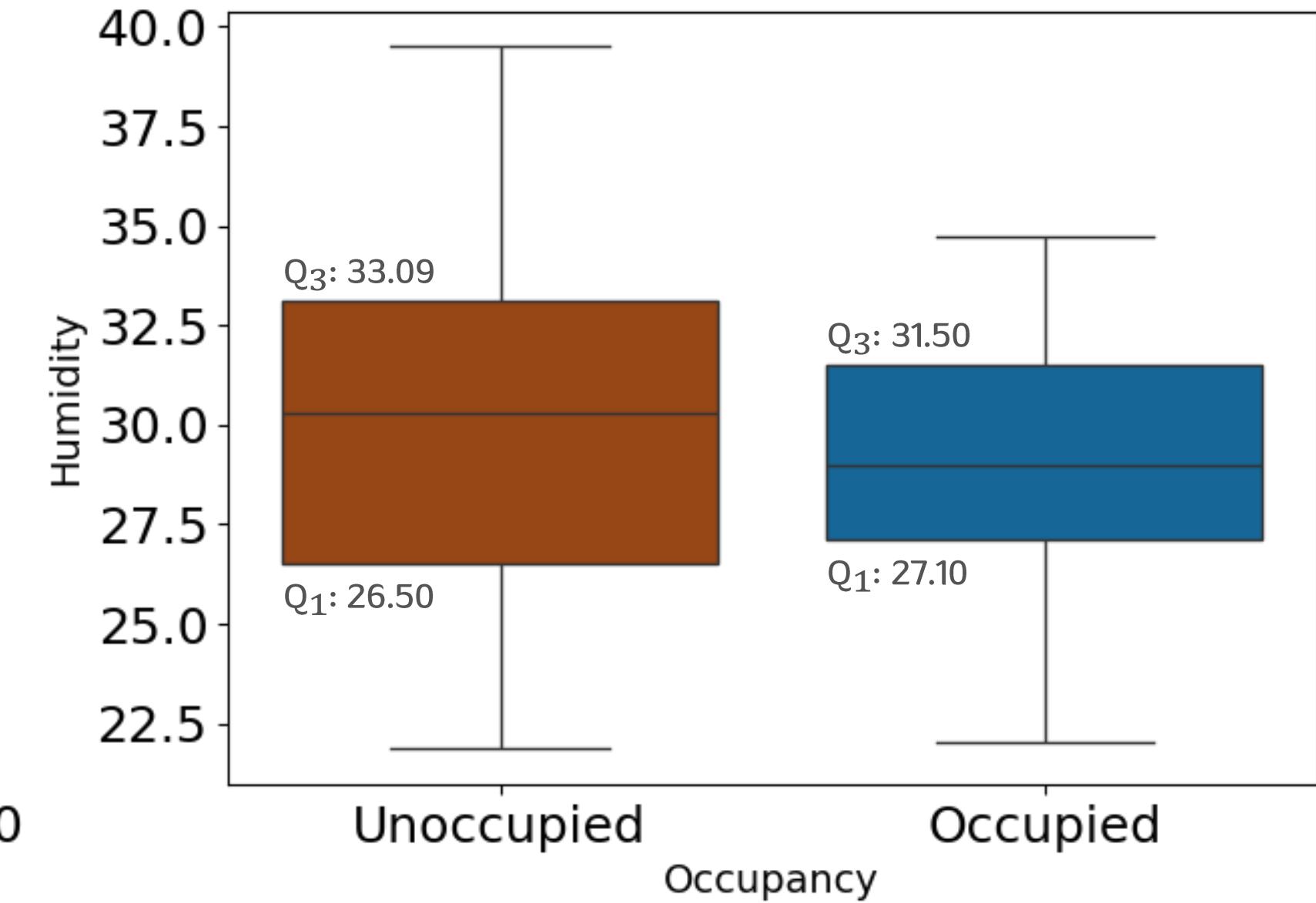
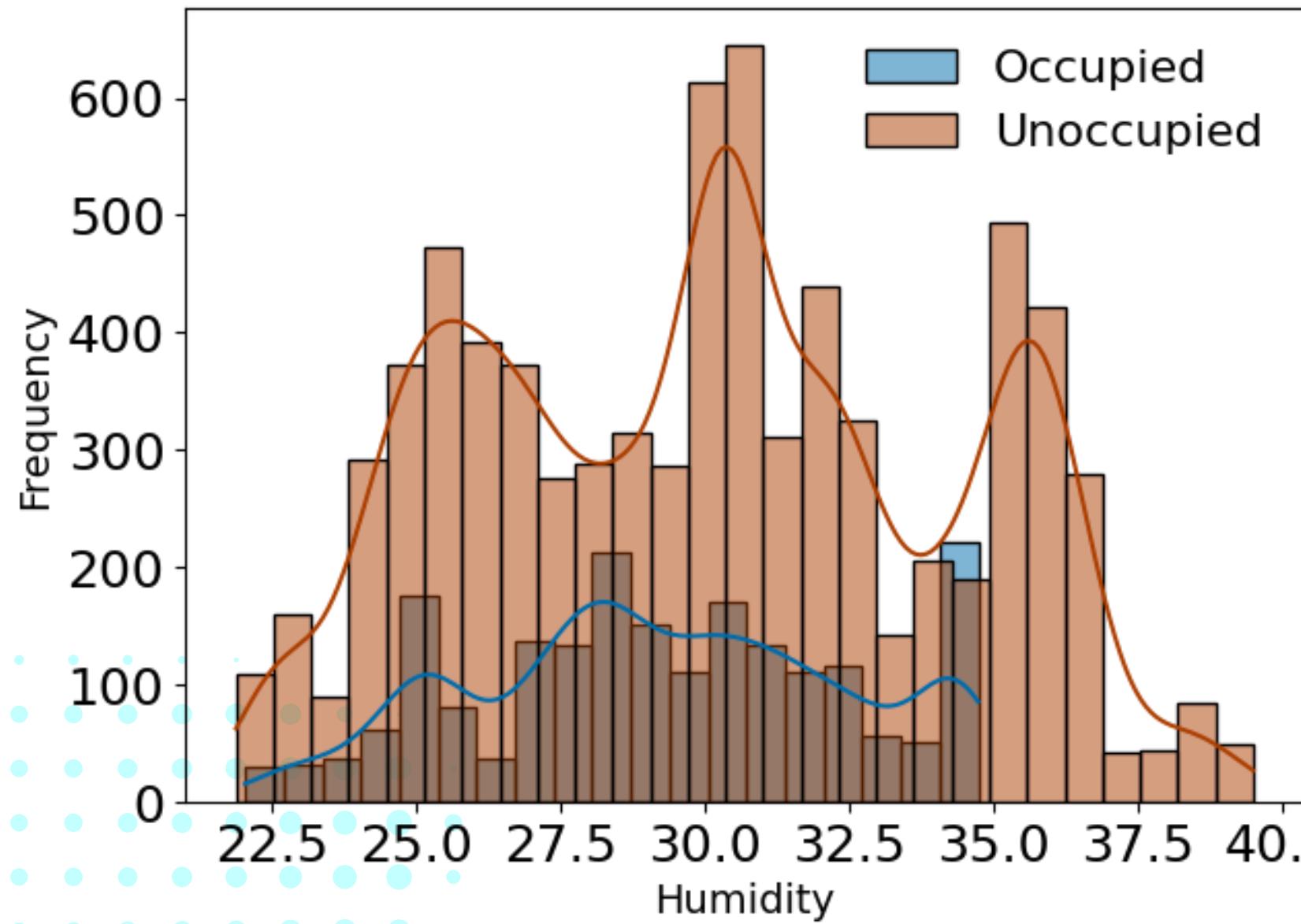
Humidity

- ข้อมูลในชุด Data Test 1 มีลักษณะเบ้าไปทางขวาในระดับปานกลาง (Right-skewed)
- ความชื้นสัมพัทธ์ในช่วง 22.55% - 25.05% จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- ความชื้นสัมพัทธ์ในช่วง 25.50% - 28.50% จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



Humidity

- ข้อมูลในชุด Data Test 2 มีลักษณะเป้าไปทางขวาเล็กน้อย (Right-skewed)
- ความชื้นสัมพัทธ์ในช่วง 26.50% - 33.09% จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- ความชื้นสัมพัทธ์ในช่วง 27.10% - 31.50% จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



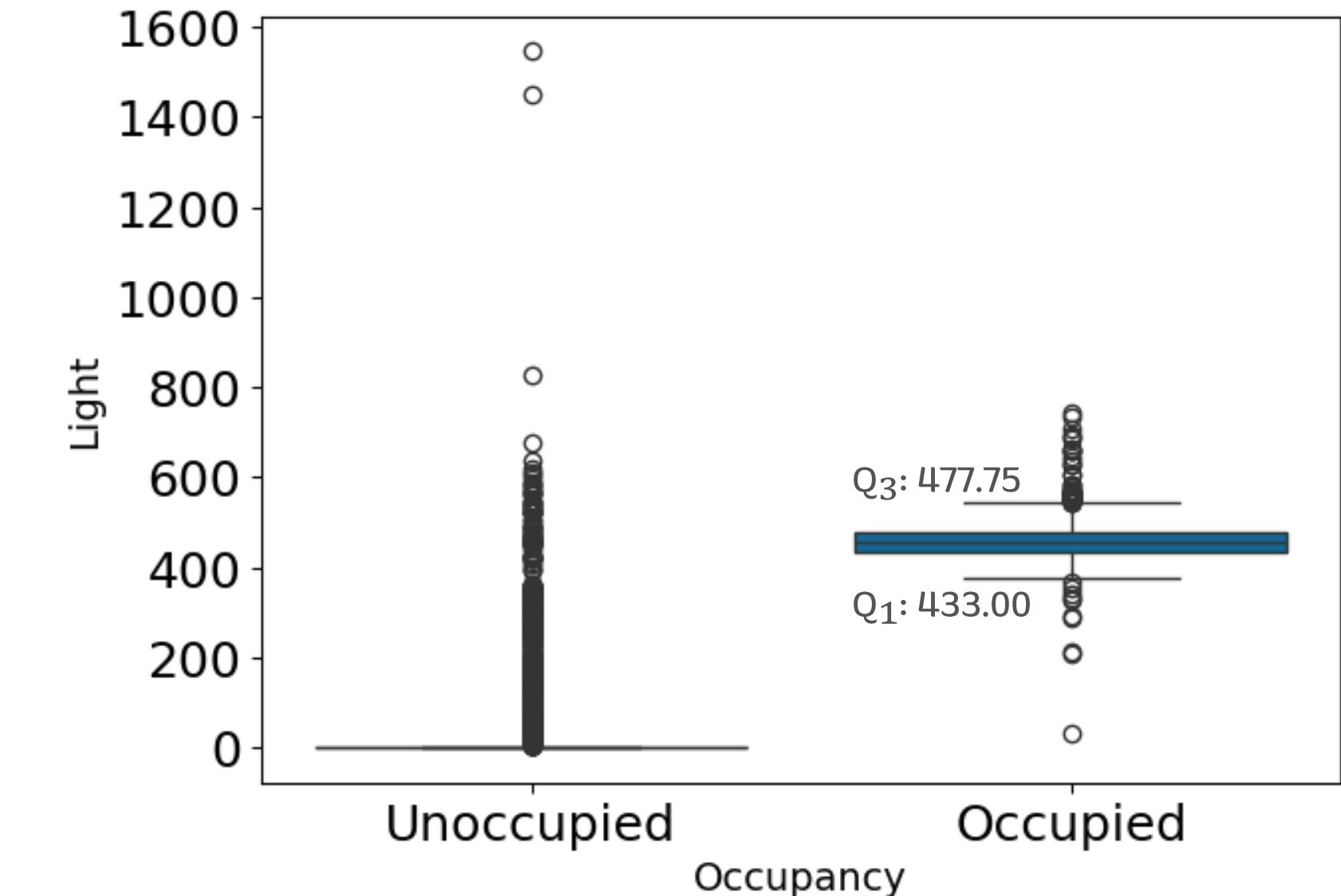
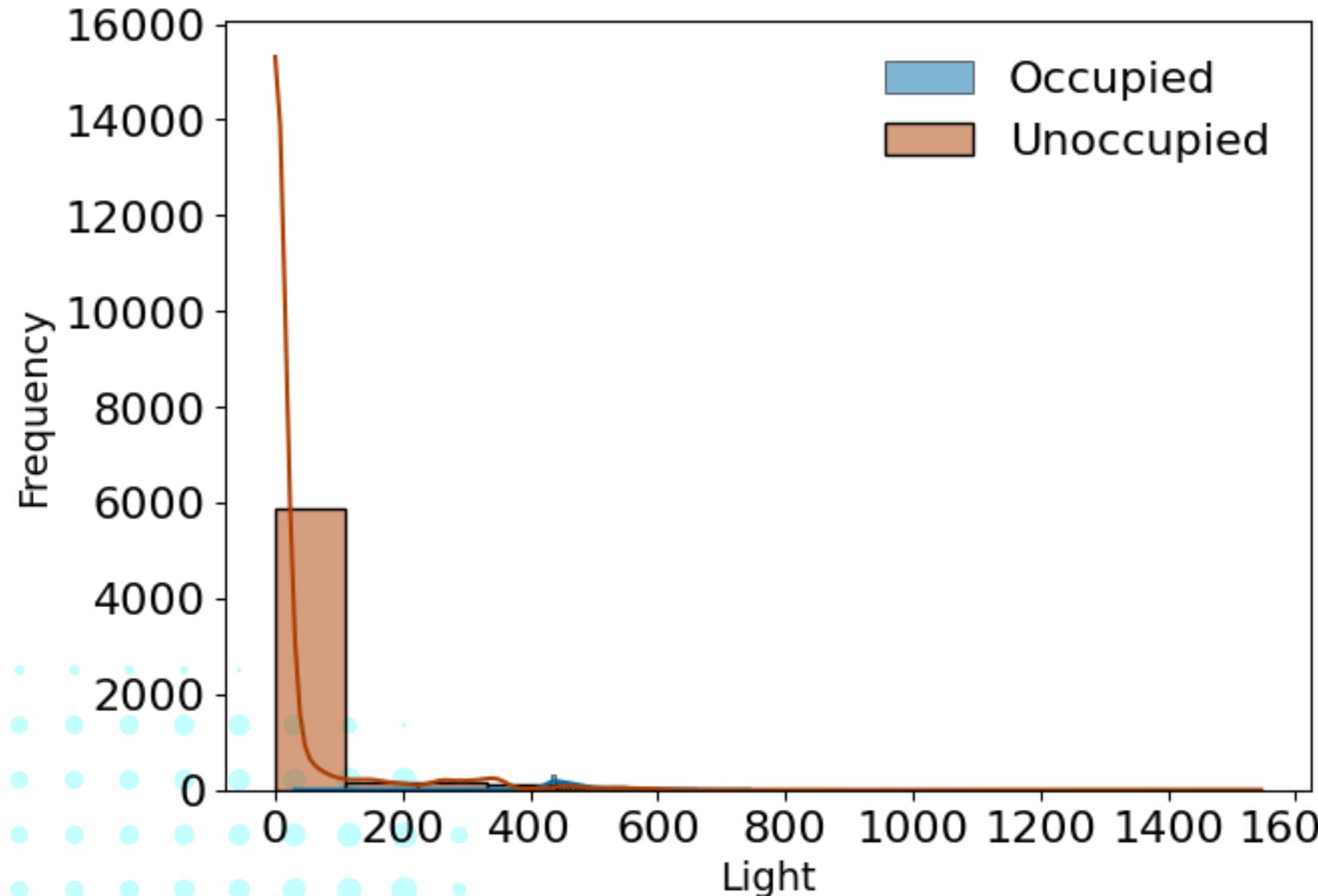
Light

- เป็น Feature ประเภทเชิงปริมาณ (Numerical) ชนิดอัตราส่วน (Ratio)
- กำหนดให้ค่าความสว่างภายในห้อง ในหน่วย Lux

	Training	Test 1	Test 2		Training	Test 1	Test 2
Mean	119.52	193.23	123.07	Q₁	0.00	0.00	0.00
Median	0.00	0.00	0.00	Q₂	0.00	0.00	0.00
S.D.	194.76	250.21	208.22	Q₃	256.36	442.5	208.25
Min	0.00	0.00	0.00	Max	1,546.33	1,697.25	1,581.00

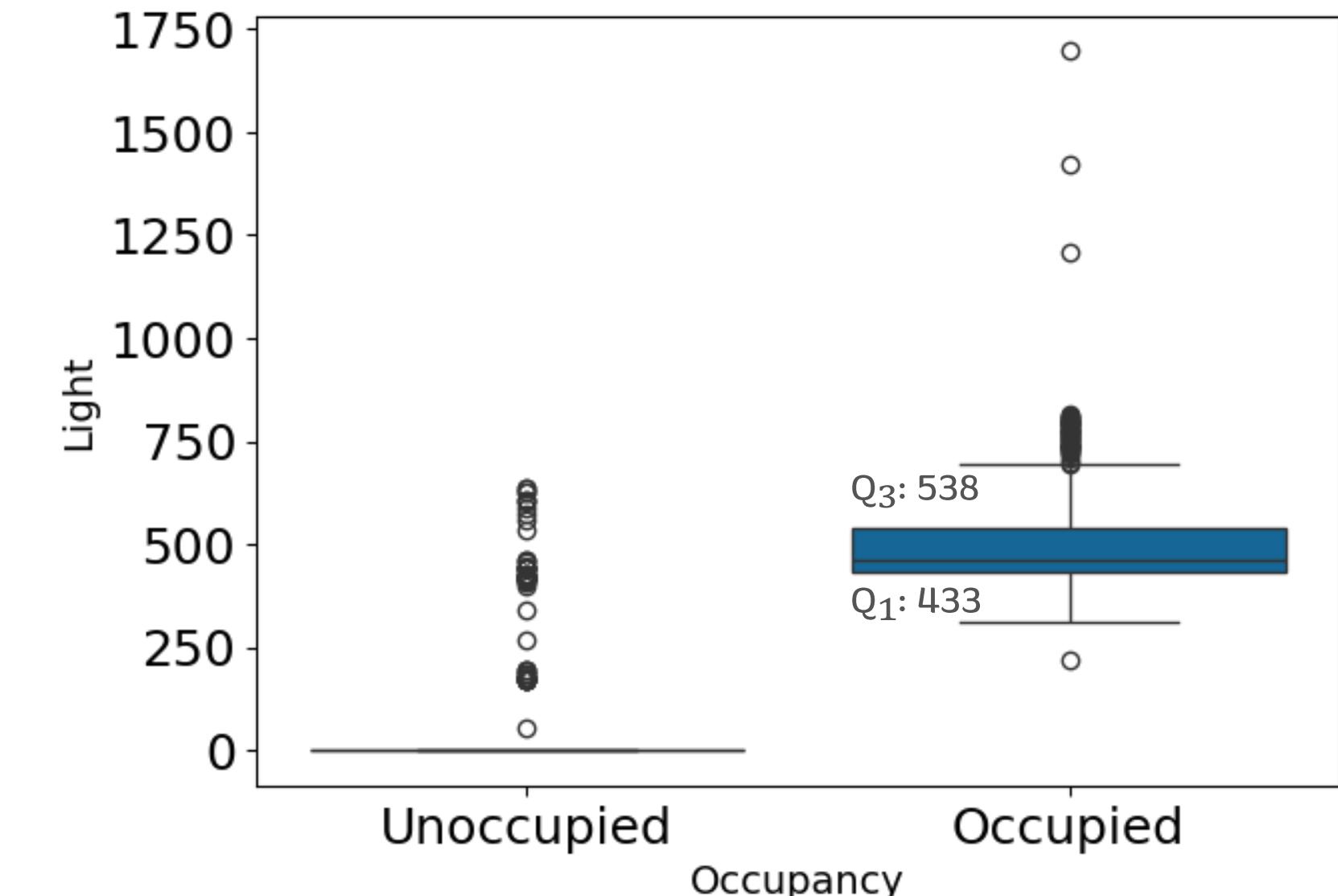
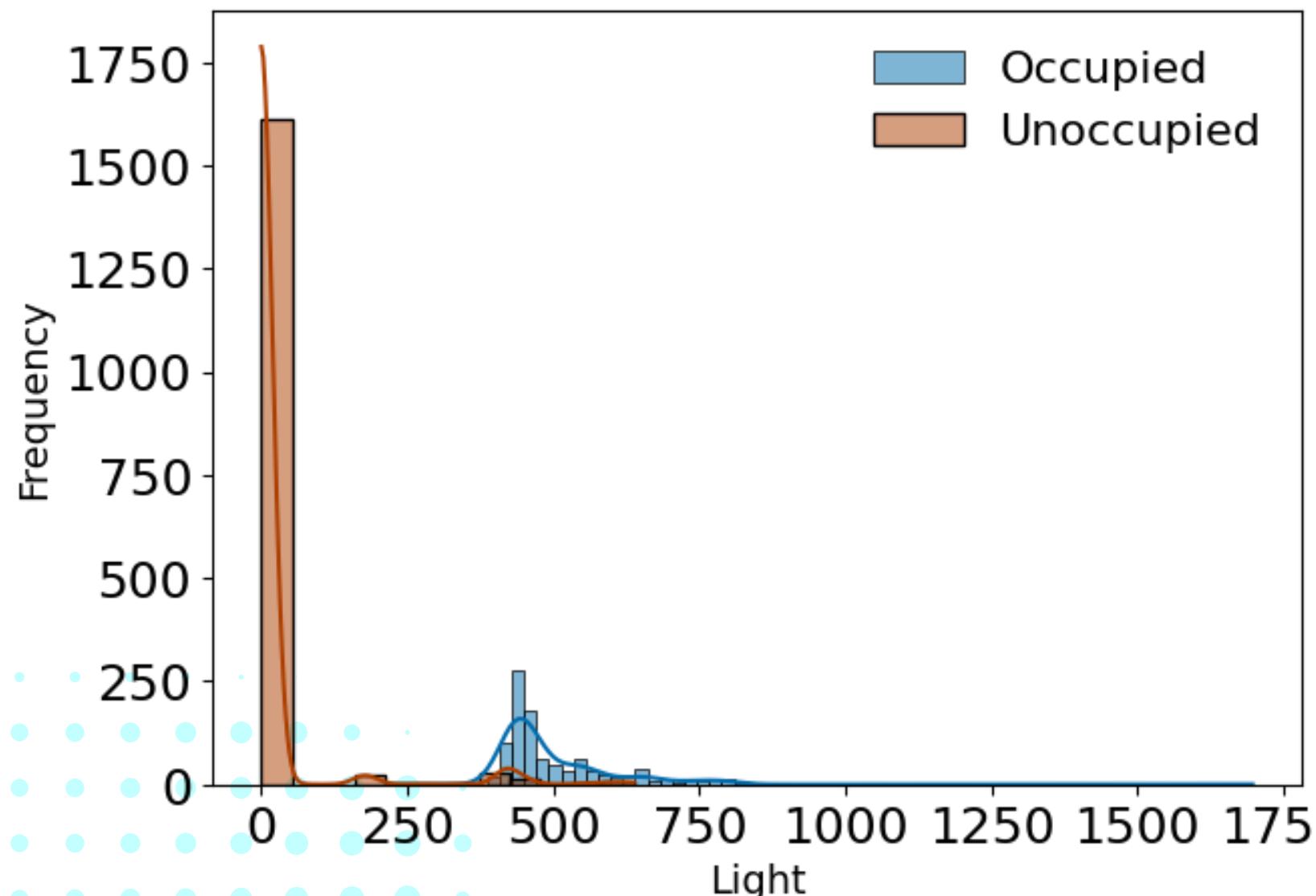
Light

- ข้อมูลในชุด Data Training มีลักษณะเป้าไปทางขวาในระดับสูง (Right-skewed)
- ค่าความสว่างที่ 0 Lux จะเป็นช่วงที่ไม่มีคนอยู่
- ค่าความสว่างในช่วง 433 - 477.75 Lux จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



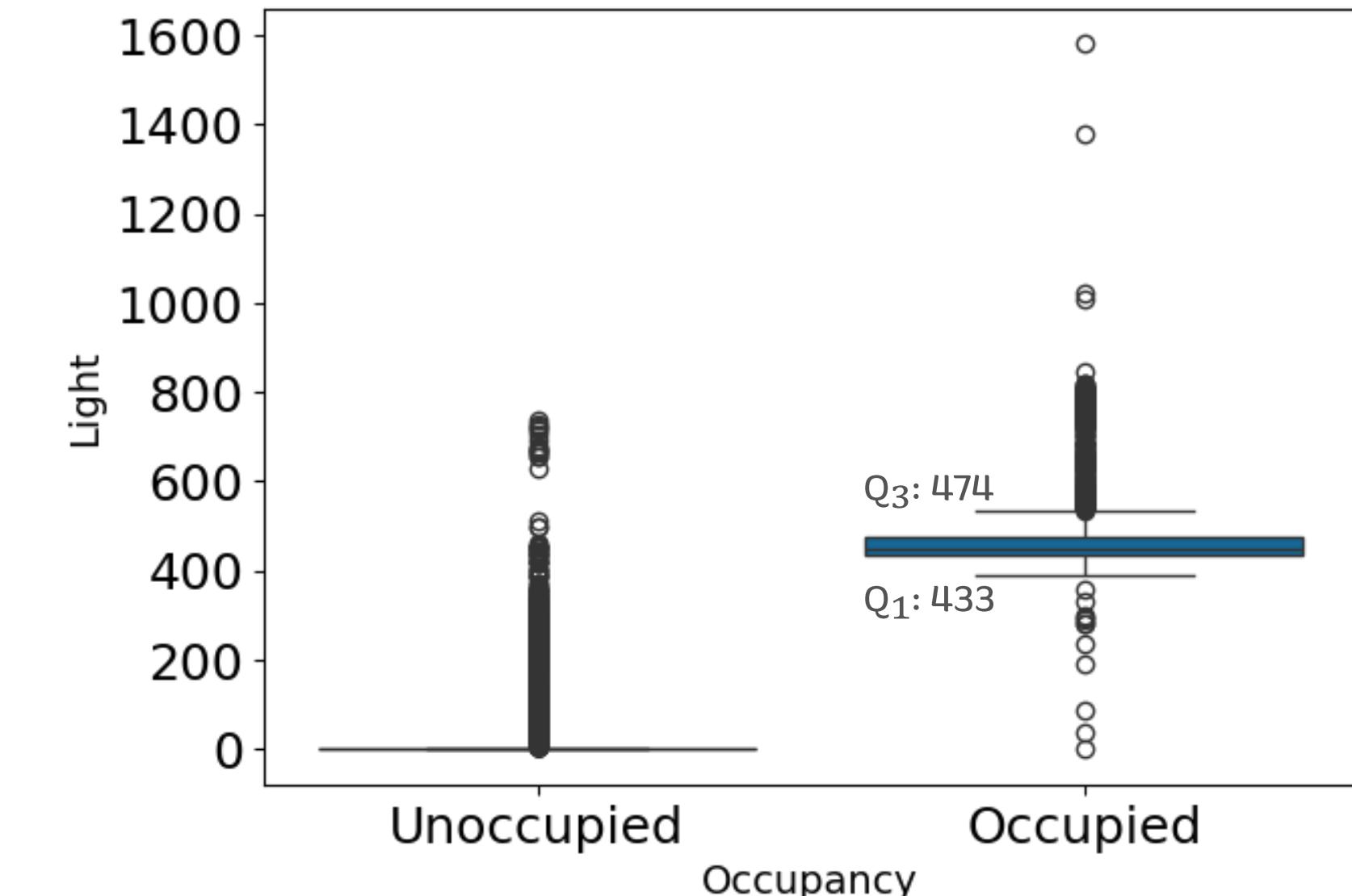
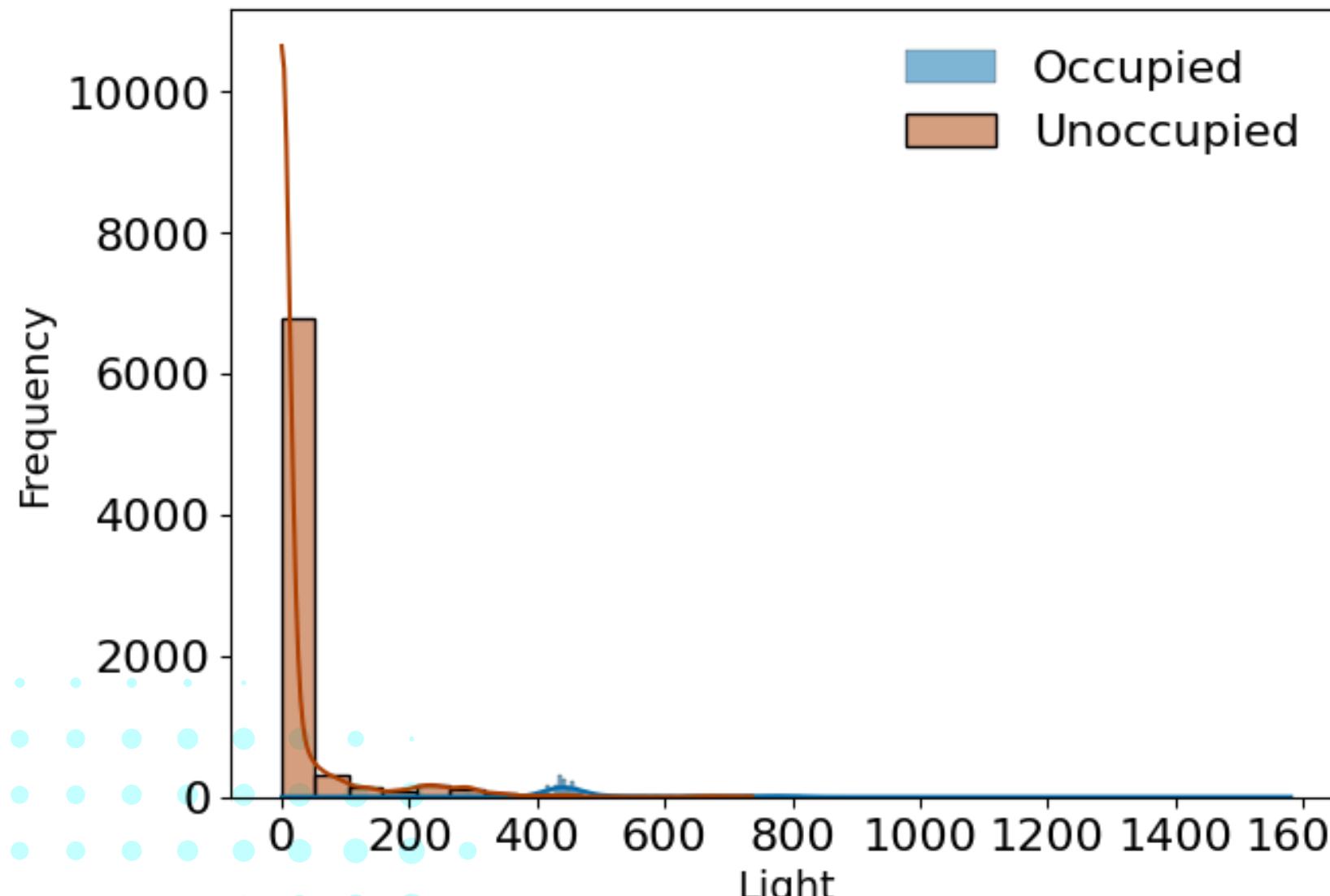
Light

- Data Test 1 มีลักษณะเบ้าไปทางขวาในระดับปานกลาง (Right-skewed)
- ค่าความสว่างที่ 0 Lux จะเป็นช่วงที่ไม่มีคนอยู่
- ค่าความสว่างในช่วง 433 - 538 Lux จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



Light

- ข้อมูลในชุด Data Test 2 มีลักษณะเบ้าไปทางขวาในระดับสูง (Right-skewed)
- ค่าความสว่างที่ 0 Lux จะเป็นช่วงที่ไม่มีคนอยู่
- ค่าความสว่างในช่วง 433 - 474 Lux จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



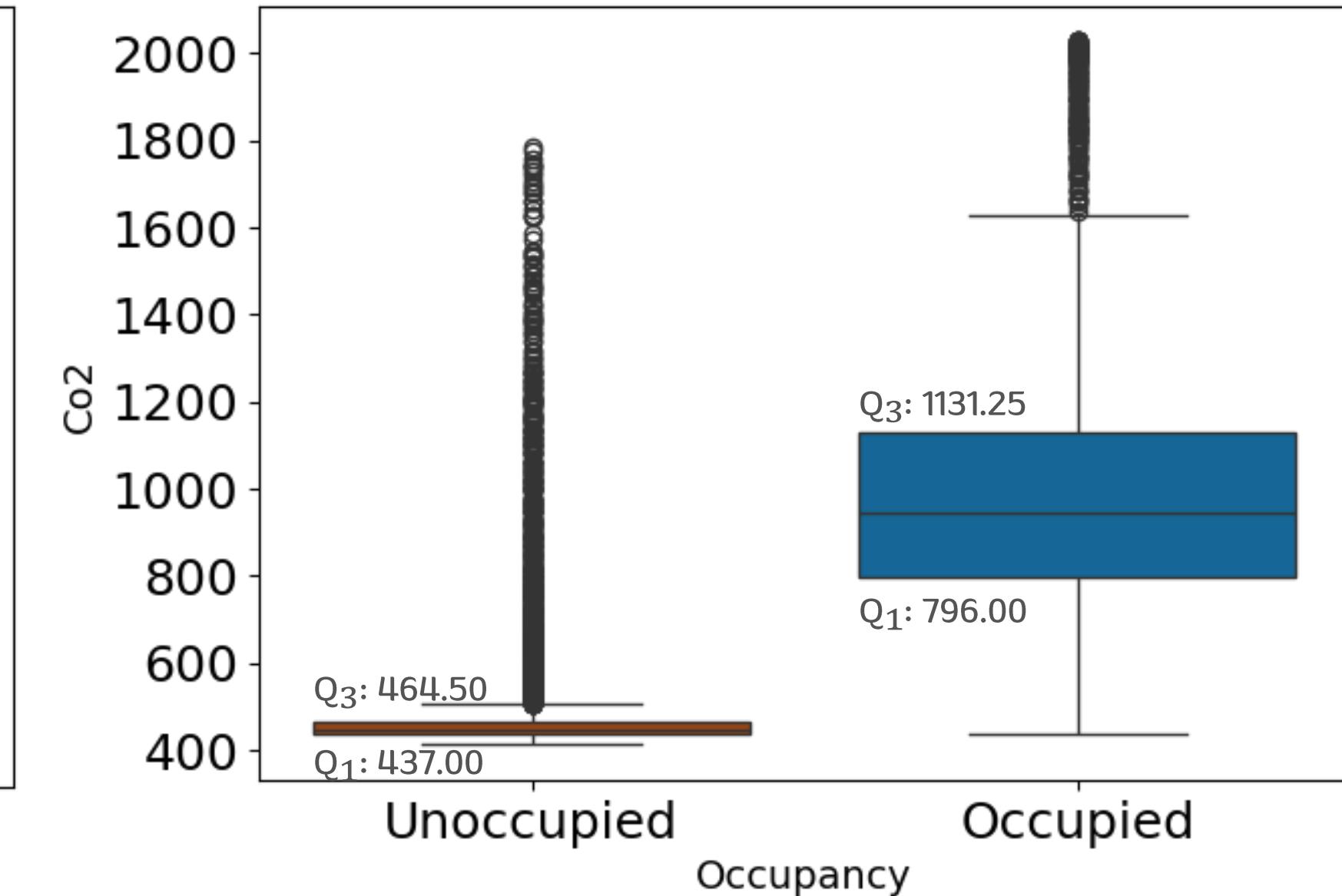
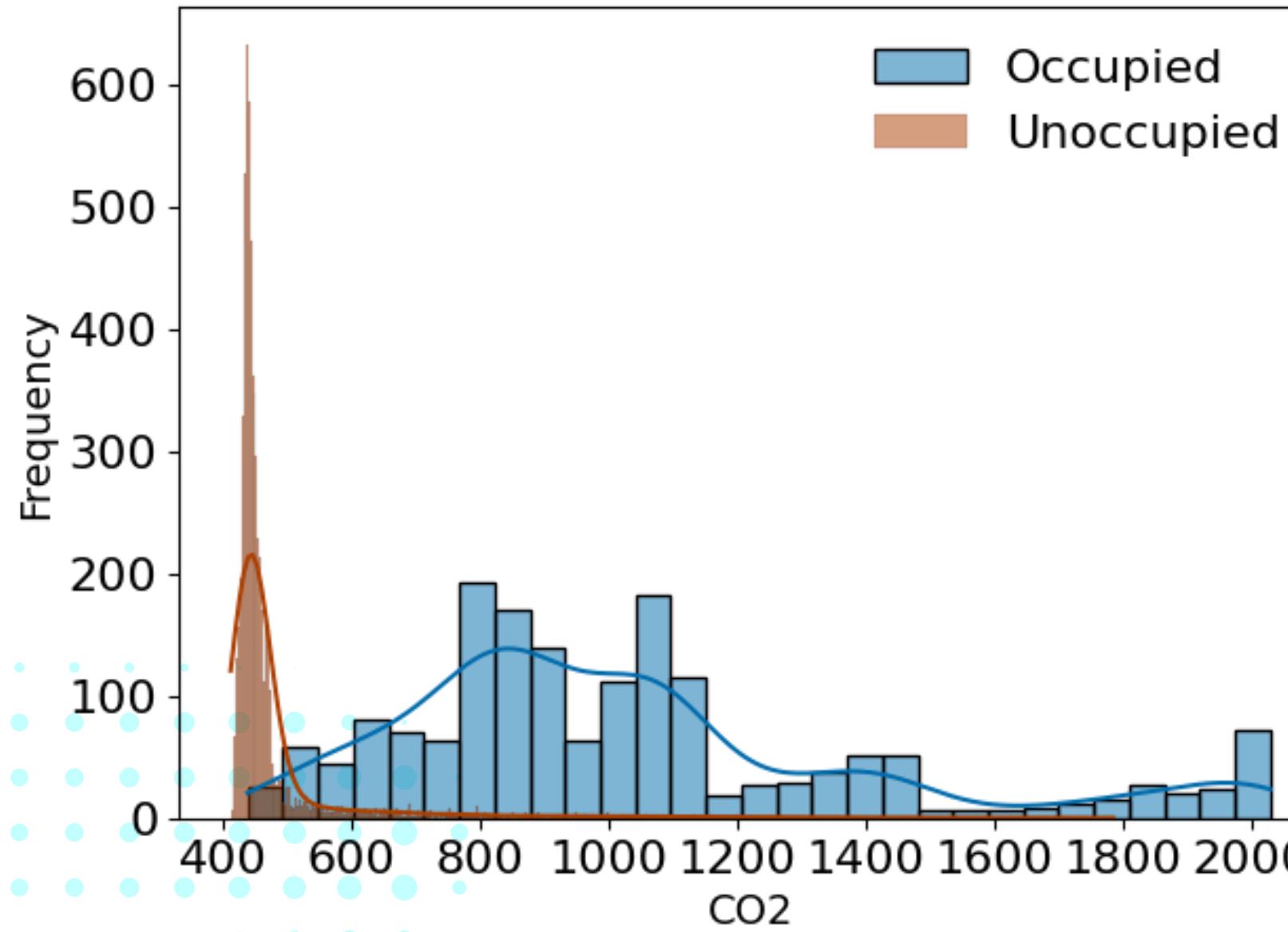
CO₂

- เป็น Feature ประเภทเชิงปริมาณ (Numerical) ชนิดอัตราส่วน (Ratio)
- กำหนดให้ค่าระดับก๊าซคาร์บอนไดออกไซด์ภายในห้อง ใหม่นิวย ppm

	Training	Test 1	Test 2		Training	Test 1	Test 2
Mean	606.55	717.91	753.22	Q₁	439.00	466.00	542.31
Median	453.50	580.50	639.00	Q₂	453.50	580.50	639.00
S.D.	314.32	292.68	297.10	Q₃	638.83	956.33	831.13
Min	412.75	427.50	484.67	Max	2028.50	1,402.25	2,076.50

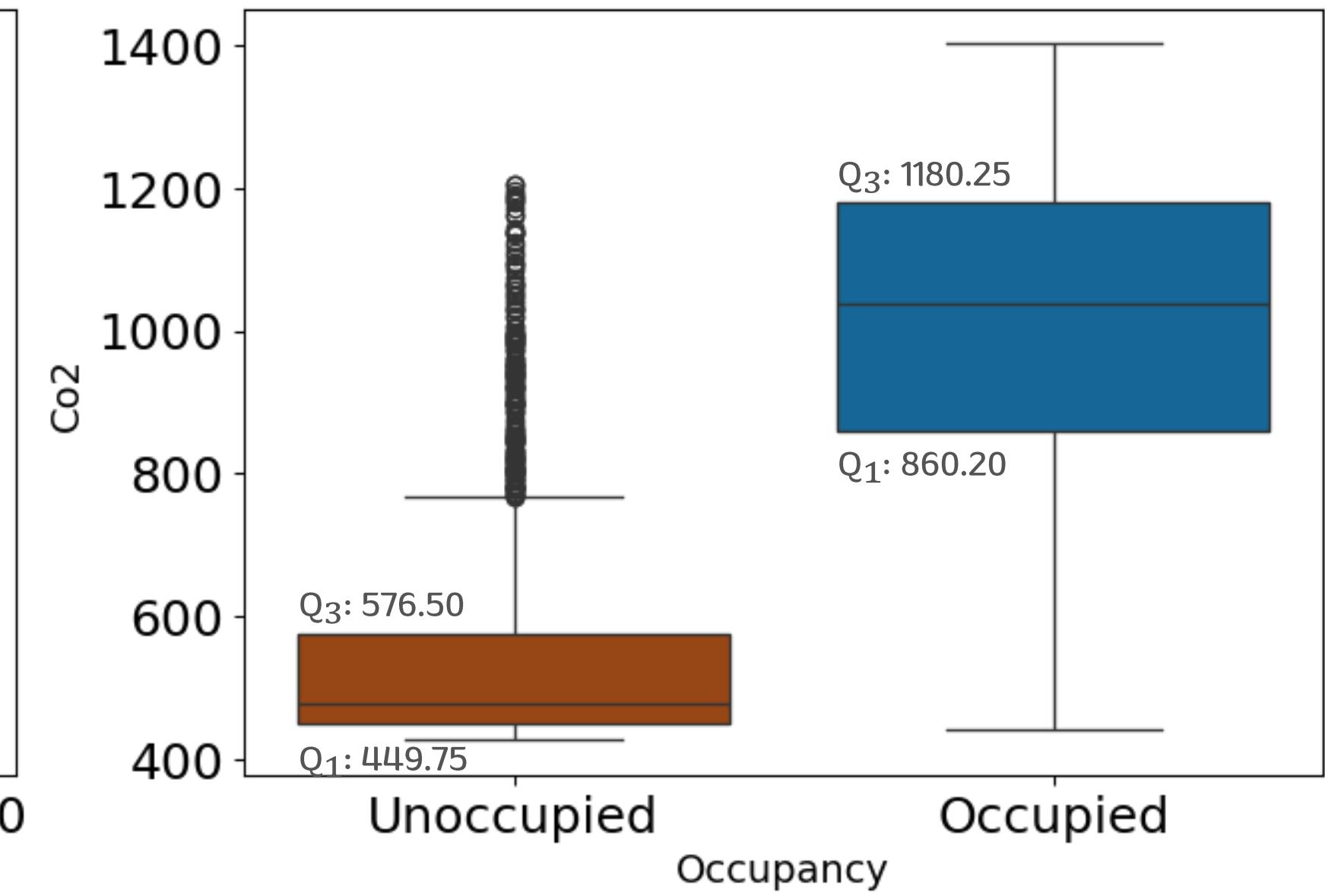
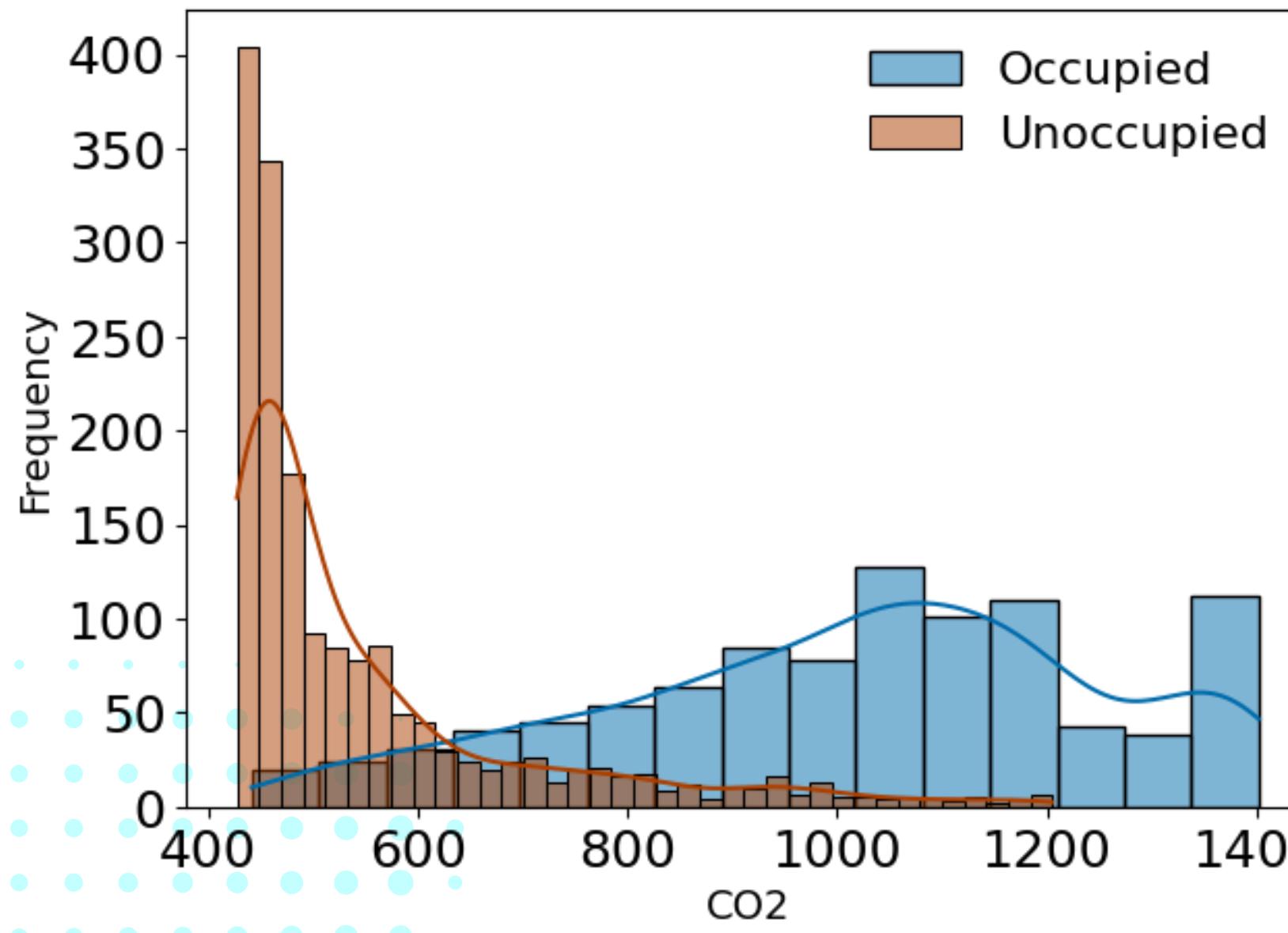
CO₂

- ข้อมูลในชุด Data Training มีลักษณะเป้าไปทางขวาในระดับสูง (Right-skewed)
- ระดับกําช CO₂ ในช่วง 437.00 - 464.50 ppm จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- ระดับกําช CO₂ ในช่วง 796.00 - 1,131.25 ppm จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



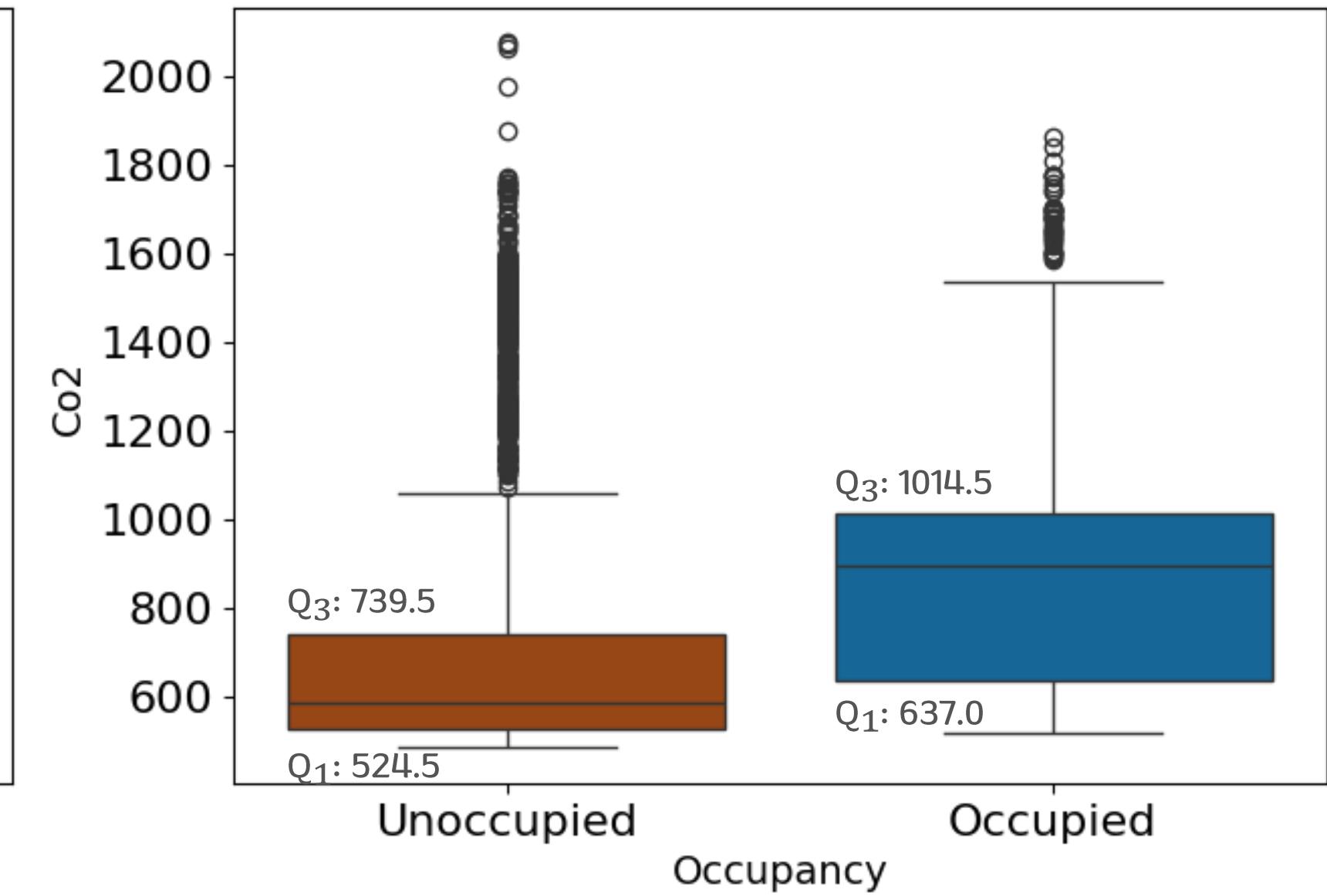
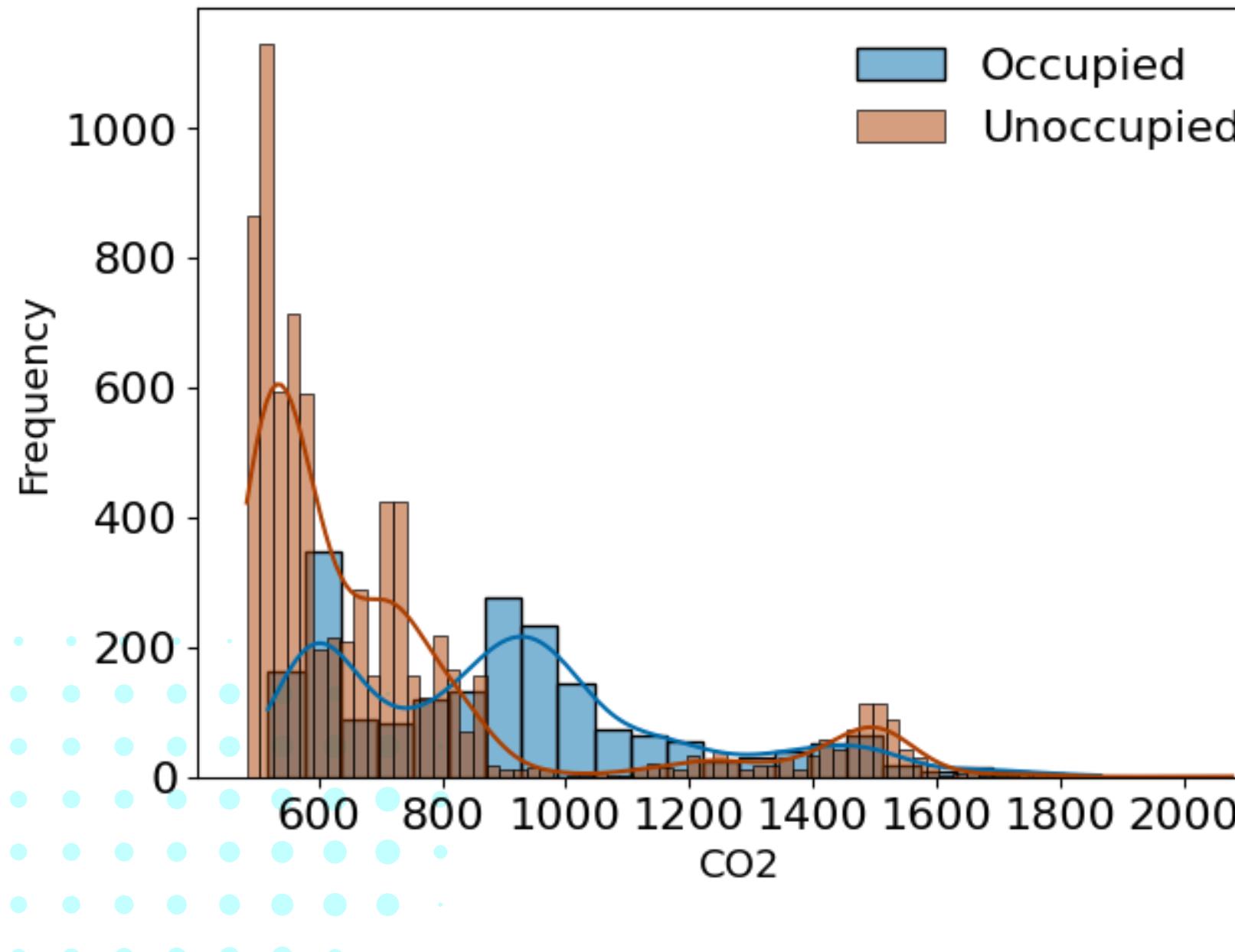
CO₂

- ข้อมูลในชุด Data Test 1 มีลักษณะเบ้าไปทางขวาในระดับปานกลาง (Right-skewed)
- ระดับกําช CO₂ ในช่วง 449.75 - 576.50 ppm จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- ระดับกําช CO₂ ในช่วง 860.20 - 1,180.25 ppm จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



CO₂

- ข้อมูลในชุด Data Test 2 มีลักษณะเป้าไปทางขวาในระดับสูง (Right-skewed)
- ระดับกําช CO₂ ในช่วง 524.5 - 739.5 ppm จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- ระดับกําช CO₂ ในช่วง 637.0 - 1,014.5 ppm จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



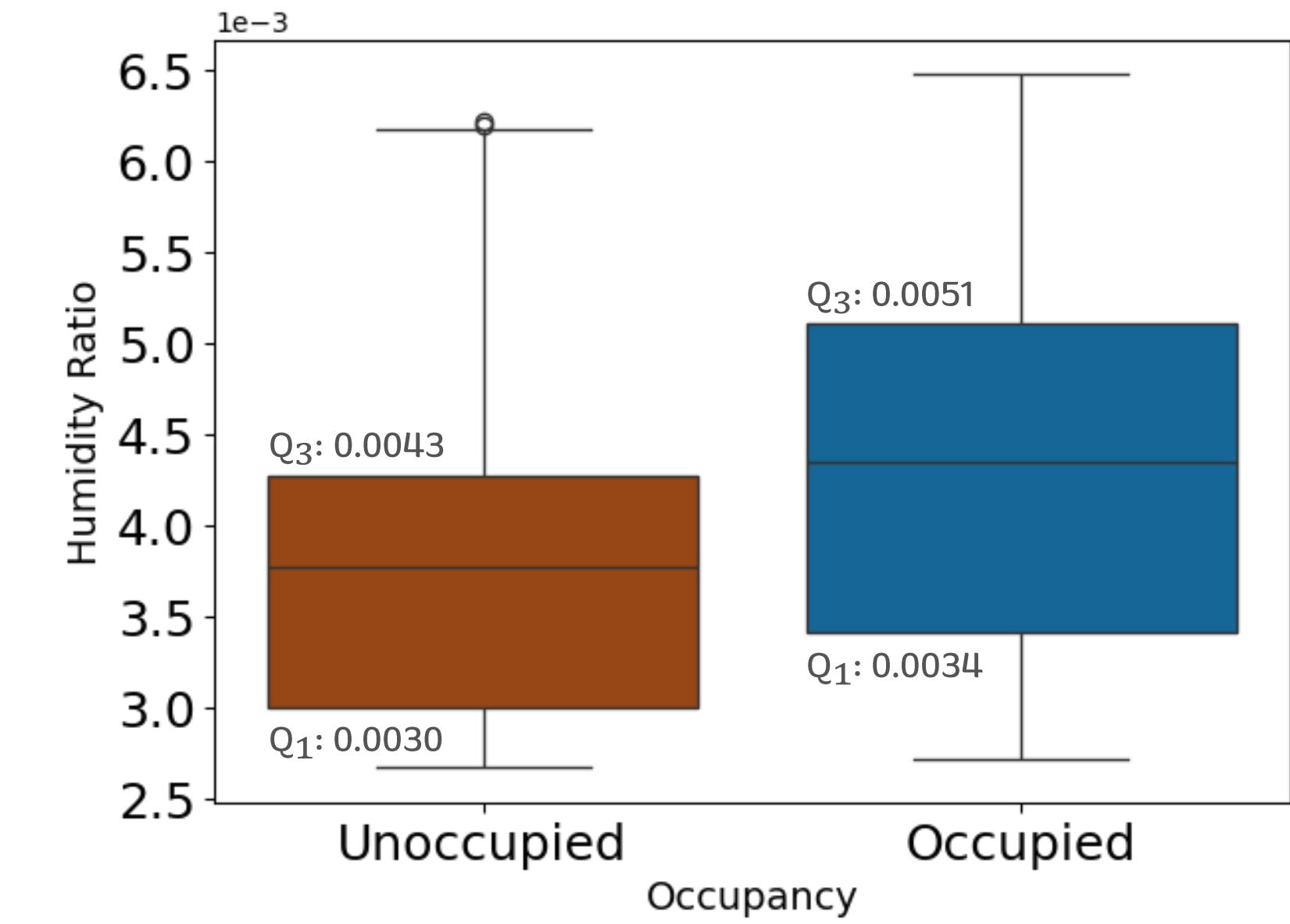
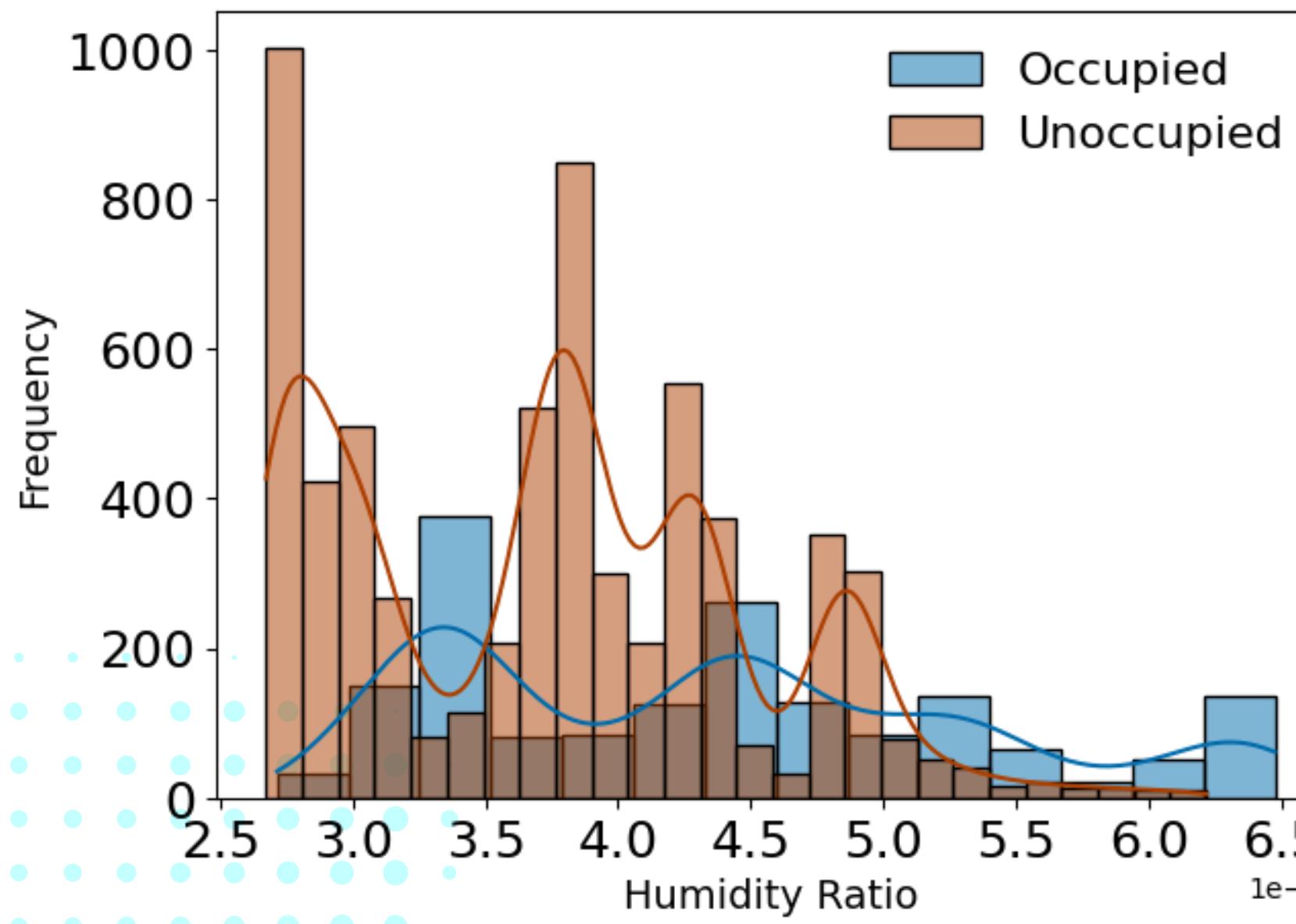
Humidity Ratio

- เป็น Feature ประเภทเชิงปริมาณ (Numerical) ชนิดอัตราส่วน (Ratio)
- กำหนดให้ค่าสัดส่วนของไอน้ำต่ออากาศแห้ง ในหน่วย kg-water-vapor/kg-air

	Training	Test 1	Test 2		Training	Test 1	Test 2
Mean	0.0039	0.004	0.0046	Q₁	0.0031	0.0035	0.0042
Median	0.0038	0.0038	0.0046	Q₂	0.0038	0.0038	0.0046
S.D.	0.0009	0.0006	0.0005	Q₃	0.0044	0.0045	0.005
Min	0.0028	0.0033	0.0033	Max	0.0065	0.0054	0.0058

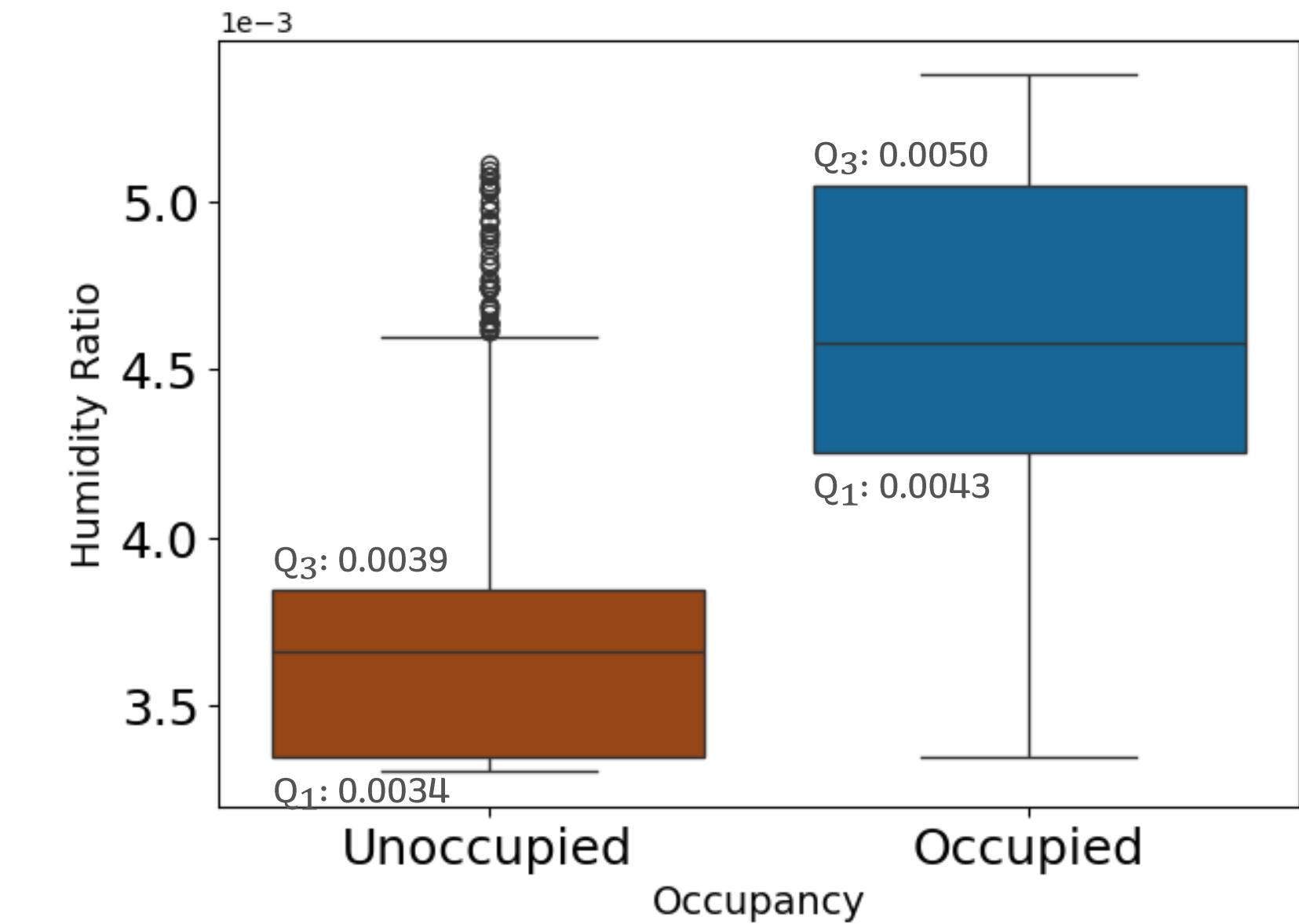
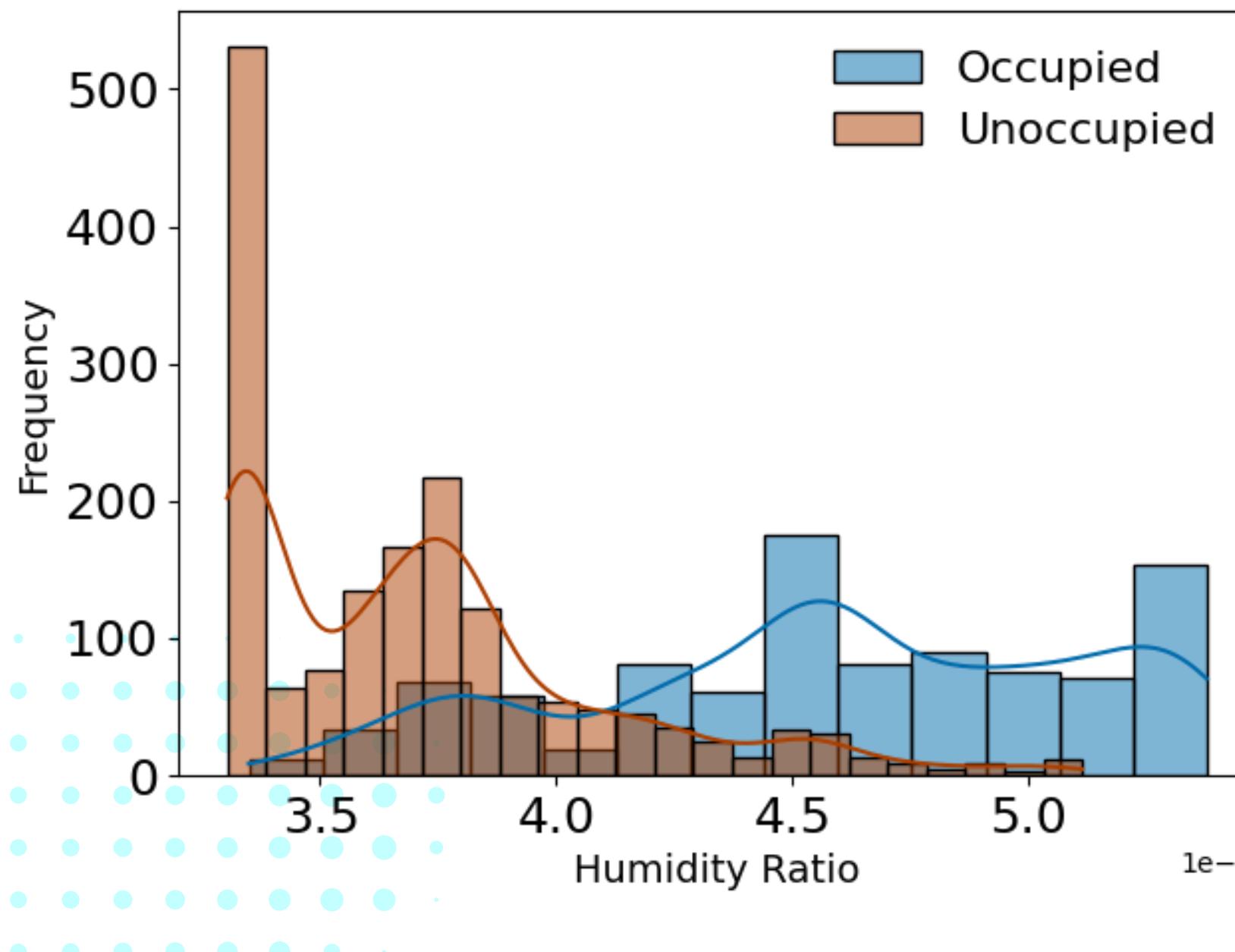
Humidity Ratio

- ข้อมูลในชุด Data Training มีลักษณะเป้าไปทางขวาในระดับปานกลาง (Right-skewed)
- Humidity Ratio ในช่วง 0.0030 - 0.0043 จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- Humidity Ratio ในช่วง 0.0034 - 0.0051 จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



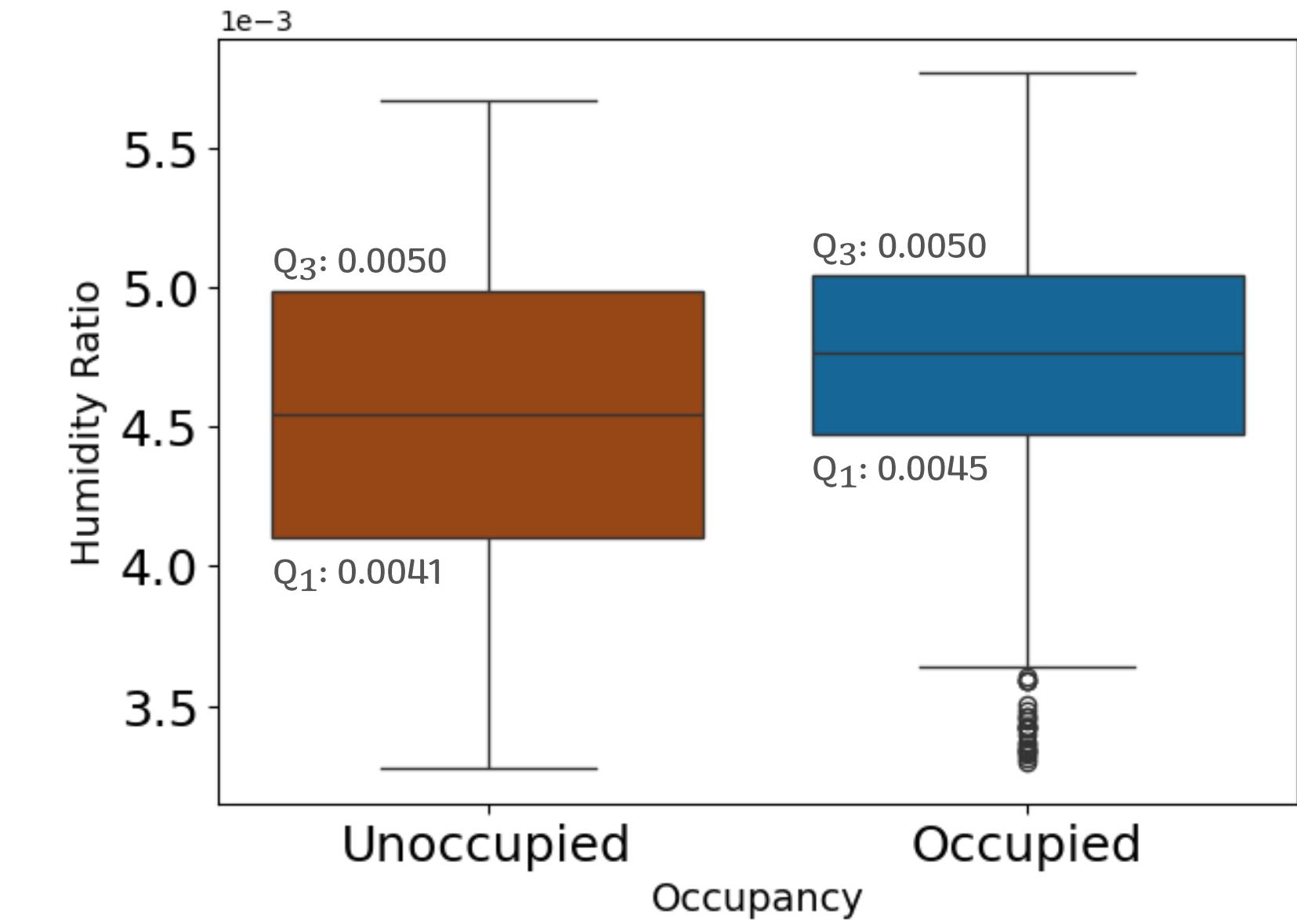
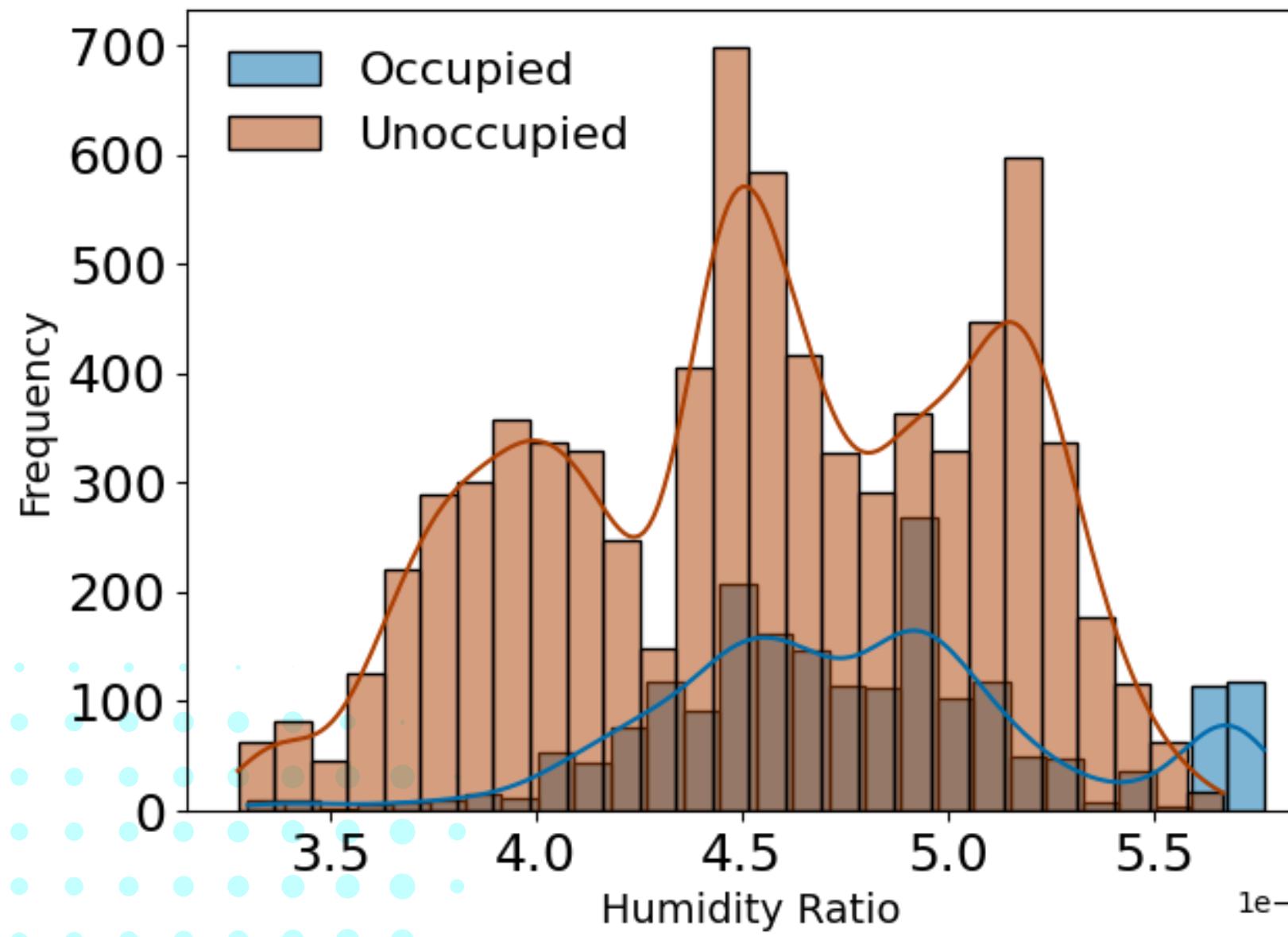
Humidity Ratio

- ข้อมูลในชุด Data Test 1 มีลักษณะเบ้าไปทางขวาในระดับปานกลาง (Right-skewed)
- Humidity Ratio ในช่วง 0.0034 - 0.0039 จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- Humidity Ratio ในช่วง 0.0043 - 0.0050 จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่

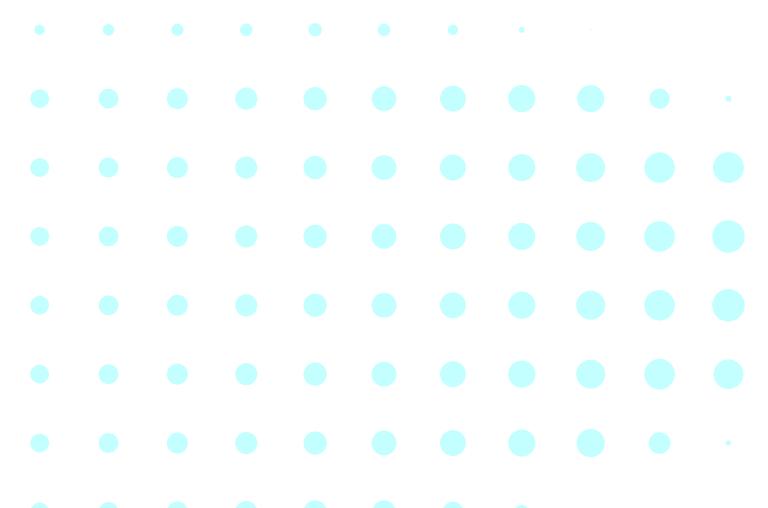


Humidity Ratio

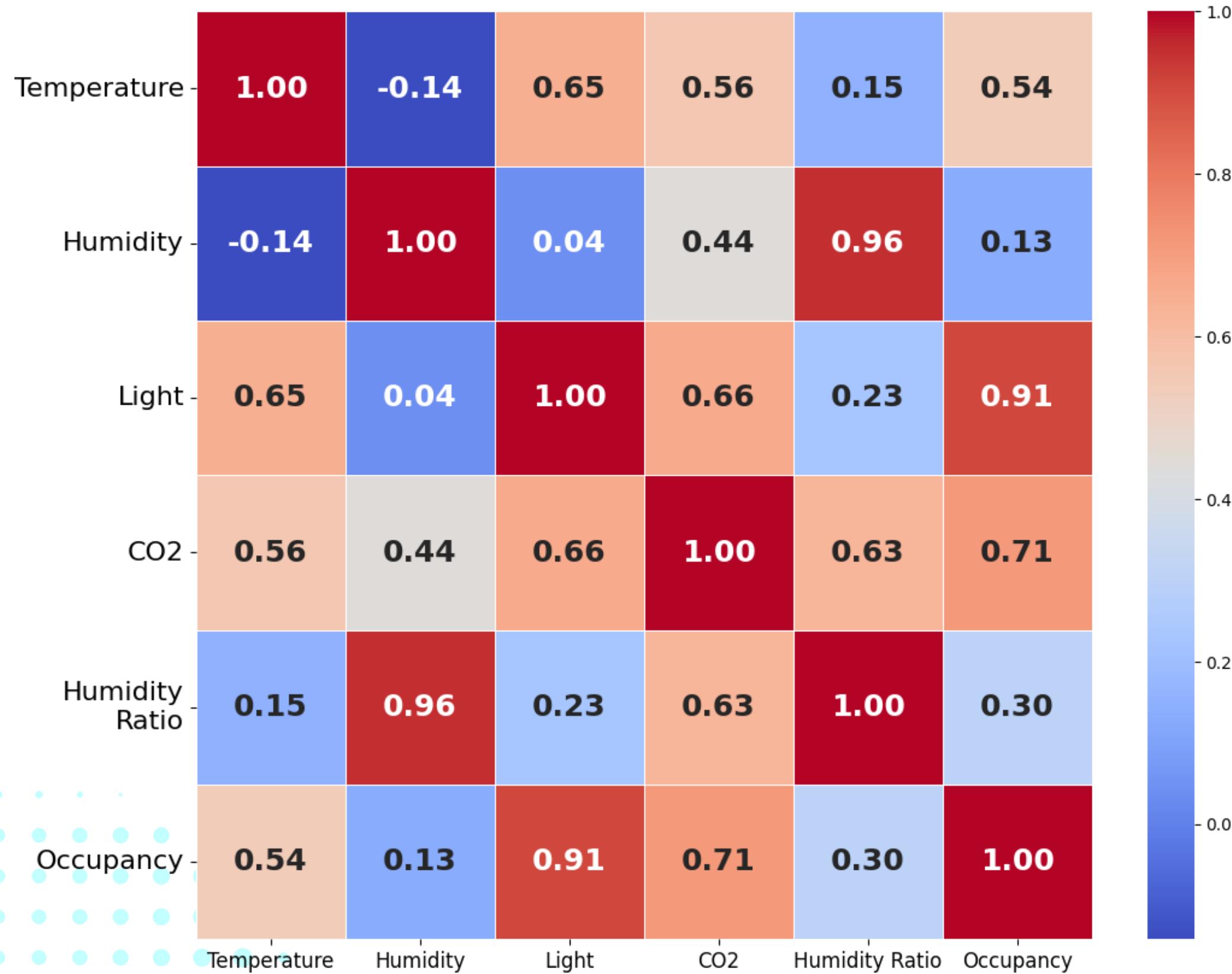
- ข้อมูลในชุด Data Test 2 มีลักษณะเป้าไปทางซ้ายเล็กน้อย (Left-skewed)
- Humidity Ratio ในช่วง 0.0041 - 0.0050 จะเป็นช่วงที่ส่วนใหญ่จะไม่มีคนอยู่
- Humidity Ratio ในช่วง 0.0045 - 0.0050 จะเป็นช่วงที่ส่วนใหญ่จะมีคนอยู่



ความสัมพันธ์ของแต่ละ Feature

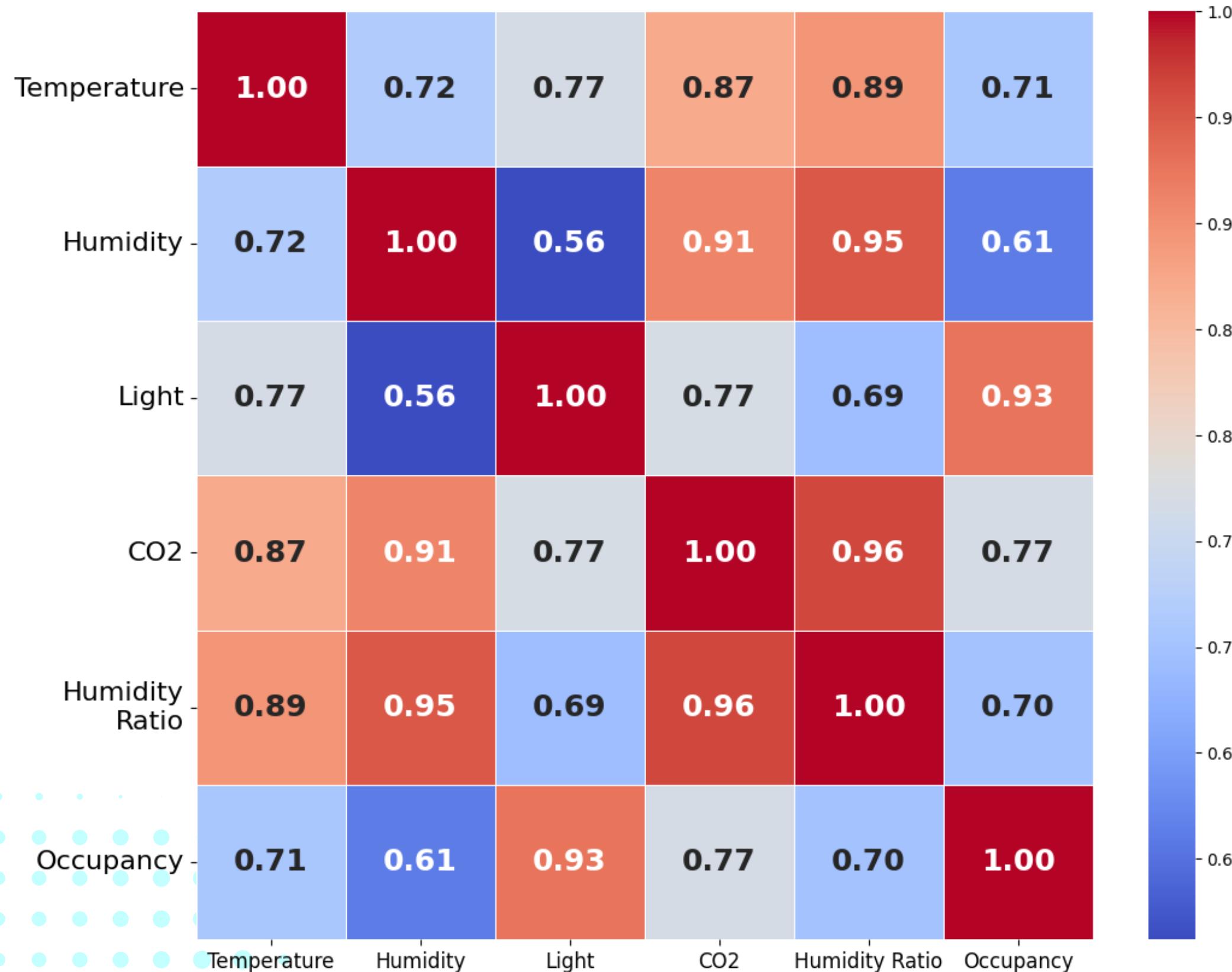


Data Training



- มีความสัมพันธ์ในเชิงบวก
(0.6, 1)
- มีความสัมพันธ์ต่ำหรือไม่มี
(0.2, 0.6)
- มีความสัมพันธ์ในเชิงลบ
(-0.2, 0.2)

Data Test 1



- มีความสัมพันธ์ในเชิงบวก
(0.85, 1)
- มีความสัมพันธ์ต่ำหรือไม่มี
(0.70, 0.85)
- มีความสัมพันธ์ในเชิงลบ
(0.55, 0.70)

Data Test 2



- มีความสัมพันธ์ในเชิงบวก
(0.467, 1)
- มีความสัมพันธ์ต่ำหรือไม่มี
(-0.067, 0.467)
- มีความสัมพันธ์ในเชิงลบ
(-0.6, -0.067)

Data Preprocess in Data Training

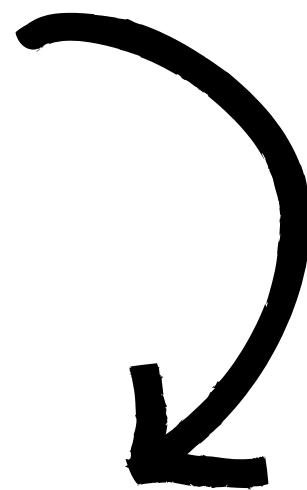


Drop Feature

- Drop Feature ที่ไม่มีผลต่อการคำนายของโมเดล ซึ่งได้แก่ ID และ Date
- Drop Feature ที่มีค่า Correlation ระหว่างกันที่มากกว่า 0.9 (ยกเว้น Target Feature)
เพื่อลดโอกาสที่โมเดลจะเรียนรู้ข้อมูลซ้ำซ้อนและทำให้เกิดการ Overfitting
 - Feature ที่เข้าเงื่อนไขในการ Drop จะคือ Feature Humidity กับ Feature Humidity Ratio
 - โดยจะเลือก Drop Feature Humidity เนื่องจากมีค่า Correlation กับ Occupancy น้อยกว่า Humidity Ratio

Drop Feature

	id	date	Temperature	Humidity	Light	CO2	Humidity Ratio	Occupancy
0	1	2015-02-04 17:51:00	23.18	27.2720	426.0	721.25	0.004793	1
1	2	2015-02-04 17:51:59	23.15	27.2675	429.5	714.00	0.004783	1
2	3	2015-02-04 17:53:00	23.15	27.2450	426.0	713.50	0.004779	1
3	4	2015-02-04 17:54:00	23.15	27.2000	426.0	708.25	0.004772	1
4	5	2015-02-04 17:55:00	23.10	27.2000	426.0	704.50	0.004757	1



ชุดข้อมูลก่อนการ Drop



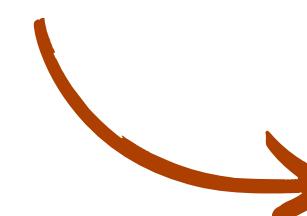
ชุดข้อมูลหลังการ Drop

	Temperature	Light	CO2	Humidity Ratio	Occupancy
0	23.18	426.0	721.25	0.004793	1
1	23.15	429.5	714.00	0.004783	1
2	23.15	426.0	713.50	0.004779	1
3	23.15	426.0	708.25	0.004772	1
4	23.10	426.0	704.50	0.004757	1

Feature Scaling

- ทำให้ข้อมูลในแต่ละ Feature มีความใกล้เคียงกัน โดยทำให้ข้อมูลอยู่ในช่วง 0 - 1 โดยใช้ Min-Max Scaling

ชุดข้อมูลก่อนการทำ Min-Max Scaling



	Temperature	Light	CO2	Humidity Ratio	Occupancy
0	23.18	426.0	721.25	0.004793	1
1	23.15	429.5	714.00	0.004783	1
2	23.15	426.0	713.50	0.004779	1
3	23.15	426.0	708.25	0.004772	1
4	23.10	426.0	704.50	0.004757	1

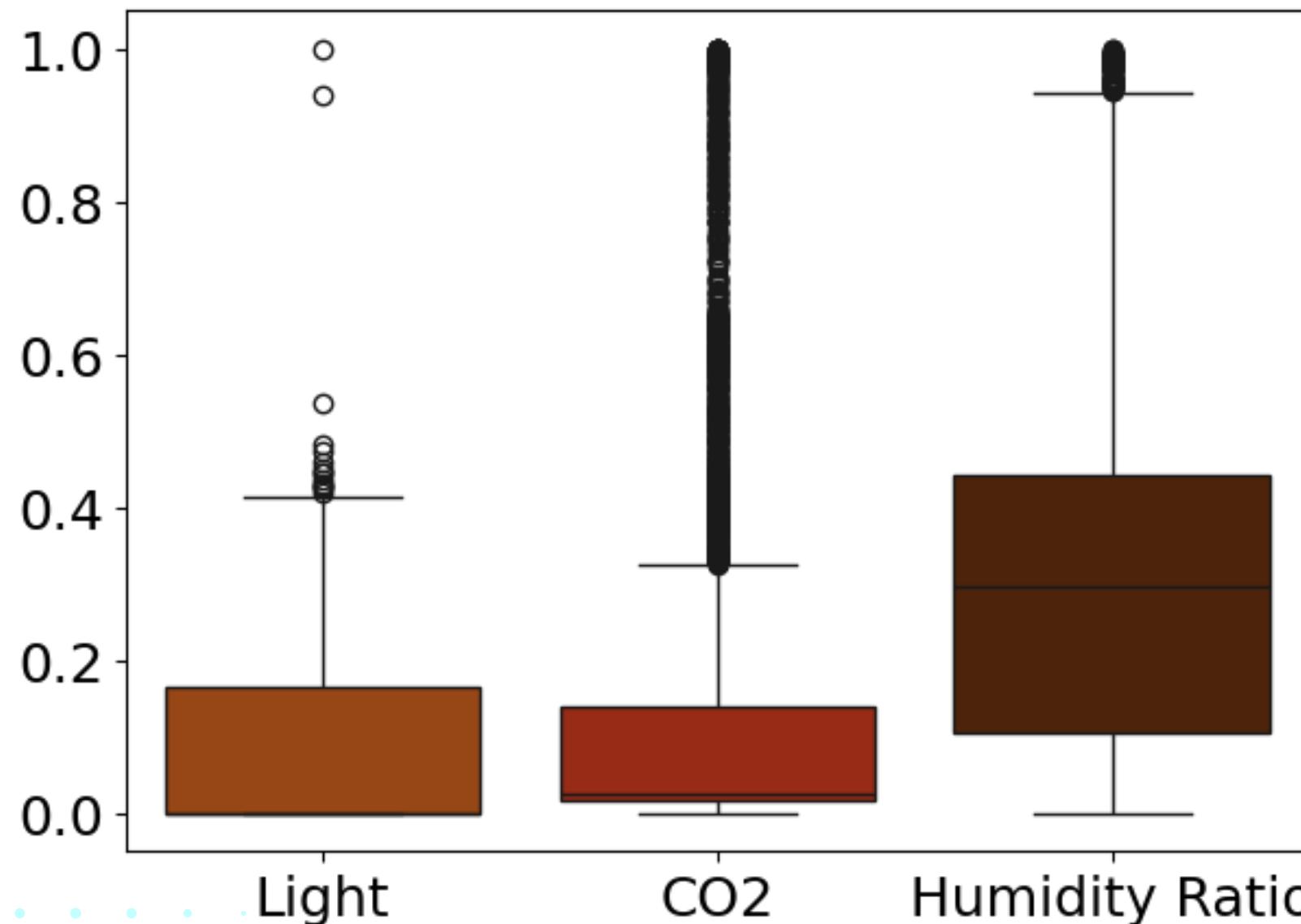
	Temperature	Light	CO2	Humidity Ratio	Occupancy
0	1.000000	0.275490	0.190933	0.557318	1
1	0.992823	0.277754	0.186446	0.554807	1
2	0.992823	0.275490	0.186136	0.553761	1
3	0.992823	0.275490	0.182887	0.551669	1
4	0.980861	0.275490	0.180566	0.547851	1

ชุดข้อมูลหลังการทำ Min-Max Scaling



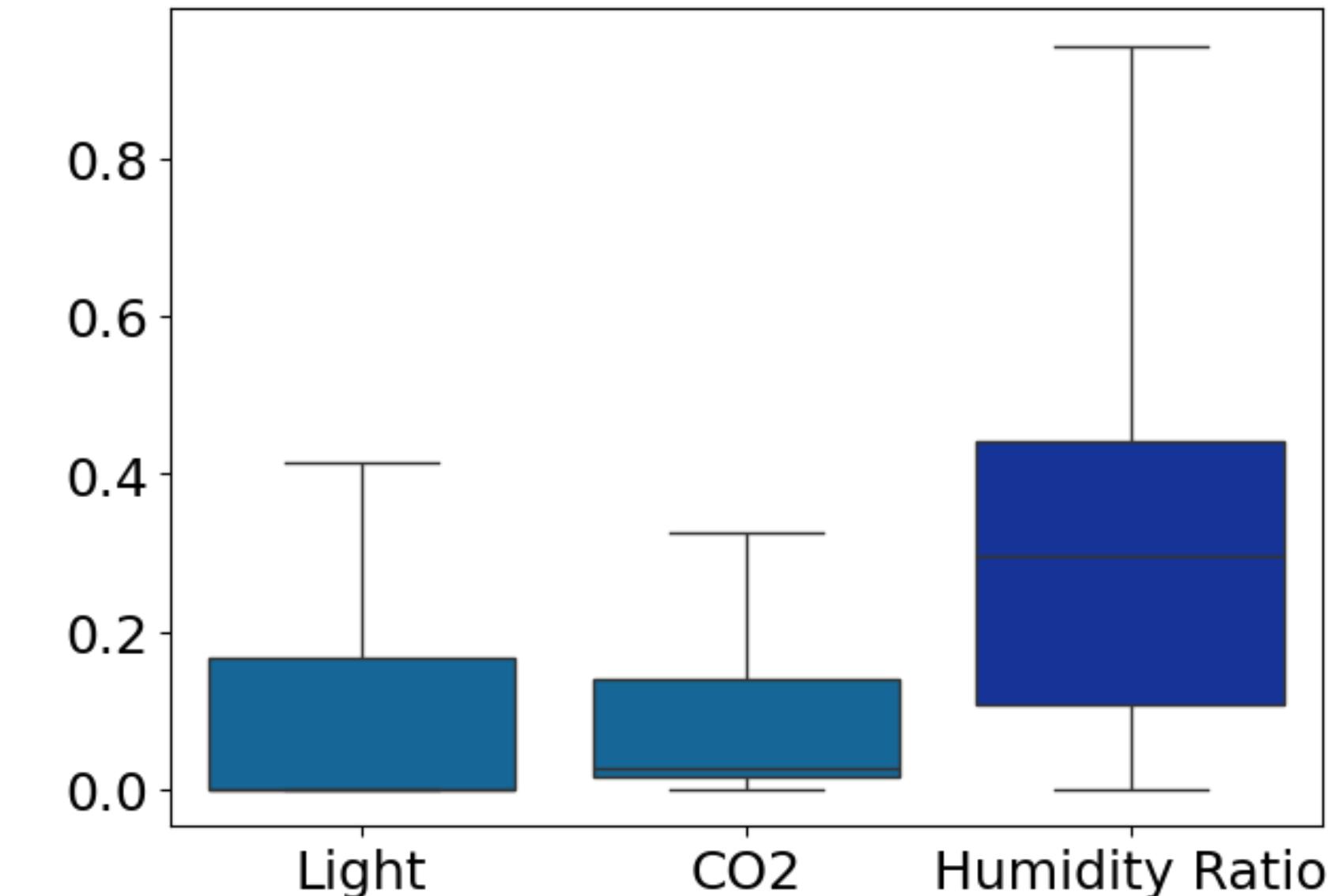
Handling Outliers

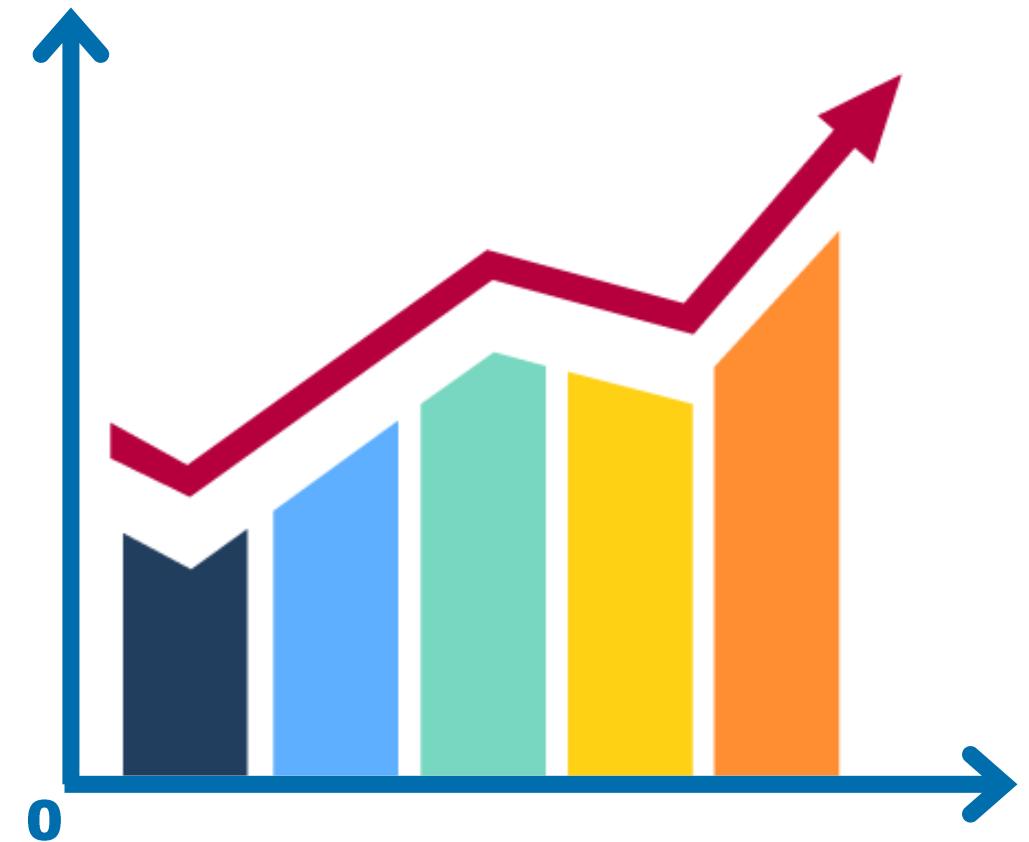
- กำจัด Outliers ใน Feature : Light, CO2 และ Humidity Ratio



boxplot ก่อนการกำจัด Outliers

boxplot หลังการกำจัด Outliers





THANK YOU

นายศุภนัช แซ่เตี้ย

ID: 6505000270

อ้างอิง

- Luis Candanedo. (2559). **Occupancy Detection**. สืบค้นเมื่อ 9 กุมภาพันธ์ 2568, จาก <https://archive.ics.uci.edu/dataset/357/occupancy+detection>
- L. Candanedo, V. Feldheim. (2558). **Accurate occupancy detection of an office room from light, temperature, humidity and CO₂ measurements using statistical learning models**. สืบค้นเมื่อ 9 กุมภาพันธ์ 2568, จาก <https://www.sciencedirect.com/science/article/pii/S0378778815304357>
- google scholar. (2568). **Luis Miguel Candanedo Ibarra**. สืบค้นเมื่อ 10 กุมภาพันธ์ 2568, จาก https://scholar.google.be/citations?user=SHQAn_8AAAAJ
- OpenAI. (2567). **ChatGPT (Version 4) [Large language model]**. สืบค้นเมื่อ 28 มกราคม 2568 , จาก <https://chatgpt.com/>