รายงานปัญหาที่พบ

ชุดข้อมูลเซนเซอร์ IoT: iot_sensor_raw_data_extended.csv (480 แถว)

1. ภาพรวมปัญหาที่พบ

จากการตรวจสอบข้อมูลดิบจากเซนเซอร์ IoT จำนวน 480 แถว ก่อนทำความสะอาด พบปัญหาคุณภาพข้อมูล 4 กลุ่มหลัก ได้แก่ Duplicate, Outlier, Missing Value และ Error (ค่าที่ไม่ถูกต้องตามรูปแบบที่กำหนด) โดยสรุปจำนวนได้ดังตาราง (นับตามเงื่อนไขแต่ละประเภท แถวเดียวกันอาจเข้าเงื่อนไขมากกว่าหนึ่งประเภทพร้อมกันได้):

ประเภทปัญหา	กฎที่เกี่ยวข้อง	จำนวนที่พบ (แถว)
Duplicate rows (device_id + timestamp ซ้ำ)	R3	50 แถว (25 คู่)
Temp outlier (นอกช่วง 0–50 °C)	R1	10
Battery outlier (นอกช่วง 0–100)	R4	5
Missing device_id	R2	2
Missing temp_c	-	7
Missing battery	-	4
Motion ค่าไม่ถูกต้อง (error)	R5	5
Motion ว่าง (missing)	-	4

2. รายละเอียดปัญหาแต่ละประเภท พร้อมตัวอย่างข้อมูล

2.1 Duplicate Rows — ข้อมูลซ้ำ (ผิดกฎ R3)

พบทั้งหมด 50 แถว (คิดเป็น 25 คู่ซ้ำ) ที่มี device_id และ timestamp เดียวกัน ซึ่งขัดกับกฎที่ว่าห้ามมีข้อมูลซ้ำที่ device_id และ timestamp ตรงกัน สาเหตุที่เป็นไปได้คือการส่งข้อมูลซ้ำจากเซนเซอร์ (retry) หรือความผิดพลาดในการบันทึกข้อมูล ตัวอย่าง:

timestamp	device_id	temp_c	motion	battery
2026-07-12 14:53:00	SN-9982	29.4	true	88
2026-07-12 14:53:00	SN-9982	29.4	true	88
2026-07-12 14:53:40	SN-9982	29.4	false	88
2026-07-12 14:53:40	SN-9982	29.4	false	88

2.2 Temp Outlier — อุณหภูมินอกช่วงที่กำหนด (ผิดกฎ R1)

พบ 10 แถว ที่ temp_c อยู่นอกช่วง 0–50 องศาเซลเซียส ส่วนใหญ่เป็นค่า -999 (ค่าที่มักใช้แทนการอ่านค่าไม่สำเร็จของเซนเซอร์ / sentinel value) และค่า 85.0 องศา ซึ่งเกินขอบเขตทางกายภาพที่เป็นไปได้ของอุณหภูมิห้อง/พื้นที่ใช้งานทั่วไป ตัวอย่าง:

timestamp	device_id	temp_c	motion	battery
2026-07-12 14:52:40	SN-9983	-999.0	true	94
2026-07-12 14:54:00	SN-9984	-999.0	true	75
2026-07-12 14:55:40	SN-9984	85.0	true	75

2.3 Battery Outlier — ค่าแบตเตอรี่นอกช่วงที่กำหนด (ผิดกฎ R4)

พบ 5 แถว ที่ battery อยู่นอกช่วง 0–100% เช่น ค่าติดลบ (-5, -1) และค่าที่เกิน 100% (999, 115, 108)
ซึ่งไม่สอดคล้องกับความหมายทางกายภาพของเปอร์เซ็นต์แบตเตอรี่ ตัวอย่าง:

timestamp	device_id	temp_c	motion	battery
2026-07-12 14:51:20	SN-9983	28.3	true	-5
2026-07-12 14:57:30	SN-9983	27.2	false	999
2026-07-12 14:58:00	SN-9984	28.1	false	115

2.4 Missing device_id — ไม่มีรหัสอุปกรณ์ (ผิดกฎ R2)

พบ 2 แถว ที่ device_id เป็นค่าว่าง ทำให้ไม่สามารถระบุได้ว่าแถวข้อมูลนั้นมาจากเซนเซอร์ตัวใด
ถือเป็นปัญหาที่ร้ายแรงที่สุดเพราะไม่สามารถเติมคืนค่าที่ถูกต้องได้อย่างน่าเชื่อถือ ตัวอย่าง:

timestamp	device_id	temp_c	motion	battery
2026-07-12 14:59:40	(ว่าง)	26.4	true	79
2026-07-12 15:02:10	(ว่าง)	28.4	true	85

2.5 Missing temp_c — ค่าอุณหภูมิว่าง

พบ 7 แถว ที่ไม่มีค่า temp_c บันทึกไว้ (อาจเกิดจากเซนเซอร์อ่านค่าไม่สำเร็จหรือสัญญาณขาดหาย) ตัวอย่าง:

timestamp	device_id	temp_c	motion	battery
2026-07-12 14:54:00	SN-9986	(ว่าง)	true	68
2026-07-12 14:58:50	SN-9984	(ว่าง)	false	74

2.6 Missing battery — ค่าแบตเตอรี่ว่าง

พบ 4 แถว ที่ไม่มีค่า battery บันทึกไว้ ตัวอย่าง:

timestamp	device_id	temp_c	motion	battery
2026-07-12 14:56:40	SN-9983	27.5	false	(ว่าง)
2026-07-12 14:57:30	SN-9986	29.1	false	(ว่าง)

2.7 Motion Error — ค่าการตรวจจับการเคลื่อนไหวไม่ถูกต้อง (ผิดกฎ R5)

พบ 5 แถว ที่ค่า motion ไม่ใช่ true/false ตามที่กำหนด โดยพบค่าที่หลากหลาย เช่น '0', 'yes', 'unknown', 'TRUEE', 'FALSEE' ซึ่งน่าจะเกิดจากความผิดพลาดในการบันทึก (พิมพ์ผิด) หรือรูปแบบข้อมูลจากอุปกรณ์ที่ไม่สอดคล้องกัน ตัวอย่าง:

timestamp	device_id	temp_c	motion	battery
2026-07-12 14:51:50	SN-9982	29.3	0	88
2026-07-12 14:57:50	SN-9982	27.9	yes	87
2026-07-12 14:59:20	SN-9984	28.4	unknown	74
2026-07-12 15:01:00	SN-9984	29.1	TRUEE	74

2.8 Motion Missing — ค่าการตรวจจับการเคลื่อนไหวว่าง

พบ 4 แถว ที่ไม่มีค่า motion บันทึกไว้ ตัวอย่าง:

timestamp	device_id	temp_c	motion	battery
2026-07-12 14:53:50	SN-9982	29.2	(ว่าง)	88
2026-07-12 14:56:40	SN-9985	26.2	(ว่าง)	80

3. สาเหตุที่เป็นไปได้ของปัญหา

- Duplicate: เซนเซอร์หรือระบบส่งข้อมูลซ้ำ (retry mechanism) เมื่อเครือข่ายไม่เสถียร
- Outlier (-999, 999 เป็นต้น): เป็นค่า sentinel/error code ที่อุปกรณ์ส่งมาแทนการอ่านค่าที่ล้มเหลว มากกว่าจะเป็นค่าจริงทางกายภาพ
- Missing values: สัญญาณขาดหาย, เซนเซอร์ตัดการเชื่อมต่อชั่วคราว, หรือข้อผิดพลาดในการส่ง/บันทึกข้อมูล
- Motion error: ความไม่สอดคล้องของรูปแบบข้อมูล (data format) จากอุปกรณ์ต่างรุ่น/เฟิร์มแวร์ หรือข้อผิดพลาดของมนุษย์ในการป้อนข้อมูลทดสอบ

4. แนวทางการแก้ไข (สรุป — รายละเอียดอยู่ในรายงาน Data Cleaning)

- Duplicate ⟶ ลบด้วย drop_duplicates(subset=['timestamp','device_id'], keep='first')
- Temp/Battery outlier ⟶ แปลงเป็น NULL แล้ว interpolate ตามลำดับเวลาในแต่ละ device
- Motion error ⟶ แปลงเป็น NULL ก่อน แล้วเติมด้วยค่าฐานนิยม (mode) ของ device นั้น
- Missing temp_c / battery ⟶ interpolate ตาม device, ถ้ายังว่างเติมด้วยค่ามัธยฐานรวม
- Missing motion ⟶ เติมด้วยค่าฐานนิยม (mode) ของ device นั้น
- Missing device_id ⟶ ไม่เติมค่า เนื่องจากไม่สามารถระบุที่มาได้ ทำเครื่องหมายเป็น invalid_missing_device_id และไม่รวมในชุดข้อมูลสะอาดสุดท้าย

5. ผลลัพธ์หลังแก้ไขปัญหา

หลังดำเนินการทำความสะอาดตามแนวทางข้างต้น จากข้อมูลดิบ 480 แถว เหลือชุดข้อมูลสะอาดพร้อมใช้งาน 453 แถว (ตัดข้อมูลซ้ำ 25 แถว และแถวที่ device_id ว่าง 2 แถวออก) โดยยังคงเก็บไฟล์ที่มีแถวถูกตีธง (flagged) ไว้ทั้งหมดเพื่อการตรวจสอบย้อนหลัง