# Qurity report e_commerce_raw.csv

จากข้อมูล

| Metric | | Before | After |
|---|---|---|---|
| 0 | Total rows | 630 | 612 |
| 1 | Exact duplicate rows | 18 | 0 |
| 2 | Missing Customer_Email | 48 | 0 |
| 3 | Missing Province | 12 | 0 |
| 4 | Gender format variants | 10 | 2 |
| 5 | Payment_Method format variants | 14 | 5 |
| 6 | Order_Date unparseable/invalid | 0 | 0 |
| 7 | Quantity outside business rule (1-20) | 3 | 3 |
| 8 | Accuracy mismatch (non-promo, >1 baht) | 43 | 43 |
| 9 | Order_Status non-standard rows | 320 | 308 |
| 10 | Outliers outside IQR bound (non-promo) | 19 | 0 |
| 11 | Late loads (Load_Delay_Hours > 24h) | 48 | 48 |

## บทสรุปภาพรวม (Executive Summary)

การทำ Data Cleaning ครั้งนี้มีวัตถุประสงค์เพื่อยกระดับคุณภาพข้อมูล (Data Quality) และเตรียมข้อมูลให้พร้อมสำหรับการนำเข้าสู่ Data Warehouse โดยยึดหลัก

รักษาตัวเลขยอดขาย (Revenue Preservation), จัดรูปแบบมาตรฐาน (Standardization) และ

สร้างระบบย้อนตรวจสอบได้ (Auditability)

จำนวนแถวรวม: ลดลงจาก 630 เหลือ 612 แถว (ลบแถวซ้ำสมบูรณ์ 18 แถว)

Missing Values: จัดการครบ 100% (Email 48 รายการ, Province 12 รายการ) โดยไม่เสีย Fact Table

Standardization: ปรับรูปแบบ Gender (เหลือ M, F) และ Payment Method (เหลือ 5 ช่องทางหลัก) ให้เป็นมาตรฐาน

Outliers: ทำ Capping รายการที่ไม่ใช่โปรโมชัน 19 รายการ ด้วยวิธี IQR

**ด้านความครบถ้วน (Completeness)**
Customer_Email (48 รายการ): เติมด้วย unknown_customer@placeholder.com พร้อมติด Flag ไม่ใช้ dropna() เพราะจะทำให้ยอดขายรวมใน Fact Table สูญหาย
Province (12 รายการ): เติมด้วย 'Unknown' พร้อมติด Flag เพื่อย้อนมาอัปเดตข้อมูลทีหลังได้
**ด้านความสอดคล้อง (Consistency)**
Gender: Map คำสะกดต่างๆ ให้เหลือ M, F และ Unknown
Payment_Method: Map รูปแบบคำสะกด (เช่น CC, CreditCard) ให้เหลือ 5 กลุ่มมาตรฐาน (Credit Card, PromptPay, Cash on Delivery, Bank Transfer, Unknown) Order_Date & Order_Status: แปลงวันที่เป็น YYYY-MM-DD โดยใช้ Load_Timestamp ช่วยตัดสินความถูกต้อง และจัดรูปแบบสถานะคำสั่งซื้อเป็น Title Case

**ด้านความซ้ำซ้อน (Uniqueness)**
Exact Duplicates (18 รายการ): ลบแถวที่ซ้ำกันทุกคอลัมน์โดยเก็บเฉพาะแถวแรกไว้ (keep='first')
ด้านกฎธุรกิจและความถูกต้อง (Accuracy & Validity): Quantity & Accuracy Mismatch: แถวที่จำนวนสินค้าอยู่นอกช่วง 1–20 ชิ้น หรือยอดเงินไม่ตรงกับสูตรคำนวณ จะถูกติด Quality Flag ไว้ เพื่อไม่ให้กระทบโครงสร้างข้อมูล แต่ใช้ส่งต่อให้ทีมที่เกี่ยวข้องตรวจสอบ
**ด้านค่าผิดปกติ (Outliers & IQR)**
Outliers (19 รายการ): คำนวณ IQR เฉพาะกลุ่ม Is_Promotion != 'Y' แล้วทำ Capping ขอบบน/ลบด้วย .clip() โดยไม่ลบรายการโปรโมชันทิ้งเพราะเป็นยอดขายจริงจากแคมเปญ
**ด้านความทันเวลา (Timeliness)**
Late Loads (48 รายการ): คำนวณระยะเวลาโหลดข้อมูล หากเกิน 24 ชั่วโมงจะติด Is_Late_Load = True เพื่อประเมินประสิทธิภาพ Data Pipeline