PORTFOLIO PROJECT | NOVITA ANGGRAINI

Data Quality Audit

& Revenue Integrity Report

Amazon India E-Commerce Dataset [Kaggle] | April 2022

DATASET	RECORDS	TOOLS	TYPE
Amazon India	**128,975 orders**	**Python · Pandas**	**Data Quality Audit**

Ketika Angka Penjualan Tidak Bisa Dipercaya dan Cara Memperbaikinya

AGENDA

Apa yang Akan Kita Bahas

01 **Dataset & Konteks**

Sumber data, struktur, dan use case nyata

02 **Data Validation**

6 business rules untuk mendeteksi anomali

03 **Revenue Integrity Check**

Dampak finansial dari data yang tidak divalidasi

04 **Data Cleaning Process**

5 langkah Data Cleaning

05 **Analisis Deskriptif**

Gambaran umum data bersih: order, revenue, status

06 **Analisis Segmentasi**

Kategori, fulfillment, ukuran, dan segmen B2B

07 **Kesimpulan & Rekomendasi**

Temuan dan langkah yang harus diambil

01 / DATASET

Mengenal Data Sebelum Menyentuhnya

Dataset transaksi e-commerce Amazon India. Berisi 128,975 order dari berbagai kategori produk fashion dengan informasi lengkap mulai dari identitas order, status pengiriman, nilai transaksi, hingga segmen pelanggan.

Total Orders

128,975

transaksi

Total Kolom

24

atribut

Raw Revenue

INR 78.6M

sebelum validasi

Valid Revenue

INR 67.0M

setelah cleaning

Kolom Utama Dataset

Order ID	Identitas unik transaksi
Date	Tanggal order (format MM-DD-YY)
Status	13 status: Shipped, Cancelled, dll
Fulfilment	Amazon atau Merchant
Category	9 kategori produk fashion
Qty	Jumlah unit dipesan
Amount	Nilai transaksi (INR)
B2B	Segmen korporat vs konsumen

01 / USE CASE

Mengapa Kita Tidak Bisa Langsung Analisis?

SKENARIO

Tim finance menerima file CSV transaksi April 2022 dari sistem. Tanpa dicek, angka raw revenue INR 78.6M langsung dipakai untuk laporan bulanan dan proyeksi target Q2.

Masalah: 38.7% data di file tersebut sebenarnya tidak layak dihitung.

18,332 Cancelled Orders

Masih ikut dihitung sebagai revenue. Padahal order dibatalkan = tidak ada uang masuk.

12,807 Qty Tidak Valid

Order dengan jumlah ≤0. Tidak mungkin ada transaksi tanpa kuantitas produk.

8,597 Duplicate Order ID

Order yang sama terhitung dua kali. Menggembungkan jumlah transaksi dan revenue.

Proses di deck ini: validasi → bersihkan → baru analisis.

Pertanyaan yang Harus Dijawab

Kualitas Data

Q1

Seberapa bersih data mentah dari sistem?

Q2

Apakah angka revenue bisa dipercaya?

Q3

Aturan bisnis apa yang harus jadi filter baku?

Bisnis & Strategi

Q4

Kategori mana yang menggerakkan revenue terbesar?

Q5

Metode fulfillment mana yang paling bermasalah?

Q6

Segmen B2B vs konsumen: adakah potensi yang belum digarap?

6 pertanyaan ini menjadi panduan pada seluruh proses analisis dari validasi hingga rekomendasi akhir.

6 Business Rules untuk Mendeteksi Anomali

Rule	Kondisi	Alasan Bisnis	Ditemukan	Status
R1	Qty <= 0	Tidak ada transaksi tanpa kuantitas produk	12,807 (9.9%)	FAIL
R2	Amount null/nol	Transaksi tanpa nilai = noise laporan keuangan	10,138 (7.9%)	FAIL
R3	SKU missing	Produk tanpa identitas tidak bisa dilacak inventory	0 (0.0%)	PASS
R4	Date invalid	Tanggal di masa depan menunjukkan error sistem	0 (0.0%)	PASS
R5	Duplicate Order ID	Order sama tidak boleh terhitung dua kali	8,597 (6.7%)	FAIL
R6	Cancelled orders	Order batal tidak boleh masuk laporan pendapatan	18,332 (14.2%)	FAIL

Total 49,874 record bermasalah dari 128,975 transaksi (38.7%) hampir 2 dari 5 baris tidak layak langsung dianalisis.

04 / REVENUE INTEGRITY

Berapa Besar Angka yang Salah?

RAW REVENUE

INR 78.6M

angka yang beredar
tanpa validasi

VALID REVENUE

INR 67.0M

angka yang seharusnya
dilaporkan

SELISIH

+14.7%

revenue lebih tinggi
dari data real

Sumber Overstatement:

1. 18,332 cancelled orders → INR 6,919,284 revenue salah terhitung
2. 2,466 Amount null + 106 Qty invalid → transaksi fiktif/error
3. 7,344 duplikat order ID → penghitungan ganda volume transaksi

05 / DATA CLEANING

5 Langkah Pembersihan yang Terdokumentasi

1	**Hapus Kolom Artefak** Kolom Unnamed: 22 yang sepenuhnya tidak berguna dihapus. Artefak dari export Excel.	1 col
2	**Filter Cancelled Orders** 18,332 order berstatus Cancelled dihapus. Kontributor utama revenue overstatement INR 6.9M.	-18,332
3	**Filter Qty Tidak Valid** 106 record dengan Qty ≤ 0 dihapus. Setelah cancelled difilter, jumlah ini jauh berkurang.	-106
4	**Filter Amount Null/Nol** 2,466 record dengan nilai Amount kosong dihapus. Tidak memiliki nilai bisnis yang bisa dianalisis.	-2,466
5	**Deduplikasi Order ID** 7,344 duplikat Order ID dideteksi dan dihapus. Ini angka besar yang bisa mendistorsi volume dan revenue.	-7,344

Hasil: 128,975 → 100,727 valid | Revenue terkoreksi INR 78.6M → INR 67.0M | Bias -14.7%

Data sudah bersih.

Sekarang baru kita analisis.

100,727	INR 67.0M	28,248
Valid Orders	Valid Revenue	Record dihapus

Bagian selanjutnya: analisis deskriptif dulu, lalu segmentasi mendalam.

Gambaran Umum Data yang Sudah Bersih

Total Valid Orders
100,727
setelah cleaning

Total Valid Revenue
INR 67.0M
dapat dilaporkan

Average Order Value
INR 665
per transaksi valid

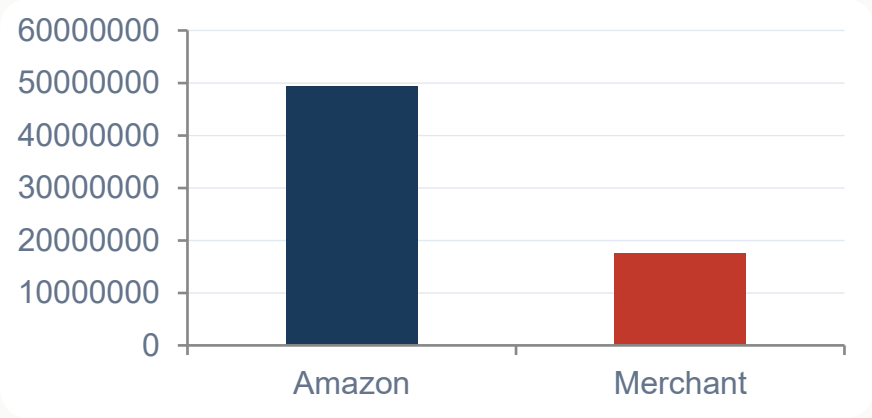
Revenue Diselamatkan
INR 11.6M
dari kesalahan perhitungan

Split Fulfillment

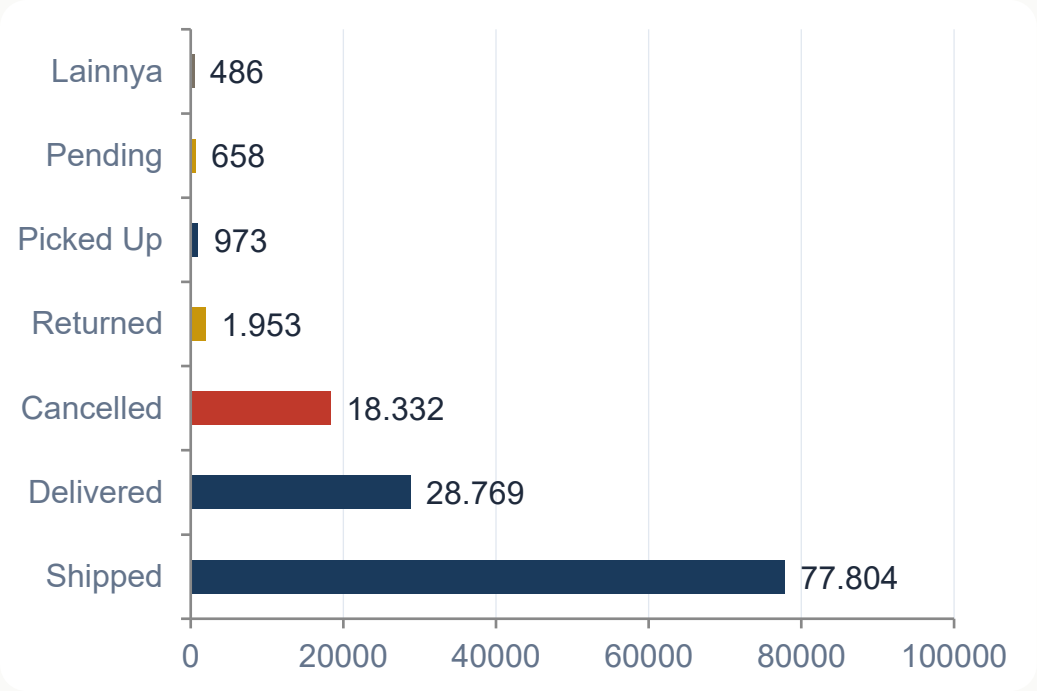
69.5% **Amazon**
89,698 orders

30.5% **Merchant**
39,277 orders

Revenue per Fulfillment



Distribusi Status Order (Sebelum Cleaning)



82.9%
Shipped (delivered/in transit)

14.2%
Cancelled tidak boleh dihitung revenue

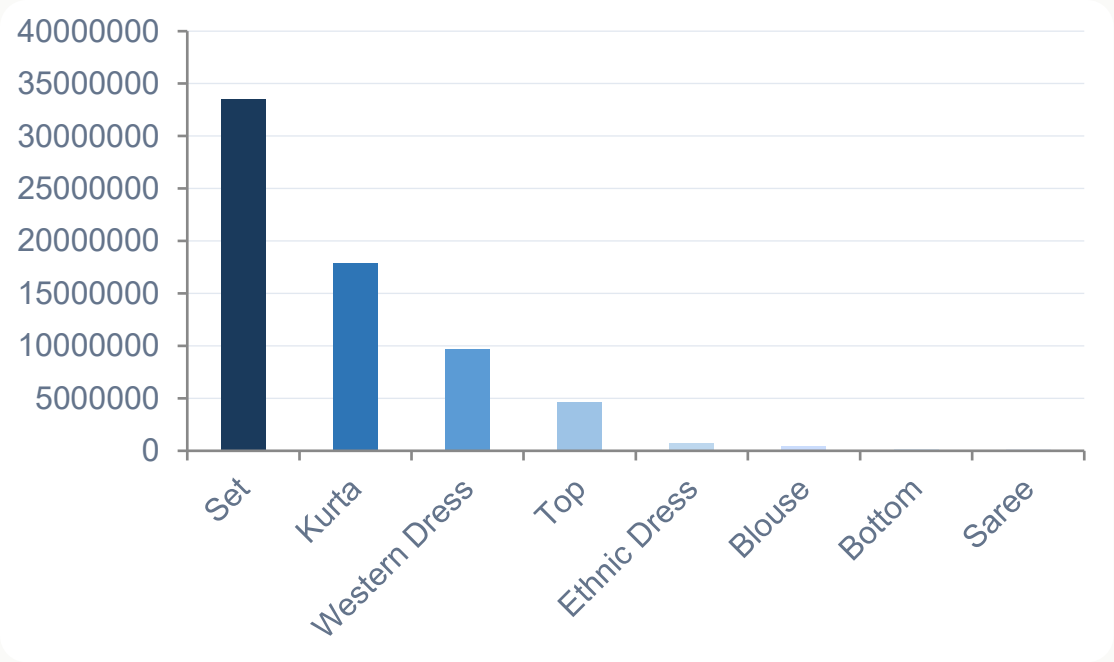
1.5%
Returned revenue sudah hangus

0.8%
Pending / belum ada status final

Status 'Shipped - Delivered to Buyer' dan 'Shipped' bersama mencakup 82.9% orders. Hanya 3 status utama yang perlu perhatian lebih: Cancelled, Returned, Pending.

07 / SEGMENTASI KATEGORI

Kategori Produk: Siapa yang Menggerakkan Revenue?



50.0% Set
INR 33.5M
Satu kategori = separuh revenue

26.7% Kurta
INR 17.9M
Runner-up kuat, bukan sekadar pelengkap

14.4% Western Dress
INR 9.7M
Gap jauh dari Top 2, tapi #3 solid

6.9% Top
INR 4.6M
Volume tinggi, revenue lebih kecil

Set + Kurta = 76.7% total revenue. Keduanya adalah tulang punggung (asik) bisnis yang harus dijaga supply chain dan promosinya.

07 / SEGMENTASI FULFILLMENT

Di Mana Revenue Menguap Sebelum Terwujud?

Amazon Fulfillment

12.8%

CANCELLATION RATE

Lebih Baik

Order diproses langsung oleh Amazon. Standar operasi ketat dan sistem monitoring lebih baik. Cancellation rate yang lebih terkontrol.

Merchant Fulfillment

17.5%

CANCELLATION RATE

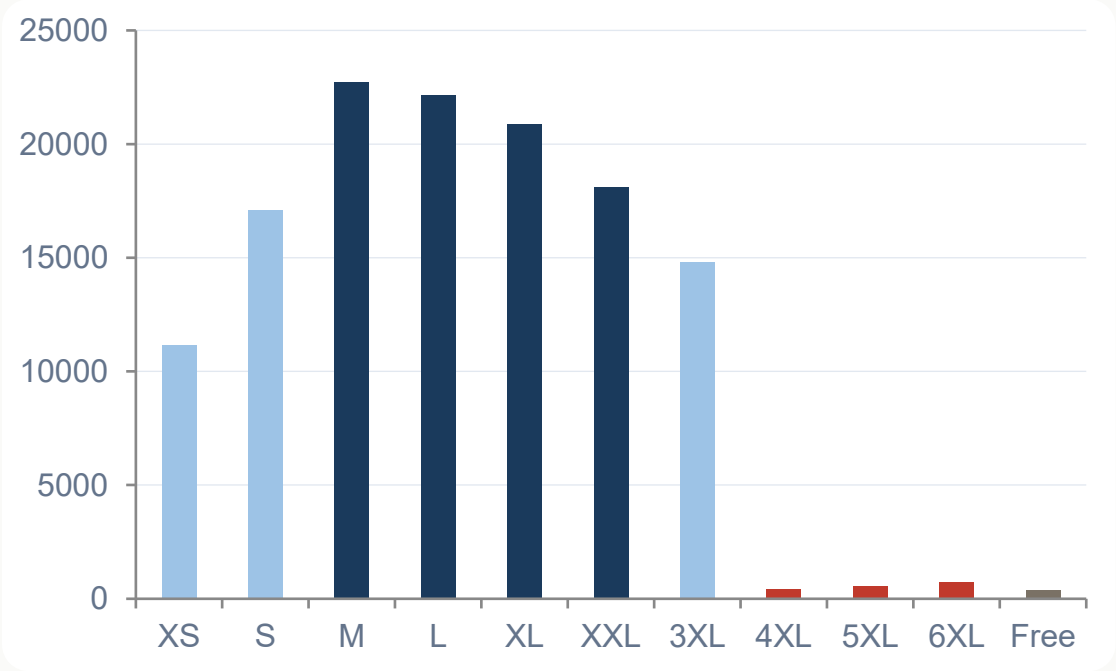
Perlu Audit

Hampir 1 dari 5 order berakhir dibatalkan. Gap 4.7% vs Amazon adalah sinyal masalah operasional yang butuh investigasi segera.

Revenue hilang akibat cancelled orders: INR 6,919,284

Gap 4.7% antara Merchant vs Amazon = peluang recovery jika operasi Merchant diperbaiki

Distribusi Ukuran: Siapa yang Paling Laku?



M–3XL adalah range utama. 5 ukuran ini menyerap >90% volume order. Fokuskan stok dan promosi di range ini.

Top 3 Revenue

01	M	INR 11.9M
02	L	INR 11.3M
03	XL	INR 10.7M

Temuan Penting

▲ M, L, XL, XXL, 3XL mendominasi distribusi cukup merata di range tengah

▼ 4XL, 5XL, 6XL sangat kecil stok perlu dikurangi

07 / SEGMENTASI B2B

Segmentasi Pelanggan: Korporat vs konsumen

Non-B2B (konsumen)

99.2%

REVENUE SHARE

Revenue	Orders	Avg Order Value
INR 66.5M	~99,900	INR 665

B2B (Korporat)

0.8%

REVENUE SHARE

Revenue	Orders	Avg Order Value
INR 535K	~827	INR 704

Insight B2B yang Menarik:

AOV B2B (INR 704) lebih tinggi dari konsumen (INR 665). Meski hanya 0.8% revenue saat ini, segmen korporat membeli dengan nilai lebih besar per transaksi. Ada potensi yang belum digarap terutama jika ada program khusus B2B.

Apa yang Kita Pelajari?

01

38.7% data mentah tidak layak dianalisis langsung

49,874 record bermasalah dari 128,975 total transaksi. Cancelled 14.2%, Qty invalid 9.9%, Duplikat 6.7%, Amount null 7.9%. Tanpa validasi, angka apapun dari dataset ini tidak bisa dipercaya.

02

Revenue overstated 14.7% INR 11.6M selisih

Raw data INR 78,592,678 vs valid revenue INR 67,030,256. Selisih INR 11,562,422 angka ini berasal dari 18,332 cancelled orders, 2,466 Amount null, 106 Qty invalid, dan 7,344 duplikat Order ID.

03

Merchant Fulfillment cancel rate 4.7% lebih tinggi

Notebook mengonfirmasi: Merchant cancel rate 17.5% vs Amazon 12.8%. Revenue hilang dari pembatalan: INR 6,919,284. Gap 4.7% ini konsisten dan harus diinvestigasi, bukan diabaikan.

04

Set + Kurta = 76.7% revenue konsentrasi tinggi

Set: INR 33.5M (50.0%), Kurta: INR 17.9M (26.7%). Dua kategori ini menopang 76.7% total valid revenue. Gangguan supply chain keduanya berdampak sangat besar bagi bisnis.

05

B2B AOV lebih tinggi dari konsumen

B2B hanya 827 orders (0.8% revenue = INR 535K) tapi AOV INR 704 lebih tinggi dari konsumen INR 665. Segmen korporat membeli lebih besar per transaksi peluang yang belum dioptimalkan.

06

M–3XL adalah range ukuran yang perlu prioritas stok

Top 3 revenue: M (INR 11.9M), L (INR 11.3M), XL (INR 10.7M). Range M–3XL menyerap >90% volume order. 4XL ke atas (4XL/5XL/6XL) sangat kecil buffer stok bisa dikurangi dan dialihkan.

09 / REKOMENDASI

Langkah Selanjutnya

HIGH 01

Bangun Validation Pipeline Otomatis

Script Python terjadwal: filter cancelled, Amount null, dan 7K+ duplikat sebelum masuk dashboard.

HIGH 02

Investigasi Sumber Duplikat Order ID

7,344 duplikat bukan angka kecil. Cari tahu titik di sistem order di mana duplikat terjadi.

HIGH 03

Audit & Perbaiki Merchant Fulfillment

Identifikasi merchant dengan cancellation rate tertinggi. Berikan SLA yang lebih ketat atau pertimbangkan churn.

MEDIUM 04

Jaga Supply Chain Set & Kurta

76.7% revenue bergantung pada dua kategori ini. Prioritaskan inventory dan promosi keduanya.

MEDIUM 05

Optimalkan Stok Range M–3XL

Kurangi buffer stok 4XL ke atas. Realokasi ke M, L, XL yang terbukti dominan.

LOW 06

Program Akuisisi B2B

AOV B2B lebih tinggi. Buat program khusus: volume discount, credit terms, atau account manager dedicated.

Terima Kasih

Data yang bersih adalah fondasi dari keputusan yang baik.

49,874	14.7%	7,344	100,727
Record Anomali	Revenue Overstated	Duplicate Orders	Valid Records

Amazon India E-Commerce | Data Quality Audit | April 2022