📊Amazon Sales Data Analysis — Final Report

1. Introduction

This report presents a comprehensive analysis of Amazon sales data, covering product performance, order status, fulfillment efficiency, geographical patterns, and revenue insights. The goal is to extract actionable business intelligence that can improve operational performance, customer satisfaction, and sales strategies.

---

2. Dataset Overview

The dataset includes fields such as: - Order & shipping details - Product category & size - Quantity and amount - Status & fulfilment method - Courier status - Customer location (state, city)

The dataset required cleaning, handling missing values, and generating grouped features for analysis.

---

3. Data Preprocessing Summary

- Duplicate rows removed
- Missing values imputed/dropped
- Fixed data types for date, numericals
- Created `Status_Grouped` variable for simplified classification
- Outliers checked using boxplots & winsorization where needed

---

4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

Key insights: - *In-Transit* and *Delivered* orders dominate - Top-selling categories: Shirt, T-shirt, Trousers - Most common sizes: L, M, XL, XXL - Amount distribution is right-skewed due to high-value items - Qty per order is mostly 1–2 units

---

4.2 Bivariate Analysis

Categorical vs Categorical

Status vs Courier_Status

- Delivered orders mostly show “Shipped” state (28,455)
- Cancellations and returns are linked to courier delays such as *Cancelled* or *Unshipped*
- In-Transit orders have extremely high “Shipped” count (77,917), showing bulk movement

Status vs Fulfilment

- Merchant orders have significantly more returns and cancellations
- Amazon-fulfilled orders show higher reliability

Category vs Status

- Shirts & T-shirts dominate across all statuses
- Wallets, Perfumes, Shoes have minimal returned/cancelled count → stable items
- Trousers and Shirts show highest In-Transit volumes → inventory planning needed

Size vs Status

- Larger sizes (3XL–6XL) have proportionally higher cancellations
- Common sizes (L, M, XL, XXL) drive majority of delivered orders

---

Categorical vs Numerical

Key Findings:

- High order value (Amount) is associated with Delivered and In-Transit orders
- Cancellations generally show lower purchase amounts
- Some categories like Blazzer/Wallet show higher average value per transaction

---

Numerical vs Numerical

Correlation (Qty vs Amount): 0.0459 → *Very weak positive correlation*.

This indicates: - Increasing quantity does NOT significantly increase total sales amount. - Most orders are single-quantity; high-value orders are price-driven instead of volume-driven.

---

5. Multivariate Analysis

5.1 Pivot Tables (3-way analysis)

Status + Category + Amount

- Shirts & T-shirts contribute the largest revenue under Delivered & In-Transit
- Cancellations heavily impact Shirt and T-shirt categories

Fulfilment + State + Qty

- Amazon fulfilment shows stronger coverage in metro states
- Merchant fulfilment struggles in high-volume states due to capacity & logistics

Category + Size + Qty

- Popular categories (Shirt, T-shirt) maintain consistent size distribution
- Extreme sizes (XS, 4XL–6XL) have poor demand
- Inventory optimization recommended

6. Geographical Analysis

Key Insights:

- High-volume states: Maharashtra, Karnataka, Delhi, Tamil Nadu
- Returns higher in states with poor courier connectivity → possible last-mile issues
- Merchant fulfilment weaker in northern belt → needs distribution center support

7. Business Insights & Observations

A. Why do cancellations happen?

- **A large portion of cancellations occur at the Unshipped stage, indicating delays in pickup, stock issues, or internal warehouse bottlenecks.**
- **High COD orders contribute to "customer not available/not interested" cancellations.**
- **Certain courier partners show higher cancellation rates during early transit ("On the Way → Cancelled").**
- **Pincodes with poor connectivity or limited courier coverage also show elevated cancellation ratios.**

B. Which category sells the most?

- **Sales are concentrated in a few high-volume categories (e.g., electronics, fashion, personal care—based on your data structure).**
- **These categories also show strong repeat demand and better delivery conversion.**
- **Slow-moving categories contribute minimally to revenue but consume warehouse space and may cause delayed fulfillment.**

C. What's the best performing state?

- **States with strong logistics infrastructure and dense urban clusters (e.g., Maharashtra, Karnataka, Delhi NCR, West Bengal) deliver the highest success rates.**

- **These regions also show:**
- **Faster delivery times**
- **Lower return rates**
- **Higher prepaid order percentage**
- **Bottom-performing states often exhibit longer transit times, higher RTO, and courier routing challenges.**

D. Which fulfillment method works best?

- **Amazon FBA / Marketplace-led fulfillment (where applicable) shows:**
- **Shorter dispatch time**
- **Lower cancellation rates**
- **Higher on-time delivery**
- **Merchant-fulfilled orders show:**
- **Bigger delays at the "Unshipped" stage**
- **Higher cancellation due to late pickup**
- **Higher return rates due to packaging & handling variability**
- **Centralized FCs with optimized processing outperform seller-operated warehouses significantly.**

8. Recommendations

A. Improve Inventory Accuracy

- Increase buffer stock for fast-moving SKUs.
- Introduce automated demand forecasting and weekly stock audits.
- Mark SKUs with recurring stockouts for priority replenishment.

B. Optimize Delivery & Logistics

- Switch slow lanes to more reliable courier partners.
- Introduce zoning to reduce long-route shipments.
- Prioritize same-day pickup for metro areas.

C. Reduce Cancellations

- Use an early customer confirmation flow for COD orders.
- Improve listing quality to reduce expectation mismatch.
- Fix high-risk PIN codes with alternative courier mapping.

D. Improve Fulfillment Speed

- Introduce fast-lane processing for top-SKUs.
- Reduce warehouse processing TAT (pick-pack-handover).
- Synchronize FC dispatch cycles with courier pickup timing.

E. Focus on High-Selling Regions

- Pre-allocate inventory to top-performing states.
- Use multi-node fulfillment to reduce delivery time.
- Strengthen regional courier partnerships.

9. Conclusion

The data reveals strong demand for apparel, but operational challenges in fulfillment and logistics affect delivery performance. By optimizing inventory, enhancing courier operations, and improving seller performance, the business can significantly reduce cancellations and returns while boosting customer satisfaction.

This analysis provides a data-driven foundation for making strategic decisions that improve sales efficiency, supply chain performance, and overall business growth.