Customer & Sales Insights from Amazon Clothing Transactions

## **Learning Objectives**

By completing this project, you will be able to:

1. **Clean and preprocess** a real-world messy dataset (missing values, inconsistent types, outliers).
2. Formulate and test **business-oriented hypotheses**.
3. Perform **univariate, bivariate, and multivariate** exploratory analysis.
4. Derive **actionable insights** from sales, returns, customer behavior, pricing, and delivery performance.
5. **Visualize** findings effectively with appropriate charts and narrative.
6. Structure code for reproducibility and shareable storytelling via **GitHub** (README, notebooks, summary).
7. Communicate data-driven recommendations as if advising a **product/operations team**.

## **Dataset**

Amazon Clothing Sales 2025 [ [link](https://docs.google.com/spreadsheets/d/1XU8dWfq_MbKqgRsXFcUNZCtpnk2fWmNlMX-Hq1IO4mc/edit?usp=sharing) ]

This dataset contains 25,000+ Amazon clothing sales **records scrapped in 2024–2025** from publicly available product pages, seller listings, and transactional overlays.  
  
It features purchase-level data across Men’s, Women’s, Kids’, and Baby categories. Each record includes product-level attributes, pricing and discount metrics, customer demographics, purchase details, and device-level access behavior.

**Schema & Feature Descriptions**

| Column Name | Description |
| --- | --- |
| order\_id | Alphanumeric Amazon-style Order ID |
| customer\_id | Unique customer reference |
| product\_id | Product identifier (ASIN format) |
| product\_name | Product brand and title |
| main\_category | Clothing category: Men, Women, Kids, Baby |
| sub\_category | Specific item type (e.g., Shirts, Dresses, Jackets) |
| brand | Product brand (e.g., Nike, Zara, Carter’s) |
| price | Unit price of the product in USD |
| quantity | Quantity purchased |
| discount\_percent | Discount offered (%) |
| final\_price | Final total paid after discount |
| payment\_method | Payment method used (e.g., Credit Card, PayPal) |
| review\_rating | Customer rating on a 1–5 scale |
| order\_date | Purchase date |
| delivery\_days | Delivery time in days |
| is\_returned | Return status: 1 = returned, 0 = not returned |
| region | US region of delivery |
| customer\_age\_group | Age bucket of the customer |
| device\_type | Device used to complete the purchase |

# 

## **Step-by-Step EDA Project Guideline**

### **A. Initial Data Exploration & Cleaning**

1. **Schema & Types**: Check data types. Convert order\_date to datetime.
2. **Missing Values**:  
   * Quantify missingness per column.
   * Decide strategy: impute, flag, or drop
3. **Duplicates & Integrity**:  
   * Are there duplicates?
   * Consistency between price, discount\_percent, and final\_price (validate formula).
4. **Outliers**:  
   * Check and treat outliers.
5. **Feature Engineering**:  
   * Extract order\_month, order\_weekday, order\_year from order\_date.
   * Compute discount\_amount = price \* discount\_percent / 100.
   * Create unit\_price = final\_price / quantity.
   * Flag “fast” vs “slow” delivery (e.g., delivery\_days threshold).
   * Customer-level aggregates: total spend, total orders, average rating given, return rate.

### **B. Univariate Analysis**

* Distribution of price, final\_price, discount\_percent, review\_rating, delivery\_days.
* Counts / frequency of main\_category, sub\_category, brand, payment\_method, region, customer\_age\_group, device\_type.
* Returned vs non-returned proportion.

### **C. Bivariate & Multivariate Analysis**

1. **Sales Performance**:  
   * Revenue over time (monthly / weekly trend).
   * Which categories/brands drive most revenue?
   * Price vs quantity sold (does higher price suppress quantity?).
2. **Discount Effectiveness**:  
   * Relationship between discount percent and quantity sold / revenue.
3. **Returns Analysis**:  
   * Return rate by category, brand, customer age group, device type.
4. **Customer Behavior**:  
   * High-value customers vs one-time buyers (segmentation).
   * Average order value (AOV) by region / age group.
5. **Delivery Performance**:  
   * Delivery days distribution and its effect on review rating / return probability.
6. **Rating Signal**:  
   * Average rating by brand, category, region.
   * Correlate low ratings with subsequent returns.
7. **Payment Method Insights**:  
   * Does payment method influence return likelihood or average spend?
8. **Cross-analysis**:  
   * E.g., return rate for mobile users vs desktop, region vs delivery performance, age group vs discount sensitivity.

### **D. Hypothesis Ideas (picks/answers at least 5)**

1. Higher discounts lead to higher quantity sold but also higher return rates.
2. Longer delivery times negatively impact review ratings and increase returns.
3. Certain age groups (e.g., 25–34) have higher average order values.
4. Customers ordering from specific regions have systematically different return behavior.
5. Mobile shoppers behave differently in terms of average spend or returns compared to desktop.
6. Premium brands (e.g., Nike) have lower return rates despite higher prices.
7. Orders with low review ratings are predictive of being returned.
8. Payment method “Gift Card” orders have distinct patterns in rating or returns.

### **E. Advanced Analyses**

1. **Cohort Analysis**: Group customers by their first purchase month and track retention (repeat orders) over time.
2. **Customer Segmentation**: Cluster customers based on RFM (Recency, Frequency, Monetary) and behavior (returns, rating).
3. **Price Elasticity Estimation**: Approximate sensitivity of quantity sold to price/discount changes.
4. **Return Prediction Feature Ideas**: Build a simple logistic regression or decision tree prototype to predict whether an order will be returned (feature importance discussion suffices for EDA).
5. **Anomaly Detection**: Identify suspicious orders (e.g., extremely high discount with high quantity + immediate return).
6. **Time-to-Delivery Hotspots**: Identify regions or product types with consistent delivery delays.

### **F.Visualization Expectations**

Must include appropriate visualizations; examples:

* Time series line charts (revenue trend, order volume)
* Boxplots / violin plots (price distribution by category)
* Heatmaps (correlation matrix, return rate by category × region)
* Bar charts (top brands, top regions, payment method breakdown)
* Scatter plots with regression lines (price vs quantity, delivery\_days vs rating)
* Customer segmentation visuals (cluster scatter, RFM buckets)
* Highlight comparisons (returned vs non-returned) with side-by-side plots

*Each plot must have: title, axis labels, legend (if needed), and a one-sentence takeaway caption.*

## **Project Structure / Deliverables**

Each participant must submit the following on GitHub:

1. **Cleaned Jupyter Notebook or well-structured Python script** performing the full EDA.
2. **README.md** explaining the project, methodology, key findings, and instructions to reproduce.
3. **Data Quality Report** (could be a section in notebook) listing missingness, anomalies, and cleaning steps.
4. **Insight Summary**: Top 5–8 business insights with visual evidence and recommended actions.
5. **Hypothesis Testing Section**: At least 5 hypotheses examined with supporting analysis.
6. **Visualizations**: Clearly labeled plots (static or interactive) embedded in the notebook.
7. **GitHub Repository** with:  
   * Proper folder structure
   * Requirements file (requirements.txt)
   * Code modularization (e.g., separate data cleaning, analysis, utility functions)