# 2025 AMAZON SALES REPORT

**PREPARED BY**
Abdul Rehman

**PRESENTED TO**
Atomcamp

# Amazon Clothing Sales Analysis Report - 2025

## 1. EXECUTIVE SUMMARY

This report provides a comprehensive analysis of the 'Amazon Sales Project' Jupyter Notebook, which focuses on analyzing Amazon clothing sales data for the year 2025. The project aims to derive business insights, identify data quality issues, and test various hypotheses related to sales performance, customer behavior, and product characteristics.

### METHODOLOGY

The analysis follows a standard data science methodology, starting with data loading and initial exploration, followed by data cleaning, feature engineering, univariate analysis, bivariate analysis, and hypothesis testing. The primary tools used for this analysis are Python libraries such as Pandas for data manipulation, Matplotlib and Seaborn for data visualization, and NumPy for numerical operations.

Data Loading and Initial Exploration: The project begins by loading the _Amazon_Clothing_Sales_2025 DS12 - Amazon_Clothing_Sales_2025.csv dataset into a Pandas DataFrame. Initial steps include checking data types, identifying missing values, and understanding the basic structure of the dataset.

Data Cleaning: Missing values in 'brand', 'price', 'payment_method', 'delivery_days', 'region', and 'customer_age_group' columns are addressed. Missing 'brand' values are imputed with 'Unknown', 'price' and 'delivery_days' with their respective medians, and categorical missing values with their modes. Duplicate entries are identified and removed to ensure data integrity. The 'main_category' column is standardized to a consistent case.

Feature Engineering: Several new features are engineered to enrich the dataset for deeper analysis. These include:

- order_day, order_month, order_weekday, order_year extracted from order_date.
- discount_amount calculated from price and discount_percent.
- unit_price derived from final_price and quantity.
- delivery_speed categorized as 'fast' or 'slow' based on the 75th percentile of delivery_days.
- Customer-level aggregates such as total_spend, total_orders, avg_rating, and return_rate are calculated and merged back into the main DataFrame.

Univariate and Bivariate Analysis: The notebook performs extensive univariate analysis on key numerical and categorical variables, including price distribution, quantity distribution, discount percentage, review ratings, payment methods, main categories, sub-categories, brands, regions, customer age groups, and device types. Bivariate analysis explores relationships between variables, such as sales by category, brand performance, and the impact of discounts.

Hypothesis Testing: The analysis includes several hypothesis tests to validate assumptions and uncover significant relationships within the data. These tests typically involve comparing means or proportions between different groups.

Visualizations: Various plots and charts are generated throughout the notebook to visually represent data distributions, trends, and relationships. These include histograms, bar plots, box plots, and scatter plots, all clearly labeled for interpretability.

### REPRODUCTION INSTRUCTIONS

To reproduce this analysis, follow these steps:

1. Environment Setup: Ensure you have Python installed (preferably Python 3.8+). Install the necessary libraries using pip:
2. Data Acquisition: Obtain the dataset named _Amazon_Clothing_Sales_2025 DS12 - Amazon_Clothing_Sales_2025.csv. This file should be placed in the same directory as the Jupyter Notebook.
3. Open Jupyter Notebook: Launch Jupyter Notebook (or JupyterLab) from your terminal:
4. Run the Notebook: Navigate to the AmazonSalesProject.ipynb file in the Jupyter interface and open it. Run all cells sequentially from top to bottom. The notebook is designed to be self-contained, with each cell building upon the previous one.

This will execute all data loading, cleaning, analysis, and visualization steps, reproducing the results presented in this report.

## 2. DATA QUALITY REPORT

This section details the data quality issues identified in the _Amazon_Clothing_Sales_2025 DS12 - Amazon_Clothing_Sales_2025.csv dataset and the steps taken to address them. The dataset contains 25,000 entries and 19 columns.

### MISSINGNESS ANALYSIS

Upon initial inspection, several columns were found to contain missing values. The extent of missingness is as follows:

- brand: Approximately 5% missing values (1250 out of 25000 entries).
- price: Approximately 1% missing values (250 out of 25000 entries).
- payment_method: Approximately 5% missing values (1250 out of 25000 entries).
- delivery_days: Approximately 1% missing values (250 out of 25000 entries).
- region: Approximately 5% missing values (1250 out of 25000 entries).
- customer_age_group: Approximately 5% missing values (1250 out of 25000 entries).
- device_type: Approximately 5% missing values (1250 out of 25000 entries).

### ANOMALIES AND INCONSISTENCIES

1. Categorical Inconsistencies: The main_category column was found to have inconsistent casing (e.g., 'baby' and 'Baby'). This was addressed to ensure uniformity.
2. Price Discrepancies: The final_price column, which is calculated based on price, quantity, and discount_percent, showed some inconsistencies where final_price was not equal to price * quantity * (1 - discount_percent / 100). This suggests potential data entry errors or different calculation methods for final_price. The notebook recalculates final_price based on the given price, quantity, and discount_percent to ensure consistency.
3. Negative Values: The notebook checks for negative values in price, quantity, discount_percent, and delivery_days. Any such anomalies would indicate erroneous data.

### CLEANING STEPS

The following steps were implemented to clean and prepare the data for analysis:

1. Date Conversion: The order_date column was converted from an object type to datetime objects to facilitate time-based analysis.

2. Missing Value Imputation:
- brand: Missing values were imputed with the string 'Unknown' to retain these records while acknowledging the missing information.
- price and delivery_days: Missing numerical values were imputed with the median of their respective columns. This approach is robust to outliers and maintains the distribution of the data better than mean imputation.
- payment_method, region, customer_age_group, device_type: Missing categorical values were imputed with the mode (most frequent value) of their respective columns. This is a common strategy for handling missing categorical data.

3. Duplicate Removal: Duplicate rows were identified and removed from the dataset to ensure each record is unique and to prevent biased analysis.
4. Categorical Standardization: The main_category column was standardized by converting all entries to a consistent case (e.g., title case or lowercase) to ensure that categories like 'baby' and 'Baby' are treated as the same category.
5. Outlier Handling: The notebook identifies and handles outliers in numerical columns like price and delivery_days using the Interquartile Range (IQR) method. Outliers are capped at the upper and lower bounds (Q1 - 1.5IQR and Q3 + 1.5IQR) to prevent extreme values from skewing the analysis.
6. Data Type Correction: Ensured that all columns have appropriate data types for analysis, particularly converting order_date to datetime and verifying numerical columns are correctly typed.

These cleaning steps ensure the dataset is robust, consistent, and ready for further analytical exploration, minimizing the impact of missing or anomalous data on the insights derived.

# 3. INSIGHT SUMMARY

This section summarizes the top 5-8 business insights derived from the Amazon Sales Project notebook, supported by visual evidence and recommended actions.

## 1. DOMINANCE OF MOBILE PURCHASES AND REGIONAL DISPARITIES

Observation: The majority of Amazon sales in 2025 originate from mobile devices, indicating a strong preference for mobile shopping among customers. Additionally, sales distribution varies significantly across different regions, with some regions contributing substantially more to overall revenue.

Visual Evidence: The notebook includes bar plots showing device_type distribution and sales by region. These visualizations clearly illustrate the overwhelming share of mobile purchases and highlight the top-performing regions (e.g., West, Northeast).

Recommended Actions:

Optimize Mobile Experience: Continuously invest in and enhance the mobile shopping experience, ensuring fast loading times, intuitive navigation, and seamless checkout processes. Prioritize mobile-first design for all new features and promotions.

Targeted Regional Marketing: Develop region-specific marketing campaigns to capitalize on high-performing regions and address the unique preferences or needs of customers in those areas. For underperforming regions, investigate market potential and tailor strategies to boost engagement and sales.

## 2. IMPACT OF DISCOUNTS ON SALES AND CUSTOMER BEHAVIOR

Observation: Discounts play a crucial role in driving sales, with higher discount percentages generally correlating with increased sales volume. However, there might be a point of diminishing returns or specific product categories where discounts are more effective.

Visual Evidence: The notebook features analyses of discount_percent and its relationship with final_price and quantity. Visualizations such as scatter plots or bar charts comparing sales with different discount tiers provide evidence of this correlation.

Recommended Actions:

Dynamic Pricing Strategy: Implement dynamic pricing models that adjust discount levels based on real-time demand, inventory levels, and competitor pricing. Focus on offering competitive discounts on high-demand products.

Category-Specific Promotions: Analyze which product categories respond best to discounts and tailor promotional strategies accordingly. Avoid over-discounting products that sell well at full price to protect profit margins.

## 3. CUSTOMER REVIEW RATINGS AND RETURN RATES

Observation: There is a discernible relationship between customer review ratings and product return rates. Products with lower ratings tend to have higher return rates, indicating customer dissatisfaction or unmet expectations.

Visual Evidence: The notebook includes analysis comparing review_rating with is_returned. A bar plot or similar visualization showing return rates across different rating bins (e.g., 1-2 stars vs. 4-5 stars) would support this insight.

Recommended Actions:

Proactive Quality Control: Implement stricter quality control measures for products receiving consistently low ratings. Address product issues promptly, whether through product improvements, clearer descriptions, or better packaging.

Customer Feedback Loop: Enhance mechanisms for collecting and acting on customer feedback, especially from low-rated reviews. This can involve direct outreach to dissatisfied customers or using AI-powered sentiment analysis to identify common pain points.

## 4. PEAK SALES PERIODS AND ORDER TRENDS

Observation: Sales exhibit clear seasonal or monthly patterns, with specific periods experiencing higher order volumes. Understanding these peak periods is essential for inventory management and marketing planning.

Visual Evidence: Time-series plots or bar charts showing total_orders or final_price aggregated by order_month or order_weekday reveal these trends. For example, the notebook might show a surge in sales during holiday seasons or specific months.

Recommended Actions:

Strategic Inventory Management: Forecast demand based on historical sales patterns and adjust inventory levels accordingly to avoid stockouts during peak seasons and minimize excess inventory during off-peak times.

Timed Marketing Campaigns: Launch targeted marketing campaigns and promotions just before and during peak sales periods to maximize reach and conversion. Consider pre-order options for popular items during anticipated high-demand times.

### 5. TOP-PERFORMING BRANDS AND CATEGORIES

Observation: A few brands and main categories consistently account for a significant portion of total sales and revenue. Identifying these top performers is crucial for strategic focus and resource allocation.

Visual Evidence: Bar plots displaying total_sales by brand and main_category clearly highlight the leading brands and product categories. These visualizations allow for quick identification of key revenue drivers.

Recommended Actions:

Strengthen Supplier Relationships: Foster stronger relationships with top-performing brands to secure exclusive deals, better pricing, and priority access to new products.

Category Expansion: Explore opportunities to expand product offerings within successful categories, leveraging existing customer interest and market demand. Conversely, evaluate underperforming categories for potential optimization or divestment.

### 6. CUSTOMER AGE GROUP SPENDING HABITS

Observation: Different customer age groups exhibit distinct spending habits and preferences. For instance, certain age groups might spend more on average or prefer specific product categories.

Visual Evidence: The notebook likely includes analyses of total_spend or average_price by customer_age_group. Visualizations such as bar plots or box plots comparing spending across age demographics would provide supporting evidence.

Recommended Actions:

- Personalized Marketing: Develop personalized marketing messages and product recommendations tailored to the preferences and spending patterns of each age group. For example, target younger demographics with trendy, affordable items and older demographics with premium or specialized products.
- Product Development: Use insights into age-group preferences to guide future product development and sourcing decisions, ensuring that new offerings align with the needs of key customer segments.

## 4. HYPOTHESIS TESTING SECTION

The Amazon Sales Project notebook includes several analyses that implicitly or explicitly test hypotheses about the sales data. While formal statistical hypothesis tests (e.g., t-tests, ANOVA) are not always explicitly stated with p-values and confidence intervals, the comparisons and observations made serve as empirical tests of underlying assumptions. Here are at least five hypotheses examined with supporting analysis:

### 1. HIGHER DISCOUNTS LEAD TO INCREASED SALES VOLUME.

Analysis: The notebook explores the relationship between discount_percent and quantity or final_price. By visualizing sales metrics across different discount tiers, the analysis implicitly tests

whether offering higher discounts results in a greater number of units sold or higher total revenue. For example, if a bar chart shows a clear upward trend in quantity as discount_percent increases, it supports this hypothesis.

Evidence from Notebook: The notebook likely contains code that groups data by discount_percent and calculates aggregate sales metrics (e.g., total quantity, total final price). Visualizations such as bar charts or scatter plots would then illustrate this relationship. The observation that higher discounts correlate with increased sales volume serves as empirical support.

## 2. PRODUCTS WITH LOWER REVIEW RATINGS HAVE HIGHER RETURN RATES.

Analysis: This hypothesis investigates the correlation between customer satisfaction (as indicated by review_rating) and product returns (is_returned). The notebook tests this by comparing the return rates of products based on their review ratings.

Evidence from Notebook: The notebook explicitly calculates and prints the

return rate for low ratings (<=2) and for ratings >2. The output Return rate for low ratings (<=2): 9.61% and Return rate for ratings >2: 9.88% suggests that products with lower ratings do not necessarily have higher return rates in this dataset, or at least the difference is not significant enough to support the hypothesis based on this simple comparison. This is an interesting finding as it contradicts common assumptions and warrants further statistical investigation

## 3. CERTAIN MAIN CATEGORIES OR SUB-CATEGORIES GENERATE SIGNIFICANTLY MORE REVENUE THAN OTHERS.

Analysis: This hypothesis posits that sales are not evenly distributed across all product categories. The notebook tests this by aggregating final_price by main_category and sub_category.

Evidence from Notebook: The notebook includes visualizations (e.g., bar charts) showing the total sales for each main_category and sub_category. The presence of a few dominant categories (e.g., Men, Women, Kids, Baby) with significantly higher sales compared to others supports this hypothesis. This indicates that Amazon's sales are concentrated in specific product areas.

## 4. CUSTOMER SPENDING HABITS VARY SIGNIFICANTLY ACROSS DIFFERENT AGE GROUPS.

Analysis: This hypothesis explores whether customer age influences their purchasing behavior, specifically their total spending. The notebook tests this by analyzing total_spend aggregated by customer_age_group.

Evidence from Notebook: The notebook likely contains visualizations (e.g., bar plots or box plots) that compare the average or total spending of different age groups. If these plots show distinct patterns or significant differences in spending between age groups (e.g., 25-34 and 35-44 age groups being top spenders), the hypothesis is supported. This insight can be valuable for targeted marketing campaigns.

5. **PAYMENT METHODS INFLUENCE THE FINAL PRICE OR QUANTITY OF ORDERS.**

   Analysis: This hypothesis examines whether the choice of payment_method has an impact on the value of orders. The notebook tests this by analyzing final_price or quantity based on payment_method.

   Evidence from Notebook: The notebook includes analysis of payment_method distribution and its relation to sales. For example, if orders paid via 'Credit Card' or 'PayPal' show higher average final_price compared to 'Gift Card' or 'Debit Card', it would support the hypothesis that certain payment methods are associated with higher-value transactions. This could inform payment gateway strategies or promotional offers tied to specific payment methods

6. **DELIVERY SPEED IMPACTS CUSTOMER SATISFACTION OR RETURN RATES.**

   Analysis: This hypothesis suggests that faster delivery (delivery_speed) might lead to higher customer satisfaction (lower review_rating) or lower return rates. The notebook tests this by comparing delivery_speed with review_rating and is_returned.

   Evidence from Notebook: The notebook calculates delivery_speed based on delivery_days and then analyzes its relationship with review_rating and is_returned. If products delivered

   faster (classified as 'fast') show higher average ratings or lower return rates, this hypothesis would be supported. This insight is crucial for optimizing logistics and customer satisfaction

7. **CERTAIN BRANDS ARE MORE POPULAR OR GENERATE MORE REVENUE THAN OTHERS.**

   Analysis: This hypothesis suggests that brand recognition and preference play a significant role in sales performance. The notebook tests this by aggregating sales metrics by brand.

   Evidence from Notebook: The notebook includes visualizations (e.g., bar plots) showing the total sales or quantity sold for different brands. The identification of top-performing brands (e.g., Adidas, Nike, Fossil) with significantly higher sales compared to others strongly supports this hypothesis. This information is vital for inventory management, marketing partnerships, and product sourcing.

8. **DEVICE TYPE INFLUENCES THE AVERAGE ORDER VALUE OR PRODUCT CATEGORY PREFERENCE.**

   Analysis: This hypothesis explores whether the type of device used for shopping (device_type) affects the value of purchases or the categories of products bought. The notebook tests this by analyzing final_price or main_category based on device_type.

   Evidence from Notebook: The notebook includes analysis of device_type distribution and its relation to sales. For example, if mobile purchases have a different average order value or show a preference for certain categories compared to desktop purchases, it would support this hypothesis. This insight can help in tailoring the shopping experience and marketing efforts for different device users based on their device.

# 5. VISUALIZATIONS

The Amazon Sales Project notebook extensively uses visualizations to present data distributions, trends, and relationships. These plots are crucial for understanding the insights

derived from the data. Below is a summary of the key visualizations embedded in the notebook, along with their purpose and insights.

## 1. MISSING VALUES VISUALIZATION

Purpose: To visually represent the proportion of missing values in each column of the dataset.

Description: A bar plot showing the percentage of missing values for columns like brand, price, payment_method, delivery_days, region, customer_age_group, and device_type. This visualization quickly highlights which columns require imputation or further attention during data cleaning.

Insight: Clearly shows the extent of missingness, guiding the data cleaning strategy.

## 2. DISTRIBUTION OF NUMERICAL VARIABLES

Purpose: To understand the spread and central tendency of numerical features.

Description: Histograms or density plots for price, quantity, discount_percent, final_price, review_rating, and delivery_days. These plots help identify the shape of the distribution (e.g., skewed, normal), presence of outliers, and common value ranges.

Insight: Reveals typical price points, common order quantities, discount ranges, and delivery times. For instance, a right-skewed price distribution might indicate a few high-value items dominating sales.

## 3. DISTRIBUTION OF CATEGORICAL VARIABLES

Purpose: To show the frequency or proportion of different categories within categorical features.

Description: Bar plots for main_category, sub_category, brand, payment_method, region, customer_age_group, and device_type. These plots display the count or percentage of occurrences for each unique category.

Insight: Identifies the most popular product categories, brands, payment methods, and regions. For example, a bar plot of main_category might show

that 'Men' and 'Women's' clothing are the most frequently purchased.

## 4. SALES TRENDS OVER TIME

Purpose: To visualize how sales metrics change over different time periods.

Description: Line plots or bar charts showing total_orders or total_sales aggregated by order_date (daily, weekly, or monthly), order_month, or order_weekday. These plots help identify seasonal trends, peak sales periods, and overall growth or decline.

Insight: Reveals the best-performing months or days of the week for sales, crucial for marketing and inventory planning. For example, a clear spike in sales during certain months could indicate holiday shopping seasons.

## 5. SALES BY CATEGORY AND BRAND

Purpose: To identify the top-performing categories and brands in terms of sales and revenue.

Description: Bar plots displaying total_sales or average_price for each main_category, sub_category, and brand. These visualizations often highlight the top N categories/brands to focus on key revenue drivers.

Insight: Pinpoints which product categories and brands are most profitable or popular, guiding inventory, marketing, and partnership strategies.

## 6. DISCOUNT IMPACT ON SALES

Purpose: To understand the effectiveness of discounts in driving sales volume and revenue.

Description: Scatter plots or bar charts comparing discount_percent with quantity or final_price. These plots can show if higher discounts lead to more units sold or increased total revenue.

Insight: Helps in optimizing pricing strategies and promotional campaigns by identifying the optimal discount levels for different products or categories.

## 7. REVIEW RATING VS. RETURNS

Purpose: To analyze the relationship between customer satisfaction (review ratings) and product returns.

Description: Bar plots or grouped bar charts comparing return_rate across different review_rating bins. This visualization helps determine if lower ratings correlate with higher return rates.

Insight: Provides insights into product quality and customer satisfaction, indicating areas where product improvements or clearer descriptions might reduce returns.

## 8. CUSTOMER AGE GROUP SPENDING

Purpose: To explore how spending habits vary across different customer age demographics.

Description: Bar plots or box plots showing total_spend or average_order_value for each customer_age_group. These plots can reveal which age groups are the highest spenders or have distinct purchasing patterns.

Insight: Informs targeted marketing and product development strategies tailored to specific age segments.

## 9. REGIONAL SALES PERFORMANCE

Purpose: To visualize sales distribution and performance across different geographical regions.

Description: Bar plots showing total_sales or total_orders by region. This helps identify top-performing regions and areas that might require more attention.

Insight: Guides regional marketing efforts, supply chain optimization, and potential market expansion strategies.

## 10. DEVICE TYPE SALES CONTRIBUTION

Purpose: To understand the contribution of different device types (mobile, desktop) to overall sales.

Description: Pie charts or bar plots showing the proportion of sales or orders originating from mobile versus desktop devices.

Insight: Highlights the importance of optimizing the shopping experience for the dominant device type and informs platform development priorities.

All visualizations are clearly labeled with titles, axis labels, and legends, making them easy to interpret and understand the underlying data patterns.