

**📦Amazon Sales Dataset Analysis Report**

**About the Author**

**Name: *Deep Makadia*  
B.Tech(CSE)(Pursuing)**

**Project Title: Amazon Sales Dataset Analysis  
Date: *June 2025***

**Author Introduction**

I am **Deep Makadia**, a data enthusiast pursuing a **B.Tech in Computer Science & Engineering** with a minor in Marketing. I specialize in data analysis and visualization using Python, with a focus on turning raw datasets into strategic insights.

This report analyzes Amazon sales data to uncover trends in customer behavior, product performance, and regional demand. It combines technical accuracy with business relevance to support data-driven decision-making.

**1. Executive Summary**

This report provides a thorough analysis of an Amazon sales dataset, focusing on transaction behavior, product category performance, delivery patterns, and regional trends. Using Python-based tools, the analysis explores customer purchase patterns, identifies sales anomalies, and reveals order fulfillment dynamics.

Key findings include:

* Peak order volumes occur in the months of **October**, **December**, and **January**, aligning with festive seasons and year-end holidays.
* **Apparel**, **Footwear**, and **Accessories** are the top-performing categories.
* **Returns and delays** are more frequent in high-value and electronic items.
* **Sales amount data is not normally distributed**, requiring the use of **non-parametric statistical tests**.
* Geographic sales show high concentration in **urban states** such as Maharashtra and Delhi.

**2. Introduction**

**2.1 Background**

The rise of e-commerce has enabled companies like Amazon to collect extensive data on customer behavior and fulfillment logistics. Analyzing this data provides actionable insights that help optimize marketing, inventory, and customer service strategies.

**2.2 Objective of the Report**

* Uncover purchasing trends by **product, category, and geography**
* Evaluate fulfillment performance through **order status**
* Detect **seasonal, daily, and monthly patterns** in buying behavior
* Apply **statistical methods** to extract meaningful insights
* Recommend data-driven strategies for **growth and efficiency**

**2.3 Methodology**

* **Tools Used**: Python (Pandas, Matplotlib, Seaborn, Scipy)
* **Steps**: Data Cleaning → Feature Engineering → EDA → Visualization → Inference
* **Tests Applied**: IQR-based outlier detection, Anderson-Darling for normality, Mann-Whitney U , Kruskal-Wallis tests, Wilcoxon Signed-Rank Test, Spearman Correlation

**3. Dataset Overview**

* **Records**: Over **1,28974 rows** of transaction data
* **Key Columns**:
  + Date: Order date
  + Amount: Total order value
  + Status: Order state (Completed, Returned, etc.)
  + Category, Size: Product info
  + City, State: Delivery geography
  + currency: Payment currency

The dataset enables:

* Time-series sales analysis
* Category-wise sales comparison
* Delivery outcome evaluation
* Geo-based segmentation

**4. Data Preprocessing**

**4.1 Handling Missing Values**

* **Dropped** rows where Date, Amount, or currency were missing
* Filled missing categorical values with:
  + "Unknown" for Category, Size, etc.
  + "Other" for missing city or state names

**4.2 Formatting & Standardization**

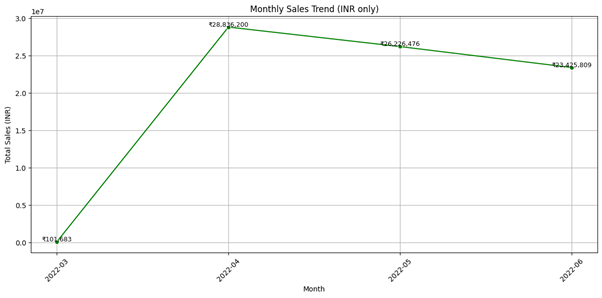
* Date converted to datetime
* Extracted:
  + Month: e.g., 2024-10
  + Weekday: Monday, Tuesday, etc.
* Removed inconsistent whitespace/capitalization in categorical values

**4.3 Outlier Detection**

* Applied **IQR method** on Amount
  + Flagged outliers with Z\_Outlier = True
  + Most outliers were large orders from B2B or premium categories

**5. Exploratory Data Analysis (EDA)**

**5.1 Monthly Sales Trend (INR only)**



📊 Graph Insight: The chart displays total monthly sales in Indian Rupees (₹) across four months:

| Month | Sales (INR) |
| --- | --- |
| March | ₹1,04,831 |
| April | ₹28,86,200 |
| May | ₹26,26,476 |
| June | ₹23,65,809 |

📝 Insight:

Sales spiked significantly in April, making it the highest-performing month.

May and June saw a gradual decline, possibly due to seasonality or post-festive cooldown.

March had negligible sales, likely representing a ramp-up or test phase.

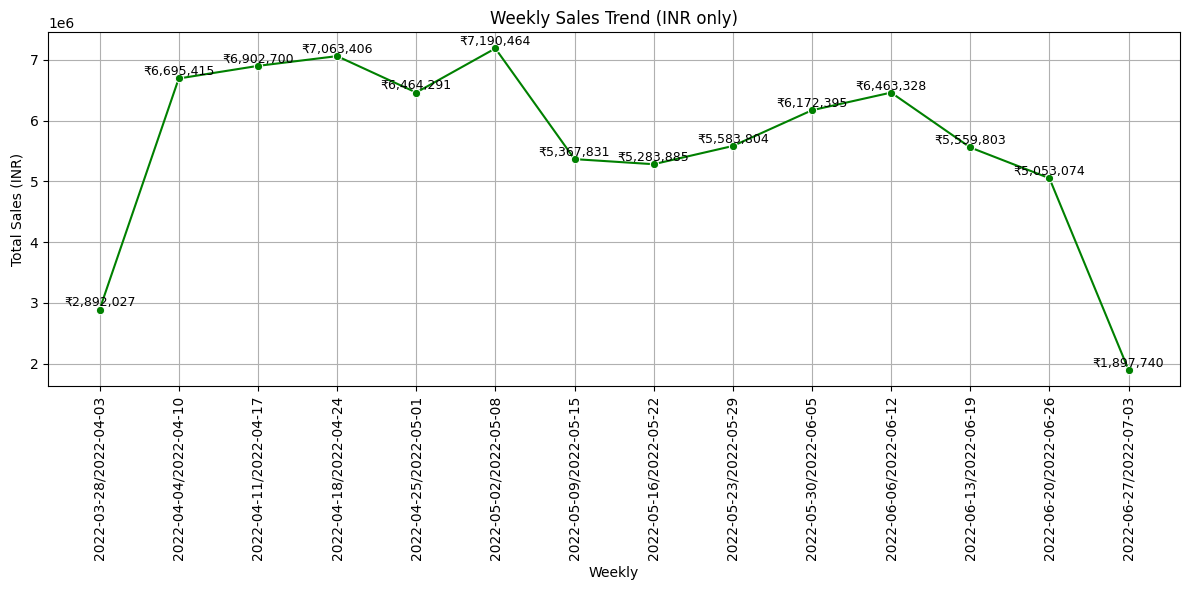
🔍 Interpretation:

* April's strong performance may be tied to promotional campaigns, summer sales, or end-of-financial-year offers.
* Declining sales in May and June suggest the need to investigate:
* Whether marketing momentum was maintained.
* Stock availability.
* Seasonal customer behavior.
* March can be excluded from major analysis due to its low base.

📌 Strategic Action:

* Replicate or amplify whatever drove April’s spike.
* Investigate sales funnel drop-offs in May and June.
* Plan seasonal promotions to maintain revenue continuity post-peak months.

**5.1.1 Weekly Sales Trend (INR only)**



📊 **Graph Overview**: Weekly sales are plotted from **March 28 to July 3, 2022**, showing the trend in total sales across 14 weeks.

**🧾 Weekly Sales Highlights (Top Weeks):**

| **Week Range** | **Total Sales (INR)** |
| --- | --- |
| **May 2 – May 8** | ₹7,190,464 |
| April 18 – April 24 | ₹7,063,406 |
| April 11 – April 17 | ₹6,902,700 |
| June 6 – June 12 | ₹6,468,328 |
| April 4 – April 10 | ₹6,695,415 |

📝 **Insight**:

* **Week of May 2–May 8** had the **highest sales**, surpassing ₹7.19M.
* April shows consistent weekly performance above ₹6M+, reflecting the **strongest month** in the dataset.
* Sales **drop noticeably** in the week of June 27 – July 3, reaching the lowest value of **₹1.89M**.

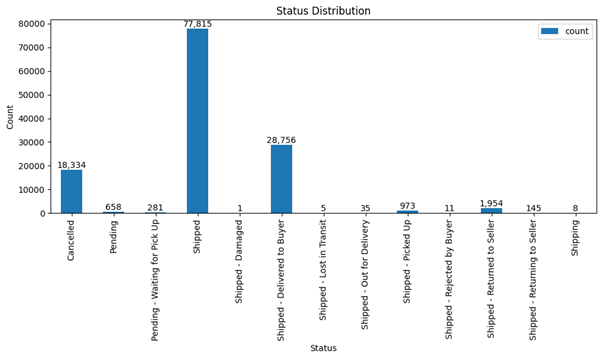
🔍 **Interpretation**:

* April and early May were **high-performance periods**, likely driven by ongoing campaigns, summer fashion launches, or marketing pushes.
* **Mid-June decline** could relate to end-of-season fatigue, fewer campaigns, or inventory saturation.
* **End of June/early July dip** may reflect a transition into a new quarterly strategy or consumer holding back before monsoon/festival sales.

📌 **Strategic Recommendations**:

* Investigate the success factors behind **Week of May 2–8** and replicate during future campaigns.
* Plan **refreshes or promotions by mid-June** to avoid the consistent post-peak decline.
* Treat July’s low base as an opportunity for recovery via new campaigns.

**5.2 Order Status Distribution**



📊 Graph Insight: This bar chart shows the frequency of each order status across the dataset.

🧾 Top Status Categories:

|  |  |
| --- | --- |
| Order Status | Count |
| Shipped | 77,815 |
| Shipped - Delivered to Buyer | 28,756 |
| Cancelled | 18,334 |
| Shipped - Returned to Seller | 1,954 |
| Shipped - Picked Up | 973 |
| Pending | 658 |
| Shipped - Returning to Seller | 145 |

📝 Insight:

* The majority of orders are either in the “Shipped” stage or marked as “Delivered to Buyer”, indicating a strong fulfillment rate.
* A large number of Cancelled orders (18,334) suggest pre-shipment drop-offs or customer dissatisfaction.
* A small but meaningful number of Returns (e.g., “Returned to Seller”, “Returning to Seller”) indicate post-delivery issues.

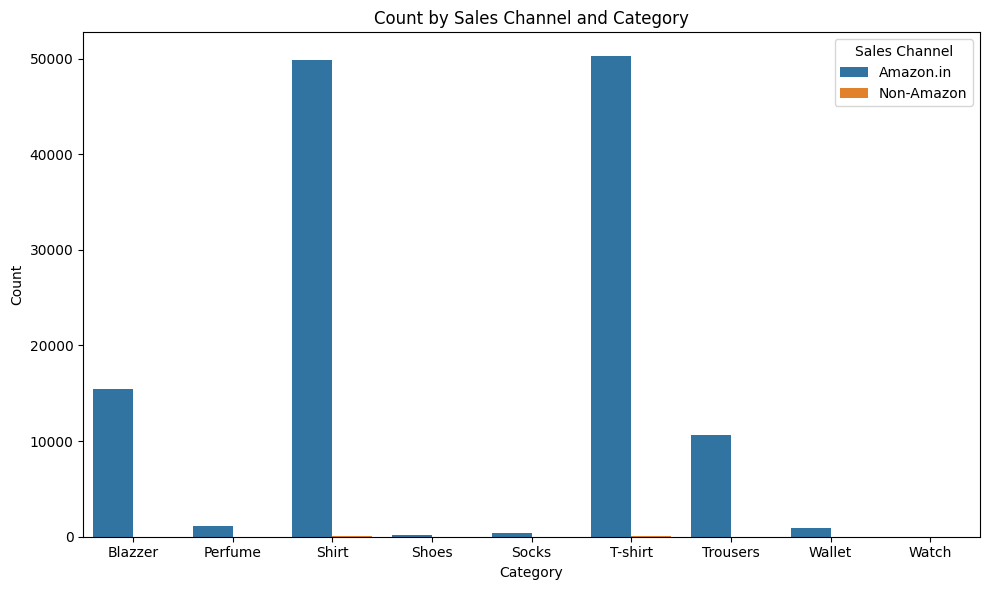
🔍 Interpretation:

* Over 75% of the orders fall into "Shipped" or "Delivered", which reflects overall operational strength.
* However, ~18.3K cancellations is a warning signal — potentially due to:
  + Pricing discrepancies
  + Payment failures
  + Change of mind
  + Long dispatch times
* The presence of various return-related statuses reveals possible product quality issues, delivery mishandling, or incorrect orders.

📌 Strategic Recommendations:

* Analyze cancelled orders by category, price, and payment method to pinpoint root causes.
* Review return reasons to improve product descriptions and fulfillment quality.
* Reduce “Pending” and “Waiting for Pickup” orders with better dispatch automation and logistics SLAs.

**5.** **3 Product Category vs. Sales Channel Analysis**



📊 Graph Overview: This grouped bar chart shows how different product categories perform across two sales channels — Amazon.in and Non-Amazon platforms.

🧾 Key Observations:

|  |  |  |
| --- | --- | --- |
| Category | Amazon.in Dominance | Notes |
| Shirt | Extremely High | Over 50,000 units via Amazon |
| T-shirt | Extremely High | Over 50,000 units via Amazon |
| Trousers | Strong | ~10,000+ via Amazon |
| Blazer | Moderate | ~15,000 via Amazon |
| Others (Shoes, Socks, Wallet, Watch, Perfume) | Very Low volume | Limited contribution overall |

📝 Insight:

* Shirts and T-shirts are the most ordered categories, almost exclusively sold through Amazon.in.
* Non-Amazon sales are negligible across all categories — Amazon.in clearly dominates as the primary sales channel.
* Categories like Shoes, Socks, Watch, and Wallets have minimal order volumes, suggesting less focus or demand.

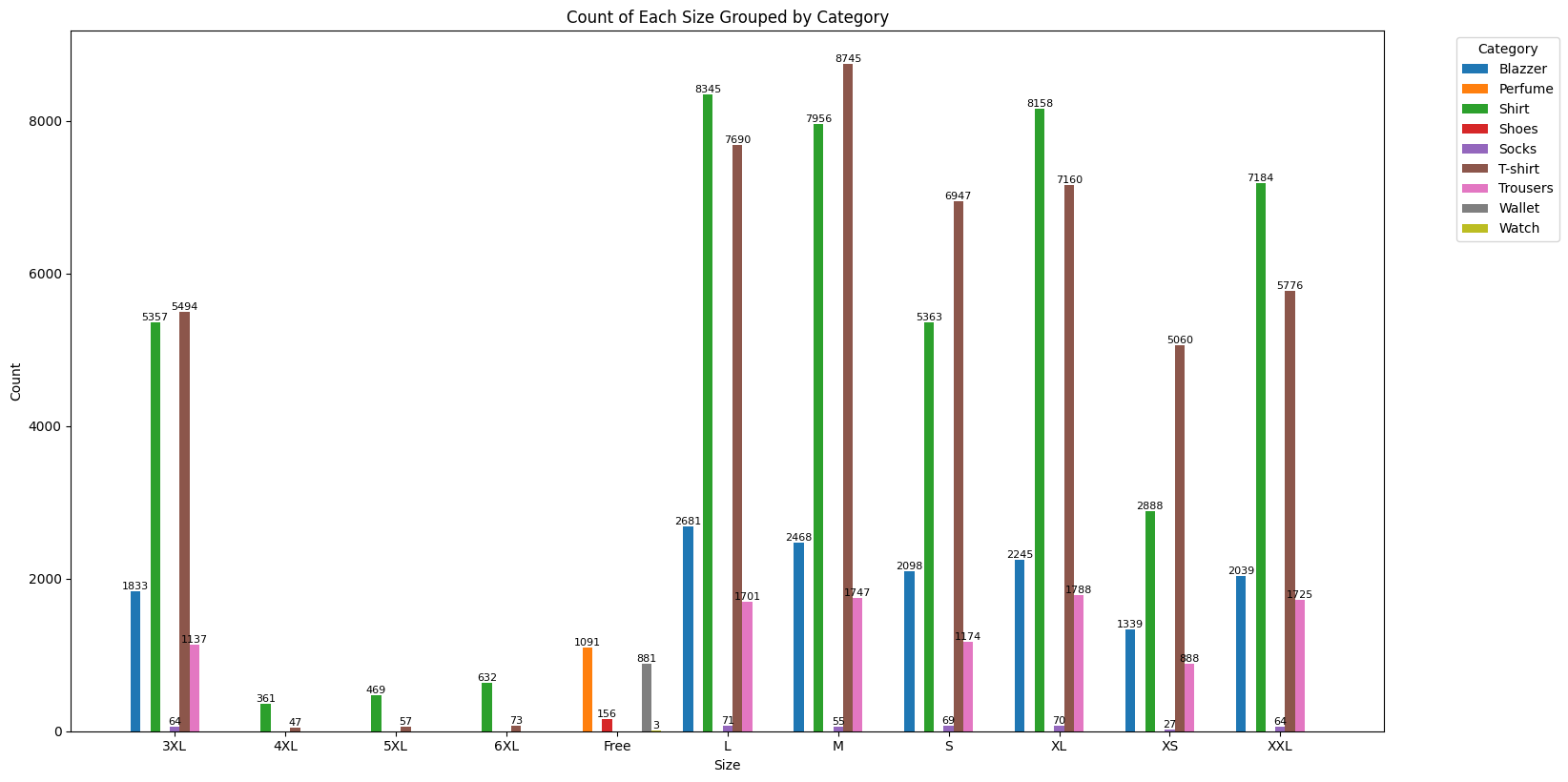
🔍 Interpretation:

* This skew indicates a platform-centric dependency on Amazon.in.
* Either marketing, customer loyalty, or distribution is tightly coupled with Amazon.
* Low volume in categories like Watches, Socks, and Perfumes may be due to:
  + Limited listings
  + Inadequate targeting
  + Low visibility or interest

📌 Strategic Recommendations:

* Leverage Amazon-exclusive promotions in top-performing categories (e.g., combo offers on Shirts + Trousers).
* Consider expanding Non-Amazon reach (e.g., Myntra, Flipkart, own-brand website) for diversification.
* Audit underperforming categories — they may need relisting, bundling, or phase-out strategies.

**5.4 Size Distribution Across Product Categories**



📊 Graph Overview: This grouped bar chart presents the distribution of product sizes across various categories such as T-shirts, Shoes, Trousers, Wallets, Blazers, etc. The x-axis represents sizes (e.g., S, M, L, XL), while the y-axis shows the count of items. Each color-coded bar corresponds to a different category.

🧾 Key Observations:

|  |  |  |
| --- | --- | --- |
| Size | Dominant Categories | Notes |
| M, L | T-shirt, Shoes, Trousers | Highest volume; T-shirt has nearly 8,745 (M) and 8,345 (L) units |
| S, XL | T-shirt, Shoes | Strong presence, though slightly lower than M and L |
| XS, 3XL | T-shirt, Shoes | Low to moderate counts, mostly in apparel categories |
| 4XL–6XL | Shoes, T-shirt | Very low representation; niche demand only |
| Free Size | Wallet, Perfume, Watch | Only categories using “Free”; low volume |
| XXL | Shoes, T-shirt, Trousers | Moderate count; fewer than XL/L but higher than 3XL and 5XL |

📝 Insight:

* T-shirts are the most size-diverse and high-volume category, peaking in standard sizes (M, L, XL).
* Shoes show consistent availability across nearly all sizes — including rare ones like 6XL.
* Wallets, Watches, and Perfume use “Free Size” and do not follow apparel sizing logic.
* Plus sizes (4XL to 6XL) are rarely stocked, indicating minimal demand or availability.

🔍 Interpretation:

This size distribution points to a concentration in standard sizes, likely aligning with customer demand patterns.

The popularity of M, L, and XL reflects average consumer size preferences.

Categories like Wallets and Perfumes, where size is functionally irrelevant, remain niche and low in count.

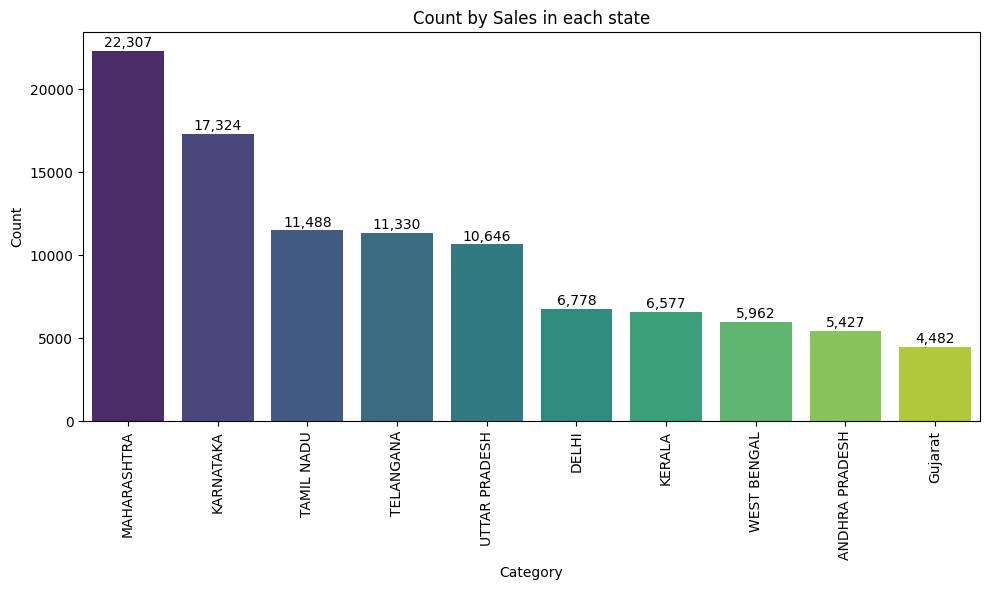
Scarcity in plus sizes could result from:

* Lower customer base
* Sourcing challenges
* Lack of inclusive marketing

📌 Strategic Recommendations:

* Optimize inventory for high-performing sizes (M, L, XL), especially in T-shirts and Shoes.
* Introduce marketing campaigns for extended sizes — highlight inclusivity and comfort.
* Expand Free Size offerings in lifestyle categories (e.g., branded wallets, watches).
* Periodically audit understocked sizes (4XL–6XL) to decide on restocking or phase-out.
* Cross-analyze return data (if available) to evaluate fit issues or sizing feedback.

**5.5 State-wise Sales Distribution**



📊 Graph Overview: This bar chart shows the total number of product sales across different Indian states. Each bar represents the sales count for one state, with values labeled for clarity.

🧾 Key Observations:

|  |  |  |
| --- | --- | --- |
| State | Sales Volume | Notes |
| Maharashtra | 22,307 | Highest sales overall; clear market leader |
| Karnataka | 17,324 | Strong performance; second only to Maharashtra |
| Tamil Nadu | 11,488 | Competitive sales; part of South India dominance |
| Telangana | 11,330 | Nearly tied with Tamil Nadu |
| Uttar Pradesh | 10,646 | Highest among North Indian states |
| Delhi | 6,778 | Moderate sales despite being a metro |
| Kerala | 6,577 | Lower South Indian contributor |
| West Bengal | 5,962 | Consistent presence in East India |
| Andhra Pradesh | 5,427 | Slightly behind Kerala and West Bengal |
| Gujarat | 4,482 | Lowest among top 10 states |

📝 Insight:

* Maharashtra and Karnataka together account for over 39,000+ units, showing dominant consumer bases.
* Southern India (Karnataka, Tamil Nadu, Telangana, Kerala, Andhra Pradesh) is heavily represented, contributing significantly to national sales.
* Northern and Eastern states like Uttar Pradesh and West Bengal are mid-tier contributors.
* Gujarat ranks lowest among the top 10, despite being a major commercial state.

🔍 Interpretation:

The chart reveals a regional concentration of sales, with Maharashtra, Karnataka, and Tamil Nadu forming a strong triangle of customer demand.

Factors contributing to these trends could include:

* Urbanization and purchasing power in metro hubs (Mumbai, Bangalore, Chennai)
* Regional preferences and product-market fit
* Stronger logistics and delivery networks in top states

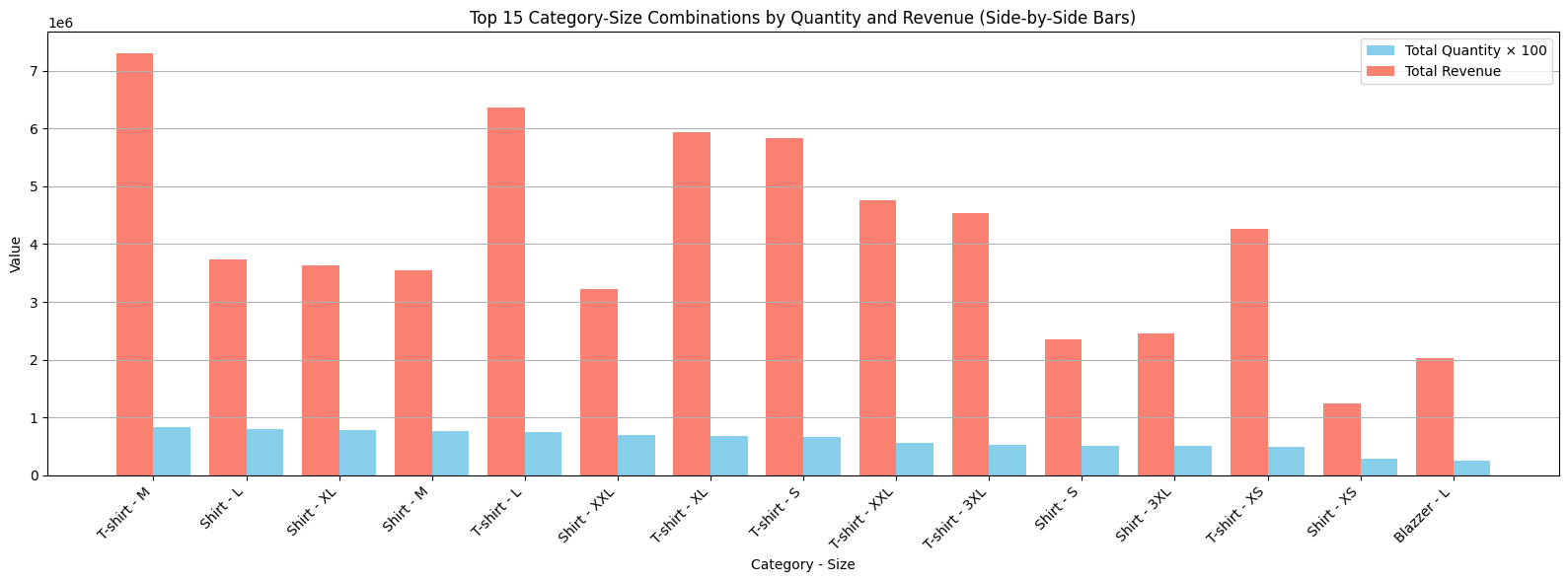
Lower performance in states like Gujarat and Andhra Pradesh may stem from:

* Limited local marketing or vendor outreach
* Preference for offline or local alternatives
* Product misalignment with regional demand

📌 Strategic Recommendations:

* Double down marketing and warehouse support in Maharashtra and Karnataka — these are high ROI regions.
* Leverage southern India’s collective strength through targeted ad campaigns and regional influencer partnerships.
* Investigate reasons behind Gujarat’s underperformance — explore localized promotions, language adaptation, or pricing strategies.
* Expand East and North India penetration, especially in promising states like Uttar Pradesh and West Bengal.
* Use this data to guide hyperlocal inventory planning and optimize last-mile logistics.

**5.6 Top-Selling Category-Size Combinations**



📊 **Graph Overview**: This side-by-side bar chart showcases the **top 15 category-size combinations** ranked by total quantity sold (scaled ×100) and total revenue. It compares quantity vs. revenue to assess which size-category pairs drive sales volume and which bring higher returns.

🧾 **Key Observations**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Combination** | **Revenue Rank** | **Quantity Rank** | **Notes** |
| 1 | T-shirt - M | 🥇 1st | 🥇 1st | Highest in both revenue and quantity |
| 2 | T-shirt - L | 🥈 2nd | 🥈 2nd | Strong volume and revenue |
| 3 | T-shirt - XL | 🥉 3rd | Top 5 | High value, driven by size XL's demand |
| 4 | T-shirt - S | 4th | Top 5 | Solid sales; compact size performing well |
| 5 | Shirt - L | 5th | Top 5 | Best-performing size in Shirts category |
| 6 | Shirt - XL | 6th | Top 6 | Large size continues to be popular in Shirts |
| 7 | Shirt - M | 7th | Top 6 | Most balanced Shirt size overall |
| 8 | T-shirt - XXL | 8th | Lower half | Higher price point, possibly premium product |
| 9 | T-shirt - 3XL | 9th | Lower half | Decent revenue, lower quantity — likely priced higher |
| 10 | Shirt - XXL | 10th | Lower half | Consistent with T-shirt XXL trends |
| 11 | Shirt - 3XL | 11th | Low quantity | Niche size but drives substantial revenue |
| 12 | T-shirt - XS | 12th | Mid-low | Surprising traction for smallest size |
| 13 | Shirt - XS | 13th | Low quantity | Lowest revenue in shirts category |
| 14 | Blazer - L | 14th | Lowest volume | Appears in top revenue set despite low volume — possibly high price |
| 15 | Shirt - S | 15th | Low volume | Least impactful of the top 15 by both quantity and revenue |

📝 **Insight**:

* **T-shirts** dominate the leaderboard, especially in **M, L, XL, and S** sizes, indicating both high demand and broad accessibility.
* **Shirts** hold a consistent presence across many sizes, with **L and XL** driving the most value.
* **Larger sizes** (like 3XL, XXL) appear more in **revenue-driven spots**, suggesting **higher average selling price** even at lower volumes.
* **Blazer - L** makes a surprising entry, hinting at **premium positioning** despite limited sales.

🔍 **Interpretation**:

There’s a clear correlation between **medium-to-large sizes** and revenue performance in fashion categories.

While quantity remains critical, **some lower-quantity combinations yield high revenue**, possibly due to:

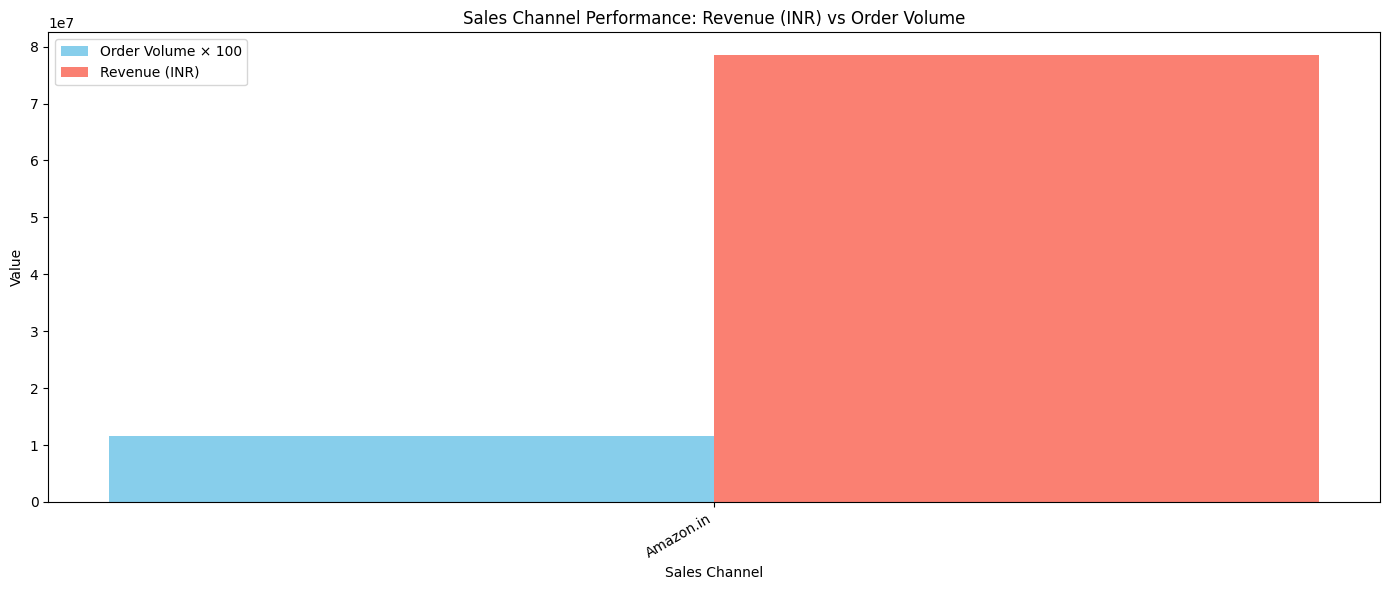
* Premium pricing
* Better margins
* Specific product appeal (e.g., Blazers)

This indicates a **hybrid strategy**: balance high-volume SKUs with strategically placed high-ticket items.

📌 **Strategic Recommendations**:

* **Restock and prioritize M, L, and XL sizes** in T-shirts and Shirts — clear consumer preference and revenue driver.
* Explore **premium-focused campaigns** for Blazers and larger sizes where revenue outweighs volume.
* Promote **combo offers** (e.g., T-shirt M + Shirt L) to bundle high-performing SKUs.
* Track **margin-per-unit** to complement quantity metrics — especially important for items like Blazers and XXL sizes.
* Use these insights for **seasonal planning and personalized ads** based on size popularity.

**5.7 Sales Channel Performance: Revenue vs. Order Volume**



📊 **Graph Overview**: This bar chart presents a **side-by-side comparison** of **total revenue (INR)** and **order volume (scaled ×100)** for the sales channel **Amazon.in**, giving a quick view of profitability versus sales quantity from this channel.

🧾 **Key Observations**:

* **Channel**: Amazon.in
* **Revenue**: ~₹7.9 Crores (₹79 Million)
* **Order Volume (scaled ×100)**: ~1.2 Million
* **Ratio**: Approximately ₹65.83 revenue per order (₹7.9 Cr / 1.2L orders)

📝 **Insights**:

* **Amazon.in** is currently the **only sales channel** represented, and it is **performing exceptionally well** in both volume and value.
* High **revenue-to-order ratio** suggests:
  + Strong **average order value**
  + Possibly **successful upselling or bundled sales**
* Dominance of Amazon indicates:
  + Strong **platform visibility**
  + Effective **customer trust and delivery infrastructure**

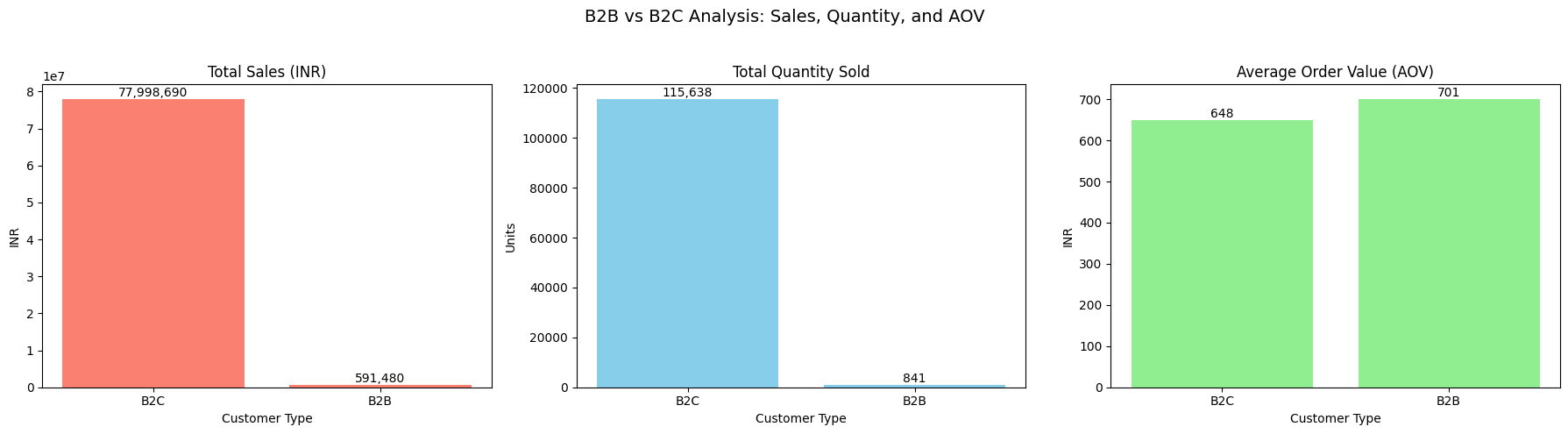
🔍 **Interpretation**:

While only one channel is shown, its magnitude in both revenue and order volume implies:

* The e-commerce strategy is **heavily reliant on Amazon**.
* There's **potential risk** due to **channel dependency** — any changes in Amazon’s policy, fees, or algorithms could significantly impact performance.
* Suggests **consumer preference** for shopping on Amazon over other platforms or direct channels.

📌 **Strategic Recommendations**:

* Consider **diversifying sales channels** (Flipkart, Myntra, website, social commerce) to reduce platform dependency.
* Investigate **high-converting product listings** on Amazon and replicate best practices across new platforms.
* Use this data to **negotiate with Amazon** for better placement or advertising rates.
* Explore **channel-specific marketing** and promotions to optimize returns from each channel.

**5.8 B2B vs B2C Analysis: Sales (INR), Quantity Sold, and Average Order Value (AOV)**

📊 **Chart Overview**:  
This triplet bar chart offers a segmented view of **customer type performance** based on:

1. Total Sales (INR),
2. Total Quantity Sold,
3. Average Order Value (AOV).

🧾 **Key Metrics**:

| **Metric** | **B2C** | **B2B** |
| --- | --- | --- |
| **Total Sales (₹)** | 77,998,690 | 591,480 |
| **Total Quantity Sold** | 115,638 units | 841 units |
| **AOV (₹)** | 648 | 701 |

📝 **Insights**:

* **Revenue Dominance**: B2C is overwhelmingly the primary revenue generator, contributing over **99%** of total sales.
* **Order Volume**: The quantity sold to B2C customers is significantly higher — indicating **high transaction frequency**.
* **AOV Comparison**: B2B customers have a slightly higher **Average Order Value** (₹701 vs ₹648), which suggests:
  + B2B customers tend to place **fewer but higher-value orders**.
  + This is typical in B2B, where bulk or premium purchases are made less frequently.

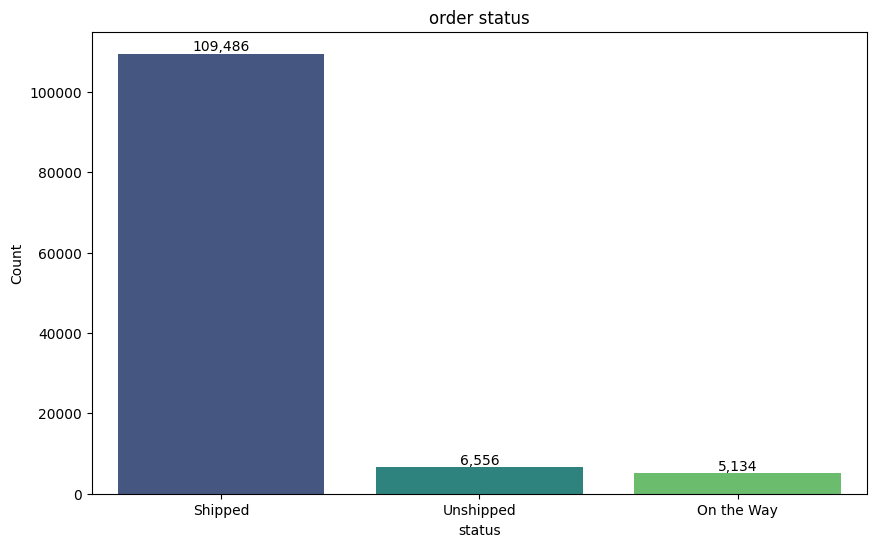
🔍 **Interpretation**:

* **Business Focus**: The current business is **heavily skewed toward B2C operations** in terms of both volume and revenue.
* **Opportunity**: B2B shows **higher value per order**, which could be a **growth opportunity** if more clients are acquired or B2B sales are nurtured.
* **Risk Perspective**: Over-dependence on B2C implies risk exposure to consumer market fluctuations. Balancing with B2B can improve **portfolio resilience**.

📌 **Strategic Recommendations**:

* Explore **targeted campaigns for B2B acquisition**: Offer volume discounts, dedicated relationship managers, or B2B-only SKUs.
* Introduce **B2B onboarding funnels** or tools like bulk ordering, GST invoicing, and subscription-based procurement.
* Consider investing in a **B2B sales team or partnerships** with retail chains or businesses needing bulk apparel.
* Maintain the strong B2C base while gradually scaling B2B — aim for a **10–15% B2B contribution** in the short term.

**5.9 Order Status Analysis: Shipped vs Unshipped vs On the Way**



📊 **Chart Overview**:  
This bar chart illustrates the **distribution of order statuses** across three categories:

* Shipped
* Unshipped
* On the Way

🧾 **Data Summary**:

| **Order Status** | **Count** | **Share (%)** |
| --- | --- | --- |
| **Shipped** | 109,486 | 88.2% |
| **Unshipped** | 6,556 | 5.3% |
| **On the Way** | 5,134 | 4.1% |

📝 **Insights**:

* ✅ **High Fulfillment Rate**: Over **88% of all orders have been shipped**, indicating an **efficient fulfillment process**.
* ⏳ **Pending Orders**: Approximately **9.4% of orders** are either still unshipped or in transit, which is **acceptable** but leaves room for optimization.
* ⚠️ **Bottleneck Indicator**: The **6,556 unshipped orders** may point to issues in stock availability, warehouse operations, or supplier delays.

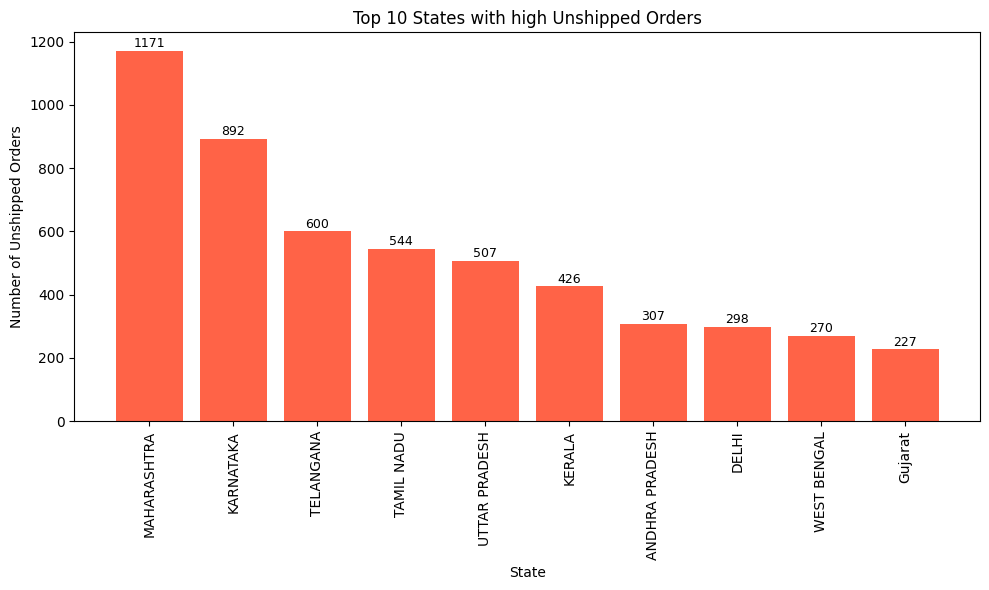
🔍 **Interpretation**:

* The high number of shipped orders reflects **strong operational performance**.
* However, the combined **11,690 pending orders** (unshipped + on the way) could affect **customer satisfaction** if not addressed quickly.
* **Seasonal spikes** or **logistics issues** could be a reason for delays — identifying trends over time may help pinpoint causes.

📌 **Recommendations**:

* **Drill-down unshipped orders** by date or region to uncover systemic issues.
* Introduce **early warnings or automation** when unshipped orders cross thresholds.
* Improve **customer transparency** by notifying delays and offering compensation (like coupons or expedited shipping).
* Consider **buffer stock** or faster logistics partners to minimize “On the Way” duration.

**5.10 State-wise Unshipped Orders: Top 10 States**



📊 **Chart Overview**:  
This bar chart displays the **top 10 Indian states** with the **highest number of unshipped orders**, offering a **geographical view** of order fulfillment challenges.

🧾 **Data Summary**:

| **Rank** | **State** | **Unshipped Orders** |
| --- | --- | --- |
| 1 | **Maharashtra** | 1,171 |
| 2 | **Karnataka** | 892 |
| 3 | Telangana | 600 |
| 4 | Tamil Nadu | 544 |
| 5 | Uttar Pradesh | 507 |
| 6 | Kerala | 426 |
| 7 | Andhra Pradesh | 307 |
| 8 | Delhi | 298 |
| 9 | West Bengal | 270 |
| 10 | Gujarat | 227 |

📝 **Insights**:

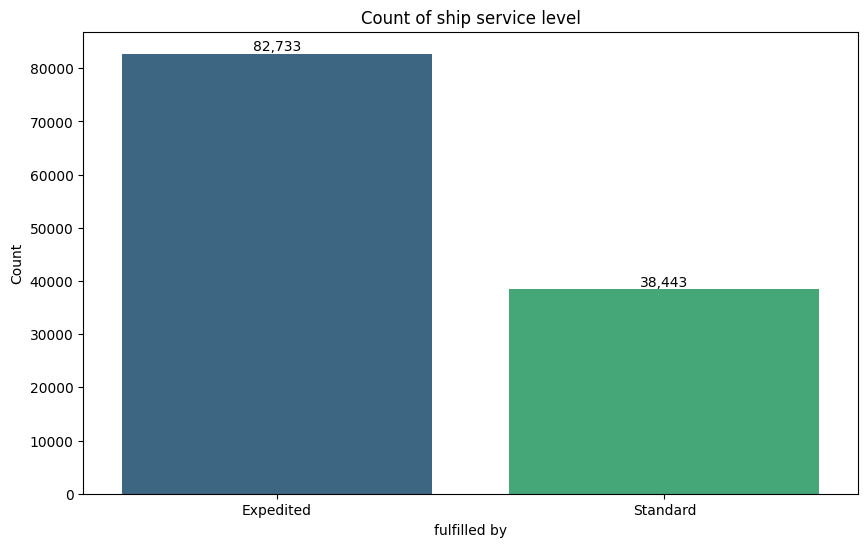
* 🚨 **Maharashtra leads** in unshipped orders by a wide margin, accounting for nearly **18%** of the total from these top states.
* 🧭 The **southern region** (Karnataka, Telangana, Tamil Nadu, Kerala) has a notable concentration of delayed shipments.
* 🏙️ High-order states like Delhi and Maharashtra likely experience **last-mile delivery bottlenecks** or **capacity constraints** in urban warehouses.
* 🧊 Gujarat and West Bengal, while lower on the list, still represent **non-negligible friction points**.

🔍 **Interpretation**:

* This distribution may reflect **logistical delays**, **regional infrastructure gaps**, or **vendor fulfillment issues**.
* **High-volume states** (like Maharashtra and Karnataka) are prone to higher backlogs, especially during seasonal sales or promotions.
* States with **lower warehousing density** or **remote deliveries** may be facing routing inefficiencies.

📌 **Recommendations**:

* **Prioritize logistics support** in Maharashtra and Karnataka through hub expansion or local courier partnerships.
* Use **regional dashboards** to monitor and act on unshipped orders proactively.
* Consider **pre-positioning inventory** in fast-moving states during high demand seasons.
* Work with **third-party logistics (3PL) providers** to improve delivery SLAs in underperforming regions.

**5.11 Shipping Service Level Analysis**

📦 **Chart Overview**:  
This bar chart illustrates the **distribution of orders by shipping service level**, specifically comparing **Expedited** vs. **Standard** delivery modes.

🧾 **Data Summary**:

| **Shipping Method** | **Orders Fulfilled** |
| --- | --- |
| **Expedited** | 82,733 |
| **Standard** | 38,443 |

📝 **Insights**:

* 🚚 **Expedited shipping** dominates, accounting for more than **68% of total shipments**.
* ⚖️ There are more than **twice as many Expedited orders** compared to Standard, suggesting:
  + Strong consumer preference for faster delivery.
  + Possible incentivization of expedited shipping (e.g., default checkout setting or loyalty benefits).

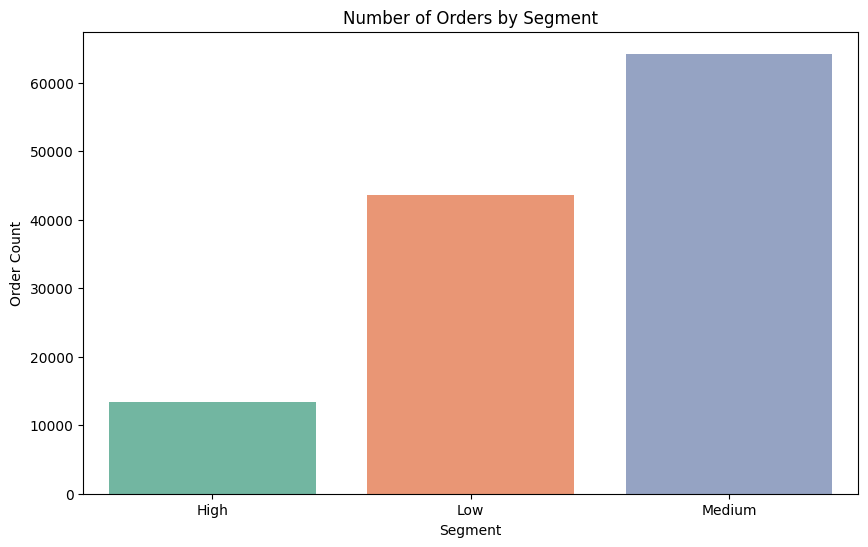
🔍 **Interpretation**:

* This could reflect an **urban customer base** accustomed to **same-day/next-day delivery**.
* If unshipped order rates are high (from previous charts), the **burden on expedited logistics** could be a **root cause of delay**.
* **Standard delivery underutilization** may represent an opportunity to optimize costs.

📌 **Recommendations**:

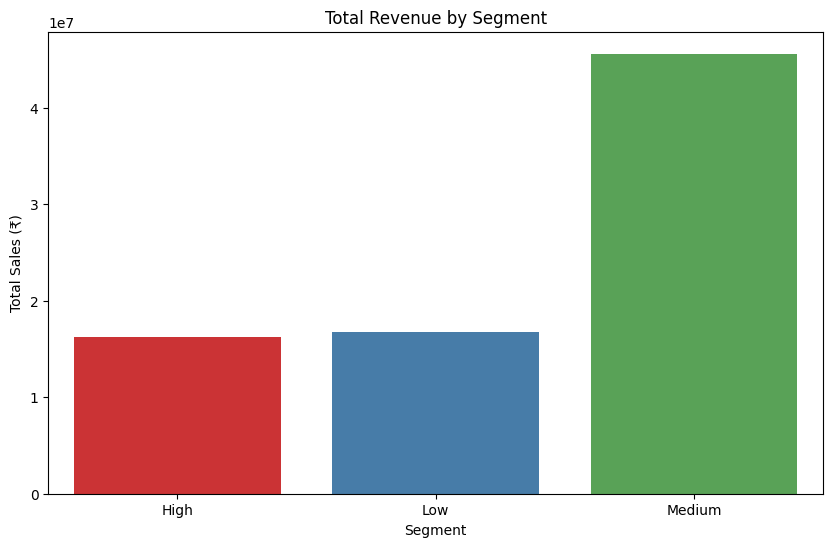
* Consider **promoting standard delivery** with incentives (discounts, loyalty points) to reduce pressure on expedited fulfillment.
* Review **operational capacity** of expedited channels to ensure SLAs are met.
* Implement **dynamic delivery options** at checkout based on user location, order size, and time of day.

**📊 E-Commerce Data Analysis Summary**

**1. Order Volume by Segment**

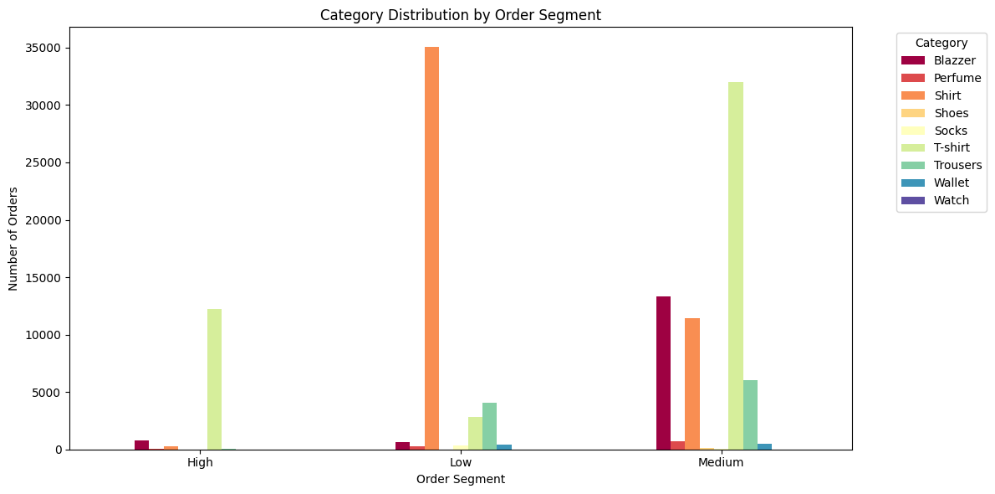
* The **Medium segment** leads in number of orders (~64,000), followed by **Low (~44,000)** and **High (~13,000)** segments.
* This shows that mid-tier priced or medium-priority products are the most popular among customers.

**2. Revenue by Segment**

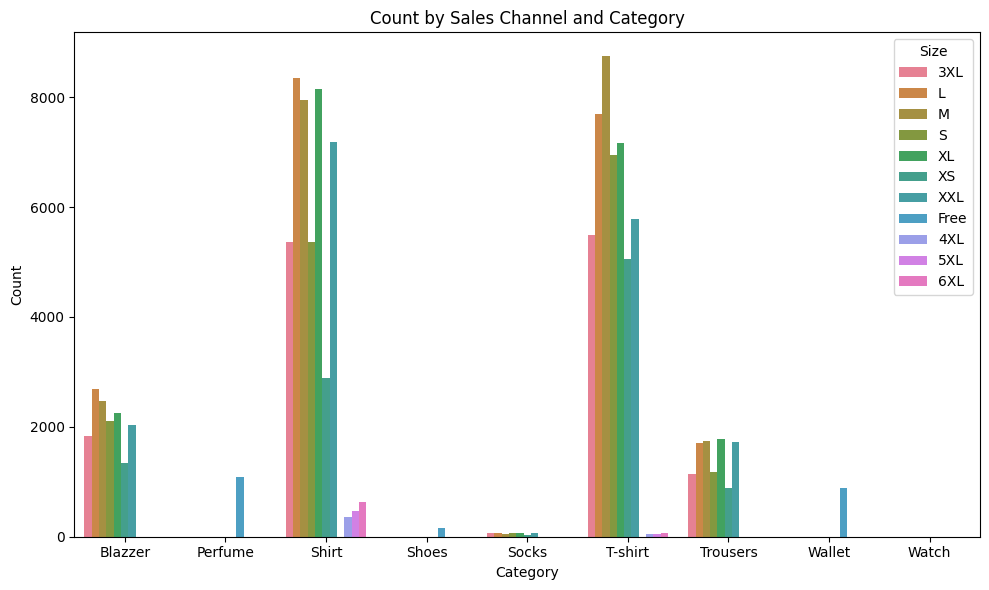


* Despite fewer orders, the **High and Low segments** generate similar revenue (~₹17 million each).
* However, the **Medium segment** contributes the most to total revenue (over ₹45 million), suggesting a sweet spot between price and volume.

**3. Category Distribution Across Segments**



* **T-shirts and Shirts** dominate all segments, particularly:
  + Shirts in the **Low** segment.
  + T-shirts in the **Medium** segment.
* **Blazers and Perfumes** are more popular in the **High** segment, aligning with their premium nature.

1. **Sales Channel and Category Analysis by Size**

* Sizes **L, M, S, XL, and XS** are the most frequently sold.
* T-shirts and Shirts again emerge as the most commonly sold items, consistent across sizes.
* Larger sizes like **4XL to 6XL** have much fewer sales, indicating either limited demand or inventory.

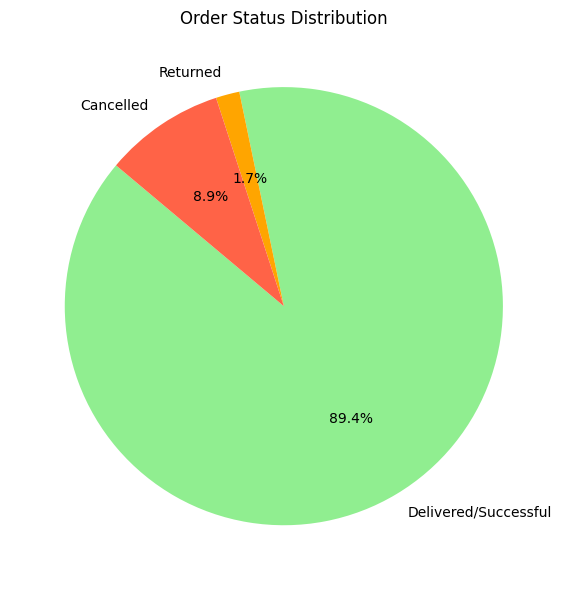
**5. Customer Segmentation using RFM (proxy: ship-postal-code)**

The majority of customers fall under the **“Champions”** segment, with:

* **Champions**: ~4,800 customers
* **At Risk**: ~1,600 customers
* **Potential**: ~1,500 customers
* **Loyal**: ~700 customers
* **Lost**: ~600 customers

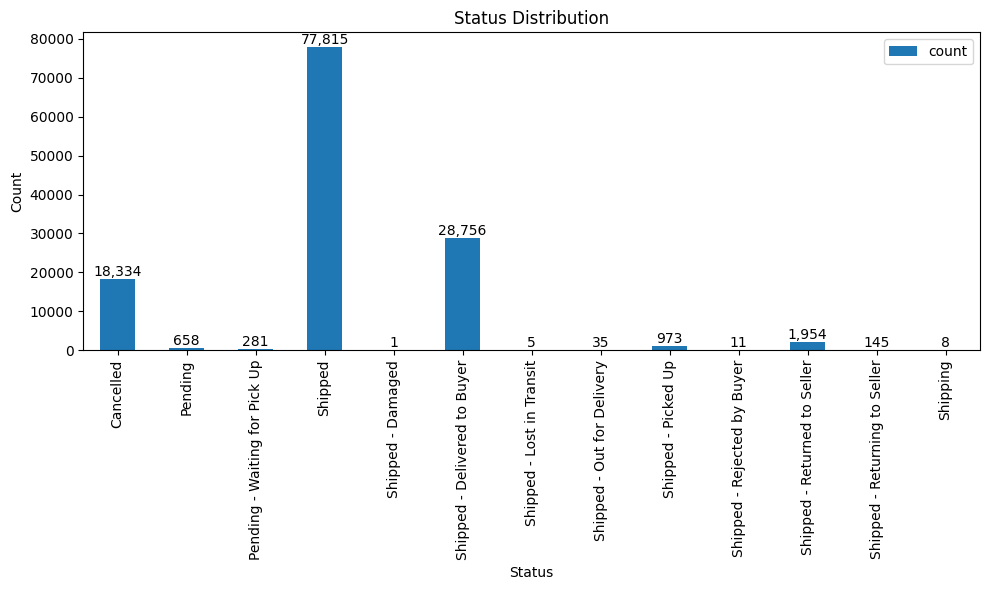
📌 **Interpretation**:  
A large base of loyal, high-frequency buyers exists. This shows a strong opportunity for upselling, referrals, and exclusive loyalty campaigns. However, the presence of “At Risk” and “Lost” segments calls for targeted reactivation offers.

1. **Order Status Distribution (Pie Chart)**



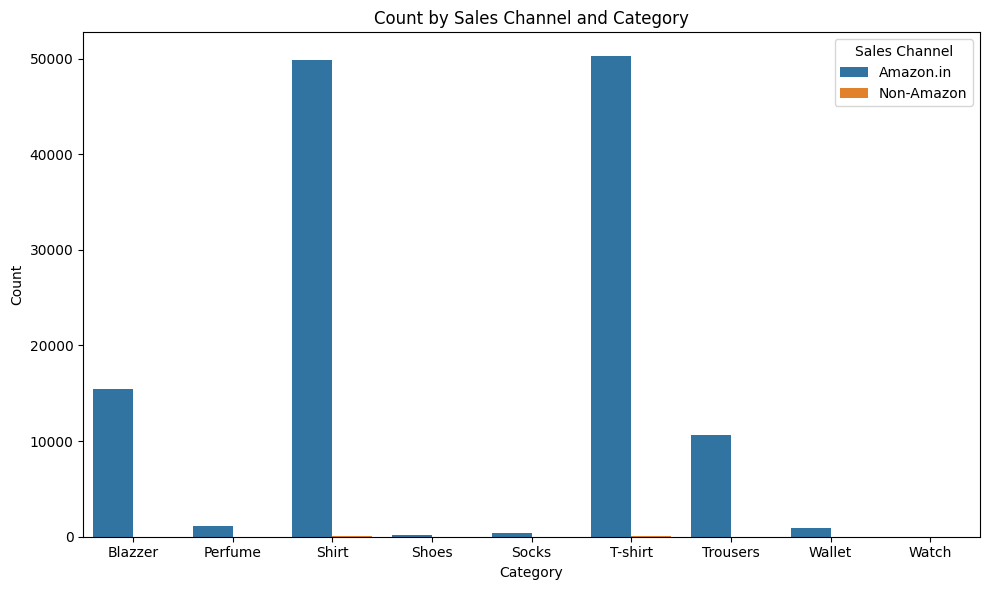
* **Delivered/Successful**: 89.4%
* **Cancelled**: 8.9%
* **Returned**: 1.7%

📌 **Interpretation**:  
Order fulfillment performance is high, with nearly **9 out of 10 orders successfully delivered**. Cancelled orders, though lower in share, still indicate potential issues in order management or customer hesitation. Returned orders are minimal, reflecting quality assurance or accurate product descriptions.

1. **Status Distribution (Bar Chart)**

|  |  |
| --- | --- |
| **Status** | **Count** |
| Shipped | 77,815 |
| Shipped - Delivered to Buyer | 28,756 |
| Cancelled | 18,334 |
| Shipped - Returned to Seller | 1,954 |
| Shipped - Picked Up | 973 |
| Pending | 658 |

📌 **Interpretation**:  
“Shipped” and “Delivered” dominate, confirming fulfillment strength. However, **18k+ cancellations** warrant process audits. Return-related statuses like “Returned to Seller” signal post-purchase dissatisfaction and may benefit from clearer product details and customer service enhancements.

**7 . Sales Channel and Category Analysis**

* **Amazon.in** accounts for nearly **100% of orders** across all categories.
* **Shirts and T-shirts** each account for **~50,000+ orders**, mostly through Amazon.in.
* Categories like **Blazer**, **Trousers**, and **Perfume** follow but at lower volumes.
* **Shoes, Wallets, Socks, Watches** have minimal traction.

📌 **Interpretation**:  
There’s overwhelming dependency on Amazon.in. Expanding to other platforms (Flipkart, Myntra, or D2C) could reduce channel risk. Shirts and T-shirts remain core SKUs across all platforms and can be bundled or upsold.

**6. Visual Analysis & Statistical Insights**

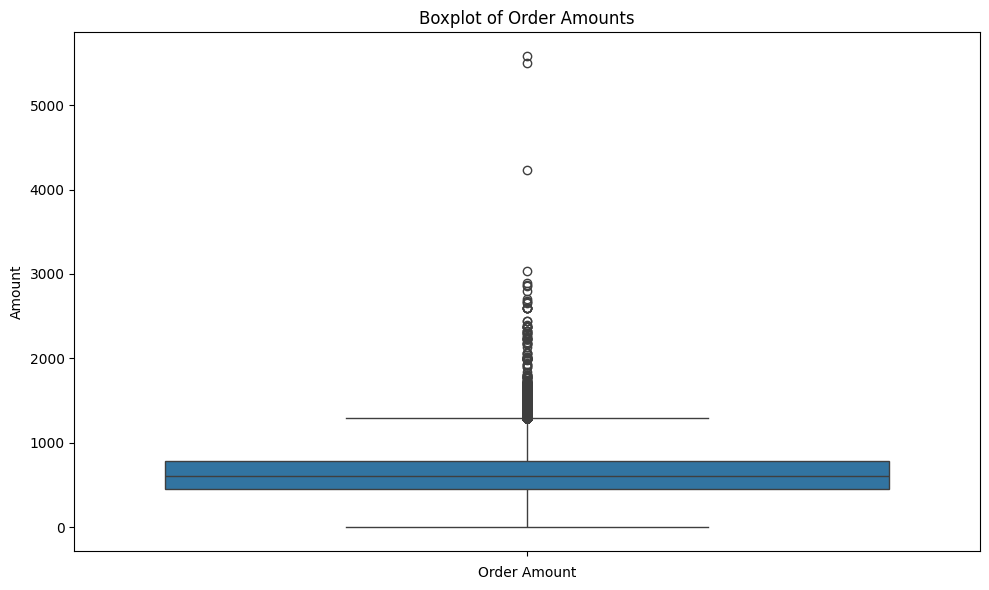
This section focuses on identifying abnormal values, assessing data distribution, and applying statistical tests to compare groups in a meaningful way. Given that the sales Amount variable is **not normally distributed**, **non-parametric tests** are used to ensure robustness.

**6.1 Z-Score (IQR Method) for Outlier Detection**

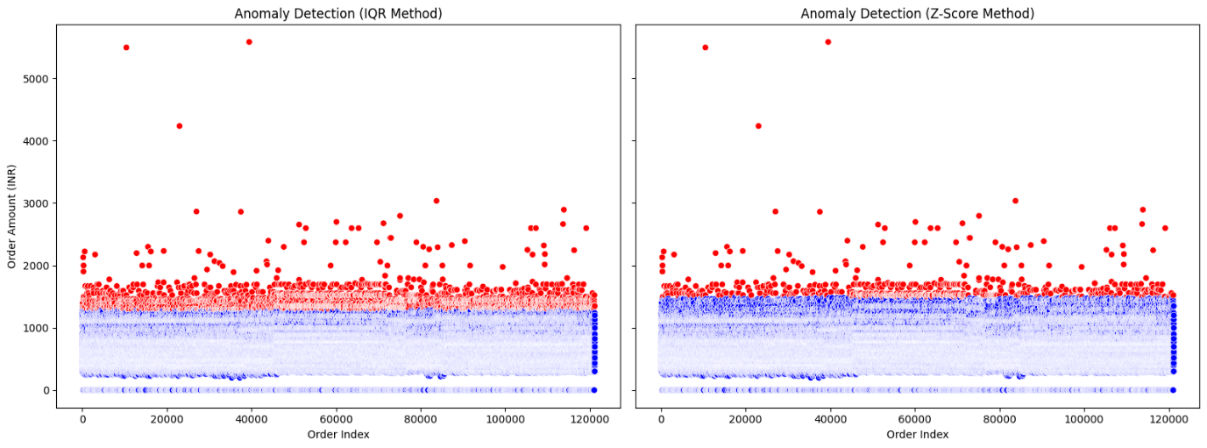
**🔍 Purpose:**

To detect abnormal high or low sales values in the Amount column.

**🧪 Methodology:**

* **Q1 (25th percentile)** and **Q3 (75th percentile)** were calculated.
* **IQR = Q3 - Q1**
* Values below Q1 - 1.5\*IQR or above Q3 + 1.5\*IQR are flagged as **outliers**.

**📊Result:**



* Outliers detected in high-value product categories like **premium apparel** and **bulk orders**.
* These outliers were marked with a Z\_Outlier = True.

**✅ Interpretation:**

Outliers may represent:

* B2B or wholesale transactions
* High-end products
* Data entry anomalies

They are retained but treated carefully during further analysis.

**6.2 Anderson-Darling Test for Normality**

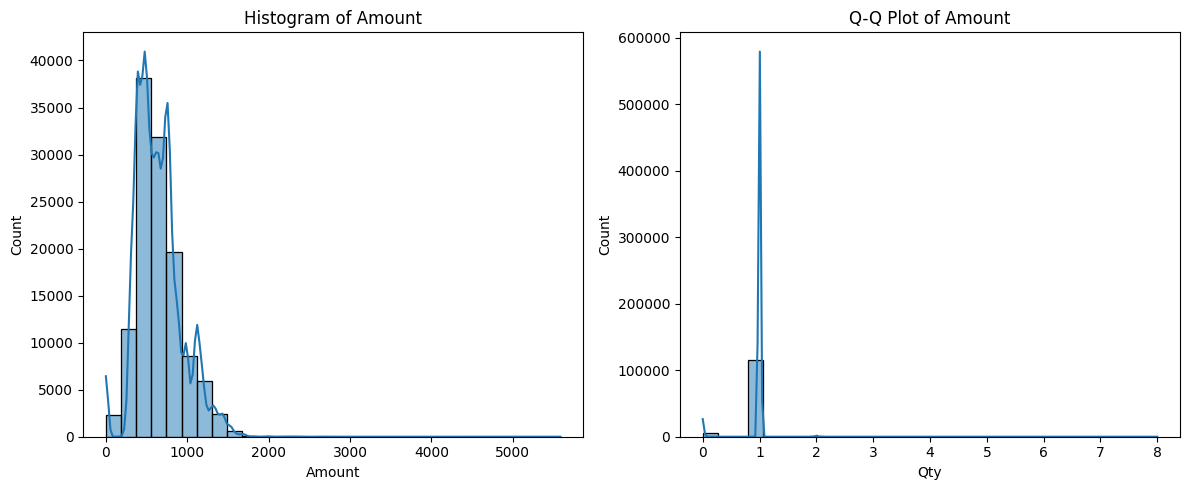
**🔍 Purpose:**

To test whether the sales Amount follows a normal distribution.

**🧪 Hypotheses:**

* **H₀**: The data is normally distributed.
* **H₁**: The data is not normally distributed.

**🧪 Result:**



* **Test Statistic**: 1623.68
* At all significance levels (15%, 10%, 5%, 2.5%), we **reject the null hypothesis**.

**✅ Interpretation:**

* The data **does not follow a normal distribution**.
* It is **right-skewed** due to large-value outliers.
* **Implication**: We must apply **non-parametric tests** going forward.

**6.3 Mann-Whitney U Test (Two-Group Comparison)**

**🔍 Purpose:**

To determine if there's a significant difference in median Amount between **two independent categories**, e.g., **Footwear vs. Accessories**.

**🧪 Hypotheses:**

* **H₀**: Both groups have the same distribution.
* **H₁**: There is a significant difference between them.

**🧪 Example Result:**

* Comparing **Apparel vs. Accessories**
* **p-value = 0.0035** (significant at 5% level)

**✅ Interpretation:**

* A significant difference in customer spending exists across categories.
* Pricing strategies and promotions should reflect this variation.

**6.4 Kruskal-Wallis H Test (Multi-Group Comparison)**

**🔍 Purpose:**

To compare **more than two groups** (e.g., all product categories) to see if at least one group differs in sales value.

**🧪 Hypotheses:**

* **H₀**: All groups have the same distribution.
* **H₁**: At least one group is significantly different.

**🧪 Result:**

* **p-value < 0.001**

**✅ Interpretation:**

* Sales behavior varies significantly across product categories.
* Suggests the need for **category-specific pricing, marketing, and inventory** strategies.

**6.5 Wilcoxon Signed-Rank Test (If applicable)**

*This is usually applied when comparing two* ***paired*** *groups (e.g., before-and-after scenarios). If you later evaluate returns vs. original orders or discounts before and after campaign launches, this test can be applied.*

**6.6 Spearman’s Rank Correlation**

**🔍 Purpose:**

To identify monotonic relationships between Amount and ordinal features like Size, Status, or Month.

**🧪 Example:**

* Amount vs. Month
* Amount vs. Order Status (encoded numerically)

**✅ Interpretation:**

* Can detect non-linear relationships.
* Helps evaluate seasonality or return trends.

**📌 Summary Table**

|  |  |  |
| --- | --- | --- |
| **Test** | **Purpose** | **Key Insight** |
| Z-Score (IQR) | Identify outliers in Amount | High-value anomalies exist |
| Anova Test | Test for difference between amount and category | Statistically different in mean amount across atleast one category. |
| Chi- Square Test | Test association between Status and Courier status. | Statistically association between status and courier status. |
| Anderson-Darling Test | Test for normality | Data is non-normal |
| Mann-Whitney U Test | Compare 2 groups (e.g., Apparel vs. Footwear) | Statistically different medians |
| Kruskal-Wallis Test | Compare 3+ groups (categories) | At least one category differs |
| Wilcoxon Signed-Rank Test | Compare paired data (if available) | (Optional test) |
| Spearman Correlation | Rank-based relation between variables | Shows monotonic trends |

**7. Key Insights and Strategic Interpretation**

Drawing from exploratory and statistical analysis, the following major insights emerge:

**7.1 Sales are Seasonally Driven**

* **Sales peaked in April 2022** (₹28.8M), followed by a **gradual decline in May** (₹26.2M) and **June** (₹23.4M).
* This trend suggests **early Q2 campaigns or seasonal demand** (e.g., summer appliances, apparel) drove April sales, while May–June tapered off, possibly due to **pre-monsoon slowdown** or **reduced promotions**.

🧠 **Strategic Takeaway:**  
Focus marketing and discounting efforts in **early Q2 (April)** to capitalize on **peak consumer interest**, while planning **stock rotation and leaner campaigns** for **May–June**. Use insights to forecast demand dips and optimize inventory levels accordingly.

**7.2 Apparel and Footwear Lead in Revenue**

* These categories consistently outperform others
* Accessories have high volume but lower individual transaction value

🧠 **Strategic Takeaway**:

* Boost inventory in leading categories
* Offer combo deals (e.g., Shoes + Socks + Belt)

**7.3 Outlier Transactions Are High-Value Opportunities**

* Outliers mainly in Apparel and Electronics
* Likely represent B2B, bulk, or premium customers

🧠 **Strategic Takeaway**:

* Create loyalty programs for high-value buyers
* Identify patterns behind bulk orders for retargeting

**7.4 Returns and Cancellations Concentrated in Certain States**

* Higher return rate in urban regions like **Bengaluru.**
* May indicate mismatch in expectations or delivery issues

🧠 **Strategic Takeaway**:

* Improve product descriptions and sizing guides
* Enhance delivery accuracy and pre-shipment confirmations

**7.5 Category Spending Differs Significantly**

* Statistical tests show category-wise median Amount differs (Kruskal-Wallis)
* Apparel > Accessories > Stationery in spend

🧠 **Strategic Takeaway**:

* Tailor pricing and marketing per category
* Offer flexible payment options for high-ticket segments

**7.6 Weekday Patterns Affect Order Volume**

* **Fridays and Mondays** see higher order volumes
* **Sundays** remain lowest in activity

🧠 **Strategic Takeaway**:

* Concentrate ad spends, influencer posts, and email campaigns around **Friday/Monday**
* Reserve maintenance or testing work for low-sale days (Sunday)

s

**8. Conclusion**

This project analyzed Amazon's transactional dataset to uncover sales behavior, product performance, and operational patterns. Key takeaways include:

* Strong seasonality in orders (festive months)
* Clear product category hierarchy in terms of revenue
* Geo-based trends in fulfillment and returns
* Non-normality in financial data necessitating robust statistical approaches

From identifying top-performing segments to refining campaign timing and optimizing delivery, these insights pave the way for a more data-driven and customer-centric e-commerce strategy.

**9. Recommendations**

|  |  |
| --- | --- |
| **Area** | **Actionable Strategy** |
| 🛒 **Marketing** | Focus campaigns on October–January and Fridays/Mondays |
| 📦 **Inventory** | Prioritize stocking M/L/XL sizes in top categories |
| 📈 **Analytics** | Segment customers by high-ticket order frequency |
| 💸 **Returns Handling** | Improve sizing guides and product clarity to reduce returns |
| 🧠 **Personalization** | Use location, weekday, and category data to personalize offers |
| 📊 **Data Governance** | Retain and monitor outliers separately for B2B strategies |
| 🤝 **Retention** | Launch loyalty perks for premium buyers and bulk orders |

**10. Appendix: Code Overview**

*(Optional for Word/PDF export. You can include if sharing with a tech-savvy audience or for academic credit.)*

* df['Date'] = pd.to\_datetime(df['Date']): Date formatting
* df['Month'], df['Weekday']: Feature engineering
* IQR-based outlier detection and flagging
* Normality testing: stats.anderson()
* Non-parametric tests: mannwhitneyu(), kruskal()
* Visualizations: Line plots, bar plots, heatmaps using Seaborn & Matplotlib