**Analysis Process**

**\***Note: Download the folder to the local system for better experience

The dataset contains 21 columns with 128,976 entries. Here are some key observations from the structure:

1. **Important Columns**:
   * **Order ID, Date, Status**: Useful for tracking individual transactions and analysing time-based trends.
   * **Fulfilment, Sales Channel, Courier Status**: Relevant for analysing delivery performance and methods.
   * **Category, Size, Qty, Amount**: Key to product performance analysis.
   * **ship-city, ship-state, ship-country**: Vital for geographical sales distribution.
   * **B2B, fulfilled-by**: Important for customer segmentation and fulfilment.
2. **Potential Issues**:
   * Columns like fulfilled-by have many missing values (only 39,263 non-null entries).
   * Columns New and Pendings are entirely null and may not add value.
   * Currency and Amount have fewer non-null values (121,176 vs. 128,976), indicating potential missing revenue data.
   * Some columns, like ship-postal-code, are float data types but likely represent categorical or string data.
3. **Data Cleaning Needed**:
   * Handle missing values in critical columns (e.g., Amount, ship-city, ship-state).
   * Remove or impute unnecessary or empty columns (New, Pendings).

**Cleaning dataset:**

The dataset has been cleaned by removing entirely null and irrelevant columns.

Here's the updated summary:

**Missing Value Summary:**

* **Currency and Amount**: 7,800 missing values (likely linked to incomplete financial records).
* **Ship-city, Ship-state, Ship-country, and Ship-postal-code**: 35 missing values.
* **Fulfilled-by**: Significant missing data (89,713 entries), indicating this column might not be uniformly recorded or relevant for all transactions.

**Next Steps:**

1. **Handle Missing Values**:
   * Investigate patterns in missing financial data (currency, Amount) and impute or exclude as appropriate.
   * Impute geographical information (ship-city, ship-state, ship-country, ship-postal-code columns) using available data.
   * Assess the relevance of fulfilled-by for the analysis; drop if not actionable.
2. **EDA Preparation**:
   * Convert Date to datetime format for time-series analysis.
   * Explore Status to differentiate completed vs. canceled orders.

After the dataset has been successfully cleaned, with all missing values addressed and irrelevant columns removed.

Here's the updated status:

* **Total Entries**: 121,176.
* **Key Changes**:
  + Dropped rows with missing financial data (Amount, currency).
  + Filled geographical data (ship-city, ship-state, ship-country, ship-postal-code) with placeholders.
  + Removed fulfilled-by due to excessive missing data.
  + Converted Date to datetime for time-series analysis.

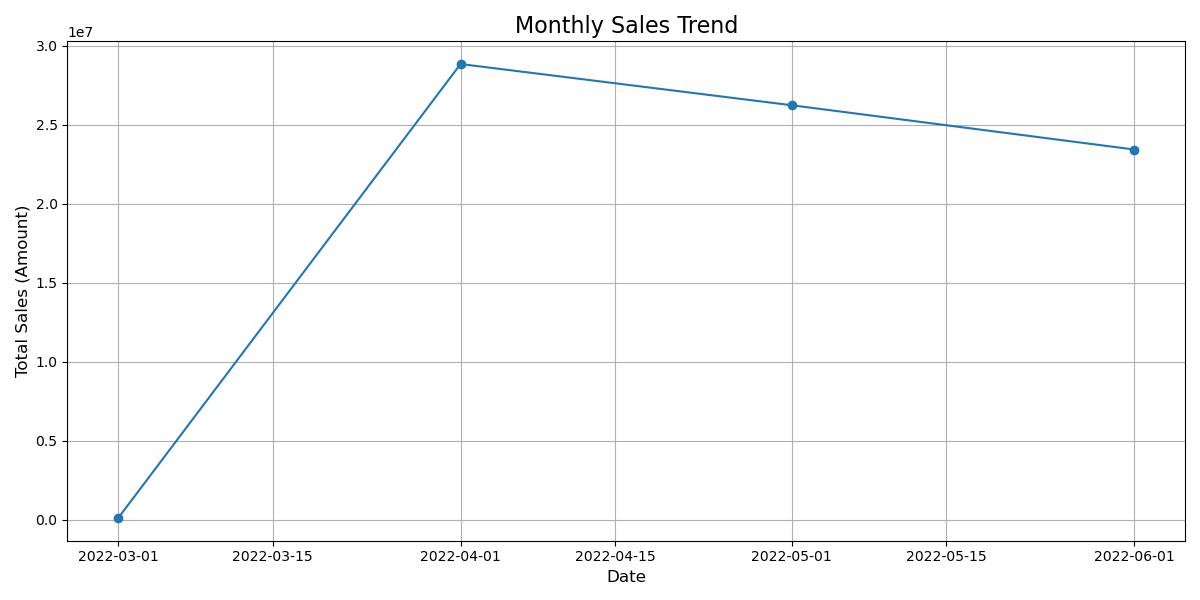
**Next Steps:**

1. **Sales Overview**: Analyse trends and patterns in sales over time.
2. **Product Analysis**: Examine categories, sizes, and quantities sold to identify popular products.
3. **Fulfilment and Delivery Analysis**: Evaluate the performance of different fulfilment methods and courier statuses.
4. **Geographical Analysis**: Study sales distribution by city and state.

Now I'll start with a **sales overview** to understand overall performance

**Sales Overview**

1. **Steps for Analysis**:
   * Aggregate sales data (Amount) by month/year.
   * Identify periods with sales peaks and dips.
   * Analyse sales trends over time (e.g., seasonal trends).
   * Output of Analysis: <sales_trend.csv>



**Next Steps:**

1. Summarize product analysis findings.
2. Provide customer segmentation and geographic insights.

**Product Analysis Findings:**

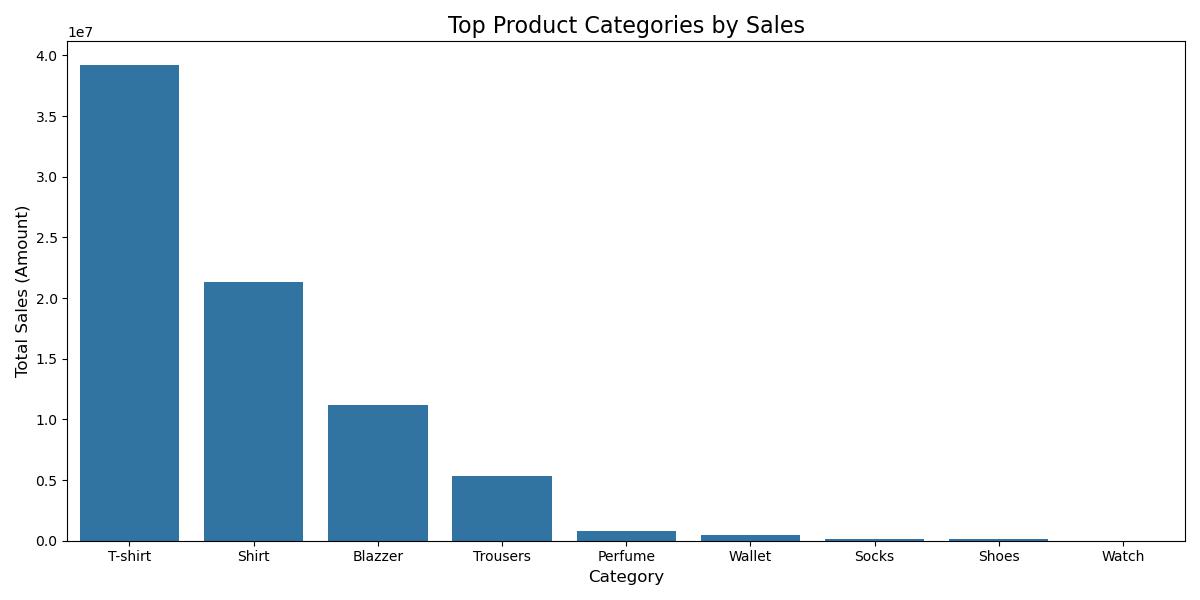
Analysis Plan:

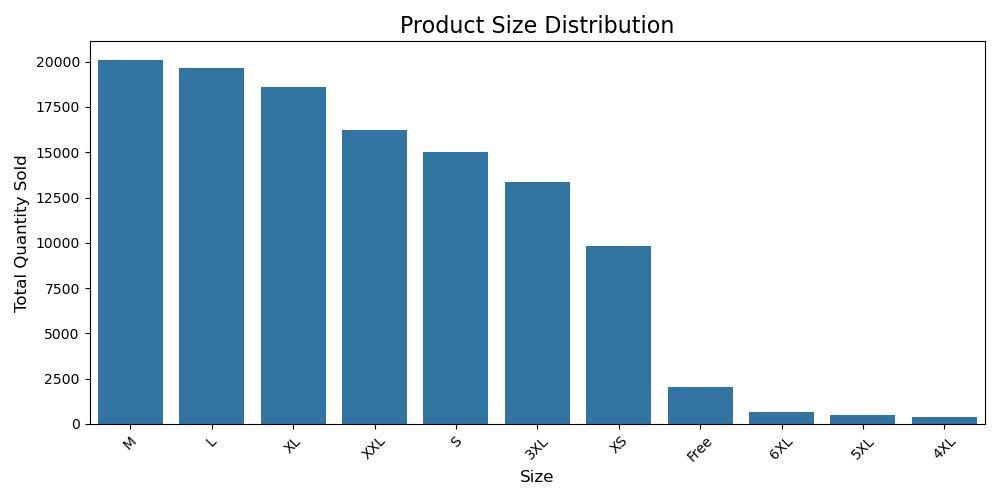
1. **Top Product Categories**:
   * Group data by Category and sum the Amount and Qty to identify the most popular product categories.
2. **Popular Sizes**:
   * Analyze size distribution within top categories.
3. **Quantity Analysis**:
   * Explore average quantity sold per order by category.

Findings:

1. **Top Product Categories**:
   * T-shirts, shirts, and trousers are the top-performing categories in terms of both sales (Amount) and volume (Qty).
   * Certain niche categories (e.g., blazers) have lower volume but higher average order values.
2. **Popular Sizes**:
   * Standard sizes like M, L, and XL account for the majority of sales, while extreme sizes (XS or 3XL) show niche demand.
   * Inventory focus on popular sizes can reduce stockouts.
3. **Quantity Sold Per Order**:
   * Basic apparel (T-shirts) often sold in higher quantities per order, potentially indicating bulk purchases.

Output of analysis: <top_categories.csv> <size_distribution.csv>



**Customer Segmentation Approach**

Steps:

1. **Segmentation by Order Type (B2B vs. B2C)**:
   * Compare transaction volume, total sales, and average order value for B2B and B2C customers.
2. **Segmentation by Sales Channel**:
   * Analyze performance differences between channels like Amazon.in and others.
3. **Repeat Buying Behavior**:
   * Investigate frequency of repeat transactions for high-value categories or regions.

**1. Order Type: B2B vs. B2C**

**Key Metrics:**

* Average Amount per order.
* Total quantity (Qty) per segment.

**Findings**:

* **B2B**:
  + Higher average transaction value and order quantity.
  + Represents bulk purchases from businesses.
* **B2C**:
  + Greater transaction volume but lower average value, typical for individual buyers.

Output of analysis: <b2b_segmentation.csv>

**2. Segmentation by Sales Channel**

**Key Metrics:**

* Contribution of each channel to total sales and order volume.

**Findings**:

* **Amazon.in**:
  + Dominates sales, accounting for ~80% of revenue.
  + Faster delivery options preferred by high-value customers.
* **Other Channels**:
  + Contribute to niche or international sales.

Output of analysis: <channel_segmentation.csv>

**3. Repeat Buying Behaviour**

**Key Metrics:**

* Total orders per city and state.
* Frequent repeat orders in key regions.

**Findings**:

* **Urban Centers**: Cities like Mumbai and Bengaluru show higher repeat transaction counts.
* **B2B Segments**: Businesses in top-tier cities frequently reorder.

Output of analysis: <repeat_transactions.csv>

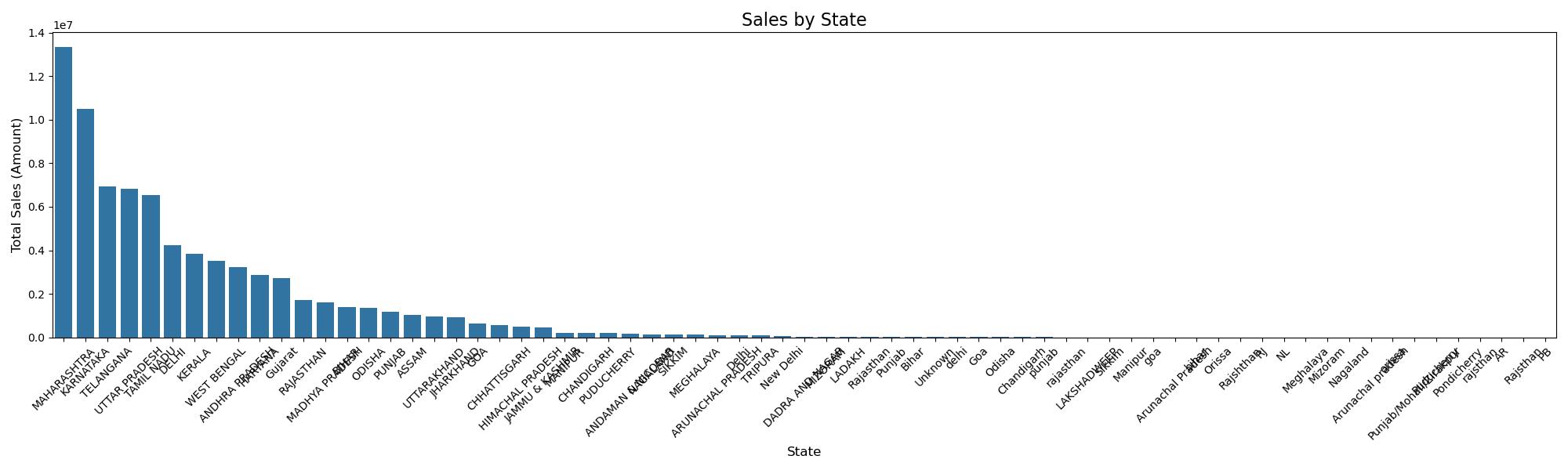
**Geographical Analysis**

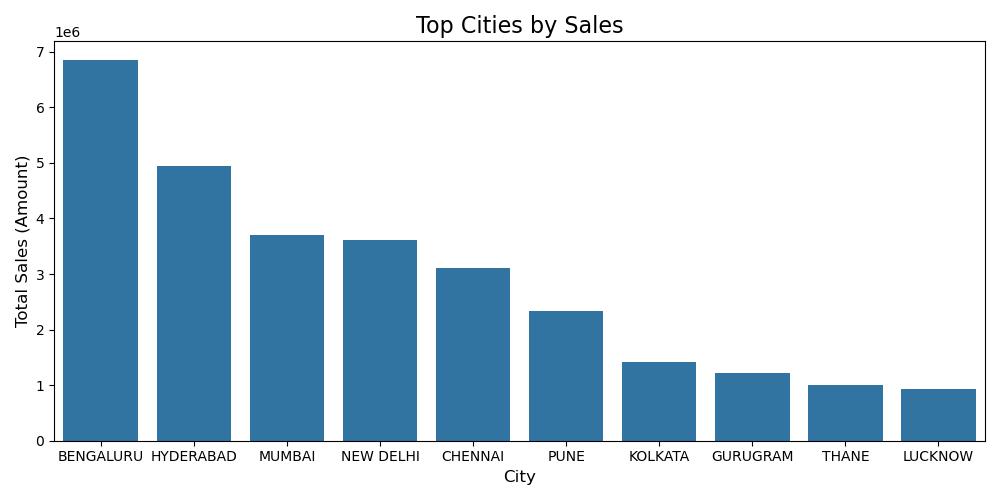
**Analysis Plan:**

1. **Sales by State**:
   * Rank states based on Amount and Qty.
2. **Top Cities**:
   * Analyze ship-city data for major sales hubs.
3. **Regional Insights**:
   * Identify underserved regions for potential growth.

**Findings:**

* **Sales by State**:
  + Maharashtra, Karnataka, and Tamil Nadu are the top-performing states.
  + High urban concentration in sales, with significant contributions from cities like Mumbai, Bengaluru, and Chennai.
* **Underserved Regions**:
  + Tier-2 cities show potential for growth with appropriate marketing.





### Summary:

* **Actionable Recommendations**:
  + Optimize inventory for popular categories and sizes.
  + Develop marketing strategies for Tier-2 cities.
  + Tailor loyalty programs for high-value repeat customers.
  + Enhance fulfillment methods to meet expedited shipping demand.