📄 Mini Report: Data Cleaning & Visualization Methodology

Project: Amazon Sales Analysis
Tools: Python (Pandas, NumPy), Power BI

---

1. Data Cleaning & Preparation (Python – Jupyter Notebook)

a. Importing & Setup

- Imported required libraries (pandas, numpy).
- Uploaded the raw CSV file (Amazon Sale Report.csv).
- Checked dataset dimensions, datatypes, and inspected sample rows for consistency.

b. Standardizing Column Names & Formats

- Converted all column names to lowercase.
- Replaced spaces and hyphens with underscores for easier handling in Python.
- Trimmed extra spaces from all text columns.

c. Date Standardization

- Converted the date column into a proper datetime format using a custom parser.
- Handled multiple formats and invalid entries with errors="coerce".
- Stored the cleaned field as order_date in DD-MM-YYYY format.
- Verified range, minimum, and maximum dates for sanity checks.

d. Location Cleaning

- Cleaned ship_city and ship_state:
    - Removed leading/trailing spaces.
    - Removed special characters/punctuation.
    - Converted into title case for uniformity (e.g., “maharashtra” → “Maharashtra”).

e. Numeric Data Cleaning

- Converted qty and amount columns to numeric datatypes.
- Handled invalid entries (NaN, text values) with coercion.
- Ensured quantities are integers, amounts are floats.

f. Column Reduction

- Dropped unnecessary fields such as:
    - Index columns.
    - Blank or irrelevant fields (new, pendings, currency, ship_postal_code, ship_country, fulfilled_by).

g. Order Status Normalization

- Converted status and courier_status columns to lowercase and stripped spaces.
- Created a **status mapping function** (map_status) to consolidate multiple raw statuses into **5 standardized categories**:
    - **Delivered** → Shipped & delivered to buyer.
    - **Cancelled** → Explicitly cancelled, or pending/shipped but courier status = cancelled.
    - **Returned** → Shipped but returned to seller/returning to seller.
    - **Problematic** → Shipped but damaged, lost in transit, or rejected by buyer.
    - **In Process** → Shipped/pending/shipping but not completed or cancelled.
- Applied this function row-wise to create a new column: status_cleaned.

h. Revenue Adjustments

- For orders marked as **Cancelled, Returned, Problematic** → set amount = 0.
- For **Delivered** orders: dropped invalid rows with missing or non-positive amount/qty.
- Ensured amount and qty were re-validated as numeric with defaults filled (0 if missing).

✅ **Output of Cleaning**

- Final cleaned dataset: fact_sales_cleaned.
- Key columns available for analysis:
    - **order_id, order_date, qty, amount, status_cleaned, courier_status, ship_city, ship_state, category, size, sales_channel, fulfilment, b2b flag**.

---

2. Visualization & Analysis (Power BI)

a. Importing Dataset

- Imported fact_sales_cleaned into Power BI.
- Created DAX measures for KPIs:
    - Total Revenue, Total Orders, Total Quantity, Delivered Revenue, Delivered Orders, Cancelled Orders, Average Order Value (AOV), Delivered AOV.

b. Graphs & Visuals Used

1. Sales Overview

- **KPI Cards**: Total Revenue, Total Orders, Total Quantity, AOV, Delivered Revenue, Delivered AOV, Cancelled Orders.

2. Monthly Trends

- **Line Chart**: Delivered vs Total Revenue by month.
- **Bar/Column Chart**: Order count by month, Revenue by Month

3. Fulfilment Performance

- **Pie Chart**: Orders by fulfilment (Amazon vs Merchant).
- **Pie Chart**: Revenue by fulfilment (Amazon vs Merchant).

4. Order Status Distribution

- **Pie Chart**: % of Delivered, Cancelled, In Process, Returned, Problematic orders.
- **Bar Chart**: In Process Distribution (Shipped, Picked Up, Pending, etc.).
- **Pie Chart**: Problematic order breakdown (Rejected, Lost in Transit, Damaged).

5. Sales Channels

- **Pie Chart**: Amazon.in vs Non-Amazon orders.

6. B2B vs B2C

- **Pie Chart**: Orders split (B2C vs B2B).
- **Pie Chart**: Revenue split (B2C vs B2B).

7. Product Analysis

- **Bar Chart**: Orders by product category (T-shirt, Shirt, Blazer, Trousers, Others).
- **Bar Chart**: Revenue by product category.
- **Bar Chart**: Orders by Product size (M, L, XL, XXL dominating).

8. Geographic Insights

- **Map Chart**: Orders by state.
- **Bar Chart**: Top 10 states by order count.
- **Bar Chart**: Top 10 states by revenue.

---

3. Deliverables

- **Cleaned Dataset** – fact_sales_cleaned with standardized formats and adjusted values.
- **Visualization Pack** – All charts in Power BI (as PNGs).
- **Final Report** – PDF document consolidating methodology, findings, and recommendations.