Project Documentation:

Walmart Sales Data EDA

---

1. Project Title

EDA (Exploratory Data Analysis) in Walmart Sales Data.

2. Project Objective

This project involves Exploratory Data Analysis (EDA) on Walmart sales data, based on a realistic case study. The primary objective is to uncover actionable business insights to improve profitability, customer targeting, and operational strategies.

3. Dataset Overview

The project used a realistic Walmart sales dataset. Before handling case study question, the first step was to understand the dataset's overall structure.

- Source: Simulated real-world Walmart dataset
- Columns: Order ID, Customer ID, Customer Name, Order Date, City, Region, Category, Quantity, Sales, Profit.
- Rows: 100

4. Data Preparation & Cleaning

- **Datetime Conversion: Order Date** was converted into datetime format for time-based analysis.
- **Feature Engineering:** Extracted new columns like **Month**, **Year**, **Day** from **Order Date** to support temporal analysis.
- **Missing Value Check:** No null values were present in the dataset.
- **Duplicates:** Checked and removed if any.
- **Data Types:** All columns were validated and corrected where necessary.

5. Case Study Question & Key Insights

1) Customer Segmentation Challenge

Identify the top 10% of customers who contributed the most to the total profit. What common characteristics (region, category, city) do they share?

Key Insights

- **Region**: Distribution is fairly even, but [East] has a slight edge.
- **Category**: [Furniture] appears more frequently.
- **City**: One or two cities like [South Megan] show up more than once, but no strong city dominance.

2) Monthly Sales Recovery Strategy

Determine which month in the past year had the lowest overall profit. What specific product category and region contributed most to this loss?

Key Insights

- Past year is **2024** – 492 records.
- **March** Month make the least amount of loss profit. Loss is **-252.22.**
- March month dissection Region wise **South** made a more amount of lose. Loss is **-486.86.**
- Category wise **Office Supplies** made a more amount of lose. Loss is **-307.96.**
- Both **Region** and **Category** wise **south** & **Office Supplies** made a more amount of lose. Loss is **-365.34.**

3) Profitability Anomaly Detection

Identify any orders with **high sales but negative profit**. What patterns do you notice in terms of region, category, or quantity?

Key Insights

- The same quantities — especially **1**, **3**, and **6** units — are showing up again and again in loss-making orders.
- This happens in all product **categories** (Furniture, Office Supplies, Technology) and **regions** (East, West, North, South).
- This tells us that small quantity orders, even when they have high sales, are still not profitable.

4) Optimizing Product Mix for Regions

For each region, find the **best-selling category by volume** and the **most profitable category**. Are they the same? What does this imply?

Key Insights

- In the East, West, and South regions, the category with the highest sales also gave the highest profit. This shows that the current product mix in these regions is working well.
- In the North region, Office Supplies had the highest sales, but Furniture made more profit.
- Even though the sales difference between **Office Supplies** and **Furniture** was small **(₹724)**, the profit difference was meaningful **(₹705).**
- On deeper analysis, **Office Supplies** in **North** had more **negative profit** orders, while **Furniture** mostly made **positive profits**.

5) Demand Prediction Case

Using historical data, identify if there is a trend or seasonal pattern in **quantity sold** for each product category over time.

Key Insights

There is both a trend and seasonality present in the quantity sold over time:

- **Trends**: Increasing demand (especially for Office Supplies).
- **Seasonality**: Regular peaks at specific months across years.

6) Loss-Leading Product Investigation

Find products or categories that have **repeatedly** shown negative profit despite high sales. Should they be discontinued or repriced?

Key Insights

- The **Technology** category accounts for the highest volume of negative-profit transactions, especially in quantity groups of **1**, **5**, and **6** units.
- These three quantity buckets alone contribute **45%** of loss transactions, indicating that these sales are frequent and significant.
- Since these products are selling well (high sales count), it's more financially sound to reprice them (increase unit price, reduce discount) rather than discontinue them.

7) Regional Sales Consistency

Which region shows the **most stable monthly sales performance** over time? Use standard deviation or coefficient of variation to support your analysis.

Key Insights

- Based on the **coefficient of variation** (CV) for monthly sales across regions, the **North** region has the most stable sales performance over time **(CV = 0.36).**
- This indicates less fluctuation in monthly sales, making it the most consistent region in terms of sales.

8) Customer Retention Analysis

Based on Customer ID, find the number of repeats vs. one-time customers. How does their average profit and sales differ?

Key Insights

- Out of all customers, only a small portion are repeat customers, and they contribute ~**9%** of sales and ~**10.5%** of profit.
- The majority of revenue is currently driven by one-time customers, showing a potential gap in customer retention.
- There is no significant profitability difference between repeat and one-time buyers, indicating a possible opportunity to re-engage one-time buyers into becoming repeat customers.

9) Bulk Buying Patterns

Are their specific cities or regions where customers **consistently buy in higher quantities** than average? What product categories are driving this?

Key Insights

- The **East** and **West** regions show slightly above-average bulk buying behaviour, with mean quantities above the overall average **(4.898).** Loyalty programs could further boost retention.
- Standard deviation across regions is relatively consistent, indicating stable purchasing patterns.
- City-level analysis was inconclusive, as most cities appeared only once in the dataset and do not provide enough volume to draw meaningful conclusions.

10) Sales Efficiency Score

Create a new metric: **Profit per Unit Sold**. Rank cities based on this efficiency. What actionable insights can Walmart take?

Key Insights

A new metric, Profit per Unit Sold, was created to evaluate the sales efficiency of each city.

- The **top 10 cities** generate **high profit** per unit, indicating strong product mix or premium customer segments.
- The **bottom 10 cities** show **low or negative** efficiency, possibly due to high volume of low-margin items or higher return rates.

Actionable insights for Walmart:

- Focus on expanding profitable categories in high-efficiency cities.
- In low-efficiency cities, re-evaluate pricing, optimize product assortment, or address operational inefficiencies.

11) Price Sensitivity Study

Is there a **negative correlation between quantity sold and profit per unit** in any region or category? What does this suggest?

Key Insights

- According to the **region segment** – Yes negative correlation between quantity sold and profit per unit is occurring and negative correlation value = **-0.36**.
- According to the **Category segment** – Yes negative correlation between quantity sold and profit per unit is occurring and occurring negative correlation value = **-0.92**.

12) Campaign Impact Simulation

Assume Walmart ran a 10% discount campaign in **August 2024**. Recalculate profit for that month and evaluate how the campaign would have affected overall profitability.

Key Insights

- **Before 10%** discount: **₹25.86 profit/order** Loyalty programs could further boost retention.
- **After discount**: **₹25.78 profit/order**

Impact = only ₹0.08 difference

- This indicates that Walmart can safely run such discount campaigns without significantly harming profitability.
- If the discount leads to even a small boost in sales volume, the overall profit may actually increase.

13) Return Risk Zones

If high-quantity orders with low profit are considered risky for returns, which region shows the **highest risk exposure**?

Key Insights

Based on the Region segmentation **East** region has **Quantity count is 270**, **total profit is 289.05**, and **percentage is 11.09%** contribution.

- **East** has high quantity but very low profit, so it is most at risk.

14) Time to Profit Threshold

Calculate how many days (based on order date) it took each region to cross a cumulative profit of ₹1,000. Who was fastest?

Key Insights

- Based The South Region was the fastest to reach **₹1,000** profit in just **34 days** from their first order date.
- This shows stronger early sales momentum or better margins in that region.

15) High-Impact Customer Recovery Plan

Identify the bottom 5% of customers by profit. Suggest a personalized sales strategy for them based on their past order behaviour.

Key Insights

- All **48 customers** in the **bottom 5%** are one-time customers.
- High-quantity buyers **(Qty > 4)** among them are responsible for:
  80% of Quantity
  60% of Sales
  56% of (Negative) Profit
- These may be discount-driven buyers → suggesting repricing to improve profit.

6. Final Conclusion

This project allows me understand the Walmart's sales patterns. Based on the Exploratory Data Analysis I highlight key decision points.

- Customer segmentation & retention
- Pricing and Product strategy
- Region – specific planning
- Campaign and Profitability planning.

7. Tools Used

- Jupiter Notebook
- Python 3.x
- NumPy
- Pandas
- Matplotlib
- Seaborn