Technical Report: E-commerce Sales Analysis Capstone

1. Introduction

This report documents the end-to-end analysis of an e-commerce sales dataset. The dataset includes order-level information for products, customers, regions, and payments. The project follows a structured workflow:

1. Data Cleaning and Preprocessing
2. Exploratory Data Analysis (EDA)
3. Advanced Analysis and Insights
4. Actionable Business Recommendations

Objective:
To provide data-driven insights on sales performance, top products, regional revenue, customer behaviour, and opportunities for business growth.

---

2. Data Description

- **Dataset:** cleaned_data.xlsx
- **Rows:** 300+
- **Columns (15):**
    1. orderid
    2. orderdate
    3. customerid
    4. customername
    5. region
    6. productname
    7. category
    8. subcategory
    9. quantity
    10. price
    11. discount
    12. tax

13. shippingcost
14. paymentmethod
15. total_sales (calculated)

- **Data Cleaning Highlights:**
  - Missing CustomerID or CustomerName replaced with 'UNKNOWN'.
  - Negative prices replaced with median.
  - Duplicates removed using OrderID.
  - Text columns standardized (title() formatting).
  - Date conversion and numeric type corrections applied.

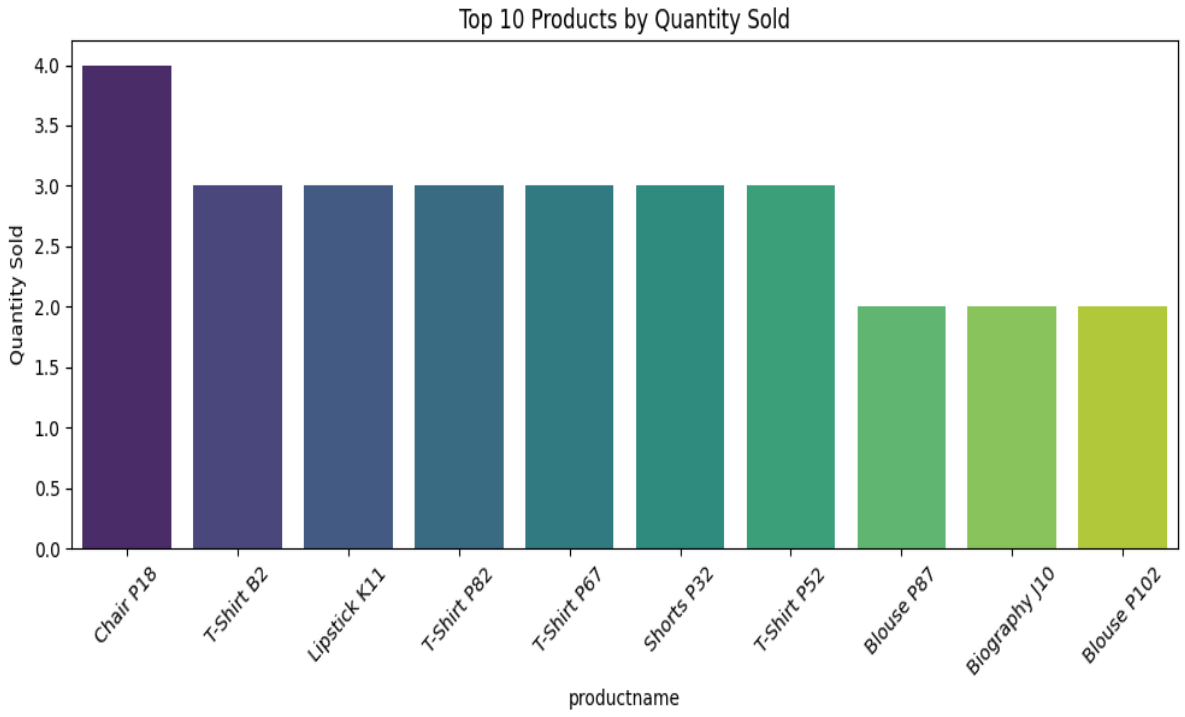
---

3. Exploratory Data Analysis (EDA)

3.1 Top 10 Products by Quantity Sold

- The top-selling products are identified using aggregated quantity.
- **Insights:** Helps focus inventory and marketing strategies on high-demand items.

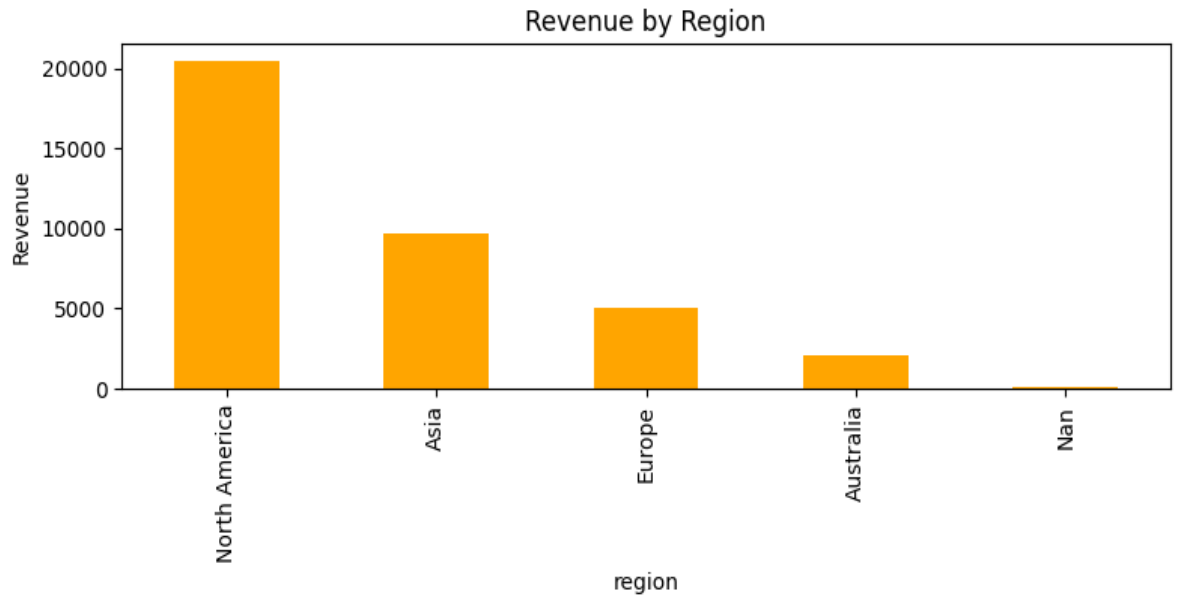
Visualization: Barplot of top 10 products by quantity sold.



3.2 Revenue by Region

- Aggregated total_sales by region to identify high-performing markets.
- **Insights:** Prioritize marketing and logistics in top revenue regions.

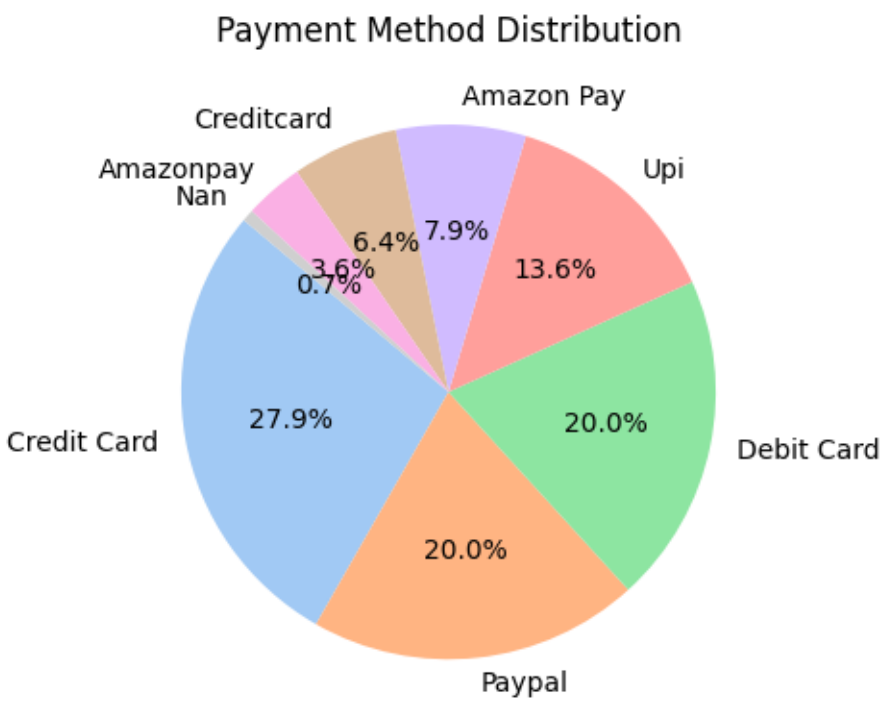
Visualization: Bar chart of revenue by region.



3.3 Payment Method Distribution

- Distribution of transactions by paymentmethod.
- **Insights:** Useful for financial planning and payment strategy optimization.

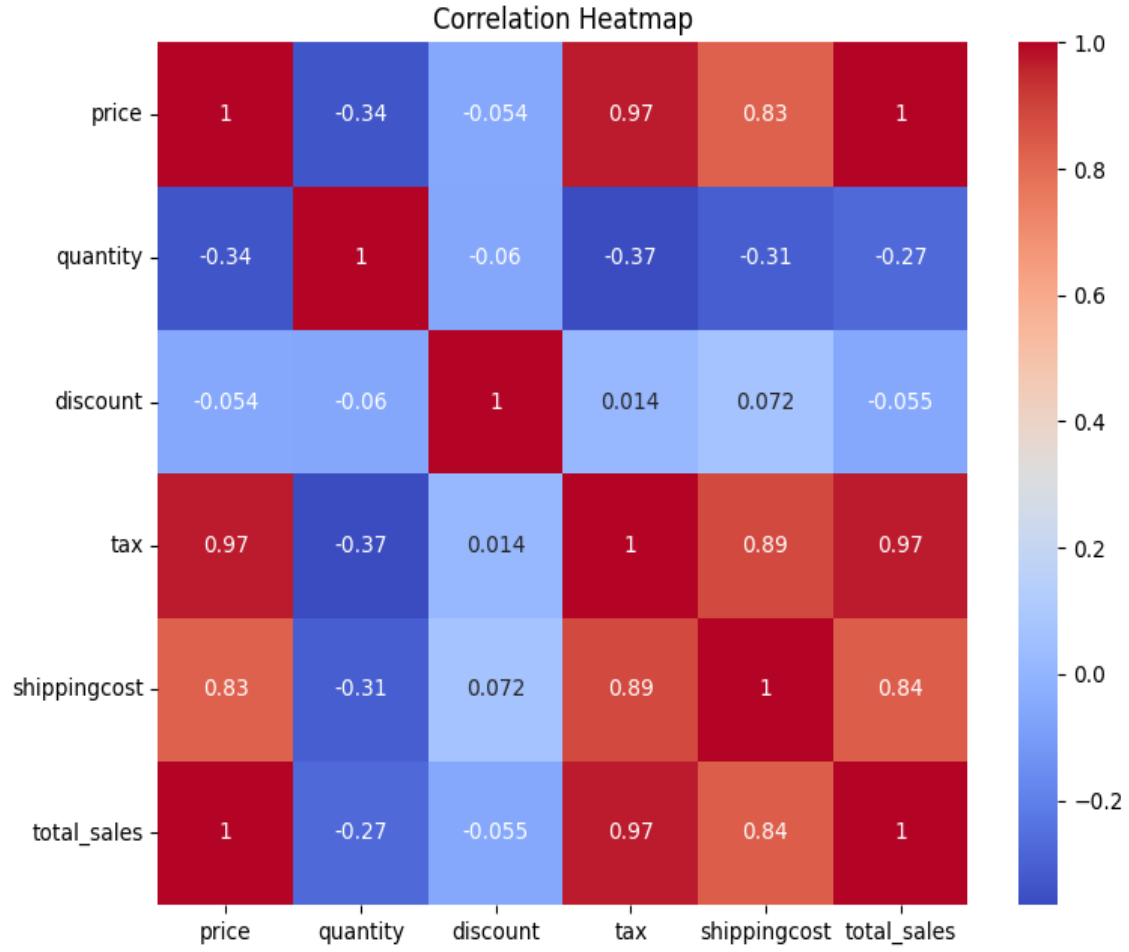
Visualization: Pie chart showing payment method distribution.



3.4 Correlation Analysis

- Correlation between numeric variables (price, quantity, discount, tax, shippingcost, total_sales).
- **Insights:** Discounts and shipping cost influence total revenue moderately.
- Helps identify variables that strongly drive sales.

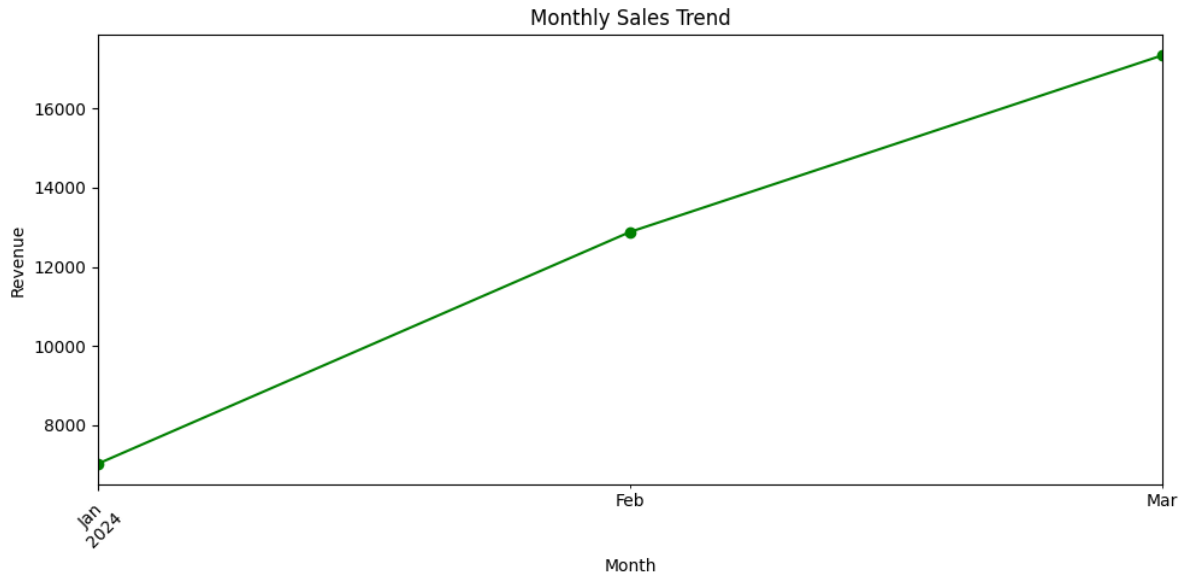
Visualization: Correlation heatmap.



3.5 Monthly Sales Trend

- Aggregated total_sales by month to identify seasonal patterns.
- **Insights:** Detect peak sales months for promotions or inventory planning.

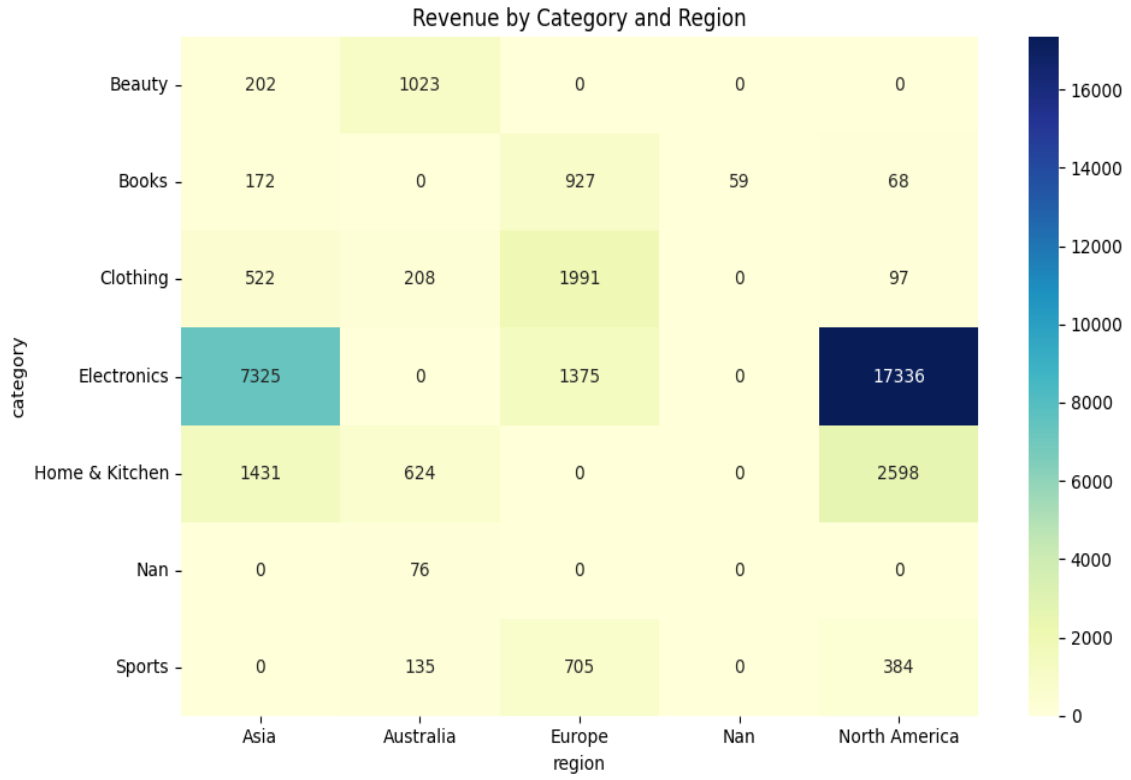
Visualization: Line chart showing monthly revenue trend.



3.6 Revenue by Category and Region

- Pivot table of total_sales across category and region.
- **Insights:** Shows which product categories perform best in each region.

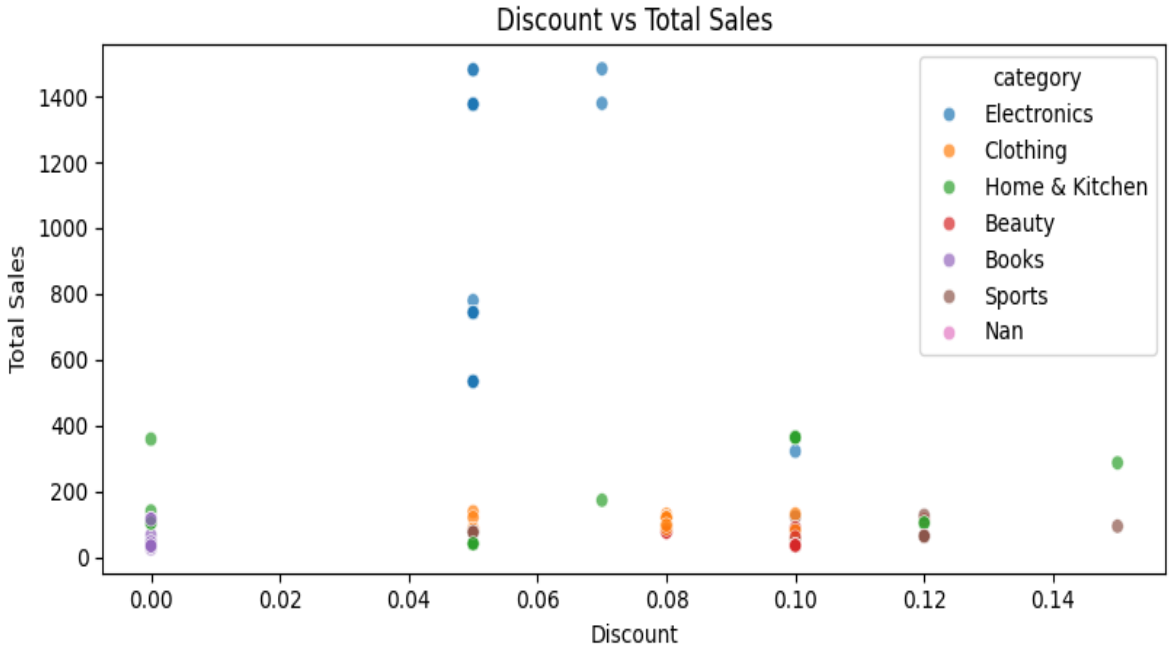
Visualization: Heatmap of revenue by category and region.



3.7 Discount Impact Analysis

- Scatter plot of discount vs total_sales by category.
- **Insights:** Evaluates effectiveness of discounts on revenue. Some categories respond more to discounts than others.

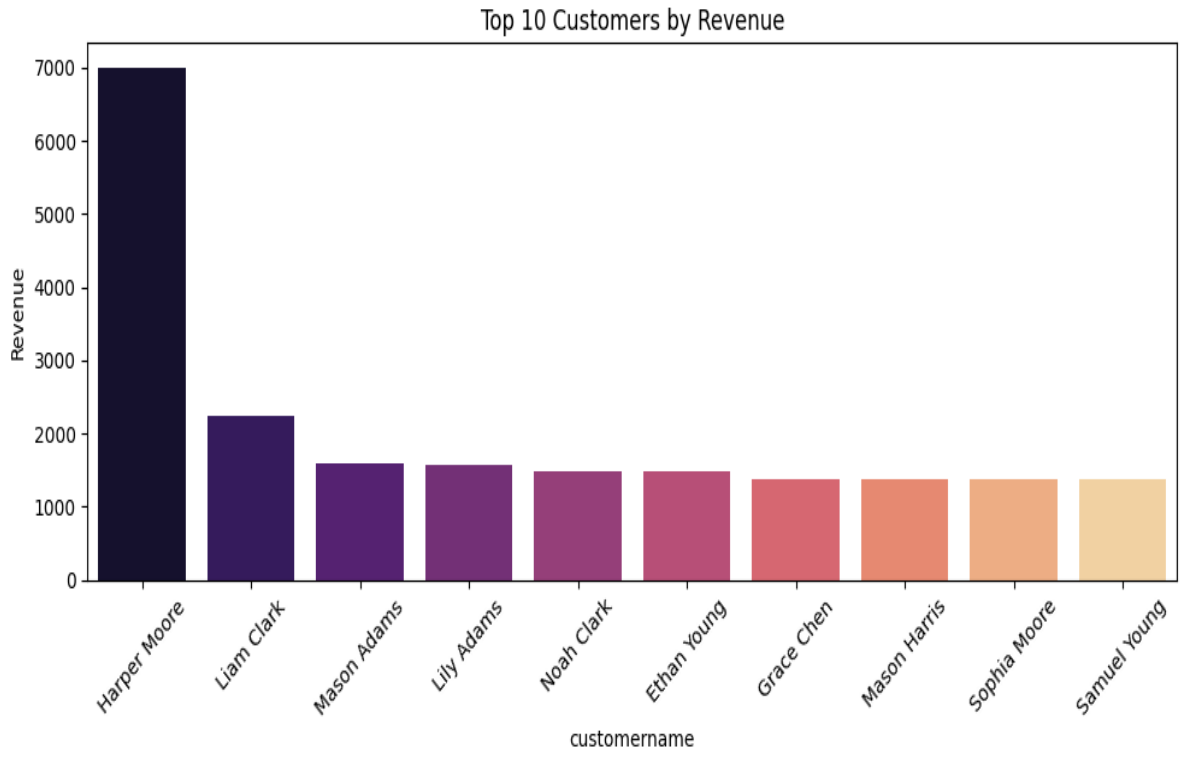
Visualization: Scatter plot of discount vs total sales.



3.8 Top Customers by Revenue

- Aggregated total_sales by customername.
- **Insights:** Identify high-value customers for loyalty programs and targeted marketing.

Visualization: Bar chart of top 10 customers by revenue.



---

4. Advanced Analysis & Recommendations

4.1 Upsell Opportunities

- Products with **high quantity but low revenue** identified for cross-selling or bundle promotions.
- **Example Logic:** total_quantity > 50 & total_revenue < 5000
- **Insights:** Suggests which products to promote to maximize revenue.

Visualization: Table of upsell opportunities.

Product Name	Total Quantity	Total Revenue
Product A	120	$3,200
Product B	95	$2,850

4.2 Recommendations Summary

1. Promote upsell opportunities via bundle deals or cross-selling.
2. Target marketing campaigns in regions with highest revenue potential.
3. Focus inventory on top-selling products to prevent stockouts.
4. Optimize discount strategies based on category performance.
5. Build loyalty programs for top customers identified in analysis.

---

5. Technical Details

- **Data Structures:**
    - Used **pandas DataFrames** for structured data manipulation.
    - Pivot tables for multi-dimensional aggregation.
- **Algorithms & Techniques:**
    - Data cleaning: missing values, duplicates, negative values, type conversions.
    - EDA: visualizations using matplotlib and seaborn.
    - Advanced analysis: groupby, aggregation, and filtering for actionable insights.

- **Architecture:** Modular scripts:

1. 1_data_cleaning.py → cleaning and preprocessing
2. 2_eda.py → exploratory visualizations
3. 3_analysis.py → advanced insights and recommendations

---

6. Testing & Validation

Test Case	Method	Result
Missing values	Fill nulls with default or median	Pass
Data type validation	Convert numeric/text types	Pass

Test Case	Method	Result
Negative prices	Replace with median	Pass
Duplicates	Drop duplicate OrderID	Pass
Revenue calculation	total_sales formula checked	Pass
Upsell opportunity identification	Filter by quantity and revenue	Pass

---

7. Conclusion

The analysis provides **comprehensive insights into product performance, regional sales, customer behavior, and discount effectiveness**. Key business recommendations include targeted upsell promotions, marketing focus on high-revenue regions, and inventory prioritization. The modular Python code ensures reproducibility and scalability for future datasets.