Bhavana Yelagandala

Retail Sales Data Analysis Report

1. Dataset Description

1.1 Source: Retail sales dataset (1,000 records).

1.2 Columns:

- **Transaction ID** – Unique identifier for each transaction
- **Date** – Transaction date (spanning Jan–Dec 2023)
- **Customer ID** – 1,000 unique customers (no duplicates)
- **Gender** – Male/Female distribution
- **Age** – 18 to 64 years (average ~41.4)
- **Product Category** – 3 categories: Beauty, Clothing, Electronics
- **Quantity** – Number of units purchased (1–5 range)
- **Price per Unit** – ₹25 to ₹1000 depending on product
- **Total Amount** – Transaction revenue (₹25 to ₹2000, avg. ₹456)

1.3 Data Quality:

- No missing values
- Clean and consistent format
- Balanced representation across genders and product categories

2. Operations Performed

2.1 Data Cleaning & Exploration

- No missing/null values found
- Parsed Date column into proper format
- Added Revenue as Quantity × Price per Unit
- Verified distinct customers and product categories

2.2 Descriptive Analytics

- Gender distribution (pie chart)

- Product category breakdown (pie chart)
- Age distribution (histogram)
- Revenue distribution (histogram)

2.3 Relationship Analysis

- Quantity vs. Revenue (scatter/hexbin plot)
- Revenue variations across categories (boxplots)
- Monthly revenue trend (line chart)
- Correlation heatmap (Quantity, Price, Revenue, Age)

3. Key Insights

3.1 Customer Demographics

- Balanced gender split: **Female 51%, Male 49%**
- Age range: **18–64 years**; average age - 41
- Core buyers: **30–50 years group**

3.2 Product Insights

- **Clothing** : highest number of transactions
- **Electronics** : fewer but high-value purchases
- **Beauty** : stable mid-range spending pattern

3.3 Revenue Insights

- Revenue spread: ₹25 – ₹2000 per transaction
- Average transaction: ₹456
- High correlation between Revenue and both Quantity & Price per Unit

3.4 Temporal Trends

- Steady monthly sales throughout 2023
- Revenue peaks in mid-year months (possible festive/seasonal effect)

4. Recommendations

4.1 Marketing Strategy

- Target mid-aged (30–50) customers with premium product promotions
- Engage younger customers (18–25) with discounts/offers to increase activity

4.2 Product Strategy

- Strengthen **Electronics marketing** as it yields higher revenue per transaction
- Maintain Clothing sales momentum through bundle offers

4.3 Revenue Growth Opportunities

- Upsell Beauty products with cross-category promotions
- Analyze repeat-purchase behavior for loyalty programs

4.4 Future Analytics Opportunities

- Predictive models for customer lifetime value
- Clustering customers by age, spending, and product preference
- Forecasting monthly revenue for inventory planning

5. Customer Purchase Trends

- **By Time of Day**: Sales peak during **afternoons and evenings**, aligning with leisure shopping hours.
- **By Day of Week**: Higher orders are recorded on **weekends**, suggesting strong leisure and family-driven purchases.
- **By Month**: Seasonal spikes appear around **festive months** (e.g., May, August, December), showing the influence of holidays and festivals.

6. Key Business Insights

- Revenue is driven by a small proportion of high-value transactions (principle: 20% customers → 80% revenue).
- Strong correlation between Quantity & Revenue and Price & Revenue, validating that both volume and pricing drive business outcomes.

- **Mid-aged buyers** represent the most valuable customer segment - a clear focus area for future marketing.

Technologies Used for Analysis

The project leverages Big Data Analytics tools and PySpark for efficient data handling:

- **Python**: Core programming language
- **PySpark**: Distributed processing and large-scale data manipulation
- **Pandas**: Additional exploration and validation
- **Matplotlib**: Visualization of trends and distributions
- **Jupyter Notebook**: Interactive development and reporting

Conclusion

The Retail Sales Data Analysis Project provides actionable insights into customer purchasing behavior and sales trends:

- Clothing dominates in transaction volume, while **Electronics generate higher-value purchases**
- **Mid-aged customers (30–50 years)** contribute the largest share of revenue
- Sales are **balanced across genders**, with females accounting for a slightly higher share
- Revenue per transaction ranges widely, highlighting both **budget and premium customer segments**
- Seasonal variations in monthly sales indicate **peaks in specific months**, useful for planning promotions
- Strong correlations exist between **Quantity, Price per Unit, and Revenue**, confirming expected business drivers

These findings can help retailers optimize product strategies, pricing, and targeted marketing campaigns. The analysis demonstrates how PySpark enables scalable big data insights for retail businesses, supporting data-driven decision-making.