Exploratory Data Analysis (EDA) Report

Kushagra Agrawal

January 25, 2025

1 Introduction

This report summarizes the key findings from the exploratory data analysis (EDA) performed on three primary datasets:

- **Customers** – 200 entries, 4 columns (*CustomerID, CustomerName, Region, SignupDate*)
- **Products** – 100 entries, 4 columns (*ProductID, ProductName, Category, Price*)
- **Transactions** – 1000 entries, 7 columns (*TransactionID, CustomerID, ProductID, TransactionDate, Quantity, TotalValue, Price*)

The main goal of the EDA was to uncover meaningful business insights related to customer information, product details, and transactional behaviors.

2 Data Overview

2.1 Customers

- 200 rows, 4 columns
- Columns: *CustomerID, CustomerName, Region, SignupDate*
- Regions include 4 distinct categories (e.g., Africa, Asia, Europe, South America).
- Preliminary analysis shows "South America" has the highest count of customers (59 out of 200), indicating a key geographical concentration.

2.2 Products

- 100 rows, 4 columns
- Columns: *ProductID, ProductName, Category, Price*
- Price ranges from approximately 16.08 (min) to 497.76 (max), with a mean of around 267.55.

- Suggests a wide spectrum of product pricing levels, potentially appealing to different market segments.

2.3 Transactions

- 1000 rows, 7 columns
- Columns: *TransactionID, CustomerID, ProductID, TransactionDate, Quantity, TotalValue, Price*
- Quantity ranges from 1 to 4, with an average of about 2.54.
- TotalValue ranges from 16.08 to 1991.04, with an average of about 690.00.
- A scatter plot of TotalValue over time revealed no significant anomalies, and Z-score checks showed no outliers.

3 Key Business Insights

3.1 Insight 1: Regional Concentration of Customers

A bar chart analysis on the "Region" column shows that South America has the highest concentration of customers (59 out of 200).

- **Implication:** South America is a strong market region for this business, presenting an opportunity to focus marketing and supply chain strategy there.
- **Recommendation:** Tailor promotional campaigns and optimize logistics/fulfillment strategies to best serve South American customers.

3.2 Insight 2: Product Pricing Distribution

Box plot analysis of the "Price" column in the Products dataset shows a wide range (16.08–497.76) and a mean of ~ 267.55.

- **Implication:** The product catalog spans from budget-friendly to high-end offerings, suggesting potential for segmenting products by price tier.
- **Recommendation:** Develop differentiated marketing strategies (e.g., premium marketing for higher-priced items) and ensure stock availability across varied price points.

3.3 Insight 3: Strong Correlation between Quantity, Price, and TotalValue

The Transactions correlation heatmap shows a high correlation between *TotalValue* and *Quantity*, as well as between *TotalValue* and *Price*. (This is logical because TotalValue = Quantity × Price.)

- **Implication:** Transaction revenue is driven directly by price and quantity; promotions that affect either metric (e.g., price discount or quantity bundling) will influence total sales value.
- **Recommendation:** Explore bundled pricing or volume discounts to stimulate purchasing multiple units; track how changes in discounts or pricing strategies affect total revenue.

3.4 Insight 4: Lack of Outliers in TotalValue

Z-score detection for outliers in “TotalValue” indicates zero transactions flagged, implying consistent transaction values with no extreme anomalies.

- **Implication:** With no glaring extremes, the revenue data is relatively stable.
- **Recommendation:** Continue monitoring on a rolling basis to catch future anomalies early (e.g., sudden spikes or dips in total transaction values).

3.5 Insight 5: Geographical Trend Identification

Although four distinct regions exist, a notable portion of customers is concentrated in South America while the other three regions collectively share the remainder.

- **Implication:** Some regions may be underpenetrated or smaller in customer base, potentially offering an untapped market.
- **Recommendation:** Investigate each underrepresented region (e.g., Asia, Africa, Europe) to determine if targeted campaigns or localized partnerships could expand market share.

4 Conclusions and Recommendations

The EDA reveals a stable and diversified product line (reflected in prices from about $16 to $498) and a geographically concentrated customer base (notably in South America). Correlation data confirms that total transaction value is primarily driven by product price and purchase quantity.

Key Recommendations

1. **Regional Focus:** Prioritize South America for targeted marketing and ensure logistical support meets demand. Investigate expansion or specialized promotions in lower-representation regions for potential growth.
2. **Pricing and Promotional Strategies:** Given the strong correlation between price, quantity, and total value, test bundle discounts or tiered pricing to increase average transaction value. Monitor changes in the correlation matrix to measure promotional effectiveness.

3. **Product Line Management:** Maintain a balanced inventory across the broad price range to cater to multiple customer segments, from budget to premium, ensuring minimal stockouts and capturing maximum market opportunity.

4. **Continuous Anomaly Monitoring:** Although no outliers were found in TotalValue, implement ongoing checks to detect unusual fluctuations promptly (e.g., sudden spikes from erroneous input or unanticipated discounts).

By implementing these data-driven actions, businesses can improve market penetration, optimize pricing strategies, enhance customer satisfaction, and maintain a stable revenue stream. Additional deeper analytics (e.g., forecasting, customer lifetime value analysis) may yield further refinements for strategic planning and sustainable growth.