Milestone 2: Exploratory Data Analysis (EDA) Report

Course: Data Analysis and Visualization (CS-352)
Milestone Lead by: Ahmad Mustafa
Group members: Malik Zaheer Abbas, Usaira Shabaz Zahid
Department: Computer Science, Namal University, Mianwali

April 2025

1 Introduction

Exploratory Data Analysis (EDA) is a crucial step in data analysis that helps in understanding the dataset's structure, identifying patterns, and detecting anomalies. This report presents an in-depth analysis of the dataset, including statistical summaries and visualizations.

2 Dataset Description

The dataset consists of **119,143 entries** and **15 columns**, including customer details, product information, pricing, payment methods, and reviews. Below is a brief description of key numerical attributes:

Column	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Price	118,310	120.65	184.11	0.85	39.90	74.90	134.90	6735.00
Payment Value	119,140	172.73	267.77	0.00	60.85	108.16	189.24	13664.08
Review Score	118,146	4.02	1.40	1.00	4.00	5.00	5.00	5.00

Table 1: Summary Statistics of Key Numerical Columns

Observations:

- Prices show a wide range, with some high outliers.
- Most payments are in the range of $60 to $189.
- Reviews are generally positive, with a **median score of 5**.

3 Visualizations and Insights

This section includes visual representations of the dataset's most insightful dimensions. Each graph is crafted to answer specific analytical questions related to sales, reviews, customer behavior, and operational performance. A total of six key graphs are discussed here—four focused on review-based analysis, and two on other crucial aspects like product categories and order statuses.

3.1 Top 10 Most Sold Product Categories

This bar chart visualizes the top 10 product categories based on the number of items sold. Categories such as "bed_bath_table," "health_beauty," and "sports_leisure" dominate the sales, indicating strong consumer demand. Understanding these trends can help the business refine inventory strategies, marketing campaigns, and pricing models to better serve customer needs.

3.2 Order Status Distribution

The order status distribution graph gives a snapshot of the entire order fulfillment lifecycle. The majority of orders fall under the "delivered" category, showcasing effective logistics. However, the presence of "canceled" and "unavailable" orders raises questions regarding stock management, vendor reliability, or customer dissatisfaction. Addressing these issues can enhance operational efficiency and customer trust.

3.3 Review Score Distribution

This bar chart reveals the frequency distribution of customer review scores ranging from 1 to 5. A majority of the reviews are either 4 or 5 stars, reflecting a generally positive customer experience. Nonetheless, a non-negligible number of lower scores (1-3 stars) suggest occasional dissatisfaction, possibly due to product quality, delays, or service issues. These reviews can be further analyzed for actionable feedback.

3.4 Average Review Score by Product Category

This visualization ranks the product categories based on their average review scores. Categories such as "watches_gifts" and "books_technical" stand out for having higher average scores, while others like "fashion_shoes" and "furniture_bedroom" exhibit lower averages. This analysis helps identify which categories excel in customer satisfaction and which require further investigation to address quality or service gaps.

3.5 Review Score by Payment Method

Different payment methods may correlate with customer sentiment. This graph shows the average review score associated with each payment method (e.g., credit card, boleto, voucher). Interestingly, credit card payments tend to result in slightly higher review scores, possibly due to smoother transaction processes or faster verification. Lower ratings linked to specific methods could be due to delays or confusion during checkout.

3.6 Review Score vs. Delivery Time

This scatter plot or line chart examines the relationship between delivery time and review score. The visualization highlights a downward trend in review scores as delivery time increases, suggesting that longer wait times may negatively impact customer satisfaction. This finding underlines the importance of optimizing shipping and fulfillment processes to enhance user experience and brand reputation.

4 Findings and Next Steps

Key Takeaways:

- Review scores are generally positive, suggesting customer satisfaction.
- Product and delivery quality appear to influence ratings.
- Some categories and payment methods show variation in feedback.

- A small percentage of orders remain canceled or pending.

Next Steps:

- Investigate negative reviews in more depth.
- Explore time-series trends of orders and reviews.
- Perform correlation analysis between review score, delivery time, and price.

5 Conclusion

This milestone provided an insightful look into the dataset's structure, key trends, and initial observations. The next phase will involve data preprocessing and correlation analysis to refine the dataset for predictive modeling.