DAV Project Phase 3: Preprocessing Report

Course: Data Analysis and Visualization (CS-352)
Phase Lead: Usaira Shahbaz (NUM-BSCS-2022-27)
Group Members:
Ahmad Mustafa (NUM-BSCS-2022-37)
Zaheer Abbas (NUM-BSCS-2022-50)
Department: Computer Science, Namal University, Mianwali

April 2025

Phase 3: Data Preprocessing Steps and Decisions

1 Introduction

In this phase, the goal was to clean and prepare the Brazilian E-Commerce dataset for analysis by addressing missing values, duplicates, and outliers. The preprocessing ensures that the data is consistent, complete, and ready for subsequent analysis and modeling.

2 Steps Taken

2.1 1. Dropping Unnecessary Columns

The `review_comment_title` column was dropped as it was deemed irrelevant to the sentiment analysis and contained a high percentage of missing values. This helps reduce noise and dimensionality in the dataset.

2.2 2. Handling Missing Values

We first checked missing values across all columns. It was found that columns like `seller_city`, `seller_state`, `product_category_name_english`, `payment_value`, and `order_delivered_customer_date` had significant missing data.

The following approach was taken:

- Rows missing critical information (`seller_city`, `seller_state`, `product_category_name_english`, `payment_value`, `payment_type`, `order_delivered_customer_date`, `review_comment_message`, `review_creation_date`) were dropped.
- The `price` column's missing values were filled with the mean price of the dataset, ensuring no bias was introduced through arbitrary value assignment.

After these steps, the dataset had no remaining missing values.

2.3 3. Handling Duplicate Values

We checked for duplicate rows using the `duplicated()` function. No duplicate rows were found, ensuring data uniqueness and integrity.

2.4 4. Outlier Detection and Handling

Outliers were detected using the Interquartile Range (IQR) method:

- For numeric columns (`price`, `payment_value`), outliers were identified outside 1.5 times the IQR.
- A significant number of outliers were detected:
  - `price`: 3553 outliers
  - `payment_value`: 3937 outliers
- Outlier values were replaced with the mean of the respective column to maintain data balance without completely discarding information.

2.5 5. Final Dataset Status

After completing preprocessing:

- The final dataset shape was **(47211 rows, 14 columns)**.
- There were **zero missing values** and **no duplicate rows**.
- Outliers were handled effectively to ensure better model performance later.

3 Rationale Behind Decisions

- **Dropping columns:** Reduces unnecessary complexity and focuses analysis on meaningful features.
- **Handling missing values:** Ensures the dataset is clean and models trained on it will not suffer from bias or errors due to incomplete information.
- **Removing duplicates:** Prevents over-representation of specific records.
- **Handling outliers:** Maintains the data range within acceptable limits without removing large portions of data, preserving dataset size.

4 Conclusion

The preprocessing phase successfully cleaned the dataset, making it ready for analysis. Strategic decisions were made to maintain data integrity, minimize information loss, and prepare a robust dataset for further sentiment analysis tasks.