Data Cleaning and Merging Report

Brazilian-Ecommerce

Data Cleaning and Merging Report for Olist Datasets

This report provides a detailed summary of the data cleaning process applied to the Olist datasets, which include **orders**, **order items**, **products**, **product category name translations**, **sellers**, **customers**, **order payments**, **order reviews**, and **geolocation** data. The objective was to ensure data quality by addressing inconsistencies, handling missing values, and standardizing formats across all files. Additionally, this report outlines the steps taken to merge these datasets into a unified dataset suitable for analysis, along with challenges encountered and recommendations for future improvements.

Overview of Key Cleaning Steps for Each Dataset

Below is a breakdown of the key cleaning steps performed on each dataset, as derived from the Jupyter notebooks and cleaned output files located in the clean/Data_Cleaned/ directory.

1. Customers Dataset (cleaned_olist_customers_dataset.csv)

- **Removed Duplicates**: Eliminated duplicate entries based on customer_id to ensure each customer is uniquely represented.
- **Handled Missing Values**: Filled missing customer_city and customer_state values with 'Unknown' to maintain data completeness.

  Data Type Standardization: Converted customer_zip_code_prefix to a string type to preserve leading zeros and ensure consistency with other datasets.

2. Geolocation Dataset (cleaned_olist_geolocation_dataset.csv)

- **Aggregation**: Grouped data by geolocation_zip_code_prefix, calculating the mean of geolocation_lat and geolocation_lng for average coordinates and selecting the most frequent geolocation_city and geolocation_state per zip code.
- **Duplicate Removal**: Removed duplicate entries to create a concise, representative dataset.
- **Missing Value Handling**: Addressed any missing values during aggregation to ensure reliable geographic data.

3. Order Items Dataset (cleaned_olist_order_items_dataset.csv)

- **Dropped Incomplete Rows**: Removed rows missing critical IDs (order_id, product_id, seller_id) to maintain data integrity.
- **Validated Numerical Data**: Ensured price and freight_value were non-negative by clipping values below zero.
- **Feature Engineering**: Added a total_cost column by summing price and freight_value for each order item.

4. Order Payments Dataset (cleaned_olist_order_payments_dataset.csv)

- **Missing Value Handling**: Filled missing payment_type entries with 'Unknown'.
- **Validated Numerical Data**: Ensured payment_value was non-negative by clipping negative values.
- **Categorized Payments**: Simplified payment_type into two categories: 'Card' (e.g., credit/debit cards) and 'Other' (e.g., vouchers, boleto) for easier analysis.

5. Order Reviews Dataset (cleaned_olist_order_reviews_dataset.csv)

- **Filled Missing Comments**: Replaced missing review_comment_title and review_comment_message with 'No Comment'.
- **Score Validation**: Confirmed review_score values were within the valid range of 1 to 5.
- **Feature Engineering**: Calculated response_time_days as the difference between review_answer_timestamp and review_creation_date to measure responsiveness.

6. Orders Dataset (cleaned_olist_orders_dataset.csv)

- **Dropped Incomplete Rows**: Removed rows with missing order_id to ensure each order is uniquely identifiable.
- **Handled Missing Timestamps**: Filled missing values in timestamp fields (e.g., order_purchase_timestamp, order_delivered_customer_date) with a placeholder future date ('2099-12-31') to maintain consistency.
- **Feature Engineering**: Calculated delivery_time_days as the difference between order_delivered_customer_date and order_purchase_timestamp.

7. Product Category Name Translation Dataset (cleaned_product_category_name_translation.csv)

- **Dropped Incomplete Rows**: Removed rows missing product_category_name or product_category_name_english to ensure complete translation mappings.
- **Text Standardization**: Converted English category names to lowercase and replaced spaces with underscores (e.g., "Home Appliances" → home_appliances) for consistency.

8. Products Dataset (cleaned_products_dataset.csv)

- **Dropped Incomplete Rows**: Removed rows with missing product_id to ensure each product is uniquely identifiable.
- **Filled Missing Numerical Values**: Replaced missing values in numerical columns (e.g., product_weight_g, product_length_cm) with their respective column medians.

- **Feature Engineering:**
  - Calculated product_volume_cm3 by multiplying product_length_cm, product_height_cm, and product_width_cm.
  - Flagged products as is_heavy if their product_weight_g exceeded the 75th percentile.

9. Sellers Dataset (cleaned_sellers_dataset.csv)

- **Dropped Incomplete Rows**: Removed rows with missing seller_id to ensure each seller is uniquely identifiable.
- **Handled Missing Values**: Filled missing seller_city and seller_state with 'Unknown'.
- **Geographic Validation**: Verified seller_state against Brazilian state codes and derived seller_region for regional analysis.

Merging Process

The merging process combined the cleaned datasets into a single, unified file (merged_olist_dataset.csv) to facilitate comprehensive analysis. The steps were as follows:

1. Base Dataset: Used the cleaned **orders dataset** as the foundation, as it contains core order information linking other datasets.

2. Sequential Merges:
    - **Order Items**: Merged with order_id to include product and seller details per order.
    - **Products**: Merged with product_id to add product-specific attributes.
    - **Category Translations**: Merged with product_category_name to incorporate English category names.
    - **Sellers**: Merged with seller_id to include seller information.
    - **Customers**: Merged with customer_id to add customer details.
    - **Order Payments**: Merged with order_id to include payment information.
    - **Order Reviews**: Merged with order_id to incorporate customer feedback.

3. Geolocation Integration:

- Aggregated the **geolocation dataset** by geolocation_zip_code_prefix to produce average coordinates and representative city/state values.
- Merged this aggregated data with:
  - **Customers** using customer_zip_code_prefix (renamed columns: e.g., customer_lat, customer_lng).
  - **Sellers** using seller_zip_code_prefix (renamed columns: e.g., seller_lat, seller_lng).

4. Output: The final merged dataset, enriched with geographic coordinates, was saved as merged_olist_dataset.csv.

Challenges and Recommendations

Challenges

- Data Type Consistency: Ensuring uniform data types (e.g., converting zip_code_prefix to strings) was critical to avoid merging errors, particularly with fields prone to formatting issues like leading zeros.
- Missing Value Handling: Balancing data preservation with integrity required careful decisions, such as using placeholders ('Unknown', future dates) versus dropping rows.
- Geographic Data Inconsistencies: The geolocation dataset had multiple entries per zip code, necessitating aggregation to create a single, reliable entry per prefix.
- Merge Complexity: Linking multiple datasets with different keys increased the risk of misalignment or duplicate records.

Conclusion

The data cleaning and merging process for the Olist datasets has resulted in a high-quality, unified dataset (merged_olist_dataset.csv) ready for analysis. Each file was meticulously cleaned to address duplicates, missing values, and inconsistencies, while the merging process integrated all relevant information, including geographic coordinates for customers and sellers. Despite challenges such as data type alignment and geographic aggregation, the resulting dataset provides a robust foundation for exploring e-commerce trends, customer behavior, and geographic insights in the Brazilian market. Implementing the recommended validations and refinements will further enhance its reliability for future use.
