Data Analysis Report

for Brazilian E-Commerce Public Dataset by Olist

Data Analysis Report for Brazilian E-Commerce Public Dataset by Olist

This report provides a comprehensive summary of the data analysis process conducted on the Brazilian E-Commerce Public Dataset by Olist. The analysis was designed to extract valuable insights into customer behavior, e-commerce performance, and financial reconciliation, utilizing multiple datasets such as orders, order items, products, payments, and reviews. The process involved several critical steps: data preparation, reconciliation criteria calculations, monthly aggregation for a financial dashboard, data export for visualization, and documentation with initial insights. Below, each step is explained in detail to outline the methodology and outcomes of the analysis.

Step 1: Data Preparation

Objective

The goal of this step was to load, verify, and clean the dataset to ensure it was ready for analysis, creating a unified dataset with all necessary columns for subsequent calculations and insights.

Process

1. **Loading the Dataset**

   - The cleaned and merged dataset, named merged_olist_dataset.csv, was imported into Python using the Pandas library.

2. **Column Verification**

   - We confirmed the presence of essential columns, including:
     - order_purchase_timestamp
     - payment_value
     - payment_type
     - order_status
     - price
     - freight_value
     - order_approved_at
     - order_delivered_customer_date
     - order_estimated_delivery_date
     - order_id

   - Missing columns were flagged to ensure compliance with analysis requirements.

3. **Data Type Handling**

   - Date columns (e.g., order_purchase_timestamp, order_delivered_customer_date) were converted to datetime format for time-based analysis.

   - Numerical columns like price, freight_value, and payment_value were verified as float types to support accurate computations.

4. **Final Cleaning**

   - Rows with missing values in critical columns (e.g., order_id, order_status, price, freight_value) were removed to maintain data integrity.

   - A new column, total_order_value, was computed as the sum of price and freight_value for each order item, providing a key metric for revenue analysis.

Outcome

- A finalized dataset was prepared with all required columns correctly formatted and present.
- The addition of total_order_value enabled revenue-based calculations in later steps.

Step 2: Reconciliation Criteria Calculations

Objective

This step focused on calculating key sales performance metrics, including total revenue, expected revenue, canceled orders, and late deliveries, to assess e-commerce efficiency and financial performance.

Process

1. **Total Revenue Calculation**

   - Total revenue was computed as the sum of total_order_value for orders with an order_status of "delivered," reflecting actual revenue from completed transactions.

2. **Expected Revenue Calculation**

- Expected revenue was calculated as the sum of total_order_value for orders where order_approved_at was not null, representing anticipated revenue from approved orders, regardless of delivery status.

3. **Canceled Orders Count**
   - The number of canceled orders was determined by counting entries with an order_status of "canceled," highlighting potential revenue losses.

4. **Late Deliveries Count**
   - Late deliveries were identified by comparing order_delivered_customer_date to order_estimated_delivery_date. Orders delivered after the estimated date were marked with a new column, is_late, and the total count was calculated.

Outcome

- **Total Revenue**: Sum of total_order_value for delivered orders.
- **Expected Revenue**: Sum of total_order_value for approved orders.
- **Canceled Orders**: Total count of orders with "canceled" status.
- **Late Deliveries**: Total count of orders delivered late.
  These metrics offer a clear picture of sales performance and delivery efficiency, critical for operational insights.

Step 3: Monthly Aggregation for Financial Dashboard

Objective

The aim was to aggregate data by month to prepare it for a financial dashboard, providing insights into financial performance, order breakdowns, and delivery metrics over time.

Process

1. **Extracting the Order Month**
   - A new column, order_month, was derived from order_purchase_timestamp to enable monthly grouping.

2. **Defining Pending Orders**
   - Orders neither delivered nor canceled (e.g., "shipped," "processing") were classified as pending using a new column, is_pending.

3. **Monthly Aggregations**

- Data was grouped by order_month, and the following metrics were calculated:
  - **Financial Overview**:
    - total_revenue: Sum of total_order_value for delivered orders.
    - total_payments_received: Sum of payment_value for delivered orders.
    - expected_revenue: Sum of total_order_value for approved orders.
  - **Order Breakdown**:
    - total_orders: Total unique orders per month.
    - canceled_orders: Count of canceled orders.
    - pending_orders: Count of pending orders.
    - delivered_orders: Count of delivered orders.
  - **Delivery Insights**:
    - late_deliveries: Count of late deliveries.
    - avg_delay_days: Average delay in days for late deliveries.

4. **Percentage Calculations**
   - Percentages of canceled, pending, and delivered orders were computed relative to total_orders per month.
   - %_revenue_reconciliation was calculated as the ratio of total_revenue to expected_revenue, indicating revenue realization efficiency.
5. **Average Order Value**
   - The average total_order_value for delivered orders was calculated per month.
6. **Handling Missing Values**
   - NaN values (e.g., from months with no late deliveries) were replaced with 0 for completeness.

Outcome

- A monthly aggregated dataset was produced, containing:
  - Financial metrics (revenue, payments, expected revenue).
  - Order statistics (total, canceled, pending, delivered).
  - Delivery performance (late deliveries, average delay).
  - Percentage metrics and reconciliation ratios.
    This dataset forms the foundation for the financial dashboard.

Step 4: Exporting Data for Visualization

Objective

To export the processed data in formats suitable for creating an interactive Power BI dashboard.

Process

1. **Exporting Monthly Data**
   - The monthly aggregated data was saved as monthly_olist_data.csv for high-level dashboard views.
2. **Exporting Detailed Data**
   - The detailed dataset with order-level information was saved as detailed_olist_data.csv for drill-down capabilities in Power BI.

Outcome

- Two files were generated:
  - monthly_olist_data.csv: For monthly trends and key metrics.
  - detailed_olist_data.csv: For detailed order exploration.
    These files are ready for Power BI import and visualization.

Step 5: Documentation and Initial Insights

Objective

To document the analysis process and provide preliminary insights based on the calculated metrics.

Process

1. **Documenting the Code**
   - The Python script was annotated with detailed comments for each step, ensuring transparency and reproducibility.
2. **Initial Insights**
   - Early observations included:

- Trends in revenue and expected revenue over time.
- The effect of canceled orders on revenue reconciliation.
- Patterns in late deliveries potentially impacting customer satisfaction.

Outcome

- A fully documented script detailing the analysis process.
- Initial insights to be expanded upon in the Power BI dashboard.

Conclusion

The data analysis process for the Brazilian E-Commerce Public Dataset by Olist was executed with precision, covering data preparation, reconciliation calculations, monthly aggregation, data export, and documentation. The resulting datasets—monthly_olist_data.csv and detailed_olist_data.csv—are primed for visualization in Power BI, enabling a comprehensive financial dashboard. This analysis provides a robust foundation for understanding e-commerce performance, with opportunities for further exploration in profitability and forecasting in subsequent phases. The process ensures actionable insights for optimizing business operations and enhancing customer experience.