Data Analysis and Exploration

Adithya Murali, Aditya Sindhavad, Ammar Mustafa, John Hwang, Manasa Maganti, Sameer Khan, Shashank Rao, Varsha Manju Jayakumar

Summary of Data Cleaning and Transformation

The dataset was subjected to various preprocessing steps to ensure quality and accuracy:

Handling Missing Values:

- Geolocation Data: The geolocation dataset contained multiple entries for the same zip code, leading to inconsistencies. To address this, the dataset was grouped by geolocation_zip_code_prefix, and the mean latitude and longitude were calculated for each group. This approach provided a representative central point for each zip code, effectively handling missing or inconsistent geolocation entries.
- Product Category Names: The dataset included product category names in Portuguese, with some entries missing. These missing values were filled with 'unknown' to maintain dataset integrity. Subsequently, the category names were translated into English using a predefined translation mapping.

Data Type Conversion:

- Datetime Conversion: Columns representing dates and times, such as order_purchase_timestamp, were converted from string formats to datetime objects. This conversion facilitated time-based analyses and ensured consistency in temporal data handling.

Feature Engineering:

- Customer Return Indicator: A new binary variable, customer_return, was created to identify repeat customers. This feature was derived by checking for duplicated entries in the customer_unique_id column, providing insights into customer loyalty and repeat purchase behavior.
- Delivery Timeliness: The datasets were merged to calculate the actual delivery time for each order. By comparing the estimated delivery date with the actual delivery date, a new feature, late_delivery, was engineered to indicate whether an order was delivered late.

Data Merging:

Multiple datasets, including orders, customers, geolocation, order items, and products, were merged using common keys. This consolidation resulted in a comprehensive dataframe

encompassing all relevant information for each order, facilitating a holistic analysis of the e-commerce platform's operations.

Key Insights from Exploratory Analysis

- **Order Trends Over Time**: A notable spike in orders was observed during Black Friday, reflecting seasonal shopping trends.
- **Orders by Day of the Week**: Mondays recorded the highest number of orders, while weekends saw a significant drop in sales.
- **Orders by Time of Day**: The majority of purchases happened in the afternoon, indicating peak shopping hours.
- **Geographic Distribution**: São Paulo had the highest order volume, followed by Rio de Janeiro and Belo Horizonte, demonstrating strong urban market demand.
- **Freight Value Fluctuations**: Shipping costs varied over time, with a peak observed in mid-2018, possibly due to logistic cost changes.
- **Revenue Growth**: A remarkable 142.15% increase in revenue was noted from January-August 2017 to the same period in 2018, showcasing business growth.
- **Product category -** Bedding and bath is the most popular category, followed by Beauty and health, while categories like office furniture have the lowest order volumes.
- **Order data based on location -** São Paulo is the dominant contributor to total purchases despite variations in income levels and population size across other states.

Justification of Data Exclusions and Assumptions

1. **Exclusions**:
    - Orders with Missing Delivery Information: Orders lacking essential delivery dates were excluded from delivery performance analyses to ensure accuracy.
    - Inactive Customers: Customers without any completed orders were removed from the analysis to focus on active user behavior.
2. **Assumptions**:
    - Geolocation Accuracy: It was assumed that the averaged latitude and longitude for each zip code accurately represented the central point of that region. While this approach mitigates individual address inaccuracies, it provides a reasonable estimation for regional analysis.
    - Review Authenticity: Customer reviews were assumed to be genuine reflections of customer satisfaction, serving as a reliable metric for assessing product and service quality.

This meticulous data preparation and exploratory analysis have established a robust foundation for subsequent modeling and strategic decision-making aimed at enhancing customer satisfaction and operational efficiency.