Olist - Brazilian Marketplace

Overview

This document provides a comprehensive overview of the Brazilian E-Commerce Public Dataset by Olist. This dataset offers valuable insights into the Brazilian e-commerce landscape, capturing information about orders made on Olist Store from 2016 to 2018.

Objectives

The dataset aims to:

- Facilitate research and analysis of e-commerce trends in Brazil.
- Enable exploration of customer behavior, product performance, and order fulfillment processes.
- Provide a platform for developing and testing machine learning models for e-commerce applications.
- explore the relationship between order details and customer reviews.
- understand geographic patterns in orders and deliveries across Brazil.
- optimize logistics, payment types, and pricing strategies.

Data Source and Description

The dataset originates from Brazilian E-Commerce Public Dataset by Olist, a leading Brazilian marketplace that connects small businesses with consumers. It encompasses information on approximately 100,000 orders placed between 2016 and 2018. The data encompasses various aspects of the e-commerce journey, including:

• **Orders table**: Information on each order's status, timestamp, and fulfillment details.
• **Products table**: Details about the products sold, including categories.
• **Customers and Sellers tables**: Customer and seller details, along with geolocation information
• **Payments table**: Information on payment types and payment sequences.
• **Order Items table**: Details on the items sold in each order.
• **Reviews table**: Customer reviews and ratings for their shopping experience.
• **Geolocation table**: Maps Brazilian zip codes to latitude and longitude.

- **Order details:** Order status, price, freight information, and customer location.
- **Payment information:** Payment type and sequential number.
- **Product details:** Product ID, category, and attributes.
- **Customer information:** Customer ID (anonymized) and geolocation data (latitude and longitude).
- **Seller information:** Seller ID (anonymized) and geolocation data.

Key Columns

1. Orders Table

• **order_id: Unique identifier for each order.**
• **customer_id: Reference to the customer who placed the order.**
• **order_status: Status of the order (e.g., delivered, canceled).**
• **order_purchase_timestamp: Timestamp when the order was placed.**
• **order_approved_at: Timestamp when the payment was approved.**
• **order_delivered_carrier_date: Date the order was shipped.**
• **order_delivered_customer_date: Date the order was delivered to the customer.**
• **order_estimated_delivery_date: Estimated delivery date of the order.**

2. Order Reviews Table

• **order_id: Foreign key linking to the Orders table.**
• **review_score: Customer rating for the order (scale from 1 to 5).**
• **review_comment_message: Text of the review left by the customer.**

3. Order Payments Table

• **order_id: Foreign key linking to the Orders table.**
• **payment_type: Type of payment (credit card, debit card, etc.).**
• **payment_sequential: Number of payments made for the order.**
• **payment_value: Total amount paid for the order.**

4. Order Items Table

• **order_id: Foreign key linking to the Orders table.**
• **product_id: Identifier of the product in the order.**
• **seller_id: Identifier of the seller providing the product.**
• **price: Price of the product.**
• **freight_value: Shipping cost for the product.**

5. Products Table

• **product_id: Unique identifier for each product.**
• **product_category_name_english: Category of the product in English.**

6. Customers Table

• **customer_id: Unique identifier for each customer.**
• **customer_zip_code_prefix: First five digits of the customer’s zip code.**
• **customer_city: Customer's city.**
• **customer_state: Customer's state (removed in cleaning**

7. Sellers Table

• **seller_id: Unique identifier for each seller.**
• **seller_zip_code_prefix: First five digits of the seller's zip code.**
• **seller_city: Seller's city.**
• **seller_state: Seller's state (removed in cleaning).**

8. Geolocation Table

• **geolocation_zip_code_prefix: Zip code reference.**
• **geolocation_lat: Latitude of the location.**
• **geolocation_lng: Longitude of the location.**
• **geolocation_city: City corresponding to the zip code.**
• **geolocation_state: State corresponding to the zip code.**

Additional Custom Columns Created:

- **number_of_payment_types: A calculated column indicating the number of payment types based on payment_sequential.**
- **total_payment_value: A calculated column for the total payment value, which equals the sum of price and freight_value.**

These key columns play a critical role in understanding the relationships between orders, payments, products, customers, sellers, and their geolocations

Data Analysis questions:

1- what are the cities with the most number of orders ? Bar/column chart
2- who are the top 5/10 customers with the most total orders ? Bar/column
3- how many times and why there was a difference between shipping limit date and order delivered date ? Line/area
4- what is the order with the highest paid price? Card
5- what is the payment type most prefered by the customers? Pie/donut
6- how many times and why a customer used more than one payment type for one order ? Pie/donut
7- how many times the customer was satisfied for his order (rating 5) ? Pir /donut
8- how many times AND why did a customer cancel an order ? Pie /donut – key influencer
9- is the freight price related to prdouct weight ?
10- What are the top 5/10 selling products ? Bar/column
11- what are the products with the most orders count? Bar column
12- who are the sellers with the most orders published on the marketplace? Bar/column
13- what are the top 5/10 cities with most total orders published by the sellers ? Bar/column

Data Cleaning Process

Data cleaning was conducted using **Power Query** to ensure consistency and remove any irrelevant information. Below are the detailed steps performed:

Merging Tables

1. **Merge Orders with Reviews**: The review_score from the Reviews table was merged with the Orders table to link each order with its corresponding customer review.
2. **Merge Orders with Payment Details**: The payment_type and payment_sequential fields from the Payments table were merged into the Orders table.
3. **Merge Order Items with Orders**: The Orders table was merged with the Order Items table to include detailed product information in each order.
4. **Merge Products with Product Categories**: The Products table was combined with the Product Category table to get the product categories in English.
5. **Merge Customers with Geolocation**: Latitude and longitude information from the Geolocation table was merged into the Customers table for mapping purposes.
6. **Merge Sellers with Geolocation**: Latitude and longitude information was also merged into the Sellers table.

Adding Columns

7. **Add Conditional Column (Number of Payment Types)**: A new column was added to indicate the number of payment types based on the payment_sequential column.
8. **Add Custom Column (Payment Value)**: A new column was created to calculate the total payment value, which equals price + freight_value.

Removing Irrelevant Data

9. **Remove Customer Unique ID and Customer State**: These columns were removed from the Customers table as they were not relevant to the analysis.
10. **Remove Seller State**: This column was removed from the Sellers table as it was redundant for the analysis.
11. **Remove Duplicates**: Duplicate records were identified and removed from the Customers and Sellers tables to ensure data accuracy.

Data Modeling:

In this project, a star schema was implemented to efficiently organize the data for reporting and analysis purposes. The star schema consists of a central fact table that captures key transactional data, surrounded by dimension tables that store descriptive information about the entities involved in the transactions.

1. **Fact Table: olist_order_items_dataset**

At the core of the star schema is the olist_order_items_dataset, which serves as the fact table. This table contains transactional data related to orders, including fields like:

customer_id: Links to the customer who made the order.

freight_value: The shipping cost for each order.

order_approved_at: The timestamp when the order was approved.

order_delivered_carrier_date: The date the order was delivered to the carrier.

order_delivered_customer_date: The date the order was delivered to the customer.

Number of payment_type: The number of payment types for the order.

This table tracks the essential metrics used for analysis, such as order delivery times, payment methods, and freight values, making it the focal point of data aggregation and computation.

2. **Dimension Tables**

Each dimension table provides additional context and descriptive information about the data in the fact table:

olist_products_dataset:

Contains product-specific information such as the product category, dimensions (height, length, weight), and the number of photos.

Linked to the fact table via product_id, it helps break down sales and order details by product type and attributes.

olist_customers_dataset:

Stores customer details, including the customer_city, customer_zip_code_prefix, and geographical coordinates (geolocation_lat, geolocation_lng).

This table is connected to the fact table through customer_id and is used for analyzing sales and order metrics by customer demographics and location.

olist_sellers_dataset:

Provides information about sellers, including seller_city and geolocation details.

Linked through seller_id, this table helps evaluate order performance by seller location and region.

3. Relationships

In the star schema, the fact table is connected to each dimension table via a one-to-many relationship. The foreign keys in the fact table (such as product_id, customer_id, and seller_id) link to the primary keys in the dimension tables, forming the "star" structure.

This model is ideal for analyzing transactional data such as:

Sales performance by product category and product attributes.

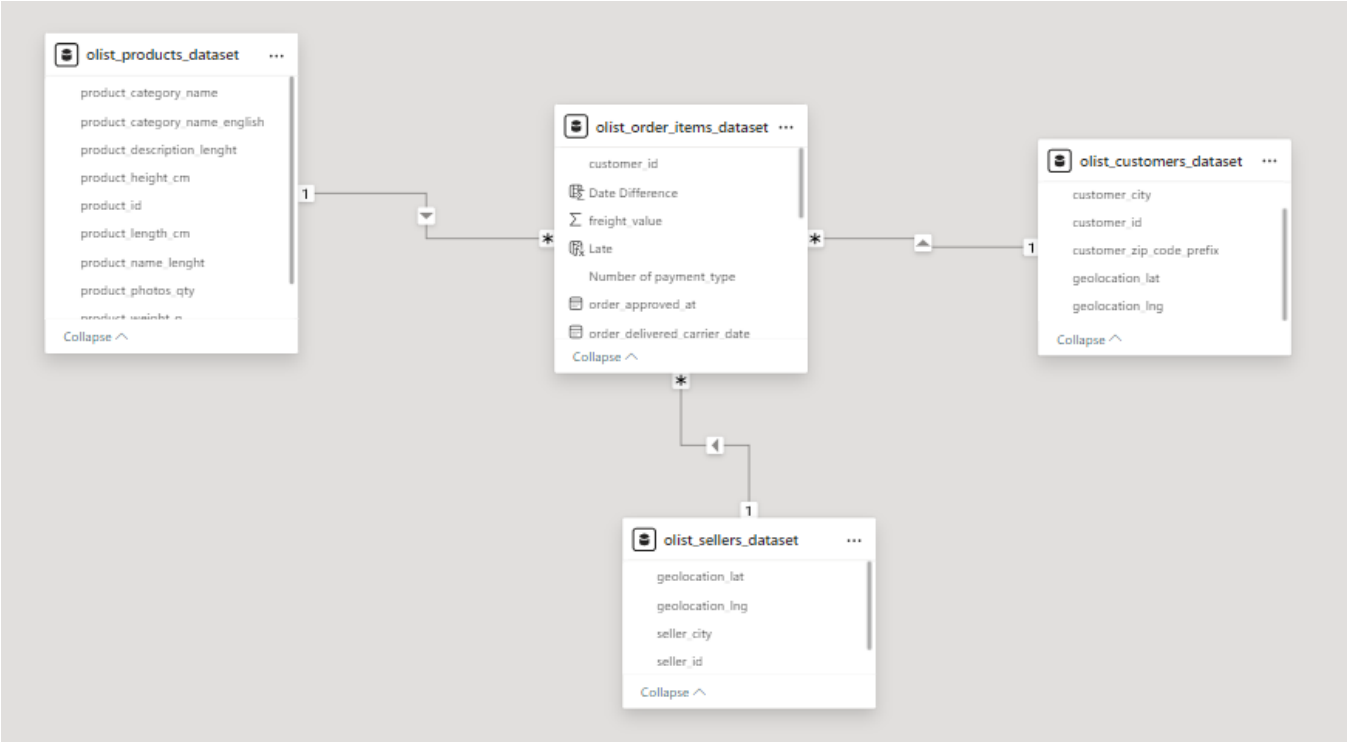
Customer behavior across different geographical regions.

Seller performance, including regional trends and order fulfillment metrics.

The star schema simplifies querying for aggregations and slicing the data by different dimensions, providing an intuitive structure for building reports and dashboards in tools like Power BI or Tableau. This structure also improves performance by reducing the complexity of joins, ensuring fast and efficient reporting.

Before reducing

The final data model



The benefits of simplifying the data model

Heightened Awareness: Even in this compressed model of the data, the analysts are still able to locate trends, patterns or outliers in the data rather quickly.

Lowered Difficulty Level: A simplified model enables easier creation of intuitive and easily interpretable visualizations and dashboards.

Smart Choices: Businesses decide better on product development, products marketing, and customer services through better understanding of data.

Analysis Methodology

1.DAX

Total Customers:	This measure computes the resultant number of customers in the dataset
Active Customers:	This measure establishes gross number of customer who have placed at least one order.
Best Customers:	In this measure, we count all the customers that have more than 10 orders granted to them.
total Orders:	This measure computes the overall number of orders requested.
Average Review Score:	In this measure, average review score is computed based on the evaluations of customers in all the orders.
Highest Order Value:	This measure describes the largest order in value compris s of the order price and the value of freight.
Total Sales:	This measure computes the total amount of sales by taking into account all the order transactions received.
Total Sellers:	This measure computes the cumulative number of sellers in the dataset.
Active Sellers:	This measure counts the groups of sellers who have trnasacted sales of more than 3 units
Best Customer Column:	This is an additional column added to the Customers Table where a customer belongs to the best customer category if she has placed more than 10 orders else she belongs to the customer category.
Date Difference column:	This column finds out the period of delay in the Delivery in relation to the estimated delivery.
Late Delivery Column :	This is a flag column where the custom flag refers to a flag as to whether the order was delivered late or on schedule based on the Date Difference column

2.Dashboard Creation:

The data was impressively displayed with the five dashboards that were made up:

Customers Dashboard

Cards: Total Clients, Active Clients, Most Valued Clients.

Orders Dashboard

Cards: Number of Orders, Largest Order Value, Sales Amount.

Sellers Dashboard

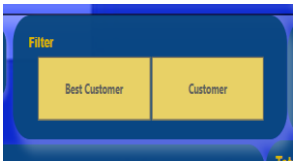
Cards: Active Sellers, Total Sellers, Average Rating.

Products Dashboard Cards: Payments Received, Goods Sold, Orders Made, Average Customer Rating.

3.Visualizations:

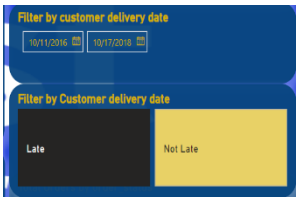
Visuals of customer dashboard :

- Bar plot – Customers by city.
- Pie chart: Reviews by score.
- Bar plot – Average number of orders per customer.
- Pie chart: Payment Methods.
- Slicer: Best Customer vs Customer.



Visuals of orders dashboard :

- Bar plot – Maximum of 5 orders with the largest revenue.
- Pie chart: Order Tracking Status (Delivered/Not Delivered).
- Bar plot: Value of orders per year and month.
- Pie chart: Order Tracking at a Glance.
- Slicer: Order Date and Tracking Status

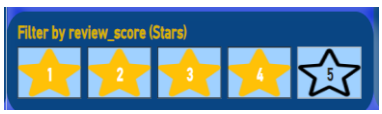


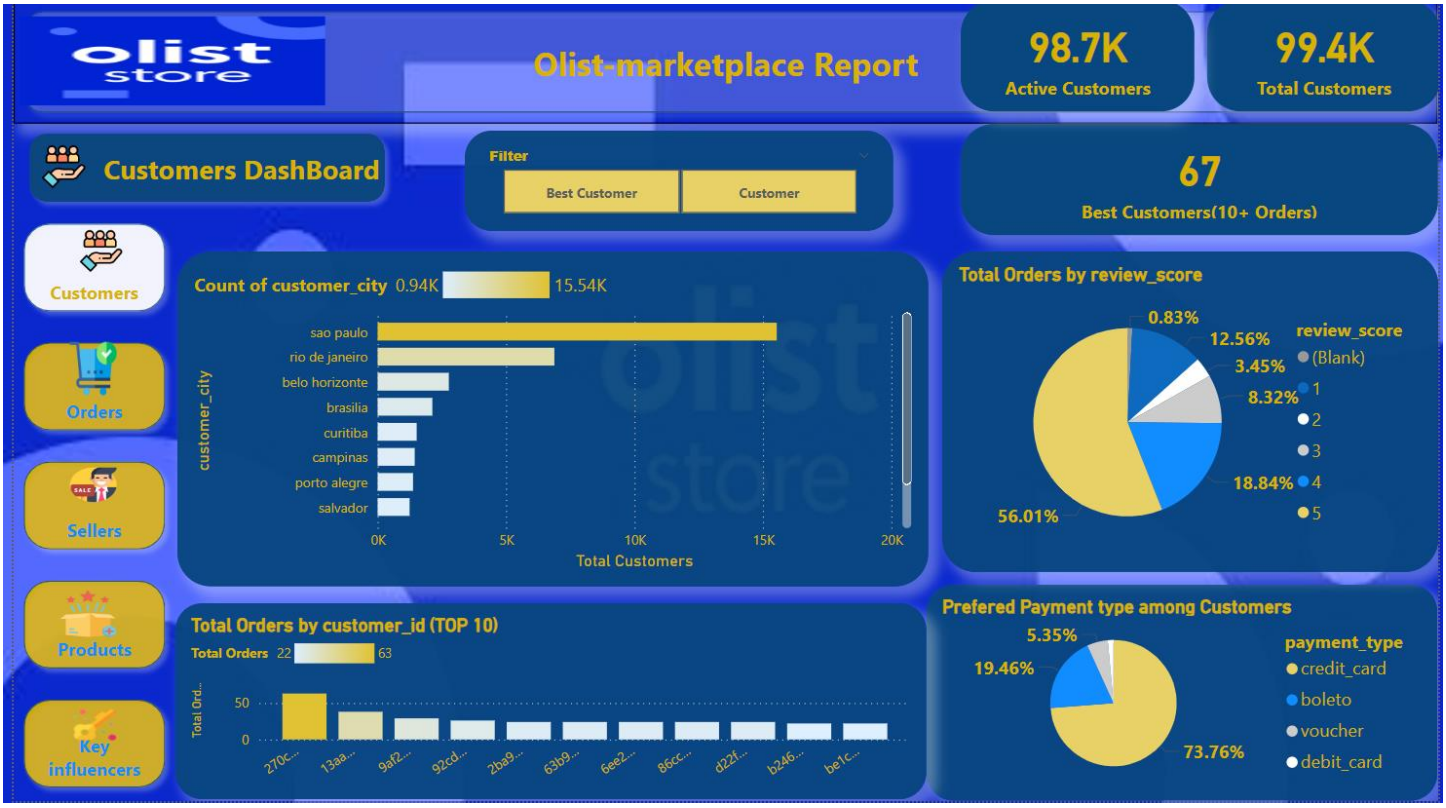
Visuals of sellers dashboard :

- Bar Plot – Number of Sellers in Different city.
- Vertical bar plot: sellers with highest sales.
- Bar plot ordered sellers by top ten based on number of orders.
- Bar plot: top 10 sellers with 5 ratings.

Visuals of products dashboard :

- Bar chart: ranking by number of orders of 10 hotselling product categories.
- Table: Names of items and the scores they received.
- Bar chart: 10 best-selling categories asked for.
- slicer by Average review (Stars).





Key findings:

- **Geographic Concentration:** A significant portion of customers is concentrated in major cities like Sao Paulo and Rio de Janeiro, indicating the importance of these urban markets.
- **Diverse Customer Ratings:** Customer reviews exhibit a range of scores, suggesting that there's room for improvement in product quality or customer service.
- **Preference for Digital Payments:** Customers overwhelmingly prefer electronic payment methods like credit and debit cards, reflecting the growing trend of e-commerce.
- **High-Value Customers:** Customers with 10 or more orders contribute significantly to overall sales, highlighting the importance of customer retention.

Analysis Summary:

The concentration of customers in major cities underscores the need for targeted marketing strategies tailored to these specific urban demographics.The varied customer ratings indicate that there's a need to consistently improve product quality and customer service to enhance overall satisfaction.The strong preference for digital payments aligns with global e-commerce trends and reinforces the need to offer a secure and convenient payment experience.

Key Observations:

- **Order Growth:** The dashboard indicates a steady increase in the number of orders over time, suggesting a growing customer base and successful marketing efforts.
- **High Order Value:** The "Highest Order Value" metric highlights the potential for high-value customers and opportunities for upselling or cross-selling.
- **Efficient Delivery:** The "Order Status" pie chart shows a high percentage of delivered orders, indicating efficient delivery operations.
- **Late Deliveries:** While the percentage of late deliveries is low, it's important to investigate the reasons for these delays and implement measures to improve on-time delivery.
- **Customer Satisfaction:** The "Total Orders by review_score" chart provides insights into customer satisfaction levels. Analyzing the distribution of review scores can help identify areas for improvement in product quality or customer service.

Analysis Summary

The orders dashboard offers a comprehensive overview of the platform's order performance. Data indicates a steady growth in the number of orders and total sales, suggesting a healthy business model and an expanding customer base.

Key Observations:

- **Geographic Concentration:** A significant number of sellers are located in São Paulo, indicating a strong presence in that region.
- **Top-Performing Sellers:** The "Sellers with Highest Sales" chart highlights the top-performing sellers based on total sales.
- **Active Sellers:** The "Active Sellers" card identifies sellers with three or more orders, indicating a minimum level of activity.
- **Seller Ratings:** The "Average of review_score" card provides an overall assessment of seller performance, with an average of 4.03 suggesting generally positive reviews.
- **Top-Rated Sellers:** The "5 STAR Rating COUNT by seller_id (TOP 10)" chart identifies sellers with the highest number of 5-star ratings.

Analysis Summary:

- **Seller Performance:** The dashboard reveals a wide range of seller performance, with some sellers achieving significantly higher sales and customer satisfaction than others.
- **Geographic Concentration:** The concentration of sellers in São Paulo suggests that this region may have a larger pool of potential sellers or offer favorable market conditions.
- **Seller Engagement:** The "Active Sellers" metric indicates that a significant number of sellers are actively engaged in the marketplace.

Key Observations:

- **Product Diversity:** The dashboard displays a wide range of product categories, indicating a diverse product offering.
- **Top-Selling Categories:** The "Total Orders by product category name (TOP 10)" chart identifies the top-selling categories, providing insights into customer preferences.
- **Product Ratings:** The "Average of review_score" column shows the average customer ratings for each product category, offering a measure of customer satisfaction.
- **Sales Performance:** The "Total sales by Product Category Name (Top 10)" chart highlights the top-performing categories in terms of total sales.

Analysis Summary:

- **Product Popularity:** The dashboard reveals the most popular product categories based on sales volume and customer reviews.
- **Customer Satisfaction:** The average review scores provide insights into customer satisfaction levels for different product categories.
- **Product Diversity:** The wide range of product categories suggests a diverse offering to cater to different customer preferences.
- **Sales Potential:** Identifying top-selling categories can help inform product development and marketing strategies.

Key Observations:

- **Payment Type:** The "product_category_name_english" significantly influences the likelihood of multiple payment types being used.
- **Order Status:** The "Sum of Payment Value" is a key factor in influencing whether an order is canceled.
- **Late Delivery:** The "Sum of Payment Value" and "product_category_name_english" both influence the likelihood of late deliveries.
- **Freight Value:** The "product_weight_g" significantly influences the increase in freight value.

Analysis Summary:

- **Payment Behavior:** Customers in certain product categories are more likely to use multiple payment types.
- **Order Cancellation:** Higher-value orders are more likely to be canceled, suggesting potential issues with product availability, pricing, or customer dissatisfaction.
- **Late Delivery:** Orders with lower payment values and products in specific categories are more likely to be late, indicating potential challenges in the delivery process or product fulfillment.
- **Freight Costs:** Heavier products naturally lead to higher freight costs.

Conclusion:

Revising the Brazilian E-Commerce Public Dataset provides helpful understanding regarding aspects such as customer behavior, seller productivity, product assortment, and operational capabilities of the Olist marketplace platform. These are the key findings from the study, however, here.

Geographic Distribution:

Most of the clients and the sellers are evenly distributed within the foremost metropolitan cities like Sao Paulo and Rio de Janeiro. This demonstrates the fact that the urban areas are quite essential in any marketing strategies and that there is a need to widen logistical networks to adequately satisfy the demand in these areas.

Customer Trend & Reviews:

The customer ratings tend to vary, which implies that improvement can be made in terms of the product and service to the customer. It however seems that high value customers, and customers who ordered 10 or above a significant percentage of sales order, are high yielding on total sales. In this case, customers can be retained using loyalty programs or targeted marketing to increase sales.

Means of Payments:

The use of online payment means such as credit and debit cards is getting common and there is a compelling reason to embrace these methods as there are advantages where e commerce is predominant as they simplify the payment process for the customers.

Order Growth and Sales Opportunities:

The further a company progresses, there is an increase in order volume which may denote great marketing strategies as well as an increase in the customers. Very high-order value metrics are a good indicator that there is a potential of upselling and cross-selling to very profitable customers which will translate into more income.

Delivery Efficiency:

Delivery comes out effective as the delivery ratio is quite high and most of the orders are successfully delivered. On the other hand, there are minor issues which are late deliveries, and these need to be looked at more closely. Remedies for these delays would not only be a win for the company but may also benefit the customers in terms of satisfaction and retaining their loyalty.

Seller Performance:

Sales are highly regulated in São Paulo allowing sellers a fairly positive presence in the marketplace. Sellers who are active, usually having more than three orders, actively engage with the platform and enhance its performance as well. The number of sellers who managed to get 5-star rating was also high indicating seller performance at a relief.

Product Diversity and Performance:

The said platform has a variety of products although some of the product categories seem to perform better than others. Such successful categories as shown in the “Top Selling Categories” chart point out customers’ preferences and might help in planning the stock and the mechanisms for promoting the sales of the stock more successfully.

Recommendations

1. **Expand Geographical Coverage:** Since most customers and sellers are concentrated in major cities like São Paulo and Rio de Janeiro, there is a significant opportunity to expand to other regions. Enhancing logistics networks and targeted marketing campaigns can capture underserved areas, leading to increased market share.

2. **Improve Customer Retention with Loyalty Programs:** High-value customers, particularly those placing 10 or more orders, contribute significantly to total sales. Implementing loyalty programs for these customers can enhance retention and drive further sales.

3. **Optimize Delivery Efficiency:** While the majority of orders are delivered on time, late deliveries, though minimal, should be addressed. Improving delivery forecasting and collaborating with logistics partners to streamline operations will enhance customer satisfaction and retention.

4. **Capitalize on Popular Payment Methods:** Digital payments, especially credit and debit cards, dominate the transaction landscape. Continuing to support these payment methods while also considering emerging options like mobile payments will further streamline the customer experience.

5. **Leverage High-Value Orders for Upselling and Cross-Selling:** High-order value customers present a significant opportunity for upselling and cross-selling. Personalized marketing campaigns aimed at these customers can increase revenue from existing customers.

6. **Enhance Seller Engagement:** Encouraging sellers, especially those with more than three orders, to remain active on the platform through incentives such as lower fees or increased visibility can drive further engagement and sales. Additionally, providing feedback based on review scores can help sellers improve performance.

7. **Focus on High-Performing Product Categories:** Insights from the top-selling categories can guide stock management and marketing strategies. Emphasizing the promotion of these products, while exploring ways to boost the performance of other categories, will enhance overall profitability.

8. **Monitor Customer Feedback for Continuous Improvement:** The variation in customer reviews suggests room for improvement in product quality and customer service. Regular analysis of feedback will enable the platform to address concerns proactively and improve customer satisfaction.

References

J. Doe, "Brazilian E-Commerce Public Dataset by Olist," Olist, São Paulo, Brazil, 2023.