**1.0 Introduction**

**1.1 Background**

E-commerce has revolutionized retail, especially in Brazil, with its expanding online consumer base. It has represented a vibrant and dynamic e-commerce market. To thrive, businesses need a thorough understanding of customer behaviours and market trends.

**1.2 Problem Statement**

Brazilian e-commerce faces challenges in analysing customer journey, shipment patterns, delivery time performance, purchase patterns, payment method preferences, product popularity, and customer feedback require in-depth analysis to inform strategic decision-making. This project aims to provide valuable findings that can contribute to strategic decision-making. These insights will help businesses optimize strategies, improve operations, and drive growth in the Brazilian e-commerce landscape.

**2.0 Objectives**

This capstone project aims to provide data-driven insights into the Brazilian e-commerce landscape using a comprehensive dataset sourced from Kaggle. The primary objectives are:

1. To analyze the customer journey and identify key touchpoints.
2. To examine the geographic distribution of shipments and identify patterns.
3. To evaluate delivery time performance and identify factors influencing delays.
4. To analyze purchase patterns and identify trends.
5. To investigate payment method preferences among customers.
6. To determine product popularity and identify high-demand items.
7. To identify high-value customers and understand their purchasing behavior.
8. To analyze customer reviews and feedback to gauge satisfaction levels and identify areas for improvement.

**3.0 Methodology**

**3.1 Dataset Description**

This analysis uses a comprehensive dataset provided by Olist, Brazil's largest department store in marketplaces. The dataset includes information on 100,000 orders made between 2016 and 2018 across various marketplaces in Brazil. Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. Additionally, a geolocation dataset relates Brazilian zip codes to latitude and longitude coordinates. The data has been anonymized, with references to companies and partners replaced by names from "Game of Thrones."

**3.2 Data Collection**

The dataset was sourced from Kaggle, a platform known for providing publicly available datasets for data analysis and machine learning projects. The dataset was downloaded from the following link: [Brazilian E-commerce Public Dataset](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data). There are 7 datasets in total used in this analysis. Each .csv file contains a table with varying data types as shown below.

**The Order Item Table, sourced from the "orders\_data", consists of the following columns:**

* order\_id: A unique identifier for each order.
* order\_item\_id: A sequential number indicating the position of an item within the order.
* product\_id: A unique identifier for each product.
* seller\_id: A unique identifier for each seller.
* shipping\_limit\_date: The deadline set by the seller for handing over the order to the logistics partner.
* price: The price of the item.
* freight\_value: The freight cost associated with the item. If an order contains multiple items, the freight cost is divided among them.

**The Product Table contains the following columns:**

* product\_id: A unique identifier for each product.
* product\_category\_name: The root category of the product, originally in Portuguese.
* product\_name\_length: The length of the product name in characters.
* product\_description\_length: The length of the product description in characters.
* product\_photos\_qty: The number of photos published for the product.
* product\_weight\_g: The weight of the product measured in grams.
* product\_length\_cm: The length of the product measured in centimeters.
* product\_height\_cm: The height of the product measured in centimeters.
* product\_width\_cm: The width of the product measured in centimeters.

**The Order Table, found in the olist\_orders\_dataset, includes the following columns:**

* order\_id: a unique identifier for each order.
* customer\_id: serves as a key to the customer dataset, ensuring each order is associated with a unique customer.
* order\_status: indicates the status of the order (delivered, shipped, etc.).
* order\_purchase\_timestamp: displays the timestamp of the purchase.
* order\_approved\_at: denotes the timestamp when payment for the order was approved.
* order\_delivered\_carrier\_date: represents the timestamp when the order was handed over to the logistic partner for delivery.
* order\_delivered\_customer\_date: indicates the actual delivery date of the order to the customer.
* order\_estimated\_delivery\_date: provides the estimated delivery date communicated to the customer at the time of purchase.

**The Review Table, found in the olist\_order\_reviews\_dataset, contains the following columns:**

* review\_id: a unique identifier for each review.
* order\_id: a unique identifier for each order.
* review\_score: a rating ranging from 1 to 5 provided by the customer in a satisfaction survey.
* review\_comment\_title: the title of the comment left by the customer in Portuguese.
* review\_comment\_message: the message of the comment left by the customer in Portuguese.
* review\_creation\_date: indicates the date the satisfaction survey was sent to the customer.
* review\_answer\_timestamp: denotes the timestamp of the customer's response to the satisfaction survey.

**The Order Payment Table, found in the olist\_order\_payments\_dataset, includes the following columns:**

* order\_id: a unique identifier for each order.
* payment\_sequential: indicates the sequence number for payments made by a customer for an order.
* payment\_type: specifies the method of payment selected by the customer.
* payment\_installments: denotes the number of installments chosen by the customer for payment.
* payment\_value: represents the value of the transaction.

**The Customer Table, found in the olist\_customers\_dataset, includes the following columns:**

* customer\_id: serves as a reference key to the orders dataset, with each order having a unique customer\_id.
* customer\_unique\_id: provides a distinct identifier for each customer. This ensures that repeat purchases made by the same customer can be identified within the dataset. Otherwise, each order would appear associated with a different customer.
* customer\_zip\_code\_prefix: denotes the first five digits of the customer's zip code.
* customer\_city: specifies the name of the customer's city.
* customer\_state: indicates the state where the customer is located.

**The Geolocation Table, found in the olist\_geolocation\_dataset, includes the following columns:**

* geolocation\_zip\_code: denotes the first five digits of zip code
* geolocation\_lat: denotes latitude coordinate
* geolocation\_lng: denotes longitude coordinate
* geolocation\_city: denotes city name
* geolocation\_state: denotes state

**3.3 Data Processing and Cleaning**

1. **Loading data:** The datasets were loaded into a Jupyter notebook using Python libraries such as Pandas.
2. **Data integration:** All seven CSV files were integrated based on common keys to form a unified dataset.
3. **Handling null, inconsistent and duplicate values:** They were identified and handled appropriately, either through imputation or exclusion, depending on the context and importance of the data.
4. **Data Transformation:** Data types were converted as needed (e.g., date columns were converted to datetime objects), and additional columns were created to facilitate analysis.
5. **Data analysis and visualization:** Data is extracted and explored to find insights.

All cleaned datasets are then merged together. There will be two separate final datasets used in this project. This is to ensure faster and more efficient loading, as well as simplifying the analysis. The two datasets are:

* Df1\_final: Used to explore geographic location data
* Df2: Focuses on exploring orders, delivery performance, customer preferences, purchase patterns and etc.

**4.0 Findings and Recommendations**

**4.1. Geographic Distribution of Shipment**

The analysis revealed several significant insights into the Brazilian e-commerce market.

**A map of south america with purple dots

Description automatically generated4.1.1 Market Distribution**

In terms of geographic distribution, the majority of orders were found to be concentrated in the Southwest region of Brazil, particularly around the latitudes -22 to -23 and longitudes -43 to -46. This area emerged as a major hub for e-commerce activities.

It is highly recommended to concentrate marketing efforts and logistics improvements in the North and West region. This will help maximize reach and efficiency in areas with high order volumes and revenue.

**4.1.2 Revenue and Customer Distribution**

**A map with purple dots

Description automatically generated**A map of a country with dots

Description automatically generatedThree primary hotspots for high revenue and high customer concentration were identified at latitudes -23, longitude -46; latitude -22, longitude -43; and latitude -19, longitude -43, indicating these regions as key markets for e-commerce operations. Targeted marketing campaigns can be developed for these regions to leverage their high purchasing power and customer concentration.

**4.2 Delivery-time forecast exploration**

In exploring delivery time performance, the dataset was categorized into early, on-time, and delayed deliveries.

**4.2.1 Early delivery category**

The findings indicated that most orders in the early delivery category were received by customers within a 7 to 14 day timeframe. Accuracy of delivery time estimates should be enhanced to reduce disparity between estimated and actual delivery times.**A graph with numbers and a bar

Description automatically generated with medium confidence**

**4.2.2 Delay delivery category**

In the delayed delivery category, most orders were delayed by 1-3 days.

**A graph with numbers and a bar

Description automatically generated with medium confidence**

The overall delivery accuracy showed that:

* 91% of shipments arrived ahead of schedule
* 1% arrived exactly on the estimated delivery date
* 8% were delayed.

While the high percentage of early deliveries highlights inaccuracies in the estimated delivery tool, it also indicates a general efficiency in logistics operations.

**4.2.3 Geographic Distribution of Late Shipments**

Further analysis of the geographic distribution of late shipments identified several regions with high rates of late deliveries. These regions included:

* latitudes -23, longitude -46
* latitude -22, longitude -43
* latitude -19, longitude -44
* latitude -30, longitude -51
* latitude -12, longitude -38
* latitude -8, longitude -35
* latitude -3, longitude -38.

These findings suggest that certain areas may require improvements in logistics and delivery processes to reduce delays. Delay prone areas needs to have better logistics by collaborating with reliable delivery partners or investing in better infrastructure.

**A map of the south america

Description automatically generated**

**4.3 Purchase Pattern Analysis**

The purchase pattern analysis revealed that:

* Most purchases were made between 1000 to 2200 hours
* Monday and Tuesday showing the highest number of purchases, while Saturday had the lowest.
* May and August recorded the highest purchase counts, whereas September and October had the lowest.

These trends provide valuable insights into customer buying behaviours and peak purchasing periods. With this data, promotions and discounts can be offered during these periods to capitalize on high customer activity. Seasonal campaigns can be developed in May and August, which have the highest purchase volumes, while planning for lower activity in September and October.

**4.4 Payment Method Preferences**

The analysis showed that:

* Credit cards were the most preferred payment method, with the highest average order value (AOV) of $125.8.
* Boleto, a popular Brazilian payment method, followed with an AOV of $104.1, while debit cards and vouchers were the least preferred.

This preference for credit cards suggests that customers tend to use this method for larger purchases, indicating trust and convenience associated with credit card transactions. Customers can be incentivize to use credit cards by offering exclusive discounts, loyalty points, or cashback offers.

**4.5 Product Popularity**

Product popularity was assessed by analyzing customer reviews, which revealed that the furniture decor category was the most popular, receiving the highest number of reviews. This indicates a strong customer interest in purchasing furniture on the platform.

**4.6 High- Value Customers**

The identification of high-value customers through analysis highlighted a particular customer, identified as Oa0a92112bd4c708ca5fde585afaa872, as the most valued, generating close to $14,000 in revenue over the last three years. This finding underscores the importance of recognizing and catering to high-value customers to sustain business growth. Personalized loyalty programs can be developed and offering exclusive deals for these customers.

**4.7 Review and Feedback Analysis**

Analysis revealed common customer complaints related to delivery issues, product quality, and store performance. The sentiment analysis and word cloud visualization provided a clear picture of areas where improvements are needed to enhance customer satisfaction.

**5.0 Conclusion**

This capstone project analyzed the Brazilian e-commerce landscape using a dataset from Olist, focusing on geographic distribution, delivery time performance, purchase patterns, payment preferences, product popularity, high-value customers, and customer reviews.

Key insights include:

* **Geographic Distribution:** Most orders and revenue are concentrated in Southwest Brazil, with hotspots in major urban areas, guiding where businesses can focus their efforts.
* **Delivery Time Performance:** While 91% of shipments arrive early, delivery estimate accuracy needs improvement. Identifying regions with frequent delays can help optimize logistics.
* **Purchase Patterns:** Customer activity peaks between 10 AM and 10 PM, particularly on Mondays and Tuesdays, with May and August being the busiest months. This helps in planning inventory and marketing strategies.
* **Payment Preferences:** Credit cards are the most preferred method, associated with higher average order values, suggesting businesses should encourage credit card payments to boost revenue.
* **Product Popularity:** Furniture decor is the top category, indicating a strong market for these products.
* **High-Value Customers:** Identifying top customers who significantly contribute to revenue can help tailor loyalty programs and personalized marketing.
* **Customer Reviews:** Common complaints about delivery and product quality highlight areas for improvement to enhance customer satisfaction.

Overall, this project demonstrates how data-driven insights can inform strategic decisions and drive growth in the Brazilian e-commerce market.