Case Study: Analysis of Brazilian E-commerce Operations (2016-2018)

1. Executive Summary

This project presents a comprehensive analysis of a large public dataset from a prominent e-commerce platform operating in Brazil. The dataset encompasses approximately 100,000 orders placed between 2016 and 2018, providing a multi-dimensional view of the company's operations.

The core objective of this case study is to delve into this rich dataset to uncover actionable insights. By examining various facets such as sales trends, logistical efficiency, customer satisfaction, and regional performance, we aim to understand the key drivers of success and identify potential areas for improvement. This analysis is designed to answer critical business questions and provide data-driven recommendations that could enhance the customer experience and optimize operational processes.

2. Business Context

The subject of this case study is a major online retailer that has established itself as a preferred shopping destination by focusing on delivering outstanding value, innovation, and an exceptional guest experience. This project specifically analyzes its operations within the dynamic and growing Brazilian market.

The dataset offers a unique opportunity to explore the intricacies of an e-commerce business at scale. It allows for a deep dive into order processing, pricing strategies, payment and shipping efficiency, customer demographics, product characteristics, and customer satisfaction levels.

3. Key Areas of Analysis & Potential Insights

This project will explore several key business questions, leveraging the provided data to build a holistic understanding of the platform's performance.

- **Sales and Revenue Analysis:**
    - What are the trends in order volume and revenue over time (daily, weekly, monthly)?
    - Which product categories generate the most revenue?
    - How do price and freight_value contribute to the total payment_value?
- **Logistics and Operational Efficiency:**
    - How accurate are the estimated delivery dates? We can calculate the delivery delta:

$$\Delta delivery = order\_delivered\_customer\_date - order\_estimated\_delivery\_date$$

    - What is the average time from purchase to shipment and from shipment to delivery?
    - Are there geographic regions (states or cities) with consistently longer shipping times?
- **Customer Behavior and Satisfaction:**
    - What is the geographic distribution of the customer base?
    - Is there a correlation between delivery delays and low review_score?
    - What are the most common payment methods (payment_type) and installment plans (payment_installments)?
    - Can we identify common themes in positive or negative reviews through analysis of review_comment_message?
- **Product and Seller Performance:**
    - Which sellers have the highest sales volume or best review scores?
    - Do product attributes (e.g., product_weight_g, product_photos_qty) correlate with sales volume or customer satisfaction?
    - What is the geographic distribution of sellers versus customers?

4. Dataset Description

The analysis is based on a relational dataset comprised of 8 individual CSV files. These files are linked by key identifiers such as order_id, customer_id, product_id, and seller_id.

- customers.csv: Contains information about unique customers.
- sellers.csv: Contains information about sellers on the platform.
- orders.csv: The central file containing details about each order.
- order_items.csv: Contains details about the specific items within each order.
- payments.csv: Contains data related to payment methods for each order.
- reviews.csv: Contains customer-submitted reviews for each order.
- products.csv: Contains attributes for each unique product.
- geolocation.csv: Contains Brazilian zip code, latitude, and longitude information.

5. Data Schema

Below is a detailed description of the columns available in each CSV file.

customers.csv

Feature	Description
customer_id	ID of the consumer who made the purchase.
customer_unique_id	Unique ID of the consumer.
customer_zip_code_prefix	Zip Code of consumer’s location.
customer_city	Name of the City from where order is made.
customer_state	State Code from where order is made (e.g., SP for São Paulo).

sellers.csv

Feature	Description
seller_id	Unique ID of the seller registered.
seller_zip_code_prefix	Zip Code of the seller’s location.
seller_city	Name of the City of the seller.
seller_state	State Code (e.g., SP for São Paulo).

order_items.csv

Feature	Description
order_id	A Unique ID of the order made.
order_item_id	A Unique ID for each item within an order.
product_id	A Unique ID for each product.
seller_id	Unique ID of the seller.
shipping_limit_date	The date before which the product must be shipped.
price	Actual price of the product.
freight_value	Shipping cost for the product.

geolocation.csv

Feature	Description
geolocation_zip_code_prefix	First 5 digits of Zip Code.
geolocation_lat	Latitude coordinate.

geolocation_lng	Longitude coordinate.
geolocation_city	City name.
geolocation_state	State name.

payments.csv

Feature	Description
order_id	A Unique ID of the order.
payment_sequential	Sequence of payments for a single order (e.g., for multiple payment methods).
payment_type	Mode of payment used (e.g., Credit Card, Boleto).
payment_installments	Number of installments for the purchase.
payment_value	Total amount paid for the order.

orders.csv

Feature	Description
order_id	A Unique ID of the order.
customer_id	ID of the consumer who made the purchase.
order_status	Status of the order (e.g., delivered, shipped).
order_purchase_timestamp	Timestamp of the purchase.
order_delivered_carrier_date	Timestamp of when the order was handed to the carrier.
order_delivered_customer_date	Timestamp of when the customer received the order.
order_estimated_delivery_date	The estimated delivery date for the order.

reviews.csv

Feature	Description
review_id	ID of the review.
order_id	A Unique ID of the order being reviewed.
review_score	Score from 1 to 5 given by the customer.

review_comment_title	Title of the review comment.
review_comment_message	The text content of the review.
review_creation_date	Timestamp of when the review was created.
review_answer_timestamp	Timestamp of when the review was responded to.

products.csv

Feature	**Description**
product_id	A Unique identifier for the product.
product_category_name	Name of the product category.
product_name_lenght	Character length of the product's name.
product_description_lenght	Character length of the product's description.
product_photos_qty	Number of available photos for the product.
product_weight_g	Weight of the product in grams.
product_length_cm	Length of the product in centimeters.
product_height_cm	Height of the product in centimeters.
product_width_cm	Width of the product in centimeters.