Oilst Marketplace Analysis

BRAZILIAN E-COMMERCE

Background

Olist is a Brazilian e-commerce site that connects multiple marketplaces through a single channel. They provide end to end comprehensive support for their merchants who are able to sell their products through the Olist Stores and ship them directly to customers. Once a customer places an order, the seller is notified. Upon receiving the order, or passing the estimated delivery date, the customer is emailed a satisfaction survey.

Data Collection

The dataset was provided by Olist, who uses their logistics partners to track order information such as order status, price, payment and freight performance to customer location, product attributes, and reviews written by customers.

This is real commercial data that has been anonymised.

Data Source

This data was released by Olist and made available on Kaggle. It has information on 100k orders placed at Olist with details on customers, orders, sellers, payments, shipping, reviews, and more.

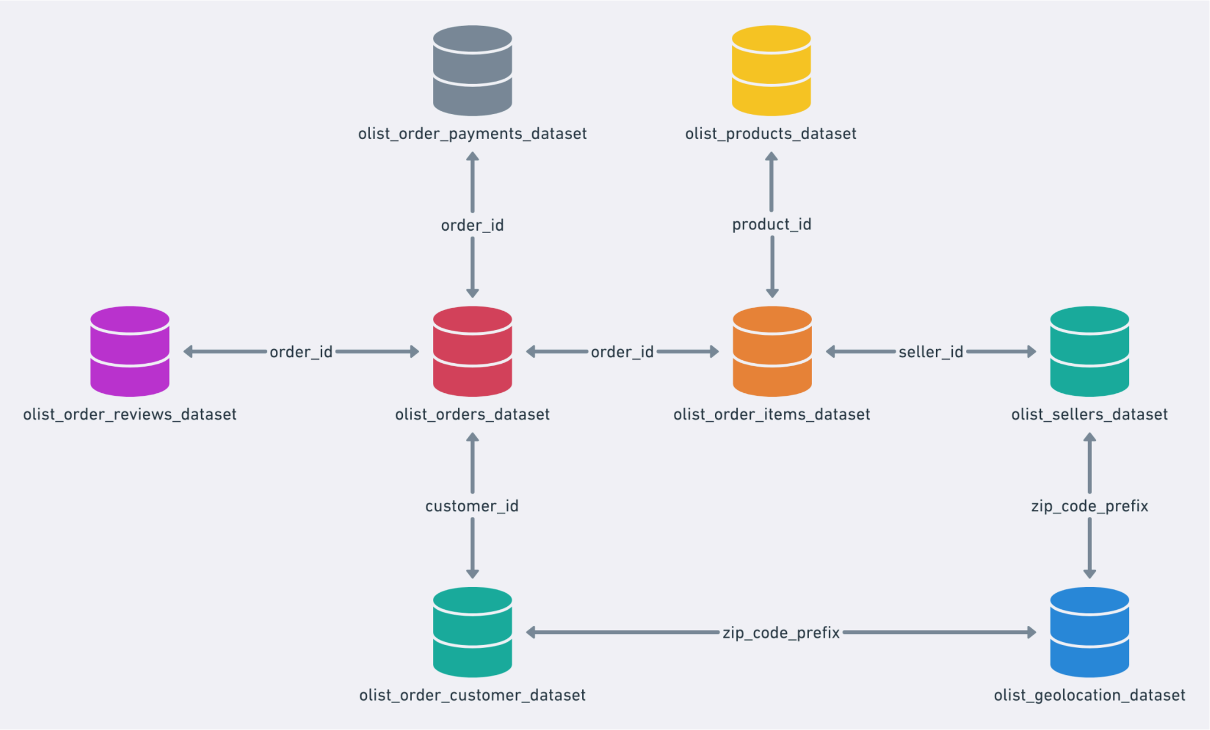
The dataset includes the following files:

1. **olist_customers_dataset.csv**
   Contains customer details including customer id, zip code, city, and state.
2. **olist_geolocation_dataset.csv**
   Contains geolocation details outlining a zip code's latitude, longitude, city, and state.
3. **olist_order_items_dataset.csv**
   Contains order details including order id, order item id, product id, seller id, shipping limit date, price, and freight value.
4. **olist_order_payments_dataset.csv**
   Contains payment details, including order id, payment sequential, payment type, payment installments, and payment value.
5. **olist_order_reviews_dataset.csv**
   Contains reviews details including review id, order id, review score, comment title, comment message, creation date, and answer timestamp.

6. **olist_orders_dataset.csv**
   Contains order and delivery details including the order id, customer id, order status, time purchased, time order approved, time delivered per carrier and per customer, and estimated delivery date.
7. **olist_products_dataset.csv**
   Contains product and listing details including the product's id, category name, name length, description length, photo quantity, weight in grams, and length in centimeters.
8. **olist_sellers_dataset.csv**
   Contains seller details including the seller's id, zip code, city, and state.
9. **product_category_name_translated.csv**
   Contains the product category name and its translation in English.

Data can be accessed here: Brazilian E-Commerce Public Dataset by Olist

See the data schema for the set below:



Geolocation data for mapping was accessed through a public Github project with updates 8 months ago at the time of access.

The dataset includes the following files:

1. **br_states.json**
   Contains a GeoJSON format of all Brazilian state parameters.

Data can be accessed here: Brazil States GeoJSON

Data Profile

Cleaning Summary

1. **olist_geolocation_dataset.csv**
   - Removed 261,831 full duplicates.
2. **olist_order_reviews_dataset.csv**
   - The `review_comment_title` and `review_comment_message` columns were changed from data type object to "string".
3. **olist_orders_dataset.csv**
   - The following columns were changed to datetime64[ns] data type: `order_purchase_timestamp`, `order_approved_at`, `order_delivered_carrier_date`, `order_delivered_customer_date`, `order_estimated_delivery_date`.
4. **olist_product_dataset.csv**
   - Typos in `product_name_lenght` and `product_description_lenght` were fixed, changing lenght to length
   - The `product_category_name` column was changed to data type "string".
5. **all_merged.pkl**
   - `Product_category` was grouped into new categories, condensing the existing ones *(for example, all fashion related categories were placed into a new category called 'Fashion')*

Column Derivations

1. **order_total:** an order id's total `price` and `freight_value` for all items in order
2. **purchase_to_approved_hrs, purchase_to_delivered_hrs:** hours from customer purchase to seller approval; hours from customer purchase to order delivered
3. **num_orders:** number of orders the customer has placed
4. **return_customer:** 0 for customers with only 1 order, 1 for customers with more than 1 order
5. **num_orders_w_seller:** number of order the customer has placed with the seller
6. **return_to_seller:** 0 for customer has placed only 1 order with the seller, 1 for customer has placed more than 1 order with seller
7. **num_items:** number of items per order (aka max `order_item_id`)
8. **on_time_flag:** order arrived 'on time', 'early', or 'late' using delivery & expected delivery dates
9. **num_orders_received:** number of orders received by a seller
10. **total_revenue:** sum of the prices of products sold by a seller
11. **num_products:** number of the unique products sold by a seller
12. **revenue_flag:** assigns 'Low', 'Medium', or 'High revenue' based on seller's `total_revenue`
13. **frequency_flag:** assigns 'New', 'Low frequency', or 'High frequency customer' based on customer's `num_orders`
14. **cluster:** assigns color of cluster that each data point belongs to ('beige', 'pink', 'light purple', 'purple', or 'dark purple')

Descriptive Analysis

	price *(of product)*	freight_value	payment_sequential	payment_installments	payment_value	review_score
Mean	$120.65	$19.99	1.09	2.85	$154.10	4.09
Standard Deviation	$183.63	$15.80	0.71	2.69	$217.49	1.34
Min	$0.85	$0.00	1	0	$0.00	1
Max	$6,735.00	$409.68	29	24	$13,664.08	5

	product_photos_qty	product_weight_g				
Mean	2.19	2,276.47				
Standard Deviation	1.74	4,282.04				
Min	1	0				
Max	20	40,425				

Why This Data

I chose to study a dataset from an e-commerce site because I'm passionate about exploring how businesses operate and identifying areas where data can shed light on performance and highlight opportunities for improvement. With my experience working with small businesses, I'm particularly interested in understanding the interconnections between various business components. This dataset offers a unique, real-life perspective on critical aspects such as cost-to-profit analysis, customer behavior, product trends, and location-based purchasing patterns over time. By analyzing these elements, I aim to uncover valuable insights that can drive business optimization and strategy.

Data Limitations

The data sets contains information on 100k orders from 2016 - 2018, providing insights from a two-year period early in Olist's history. Olist, originally a brick and mortar store called Solidarium, became a marketplace in 2011. In 2015, it officially became Olist and became an official store on major marketplaces in 2016 ("About Us"). While the dataset offers valuable insights, it primarily reflects the state of Brazilian e-commerce during that time and provides a perspective from a marketplace that was still relatively new to this status.

Ethical Considerations

As the dataset originates from an e-commerce platform, it is crucial to prioritize data privacy and confidentiality for both buyers and sellers. All personally identifiable information was removed, with only unique identification numbers retained in the publicly available dataset. When analyzing the data, it is essential to ensure it is used solely for the purpose for which it was collected, namely, to generate insights that support business objectives. Furthermore, careful consideration must be given to the potential impact of these insights on users and stakeholders, ensuring that the findings do not inadvertently reinforce inequality or contribute to unfair business practices.

Questions to Explore

- What are the types of products being sold? Are there any gaps in the market or areas that are very saturated?
- What is the profile of items being bought more versus those being bought less?
    - Are they from a “top seller” (a more frequently bought from store)?
    - Are they within a specific price range?
    - Do they have more pictures in their listing?
    - Do they have longer listing descriptions or titles?
- Who are our customers?
    - Where do they live?
    - How often do they buy?
    - How many items do they order on average?
    - Do certain areas buy certain products more/less than others?
- What is being reviewed?
    - How many reviews do we have and what is the distribution of scores?
    - Are certain products being reviewed more than others?
    - When do customers leave a comment with their review?
    - How quickly are reviews being answered by a seller?

Works Cited

“Olist Is Empowering Small Merchants to Sell Online.” *Valorcapitalgroup.com*, 2020, www.valorcapitalgroup.com/case-studies/olist-redesigned-the-marketplace-business-model-to-fit-the-realities-of-ecommerce-in-brazil. Accessed 7 Feb. 2025.

“About Us.” *Olist*, 2015, olist.com/sobre-nos/?_gl=1. Accessed 9 Feb. 2025.

“Brazilian E-Commerce Public Dataset by Olist.” *Www.kaggle.com*, www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_order_items_dataset.csv.

giuliano-macedo. (2023). GitHub - giuliano-macedo/geodata-br-states: GeoJson formatted Brazilian states perimeters. GitHub. https://github.com/giuliano-macedo/geodata-br-states/tree/main