Brazilian E-Commerce Public Dataset by Olist

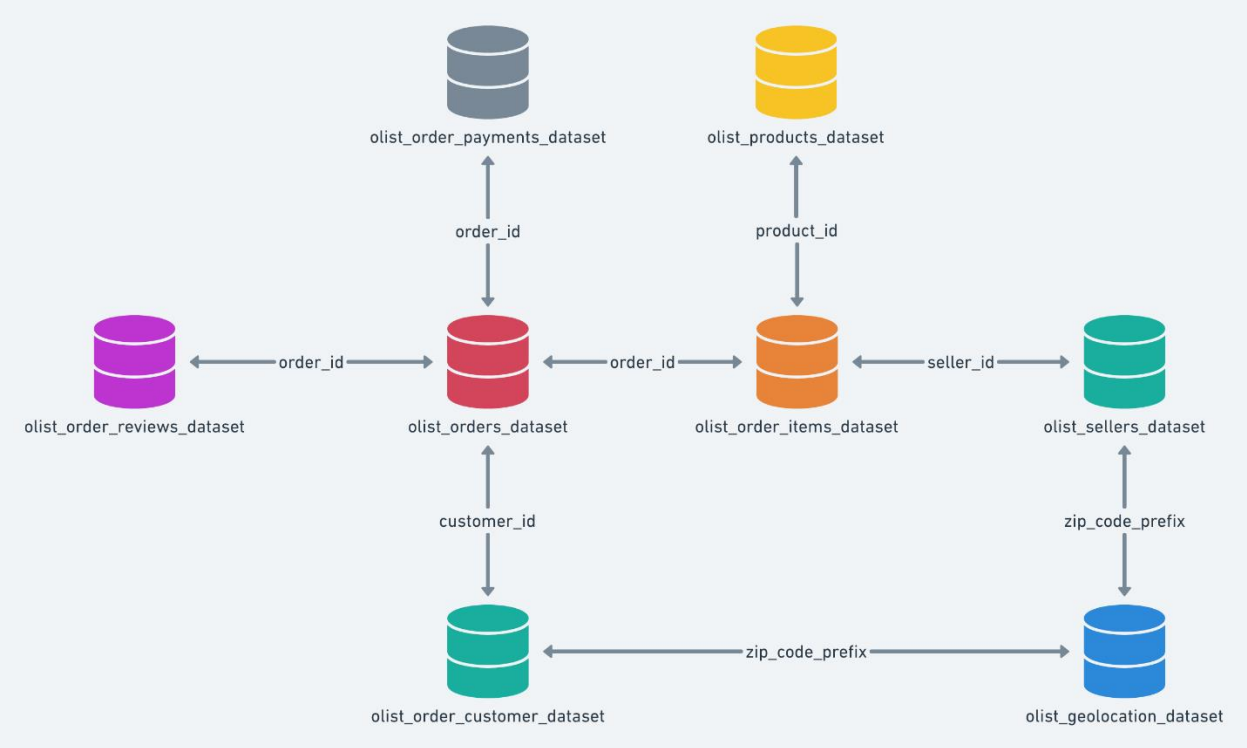
Olist is a Brazilian e-commerce platform and marketplace that connects small and medium-sized businesses (SMBs) with various e-commerce channels to help them sell their products online. It essentially acts as an intermediary between sellers and larger e-commerce platforms, providing a range of services to facilitate online sales for SMBs. The reason I chose this database is due to my personal interest and background in the fields of sales and marketing. This dataset enables me to put into practice my analytical skills within these domains.

1. Data sourcing and collection

The dataset used for this project can be found on Kaggle and was provided by Olist. The dataset has information on 100k orders placed between 2016 to 2018 through multiple marketplaces in Brazil. Because this data contains real company information, it has undergone anonymization to safeguard sensitive details and protect privacy. Companies and partners' names have been replaced by Game of Thrones references.

2. Data contents

The dataset consists of nine files, see data ERD below:



Source

3. **Tables in the dataset**

olist_order_items_dataset.csv

This table includes data about the items purchased within each order.

Colum	Description
order_id	Order unique identifier
order_item_id	Sequential number of items included in the same order
product_id	Product unique identifier
seller_id	Seller unique identifier
shipping_limit_date	Seller shipping date limit for handing the order over to the logistics partner
price	Item price
freight_value	Item freight value item (if an order has more than one item the freight value is split between items)

olist_orders_dataset.csv

This table includes data about orders.

Colum	Description
order_id	Unique order identifier
customer_id	Unique customer identifier
order_status	Order status
order_purchase_timestamp	Timestamp of when the order was placed
order_approved_at	Timestamp of when the payment was approved
order_delivered_carrier_date	Timestamp of when the order was handed over to the logistics partner
order_delivered_customer_date	Actual delivery date
order_estimated_delivery_date	Estimated delivery date

olist_customers_dataset.csv

This table has information about the customer and its location.

Colum	Description
customer_id	Each order has a unique customer id
customer_unique_id	Unique customer identifier
customer_zip_code_prefix	The first five digits of the customer's zip code
customer_city	Customer’s city
customer_state	Customer’s state

olist_geolocation_dataset.csv

This table has information on Brazilian zip codes and their lat/long coordinates.

Colum	Description
geolocation_zip_code_prefix	First five digits of the customer's zip code
geolocation_lat	Latitude
geolocation_lng	Longitude
geolocation_city	City
geolocation_state	State

olist_order_payments_dataset.csv

This table includes data about payment.

Colum	Description
order_id	Order unique identifier
payment_sequential	Number of different payment methods
payment_type	Payment method
payment_installments	Number of payment installments
payment_value	Order amount

olist_order_reviews_dataset.csv

This table includes data on reviews made by the customers.

Colum	Description
review_id	Unique review identifier
order_id	Unique order identifier
review_score	Rating from 1 to 5 given by the customer on a satisfaction survey
review_comment_title	Review title, in Portuguese
review_comment_message	Review, in Portuguese.
review_creation_date	Date when the survey was sent to the customer
review_answer_timestamp	Date when the survey was filled out and submitted by the customer

olist_products_dataset.csv

This dataset includes data about the products sold by Olist.

Colum	Description
product_id	Unique product identifier
product_category_name	Product category
product_name_lenght	Number of characters in the product name
product_description_lenght	Number of characters in the product description
product_photos_qty	Number of product photos published
product_weight_g	Product weight in grams
product_length_cm	Product length in centimeters
product_height_cm	Product height in centimeters
product_width_cm	Product width in centimeters

olist_sellers_dataset.csv

This dataset includes data about the sellers who fulfilled orders made at Olist.

Colum	Description
seller_id	Unique seller identifier
seller_zip_code_prefix	The first five digits of the seller's zip code
seller_city	Seller's city
seller_state	Seller's state

product_category_name_translation.csv

Translates the product_category_name to English.

Colum	Description
product_category_name	category name in Portuguese
product_category_name_english	category name in English

4. Data relevance

The dataset meets the requirements for this project as it is open-source and comes directly from Olist. It also contains a geospatial component and meets the size and variable requirements. The data relevance of each table will be evaluated later once the hypothesis has been defined.

5. Limitations

The dataset contains detailed geographical data related to customers, including zip codes, cities, and states, but it lacks demographic information and behavioral information. This data would allow a more detailed analysis and forecast. Also, the dataset only contains information from 100k orders placed. This raises questions on how the data was selected and how representative it is of the entire dataset. If the dataset is not representative, the analysis can't be used for decision-making.

6. Ethics

This information was gathered and provided by Olist and was anonymized to ensure that sensitive information is protected. It appears there are no ethical concerns with this dataset.

7. Wrangling steps

Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
product_category_name			Irrelevant column – category name in Portuguese
	freight_value to item_shipping_cost		Ease of comprehension
	order_item_id to order_item_count		Ease of comprehension
	order_purchase_timestamp to timestamp_oder		Ease of comprehension
	order_approved_at to timestamp_payment_approval		Ease of comprehension
	order_delivered_carrier_date to timestamp_order_dispatched		Ease of comprehension
	order_delivered_customer_date to order_actual_delivery_date		Ease of comprehension
	order_estimated_delivery_date to order_estimated_delivery_date		Ease of comprehension
	payment_sequential to payment_method_count		Ease of comprehension
	review_comment_title to review_title		Ease of comprehension
	review_comment_message to review_text		Ease of comprehension
	review_creation_date to survey_sent_date		Ease of comprehension
	review_answer_timestamp to survey_submision_date		Ease of comprehension
	payment_value to order_amount		Ease of comprehension
	Price to item_price		Ease of comprehension

Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
		seller_zip_code_prefix and customer_zip_code_prefix to strings	Zip codes are categorical values
		review_title and review_text to strings	This column had mixed types due to missing values, so it needed to be changed to a string

8. Consistency checks

Column	Missing values	Missing values treatment	Duplicates	Outliers
product_weight_g product_length_cm product_height_cm product_width_cm	There is 1 missing item in these columns	No action was taken		
timestamp_payment_approval	14 missing values	No action was taken		
timestamp_order_dispatched	1,195 missing values	No action was taken		
order_actual_delivery_date	2,400 missing values	No action was taken		
review_title	101,808 missing values	No action was taken		
review_text	66,703 missing values	No action was taken		
			There are no duplicates	
				There are no outliers

9. Data Profile of Cleaned Data

39 variables / 115,609 records

Variable	Median	Time -variant/ Invariant	Structured/ Unstructured	Qualitative/ Quantitative	Nominal/Ordinal Discrete/Continuous
order_id	n/a	Invariant	Structured	Qualitative	Nominal
order_item_count	1.19	Invariant	Structured	Quantitative	Discrete
product_id	n/a	Invariant	Structured	Qualitative	Nominal
seller_id	n/a	Invariant	Structured	Qualitative	Nominal
shipping_limit_date	n/a	Invariant	Structured	Quantitative	Continuous
item_price	120.62	Variant	Structured	Quantitative	Continuous
item_shipping_cost	20.06	Variant	Structured	Quantitative	Continuous
product_name_lenght	48.77	Variant	Structured	Quantitative	Continuous
product_description_lenght	785.81	Variant	Structured	Quantitative	Continuous
product_photos_qty	2.21	Variant	Structured	Quantitative	Continuous
product_weight_g	2,113.91	Invariant	Structured	Quantitative	Continuous
product_length_cm	30.31	Invariant	Structured	Quantitative	Continuous
product_height_cm	16.64	Invariant	Structured	Quantitative	Continuous
product_width_cm	23.11	Invariant	Structured	Quantitative	Continuous
product_category_name_english	n/a	Invariant	Structured	Qualitative	Nominal
seller_zip_code_prefix	n/a	Variant	Structured	Qualitative	Nominal
seller_city	n/a	Variant	Structured	Qualitative	Nominal
seller_state	n/a	Variant	Structured	Qualitative	Nominal
customer_id	n/a	Invariant	Structured	Qualitative	Nominal
order_status	n/a	Invariant	Structured	Qualitative	Nominal
timestamp_order	n/a	Invariant	Structured	Quantitative	Continuous

Variable	Median	Time - variant/ Invariant	Structured/ Unstructured	Qualitative/ Quantitative	Nominal/Ordinal Discrete/Continuous
timestamp_payment_approval	n/a	Invariant	Structured	Quantitative	Continuous
timestamp_order_dispatched	n/a	Invariant	Structured	Quantitative	Continuous
order_actual_delivery_date	n/a	Invariant	Structured	Quantitative	Continuous
order_estimated_delivery_date	n/a	Invariant	Structured	Quantitative	Continuous
payment_method_count	1.09	Invariant	Structured	Quantitative	Discrete
payment_type	n/a	Invariant	Structured	Qualitative	Nominal
payment_installments	2.95	Invariant	Structured	Quantitative	Discrete
order_amount	172.39	Invariant	Structured	Quantitative	Continuous
customer_unique_id	n/a	Invariant	Structured	Qualitative	Nominal
customer_zip_code_prefix	n/a	Variant	Structured	Qualitative	Nominal
customer_city	n/a	Variant	Structured	Qualitative	Nominal
customer_state	n/a	Variant	Structured	Qualitative	Nominal
review_id	n/a	Invariant	Structured	Qualitative	Nominal
review_score	4.03	Invariant	Structured	Quantitative	Discrete
review_title	n/a	Invariant	Unstructured	Qualitative	Nominal
review_text	n/a	Invariant	Unstructured	Qualitative	Nominal
survey_sent_date	n/a	Invariant	Structured	Qualitative	Continuous
survey_submision_date	n/a	Invariant	Structured	Qualitative	Continuous

10. Questions to Explore

- Is there a relationship between delivery time and customer ratings?
- Is there a relationship between the number of pictures published per product and sales?
- Is there a relationship between the length of the product description and sales?
- Is there a relationship between price and the number of payment installments?
- What is the busiest day of the week?
- Are customers loyal? How many users have placed more than one order?
- Is there a relationship between loyalty and order amount?