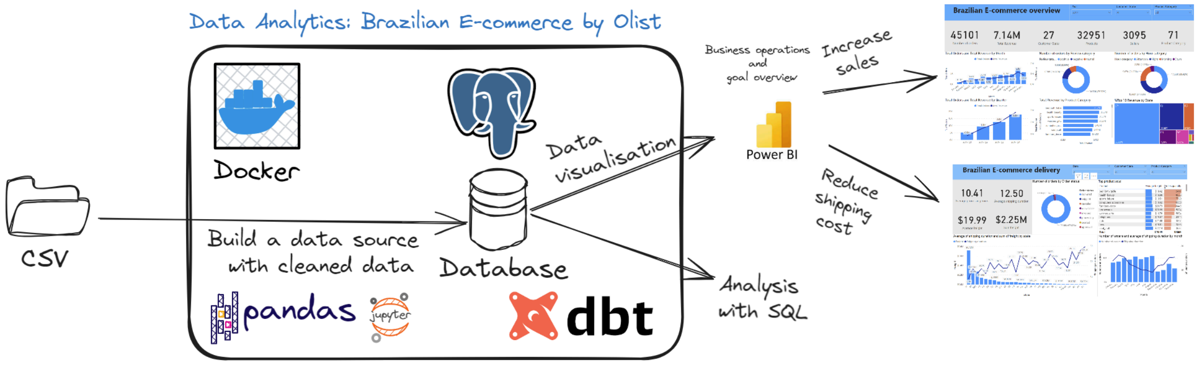
Name: Nguyen Manh Hung
Email: nguyenmanhhung04051998@gmail.com
Github: https://github.com/nmh4598/data-analytics_brazilian-ecommerce_2020

Solution overview:


Data Analytics: Brazilian E-commerce by Olist
Docker
CSV
Build a data source with cleaned data
Database
pandas
jupyter
dbt
Data visualisation
Business operations and goal overview
Power BI
Increase sales
Reduce shipping cost
Analysis with SQL
Brazilian E-commerce overview
45101
7.14M
27
32951
3095
71
Brazilian E-commerce delivery
10.41
12.50
$19.99
$2.25M

Section 1. Data exploration: explore, clean, describe dataset

The data I'll analyze comes from Olist, a Brazilian e-commerce platform that connects small businesses to larger product marketplaces. Olist published a dataset containing 99441 orders from March 2016 to August 2018. The data is anonymized, so it doesn't contain names for buyers, sellers or products. Its features allow viewing orders from multiple dimensions: from order status, price, payment, and freight performance to customer location, product attributes and finally reviews written by customers.

A. Assessment Summary

1. **Customers**
    - **Missing Value**: None.
    - **Duplicate Data**: None.
    - **Inaccurate Value**: None.
2. **Orders**
    - **Missing Value**: Three columns with missing values: order_approved_at, order_delivered_carrier_date, and order_delivered_customer_date.
    - **Duplicate Data**: None.
    - **Inaccurate Value**: None.
3. **Order Items**
    - **Missing Value**: None.
    - **Duplicate Data**: Transform order_item_id into quantity to extract unit-per-order line profile.
    - **Inaccurate Value**: Outliers in columns price and freight_value.
4. **Order Payments**
    - **Missing Value**: None.
    - **Duplicate Data**: None.
    - **Inaccurate Value**: An outlier in the column payment_value.

5. **Order Reviews**
    - **Missing Value**: Two columns with missing values: review_comment_title and review_comment_message.
    - **Duplicate Data**: None.
    - **Inaccurate Value**: None.
6. **Products**
    - **Missing Value**: Eight columns with missing values: product_category_name, product_name_lenght, product_description_lenght, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, and product_width_cm.
    - **Duplicate Data**: None.
    - **Inaccurate Value**: None.
7. **Product Category Name**
    - **Missing Value**: None.
    - **Duplicate Data**: None.
    - **Inaccurate Value**: None.
8. **Sellers**
    - **Missing Value**: None.
    - **Duplicate Data**: None.
    - **Inaccurate Value**: None.
9. **Geolocation**
    - **Missing Value**: None.
    - **Duplicate Data**: Total duplicate values: 261,831.
    - **Inaccurate Value**: None.

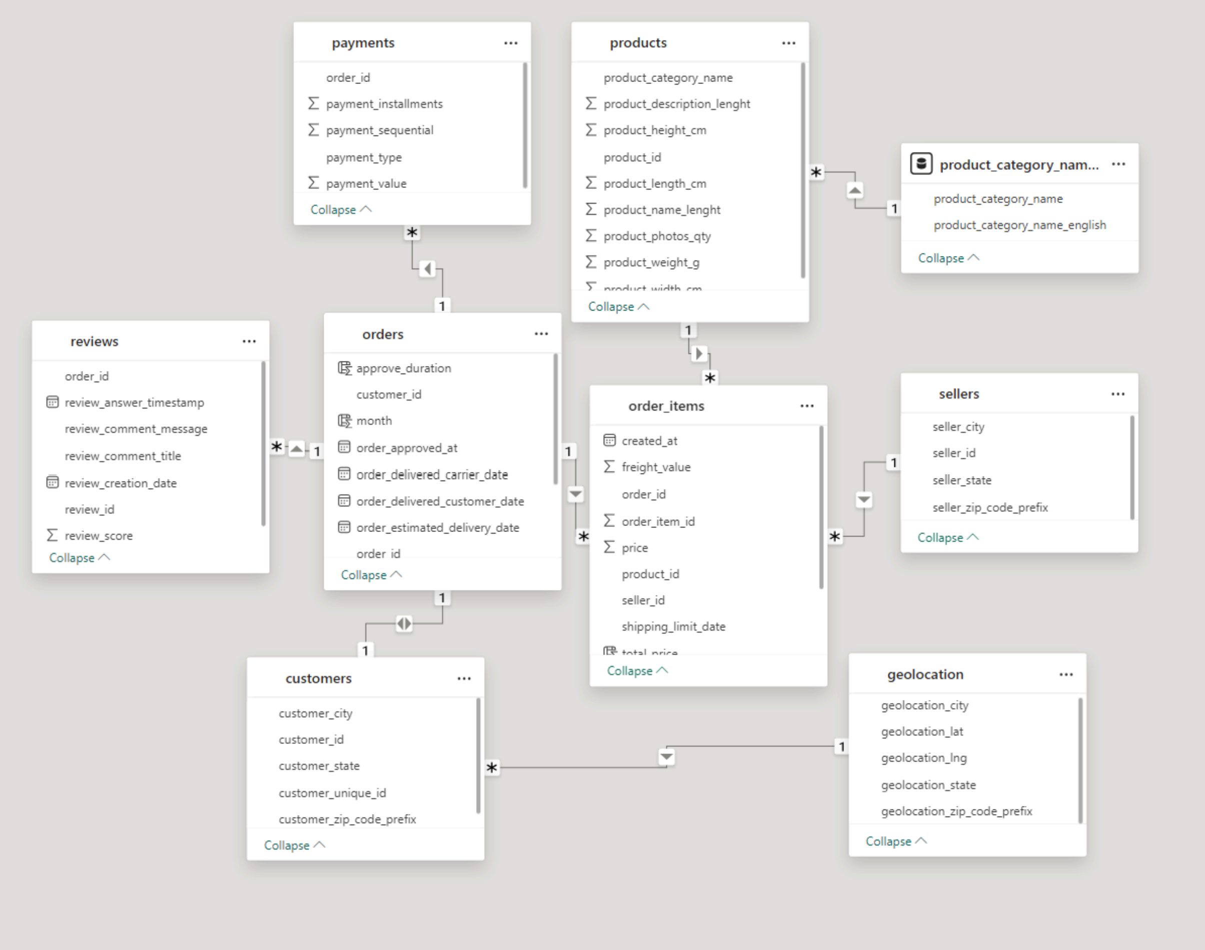
In our case, we will only clean Geolocation data with Python for customer analysis by state

B. Data Wrangling

After thorough cleaning in Python using Pandas and Numpy, the data has been imported to PostgreSQL environment using the COPY command. The data wrangling process includes:

- Handling duplicate values
- Handling data types
- Handling inconsistent data

C. Data modeling



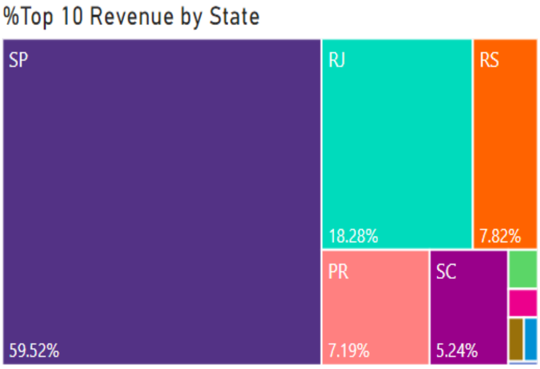
Section 2. Business Acumen

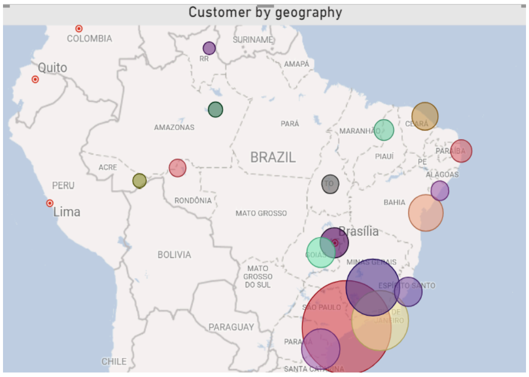
A. Analyse the company performance and deliver insights

To align with Olist’s main strategy of maximizing GMV (Gross Merchandise Volume) and optimizing spending, these are key metrics that can be used to analyze company performance and deliver insights for improvements. Each metric is explained with its rationale and how it contributes to the strategic goals:

1) % Top 10 revenue by State and Customer by geography

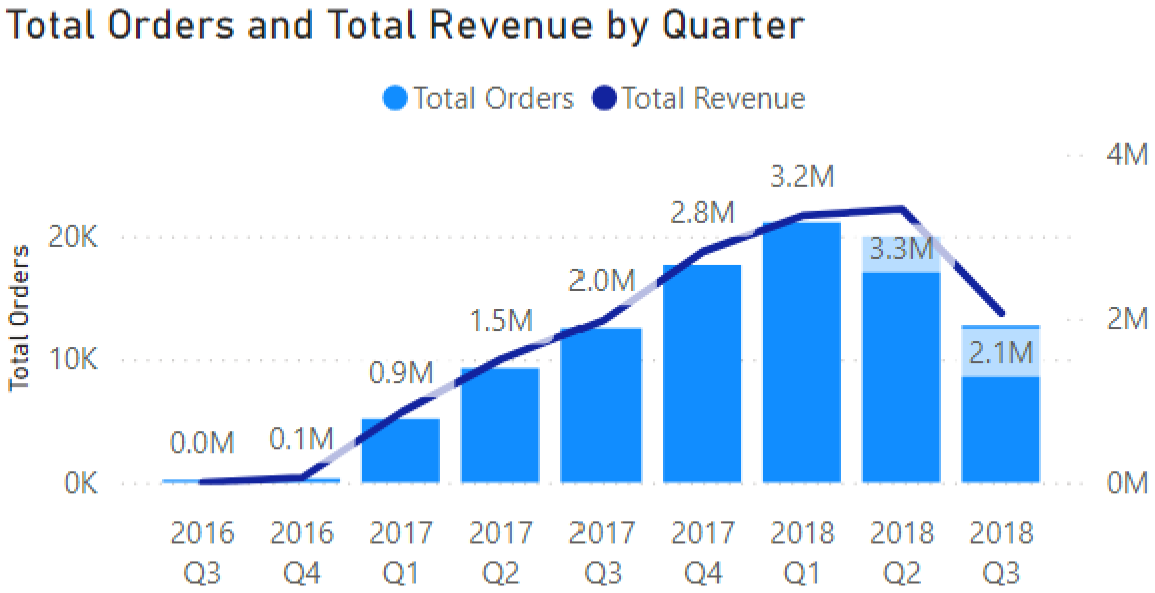
By looking at the data of these 2 charts, we can see that sao paulo city from **SP state** alone has more orders than the following 5 cities combined. This is because sao paulo is the most populous and richest state in Brazil. This indicates an opportunity for improvement in the other states. By focusing on these states, it is possible to increase the number of orders and expand the customer base.



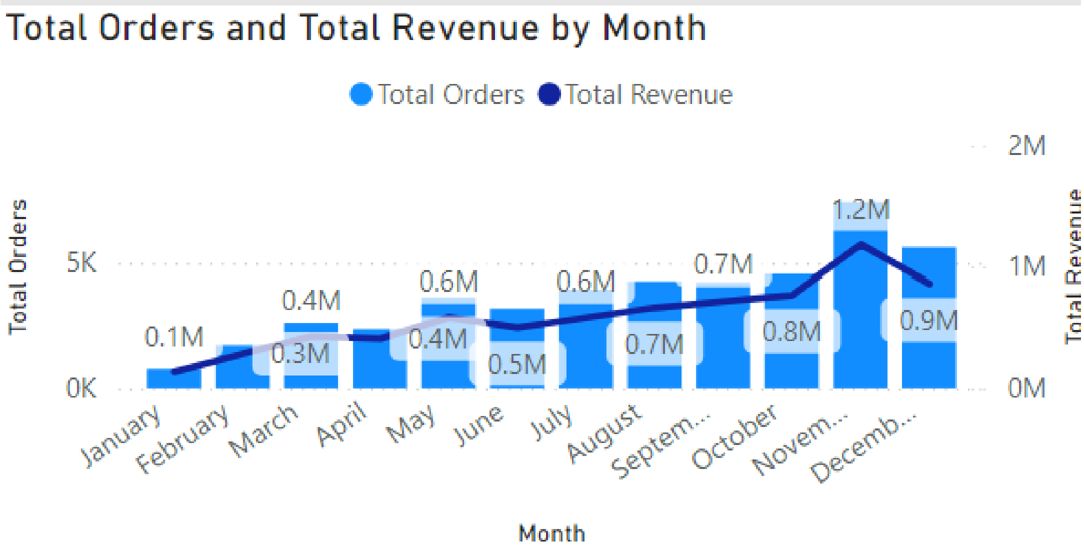


2) Total orders and total revenue by month and quarter

Given the limited dataset, drawing definitive conclusions about seasonality trends is challenging. However, the analysis and visualization in PowerBI reveal the largest total orders and total revenue were in **the first quarter of 2018**, which coincides with the Carnival season in Brazil. This finding also aligns with the previous analysis, indicating a positive correlation between the population of a state and its order count. Seasonal variations in sales are evident, with increased sales during festive periods. Businesses should tailor their marketing and sales strategies to capitalize on these peak times, enhancing customer satisfaction and driving overall sales growth.



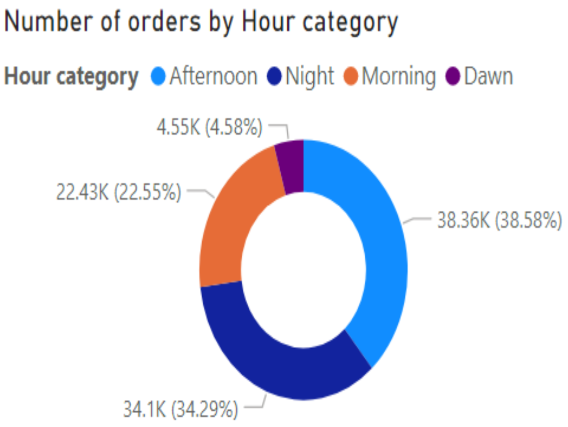
We found that each month's 2017 revenue was increasing. The revenue showed a rapid upward trend in 2017 and reached its peak in November. After that, it has remained stable and fluctuates at around 0.9M. It can be seen from this that 2017 was the peak period of Olist’s development, and the order volume increased sharply. By 2018, the sales volume stabilized and the development entered the stabilizer. Data to speculate how to design marketing strategies.



3) Number of orders by hour category : dawn, night, morning, night

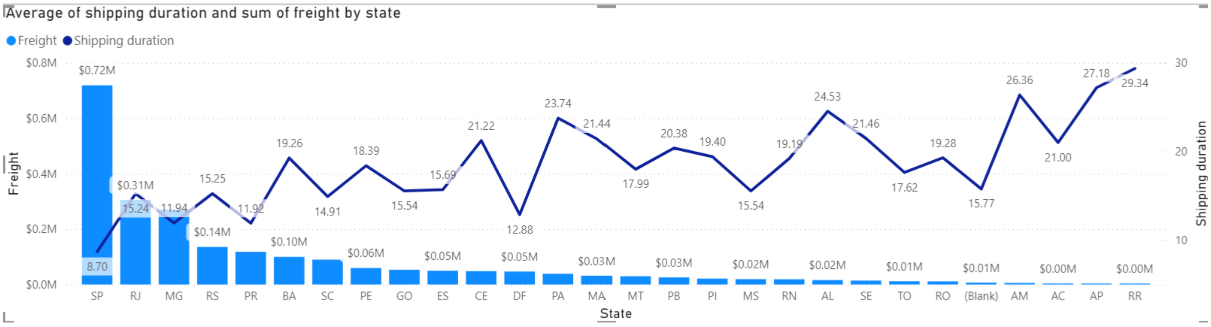
Brazilian customers predominantly place orders during the **afternoon** and **night**. This suggests a preference for online shopping during leisure time or after completing daily activities.

Understanding these buying patterns helps e-commerce businesses optimize their operations. By identifying peak buying times, companies can more effectively allocate resources, such as customer service representatives and inventory, to meet demand and ensure a seamless shopping experience.

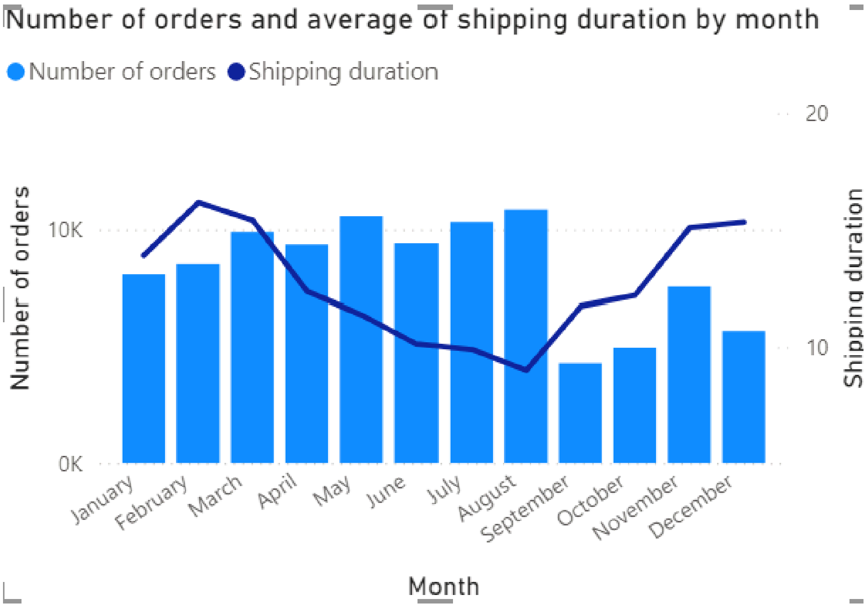


4) Average freight and number of orders of product category and number of orders and average of shipping duration by month

Here we see an inverse ratio between **average of shipping duration** and **sum of freight** by state. Obviously, freight by state is proportional to the number of orders. That also means the fewer orders a state has, the longer it takes to deliver. We can observe that **São Paulo (SP)** has the lowest average time to deliver, while **Roraima (RR)** has the highest average time to deliver. We can observe that **São Paulo (SP)** has the lowest average delivery time with 41360 orders, while **Roraima (RR)** has the highest average delivery time with only 46 orders. RJ and MG states have the 2nd and 3rd highest sum of freight, with 12,749 orders and 11,534 orders, respectively, but have quite high average delivery times. Improving delivery times in areas with longer delivery durations can have a positive impact on customer satisfaction and encourage repeat purchases. Streamlining logistics and implementing efficient shipping processes are key to achieving this.



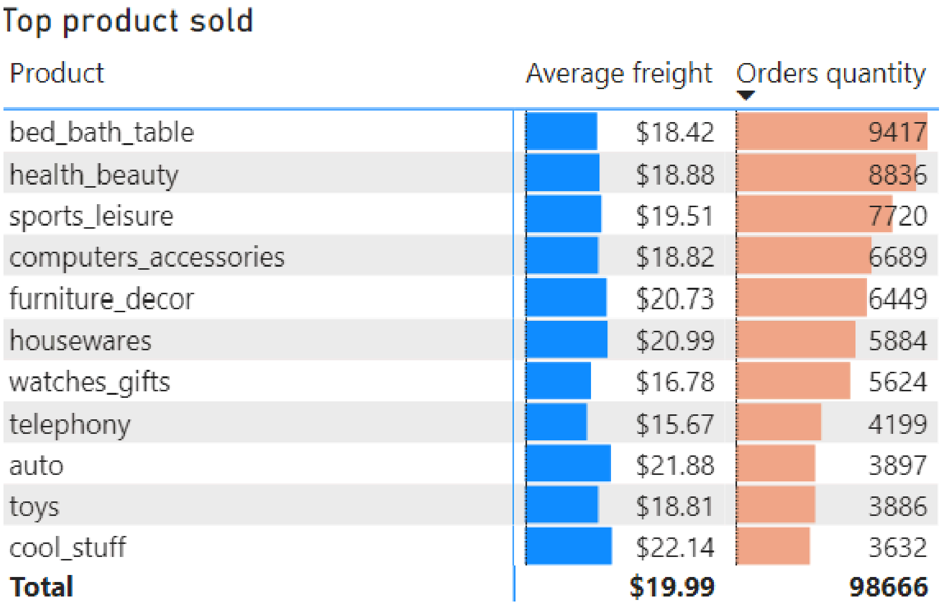
From the histogram distribution, we observed that the shipping duration in autumn and winter is significantly higher than in spring and summer. This may be due to higher demand in spring and summer compared to autumn and winter.



To sustain growth and enhance customer satisfaction, Olist must focus on strategies to boost development speed and improve logistics efficiency. To address the slowing growth of Olist, it's essential to devise strategies to reduce operating costs, particularly in shipping. This can be achieved by consolidating multiple shipments into a single one, reducing package size and weight, and shipping higher quantities of products. To do this we need the **Top product sold: average freight and order quantity**. By optimizing these logistics aspects, Olist can improve efficiency and customer satisfaction.

5) **Top product sold: average freight and order quantity** and **Top product sold: order quantity, % Order quantity, Total revenue**

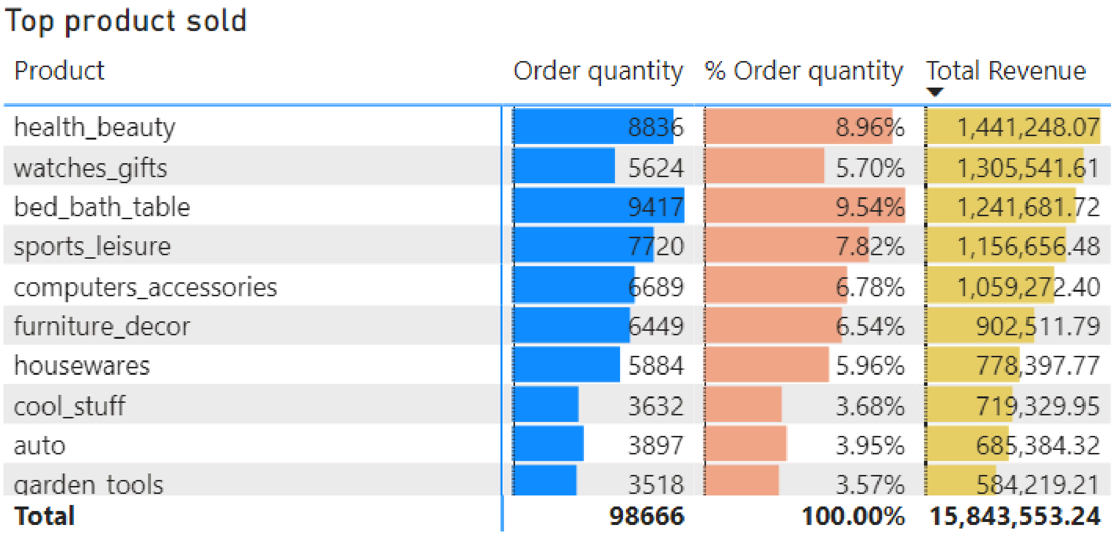
The main purpose of **Top product sold: average freight and order quantity** table is to track product categories that are ordered a lot and have high average freight. For example, bed_bath_table has 9417 orders with an average freight of 18.42. Suppose we can reduce the average freight to 18 then we can save 9417*0.42= 3955



Top product sold

Product	Average freight	Orders quantity
bed_bath_table	$18.42	9417
health_beauty	$18.88	8836
sports_leisure	$19.51	7720
computers_accessories	$18.82	6689
furniture_decor	$20.73	6449
housewares	$20.99	5884
watches_gifts	$16.78	5624
telephony	$15.67	4199
auto	$21.88	3897
toys	$18.81	3886
cool_stuff	$22.14	3632
Total	**$19.99**	**98666**

Besides, the table **Top product sold: order quantity, % Order quantity, Total revenue** is used to track the number of orders of the product category as well as revenue so we can maximize GMV. We can see watched_gifts, which has an average freight value of only 16.78 but brings the second highest revenue value of 1,305M.



Top product sold

Product	Order quantity	% Order quantity	Total Revenue
health_beauty	8836	8.96%	1,441,248.07
watches_gifts	5624	5.70%	1,305,541.61
bed_bath_table	9417	9.54%	1,241,681.72
sports_leisure	7720	7.82%	1,156,656.48
computers_accessories	6689	6.78%	1,059,272.40
furniture_decor	6449	6.54%	902,511.79
housewares	5884	5.96%	778,397.77
cool_stuff	3632	3.68%	719,329.95
auto	3897	3.95%	685,384.32
garden_tools	3518	3.57%	584,219.21
Total	**98666**	**100.00%**	**15,843,553.24**

B. Recommendations for Olist

1. Improve Order Validation Efficiency

- Strengthen communication with partner merchants to shorten order processing times.
- Implement a policy requiring general merchants to process orders within 24 hours and merchants selling customized goods to process orders within 48 hours, ensuring efficient shopping for consumers.
- Optimizing warehouse operations, refining shipping routes by consolidating multiple shipments into a single one, reducing package size and weight, and shipping higher

quantities of products. Consider increasing prices or adjusting freight fees as appropriate

2. **Attract regular customers and customers retention**

- Collect customer emails and other information to send regular product recommendations and promotional activities.
- Develop themed activities throughout the year to encourage repeat shopping on the platform.
- Enhance the customer service experience by offering chat support services and ensuring prompt and effective responses to customer inquiries.

C. Additional Data/Dataset

- **Product affinity, relationship:** find products frequently bought together. Determine correlation between purchased items: implemented as a correlation matrix. In cross-sell, we promote a relevant product at the point of purchase.
- **Monitoring stock levels**, inventory movements, and stock transactions is essential for inventory management, predicting demand, and ensuring product availability.
- **Customer transaction**. Understanding customer transaction histories and categorizations can aid in customer relationship management (CRM), personalized marketing strategies, and customer retention efforts.
- **Conversion rate**. The conversion rate, also a percentage, is the rate at which users on your ecommerce site are converting (or buying). This is calculated by dividing the total number of visitors (to a site, page, category, or selection of pages) by the total number of conversions
- **Shopping cart abandonment rate**. The shopping cart abandonment rate tells you how many users are adding products to their shopping cart but not checking out. The lower this number, the better. If your cart abandonment rate is high, there may be too much friction in the checkout process.

Source:
https://www.shopify.com/blog/7365564-32-key-performance-indicators-kpis-for-ecommerce
https://medium.com/@ongxuanhong/dataops-03-trino-dbt-spark-everything-everywhere-all-at-once-241932d27a6
https://www.analyticsvidhya.com/blog/2023/06/sql-powers-to-reveal-insights-into-brazilian-online-shopping/