CADT
IDT វិទ្យាស្ថានបច្ចេកវិទ្យាឌីជីថល
Institute of Digital Technology

PowerBi Project Report

TOPIC: “Brazilian E-Commerce Public by Olist”

Bachelor of Computer Science

Lecture by

Mr. Chan Sophal

Submitted By:

Group 4

Team member :

Sok Yongyi (IDTB080158)

Som Visal (IDTB080040)

Tet Davann (IDTB080023)

Vicheanon Norakpichit (IDTB080043)

Mak Channa

Siv Sreynoch (IDTB080133)

Academic Year: 2022-2023

I. Executive Summary

This report presents the findings from the dataset, which was provided by **Olist Store** in Brazil. From this dataset, we explored some key insights that can enhance the business performance such as improving customers satisfactions about order processing time, delivery time, review historical annual order purchases, products recommendation, and so on.

II. Introduction

A. Background

The Brazilian E-Commerce Public Dataset by Olist can be traced back to Olist, a Brazilian company that operates as a marketplace connecting small and medium-sized businesses with multiple e-commerce platforms. Olist provides these businesses with a platform to sell their products across various online marketplaces in Brazil.

To foster research and innovation in the e-commerce domain, Olist decided to release a public dataset that encompasses a significant amount of anonymized order data from their platform. The dataset covers a period of two years, from 2016 to 2018, and includes information from 100,000 orders.

B. Purpose

The purpose of gaining insights from the Brazilian E-Commerce Public Dataset by Olist is to inform and guide businesses in improving their product offerings, marketing strategies, and customer service. By analysing the dataset, businesses can extract valuable information that helps them make data-driven decisions and take actions to enhance their operations.The specific purposes of gaining insights from the dataset include:

- Product Optimization
- Marketing Strategy Enhancement
- Customer Service Improvement
- Forecasting and Planning
- Regional Targeting
- Recommender Systems
- Process Efficiency

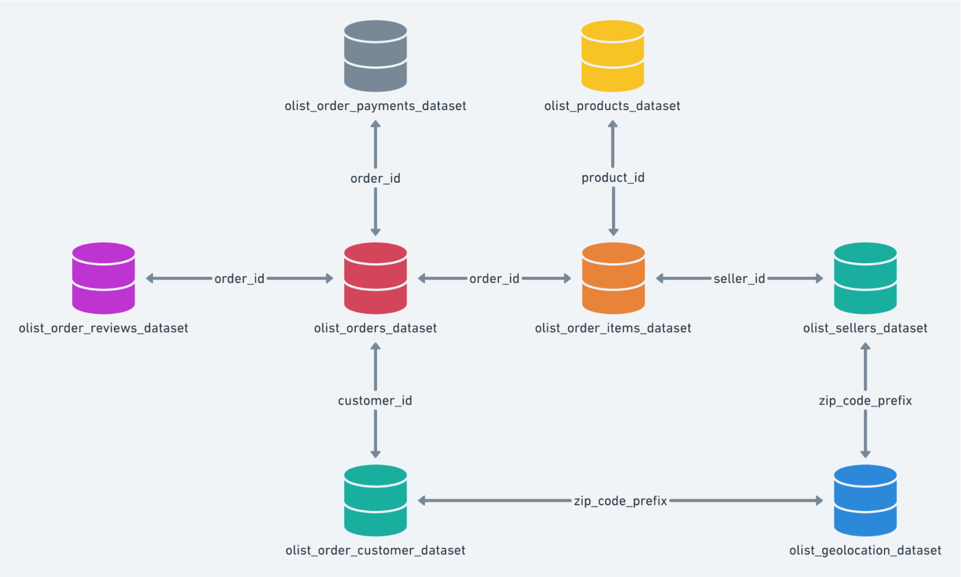
III. Methodology

A. Data Collection

We use secondary data from Kaggle platform.This dataset is a Brazilian e-commerce public dataset of orders made at Olist Store. The dataset has information on 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. 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. We also released a geolocation dataset that relates Brazilian zip codes to lat/long coordinates.

This dataset contains 8 distinct tables such as olist_customer_dataset, olist_geolocation_dataset, olist_order_items_dataset, olist_order_payments_dataset, olist_reviews_dataset, olist_orders_dataset, olist_products_dataset, olist_sellers_dataset, product_category_name_translation. All these tables are in CSV format.

Data Schema



B. Data Preprocessing

When retrieving data from MySQL on a cloud service to Power BI, all columns of the table are in text format. Therefore, it is necessary to adjust the type of each column based on the type of data it contains before proceeding with data cleaning.
During the data cleaning process, we have three main processes to clean our data:

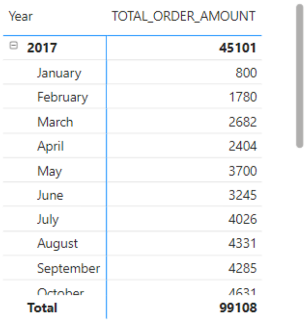
- **Handling missing data:** We can detect missing values in Power BI by observing the view column quality above the table header, which shows the percentage of missing values in each column. We found two tables with missing values, totaling six columns. Fortunately, three of those columns are not used for data visualisation. Therefore, we only need to impute three of these columns, while the unnecessary columns have been removed. The columns that need to be imputed have two types: text and date. We applied different methods to replace the missing values:
    - For text type columns, we randomly replaced the missing values with some of the top products that have high frequency.
    - For date type columns, we replaced the missing values with random values based on other related columns to avoid incorrect dates.
- **Handling outliers:** It does not have problems detecting outliers in Power BI. We can check it by clicking on columns to view the column profile (When we click,it will show at the bottom). The column profile shows other statistical information that helps to detect outliers by watching only those that like to detect missing values as well.
- **Removing duplicates:** In this step, it is very easy. By right-clicking on the column header, we can access the menu and choose "remove duplicates".

IV. Exploratory Data Analysis (EDA)

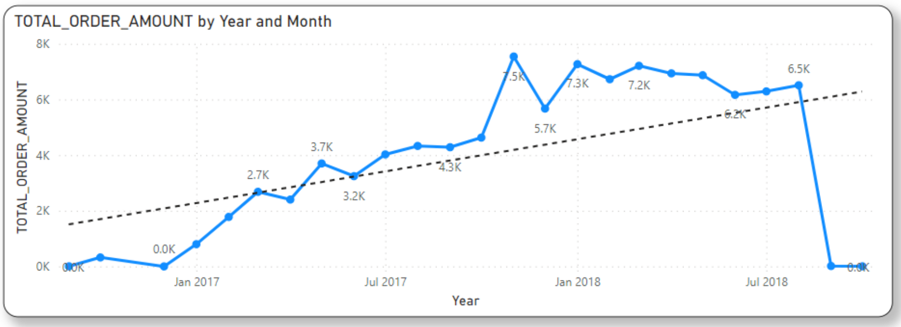
In the data presentation stage, we will show you all insights visually in order to make it easy to understand the meaningful insights from our analysis of the data.
The following visualisations illustrate the meaningful insight respectively.

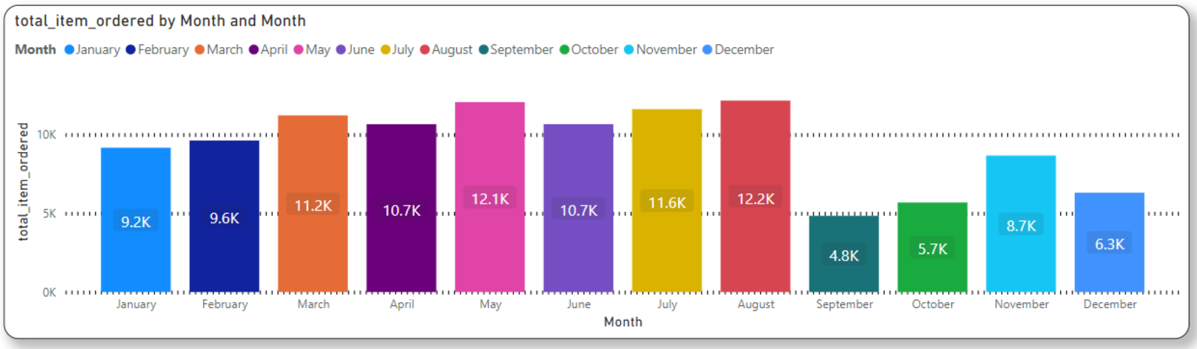
- Sales over the year [1]

These graphs below show the timeline of sales over the year from October 2016 to October 2018 and show which months have the most orders. By doing this, we can use this to help us with inventory management and identify seasonal trends. We can see that in October 2017, the amount of sales rose drastically from 4.5k to 7.6k.

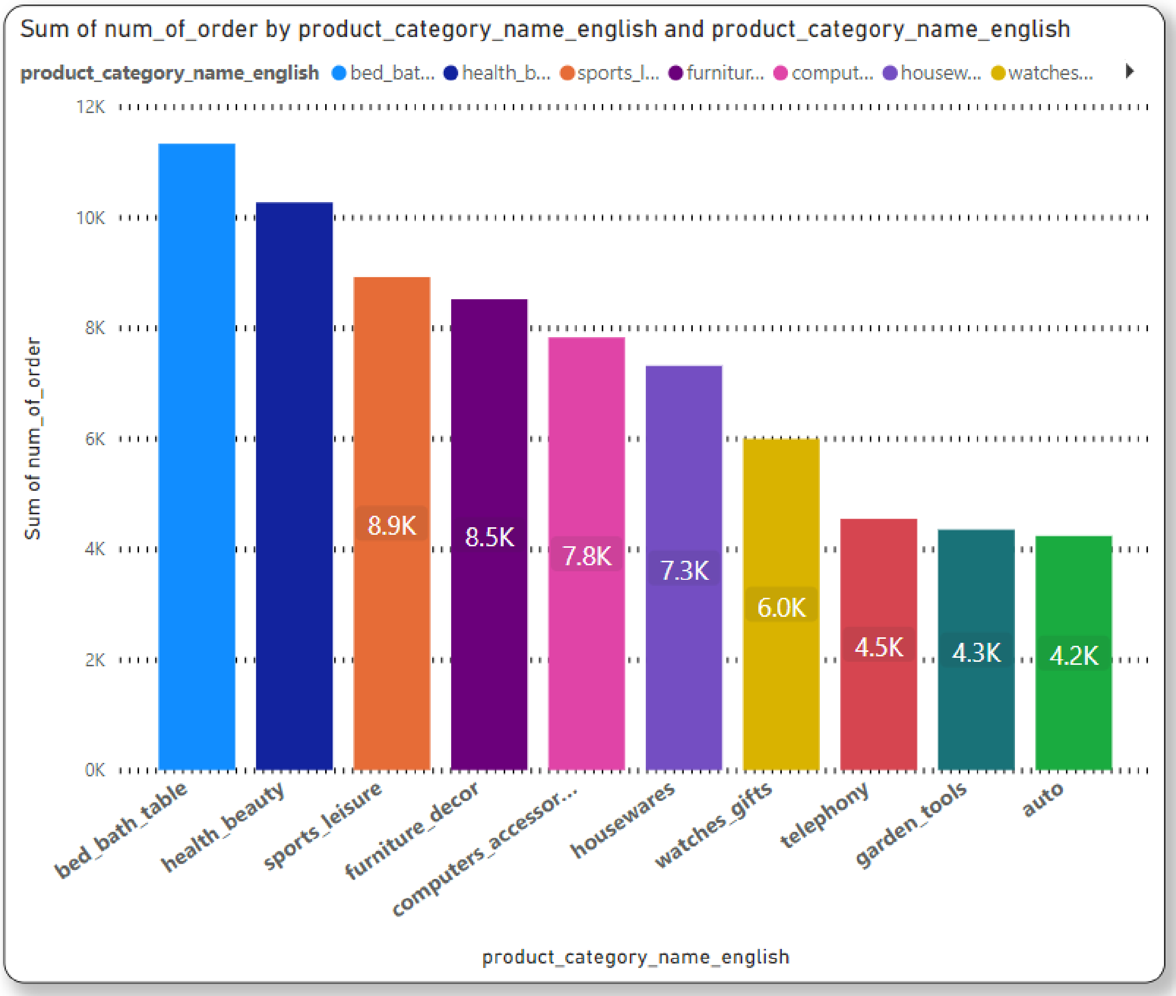


Year	TOTAL_ORDER_AMOUNT
⊟ **2017**	**45101**
January	800
February	1780
March	2682
April	2404
May	3700
June	3245
July	4026
August	4331
September	4285
October	4631
Total	**99108**





- Most popular product categories [2]

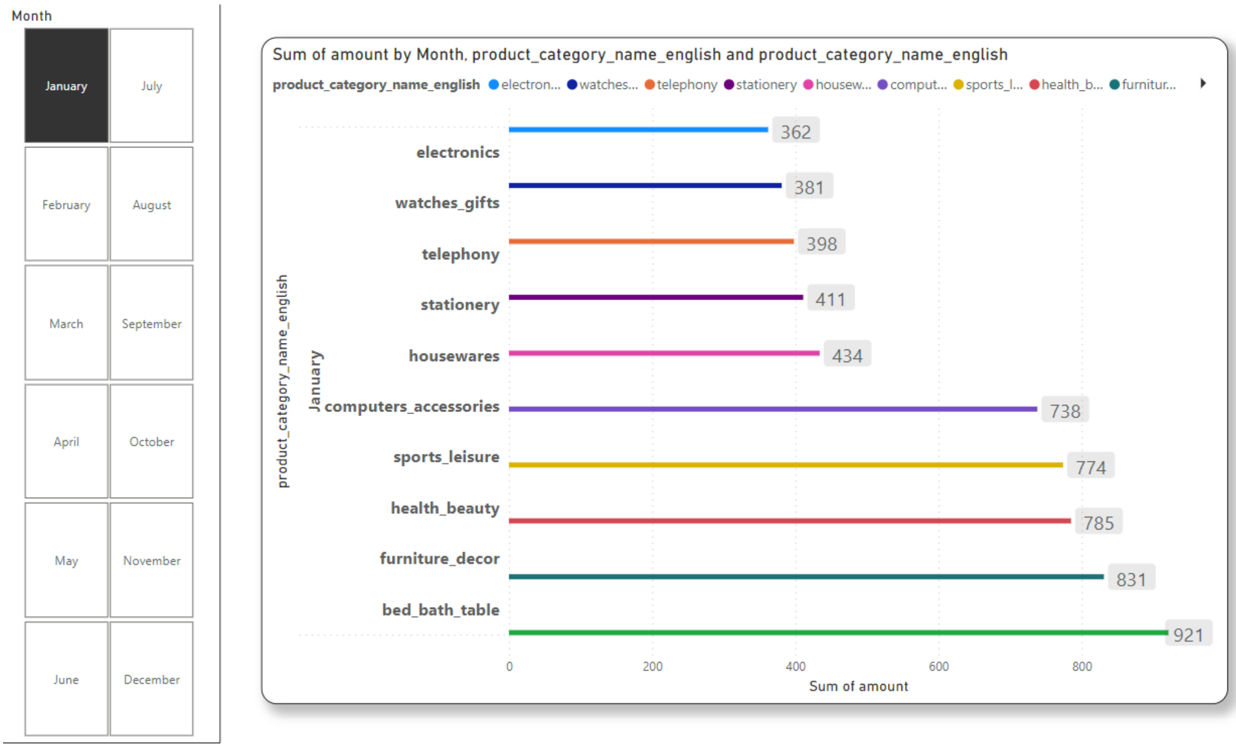


This visual tells us which categories are the top ordered items. As shown on the bara chart, we can see that the top 3 most ordered items are:

- **Bed_bath_table: 11K**
- **Health_beauty: 10K**
- **Sports_leisure: 8.9K**

With this graph, we know about the ordering behaviour of customers so that we can promote the right product categories to increase sales.

- Most popular product categories on each month [3]

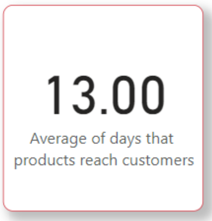


This graph illustrates the most famous categories of each month. We can see that on January, our top 5 most popular product types are:

- **Bed_bath_table: 921**
- **Furniture_decor: 831**
- **Health_beauty: 785**
- **Sports_leisure: 774**
- **Computer_accessories: 738**

We can identify the seasonal trend of customers' behaviours and we can increase sales and revenues, know the best times to launch a new product, and better inventory management.

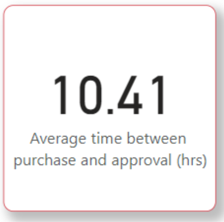
- Order Processing Time, Delivery Time, and several summary [4]



13.00

Average of days that products reach customers

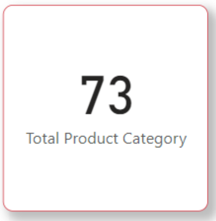
Card 1



10.41

Average time between purchase and approval (hrs)

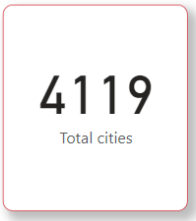
Card 2



73

Total Product Category

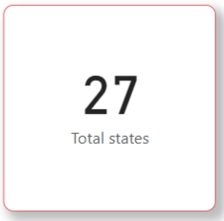
Card 3



4119

Total cities

Card 4



27

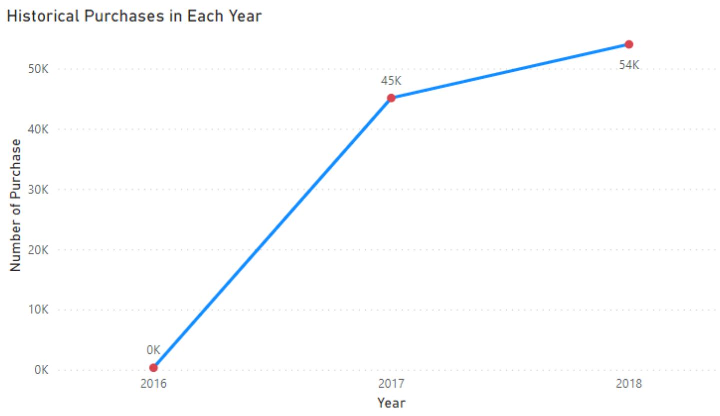
Total states

Card 5

The **Card 1** shows about the average of days the our product is delivered to the customers. As we saw in the card visualisation, it is 13-day. Hence, we can suggest the company to encourage customers to give feedback about this delivery time so we can improve delivery performance. The **Card 2** illustrates the average time of approval. As the average of delivery, we also can suggest to the company to encourage customers to give feedback about approval time. If there is any negative feedback, we will introduce the company to an automated system to make approval time faster. The **Card 3** is the summary of the total product category in the store, which are 73 categories. The **Card 4** is the summary of the total cities that our customers come from, which are 4119 cities. Finally, the **Card 5** is the summary of the total states that our customers come from, which are 27 states.

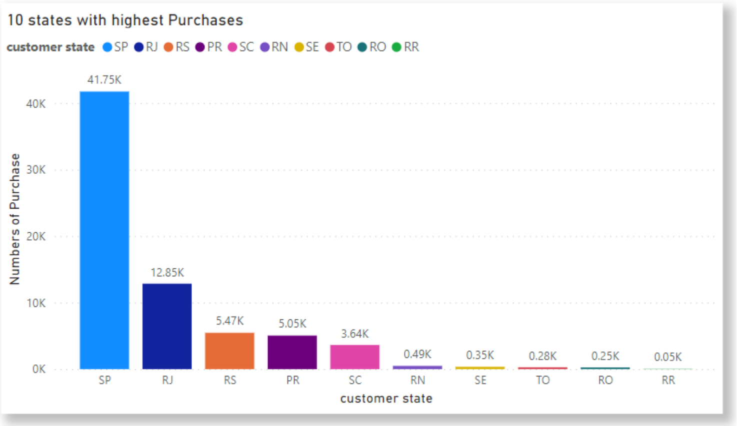
- Historical Purchases [5]

This line graph illustrates the historical number of purchases in each. As we see in the graph, in 2016, there were 329 order purchases. In 2017, there were 45101 order purchases. In 2018, there were 54011 order purchases.



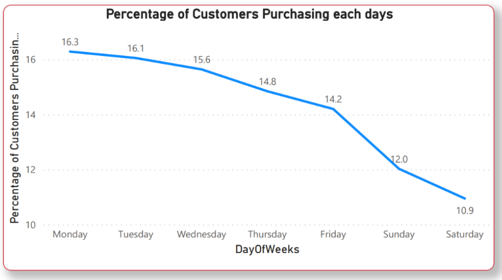
- 10 states that our customers come from the most [6]

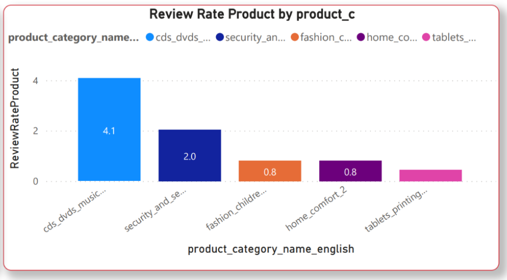
The right graph shows about 10 states that our customers come from. As we see in the graph, **SP** state is the state that has the highest customer, and then **RJ,** and so on. As a suggestion, we extend the market to the **SP** state.

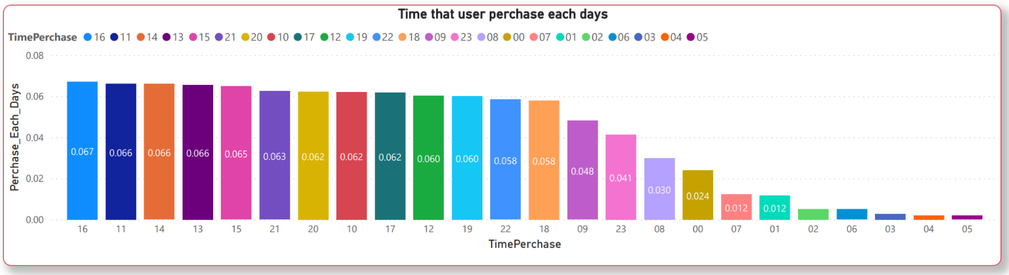


- Customer Purchase Trends and Product Recommendations [7]

Based on our analysis of the collected and visualised data, Monday is the most common day for customer purchases. Monday is 16.3 percent that is the highest percentage while Saturday is only 10.9 which is the lowest than others. Most customers make purchases between 11:00 and 23:00, with purchase patterns influenced by products, time, and location. Customers also provide review scores to indicate their satisfaction with purchased products. The products with the highest average review scores are CDs/DVDs, security and service, fashion children's clothes, home comfort, and tablet printing images. Vendors and warehouses should prioritise trending products and maintain adequate stock levels to avoid out-of-stock situations. By understanding popular products, vendors can optimise profit and develop effective marketing strategies.

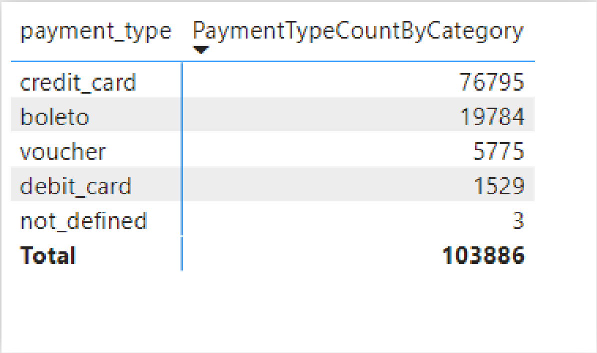




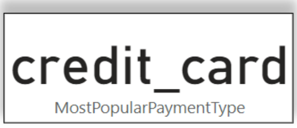


- The most popular payment method [8]

Based on our data in table (dataset olist_order_payments_dataset), column payment_type, I counted all the payment types, And the result shows that credit_card is the popular payment method of our customers.



payment_type	PaymentTypeCountByCategory
credit_card	76795
boleto	19784
voucher	5775
debit_card	1529
not_defined	3
Total	**103886**

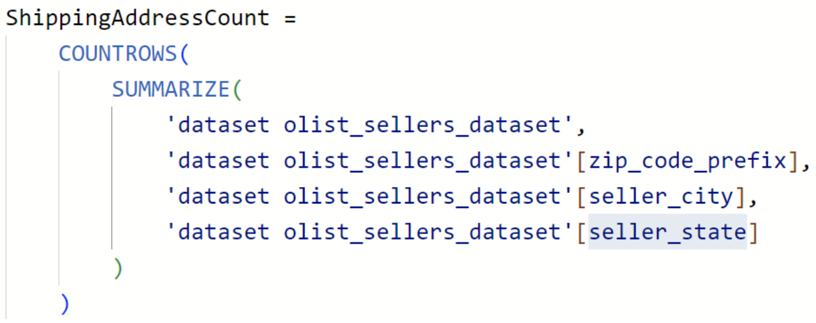


credit_card

MostPopularPaymentType

- The most common shipping addresses [9]

In this part, I using the table (dataset olist_sellers_dataset) and columns (`zip_code_prefix , seller_city , seller_state`), and then I counted the data from these column by using the DAX following below:

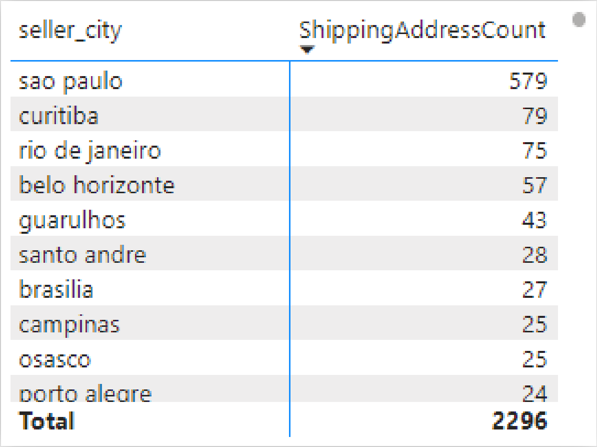


```
ShippingAddressCount =
    COUNTROWS(
        SUMMARIZE(
            'dataset olist_sellers_dataset',
            'dataset olist_sellers_dataset'[zip_code_prefix],
            'dataset olist_sellers_dataset'[seller_city],
            'dataset olist_sellers_dataset'[seller_state]
        )
    )
```

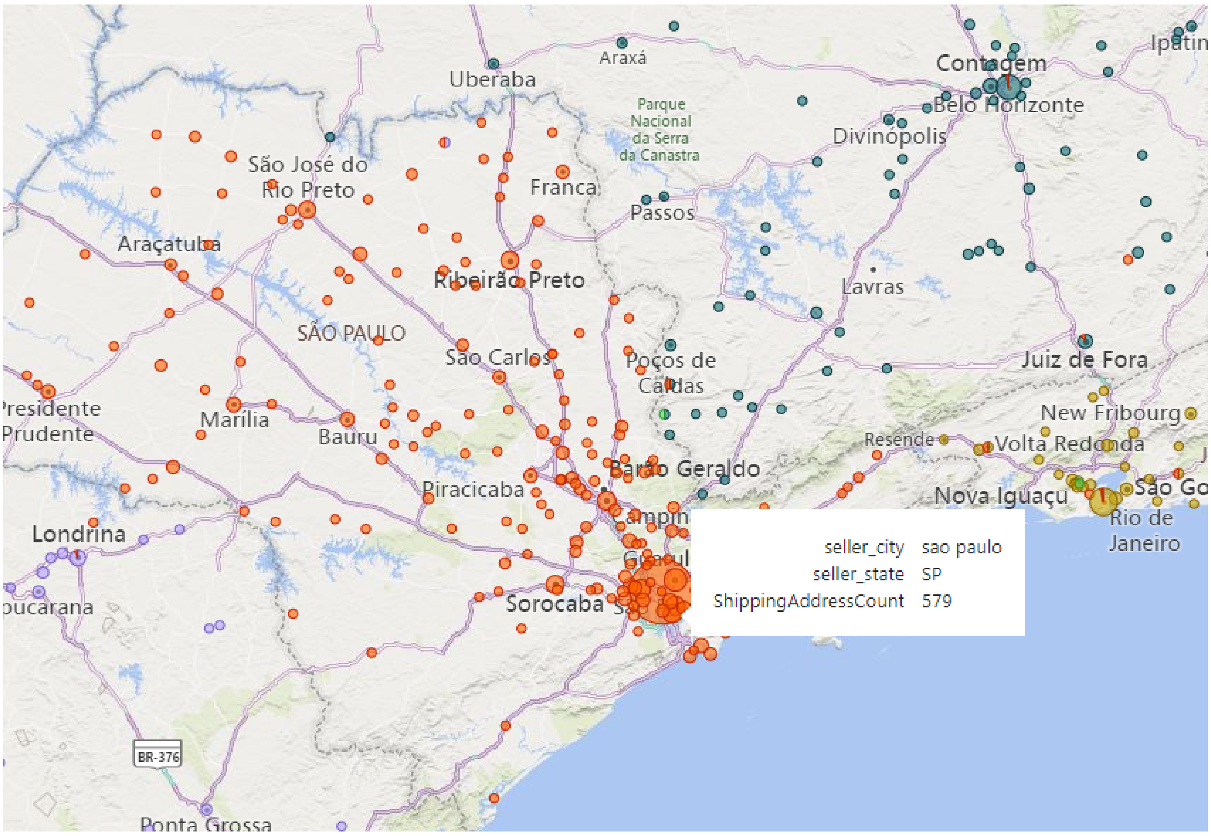
Challenge: Why did we use these data to find the most common shipping addresses?

If we imagine that a city has a lot of sellers, shipping our products is very busy in this city, So that is why we used this method to find the most common shipping addresses.

Our final result shows that Sao Paulo city has the most sellers (579 sellers) around this city in Brazil, Hence the most common shipping address is located in **Sao Paulo** city.



seller_city	ShippingAddressCount
sao paulo	579
curitiba	79
rio de janeiro	75
belo horizonte	57
guarulhos	43
santo andre	28
brasilia	27
campinas	25
osasco	25
porto alegre	24
Total	**2296**

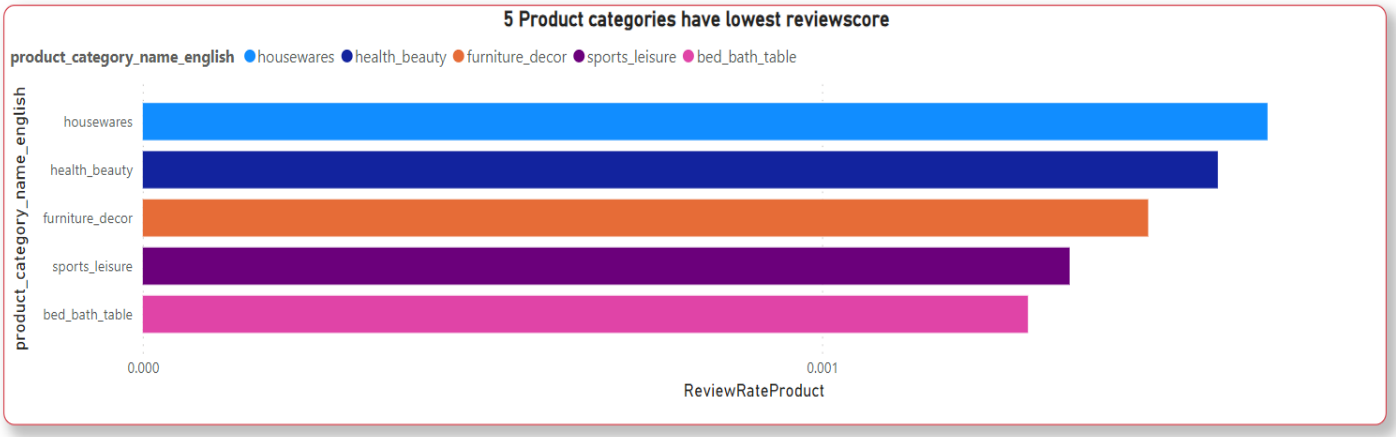


seller_city	sao paulo
seller_state	SP
ShippingAddressCount	579

In this part we can say that If the city has a lot of sellers, So this city also has a lot of customers too based on our dataset.Hence if we want to expand our customers, we should expand sellers around Brazil.

- The product categories have lowest review score [10]

Based on our dataset, we can find the insight for improving our business by decreasing which product category has the lowest review score or not sold out.



From this bar chart, We can see the lowest product review score is Housewares product, So we should not produce this product too much. Just a little bit that is enough for our customers who order this product.

V. Recommendation

- [9] Payment method: We should suggest that our cashiers must have a credit card account for customers charged by it.
- [9] Seller: Increasing seller for city that has little bit sellers for customers easy to buy our product by do not pass from one city to another to buy our products. Hence our product will increase double customers.
- [7] Discount: We should have a some discount on weekend for deal the lack of customers for the every weekend, However, we selected the lowest rate product for this discount for increasing the customers to buys this product on weekend.
- [1] Discount for seller: Based on visualization[1] we could see the our product lack of sold on September, October, November, and December. So we should have any discount for them for every the end of the years.

*Note: [9], [7]... means that these recommendation based on the visualization.

VI. Conclusion

Based on the above interpretation, We can improve Strategy for following below by:

- Improving Warehouse or stock products based on trending products.
- Analyze delivery routes: Evaluate existing routes to identify inefficiencies and challenges.
- Coordinate with local delivery services: Collaborate with local logistics partners who have expertise in these states.
- Enhance customer communication: Keep customers informed about orders and potential delays.
- By implementing these strategies and addressing the challenges specific to location you can improve delivery performance and enhance overall customer satisfaction.

References

- Dataset link: Brazilian E-Commerce Public Dataset by Olist
- Our insight note: Insights
- Tool for analysis and visualization: Power BI (https://www.microsoft.com/en-us/power-platform/products/power-bi)
- Cloud for storing data: MySQL (https://www.mysql.com/)
- Supported AI: ChatGPT (https://chat.openai.com/) , Bard (https://bard.google.com/)

Thanks you.