Enhancing Shipping Efficiency and Customer Interaction on OList

A Data-Driven Analytical Approach

[BDM - Project]

A Final report for the BDM capstone Project

Submitted by

Name: Vivek Singh Rao

Roll number:21f3002861@ds.study.iitm.ac.in

Date of submission- 23-12-2024

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai, Tamil Nadu, India, 600036

Contents

Declaration Statement

I am working on a Project titled “Brazilian Ecommerce”. I extend my appreciation to Kaggle site, for providing the necessary resources that enabled me to conduct my project.

I hereby assert that the data presented and assessed in this project report is genuine and precise to the utmost extent of my knowledge and capabilities. The data has been gathered from primary sources and carefully analyzed to assure its reliability.

Additionally, I affirm that all procedures employed for the purpose of data collection and analysis have been duly explained in this report. The outcomes and inferences derived from the data are an accurate depiction of the findings acquired through thorough analytical procedures.

I am dedicated to adhering to the principles of academic honesty and integrity, and I am receptive to any additional examination or validation of the data contained in this project report.

I understand that the execution of this project is intended for individual completion and is not to be undertaken collectively. I thus affirm that I am not engaged in any form of collaboration with other individuals, and that all the work undertaken has been solely conducted by me. In the event that plagiarism is detected in the report at any stage of the project's completion, I am fully aware and prepared to accept disciplinary measures imposed by the relevant authority.

I understand that all recommendations made in this project report are within the context of the academic project taken up towards course fulfillment in the BS Degree Program offered by IIT Madras. The institution does not endorse any of the claims or comments.

Vivek Singh

Signature of Candidate: (**Digital Signature**)

Name: Vivek Singh Rao

Date: 2024/12/03

1. Executive Summary and Title

This project analyses a Brazilian ecommerce dataset from Olist, a platform that connects small businesses across Brazil, enabling them to access multiple sales channels seamlessly under a single contract. Olist allows merchants to market their products through the Olist Store and use its logistics partners for direct shipping to customers. The dataset includes 100,000 orders from 2016 to 2018, offering valuable insights into order performance, logistics, customer demographics, product attributes, and customer feedback.

The analysis highlights two major challenges: poor customer loyalty, with only 10% of buyers returning for repeat purchases, and a significant disparity in sales across regions, with 80% of revenue concentrated in just two regions. These findings reveal missed opportunities in underperforming regions and underscore the need to strengthen customer engagement and tailor strategies for long-term business growth.

To address these issues, the project proposes conducting customer segmentation analysis to develop region-specific strategies. This involves correlating variables such as order status and freight costs in different regions to identify patterns. Additionally, sentiment analysis of customer review comments will be performed and correlated with features like review scores, order statuses, and customer locations. These analyses aim to uncover actionable insights to improve customer loyalty and reduce regional disparities in sales.

To optimize operations, the business could think of opening warehouses in strategic locations and redistributing sellers to reduce delivery times and costs, especially in underperforming regions. Poor-quality products with bad reviews should be removed to maintain high standards. Aligning seller distribution with regional demand and partnering with local couriers can further enhance delivery efficiency. Proactive competitor analysis is also recommended to identify opportunities and address gaps in pricing, product offerings, and customer satisfaction.

2. Proof of Originality

Links Dataset and G-Drive Link (BDM- Analysis)

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

https://drive.google.com/drive/folders/1rxPMgreJHHQHMqxsJMCb12FuFfwBlfLb?usp=sharing

The dataset used in this data analysis was taken from Kaggle, specifically from the extremely popular dataset titled Brazilian E-Commerce Public Dataset by Olist (link to the dataset). The dataset is presented under the CC BY-NC-SA 4.0 license and has, in fact, attracted most attention in the analysis regarding e-commerce, with 3,393 upvotes on the Kaggle platform in addition to a golden badge recognition.

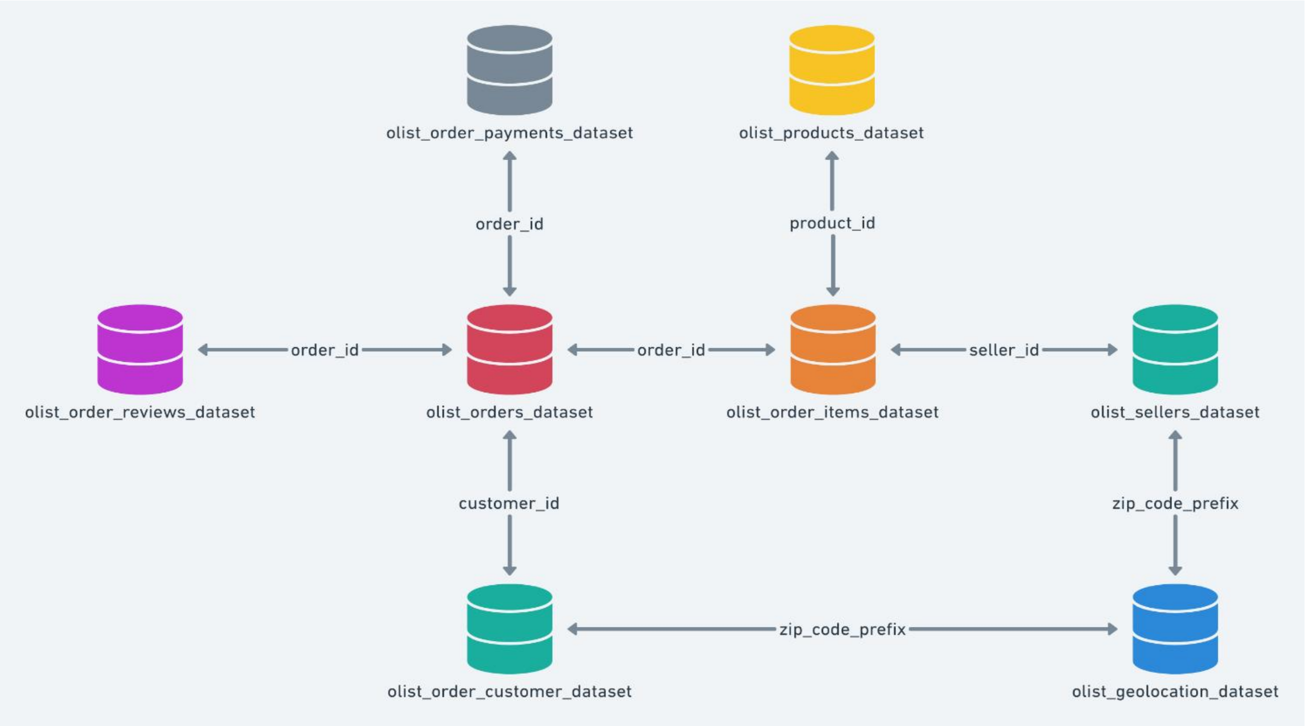
The company uses data from Olist. This is the biggest department store in Brazilian marketplaces and known for the role that makes small businesses in Brazil interact with their customers through seamless integration. It helps the Olist Store make sales through a unified channel and manages the logistics of its network of partners. More information about Olist can be found at: www.olist.com.

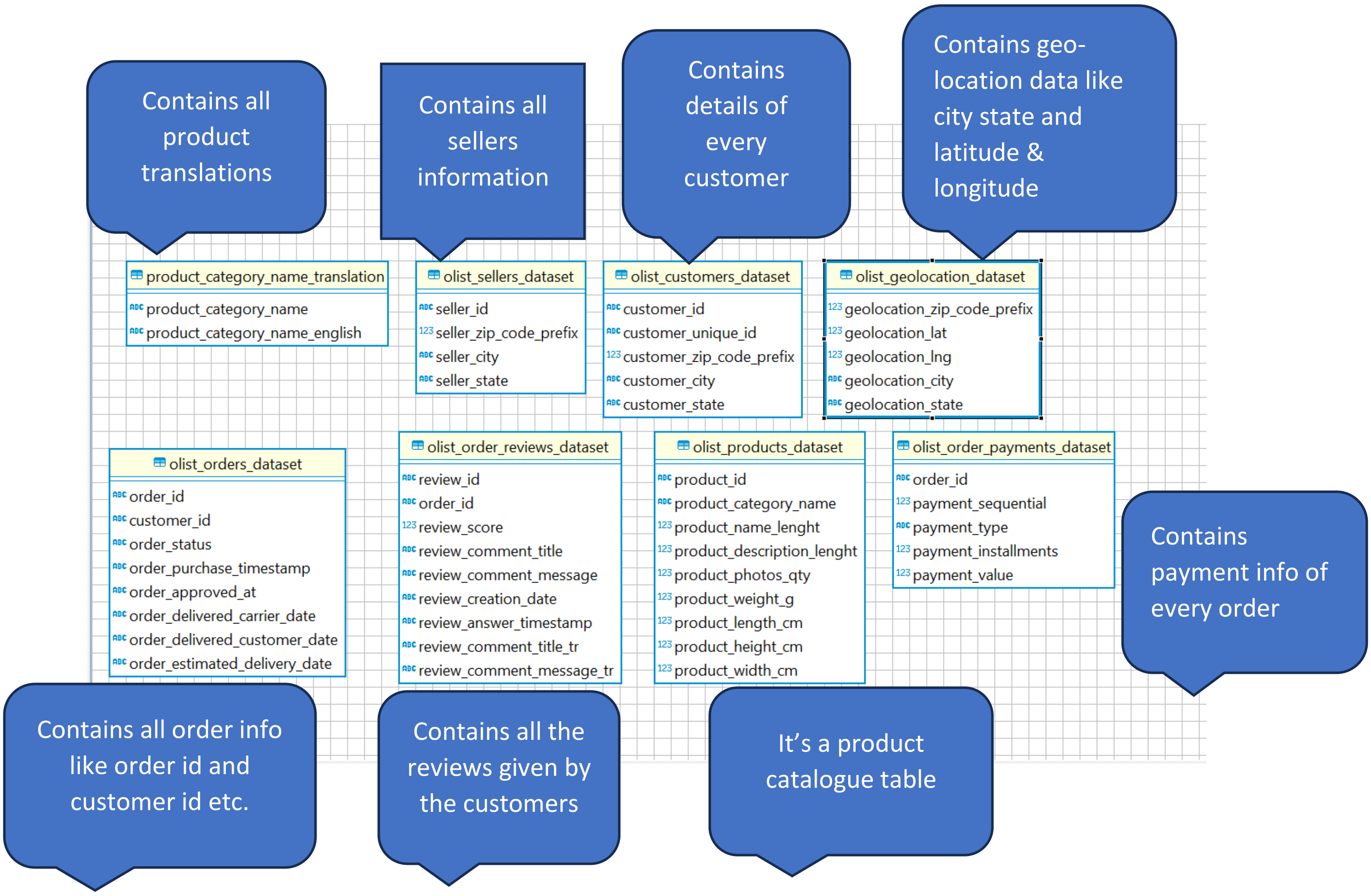
The dataset comprises the Olist Store's detailed transactional data. Once a customer buys a product, the seller receives the order. Customers are asked to provide feedback upon delivering the product or by the date of delivery, and they are given an option to do so using a satisfaction survey that is a rating with optional written comments.

This project contributes to the foundation dataset with an analysis of customer retention patterns and regional sales disparities and provides actionable insights into the problems. All analyses, visualizations, and interpretations presented in this report are independently conducted and uniquely crafted for this study.

3. Meta data and descriptive statistics

Database Schema & ER Diagram's of Tables





Contains all product translations

Contains all sellers information

Contains details of every customer

Contains geo-location data like city state and latitude & longitude

product_category_name_translation
product_category_name
product_category_name_english

olist_sellers_dataset
seller_id
seller_zip_code_prefix
seller_city
seller_state

olist_customers_dataset
customer_id
customer_unique_id
customer_zip_code_prefix
customer_city
customer_state

olist_geolocation_dataset
geolocation_zip_code_prefix
geolocation_lat
geolocation_lng
geolocation_city
geolocation_state

olist_orders_dataset
order_id
customer_id
order_status
order_purchase_timestamp
order_approved_at
order_delivered_carrier_date
order_delivered_customer_date
order_estimated_delivery_date

olist_order_reviews_dataset
review_id
order_id
review_score
review_comment_title
review_comment_message
review_creation_date
review_answer_timestamp
review_comment_title_tr
review_comment_message_tr

olist_products_dataset
product_id
product_category_name
product_name_lenght
product_description_lenght
product_photos_qty
product_weight_g
product_length_cm
product_height_cm
product_width_cm

olist_order_payments_dataset
order_id
payment_sequential
payment_type
payment_installments
payment_value

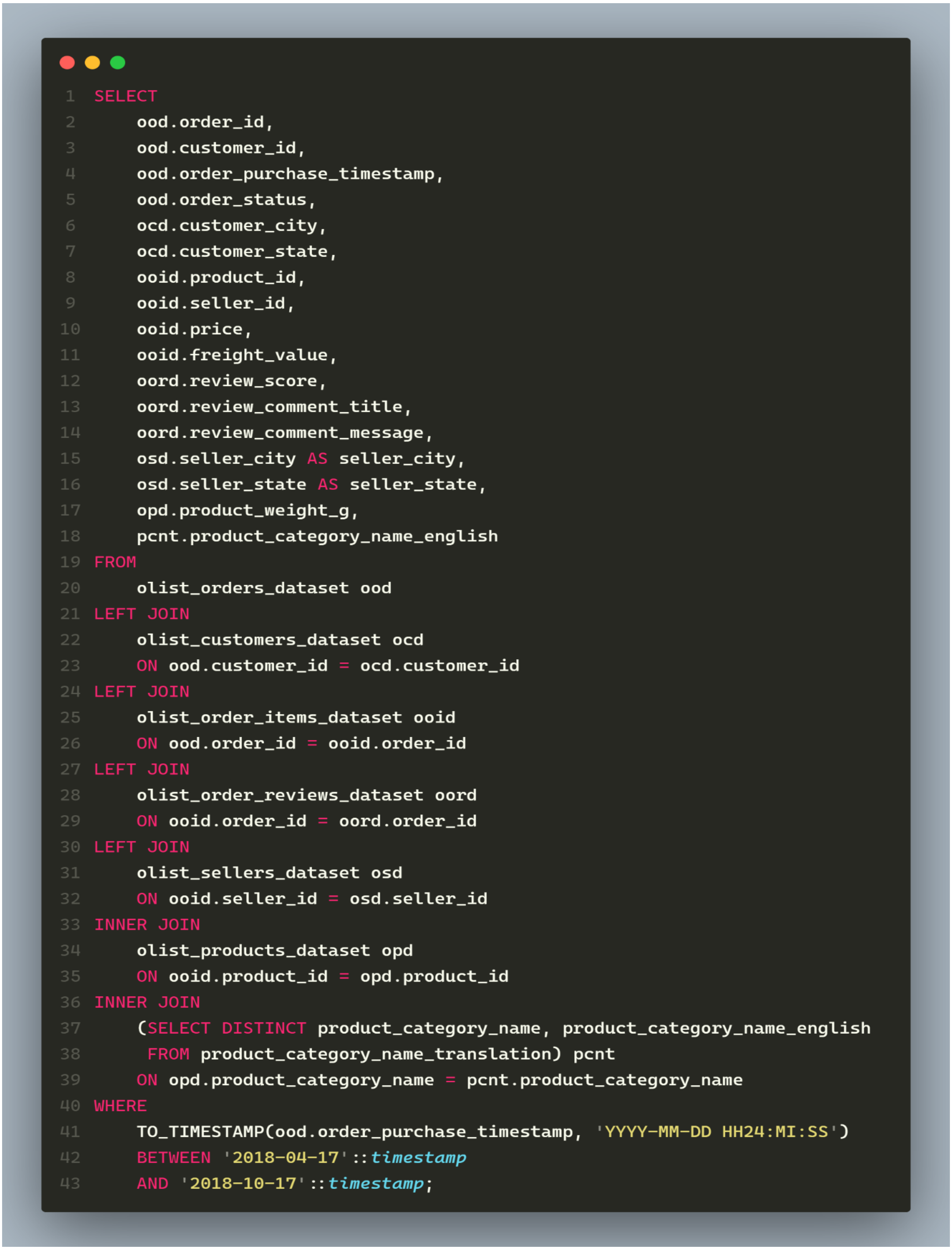
Contains payment info of every order

Contains all order info like order id and customer id etc.

Contains all the reviews given by the customers

It's a product catalogue table

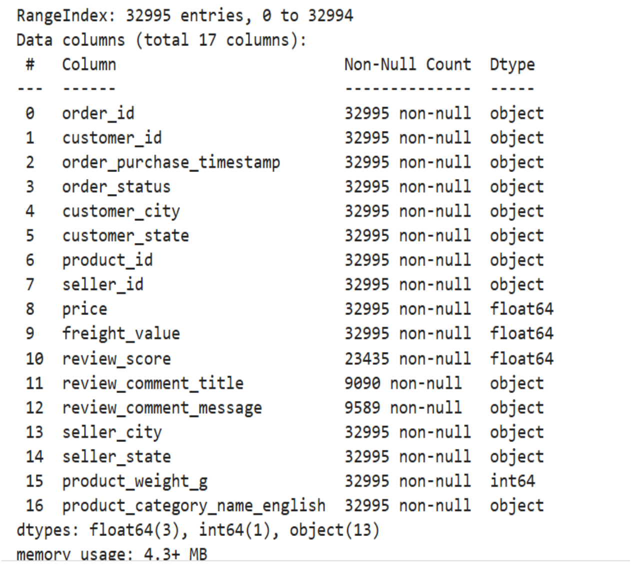
A single SQL query for unifying all database tables



```sql
SELECT
    ood.order_id,
    ood.customer_id,
    ood.order_purchase_timestamp,
    ood.order_status,
    ocd.customer_city,
    ocd.customer_state,
    ooid.product_id,
    ooid.seller_id,
    ooid.price,
    ooid.freight_value,
    oord.review_score,
    oord.review_comment_title,
    oord.review_comment_message,
    osd.seller_city AS seller_city,
    osd.seller_state AS seller_state,
    opd.product_weight_g,
    pcnt.product_category_name_english
FROM
    olist_orders_dataset ood
LEFT JOIN
    olist_customers_dataset ocd
    ON ood.customer_id = ocd.customer_id
LEFT JOIN
    olist_order_items_dataset ooid
    ON ood.order_id = ooid.order_id
LEFT JOIN
    olist_order_reviews_dataset oord
    ON ooid.order_id = oord.order_id
LEFT JOIN
    olist_sellers_dataset osd
    ON ooid.seller_id = osd.seller_id
INNER JOIN
    olist_products_dataset opd
    ON ooid.product_id = opd.product_id
INNER JOIN
    (SELECT DISTINCT product_category_name, product_category_name_english
     FROM product_category_name_translation) pcnt
    ON opd.product_category_name = pcnt.product_category_name
WHERE
    TO_TIMESTAMP(ood.order_purchase_timestamp, 'YYYY-MM-DD HH24:MI:SS')
    BETWEEN '2018-04-17'::timestamp
    AND '2018-10-17'::timestamp;
```

Query Results

Meta Data – Unified Table



```
RangeIndex: 32995 entries, 0 to 32994
Data columns (total 17 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   order_id                       32995 non-null  object
 1   customer_id                    32995 non-null  object
 2   order_purchase_timestamp       32995 non-null  object
 3   order_status                   32995 non-null  object
 4   customer_city                  32995 non-null  object
 5   customer_state                 32995 non-null  object
 6   product_id                     32995 non-null  object
 7   seller_id                      32995 non-null  object
 8   price                          32995 non-null  float64
 9   freight_value                  32995 non-null  float64
 10  review_score                   23435 non-null  float64
 11  review_comment_title           9090 non-null   object
 12  review_comment_message         9589 non-null   object
 13  seller_city                    32995 non-null  object
 14  seller_state                   32995 non-null  object
 15  product_weight_g               32995 non-null  int64
 16  product_category_name_english  32995 non-null  object
dtypes: float64(3), int64(1), object(13)
memory usage: 4.3+ MB
```

Unified Table Schema

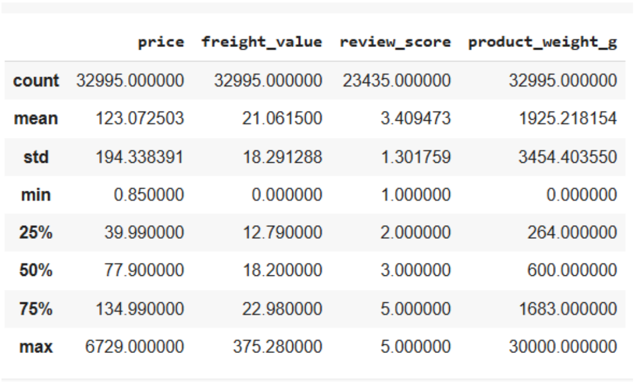
Rows=32995 Entries

Columns= 17 features

Table Information

Columns	Descriptions
Order_id	Unique ids for every order
Customer_id	Customer ids
Order purchase timestamp	Time at which customer placed order
Order_status	Status of order (delivered, shipped etc.)
Customer_city	City of the customer
Customer state	State of the customer
Product Id	Unique id of the product
Seller Id	Unique id of the seller
price	Price of the product
Freight value	Shipping Cost
Review Score	Rating given by customer to ordered item
Review_comment_title	Rating title
Review_comment_message	Rating comment message
Seller city	Seller City
Seller state	Seller State
Product_weight_g	Weight of the product in grams
Product category name english	Name of the product category in english

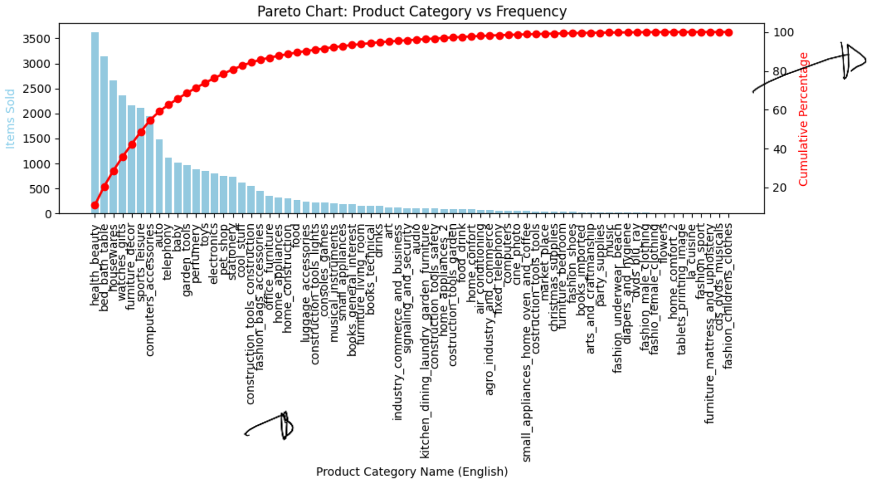
- I have consolidated all eight tables into a unified dataset, including all the essential features needed for data analysis to address the business challenges effectively.



	price	freight_value	review_score	product_weight_g
count	32995.000000	32995.000000	23435.000000	32995.000000
mean	123.072503	21.061500	3.409473	1925.218154
std	194.338391	18.291288	1.301759	3454.403550
min	0.850000	0.000000	1.000000	0.000000
25%	39.990000	12.790000	2.000000	264.000000
50%	77.900000	18.200000	3.000000	600.000000
75%	134.990000	22.980000	5.000000	1683.000000
max	6729.000000	375.280000	5.000000	30000.000000

Statistical Table

- Freight Cost has high standard deviation
- Price distribution is highly skewed
- Reviews are generally neutral around 3



Pareto Chart

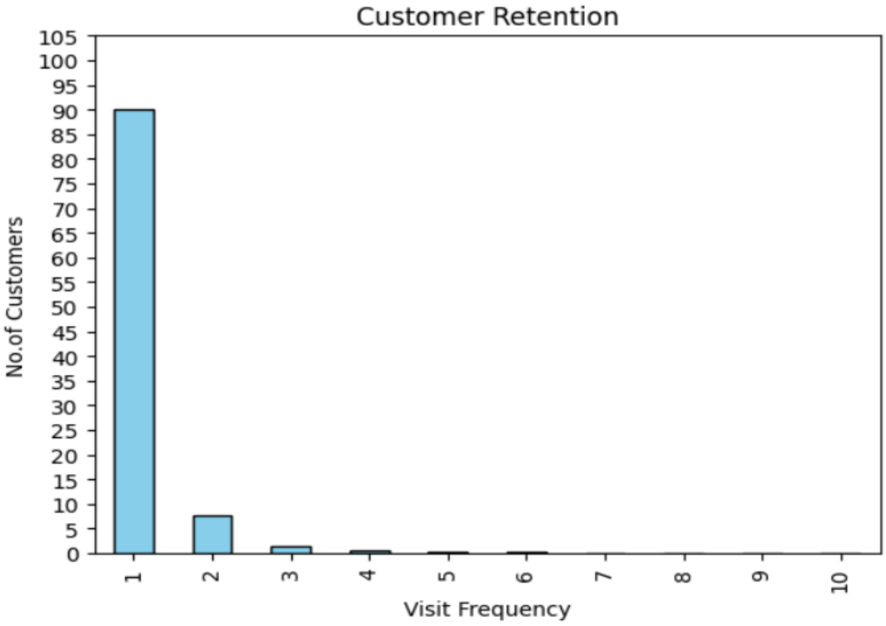
The Famous Pareto Principal

80% of the sales coming from 20% of the categories

X-axis= Product Category Name

Y-axis= Number of Items Ordered

Focus can be placed on the most influential areas by using the Pareto principle to determine the main elements that contribute to most sales. A bar graph makes it simpler to compare performance across categories or geographical areas by graphically highlighting differences and trends.



Customer Retention

X-axis= Visit Frequency

Y-axis= No. of Customers

To better understand customer retention issues, I plotted a graph showing customer visit frequency.

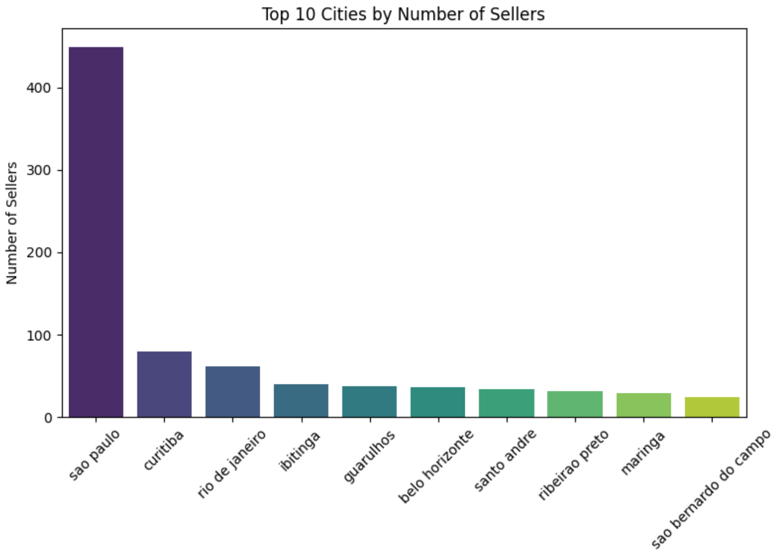
We can derive the following insights from the above graph that

- 90% of the customer only visited the store for one time

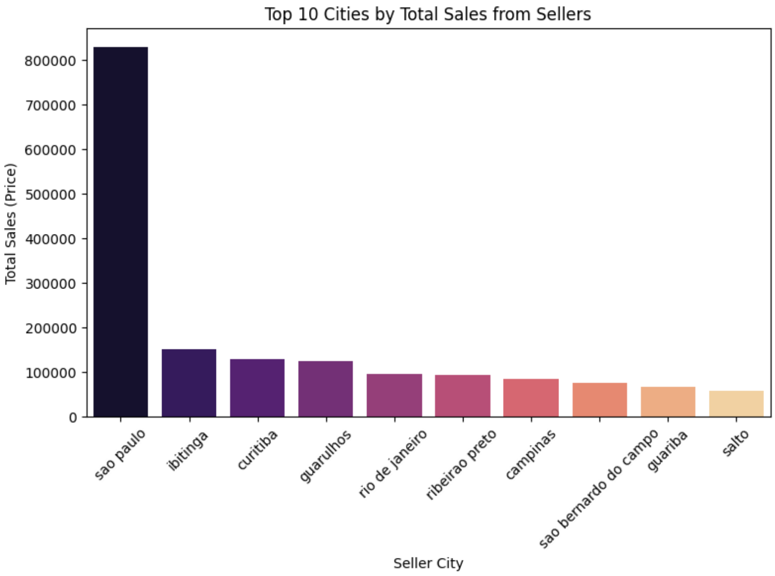
- 7.5% of the customer visited the store twice

- 1.6% of the customer visited the store only thrice

- **Customers Retention is one of the major Problems to this business**



Cities vs No. of Sellers



Cities vs Total Sales

For studying the trends of seller city across the sales and no.of sellers, Two Bar graph plots have been plotted which clearly indicates and presents the data in easy and reliable format

We can derive the following insights from the above graph that

- Sao Paulo dominates both seller count and sales: Sao Paulo has the highest number of sellers and generates the most sales, indicating its strong market presence and influence.

- Disparity between seller count and sales: While some cities have many sellers, their total sales are relatively low, suggesting potential opportunities for increasing sales per seller.

- **Non-Uniform Distribution of Sales seems to be a major problem here**

4. Detailed Explanation of the Analysis Method

Sentimental Analysis

- Analyzing customer reviews to identify factors contributing to low customer satisfaction and retention.
- Used NLP techniques and Libraries (NTLK) on the customer reviews to identify their sentiment positive or negative
- Created a Sentimental Word Cloud for Insights into customer pain points to address operational or service issues.
- Identifying recurring themes or keywords in negative reviews (e.g., issues with delivery, product quality, or customer service).
- **Implementation**: - Creating a negative sentiment word cloud using this ntlk library and then analyzing those sentiments. (Refer More EDA Notebook & 5th Section)

Cohort Analysis

- Creating the cohort to better analysis the data
- Track the behavior of each cohort, focusing on repeat purchase rates and customer lifetime value (CLV)
- Plotting graph for various analysis done on cohort
- **Implementation:** Making cohort for customer who have visited only once & for top 20% cities and 80% cities in terms of order volume comparing & comparing various features .

AOV (Average Order Value) Analysis

- Determine the purchasing power and buying habits in various regions to uncover high-performing areas and optimize sales strategies
- Compare AOV across regions to identify disparities
- Use insights to develop tailored pricing strategies or promotional offers.

Freight Cost Analysis

- Evaluate the impact of logistics on customer satisfaction and sales performance, particularly in underperforming regions.
- Analyzing the relationship between the cities and the freight cost
- Exploring options for optimizing shipping costs, such as regional warehousing or partnerships with local carriers
- **Implementation**: - Comparing the average freight cost of top 20% and rest 80% of the cities as there is sales disparity (Refer More EDA Notebook & 5th Section)

Methods of Data Collection & Cleaning

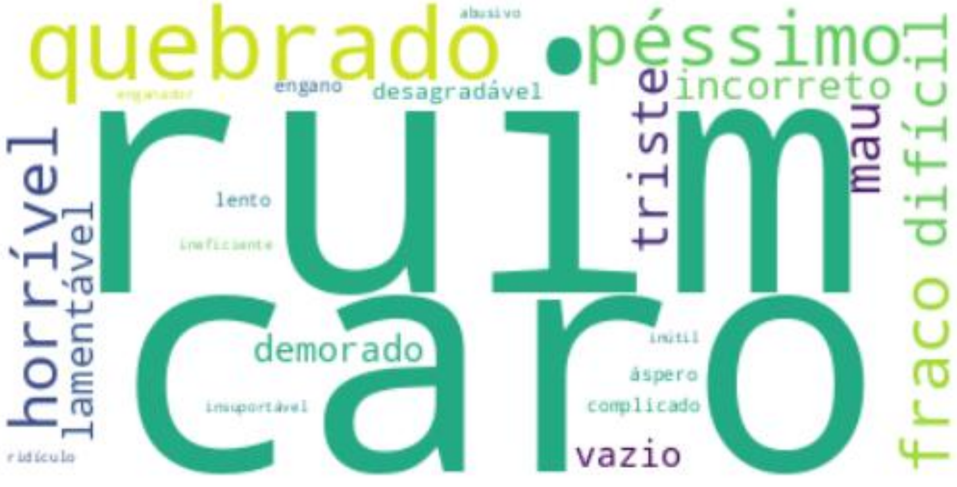
- Data Collection: - Dataset has been collected from an online site known as Kaggle
- Data Cleaning: - For data cleaning I have used SQL Queries to combine 8 table into one with relevant features and elimination or imputation various null values using python libraries internal functions

5. Results and Findings

- Here are some insights that I got while doing EDA of the dataset

Created a cohort of customers who visited or ordered once and then performed sentimental analysis

On doing **sentimental analysis on the review's comments given by the customer**, I have created two words clouds positive and negative.



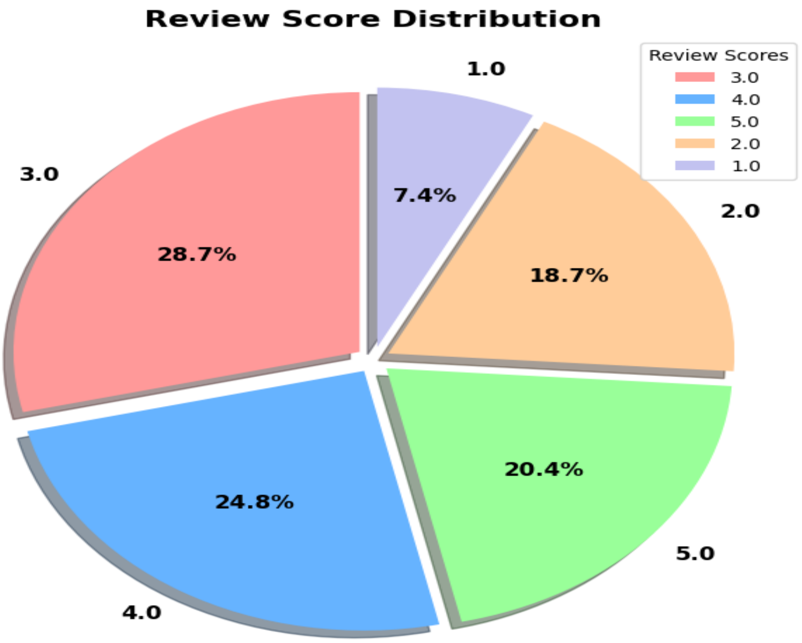
Word Cloud of Negative Words in Comments

Since these are Portuguese words, I have translated them to English

- caro – expensive - ruim – bad -

quebrado	broken
fraco	weak
difícil	difficult
lamentável	regrettable
complicado	complicated
vazio	empty
ridículo	ridiculous
inútil	useless
mau	bad
áspero	rough
péssimo	awful
lento	slow
triste	sad
incorreto	incorrect
enganador	deceptive
desagradável	unpleasant
demorado	delayed
abusivo	abusive
insuportável	unbearable

Review Score analysis for the single visit customers



Review Score Distribution of Single Visit Customers

Since we have analyzed the comments of the customers who have visited our stores once only. Let’s move to analyzing the next problem non- uniform distribution of the sales

Calculation: Order Percentage from Top 20% Cities

```
import pandas as pd

city_order_counts = city_order_counts.sort_values(
    by=["order_count"], ascending=False
)
total_orders = city_order_counts["order_count"].sum()
top_20_percent_cities = city_order_counts.head(
    int(len(city_order_counts) * 0.2)
)
top_20_percent_orders = top_20_percent_cities["order_count"].sum()
percentage_from_top_20_percent = (top_20_percent_orders / total_orders)
* 100

print(
    f"Percentage of orders from top 20% cities:
{percentage_from_top_20_percent:.2f}%"
)
```

After doing Calculation we came to know 86.83% of the sales come from top 20% percent cities.

Calculation: - Average Freight Cost in Top 20% Cities Vs Rest 80% Cities

Note: - Top in terms of order count

After doing some Calculation on Freight Cost, for calculations refer to the EDA Notebook

<u>We can conclude that</u>

- Average freight cost for top 20% cities: 19.92
- Average freight cost for rest 80% cities: 28.60

Calculation: - AOV in Top 20% Cities vs 80% Cities

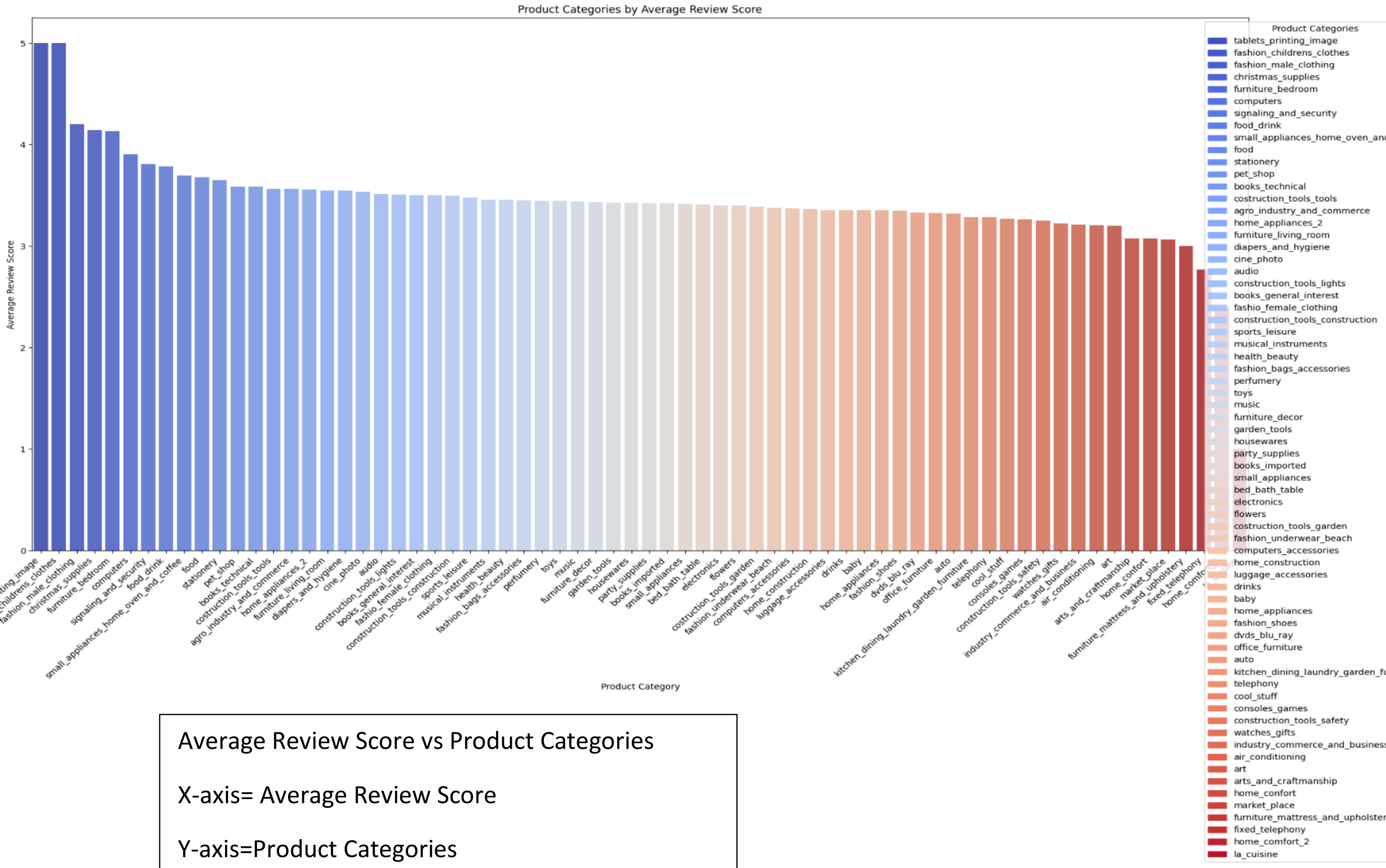
After doing some Calculation on AOV, for calculations refer to the EDA Notebook

<u>We can conclude that</u>

- Average order value for top 20% cities: 139.22
- Average order value for rest 80% cities: 148.98

The above code gives the average order value for the top 20 cities and the rest 80% cities as we can see there isn’t much difference in the order value

Product Categories by Average Score



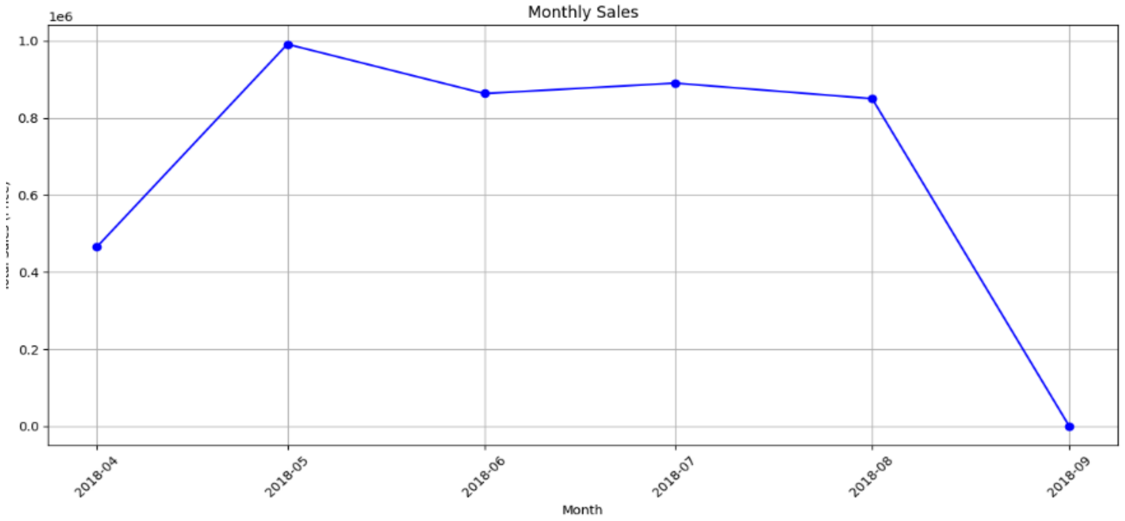
Average Review Score vs Product Categories

X-axis= Average Review Score

Y-axis=Product Categories

To get a better picture of the categories are performing a bar graph has been plotted

Sales Per Month Analysis



X-axis= Month

Y-axis= Total Sale (10^6)

- Sales seem to be okay, except for the first and last month there isn’t much variations in the sales.
- Initial peak rise in the April month and down in September may be due to data is taken after mid of April and till before mid of September. For more information, please refer to the query in the meta data & statistics section

FINDINGS

- Sentiment analysis of customer review comments revealed frequent use of negative terms such as "caro" (expensive), "ruim" (bad), "quebrado" (broken), "péssimo" (awful), "fraco" (weak), "lento" (slow), and "difícil" (difficult), indicating dissatisfaction with product pricing, quality, and performance

- The most common negative sentiments highlight key issues such as perceived high prices, defective products, slow service, and challenging user experiences, which may be contributing to low customer retention rates.

- Words such as expensive and slow highlight the shipping challenges that customers face.

- More than 50% of the product categories have a rating less than or equal to 3.

- Sales are stable overall, with April's peak and September's decline likely due to mid-month data collection, as detailed in the metadata section

6. Interpretation of results and recommendation

- Average sales seem to be consistent over the time series of 4 months data taken for the analysis.

- Categories on the right (e.g., "la_cousines_and_home_comfort_2 ", "fixed telephony", and "furniture mattress") show significantly lower review scores, with some close to 1.0. These categories may have issues related to product quality, delivery, or customer service.

- The average review scores display a clear decline from left to right, indicating substantial variability in customer satisfaction across different product categories. This suggests that specific categories outperform others in terms of quality or user experience.

- The average order value is slightly lower in the top 20% cities (139.22) compared to the rest 80% cities (148.98), suggesting that higher sales volume in top cities may come from more frequent but smaller orders.

- The average freight cost for the top 20% cities (19.92) is significantly lower than that of the rest 80% cities (28.60), indicating better logistics efficiency or infrastructure in these high-performing regions.

- Higher order volumes in the top 20% cities might enable economies of scale, reducing per-order shipping costs.

- The higher freight cost in the rest 80% of cities could highlight logistical inefficiencies, such as longer distances, poor infrastructure, or higher delivery challenges.

- Sao Paulo dominates both seller count and sales: Sao Paulo has the highest number of sellers and generates the most sales, indicating its strong market presence and influence.

- Disparity between seller count and sales: While some cities have many sellers, their total sales are relatively low, suggesting potential opportunities for increasing sales per seller.

Recommendation

- Optimize Logistics with Warehouses: Open warehouses in key strategic locations and redistribute sellers across cities to reduce delivery time and freight costs, especially in underperforming regions

- Poor-Quality Products: Identify and remove products with consistently bad reviews (e.g., *"broken," "useless,"* and *"deceptive"*), ensuring only high-quality items remain on the platform. (e.g., "la_cousines_and_home_comfort_2 ", "fixed telephony", and "furniture mattress")

- Align Seller Distribution with Regional Demand: Encourage sellers to set up operations or stock inventory closer amidst of far regions. This reduces shipping times and costs, enhancing customer satisfaction.

- Optimize Delivery Network and Partnerships: Regional warehouses or even fulfillment centers in high-demand areas reduce last mile delivery costs. Local couriers can then be collaborated with and hybrid models implemented, be it parcel lockers, crowd-sourced delivery amongst others to cut costs as well as stay flexible enough

- Proactive Competitor Awareness: Continuously monitor competitors' strategies, pricing, product offerings, and customer reviews to identify gaps and opportunities.

*********************THANK YOU******************************