EDA on Brazilian E-Commerce Dataset By Olist

By

Syed Waiz UL Hasan

Content

Dataset

Overview:

- This dataset contains real, anonymized commercial data from the Olist Store, a leading Brazilian e-commerce platform.
- It covers over 100,000 orders made between 2016 and 2018 across multiple marketplaces in Brazil.
- Offers a multi-dimensional view of e-commerce activities, including:
  - Order status, payment, and pricing
  - Delivery logistics and freight value
  - Customer reviews and ratings
  - Product details and seller information

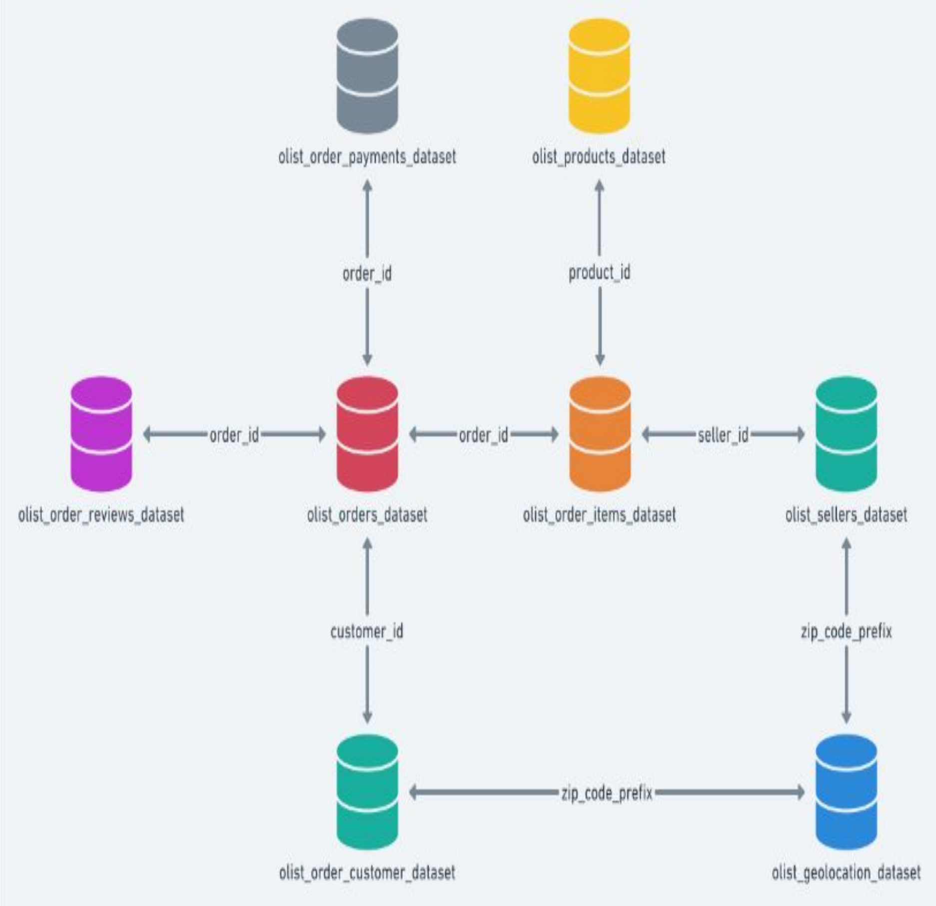
🏢 Data Provider – Olist:

- Olist connects small and medium businesses in Brazil to major marketplaces via a single platform.
- Merchants sell through the Olist Store and deliver products using Olist's logistics partners.
- After delivery, customers provide feedback through email surveys, enabling customer experience tracking.

Dataset(contd)

Included Datasets (9 Total):

1. **Customers** – Unique customer IDs and location details
2. **Orders** – Central dataset linking all order-related data
3. **Order Items** – Info about products purchased per order
4. **Payments** – Type and number of payment installments
5. **Reviews** – Review scores and written feedback from customers
6. **Products** – Product categories, dimensions, and weight
7. **Sellers** – Seller IDs, locations, and order fulfillment roles
8. **Geolocation** – Zip codes mapped to latitude & longitude
9. **Product Category Translation** – English translation of product categories



Problem Statement

- **Using Exploratory Data Analysis (EDA) techniques, we will explore and visualize the Brazilian eCommerce dataset by Olist. Our focus will be on identifying the key factors that influence customer satisfaction, operational efficiency, and overall marketplace performance. This analysis will help in deriving insights to support business decisions, enhance customer experience, optimize logistics, guide marketing strategies, and improve seller (vendor) performance on the platform.**

- **As part of the analysis, we will attempt to answer the following questions for the Brazilian E-Commerce data set:**

1. **Correlation between the columns**

2. **What is the Customer Distribution By State?**

Problem Statement(cont.)

3.What are the number of sellers in each state?

4.Does the Order Status impact the Customer Satisfaction and Review Score?

5.What key themes and terms dominate customer reviews, and what do they reveal about customer satisfaction?

6.Which payment methods are most commonly used by customers, and what does this reveal about their preferences and behavior?

7.What are the most popular product categories on Olist, and how do their sales volumes compare to each other?

Steps involved in our EDA-

1. Importing Necessary Libraries – NumPy, Pandas, Matplotlib, Seaborn,Geobr,Plotly

2. Importing Airbnb Booking csv file in Google Collab

3. Understanding the Data

4. Data Cleaning

5. Data Analysis and Visualization

Understanding the Data

- Olist dataset is huge with around 117,329 row entries(orders) and 39 columns(after merging).

- Different columns are of various data types.

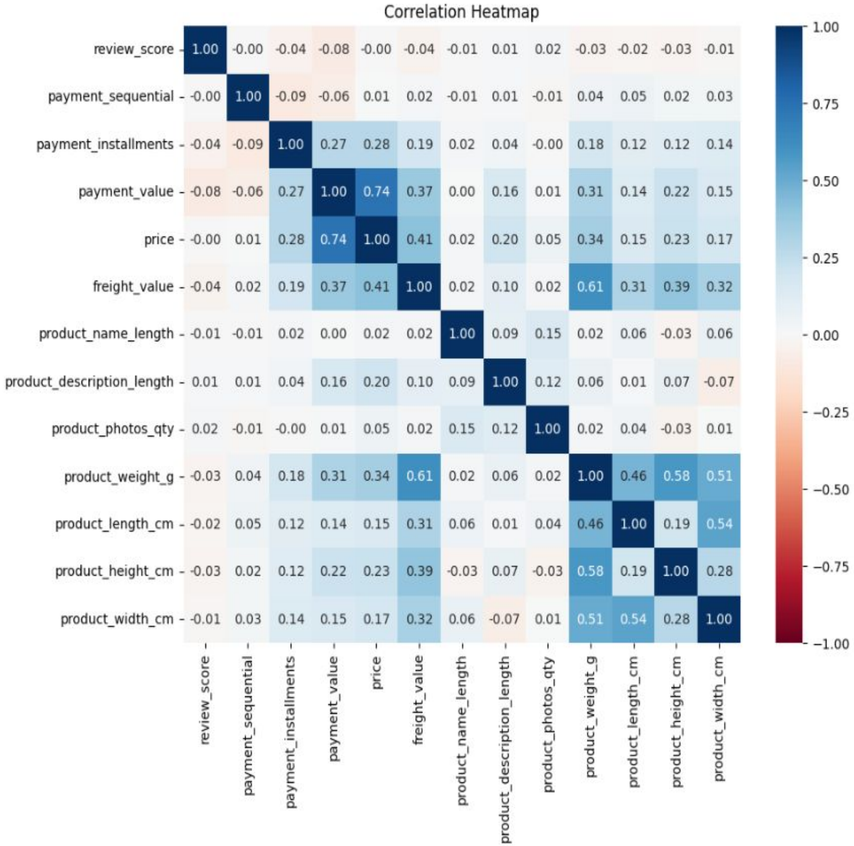
- There are significant NaN values in some columns.

- Some columns are not significant for more in depth analysis .

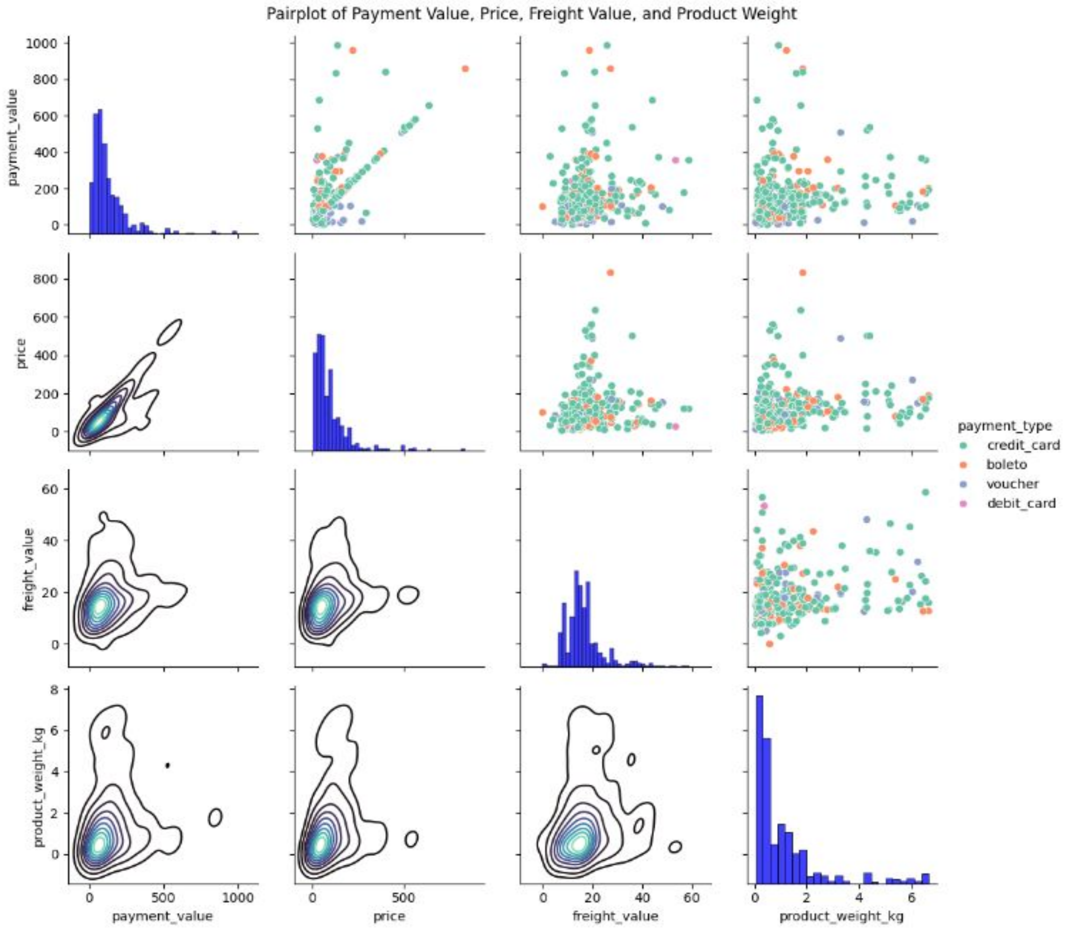
	Column	dtypes	#Missing	#Unique	Example
0	order_id	object	0	97916	e481f51cbdc54678b7cc49136f2d6af7
1	customer_id	object	0	97916	9ef432eb6251297304e76186b10a928d
2	order_status	object	0	7	delivered
3	order_purchase_timestamp	object	0	97370	2017-10-02 10:56:33
4	order_approved_at	object	15	89533	2017-10-02 11:07:15
5	order_delivered_carrier_date	object	1235	80449	2017-10-04 19:55:00
6	order_delivered_customer_date	object	2471	95021	2017-10-10 21:25:13
7	order_estimated_delivery_date	object	0	449	2017-10-18 00:00:00
8	review_id	object	0	97708	a54f0611adc9ed256b57ede6b6eb5114
9	review_score	int64	0	5	4
10	review_comment_title	object	103437	4497	NaN
11	review_comment_message	object	67650	35691	Não testei o produto ainda, mas ele veio corre...
12	review_creation_date	object	0	632	2017-10-11 00:00:00
13	review_answer_timestamp	object	0	97546	2017-10-12 03:43:48
14	payment_sequential	int64	0	29	1
15	payment_type	object	0	4	credit_card
16	payment_installments	int64	0	24	1
17	payment_value	float64	0	28831	18.12
18	customer_unique_id	object	0	94720	7c396fd4830fd04220f754e42b4e5bff
19	customer_zip_code_prefix	int64	0	14955	3149
20	customer_city	object	0	4108	sao paulo
21	customer_state	object	0	27	SP

Data Analysis and Visualization

- There is no strong correlation between most feature pairs in the dataset.
- As expected, the correlation of a column with itself is always 1.
- However, a few moderately strong correlations were observed:
  - payment_value shows a positive correlation with both price (0.74) and freight_value (0.37).
  - Product dimensions and weight are moderately correlated:
    - product_weight_g ↔ freight_value (0.61)
    - product_weight_g ↔ product_height_cm (0.58)
    - product_weight_g ↔ product_width_cm (0.51)
    - product_length_cm ↔ product_width_cm (0.54)



Correlation Analysis Of The payment Values



Breakdown of The Pairplot

Pair Plot Analysis Summary (sns.PairGrid)

🔷 **Plot Components:**

- **Diagonal** – Histograms: Distribution of each feature
- **Upper Triangle** – Scatter Plots: Relationship between feature pairs
- **Lower Triangle** – KDE Plots: Density regions between feature pairs

Scatter Plot Insights (Upper Triangle)

- **payment_value vs. price**
  → Strong positive correlation (values lie close to a diagonal)
- **payment_value vs. freight_value**
  → No clear trend; high shipping cost can occur for any payment value
- **payment_value vs. product_weight_kg**
  → No pattern; weight doesn't always influence payment value
- **price vs. freight_value**
  → Slight correlation; some costly products have high freight
- **price vs. product_weight_kg**
  → No strong relation; price doesn't depend on product weight
- **freight_value vs. product_weight_kg**
  → Clear positive correlation (heavier = more shipping cost)

Breakdown of The Pairplot(Contd)

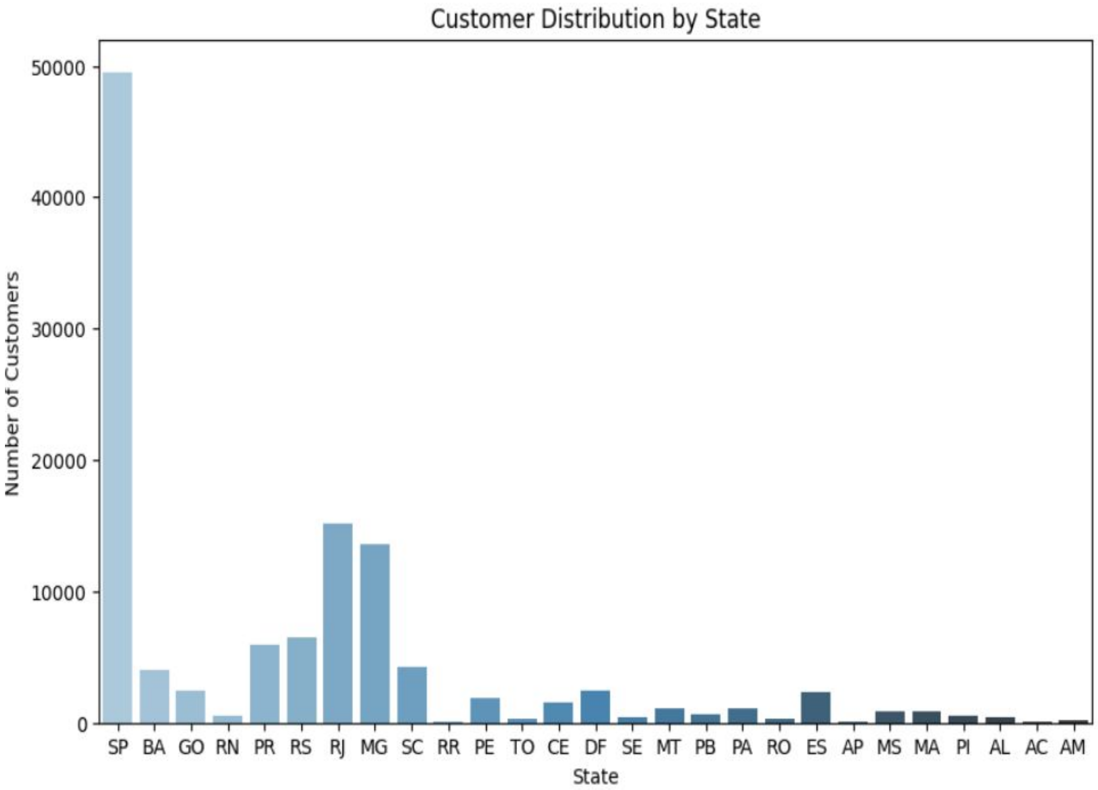
KDE Plot Insights (Lower Triangle)

- **payment_value vs. price**
  → High density along diagonal (confirms strong correlation)

- **price vs. freight_value**
  → Density concentrated in low price & low freight region

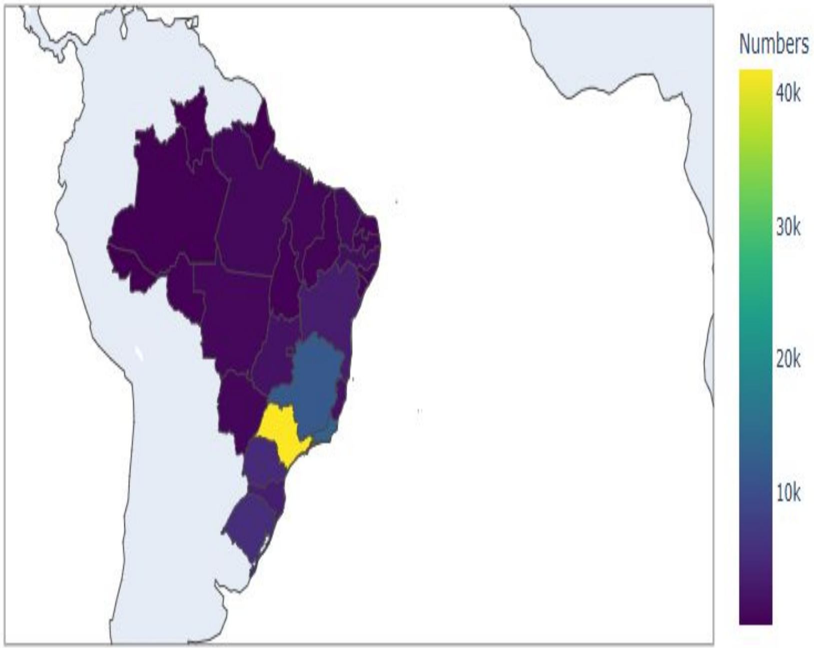
- **freight_value vs. product_weight_kg**
  → Dense regions support heavier items costing more to ship

Customer Distribution by State



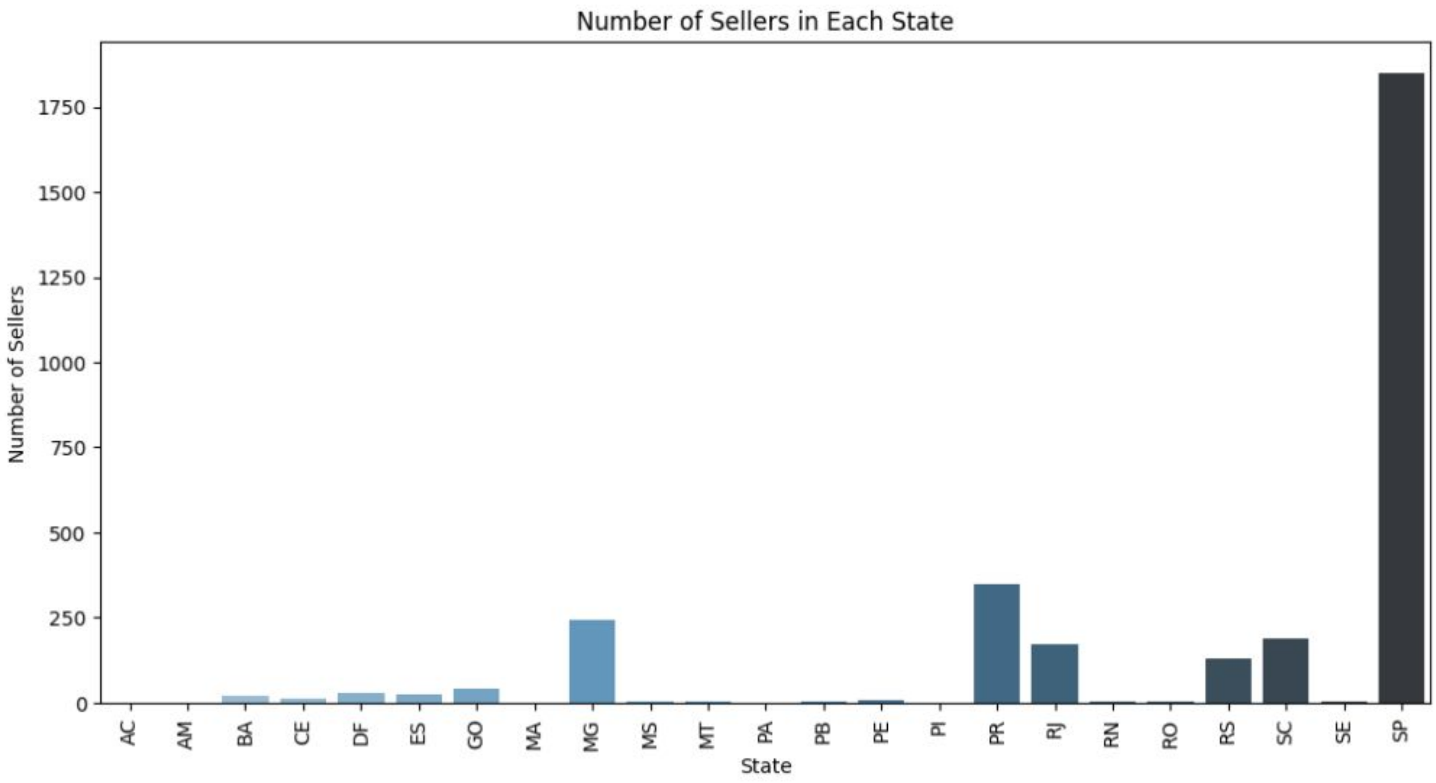
The result indicates that the customer distribution by state is not even. It suggests that some states have a higher concentration of customers than others. This can be caused by a variety of factors:

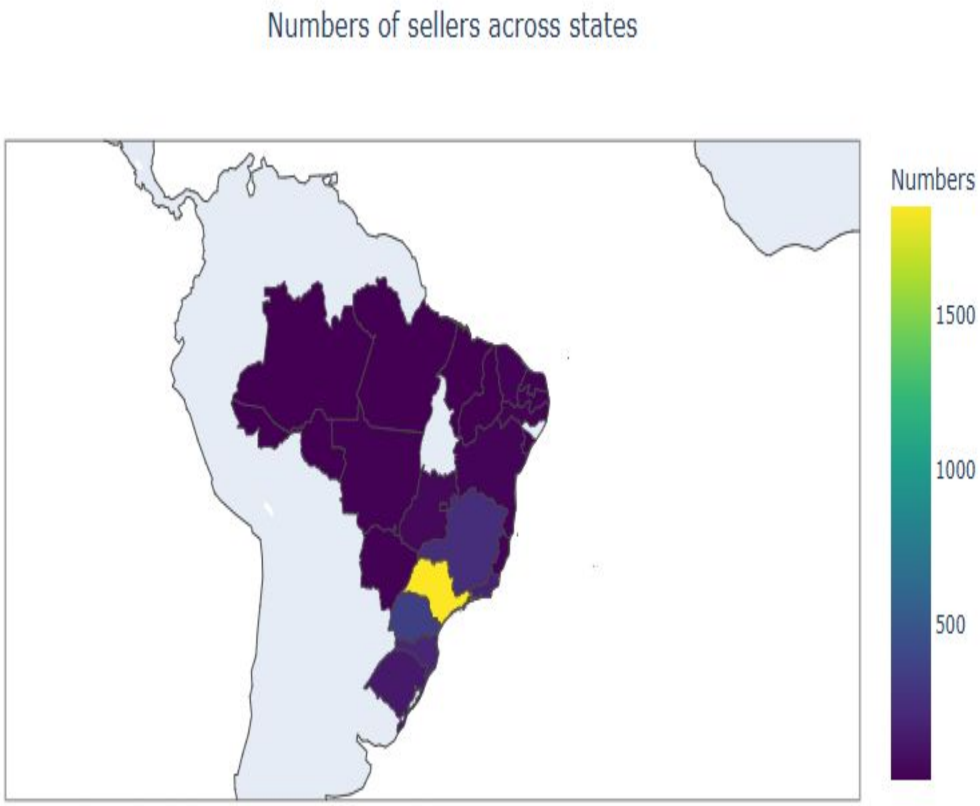
1. **Population Size or Density**: Differences in the population size or density of different states can lead to uneven distribution.
2. **Availability or Accessibility**: Variations in the availability or accessibility of the product or service being offered.
3. **Marketing or Sales Efforts**: Differences in marketing or sales efforts in different regions.
4. **Customer Preferences or Needs**: Preferences or needs of customers in different states.



- **São Paulo** leads with **~40k customers**, far ahead of other states.
- **Rio de Janeiro** and **Minas Gerais** follow with **~10k customers** each.
- Major customer clusters are found **near coastal cities and trade hubs**, likely due to better infrastructure and accessibility.
- Customer distribution across other states is **fairly uniform but lower** in volume.

Number Of Sellers in Each State





Seller-Customer Overlap: Seller distribution closely mirrors customer distribution, highlighting the influence of economic hubs.

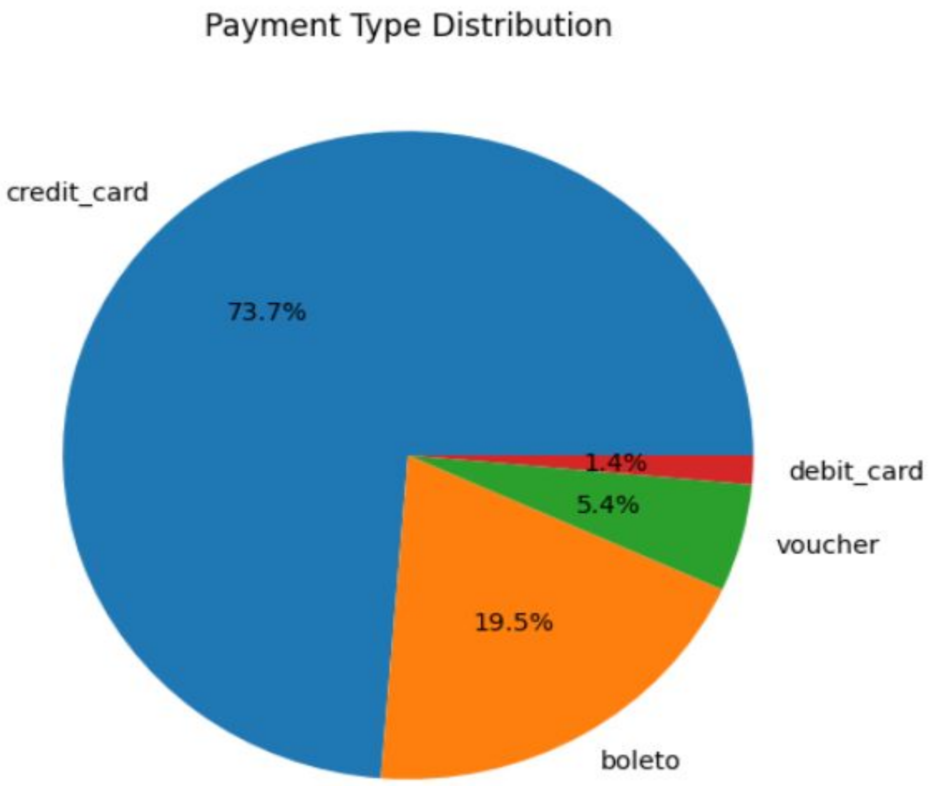
São Paulo Leads: With ~1.8K sellers, São Paulo dominates due to its strong infrastructure and high demand.

Regional Support: Rio de Janeiro and Minas Gerais also support large seller bases (~0.5K), driven by population and logistics.

Harbour Clusters: Sellers and customers are concentrated near coastal cities, indicating the importance of ports and trade routes.

Customer Spread: Customers are more evenly distributed across states compared to sellers.

Payment Type Distribution



- Credit cards dominate (73.7%), showing strong customer preference.
- Debit cards (1.4%) and vouchers (5.4%) are underused.
- Boleto (unlabeled) appears minimal, suggesting niche usage.

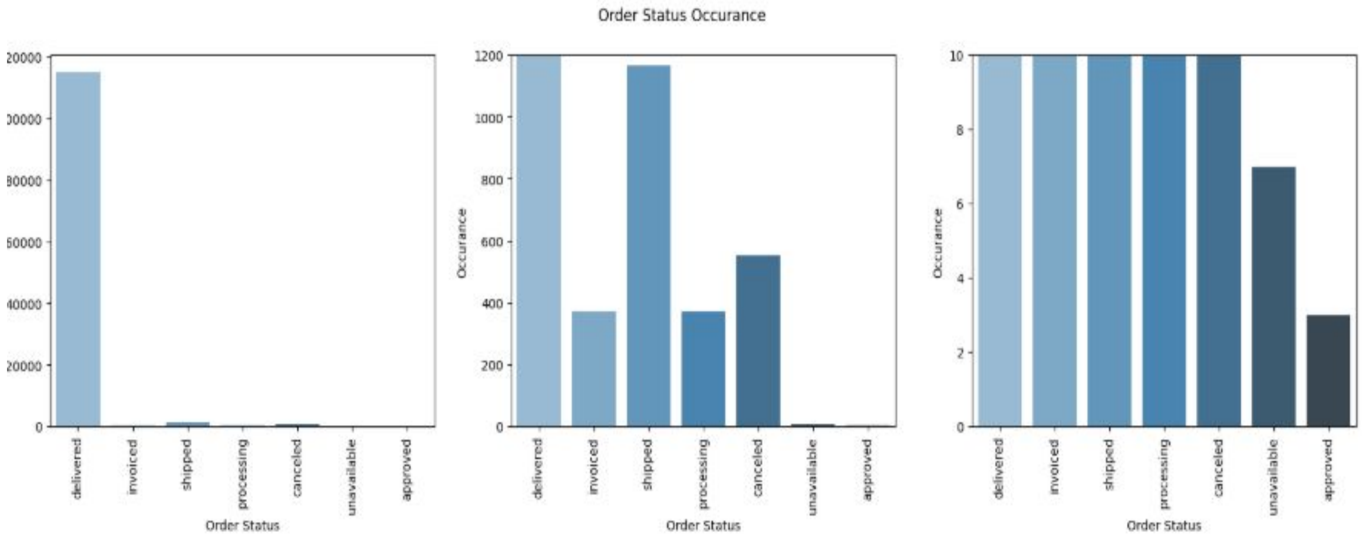
Order Status Impact On Review Ratings

```
The mean review score for "shipped" order is 1.977720651242502
The mean review score for "canceled" order is 1.5949367088607596
The mean review score for "invoiced" order is 1.654054054054054
The mean review score for "processing" order is 1.3486486486486486
The mean review score for "unavailable" order is 1.5714285714285714
The mean review score for "approved" order is 2.0
```

Non-delivered orders (like *canceled*, *unavailable*, *processing*) have **low average ratings (≤ 2)**.

These statuses may signal **failed or delayed deliveries**, affecting **customer experience**.

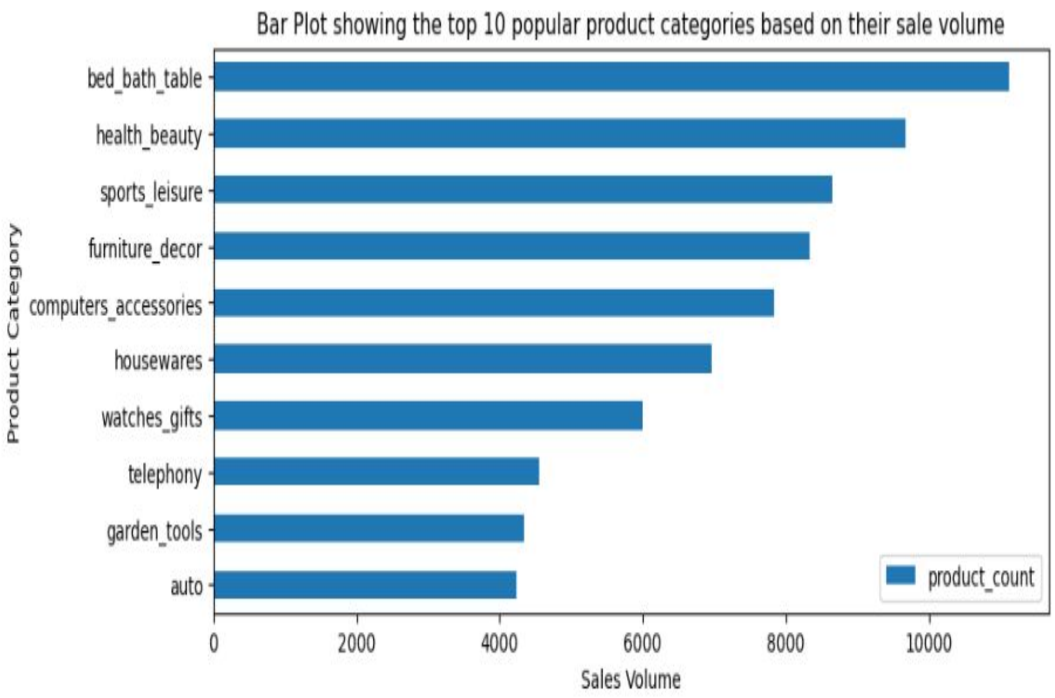
However, these special cases are **rare** in the dataset.



Word Cloud of Most Word Used(Reviews)

- The most frequently used words are centered around **product delivery and receipt**.
- Terms like **"received"**, **"arrived"**, and **"delivered"** suggest that **timely delivery** is a crucial aspect of customer satisfaction.
- Positive sentiment is reflected in words like **"recommend"**, **"quality"**, and **"good"**.
- Mentions of **"store"** and **"bought"** indicate **shopping experience** is also a key focus.

Most Popular Product Categories On Olist (by Sales Volume)



1. **Top Performers:**
   - bed_bath_table and health_beauty dominate sales, indicating strong demand for home essentials and personal care products.
   - sports_leisure and furniture_decor follow, suggesting steady interest in lifestyle and home improvement.
2. **Mid-Range Categories:**
   - computers_accessories, housewares, and watches_gifts show moderate sales, reflecting niche but consistent demand.
3. **Lower-Volume Categories:**
   - telephony, garden_tools, and auto trail significantly, hinting at either limited market interest or untapped potential.

Challenges Faced

- Understanding the meaning of some columns.

- Dealing with Null values and duplicates.

- Also, forming different graphs to show insights from the dataset and to summarize the information and communicate the results and trends to the reader successfully.

Conclusions

1.States located near the harbor—São Paulo, Rio de Janeiro, and Minas Gerais—have the highest number of customers and sellers.

2.Focus marketing and logistics efforts in São Paulo, Rio de Janeiro, and Minas Gerais to maximize reach and efficiency.To grow in less populated states, consider targeted campaigns or faster delivery options to increase engagement and trust.

3. E-commerce activity is concentrated in economically strong and logistically connected states. Strategic focus on these hubs can boost growth, while improving logistics in underrepresented states could unlock new markets.

4.Non-delivered orders (e.g., *canceled* **,** *unavailable* **,** *processing* **) are linked to significantly lower customer satisfaction (ratings ≤ 2), likely due to delivery issues. Although these cases are rare, addressing them can further enhance overall customer experience and trust.**

5.Delivery reliability is key to customer satisfaction, with most reviews showing a positive tone.Consistent delivery and quality should remain top priorities.

Conclusions(cont.)

6.Since 73.7% of customers prefer credit cards, optimizing the checkout experience for card payments is essential. However, to encourage diversity and inclusivity, offering incentives for alternative methods (like vouchers or boletos) can help reach a broader audience and increase overall sales.

7.Boost top-sellers like bed_bath_table and health_beauty with inventory and marketing.Review underperformers like garden_tools and auto for seasonal or marketing gaps.Promote mid-tier categories with bundles or discounts.

Thank You