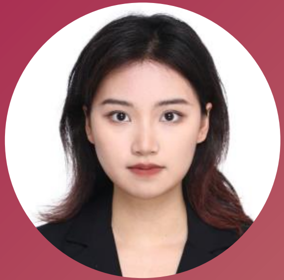
Enhancing E-Commerce Recommendation

A Tailored Shopping Experience

Dec 8, 2023

Presenters: Jessy Hu, Serena Shi,
Yiwen Song, Jiayi Deng

Meet Our Team



Yiwen Song
Data Engineer
ysong64@uchicago.edu

Jessy Hu
Business Intelligence Analyst
jiaqi722@uchicago.edu

Jiayi Deng
Data Scientist
jiayideng@uchicago.edu

Serena Shi
Data Scientist
yutong0121@uchicago.edu

AGENDA

1 Executive Summary & Research Objectives

2 Data Ingestion & Preparation

3 Data Modeling & Design

4 Insights

5 Analysis & Recommendations

6 Future Work

01

Executive Summary & Research Objectives

Executive Summary

- This project delves into the dynamic Brazilian e-commerce market, focusing on Olist, a platform that integrates small businesses into larger online marketplaces.
- Utilizing a rich Kaggle dataset alongside Brazil's macroeconomic indicators, including household income, we aim to uncover deep insights into consumer purchasing patterns and market trends.
- Our analysis includes a detailed RFM (Recency, Frequency, Monetary) assessment to enhance customer segmentation and purchasing behavior understanding.
- The objective is to derive strategic recommendations for Olist, targeting product portfolio optimization, marketing strategies, and overall sales enhancement to solidify their competitive edge in Brazil's flourishing e-commerce sector.

Research Objectives

Our research aims to conduct a comprehensive analysis of Brazil's e-commerce market, focusing on Olist's operations. Key objectives include:

Customer and Sales Data Analysis

Examine the Kaggle dataset to decode customer purchasing behaviors, sales trends, and product popularity on Olist.

Macroeconomic Integration

Assess the impact of Brazil's economic conditions, especially household income, on e-commerce trends.

RFM Analysis

Apply RFM modeling for detailed customer segmentation, identifying key groups for targeted marketing.

Strategic Recommendations

Formulate actionable strategies for Olist to optimize their product offerings, marketing tactics, and sales strategies, aiming to boost their market position and sales.

Business Use Case

Role	Incentive	Business Use
E-Commerce Platform: Olist	• Understand purchasing patterns • Optimize product recommendation algorithms • Improve customer engagement and retention	Use customer behavior data to enhance recommendation engines and personalize shopping experience
Seller	• Increase sales revenue • Reduce inventory costs • Understand product performance	Analyze customer feedback and purchasing trends to optimize stock and improve product listings
Customer	• Find products more efficiently • Receive personalized recommendations • Experience a streamlined shopping process	Leverage data to provide customers with a tailored shopping experience and improve satisfaction

02

Data Ingestion & Preparation

Data Profile

Useful Information	Data Size	Rows/Cols
Customers Data: customer id, location	8.6 MB	99k / 5
Geolocation Data: zipcode, city, state	1 MB	19k / 6
Order Items Data: order, product, seller, shipping	14.7 MB	113k / 7
Order Payments Data: payment type and value	5.5 MB	104k / 5
Order Reviews Data: review score and comment	10.9 MB	99k / 7
Orders Data: order time, status, delivery	16.1 MB	99k / 8
Products Data: product profile, category, photo quantity	2.2 MB	33k / 9
Sellers Data: seller id, location	171 KB	3k / 4
Brazil Macro-economic Data: household income, human develop index, education (state-level)	2 KB	27 / 6

Tools for Data Processing

Data Ingestion

ETL Process

Data Warehouse

Analytics & Modeling

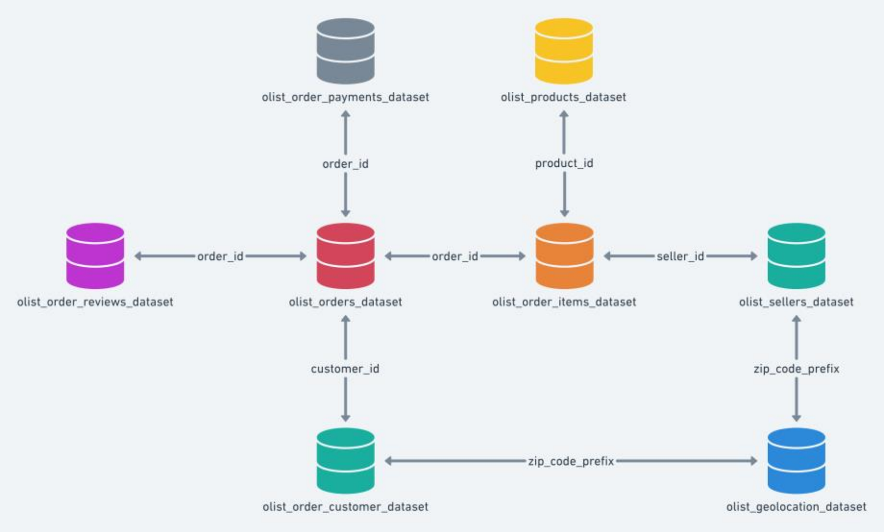
Visualization

Insights Reporting

Data Ingestion

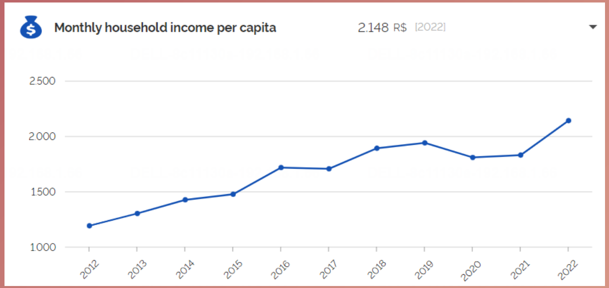
E-Commerce Data: Public Dataset by Olist

Tables connected by foreign keys



Web Scraping

Brazilian state-level macro-economic data



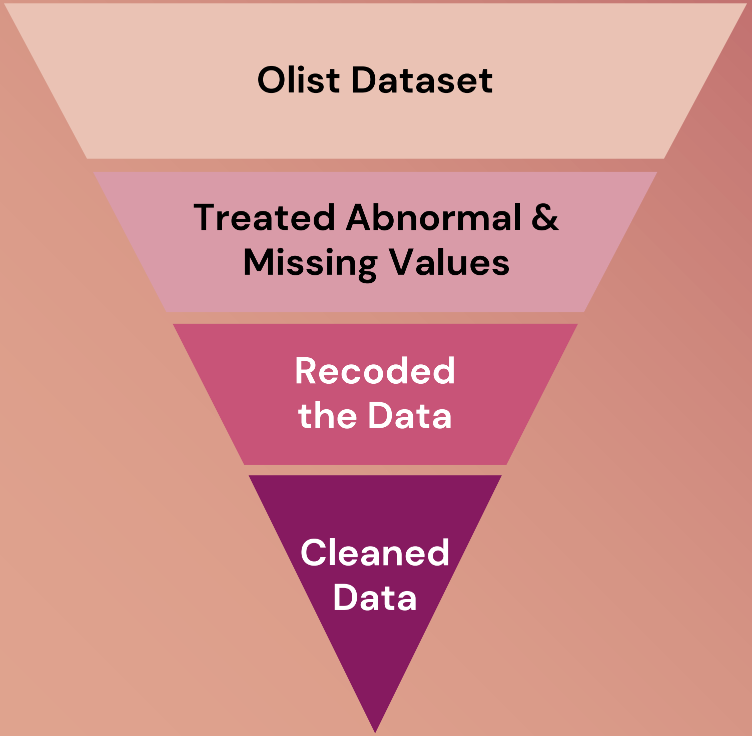
Snapshot of Column Profile

Data types, missing value, entropy of each column

	Name	dtypes	Missing	Uniques	Entropy
0	order_id	object	0	99441	16.46
1	customer_id	object	0	99441	16.46
2	order_status	object	0	8	0.25
3	order_purchase_timestamp	object	0	98875	16.44
4	order_approved_at	object	0	90734	16.27
5	order_delivered_carrier_date	object	0	81019	15.88
6	order_delivered_customer_date	object	0	95665	16.11
7	order_estimated_delivery_date	object	0	459	8.47
8	order_item_id	float64	833	21	0.72
9	product_id	object	833	32951	13.63
10	seller_id	object	833	3095	9.48
11	shipping_limit_date	object	833	93318	16.34

ETL – Process & Details

The Process



Cleaning & Transformation

Olist Data	Brazilian State-Level Macro-Economic Data
Removed emojis in customer review for loading into database	Calculated the annual average HDI, monthly household income, etc. for the years 2016-2018 (the time span of the dataset)
Supplemented missing zipcodes, cities and states (0.84% of rows) from post-code.org	Indexed all the attributes with State Abbreviations, for connecting Olist dataset
Translated product categories from Portuguese to English	Total of 27 records for all states in Brazil
Unified datetime format as yyyy/m/d hh:mm:ss	
Recoded all missing datetime values as '1900/1/1 00:00:00'	

03

Data Modeling & Design

Database Design Consideration

Database Modeling

- OLTP: Normalized physical entity-relationship model
- OLAP: Multi-dimensional snowflake model

Data Types

- Choose *VARCHAR* datatypes for primary keys and other string attributes
- Define Date attributes with datatypes *DATETIME* or *TIMESTAMP*
- Follow standard naming convention for attributes

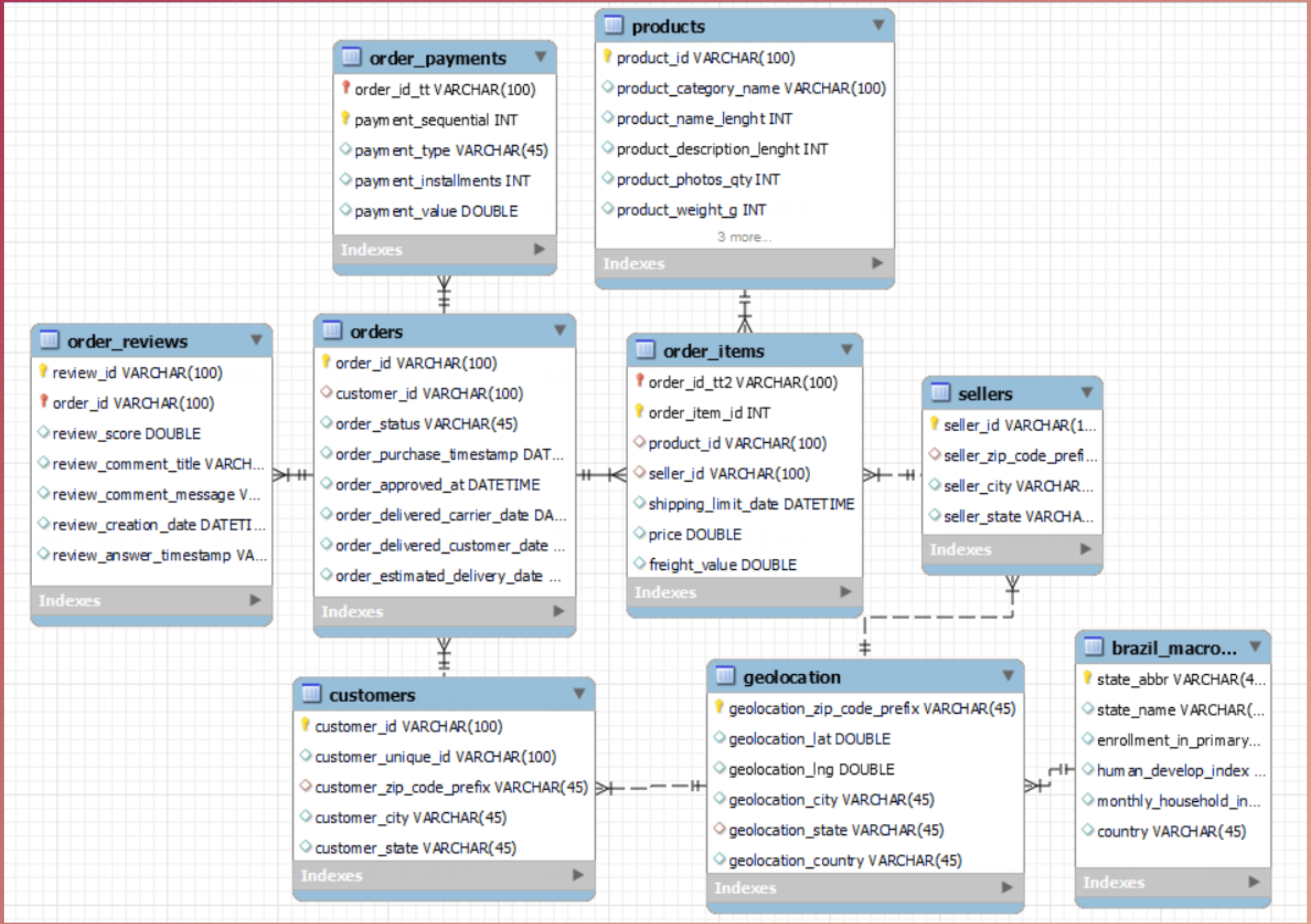
Granularity of Data

- E-commerce data is atomic in nature and can be stored in a fact table and rolled up by month/quarter

Data Integrity

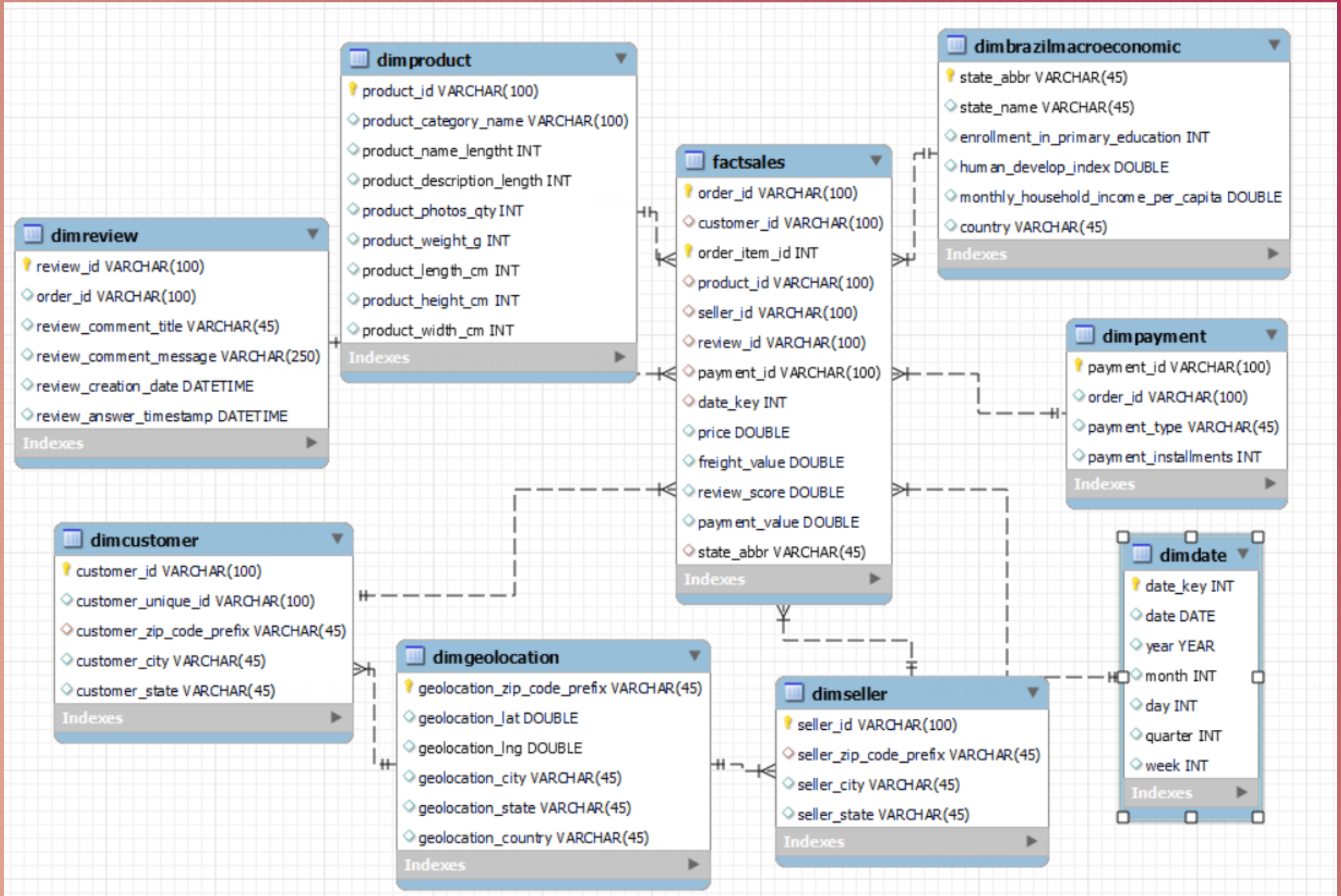
- Establish a unique primary key for each entity/table
- Set foreign key relationships and constraints (not null, unique)
- Define default values for missing values wherever applicable (e.g. -1 for *INT*, None for *VARCHAR*)

EER Diagram



order_payments
order_id_tt VARCHAR(100)
payment_sequential INT
payment_type VARCHAR(45)
payment_installments INT
payment_value DOUBLE
Indexes
products
product_id VARCHAR(100)
product_category_name VARCHAR(100)
product_name_lenght INT
product_description_lenght INT
product_photos_qty INT
product_weight_g INT
3 more...
Indexes
order_reviews
review_id VARCHAR(100)
order_id VARCHAR(100)
review_score DOUBLE
review_comment_title VARCH...
review_comment_message V...
review_creation_date DATETI...
review_answer_timestamp VA...
Indexes
orders
order_id VARCHAR(100)
customer_id VARCHAR(100)
order_status VARCHAR(45)
order_purchase_timestamp DAT...
order_approved_at DATETIME
order_delivered_carrier_date DA...
order_delivered_customer_date ...
order_estimated_delivery_date ...
Indexes
order_items
order_id_tt2 VARCHAR(100)
order_item_id INT
product_id VARCHAR(100)
seller_id VARCHAR(100)
shipping_limit_date DATETIME
price DOUBLE
freight_value DOUBLE
Indexes
sellers
seller_id VARCHAR(1...
seller_zip_code_prefi...
seller_city VARCHAR...
seller_state VARCHA...
Indexes
customers
customer_id VARCHAR(100)
customer_unique_id VARCHAR(100)
customer_zip_code_prefix VARCHAR(45)
customer_city VARCHAR(45)
customer_state VARCHAR(45)
Indexes
geolocation
geolocation_zip_code_prefix VARCHAR(45)
geolocation_lat DOUBLE
geolocation_lng DOUBLE
geolocation_city VARCHAR(45)
geolocation_state VARCHAR(45)
geolocation_country VARCHAR(45)
Indexes
brazil_macro...
state_abbr VARCHAR(4...
state_name VARCHAR(...
enrollment_in_primary...
human_develop_index ...
monthly_household_in...
country VARCHAR(45)
Indexes

Snowflake Schema



Data Quality Dimension

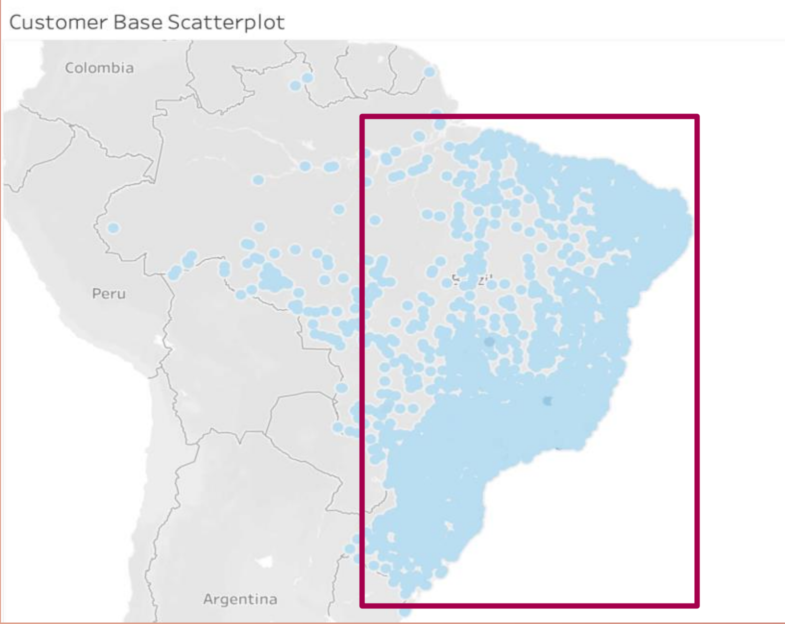
- **Completeness**: Missing values in all look-up tables are specially treated
- **Validity**: Data format and types conform to the defined business rules and constraints
- **Uniqueness**: No duplicates or redundant entries in all tables
- **Consistency**: Dimensions and data types are consistent across tables and two schemas
- **Timeliness**: Data represent reality in time as data is from relatively recent period
- **Accuracy**: Data is aggregated by summing and averaging over locations and dates, and this transformation can represent reality

04

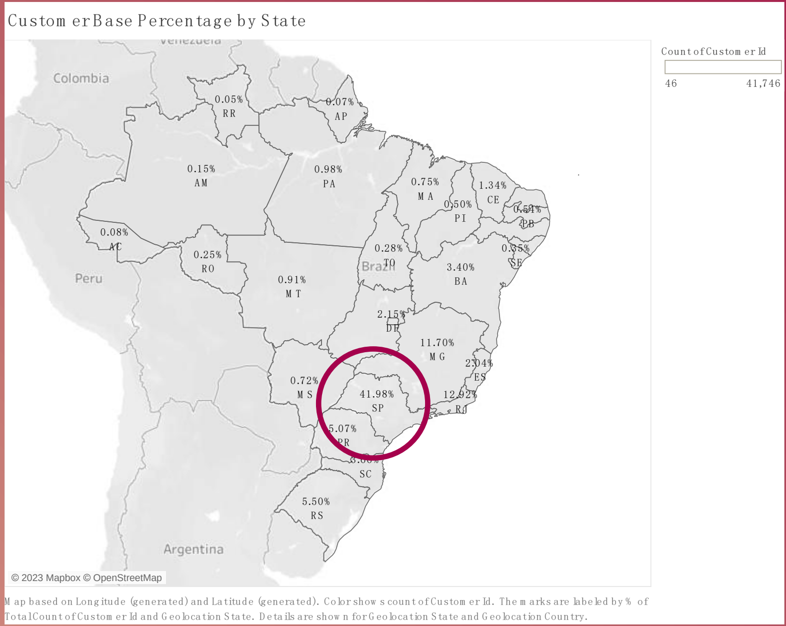
Insights

Demographics

Most Customers Locate in South

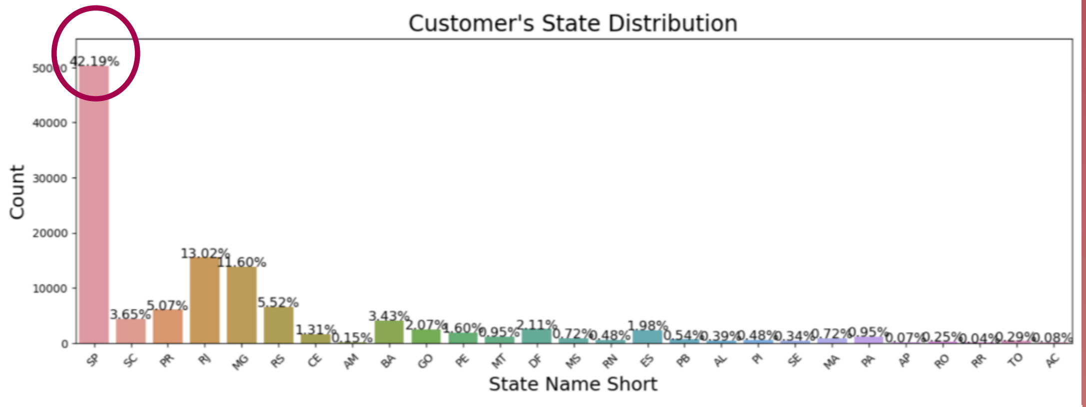


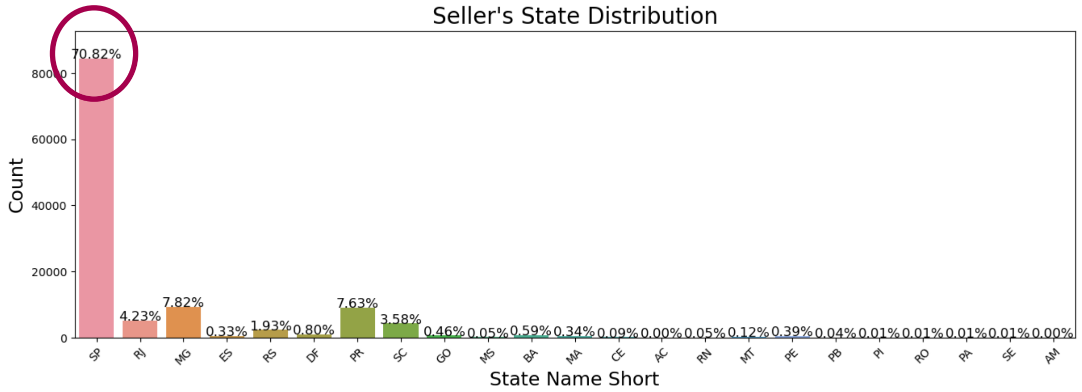
SP has 42% of Customers



SP's significant customer share offers an opportunity for localized marketing campaigns, tailored events, and regional product launches

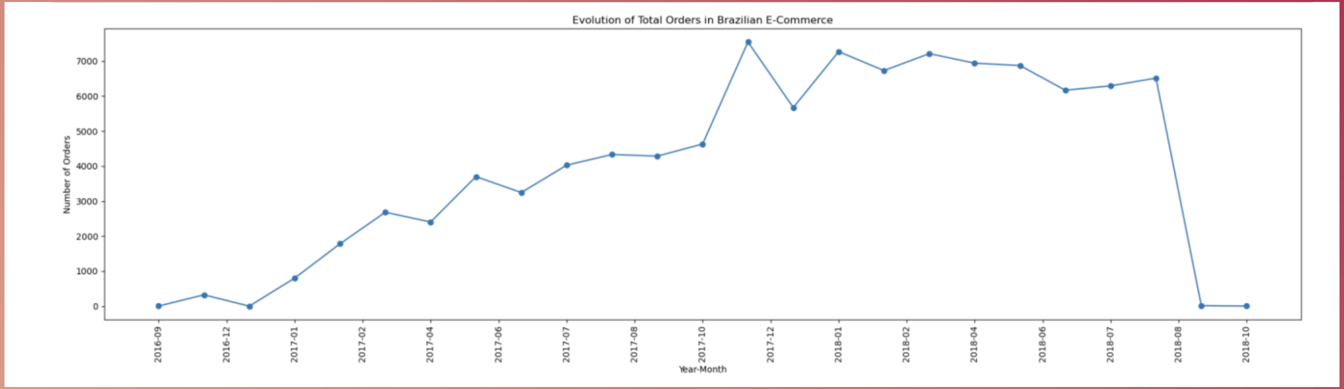
SP's massive share of both sellers and customers makes it the heart of our commercial operations

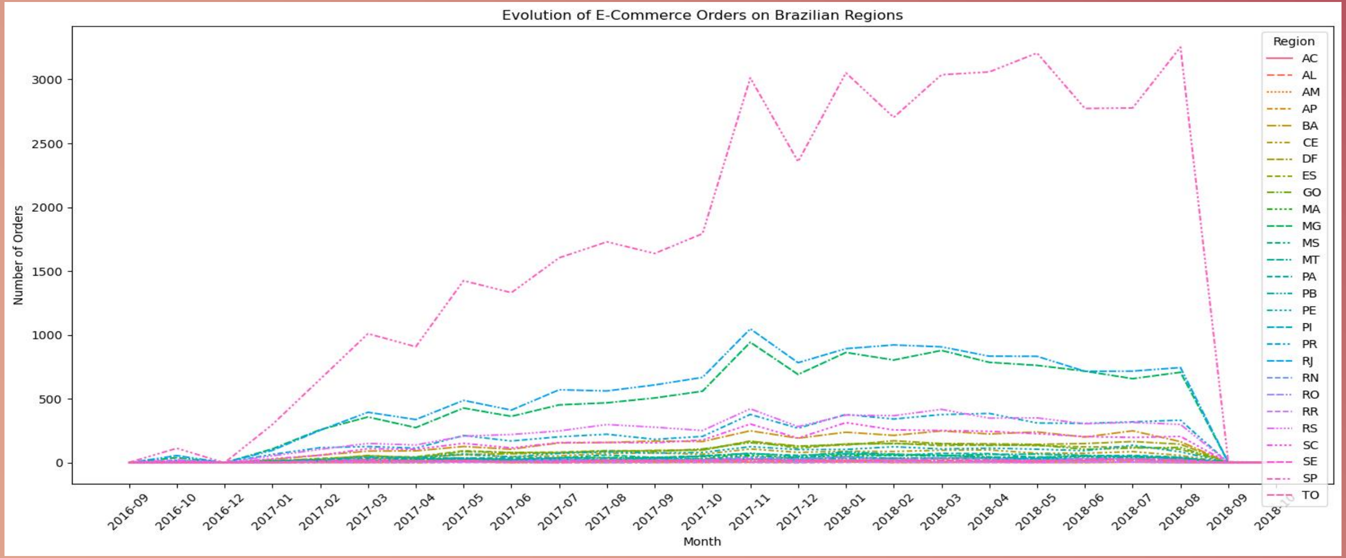




- ~42% of Customers Based in SP
- Strong regional presence
- Explore regional preferences to tailor marketing strategies and product development

- ~71% of Sellers Based in SP
- Robust supply chain efficiencies and Potential for logistics optimization
- Leverage SP's dense market to pilot new services due to the high engagement potential





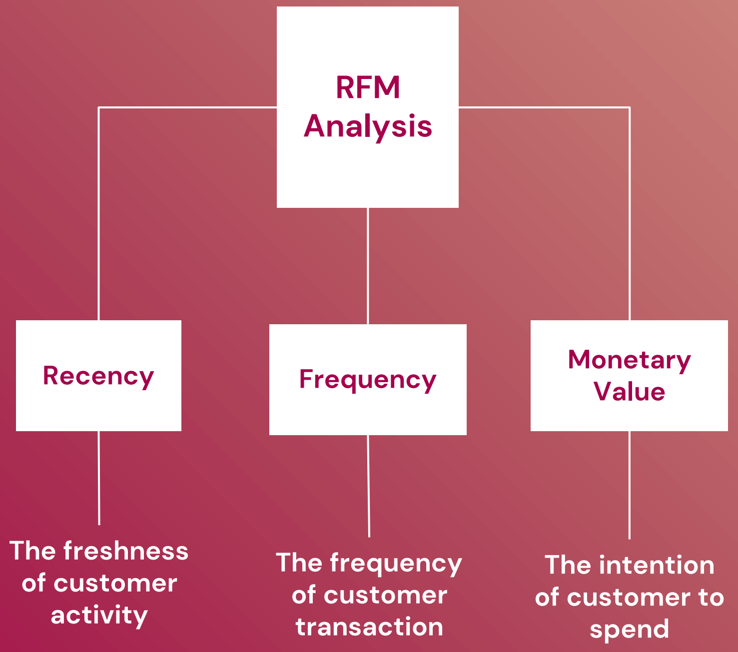
Evolution of Orders Overtime

- **Differentiated growth patterns and the market's regional diversity**
- **The varied growth rates across different regions underscore the necessity for region-specific strategies to capitalize on high-growth areas and support lagging regions**

05

Analysis & Recommendations

RFM Analysis



Filter data to contain only orders that is **delivered**

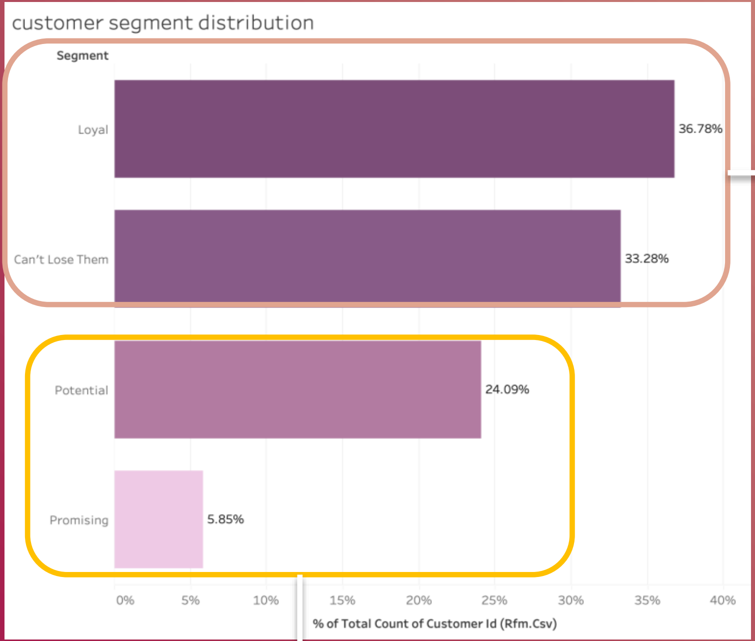
- ✓ **Recency**: Days from the most recent purchase to each order's purchase date
- ✓ **Frequency**: Total number of unique transactions of each customer
- ✓ **Monetary**: Total transaction value of each customer

```
# Calculate Recency as days from the most recent purchase to each order's purchase date
orders_df_delivered['Recency'] = (most_recent_purchase - orders_df_delivered['order_purchase_timestamp']).dt.days

# Calculate Frequency by counting unique order_ids for each customer.
frequency_df = orders_df_delivered.groupby('customer_id').size().reset_index(name='Frequency')

# Calculate Monetary value by summing the payment_value for each customer's orders.
monetary_df = order_payments_df.groupby('order_id')['payment_value'].sum().reset_index()
monetary_df = pd.merge(monetary_df, orders_df_delivered[['order_id', 'customer_id']], on='order_id')
monetary_df = monetary_df.groupby('customer_id')['payment_value'].sum().reset_index(name='Monetary')
```

Customer Segmentations



Customer Profile	Recency	Frequency	Monetary	RFM Score
Can't Lose Them	High	High	High	>9
Loyal	High	Moderate	Moderate	[7,9)
Potential Loyalists	Moderate	Moderate	Moderate	[5,7)
Promising	Low	Low	Low	[3,5)

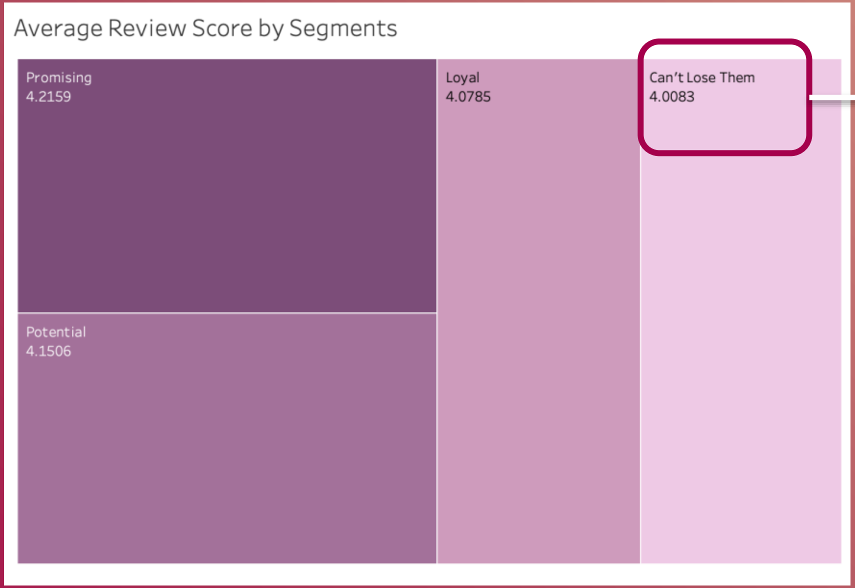
Fairly solid customer base

- With ~70% customers show their loyalty to the website

Recommendation strategy needed

- Stimulate potential loyal customer's purchase

Incentive Paired with Segmentations

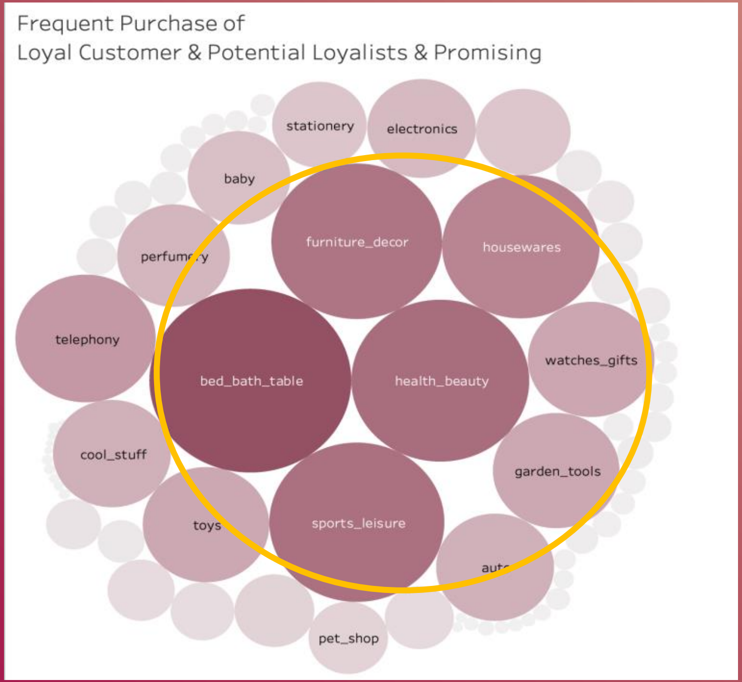


Most loyal customer segmentation has the lowest average review score

Strategy for this segmentation:

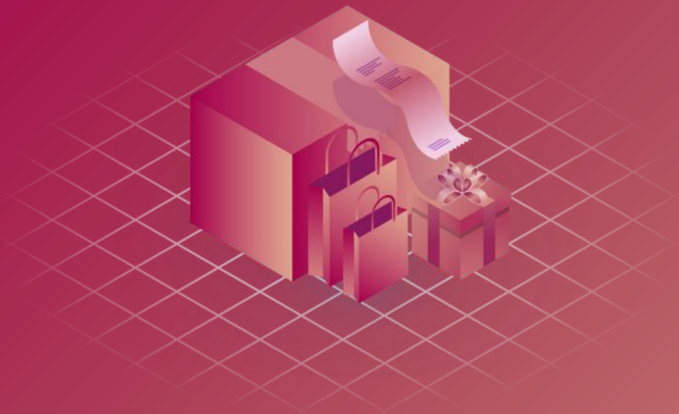
- Seek feedback on products and services to show company value their opinion.
- Offer exclusive loyalty programs or VIP status.
- Provide early access to new products or sales.

Incentive Paired with Segmentations



Strategy for Loyal & Potential Loyal & Promising Loyal Customers:

- Offer loyalty discounts or rewards for frequent purchases. (e.g. furniture décor, bed&bath, sports, etc..)
- Engage with personalized email campaigns that showcase products similar to their previous purchases.



06

Future Work & Preliminary Plan

Preliminary Plan

Request New Data of Customer Demographics

Week 1

Week 2–3

Data Analysis and Modeling

Develop Collaborative Filtering for Recommendation Engine

Week 4–6

Week 7–10

A/B Testing and Evaluate Customer Feedback

Deliver Recommendation Engine

March

Future Work & Lessons Learned

- The customer segmentation can be more accurate with additional data on customer demographics, e.g., age, race, sex, income level, etc..
- Recommendations can be more accurate with the utilization of a collaborative filtering method, which would recommend a similar product that is bought by similar customers.
- Considering the scalability and data structure, we should also implement NoSQL to cope with larger data volume and unstructured elements (e.g., customer reviews, product descriptions).
- Regular cleanup should be conducted to archive old and used data, which can help manage the growth of attributes.

References

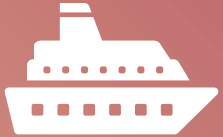
- State-level macro-economic data from the Brazilian Institute of Geography and Statistics: https://www.ibge.gov.br/en/cities-and-states/sp.html
- Olist Store official website: https://olist.com/
- Public e-commerce data from Olist Store: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data
- Most valued e-commerce and direct-to-consumer unicorns in Brazil as of October. (2022) https://www.statista.com/statistics/1282103/highest-valued-e-commerce-startup-companies-brazil/
- Leading unicorn companies based on market value in Latin America. (2023) https://www.statista.com/statistics/1028116/latin-america-unicorn-companies-market-value/

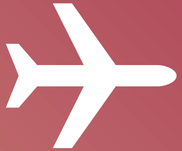
THANKS

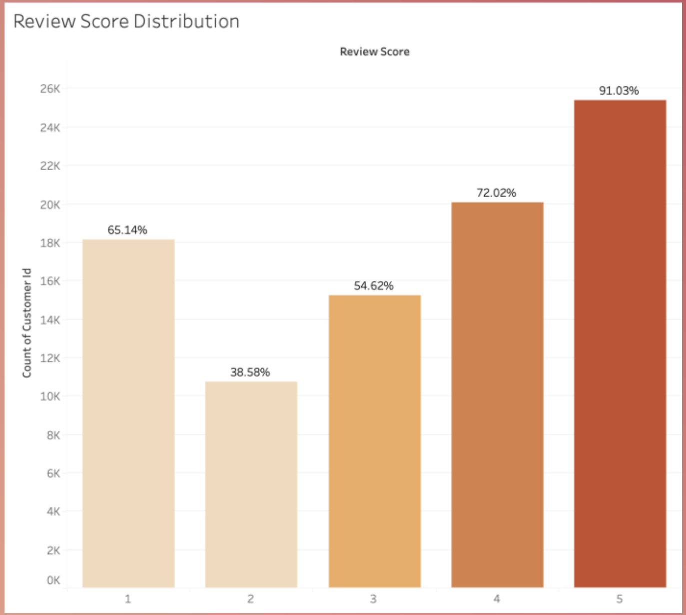
Appendix

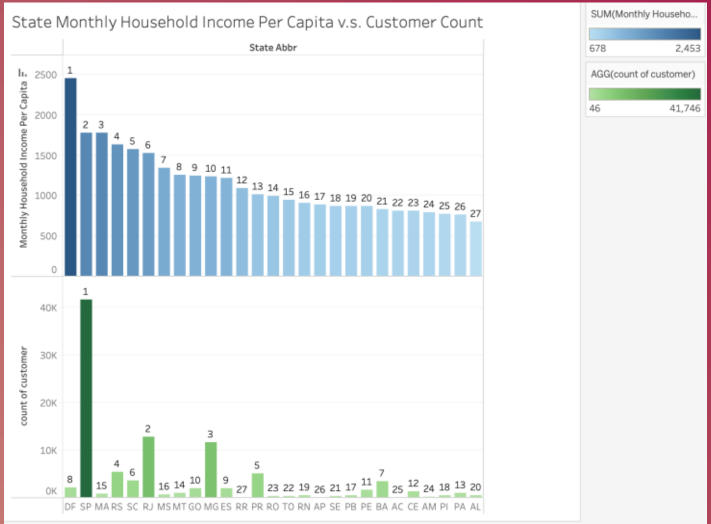
Freight Value mean from State to Regions

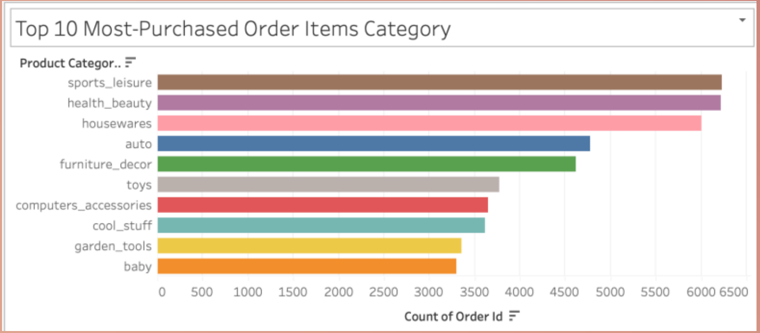
cust_Region seller_state	Midwest	North	Northeast	South	Southeast
AC	nan	nan	nan	nan	32.840000
AM	nan	nan	28.255000	nan	25.290000
BA	29.977381	46.290476	22.736355	39.643671	33.618110
CE	65.456667	85.614000	29.744054	51.445556	47.044130
DF	13.070164	58.371935	27.607364	27.096786	19.636728
ES	36.895833	68.896250	36.664464	39.894510	28.461445
GO	14.686705	26.595909	33.369841	30.535870	24.201088
MA	27.150192	28.859000	19.453333	42.660652	31.880362
MG	26.861624	41.899053	33.968349	28.904398	21.153626
MS	26.228571	21.410000	29.412308	25.600000	21.608750
MT	23.865806	27.533333	38.166667	34.652000	34.021389
PA	nan	nan	nan	35.750000	17.051429
PB	38.796667	20.930000	19.178333	nan	50.781818
PE	30.903750	41.911538	18.845508	49.931765	28.030597
PI	nan	nan	26.516000	46.855000	43.406000
PR	33.309601	47.200082	44.183779	18.748641	20.893805
RJ	20.737075	39.056667	33.177818	21.068325	16.838489
RN	22.850000	26.290000	14.346944	44.189000	37.073750
RO	nan	nan	64.570000	43.285000	49.707000
RS	33.340569	42.597143	49.271161	18.284614	25.703550
SC	35.855525	51.120175	51.117871	20.687027	24.475630
SE	36.330000	26.000000	39.500000	21.680000	33.450000
SP	21.371431	34.319450	30.620874	20.690144	15.807044





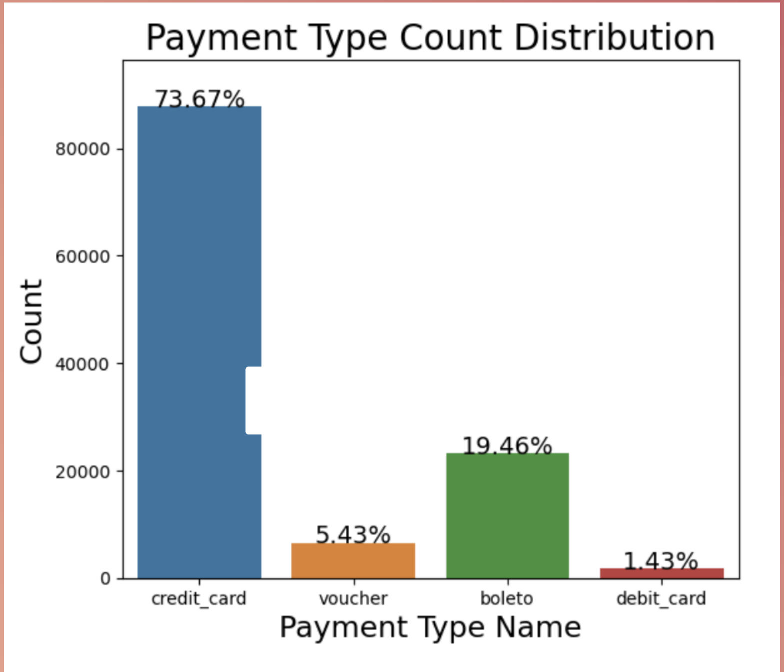

Review Score Distribution
Review Score
Count of Customer Id
26K
24K
22K
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
65.14%
38.58%
54.62%
72.02%
91.03%
1
2
3
4
5


State Monthly Household Income Per Capita v.s. Customer Count
State Abbr
Monthly Household Income Per Capita
count of customer
SUM(Monthly Househo...
678
2,453
AGG(count of customer)
46
41,746
DF SP MA RS SC RJ MS MT GO MG ES RR PR RO TO RN AP SE PB PE BA AC CE AM PI PA AL

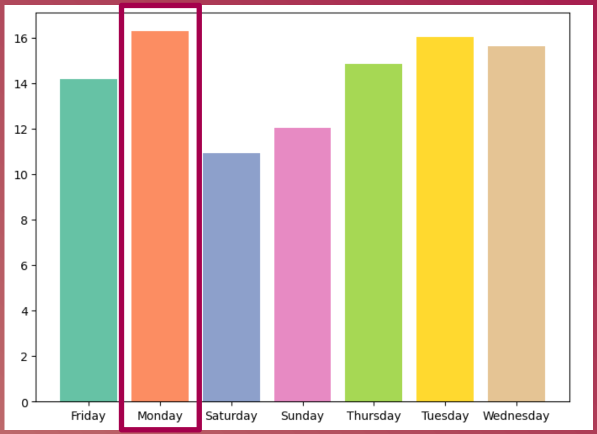

Top 10 Most-Purchased Order Items Category
Product Categor..
sports_leisure
health_beauty
housewares
auto
furniture_decor
toys
computers_accessories
cool_stuff
garden_tools
baby
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
Count of Order Id

We can see that the most common state of CUSTOMERS is SP(42.19%), followed by RJ(13.02%) and MG(11.6%).
All this states is from the southeast region of Brazil. Also, we have many sales to RS, PR,
SC (states from south region)

Payment Type Distribution with Value Labeled



Orders by Day of Week



Orders by Time of Day