E-commerce

Group members:

1. G2302794C ANG SHU WEI
2. G2406170D HTEIK TIN MIN PAING
3. G2303232C JI HONGXIAO
4. G2302707A NONG MINH HIEU
5. G2302855K ZHANG DANRUI
6. G2406019J YE SHENGHUA
7. G2404083G DING JUNCHENG

Background of the Brazilian E-Commerce Public Dataset by Olist

Context

- Olist is the largest department store in Brazilian marketplaces.
  - Connects small businesses from all over Brazil to channels without hassle and with a single contract.
  - Merchants on their platform can sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners.
- How it works?
  - After a customer purchases the product from Olist Store -> Seller gets notified to fulfill that order.
  - Once the customer receives the product or the estimated delivery date is due, the customer gets a satisfaction survey by email where he/she can give ratings and comments on his/her purchase experience.

Dataset

- ~100k+ of orders made at Olist from 2016 to 2018 made at multiple marketplaces in Brazil.
- Information of customers, order items, order payments, order reviews, orders, products, sellers, product category name translation and geolocation datasets can be found.

Note:

- An order might have multiple items.
- Each item might be fulfilled by a distinct seller.
- Real commercial data where names & references to the companies and partners in the review text are anonymized.

Before creating the ERD diagram

1

Results for file: product_category_name_translation.csv

column_name	null_values	unique
product_category_name	0	True
product_category_name_english	0	True

2

Results for file: olist_customers_dataset.csv

column_name	null_values	unique
customer_id	0	True
customer_unique_id	0	False
customer_zip_code_prefix	0	False
customer_city	0	False
customer_state	0	False
customer_id + customer_unique_id	0	True

3

Results for file: olist_geolocation_dataset.csv

column_name	null_values	unique
geolocation_zip_code_prefix	0	False
geolocation_lat	0	False
geolocation_lng	0	False
geolocation_city	0	False
geolocation_state	0	False

4

Results for file: olist_order_items_dataset.csv

column_name	null_values	unique
order_id	0	False
order_item_id	0	False
product_id	0	False
seller_id	0	False
shipping_limit_date	0	False
price	0	False
freight_value	0	False
order_id + order_item_id	0	True

5

Results for file: olist_order_payments_dataset.csv

column_name	null_values	unique
order_id	0	False
payment_sequential	0	False
payment_type	0	False
payment_installments	0	False
payment_value	0	False
order_id + payment_sequential	0	True

6

Results for file: olist_order_reviews_dataset.csv

column_name	null_values	unique
review_id	0	False
order_id	0	False
review_score	0	False
review_comment_title	87656	False
review_comment_message	58247	False
review_creation_date	0	False
review_answer_timestamp	0	False
review_id + order_id	0	True

7

Results for file: olist_orders_dataset.csv

column_name	null_values	unique
order_id	0	True
customer_id	0	True
order_status	0	False
order_purchase_timestamp	0	False
order_approved_at	160	False
order_delivered_carrier_date	1783	False
order_delivered_customer_date	2965	False
order_estimated_delivery_date	0	False

8

Results for file: olist_products_dataset.csv

column_name	null_values	unique
product_id	0	True
product_category_name	610	False
product_name_lenght	610	False
product_description_lenght	610	False
product_photos_qty	610	False
product_weight_g	2	False
product_length_cm	2	False
product_height_cm	2	False
product_width_cm	2	False

9

Results for file: olist_sellers_dataset.csv

column_name	null_values	unique
seller_id	0	True
seller_zip_code_prefix	0	False
seller_city	0	False
seller_state	0	False

- Have run checks on datasets:
  - How many null values (if any)? and
  - Whether the columns contain unique values or not
- For the geolocation dataset -> Since there are no unique values, and this table does not add much value as we are not using the geolocation latitude and geolocation longitude => Thus we will be **dropping this table**.

ER diagram

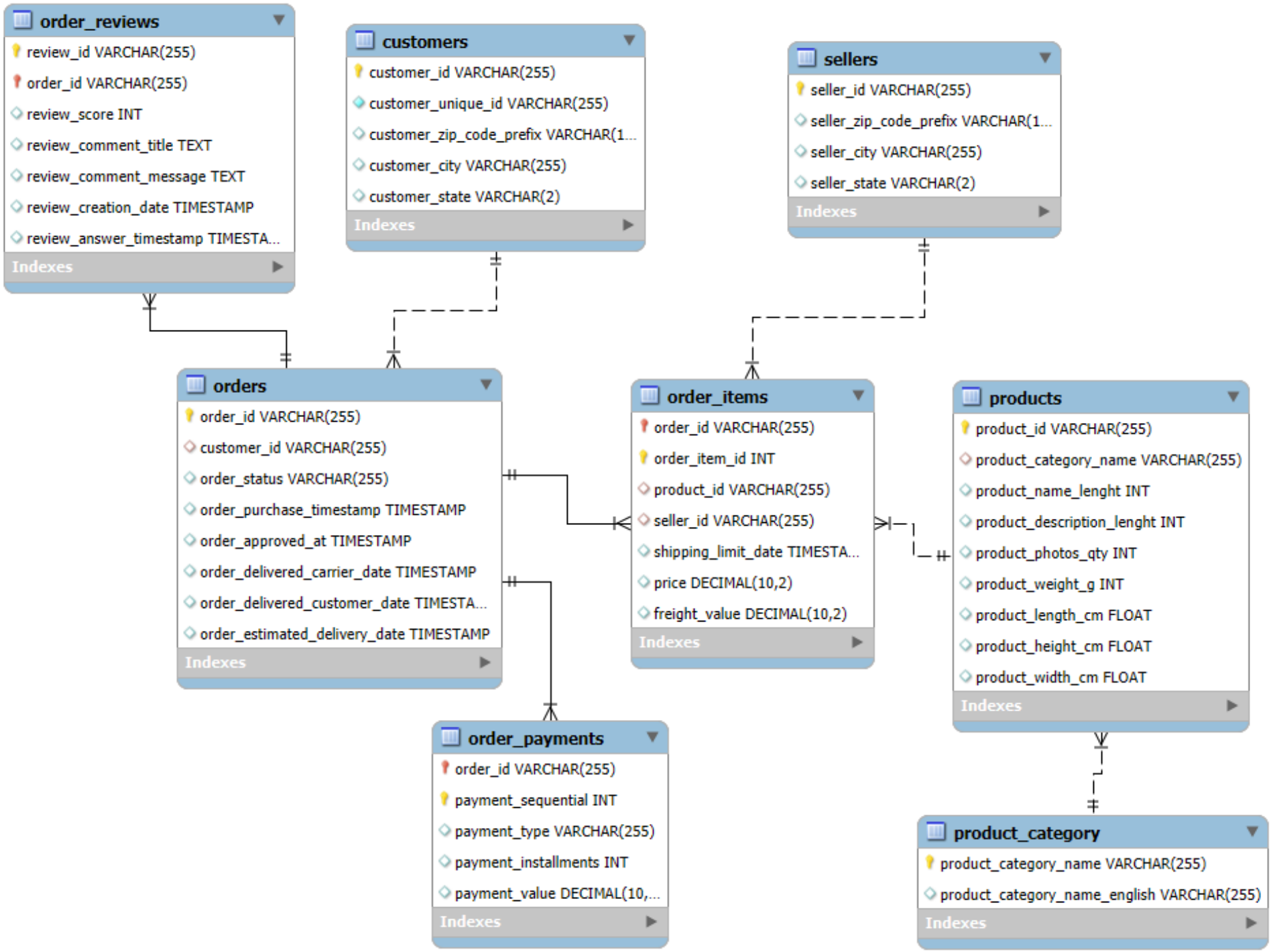

order_reviews
review_id VARCHAR(255)
order_id VARCHAR(255)
review_score INT
review_comment_title TEXT
review_comment_message TEXT
review_creation_date TIMESTAMP
review_answer_timestamp TIMESTA...
Indexes
customers
customer_id VARCHAR(255)
customer_unique_id VARCHAR(255)
customer_zip_code_prefix VARCHAR(1...
customer_city VARCHAR(255)
customer_state VARCHAR(2)
Indexes
sellers
seller_id VARCHAR(255)
seller_zip_code_prefix VARCHAR(1...
seller_city VARCHAR(255)
seller_state VARCHAR(2)
Indexes
orders
order_id VARCHAR(255)
customer_id VARCHAR(255)
order_status VARCHAR(255)
order_purchase_timestamp TIMESTAMP
order_approved_at TIMESTAMP
order_delivered_carrier_date TIMESTAMP
order_delivered_customer_date TIMESTA...
order_estimated_delivery_date TIMESTAMP
Indexes
order_items
order_id VARCHAR(255)
order_item_id INT
product_id VARCHAR(255)
seller_id VARCHAR(255)
shipping_limit_date TIMESTA...
price DECIMAL(10,2)
freight_value DECIMAL(10,2)
Indexes
products
product_id VARCHAR(255)
product_category_name VARCHAR(255)
product_name_lenght INT
product_description_lenght INT
product_photos_qty INT
product_weight_g INT
product_length_cm FLOAT
product_height_cm FLOAT
product_width_cm FLOAT
Indexes
order_payments
order_id VARCHAR(255)
payment_sequential INT
payment_type VARCHAR(255)
payment_installments INT
payment_value DECIMAL(10,...
Indexes
product_category
product_category_name VARCHAR(255)
product_category_name_english VARCHAR(255)
Indexes

Table Description

1. customers

COLUMN_NAME	COLUMN_TYPE	IS_NULLABLE	COLUMN_KEY	COLUMN_COMMENT
customer_city	varchar(255)	YES		City of the customer
customer_id	varchar(255)	NO	PRIMARY	Unique identifier per customer of each order
customer_state	varchar(2)	YES		State of the customer
customer_unique_id	varchar(255)	NO		Unique identifier for each customer
customer_zip_code_prefix	varchar(10)	YES		Prefix of zip code for customer

2. sellers

COLUMN_NAME	COLUMN_TYPE	IS_NULLABLE	COLUMN_KEY	COLUMN_COMMENT
seller_city	varchar(255)	YES		City of the seller
seller_id	varchar(255)	NO	PRIMARY	Unique identifier for each seller
seller_state	varchar(2)	YES		State of the seller
seller_zip_code_prefix	varchar(10)	YES		Prefix of zip code for seller

3. orders

COLUMN_NAME	COLUMN_TYPE	IS_NULLABLE	COLUMN_KEY	COLUMN_COMMENT
customer_id	varchar(255)	YES	FOREIGN	Foreign key to customers table
order_approved_at	timestamp	YES		Timestamp of when the order was approved
order_delivered_carrier_date	timestamp	YES		Timestamp of when the order was delivered to the carrier
order_delivered_customer_date	timestamp	YES		Timestamp of when the order was delivered to the customer
order_estimated_delivery_date	timestamp	YES		Estimated delivery date for the order
order_id	varchar(255)	NO	PRIMARY	Unique identifier for each order
order_purchase_timestamp	timestamp	YES		Timestamp of when the order was purchased
order_status	varchar(255)	YES		Current status of the order

Table Description

4. order_items

COLUMN_NAME	COLUMN_TYPE	IS_NULLABLE	COLUMN_KEY	COLUMN_COMMENT
freight_value	decimal(10,2)	YES		Freight value for the item
order_id	varchar(255)	NO	PRIMARY	Foreign key to orders table
order_item_id	int	NO	PRIMARY	Unique identifier for each item in an order
price	decimal(10,2)	YES		Price of the item
product_id	varchar(255)	YES	FOREIGN	Foreign key to products table
seller_id	varchar(255)	YES	FOREIGN	Foreign key to sellers table
shipping_limit_date	timestamp	YES		Deadline for shipping the item

5. products

COLUMN_NAME	COLUMN_TYPE	IS_NULLABLE	COLUMN_KEY	COLUMN_COMMENT
product_category_name	varchar(255)	YES	FOREIGN	Foreign key to product_category table
product_description_lenght	int	YES		Length of product description
product_height_cm	float	YES		Height of the product in centimeters
product_id	varchar(255)	NO	PRIMARY	Unique identifier for each product
product_length_cm	float	YES		Length of the product in centimeters
product_name_lenght	int	YES		Length of product name
product_photos_qty	int	YES		Quantity of product photos
product_weight_g	int	YES		Weight of the product in grams
product_width_cm	float	YES		Width of the product in centimeters

6.product_category

COLUMN_NAME	COLUMN_TYPE	IS_NULLABLE	COLUMN_KEY	COLUMN_COMMENT
product_category_name	varchar(255)	NO	PRIMARY	Category name in Portuguese
product_category_name_english	varchar(255)	YES		Category name in English

Table Description

7. order_payments

COLUMN_NAME	COLUMN_TYPE	IS_NULLABLE	COLUMN_KEY	COLUMN_COMMENT
order_id	varchar(255)	NO	PRIMARY	Foreign key to orders table
payment_installments	int	YES		Number of payment installments
payment_sequential	int	NO	PRIMARY	Sequential identifier for each payment in an order
payment_type	varchar(255)	YES		Type of payment (e.g., credit card, voucher)
payment_value	decimal(10,2)	YES		Total value of the payment

8. order_reviews

COLUMN_NAME	COLUMN_TYPE	IS_NULLABLE	COLUMN_KEY	COLUMN_COMMENT
order_id	varchar(255)	NO	PRIMARY	Foreign key to orders table
review_answer_timestamp	timestamp	YES		Timestamp when the review was answered
review_comment_message	text	YES		Full review comment
review_comment_title	text	YES		Title of the review comment
review_creation_date	timestamp	YES		Timestamp when the review was created
review_id	varchar(255)	NO	PRIMARY	Unique identifier for each review
review_score	int	YES		Rating score of the review

Why we use Database Set-up (SQLite3)?

Pros	Cons
• Easy to set up. • Light weight and fast query speed. • Portable. Which means there is no need to set up DBMS server. • Easily accessible via third party tools (DB Browser, sqlite3 library in Python, etc). • Mostly suitable for use cases where data volume is small.	• Challenging for multiple users to execute queries at the same time -> Difficult to run data operations in parallel. • Not suitable for use cases where data volume is huge. • Not so feature-rich like other DBMS like MySQL or PostgreSQL.

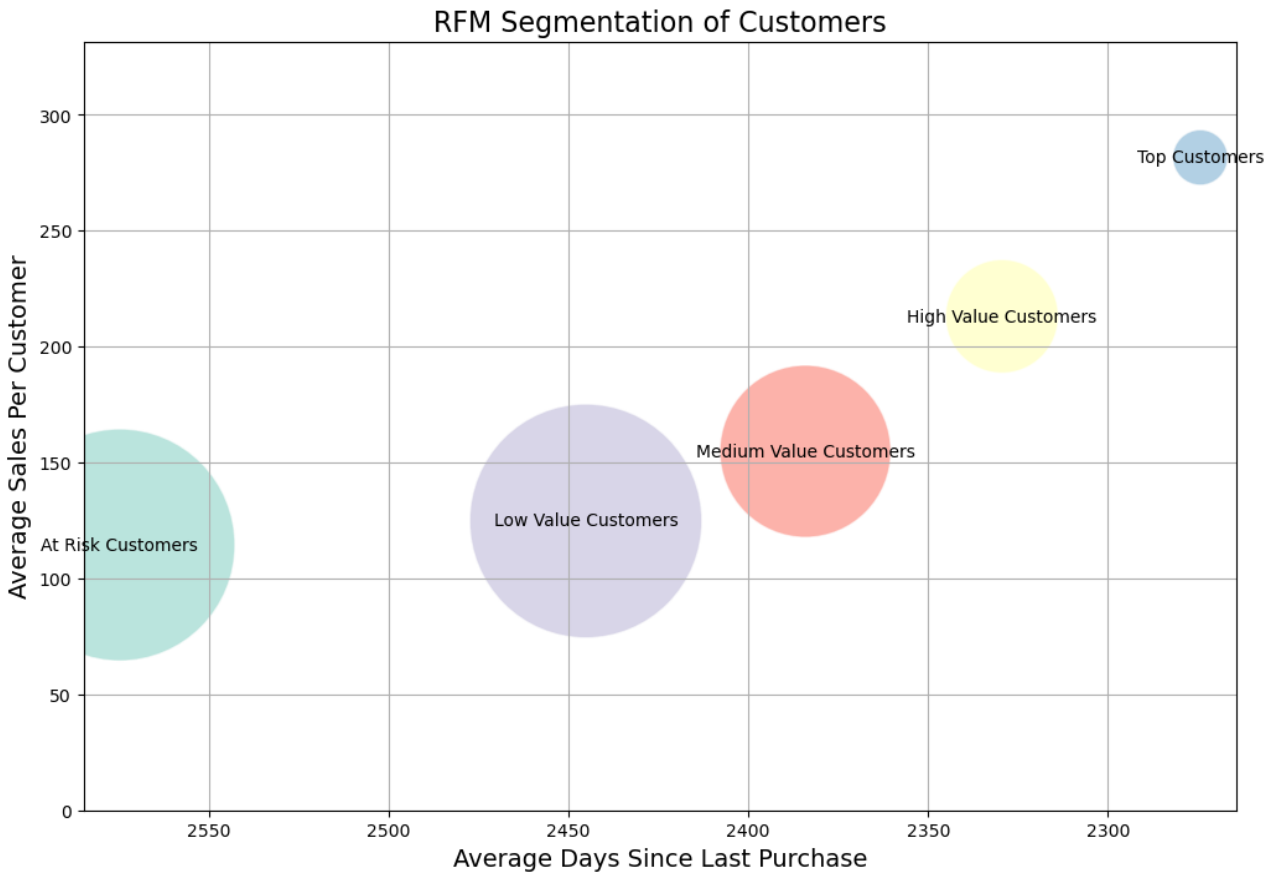
**Details for DBMS setup with SQLite3 is included in the Appendix*

Data Analysis and Visualization

- Using Python and Tableau for analysis and visualization.

- To look at the profit, sales, and quantity we deep dive into:
  - Customer segmentation
  - Product categories
  - Cities
    - Customer reviews based on delivery
    - Delivery time
    - Methods of payments

Customer Segmentation

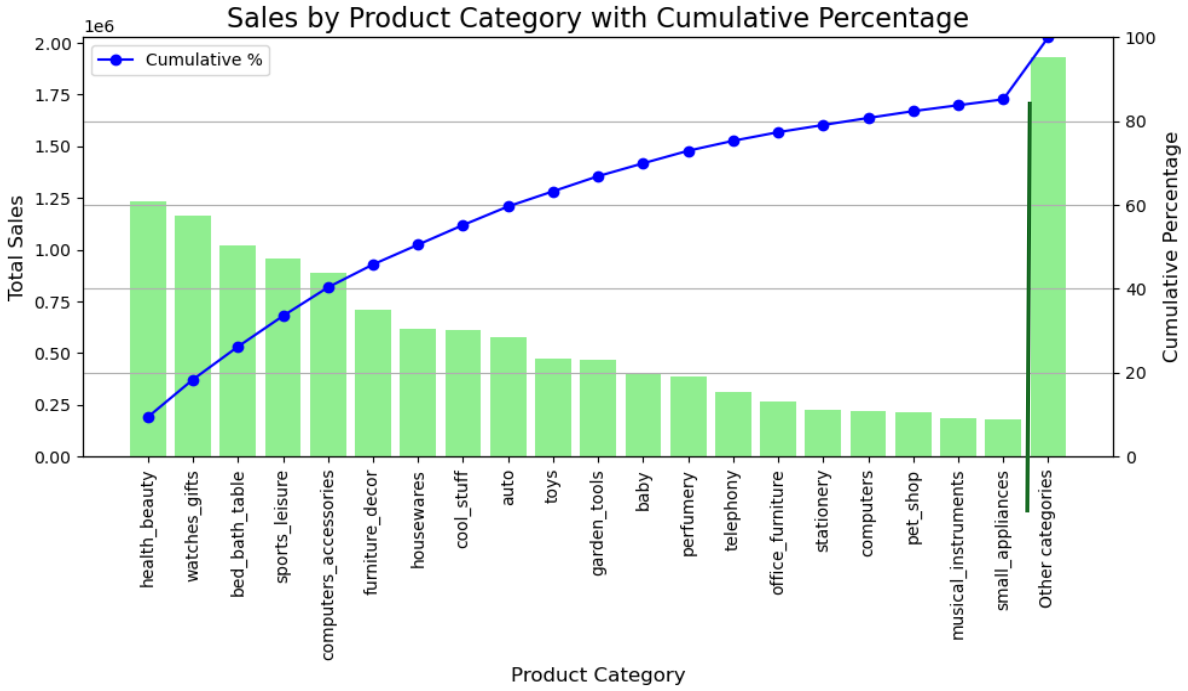


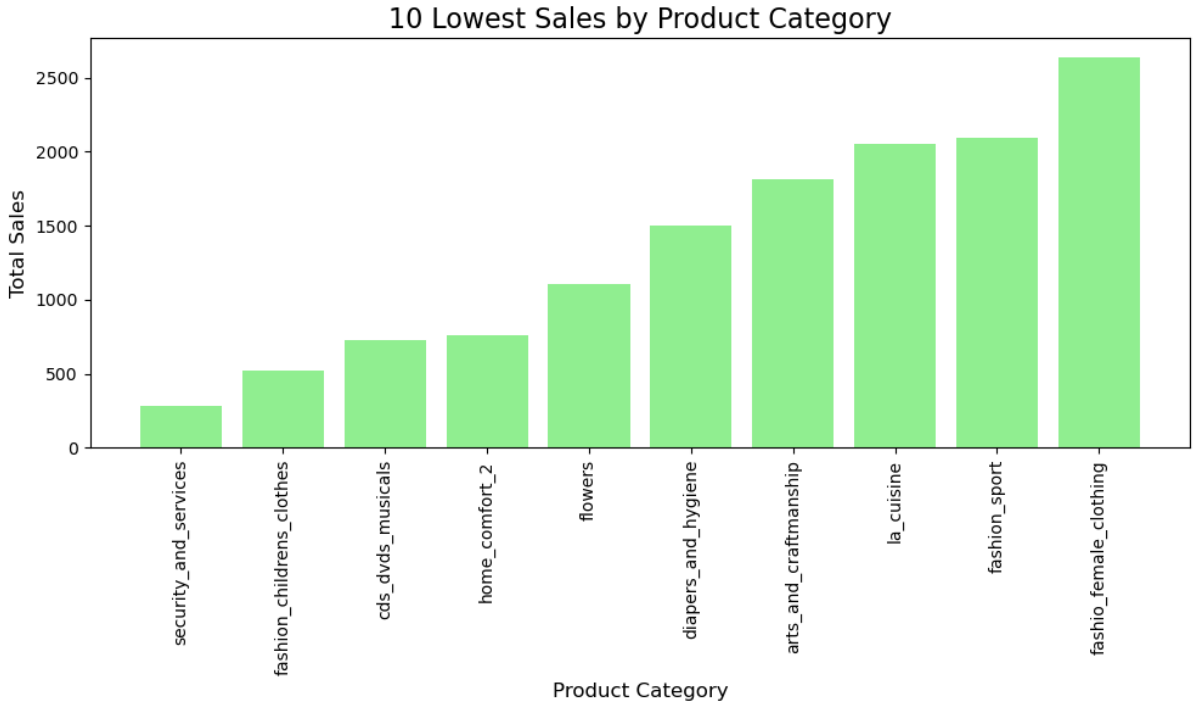
- RFM scoring was used to calculate
  - Recency: the last time a customer makes a purchase
  - Frequency: the number of times a customer has purchased from Olist
  - Monetary: the total amount of money a customer spends
- Customers were then given scoring (1 to 5) for each of the RFM metric above and grouped into 5 different segments
  - Top Customers (High R/ High M/ High F)
  - High Value (Medium to High R/ Medium to High F/ Medium to High M)
  - Medium Value (Medium R/ Medium F/ Medium M)
  - Low Value (Low to Medium R/ Low to Medium F/ Low to Medium M)
  - At Risk (Low R/ Low F/ Low M)

	RFM_Bucket	avg_days_since_purchase	avg_sales_per_customer	customer_count
0	At Risk Customers	2574.878882	114.479515	32588
1	High Value Customers	2329.475420	213.094314	7808
2	Low Value Customers	2445.241702	124.817041	33129
3	Medium Value Customers	2384.150714	154.880671	17970
4	Top Customers	2274.297894	281.629049	1863

- Understanding customer segmentation can help Olist **strategize different marketing strategies** to target each **customer segment** to maximize customer loyalty and sales.

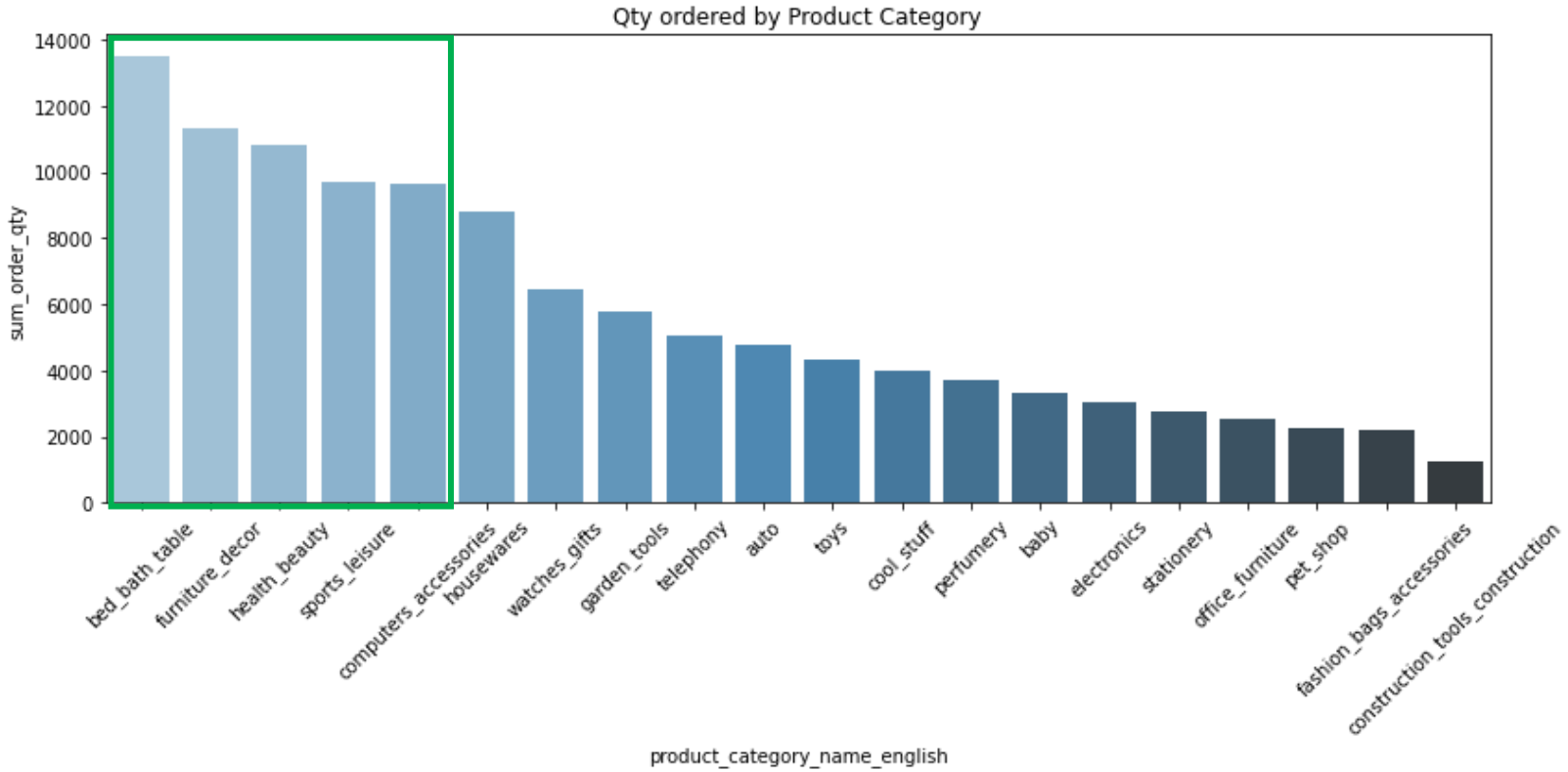
Highest and Lowest Sales by Product Category





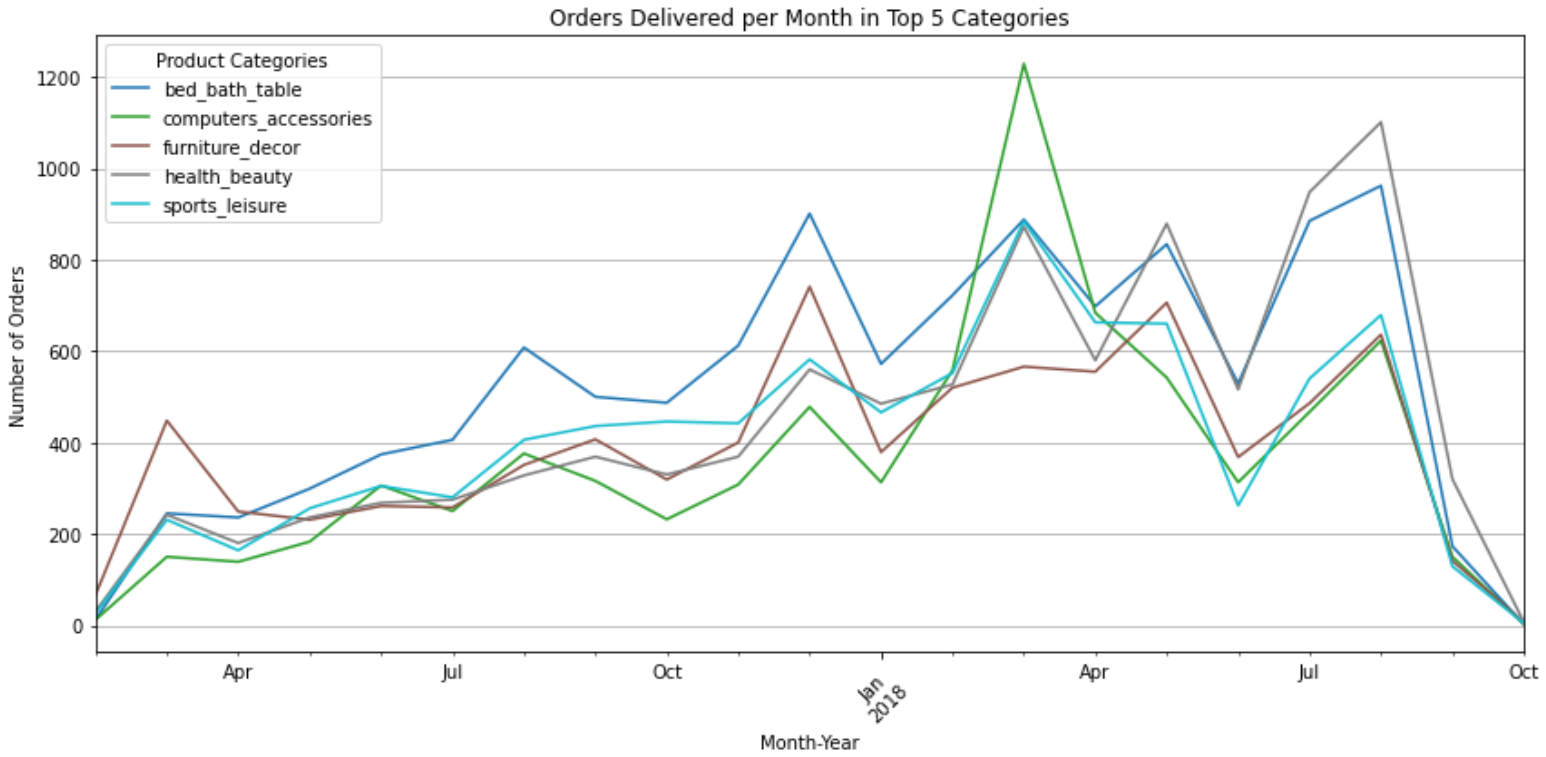
- **Top 20 product categories** contribute **85%** of the total revenue.
- Having insights of the product categories that generate the **highest** and **lowest sales** could also help us make better business decisions to **optimize the inventory** to ensure there is optimal supply of products to match the demand.
- Understand which product categories are crucial and need more attention for Olist business and channel its resources more towards marketing, promoting those products in particular
  - Such as **promoting popular products** in categories like health beauty / watches to get more traffic to the website

Top 5 quantity sales in the product category



- Top 5 quantity of orders are for products in these categories:
  - Bed bath table / furniture decor / health beauty / sports leisure / computers accessories
- Focus on these categories and deep dive into the quantity sales across the months and years to see if there is any trend.
- To see if there are any **complementary categories**
  - For example: To bring up the sales of electronics, perhaps promotion should tie with computer accessories which can lead to a win-win in terms of boasting the sales for both categories.

Trends for the Top 5 quantity sales



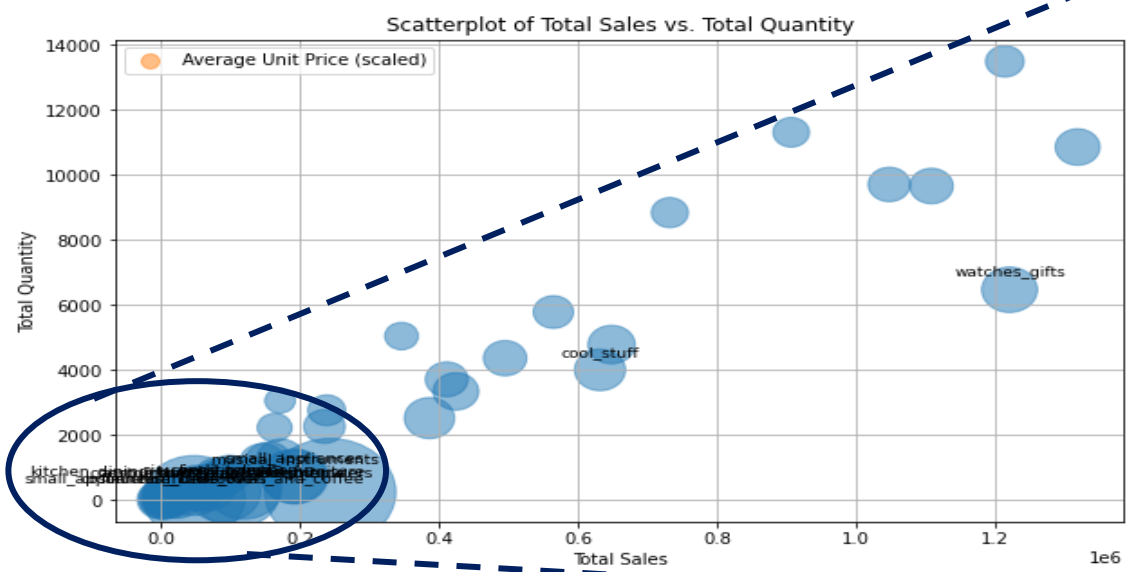
Note:

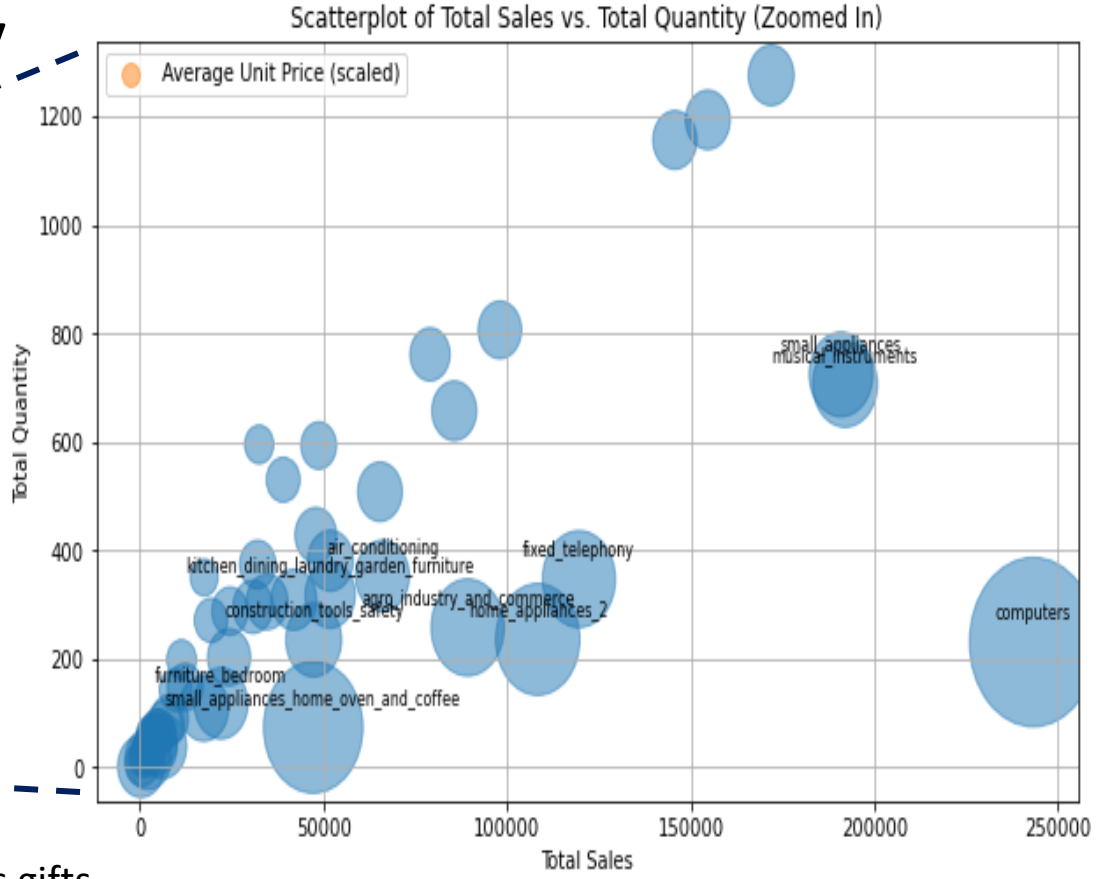
- Plotted this trend from Jan 2017 onwards due to noisy data before Jan 2017.

- From the top 5 categories, there is a greater spike in computer accessories in the period between Jan to Mar 2018 as the number of orders reached a peak of 1200+ in Mar 2018.
- An upward trend for these categories from Jan 2017 to Mar 2018.
  - Comparing Q3 in 2017 to 2018, there seems to be a sharper drop of quantities ordered in 2018 as compared to 2017.
- Important to **relook at the marketing, pricing strategies as well as customer sentiments** to see what have changed to prevent sales decline.

Relationship between total sales, total qty and average unit price of product in each category

NOTE:
- *Only Top 20% of the highest average unit price of the product in each category is annotated.*
- *The bigger the circle, the higher the average unit price.*





Segment the product categories:
- High sales, high quantity with higher-than-average unit price: watches gifts
- Mid sales, mid quantity with higher-than-average unit price: cool stuff
- Low sales, low quantity with high average unit price: computer, small appliances home oven and coffee

Recommendations:
- **Focus our marketing campaign efforts** to **promote those products** with low sales, low quantity sales but with **high average unit price to bump up sales fast**.
- A **marketing push for high-priced items can elevate brand perception**, attracting customers who associate higher prices with better quality.

Correlation between price and sales quantity

- -0.34 correlation: There is a weak negative relationship between price and total quantity ordered.
- If the prices of the products increase, there is a tendency for the number of quantity ordered to decrease.
- Higher prices may deter consumers from buying more of the products, thus **sellers are encouraged to reduce prices to clear their stocks**.

Which products are the most profitable in terms of total profit and average profit?

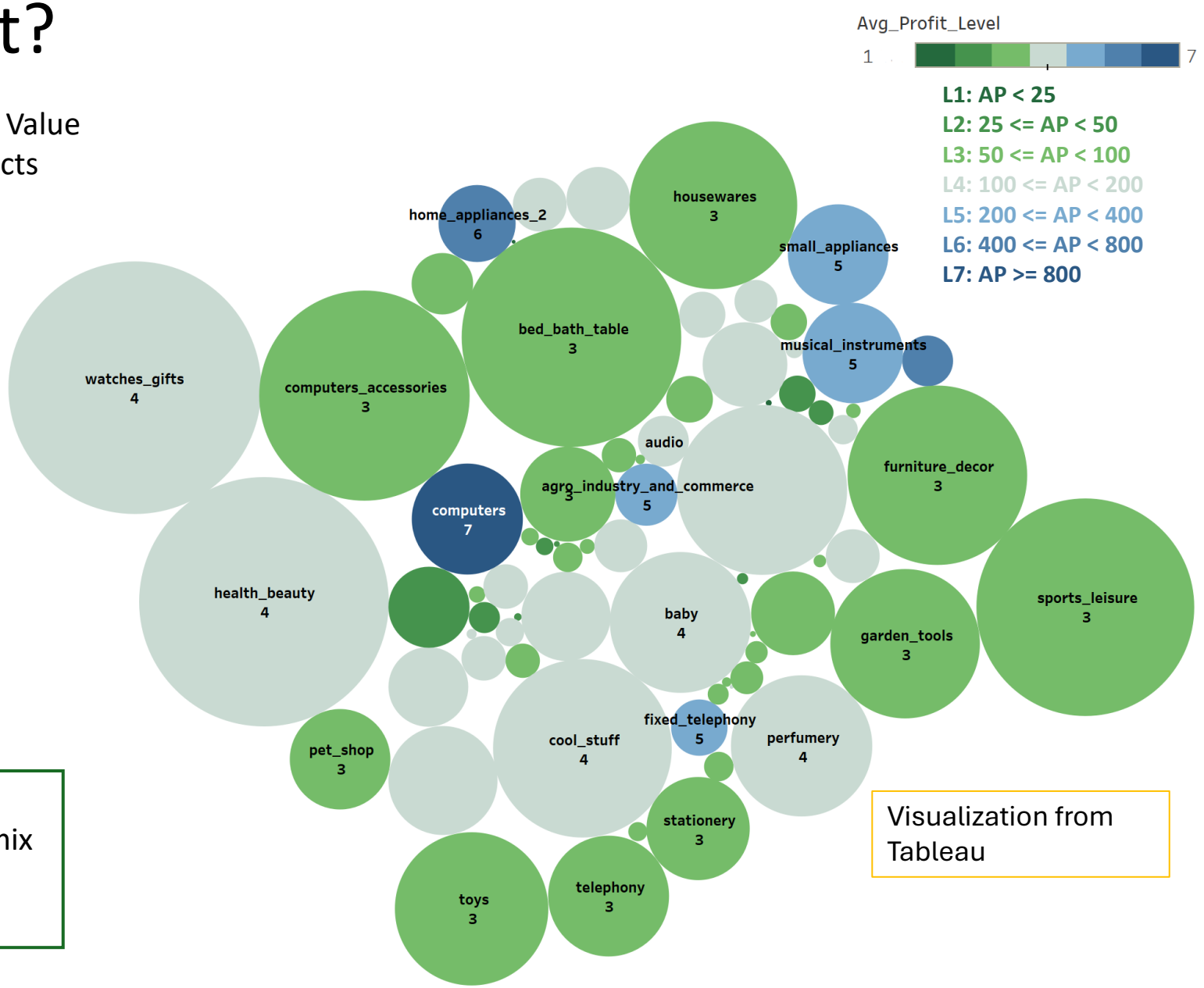
Total Profit (Bubble Size) = Total Price – Total Freight Value
Average Profit (AP)= Total Profit / Numbers of Products

- Computers have a lower volume compared to high-total-profit products, despite their highest unit profit.
  - This could be due to factors like higher pricing, niche markets, or less frequent purchases.

- Most **high-total-profit products** are in average profit **levels 3 and 4**.
  - This implies that products with a moderate unit profit and higher sales volume can often achieve higher overall profitability.

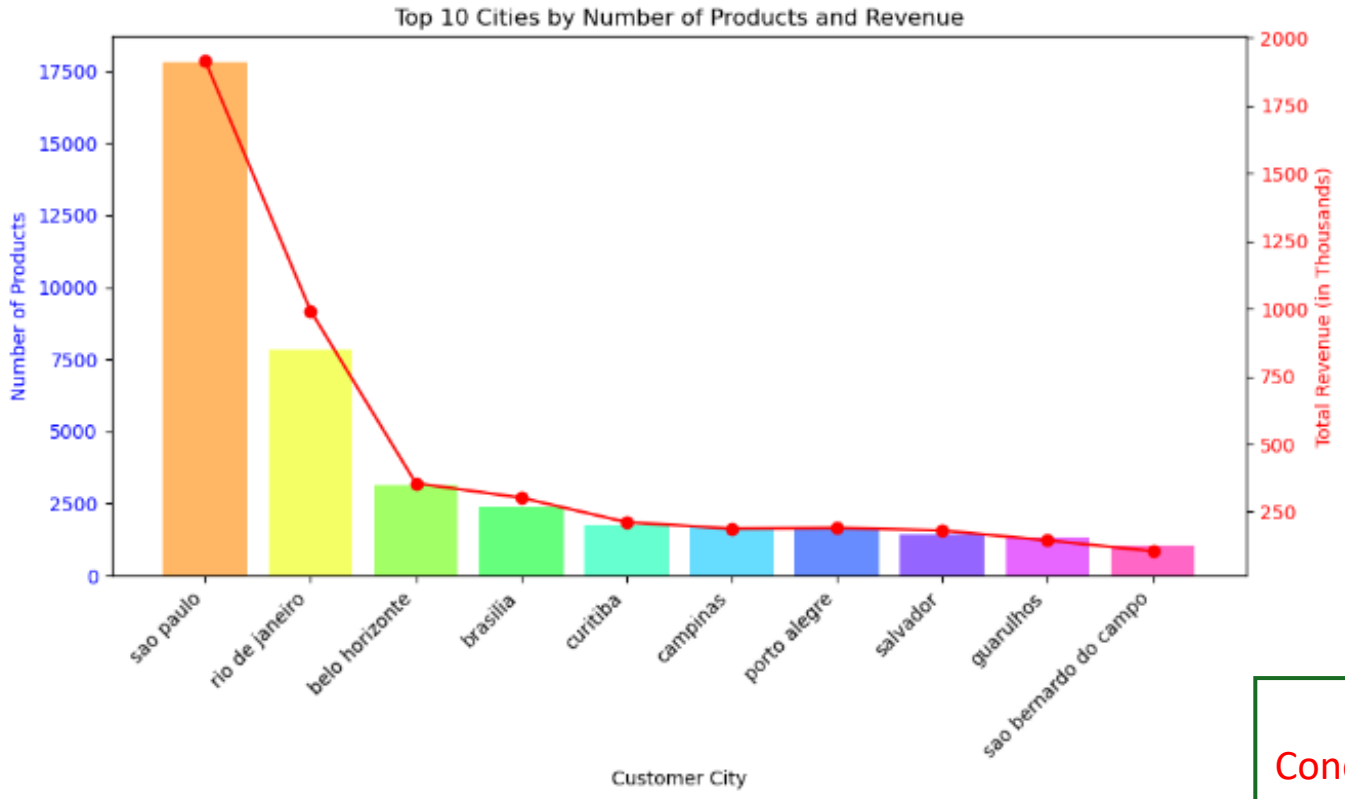
Recommendations

Product Mix Analysis: Evaluate the overall product mix and consider **adjusting the portfolio** to **focus on products with higher profit margins**.



Visualization from Tableau

Top 10 cities with the highest number of products and highest revenue

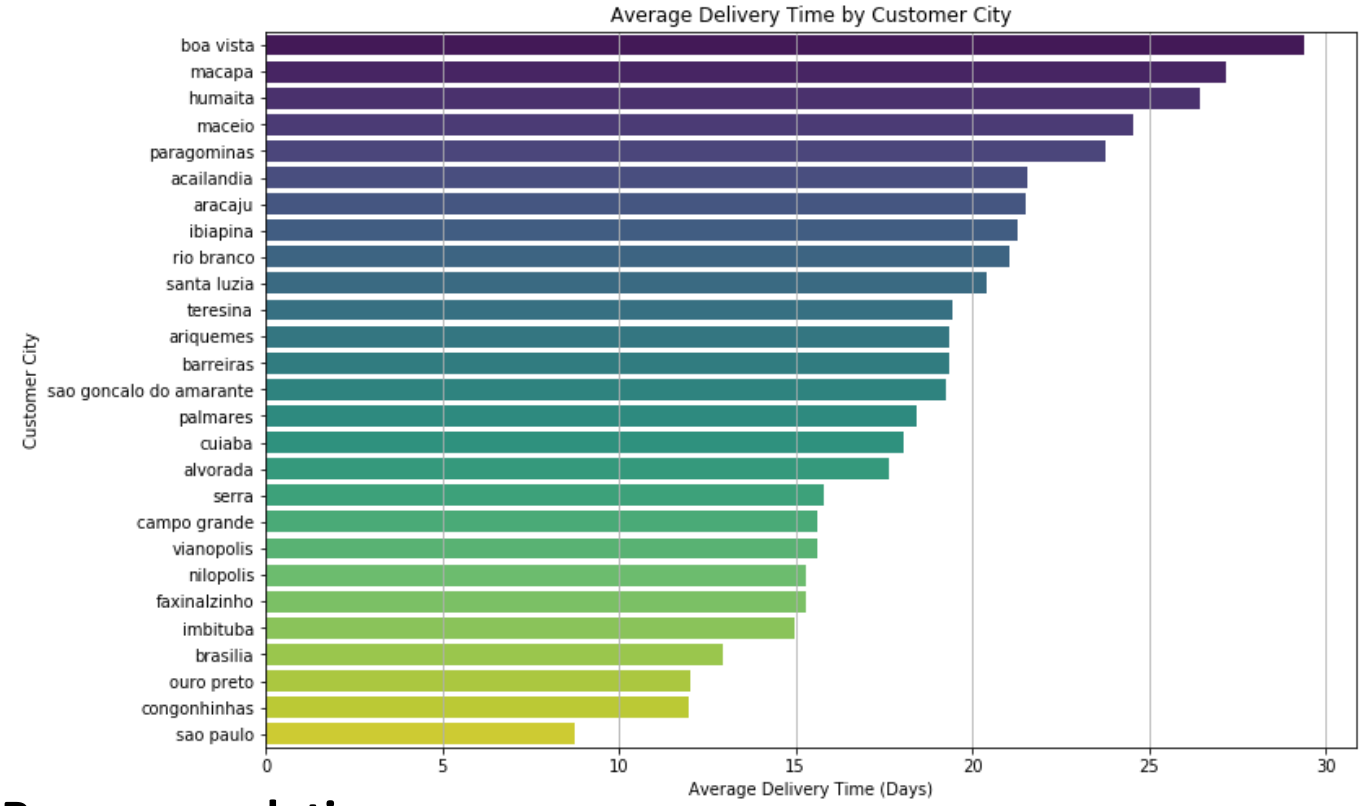


- The ranking of the Top 10 cities with the highest product quantity and highest revenue is the same.
- **Sao Paulo** leads by a wide margin in both revenue and product volume, which is almost twice as many as **Rio De Janeiro.**
- But there is a steep decline in both number of products and revenue after the top two cities, which seems to be a problem with sales outside the metropolitan areas.

Recommendations

Conduct local market research: To understand the specific needs and preferences of consumers in cities outside Sao Paulo and Rio De Janeiro. This can help tailor marketing strategies and product offerings to better align with local tastes.

How does delivery time vary based on customer location?

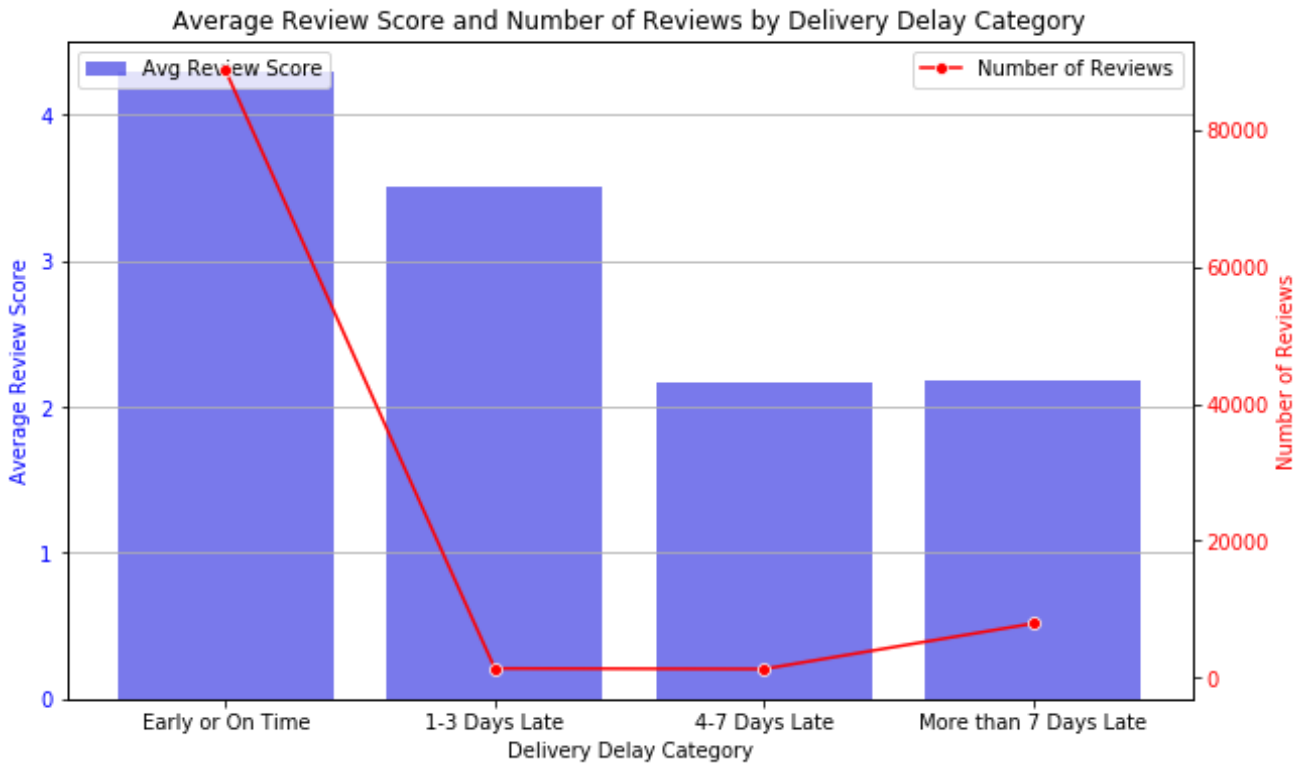


- This chart shows the average time it takes for deliveries to reach customers in various cities, which reflects logistics challenges in remote cities compared to major urban centers.
- Cities like **Boa Vista** and **Macapá** have much **longer delivery times** (close to 30 days), while cities like **São Paulo** and **Congonhinhas** experience **faster deliveries** (around 10 days or less).
- Cities like **São Paulo** and **Rio de Janeiro** perform exceptionally well in both the number of orders and revenue generation, aligning with their relatively shorter delivery times. However, cities with higher average delivery times (e.g., **Boa Vista**, **Macapá**) are not appearing in the top revenue or order counts, possibly due to the poor delivery experience.

Recommendations:

- **Improve Delivery Speed in Low-Performance Cities**:
  - Cities like Boa Vista and Macapá experience very long delivery times. **Reducing delivery delays** in these cities through better logistics, local distribution centers, or more reliable courier services could improve customer satisfaction and potentially increase repeat orders and revenue.
- **Enhance Logistics in Growing Markets**:
  - Some cities with significant potential (mid-level performers) like **Brasilia, Salvador, and Curitiba** show promising results in the delivery of order volume and revenue (i.e. able to cope with the scaling of orders thus increasing revenue). Invest in **regional warehouses or partnerships** with faster local delivery services in these cities to further boost performance.

How does the delivery delay affect customer review?



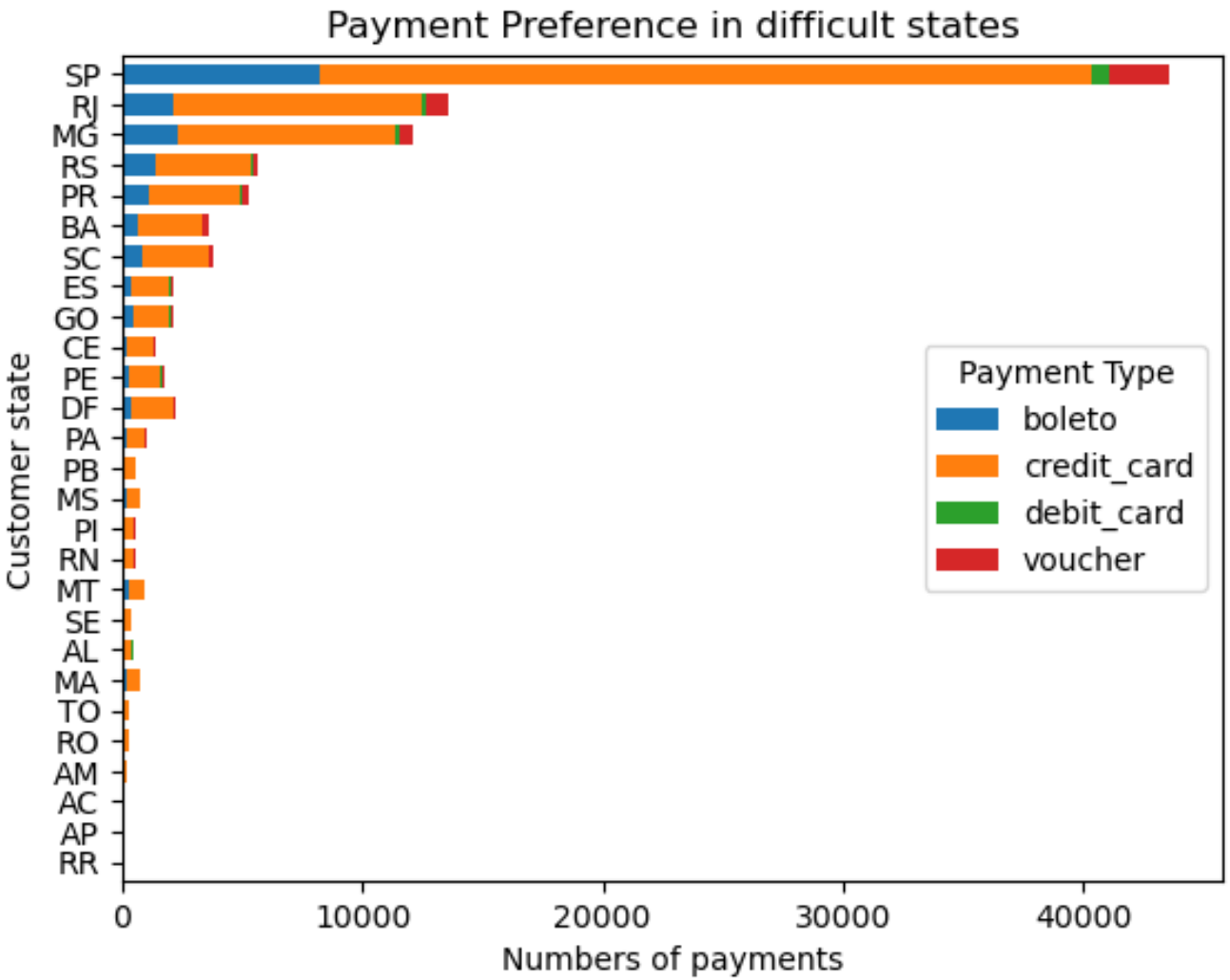
	delay_category	avg_review_score	num_reviews
0	1-3 Days Late	3.511029	1360
1	4-7 Days Late	2.174083	1281
2	Early or On Time	4.293578	88658
3	More than 7 Days Late	2.176782	7925

- This chart shows the relationship between **delivery delay categories** and their impact on **customer reviews**.
- Customers are **most satisfied** when **deliveries** are **early or on time**, giving an average review score of **4.3**, the highest rating and most willingness to give reviews.
- For deliveries that are delayed **more than 7 days**, review scores are very low **(around 2.7)** and the number of reviews increases to **7925**.
  - This suggests that long delays not only frustrate customers and are more likely to voice their opinions and express their dissatisfaction.

Recommendations

- **Improving Delivery Processes**:
  - Assess the supply chain and logistics to identify bottlenecks that contribute to delays.
  - Implementing more efficient processes could help reduce delivery times and improve customer satisfaction.
- **Transparent Communication**:
  - Keep customers informed about their order status, especially if delays are anticipated.
  - Proactive communication can mitigate frustration and enhance customer trust.

What are the preferred payment methods in different states and how do they vary in usage?



- Credit card: the most popular across all states
- Boleto: a popular Brazilian cash-based payment method, particularly welcomed in SP
- Vouncher: more concentrated presence in states like SP & RJ.
- Debit card: seems to be minimal usage

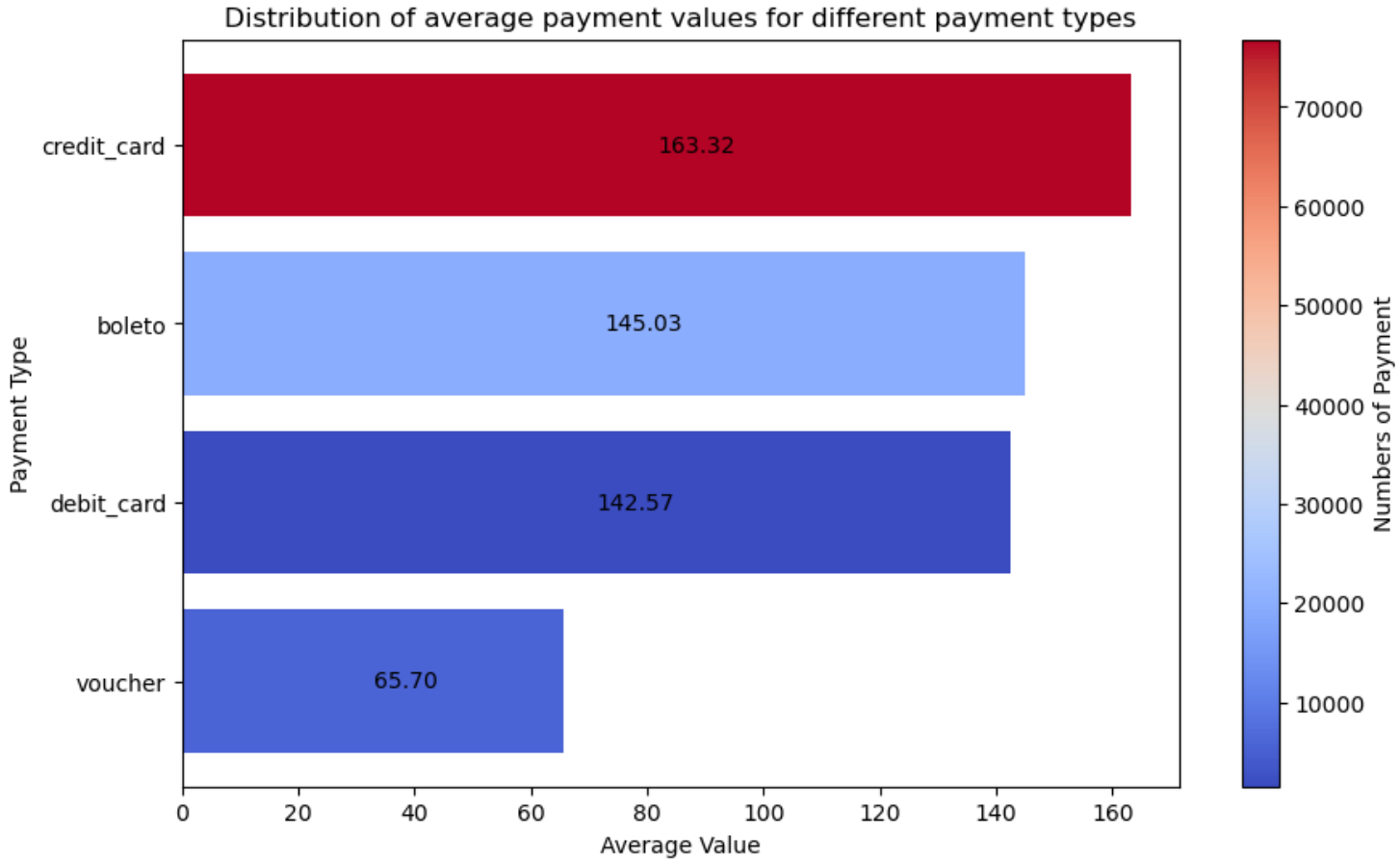
Recommendations

Optimize Payment Options: Credit card is the primary payment method. However, consider offering Boleto as an alternative option, especially in states where they have significant usage.

Promote Voucher: If there's a strategic reason to promote voucher usage, explore targeted campaigns or incentives in states where it's currently underutilized.

What is the average payment value for each payment type and how do they compare to each other?

- Number of payments is mostly positively-correlated with the average payment value.
- The higher average value for credit card payments might indicate that customers are more likely to make larger purchases or have impulse buys when using this method.
- The lower average value for voucher payments could be due to restrictions on the types of products or services purchased with vouchers.



Recommendations

Pricing Strategies: Consider adjusting pricing strategies based on the payment method used. For example, you might offer discounts or promotions for customers using Boleto or debit card.

Customer Segmentation: Segment customers based on their preferred payment methods and tailor marketing and product offerings accordingly.

Recommendations

- Based on **customer segmentation** insights,
  - Olist can implement **exclusive loyalty programs for top customers** and high value customers by offering store points, early access to new products, free delivery or exclusive discounts to retain them and encourage them to make more purchases.
  - As for medium, low value and those customers who the platform is at risk of losing them, Olist **can increase customer engagement** through personalized emails or SMS, conduct survey for feedback or provide bundle deals based on their last purchases.
- Based on product categories below and have target marketing strategies to boast sales:
  - High sales but low profit margin: Big bath table, sports leisure
  - Low sales but high profit margin: Home appliances
  - High sales and high profit margin: Computers, small appliances, musical instrument
- Cities:
  - Cities like São Paulo and Rio de Janeiro perform exceptionally well in both the number of orders and revenue generation, and they have shorter delivery times. Based on top cites (mainly the metropolitan ones such as Sao Paulo and Rio De Janeiro), the platform should **allocate more marketing resources to cities if they have already sold a significant number of product**.
  - When customers are more satisfied with when the order is delivered on time with no delays, this would encourage customers to buy more things from the platform again. It will be helpful to **implement more efficient processes to reduce delivery times and improve customer satisfaction**.
  - However, cities with higher average delivery times (e.g., Boa Vista, Macapá) are not appearing in the top revenue or order counts, possibly due to the poor delivery experience. **Improve on the delivery processes and to prevent delivery delays** so that customers can have greater satisfaction.
  - **Giving incentives to the consumers to use credit card as well as Boleto** since both payment methods are the top choices for consumers.

Future work

- Analyzing product categories and their contributions to sales can enhance predictive analytics, such as Market Basket Analysis (MBA), by allowing us to concentrate on high-performing categories and uncover valuable product relationships for further investigation.

- Explore the growth potential in terms of sales for some of the low sales cities.
  - Create targeted marketing campaigns that resonate with the community's culture and values.
  - Leverage local media channels and social media platforms to increase awareness of the platform, positioning it as a top choice for consumers when they're looking to make purchases.
  - Introduce singles day promotion like 10.10 and 11.11 to have massive discounts and thereby encourage a lot of spending and can increase the sales volume significantly.

- Look at the seller's dataset to see where are most sellers situated
  - Sales contributed by several big sellers or by many small sellers => to gauge the risk that the platform may take in case 1 big seller is removed from the platform

Appendix – DB Setup and Connection

- From the SQLite3 console, run the create.sql script:

```
~ ❯ sqlite3 ecommerce.db
SQLite version 3.37.2 2022-01-06 13:25:41
Enter ".help" for usage hints.
sqlite> .read create.sql
```

- Connect to the created database via Python sqlite3:

```python
import sqlite3
import pandas as pd

# Initialize SQLite3 connection
conn = sqlite3.connect('/path/to/your/ecommerce.db')
df   = pd.read_sql('SELECT * FROM orders;', conn)
```

Appendix

Please find the following files via the links:

- SQLite csv files for the database 'csv-20240925T144055Z-001.zip'
- DB file 'ecommerce.db'
- Python codes 'Project.ipynb'

- Information on Brazilian E-Commerce Public Dataset by Olist (kaggle.com)