



Olist Store Review Score Analysis and Customer Segmentation

Scikit-Learn Group:

- Brian Giovanni (brian.giovanni2807@gmail.com)
- Rizqi Irfan Nawwaf (R.irfannawwaf@gmail.com)

This PowerPoint is created using artworks from [freepik.com](https://www.freepik.com)



Table of Content



**EDA &
Review Score
Analysis**



**Customer
Segmentation**



**Conclusions &
Recommendations**



**Data Understanding
& Merging**



**Problem
Formulation**





Problem Formulation

Background



Why **review score** is important ?

- 01** Influence customer's decision
- 02** Strengthen company's credibility
- 03** Feedback for the company

Why **targeted marketing strategy** is important?

- 01** Enhanced advertising strategy
- 02** Build strong customer loyalty

Thus, we need to identify the target market using **customer segmentation**



Problem Statement

01

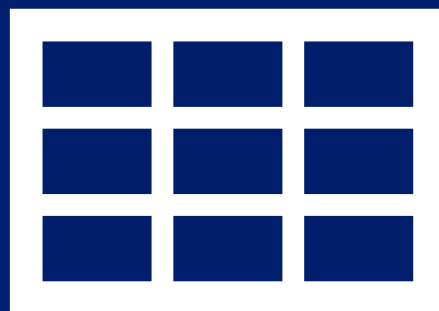


What makes a low or high **review score**?

02



What **type of customers** are using the Olist Store?



Data Understanding & Merging

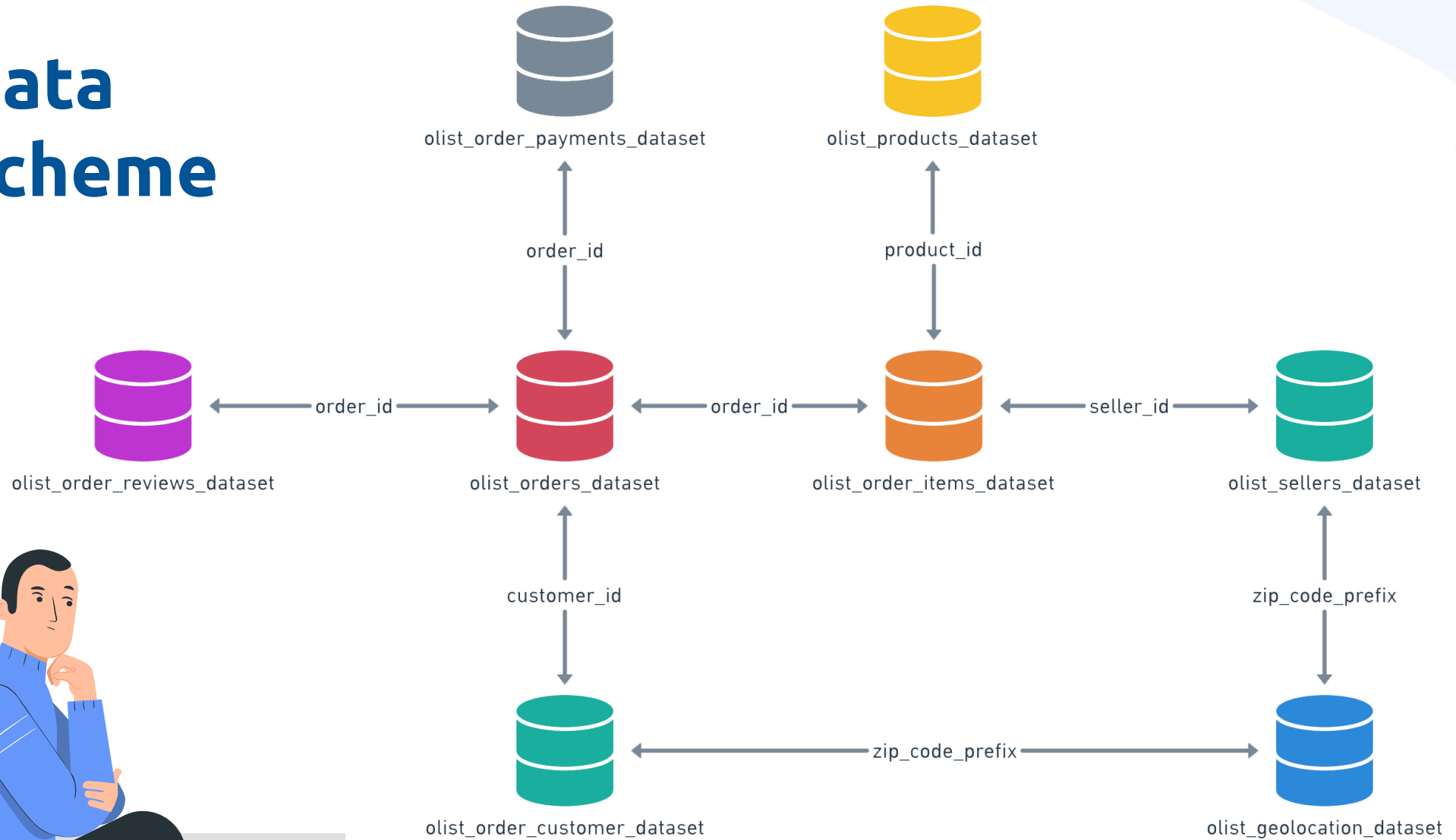
The Datasets

There are 9 datasets used in this project, which can be downloaded from [Kaggle](#).

- Customer: data about [customer's information](#) and location
- Geolocation: data about Brazilian [zip codes](#) and its [lat/lng coordinates](#)
- Order Items: data about [item purchased](#) for each order
- Order Payments: data about [payment options](#) for each order
- Order Reviews: data about [reviews](#) made by the customers
- Orders: data about [orders information](#)
- Products: data about [product specification](#)
- Sellers: data about the [sellers](#) that fulfilled orders
- Product Translations: include product's [English translation](#)



Data Scheme



Data Merging

Datasets Shape

Dataset	Rows	Columns
customers	99.441	5
geo	1.000.163	5
order_items	112.650	7
order_payments	103.886	5
order_reviews	100.000	7
orders	99.441	8
products	32.951	9
sellers	3.095	4



Merge



df_merged

Shape	Value
Rows	117.601
Columns	38



Feature Engineering

Added 3 new features:



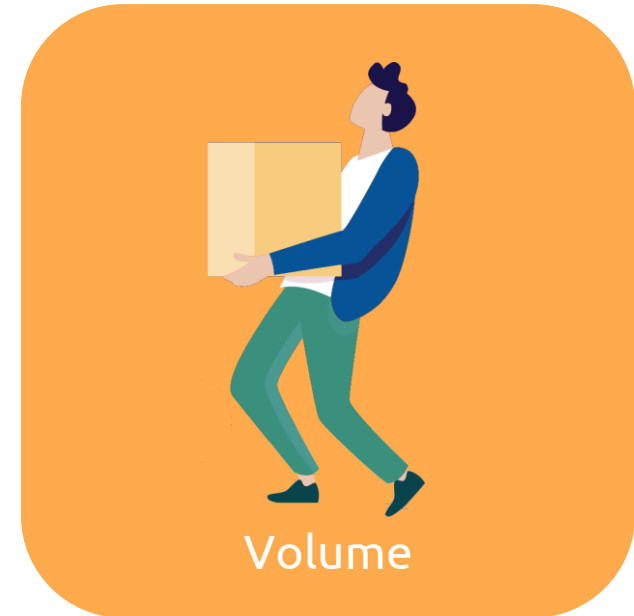
Delivery Difference

The time difference between the actual and estimated delivery date



Order Process

Time needed to process an order



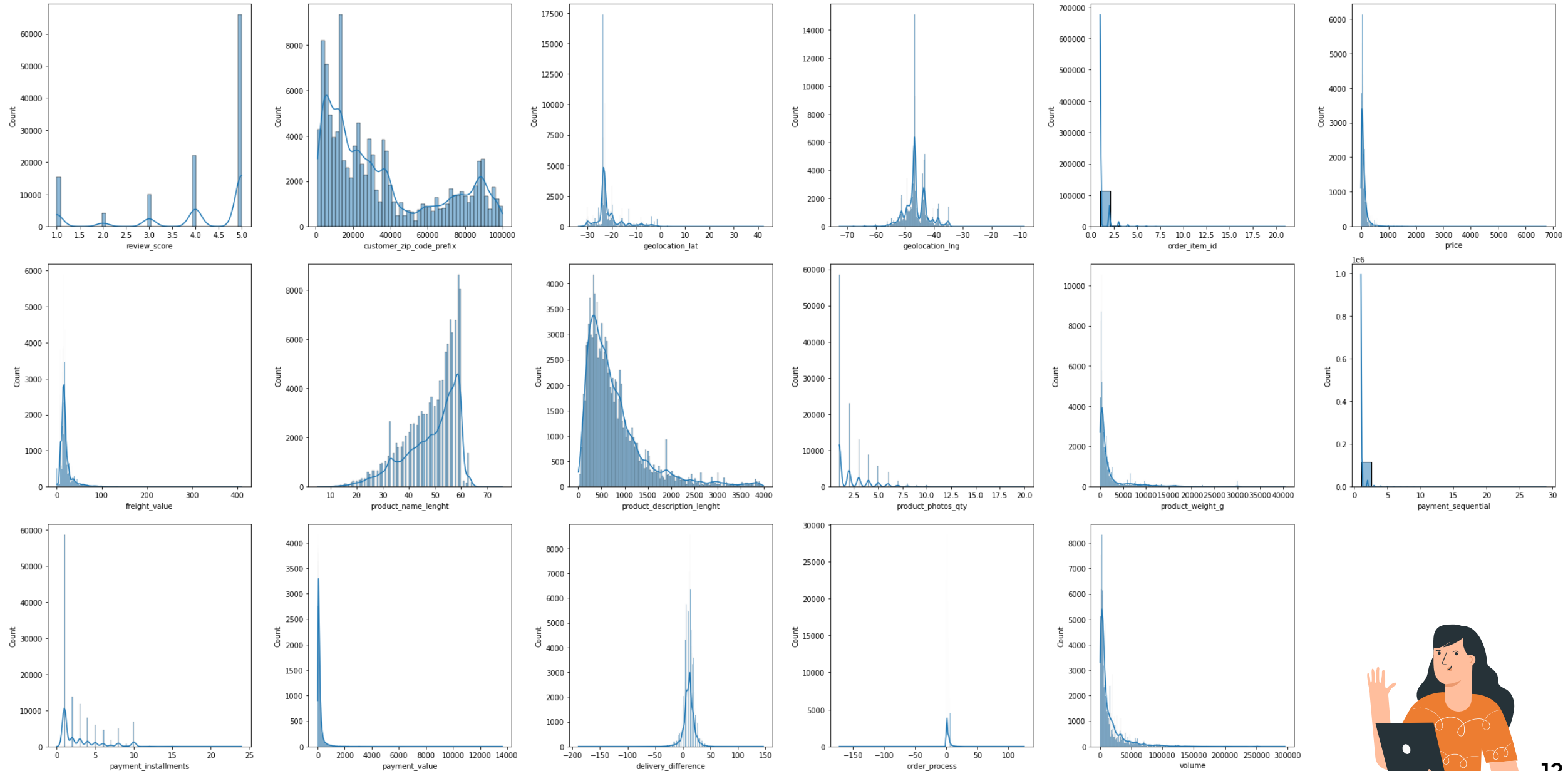
Volume

The volume of the product

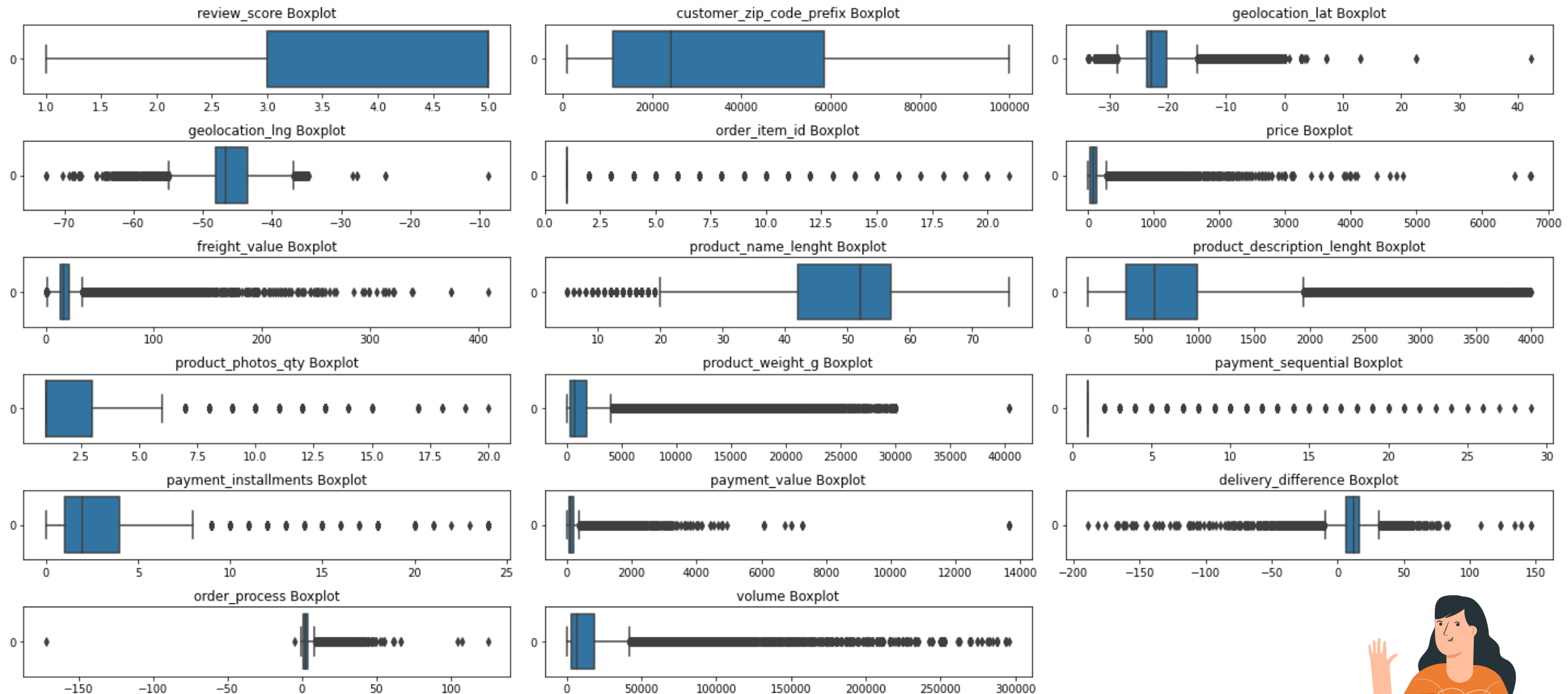


**EDA &
Review Score
Analysis**

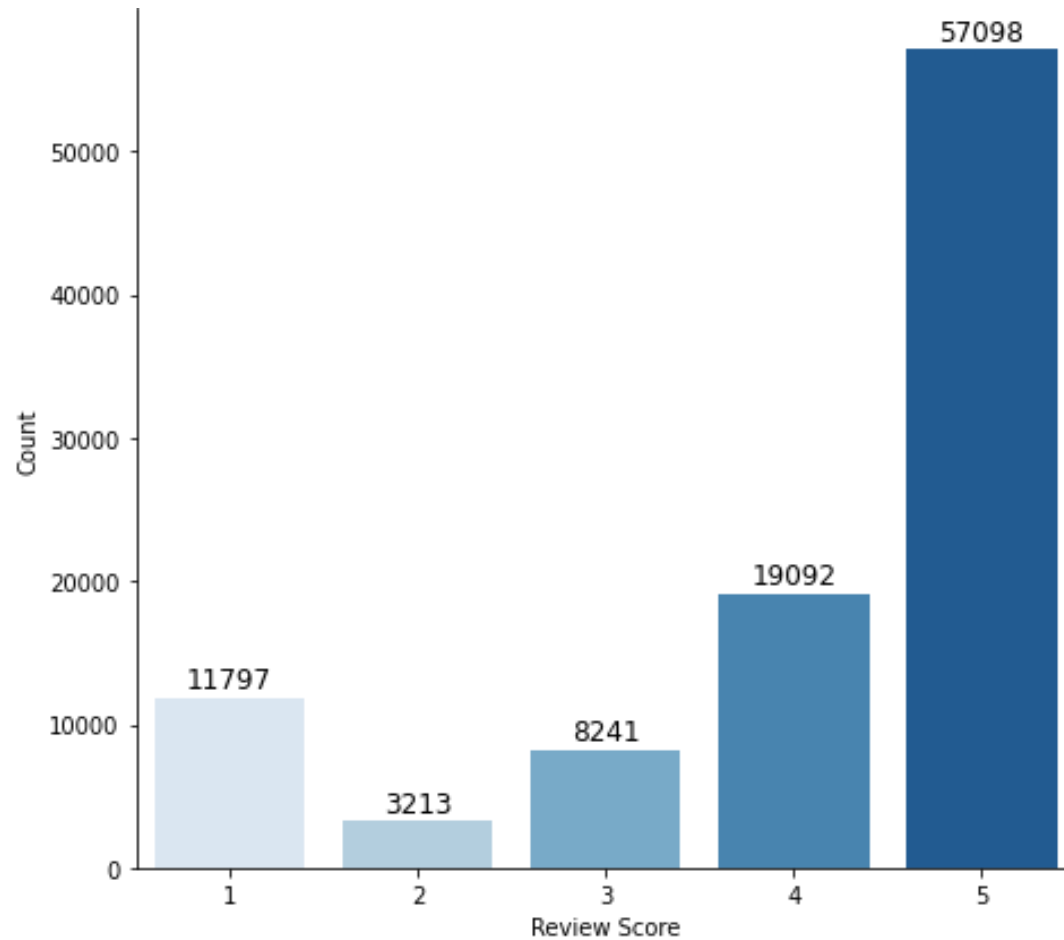
Numerical Data Distribution



Numerical Data Boxplot



Review Score

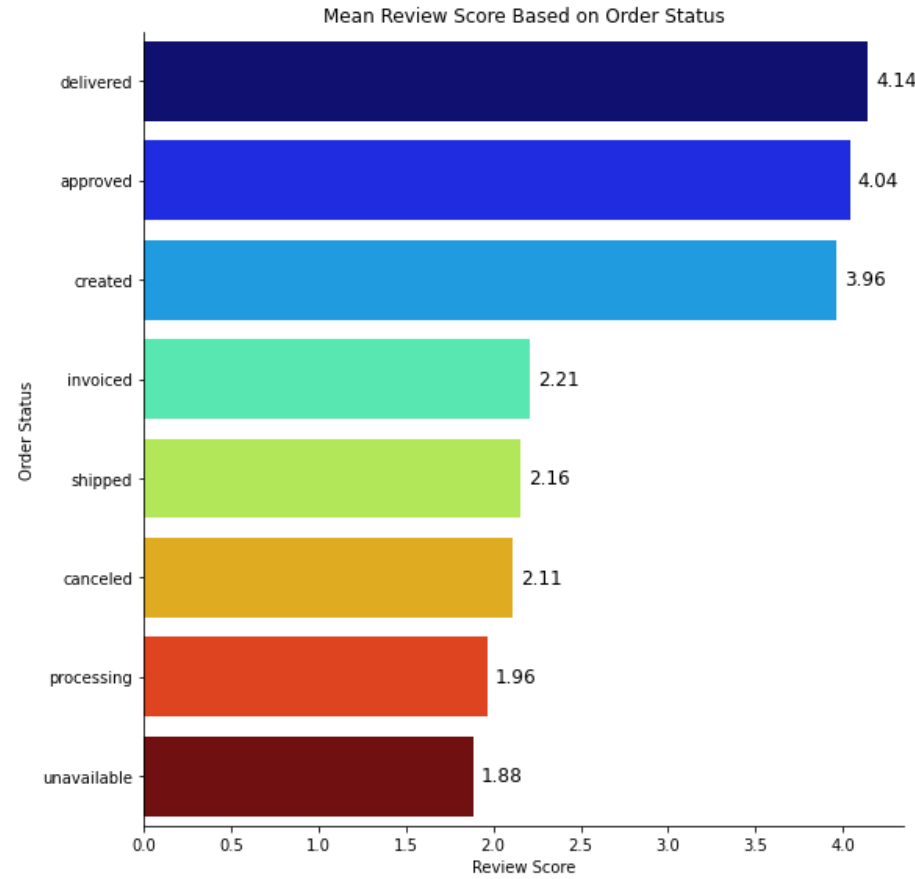
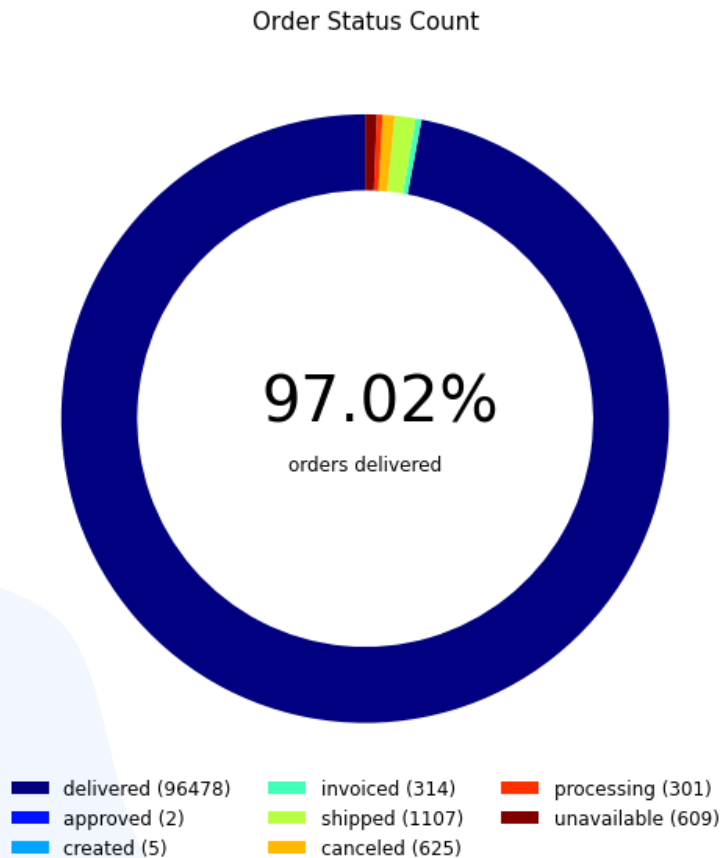


Customer Rating

4.1
(99441)



Review Score Based On Order Status



*review score is calculated using the weighted rating

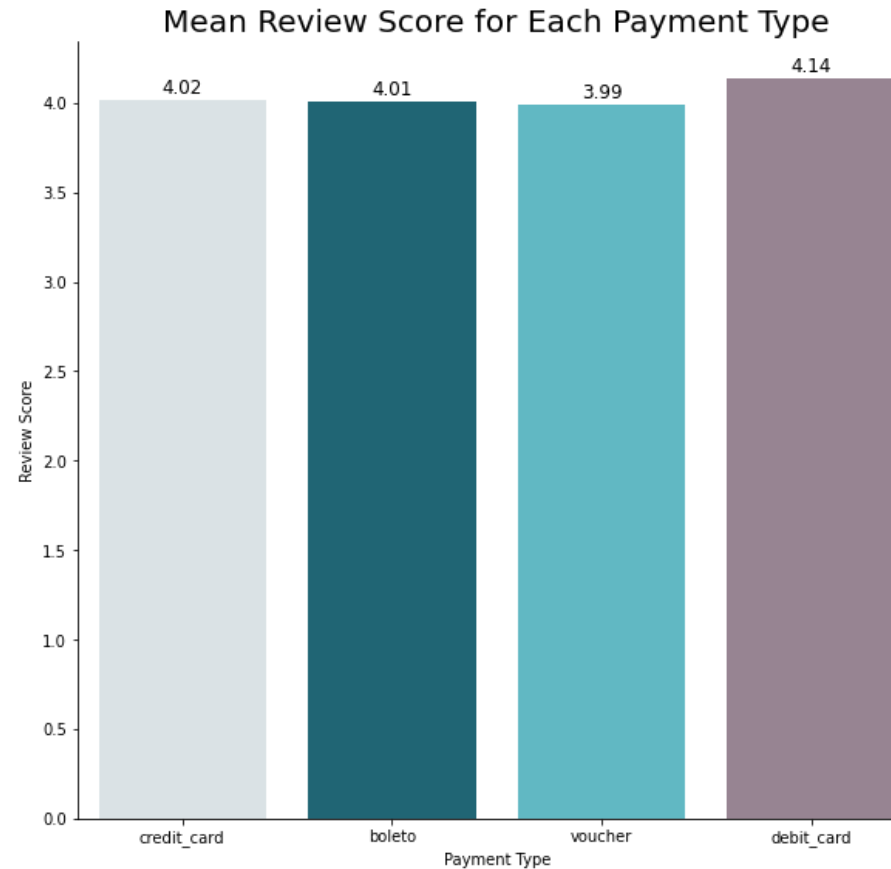
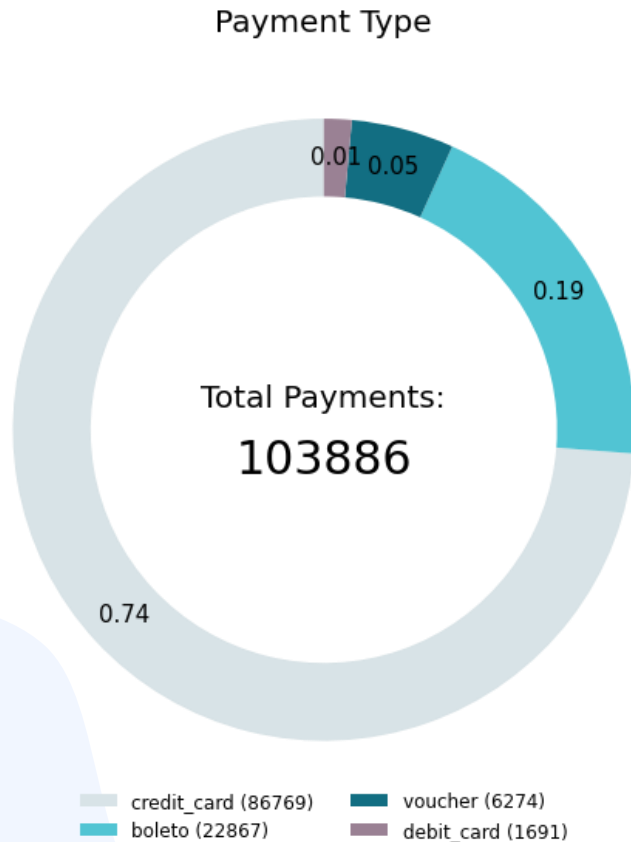
Our Recommendations

Undelivered orders tend to have a bad review score

Thus, we suggested to add a title (for example fast sellers, trusted sellers, or top-rated sellers) to fast and reliable sellers in order to reduce the amount of cancelled or a "long time to process" orders (undelivered orders)



Review Score Based On Payment Type



Our Recommendations

Majority of the customer use **credit card** as a way to pay.

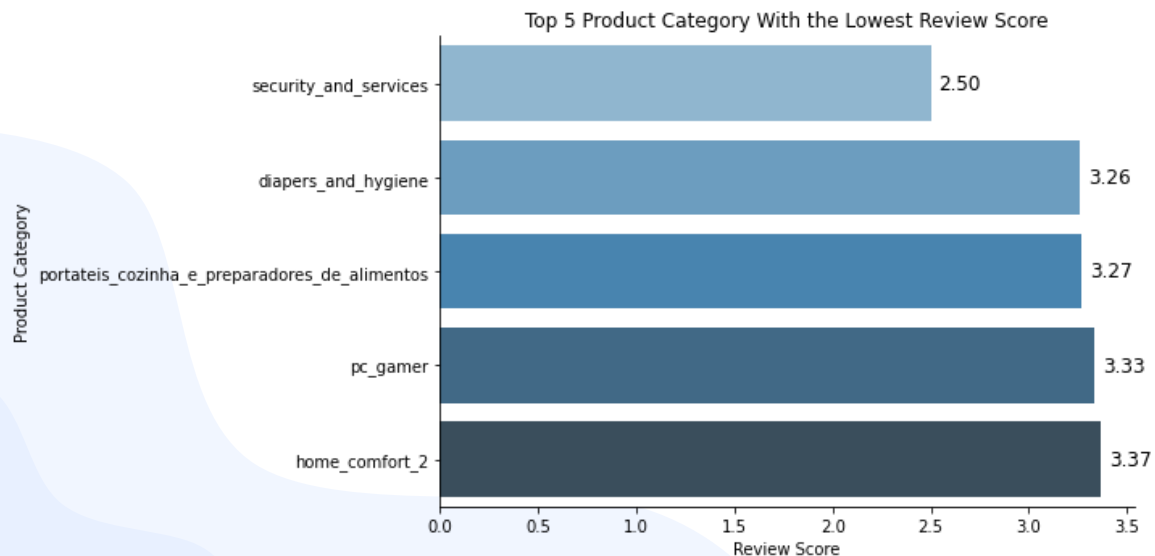
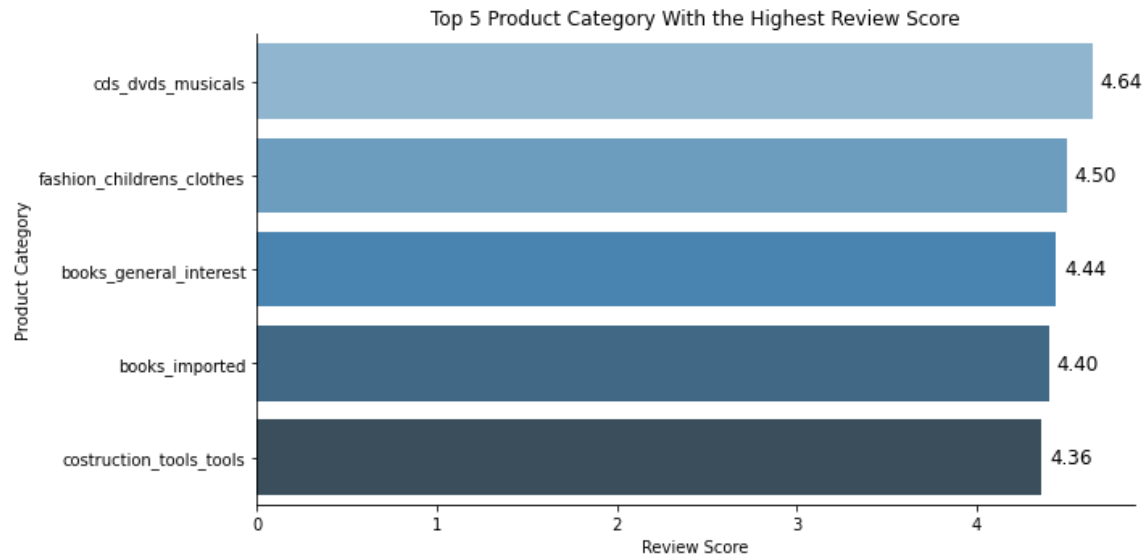
Thus, we suggest to:

- Give some **promo** when the payment type used is **credit card**
- **Encourage** the customers to also pay with **debit card** since it is **highly rated** by the customer



*review score is calculated using the weighted rating

Product Review Score

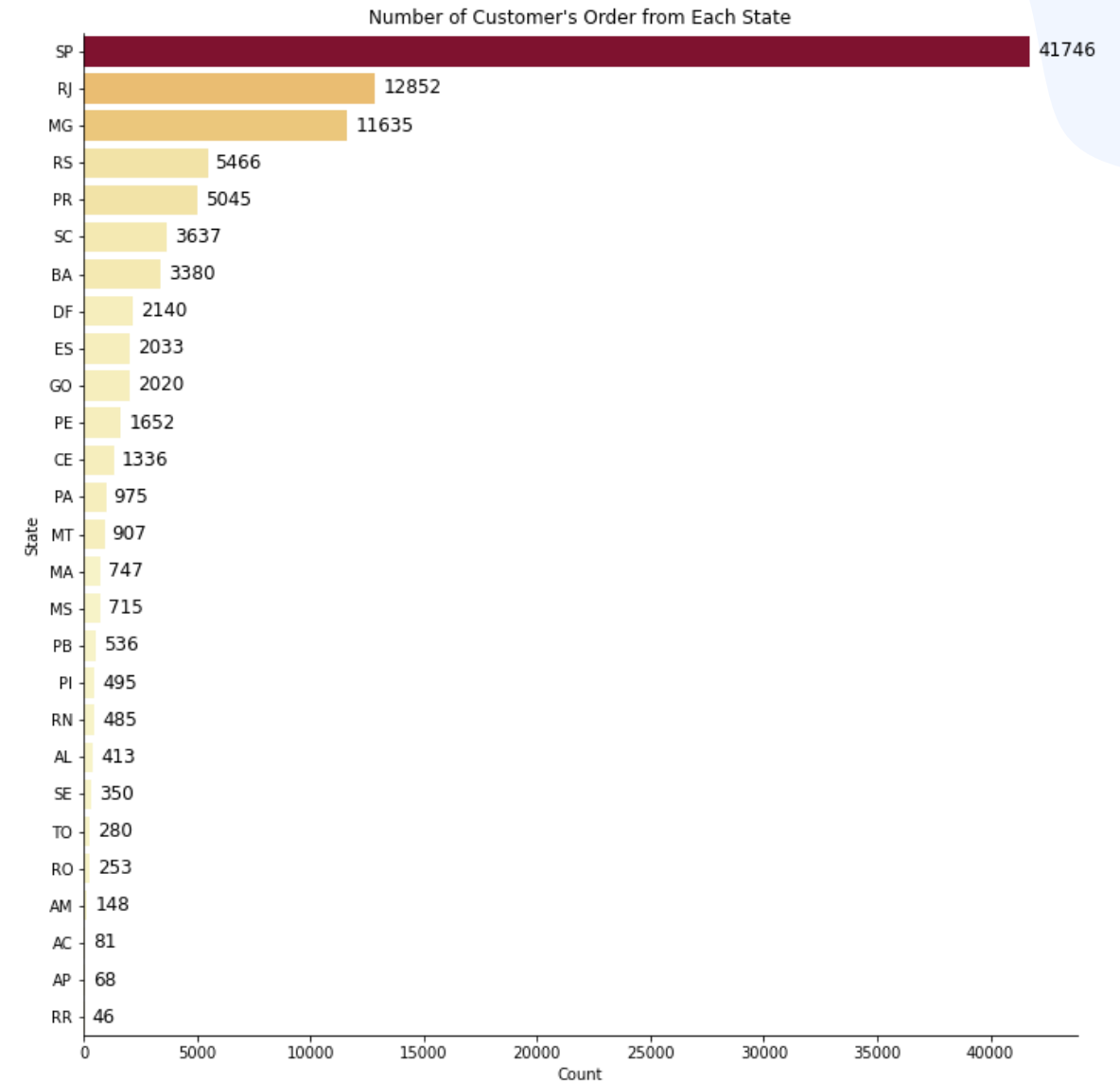
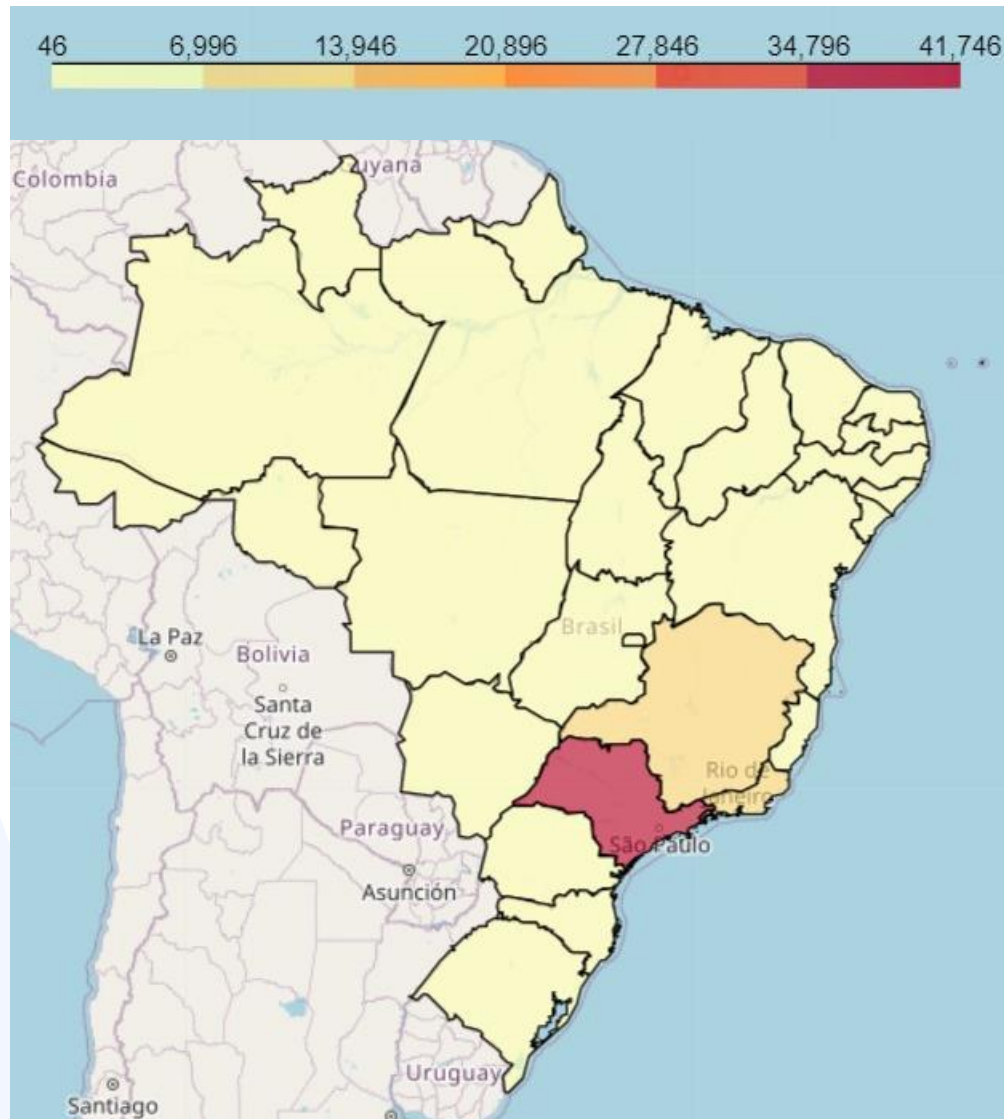


Our Recommendations

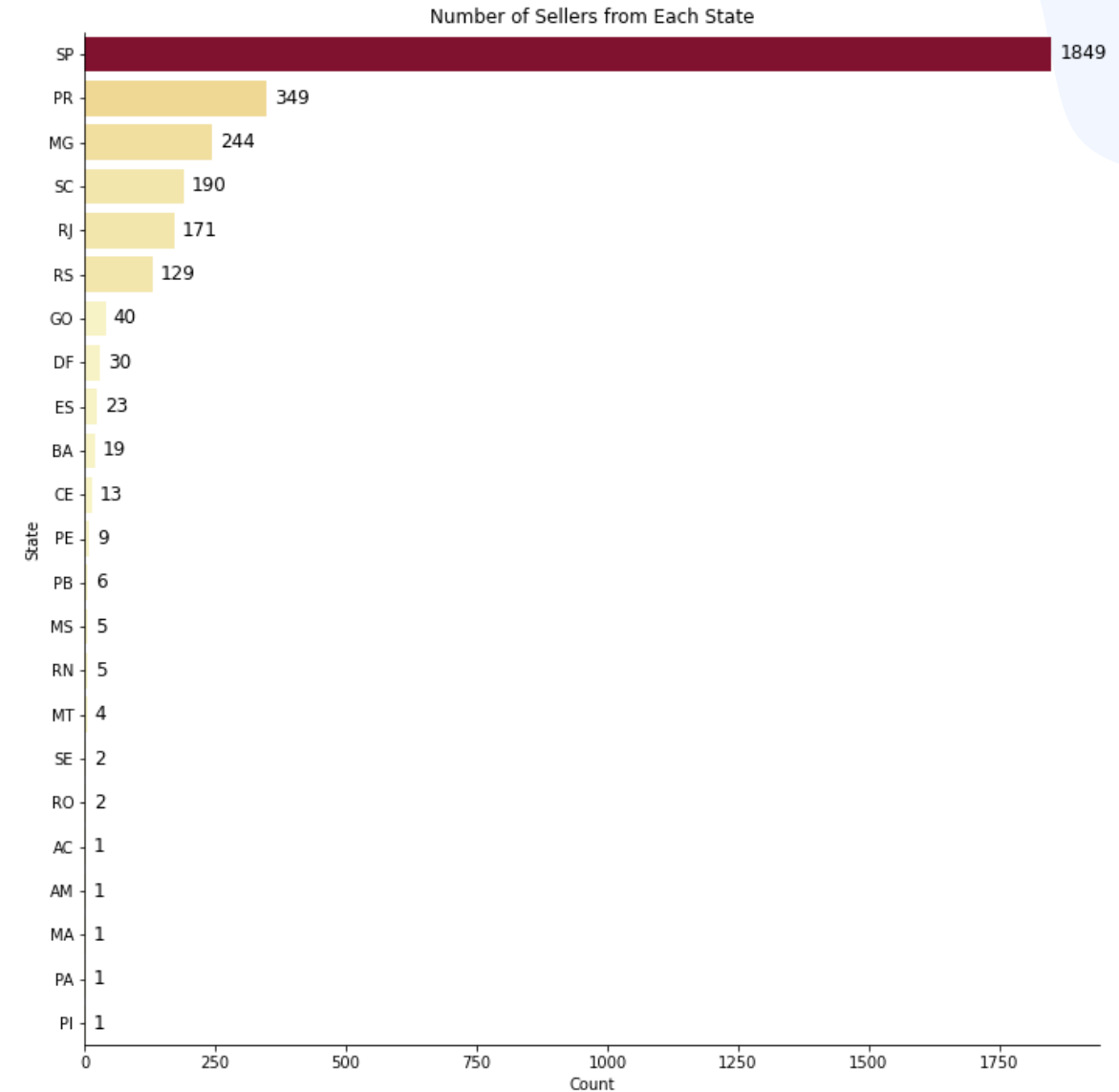
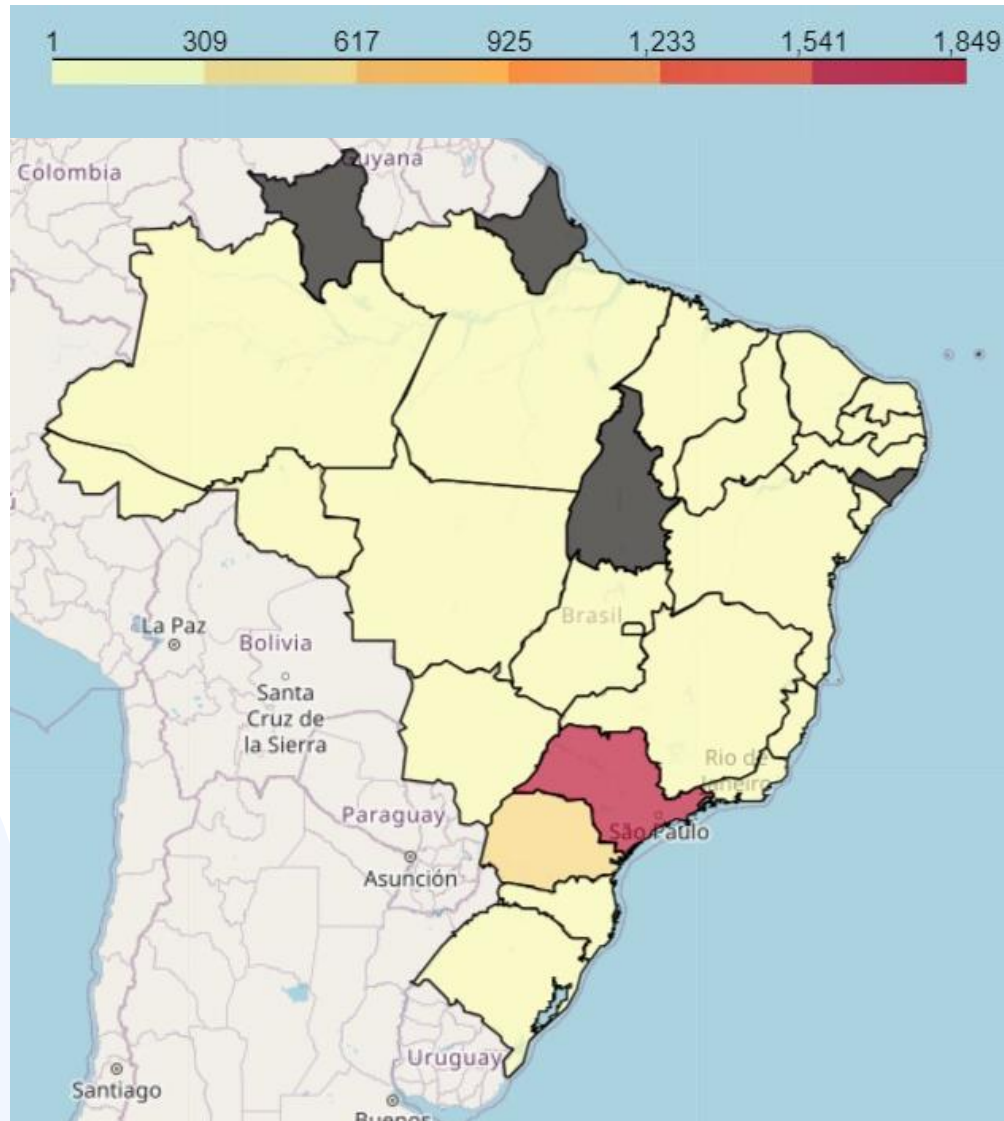
- Create more **ads and recommendations** in the Olist Store about the **highly rated** product category
- Perform further **analysis** to investigate why the **poorly rated** product category is rated that way



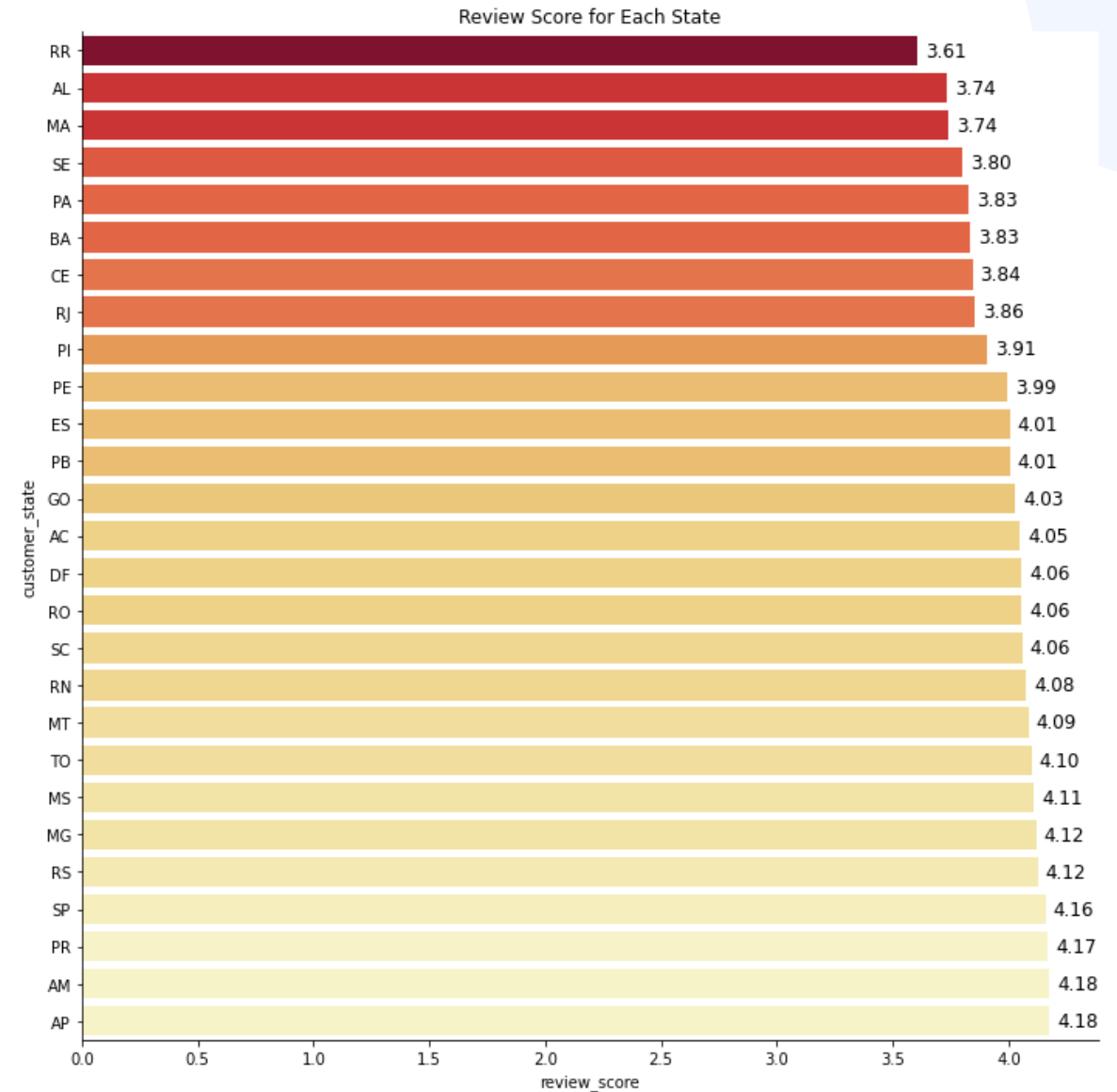
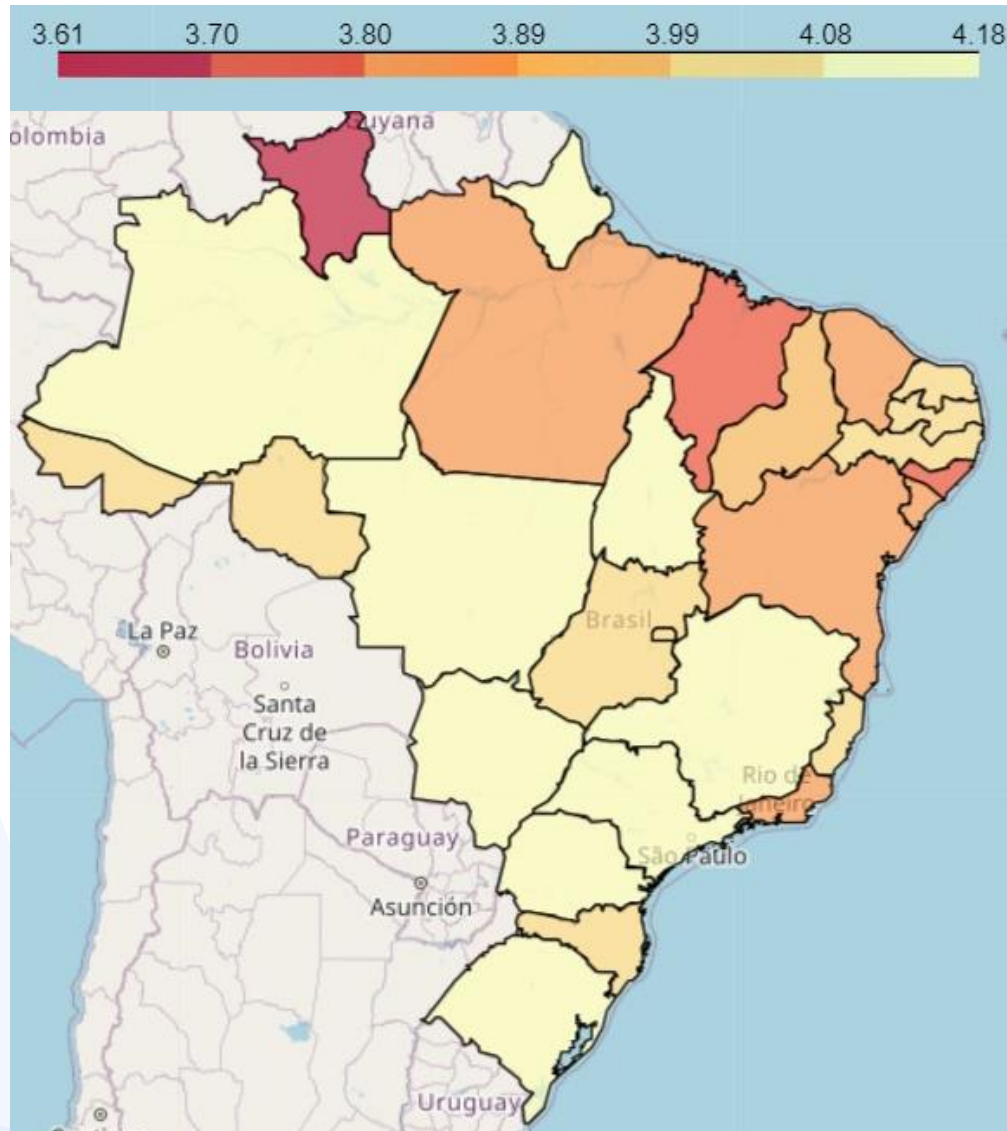
Customer Distribution



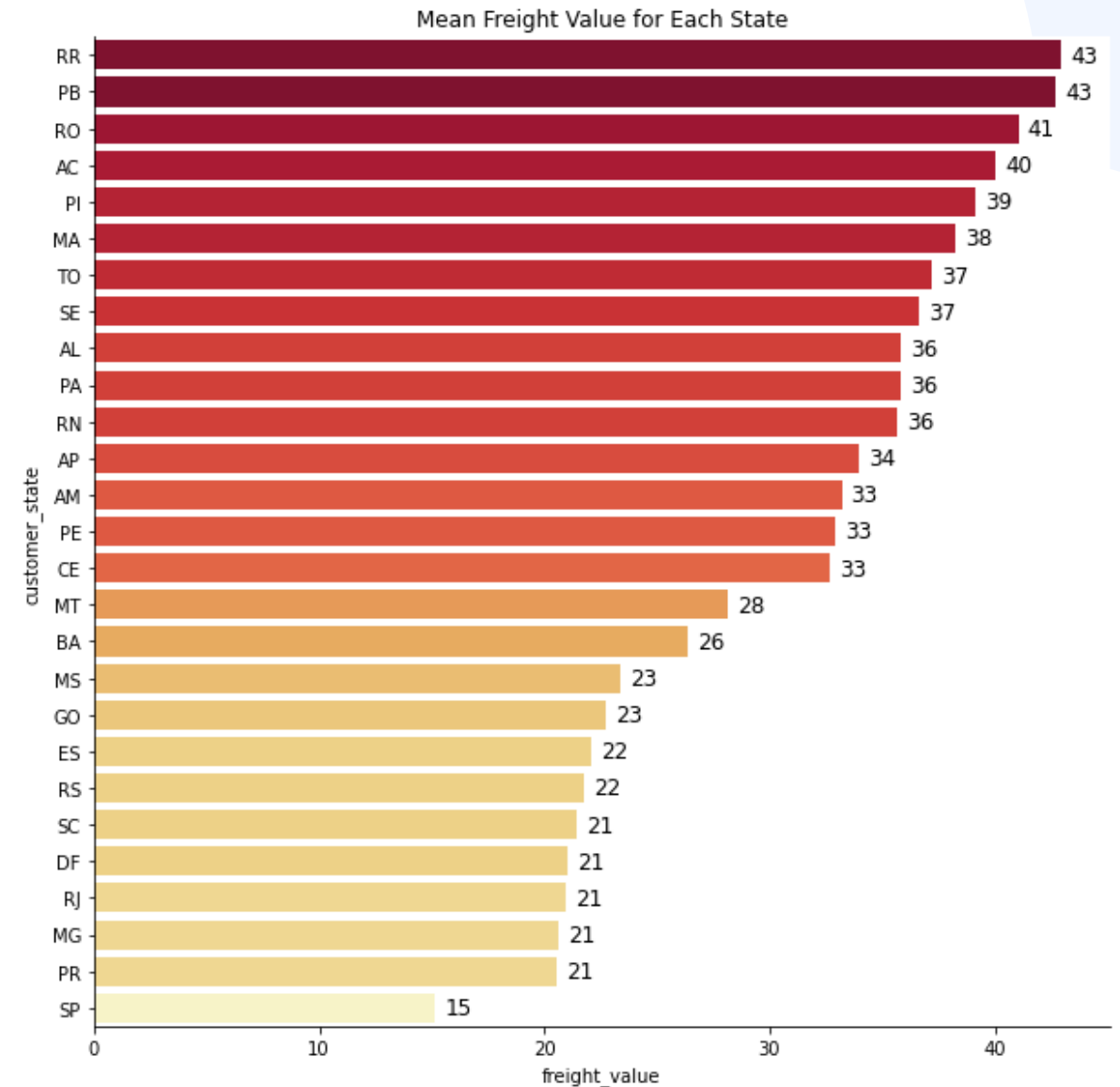
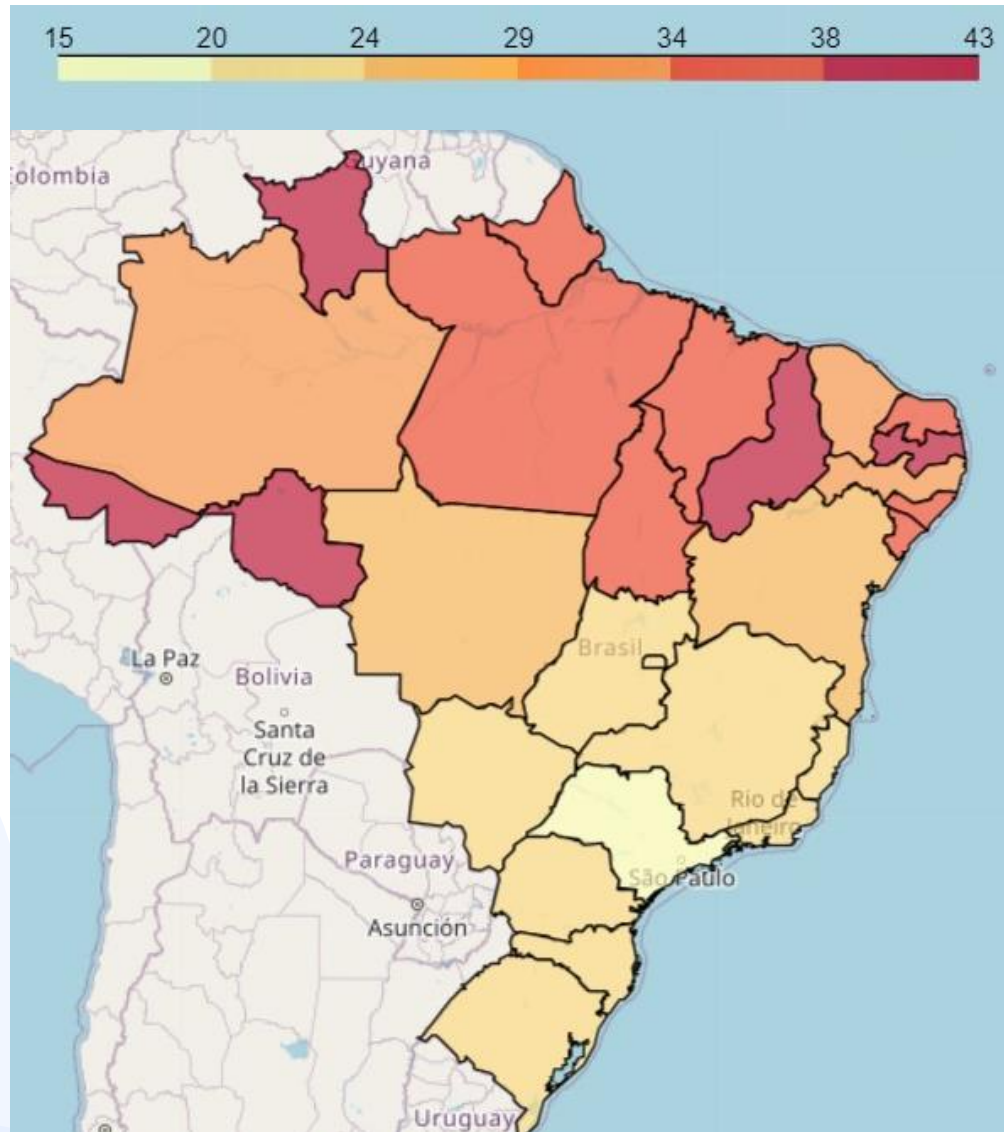
Seller Distribution



Review Score from Each State



Average Freight Price from Each State



Geolocation Review Score



The customers from the **northern side** of Brazil tend to give a **lower review score** than the customers from the **southern side** of Brazil



The customers from the **northern side** of Brazil tend to **pay higher for freight price** than the customers from the **southern side** of Brazil



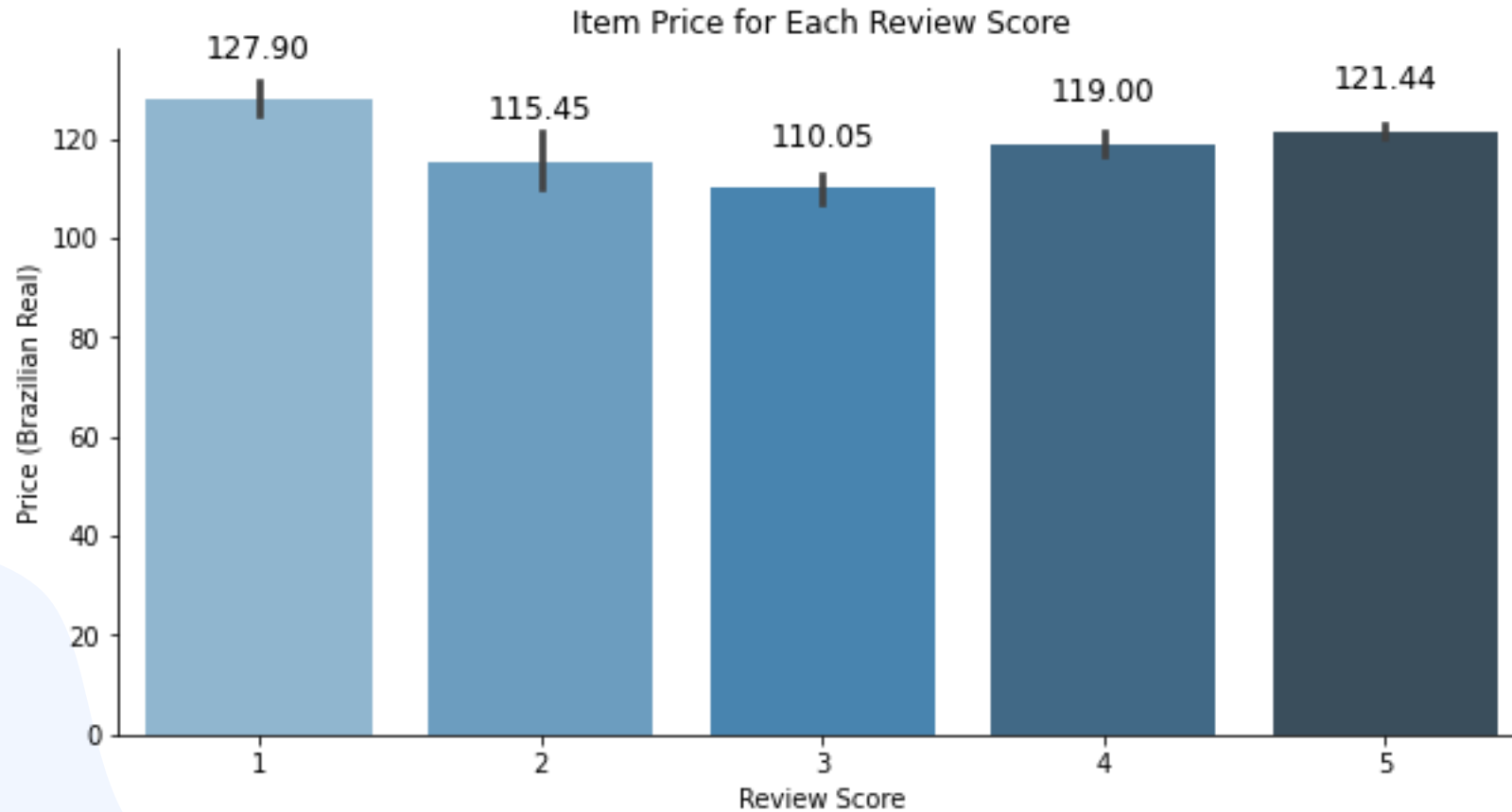
Majority of the **sellers** are in the **southern side** of Brazil

Our Recommendations

Encourage sellers from the north Brazil that **hasn't use** the Olist Store to use it so that **freight price** can be **minimized**



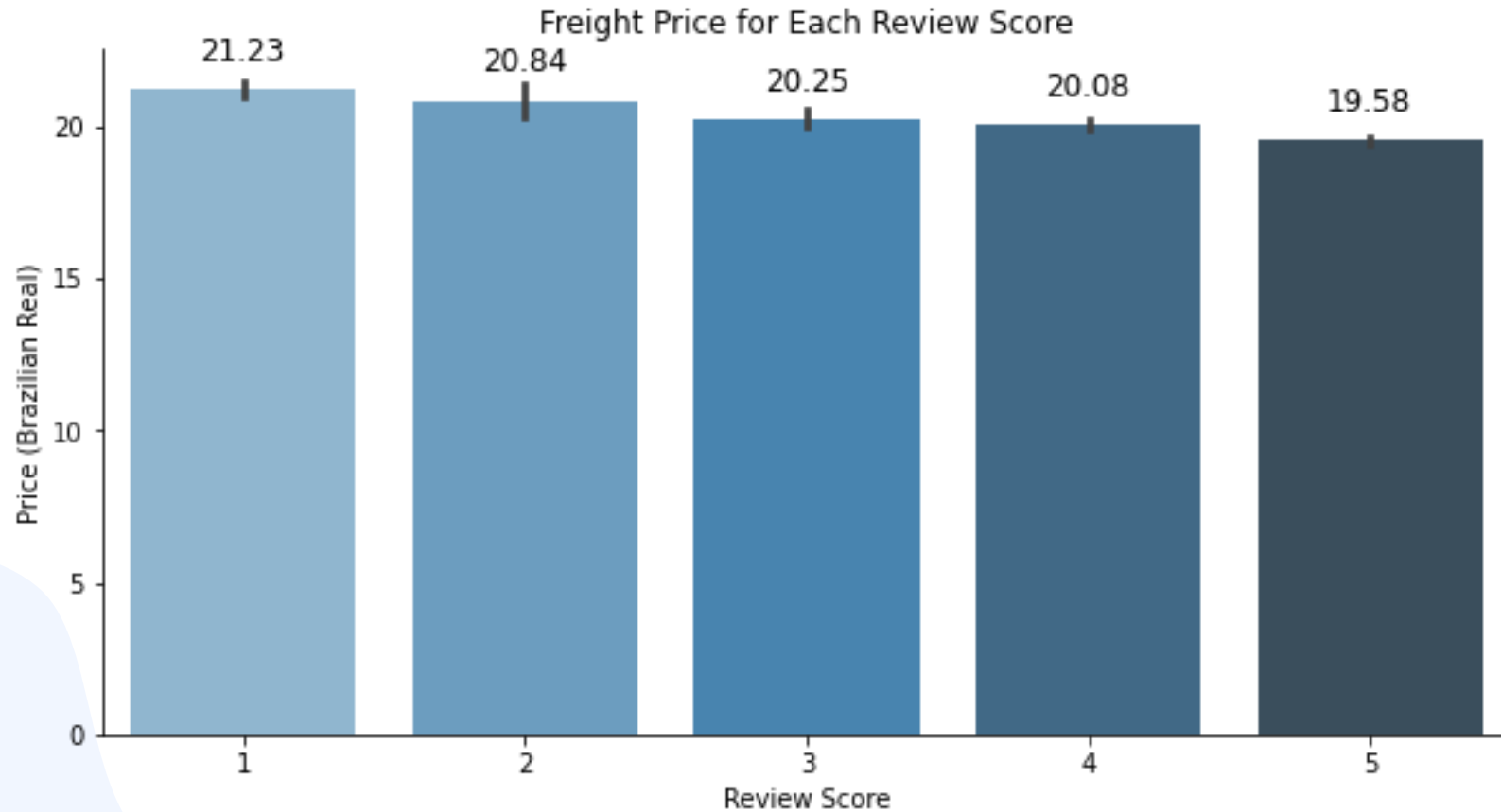
Average Price for Each Review Score



Findings

There are **no significant pattern** on price based on review score with the highest average price is when the review score is 1.

Average Freight for Each Review Score

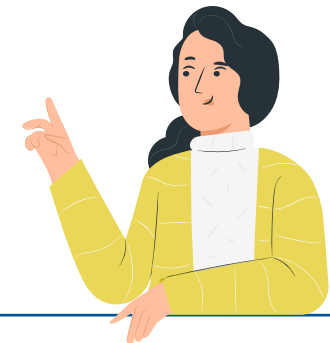


Our Recommendations

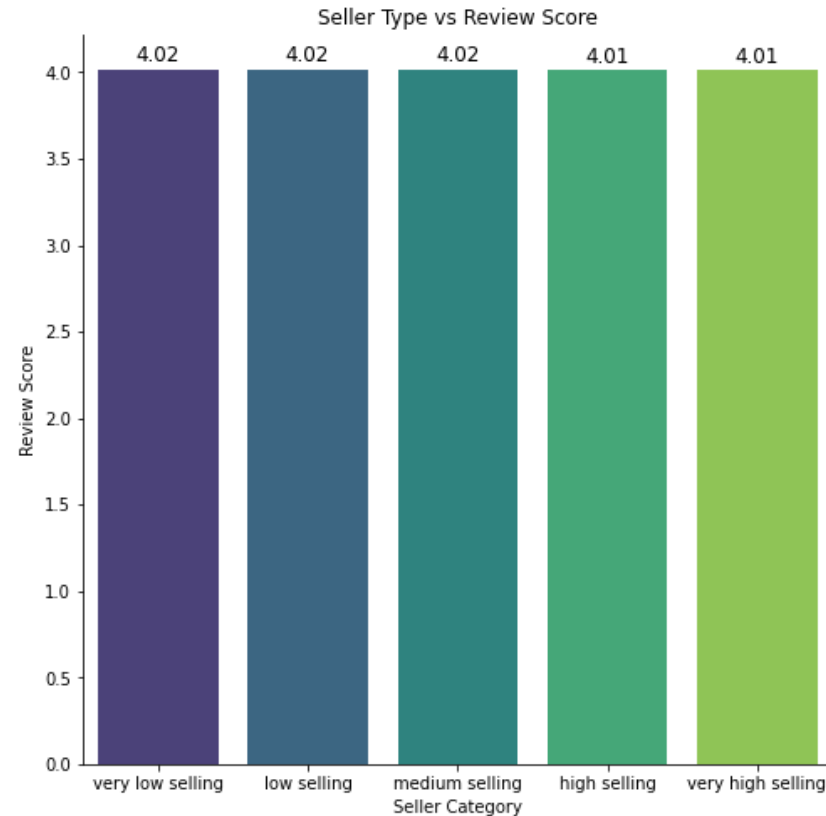
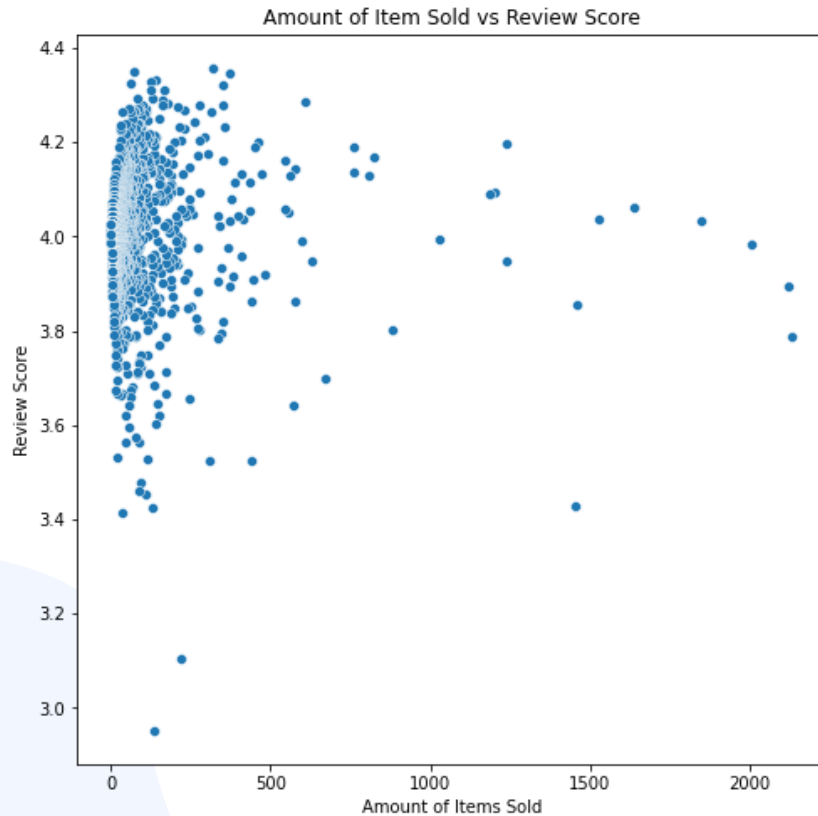
Here we see a trend where **lower freight price** tend to have a **higher review score**.

Thus, we suggest to:

- Partner with a better **logistic company** to **reduce** the freight price
- Give a **promo** about the **freight price**.



Amount of Items Sold vs Review Score

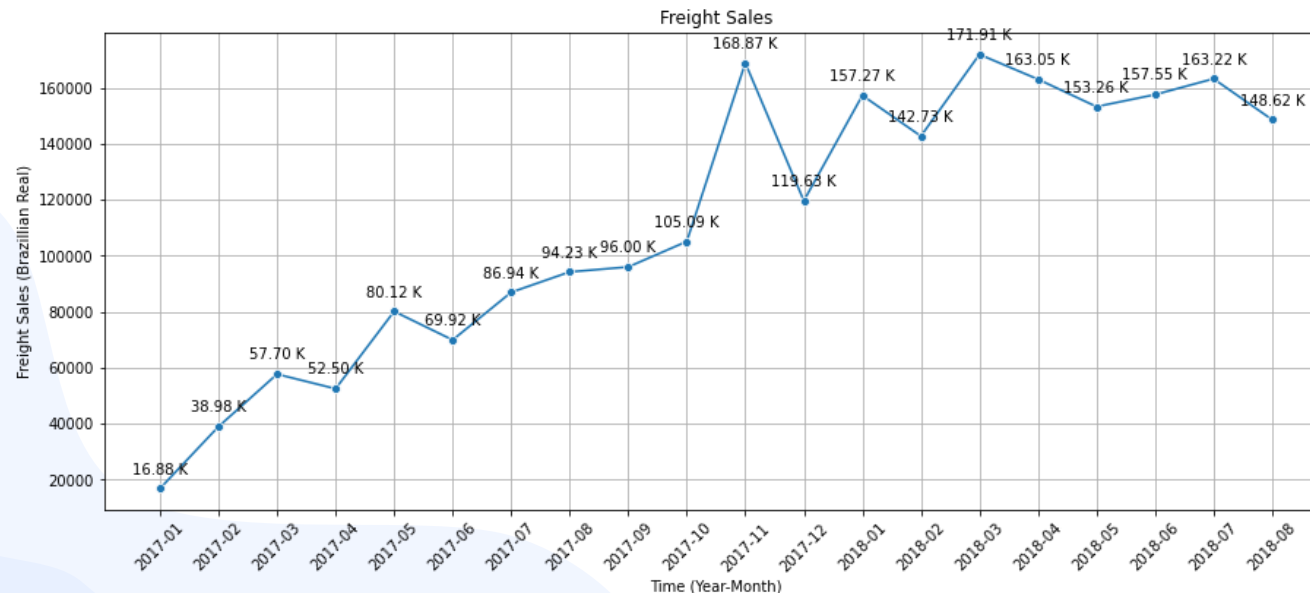
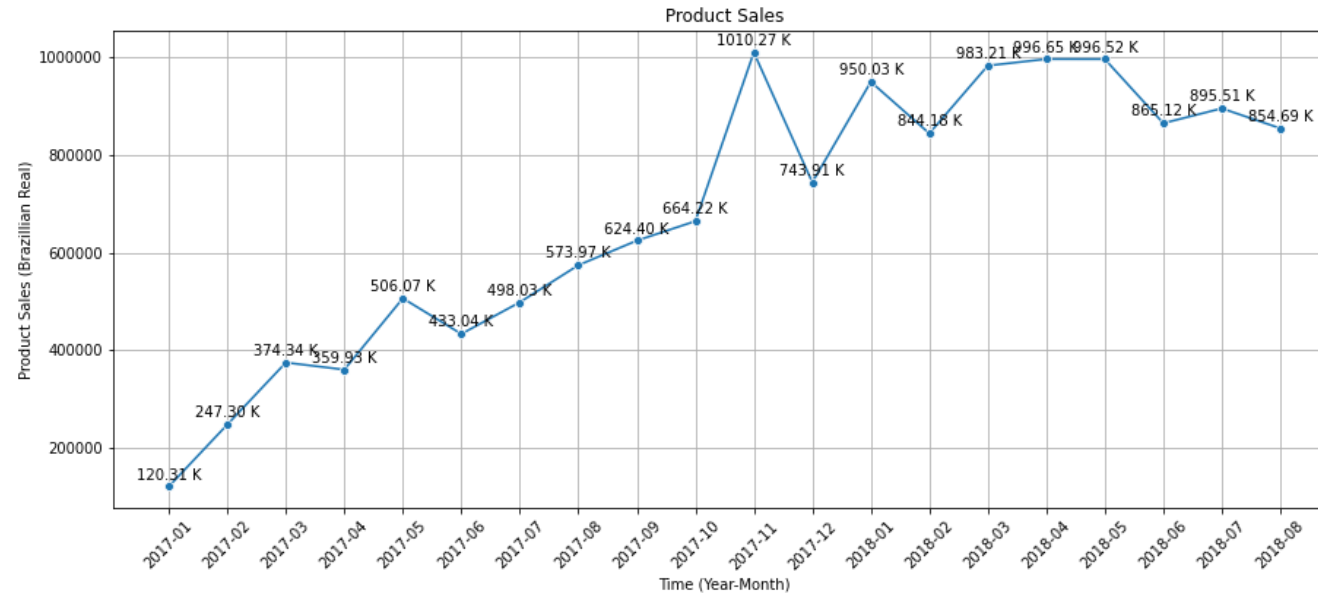


Findings

Here, we can see that the amount of items sold by a particular seller does not have any effect to the review score.

*review score is calculated using the weighted rating

Sales Over Time

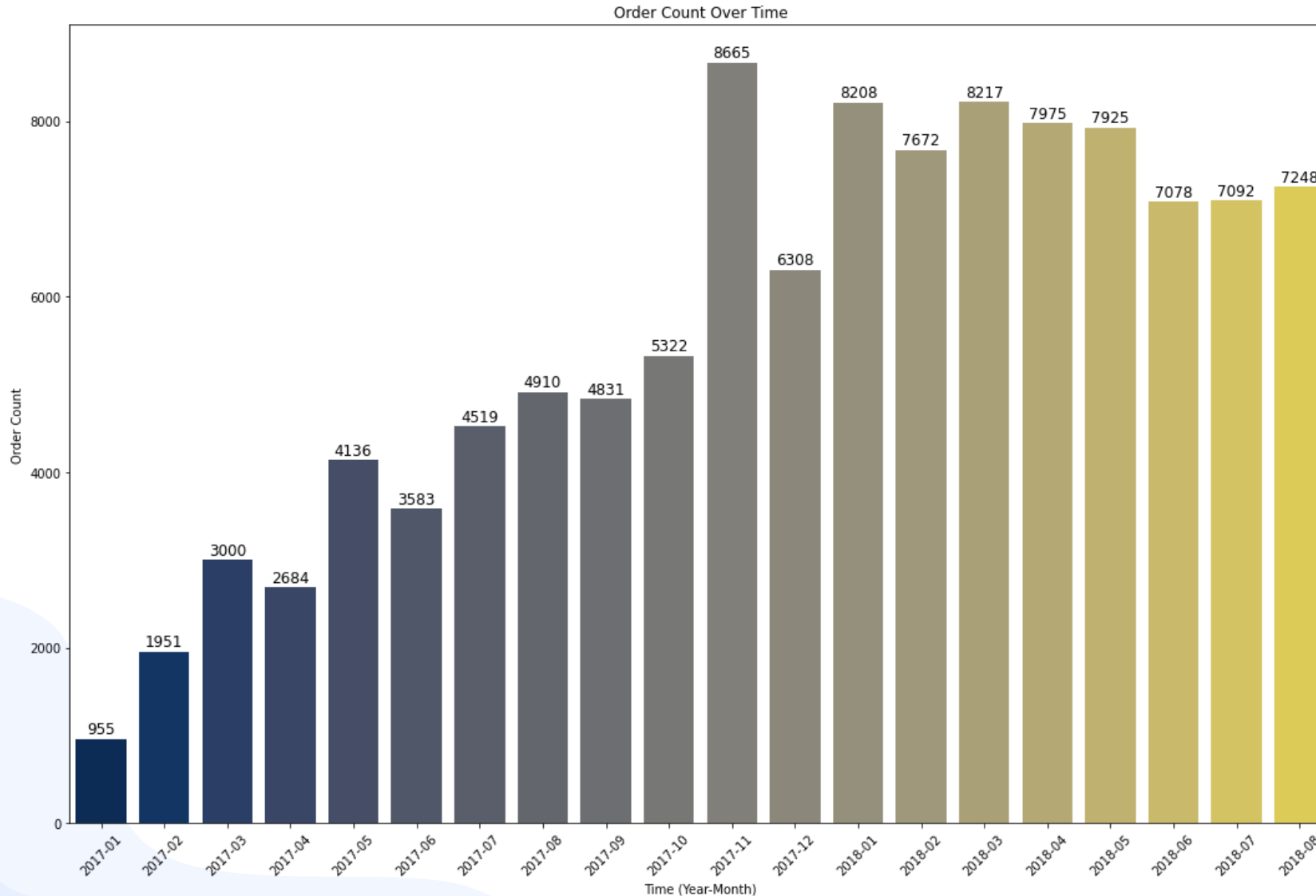


Findings

- There is a **growing trend** in **product sales** with its peak of 1.01 million Brazilian Real is achieved on **November 2017**.
- There is also a **growing trend** of **freight sales** with its peak of 171.9 K Brazilian Real is achieved on **March 2018**.

Thus, it is recommended to **apply business strategy** that was applied on **November 2017** and **March 2018**.

Orders Over Time

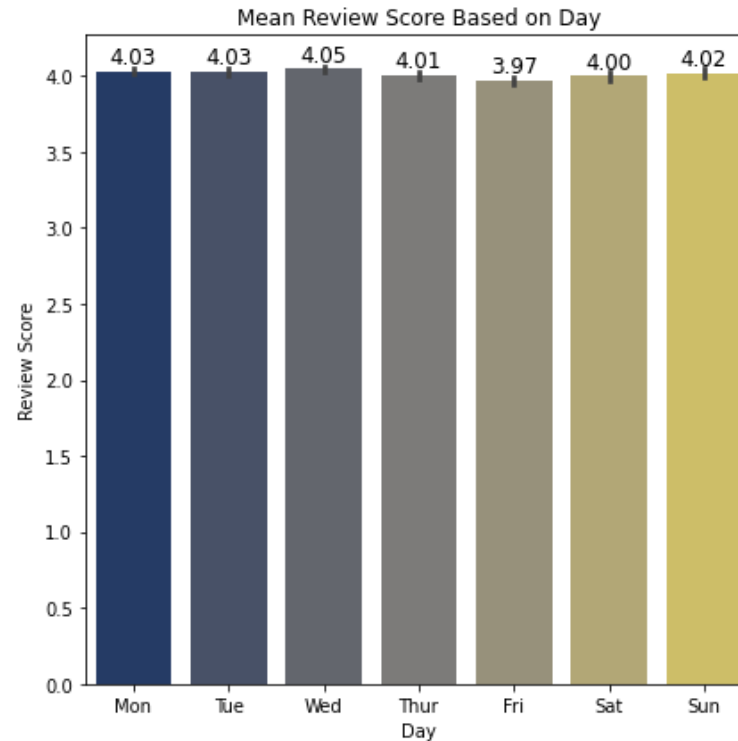
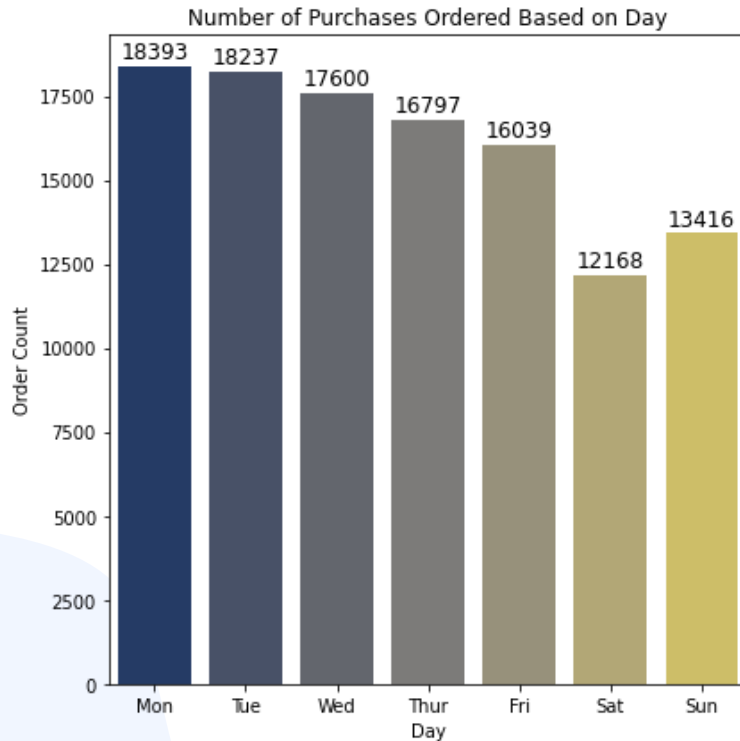


Findings

Same as product sales, there is a **growing trend** in number of order with peak order of 8665 orders a month is achieved on **November 2017**.

Thus, it is also advised to **apply business strategy** that was applied on **November 2017**.

Day of Week vs Review Score

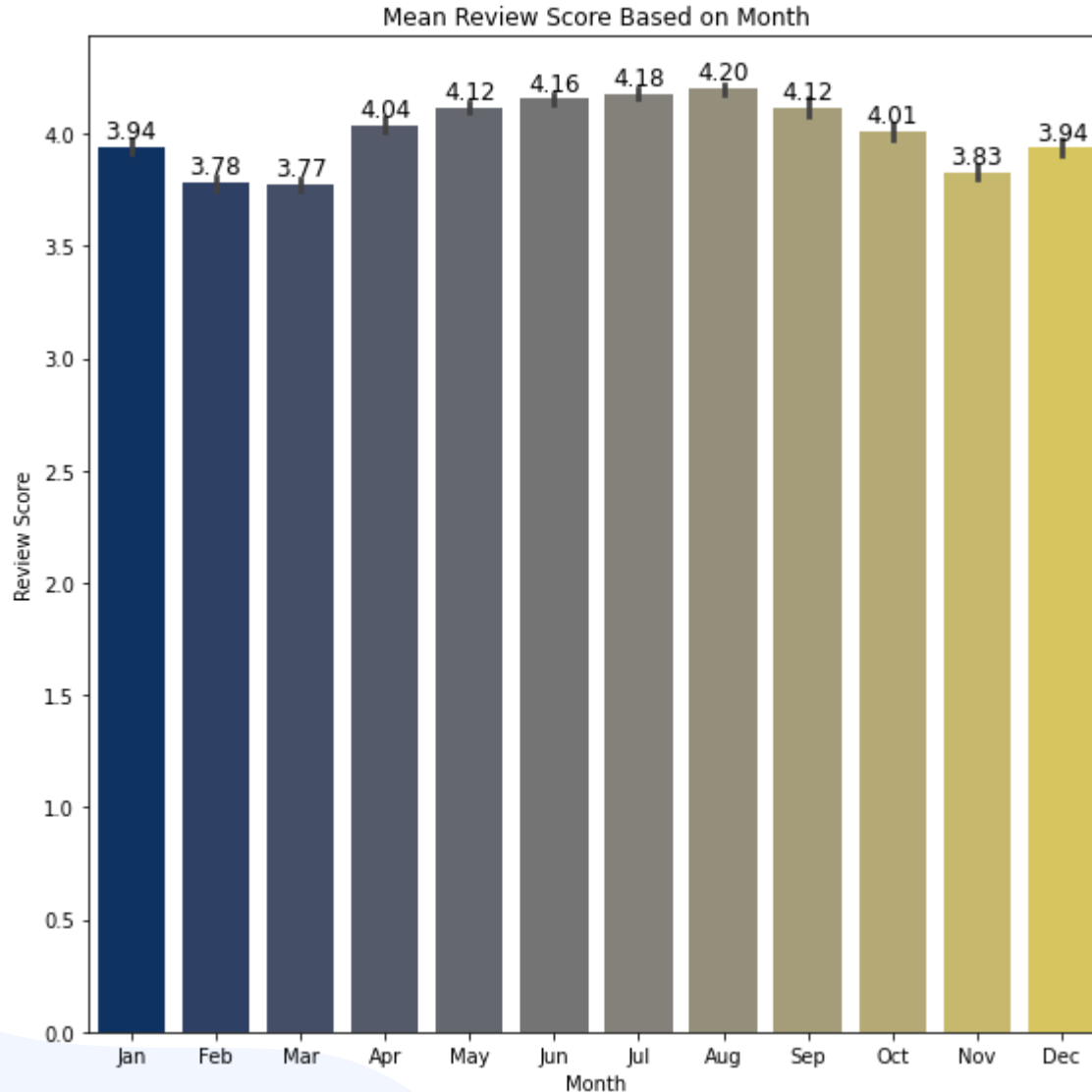


Findings

Customers in Brazil tend to purchase an order on Monday.

However, there isn't much difference in review score based on the day of purchase with the largest difference is between Wednesday and Friday.

Month of Year vs Review Score



Findings

Here, we can see that orders tend to have a [high review score](#) in the [mid year](#) (between April to October)

Delivery Difference Based on Review Score



Our Recommendations

Based on the graph, **better delivery performance** (order delivered faster than estimated) leads to **higher review score**.

Thus, we suggested to partner with a **better logistic company** to increase **delivery performance** in order to **reduce delivery delays**.



Order Processing Time Based on Review Score



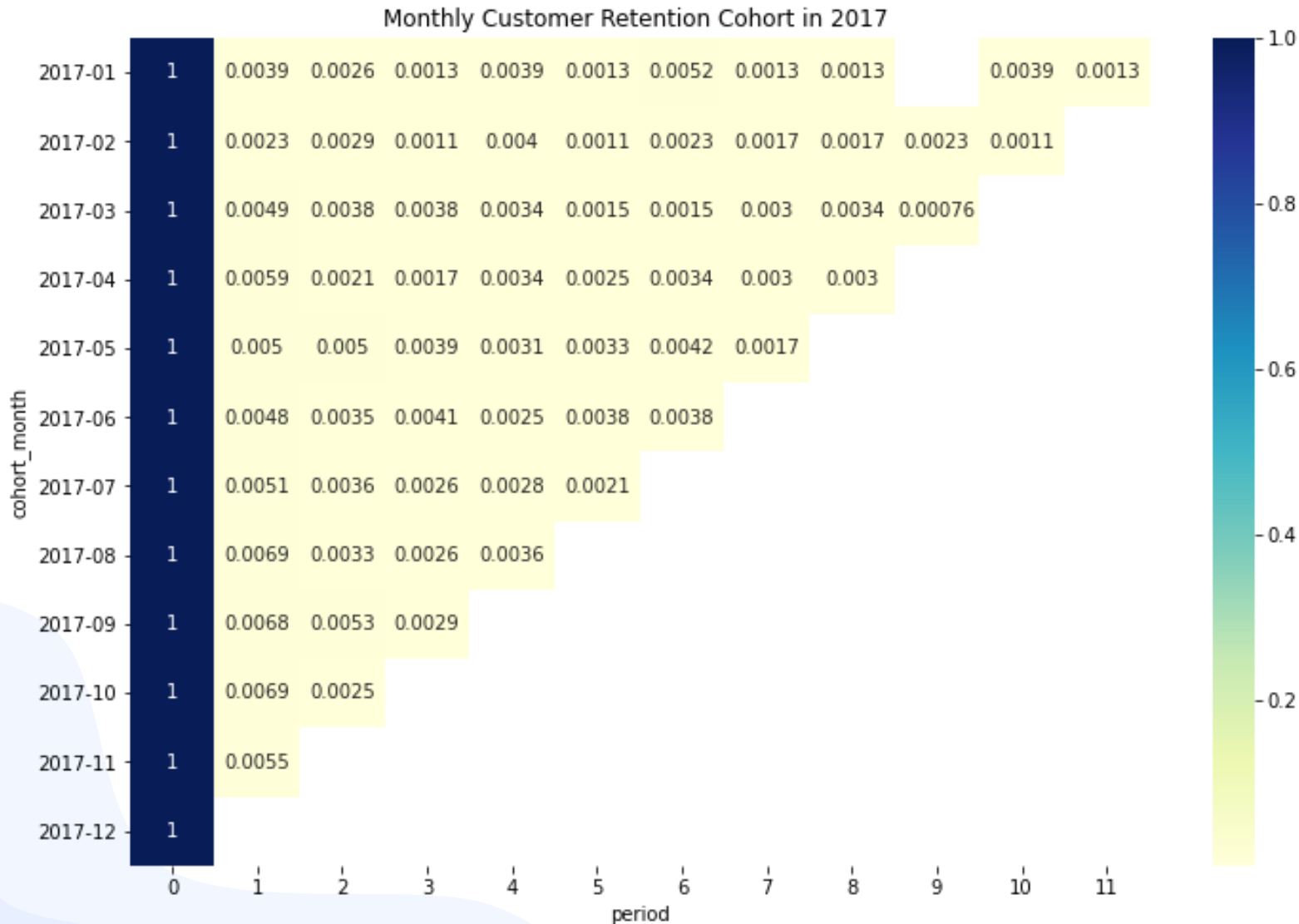
Our Recommendations

As we can see from the above graph, the **lower** the amount of **time needed to process** an order, the **higher** the **review score** given.

Thus, we suggested to **create** a new **ordering systems** that is more **faster and reliable** in order to **reduce** the order **processing time**.



Cohort Retention in 2017



Our Recommendations

Here, we can see that the retention rate in 2017 is very small. Thus, we suggested to:

- Create an event or promo for a returning customers
- Create loyalty program for customers
- Create a personalized buying experience
- Send promotions to the customers via email
- Give a great promo for the customer on their birthday





Customer Segmentation

Preprocessing

01

Drop Categorical and Irrelevant Features

Here, we drop the categorical features since all of the **categorical features** is **irrelevant** to the clustering purpose. We also drop **irrelevant features** like zip code, latitude and longitude.

02

Fill Missing Values

For the missing values, we decided to impute all missing values with the **median** since the distribution is **not normal**.

03

Scaling

The numerical data on the dataset contains **a lot of outlier**. Thus, to reduce the effect of the outliers, we will use **Robust Scaler** to scale the features.



PCA (Principal Component Analysis)

Component	Explained Variance (%)
0	46.900984
1	19.110638
2	6.907825
3	5.818597
4	4.362501
5	3.707702
6	3.391827
7	2.896662
8	2.717454
9	2.376757
10	1.809053



The preprocessed dataset contains **11 features** and **117601 rows**



It may be **space and time consuming** to process the dataset



There may also be **redundant features** in the dataset

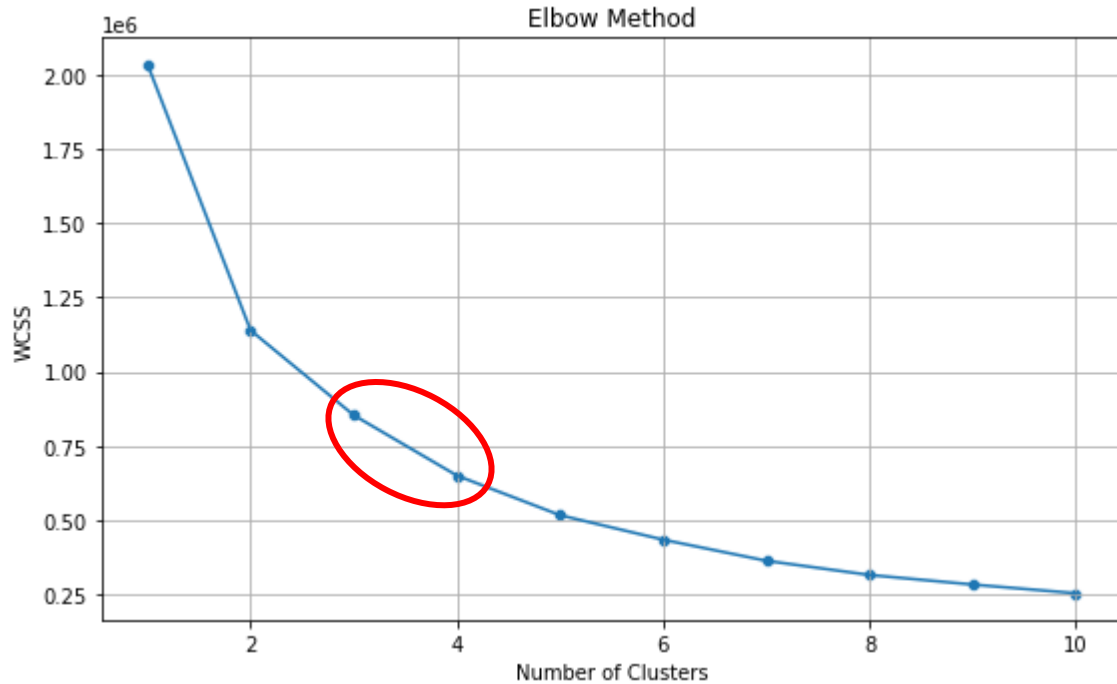


Need **dimensionality reduction**

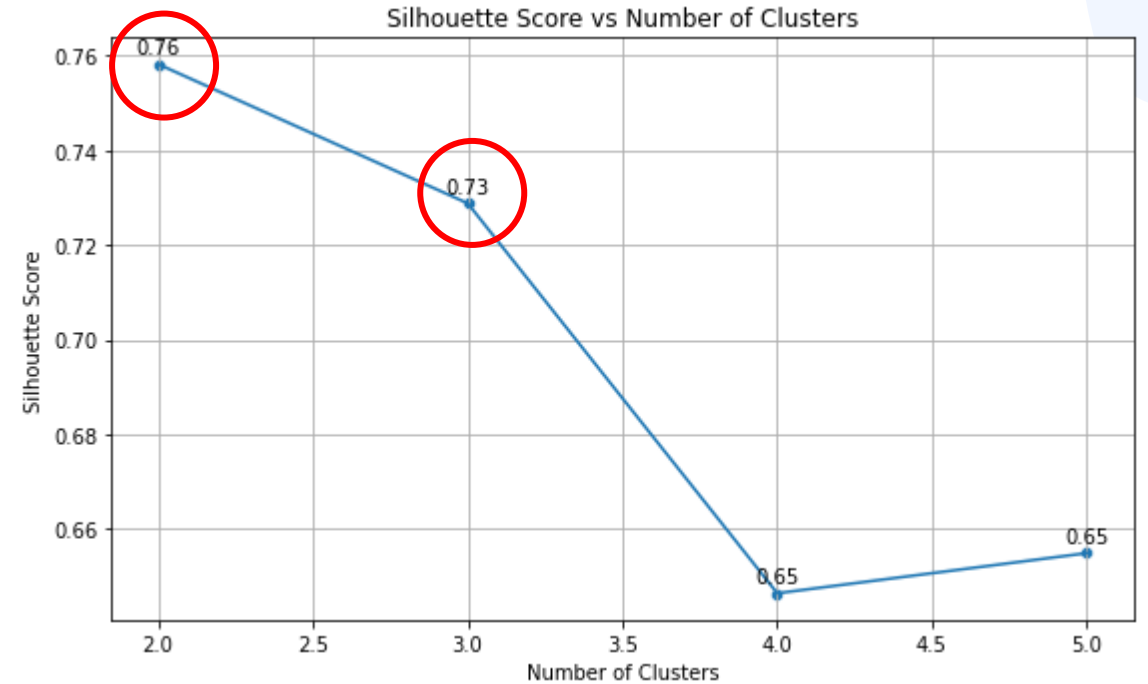
From the result, we chose **2 principal components** to perform PCA because the combined explained variance is high enough (66,01 %).

K-Means Clustering

Choosing Optimal Number of Cluster



Based on the Elbow Method, the optimal number of cluster is 3 or 4

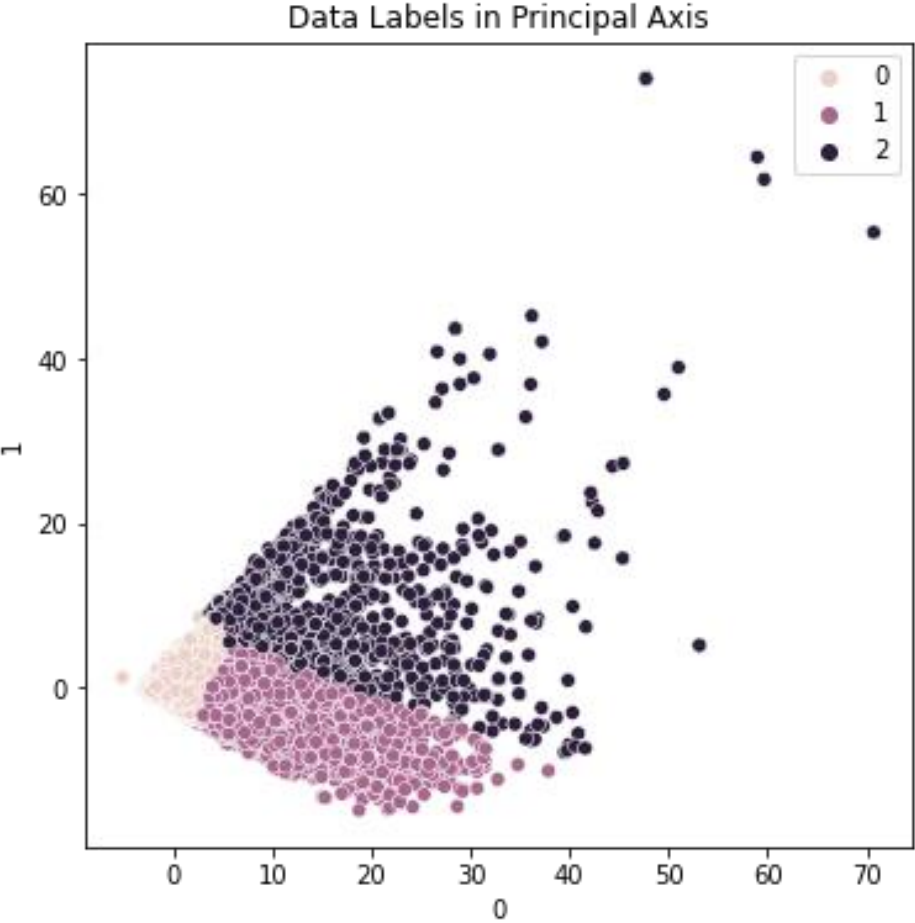


Based on the Silhouette Score, the optimal number of cluster is 2 or 3

Thus, we chose 3 as the optimal number of cluster



Customer Segmentation



Label	Price (R\$)	Freight (R\$)	Product Weight (g)	Product Volume (cm ³)	Payment Installments
0	94.61	17.10	1145.44	10295.60	2.77
1	222.26	46.16	11794.36	65785.58	4.19
2	1137.27	53.33	7041.60	43946.85	6.24

Based on the characteristic above, there are 3 types of customer:



Cheap Product Buyers



Customer who bought **cheap** products



Heavy Product Buyers



Customer who bought **large and heavy** products



Expensive Product Buyers



Customer who bought **expensive** products

Customer Segmentation

Customer Class Proportions



- Cheap Product Buyers (106114)
- Heavy Product Buyers (9720)
- Expensive Product Buyers (1767)

We can also see what **type of product** does each type of customer so that we can create a more **targeted ads** about the **corresponding product** to each type of customer.

	Cheap Product Buyers	Heavy Product Buyers	Expensive Product Buyers
1	Bed, bath, & table	Office furniture	Watches and gifts
2	Health & beauty	Housewares	Computers
3	Sports & leisure	Furniture & decorations	Computer accessories
4	Furniture & decorations	Bed, bath, & table	Health & beauty
5	Computer accessories	Garden tools	Automotive



Conclusions & Recommendations

What Makes a Low or High Review Score?

Based on the Review Score Analysis, these are the aspects that affect the review score given by the customers:

- Order status
- Geographical location
- Freight price
- Delivery difference
- Order processing time



Our Recommendations

- Give title to fast and reliable sellers in order to reduce the number of undelivered orders (cancelled, processing, unavailable)
- Encourage sellers from the north Brazil that hasn't use the Olist Store to use it so that freight price can be minimized
- Partner with a better logistic company to increase delivery performance (reduce delivery delays) and reduce freight price
- Giving a promo about the freight price could also help to increase review score, since people tend to be happy when the freight price is low
- Create a faster and more reliable ordering systems in order to reduce the time needed to process an order

Customer Segmentation

We can divide the customer into 3 categories, which are:

- Cheap product buyers
- Heavy product buyers
- Expensive product buyers



Our Recommendations



Cheap Product Buyers

- Give more ads about the most bought products for this type of customer
- Bundle products to reduce recurring pain points (such that customer can purchase multiple items in one swoop)



Heavy Product Buyers

- Give more ads about the most bought products for this type of customer
- Collaborate with local or international banks to provide zero interest installment
- Partner with logistic company that specifically great at transporting large items



Expensive Product Buyers

- Give more ads about the most bought products for this type of customer
- Collaborate with local or international banks to provide zero interest installment
- Give certain discount or promo whenever the purchase price exceeded certain value

External Sources and References

Dataset :

<https://www.kaggle.com/olistbr/brazilian-ecommerce?datasetId=55151&sortBy=voteCount>

Brazil GeoJson:

<https://www.kaggle.com/thiagobodruk/brazil-geojson>

References:

<https://medium.com/revain/why-are-customer-reviews-so-important-185b915d4e5d>

<https://www.mbaskool.com/business-concepts/marketing-and-strategy-terms/11517-targeting-strategy.html>

<https://www.helpscout.com/blog/cheap-customers/>

<https://blog.hubspot.com/service/customer-retention-rate>



 **Thank You**

