Final year project

Junior Data Engineer Program

In this project, we analyzed a Brazilian e-commerce public dataset of orders made at Olist Store, the largest department store in Brazilian marketplaces.

The dataset has orders from Oct 2016 to Nov 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions:

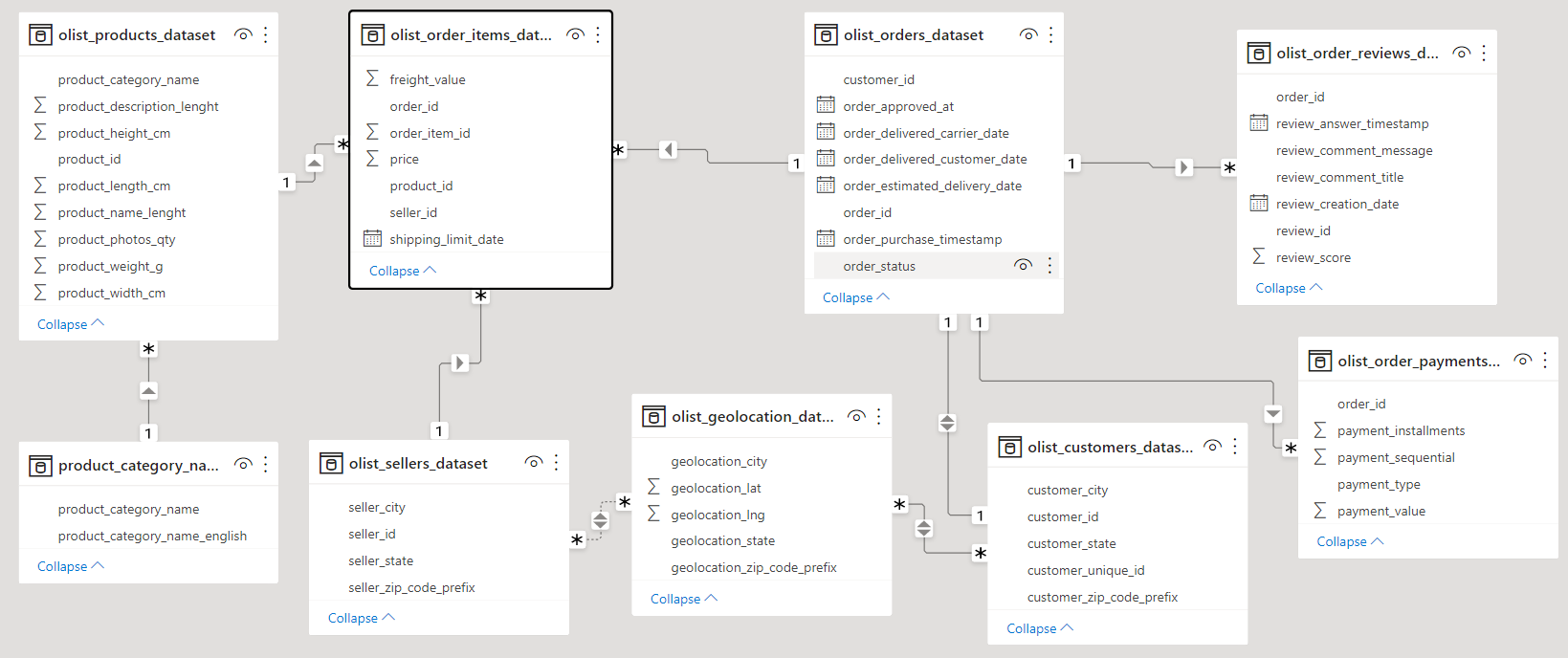
* from order status
* price
* payment and freight performance to customer location
* product attributes and finally reviews written by customers

There was also a geolocation dataset that includes Brazilian zip codes with latitude and longitude. This dataset gave us an insight into the dynamics of an e-commerce industry.

We also performed market segmentation based on the geolocation dataset through using k-means clustering and identified the different customer bases or possible market penetration locations.

We also performed NLP and Sentimental Analysis on Comments/Reviews of Customers and identified the major features affecting the customer satisfaction.

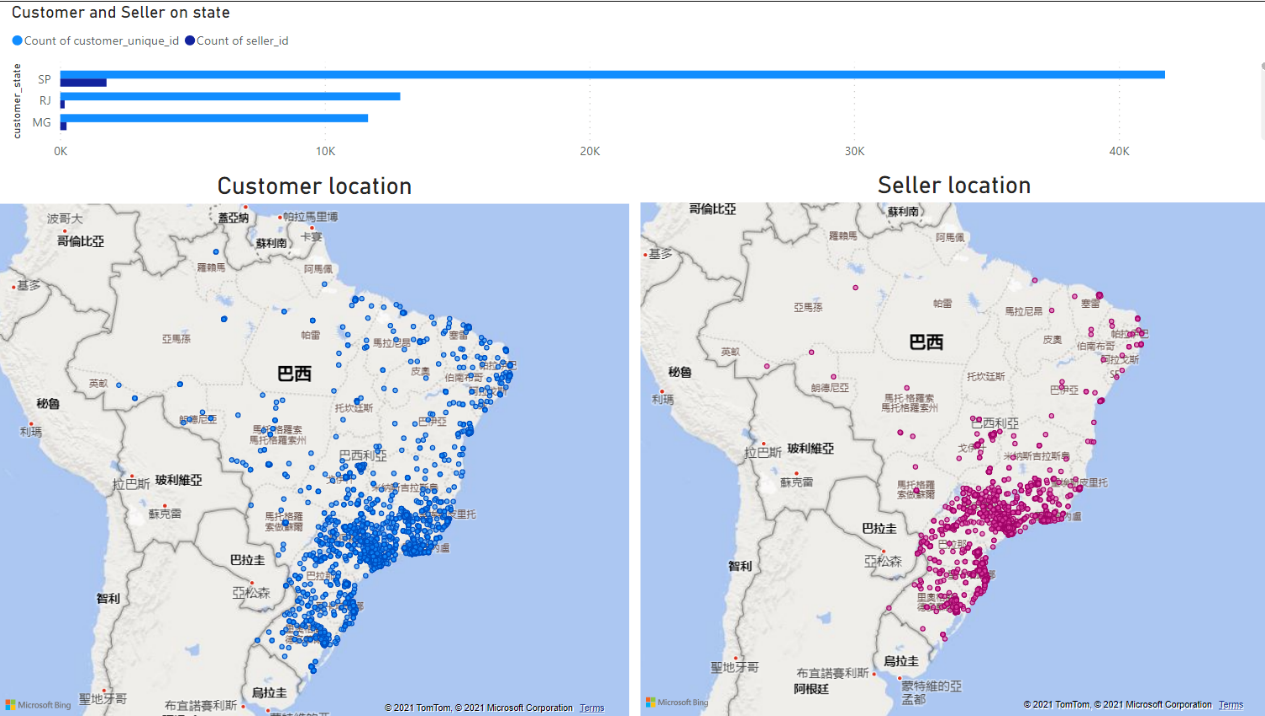
We believe our recommendations will enable Olist Store’s Operational Improvement and ultimately enable higher profits for the organization.



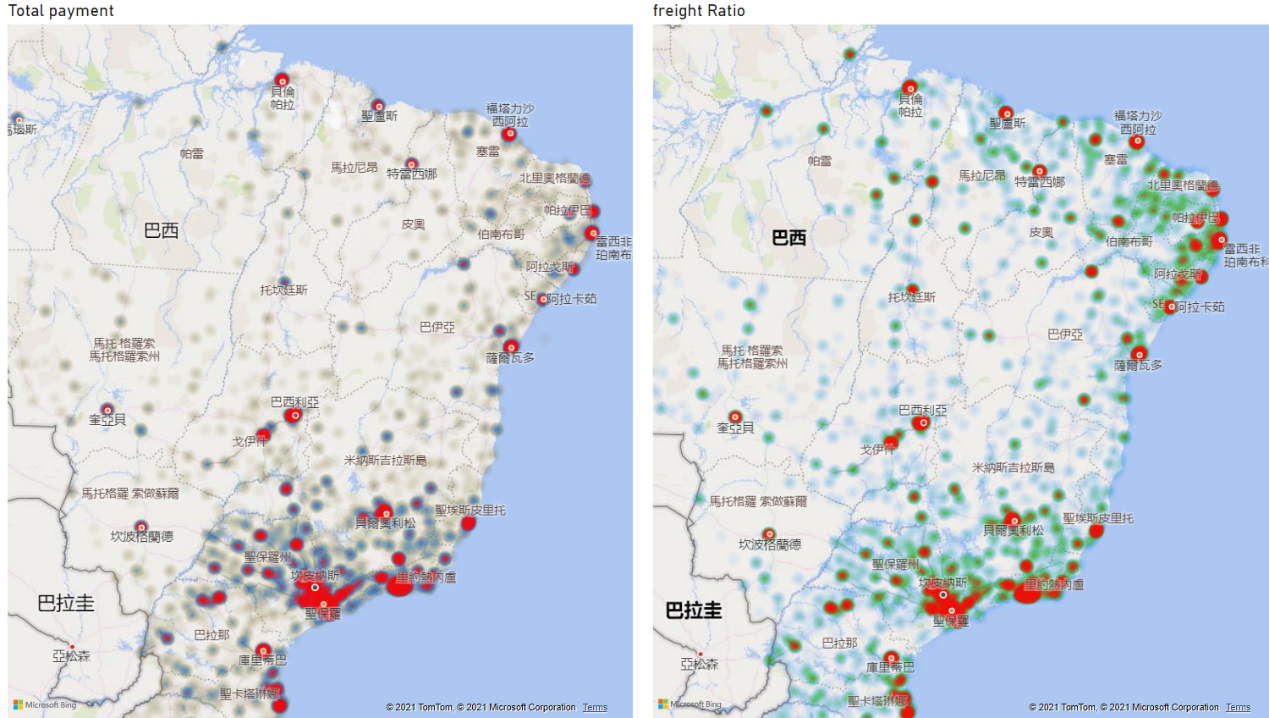
Tasks:

Some of the information or analysis that we extracted from this dataset include:

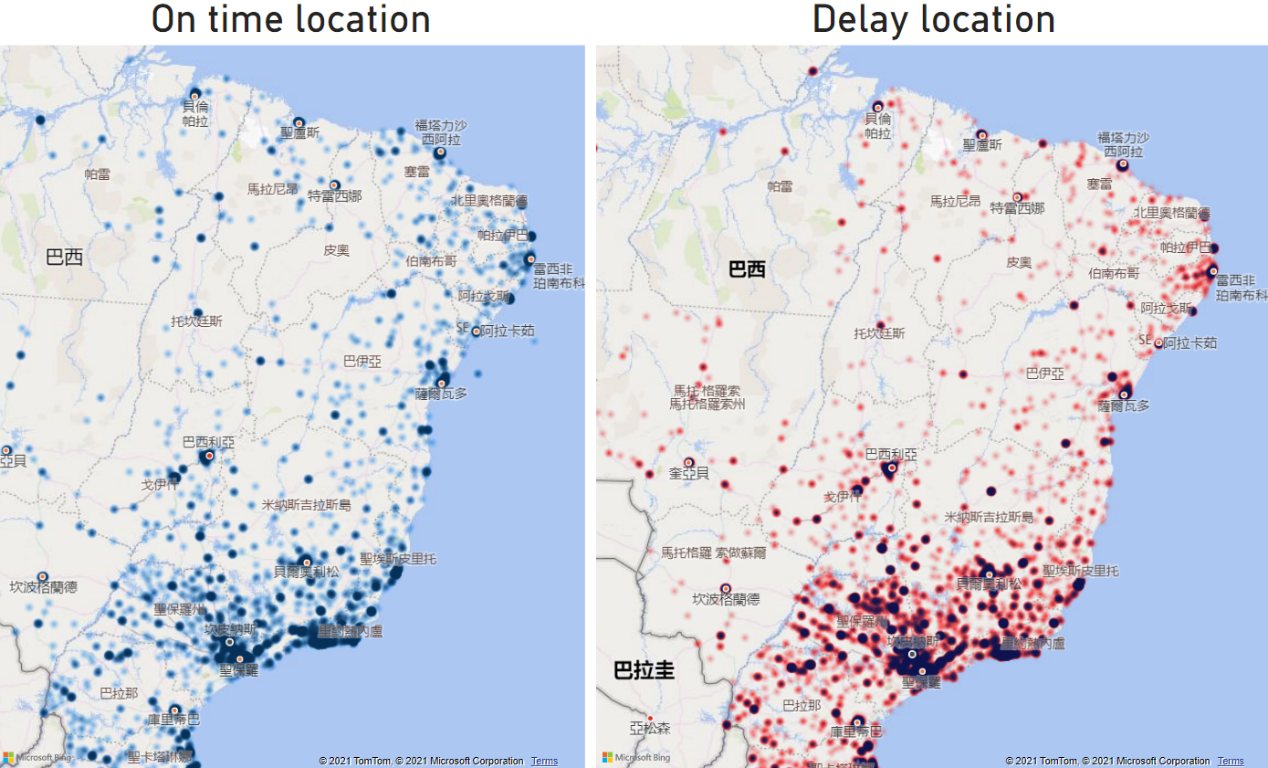
* Clustering: We performed several Clustering on the location of the customers based on various features such as Revenue earned, Freight Ratio and Carrier Delays.

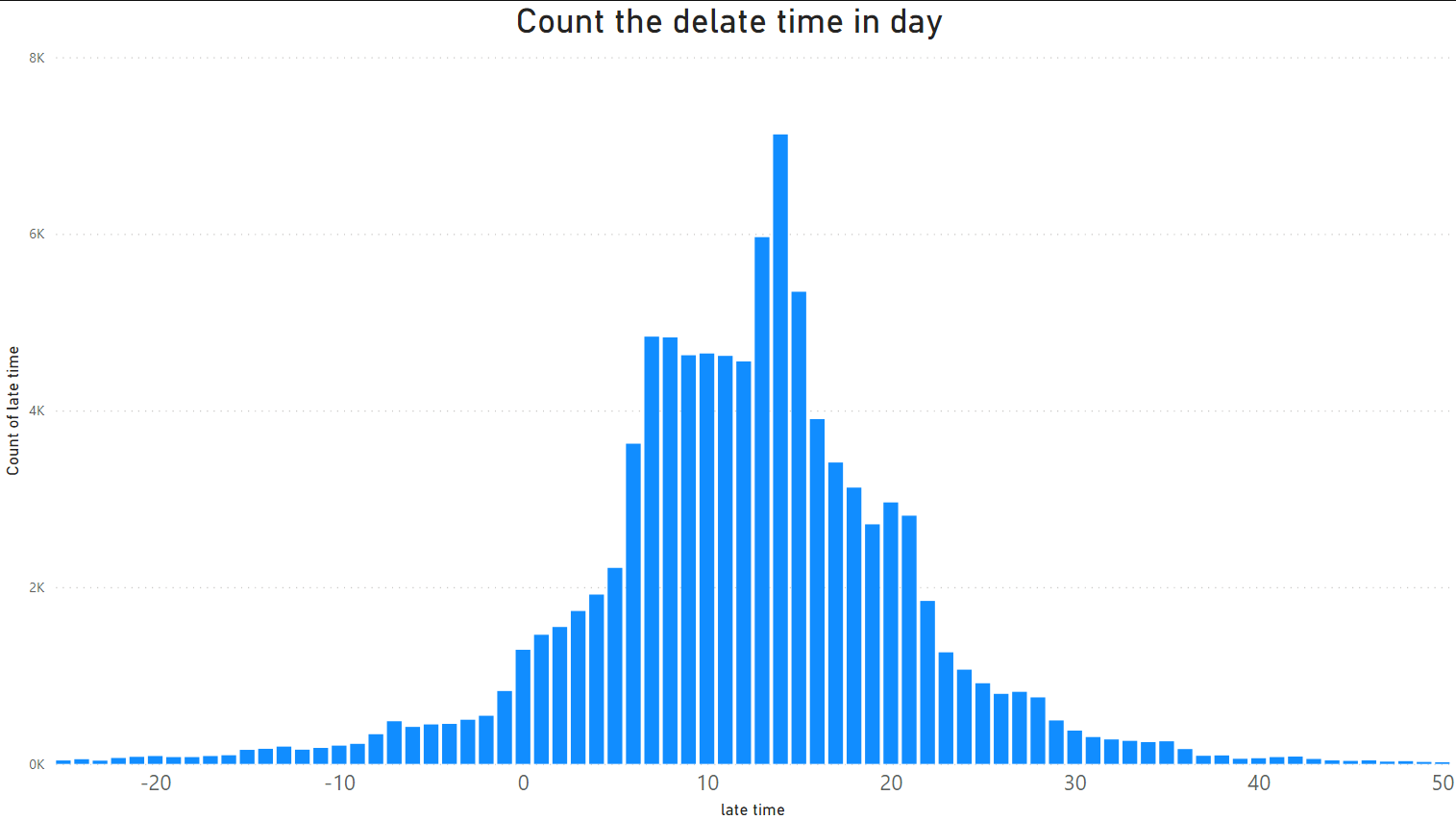


1. Revenue: We noticed that majority of the revenue came from the metropolitan developed areas of Brazil (such as Rio, São Paulo) and there was an equitable distribution in the North & North-eastern areas. We analyzed that there is very good possible market in these areas as the population density in these regions and the economic conditions of these regions were developing at a very fast pace.



1. Freight Ratio: We noticed two interesting observations. First, maximum percent of Freight ratios were seen in northern areas, this was because the sellers are mostly located in southern areas and they charge more for the delivery to the Northern customers. Second, the higher ratio in the metro areas were seen to be because of the expedited shipments, which was observed as the products were delivered to the customer much earlier as compared to the expected delivery date.



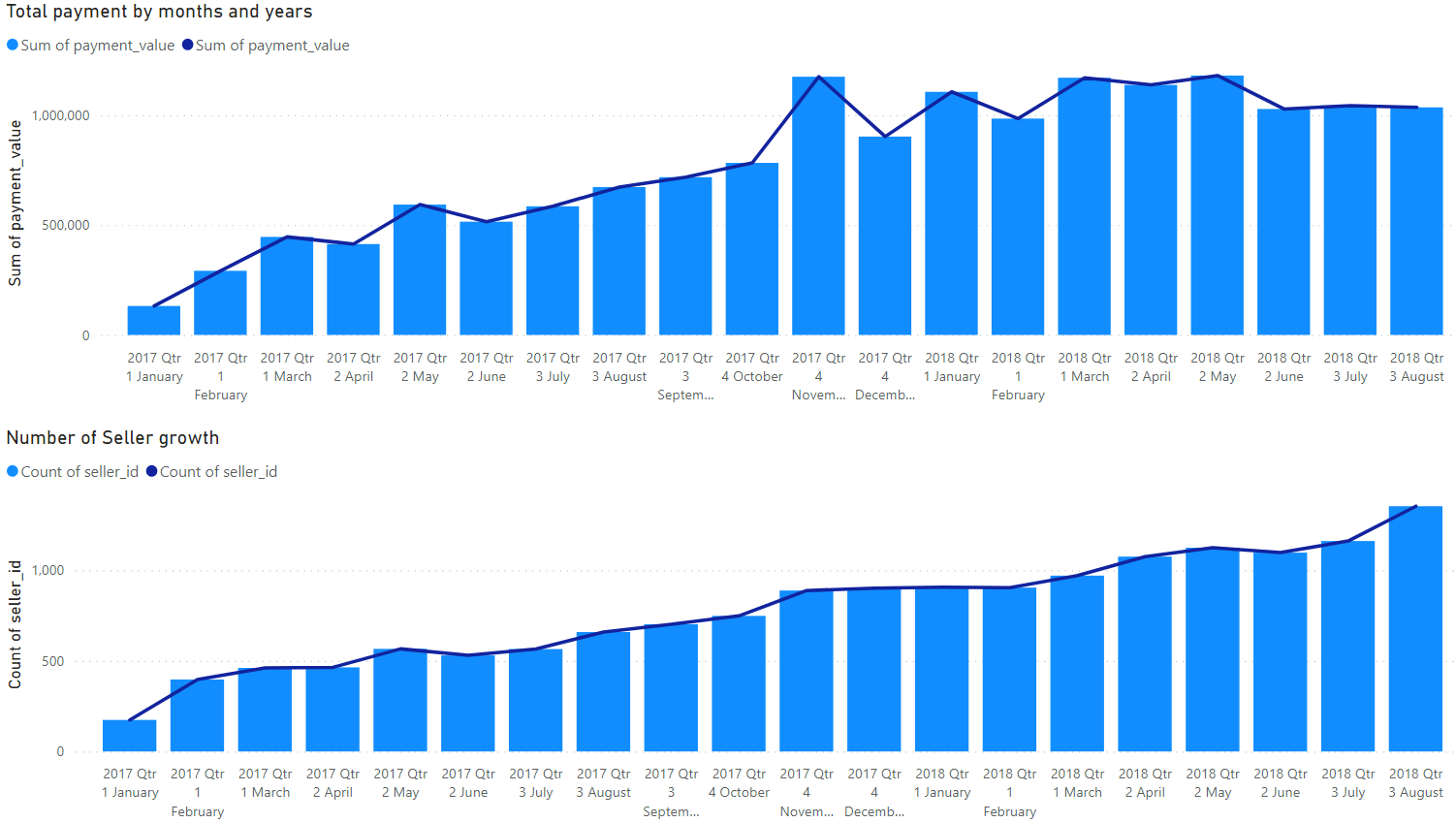


1. Carrier Delays: We observed all of delays was not related to the distance or developed metropolitan regions with infrastructure. Almost all zip code zone has delay situations.

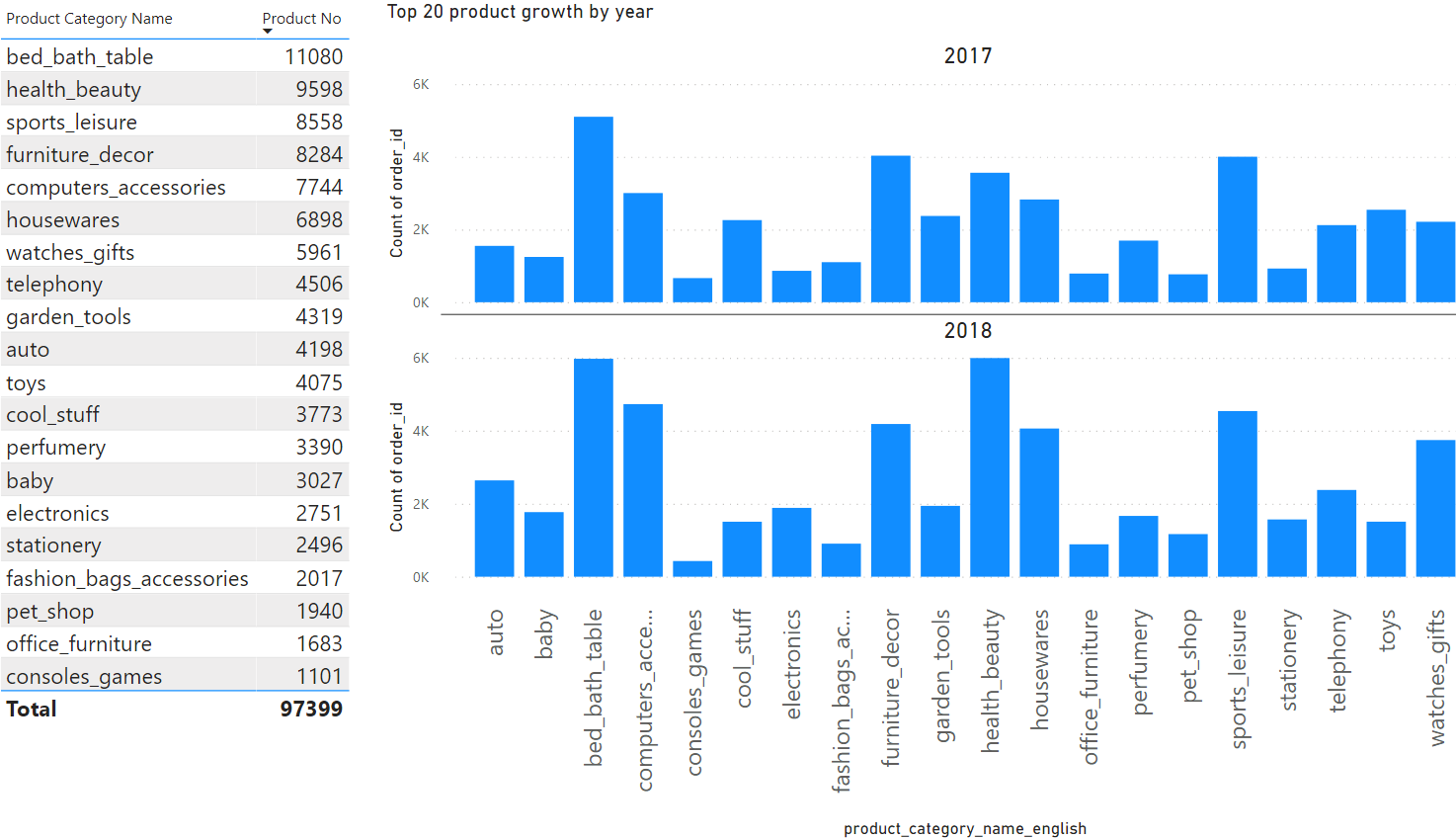
The data shows the range of the day that delay within 1 to 3 weeks. Most of the time, they delay 2 weeks.

These issues need to be addressed by the 3PL companies because the on-time delivery plays an important role in customer satisfaction (as seen from NLP & Classification analysis).

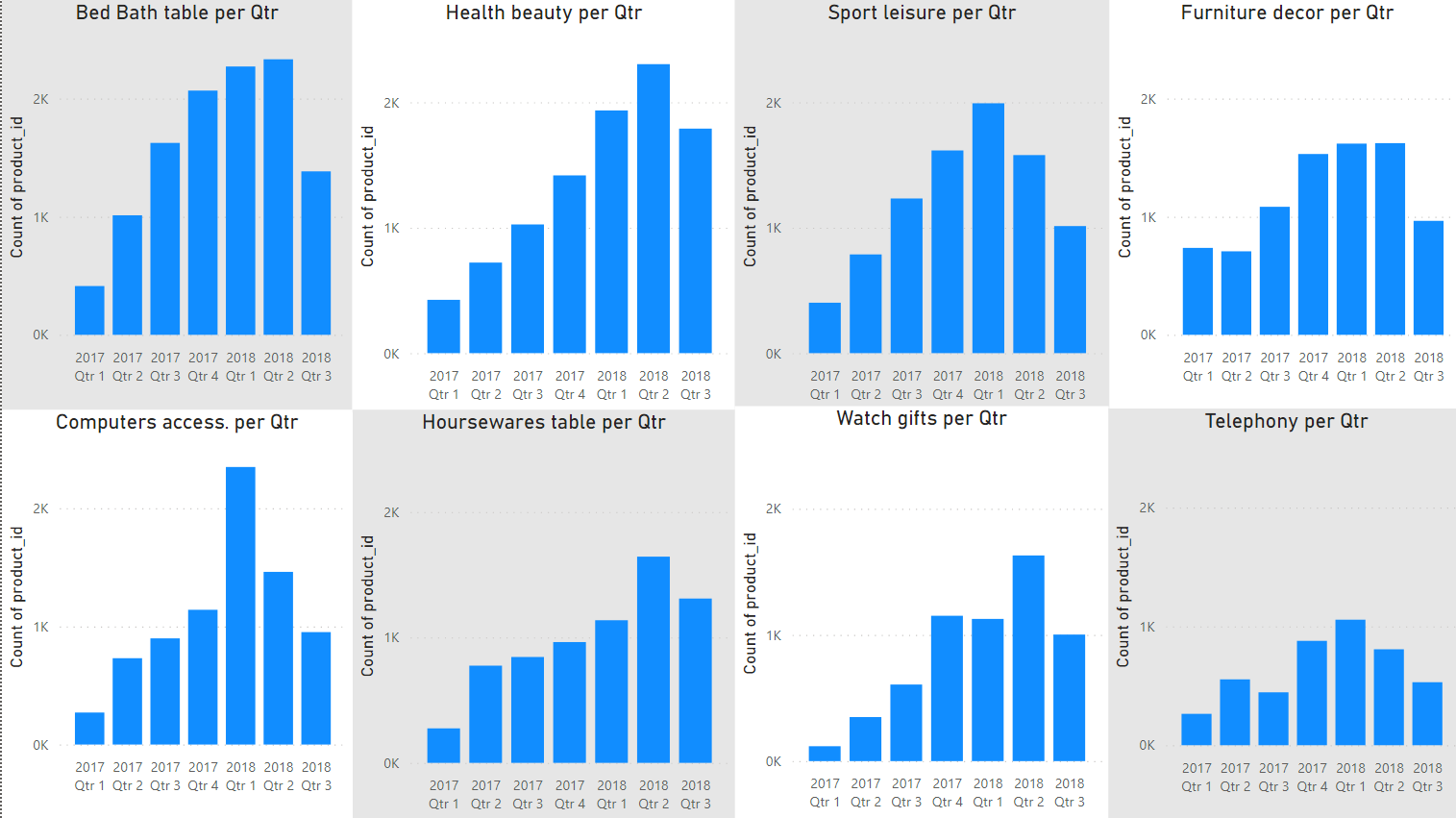
* Sales Analysis: We analyzed the performance of Olist platform from both customer side and seller side using Map-Reduce and tried to give some suggestions to the platform. In this part, we analyzed:



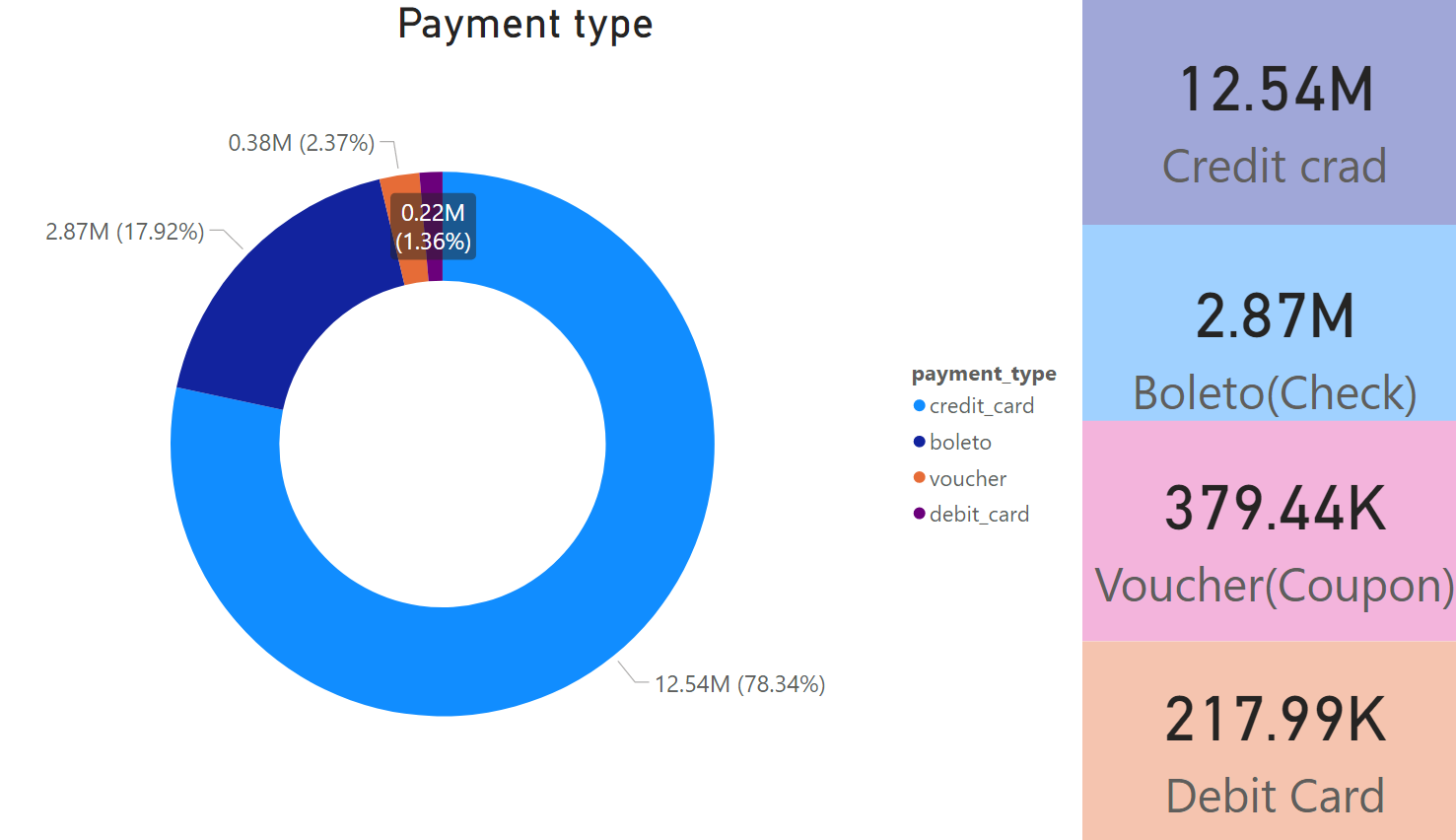
1. The trend of the total sales volume in each month the total sales volume in each month kept increasing before Oct, 2017 while started to stagnate after that.
2. The number of the seller in each month and the number of the seller kept increasing from 2016 to 2018.



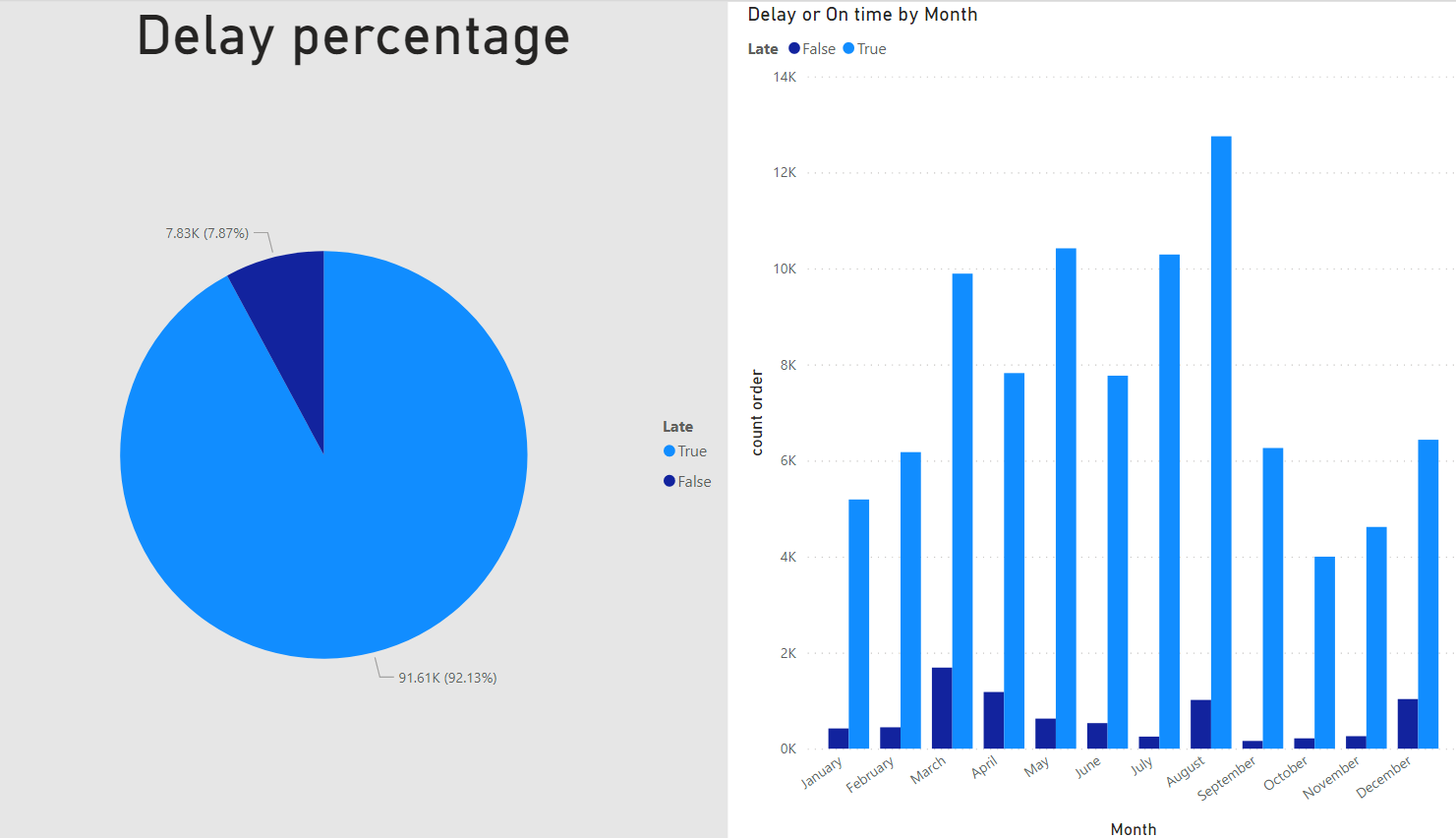
1. New product categories put on the platform in each month There was a boost in 2016, and then the increasing went stable.
2. The most popular product categories in each year the results show that the most popular product categories were Health & Beauty, Bed Bath Table, Computers Accessories, etc.



1. Frequent items bought together Because of data safety, although we found the frequent items, we just analyzed the product id and their categories but could not analyses their product names.



1. The tendency of the payment method We analyzed both payment count (because people sometimes pay by instalments) and payment value show that credit card was used mostly in each month.
2. Dominant payment method used by the customers We also created a new variable, dominant payment method, based on the payment value and we found that credit card was still the most commonly used dominant payment method.



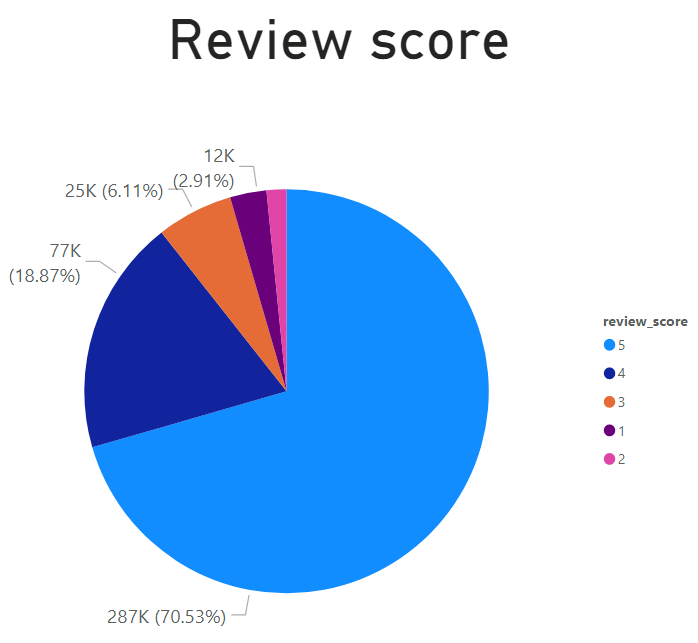
* Delivery Performance Classification:

We worked through on-time delivery and find significant features that can affect delivery times. And, we identified that customer zip codes (customer area), carrier delay, and order approval month are most significant features contributing to delivery times.

Customer zip codes can affect the delivery time because most seller locations are concentrated in metropolitan areas. Out of these areas, customer cannot access the product easily.

And, carrier delay is related with courier companies such as FedEx, USPS after shipping. If there are some issues in delivery services, the delivery can be delayed.

And, order approval month can affect the delivery time because there is a rainy season from October to March in Brazil, it can influence the delivery.



* Review Scores Classification (Multi-class): We built the multi-class classification models to predict the review scores from customers, which means the customer satisfaction and found out that on-time delivery can contribute to the satisfaction of the customer. If customers get their product on-time, they will be happy and satisfied.



* Natural Language Processing: We performed natural language processing for review dataset. After finishing cleaning, the review data, we started NLP works. We extracted top 20 words and identified the important words from the customer review.

The most meaningful words of suggestion:

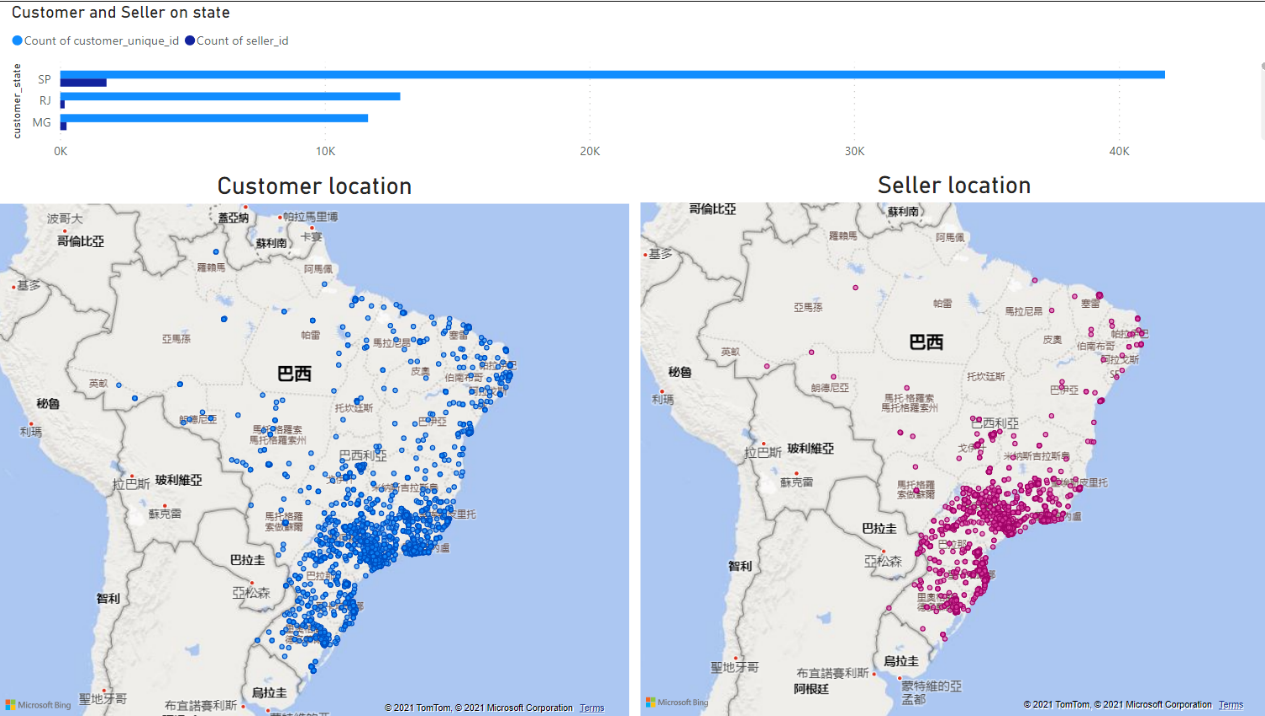
Chegou: has arrived, Entrega: delivery, qualidade: quality

That may mean: customer request the product can deliver faster.

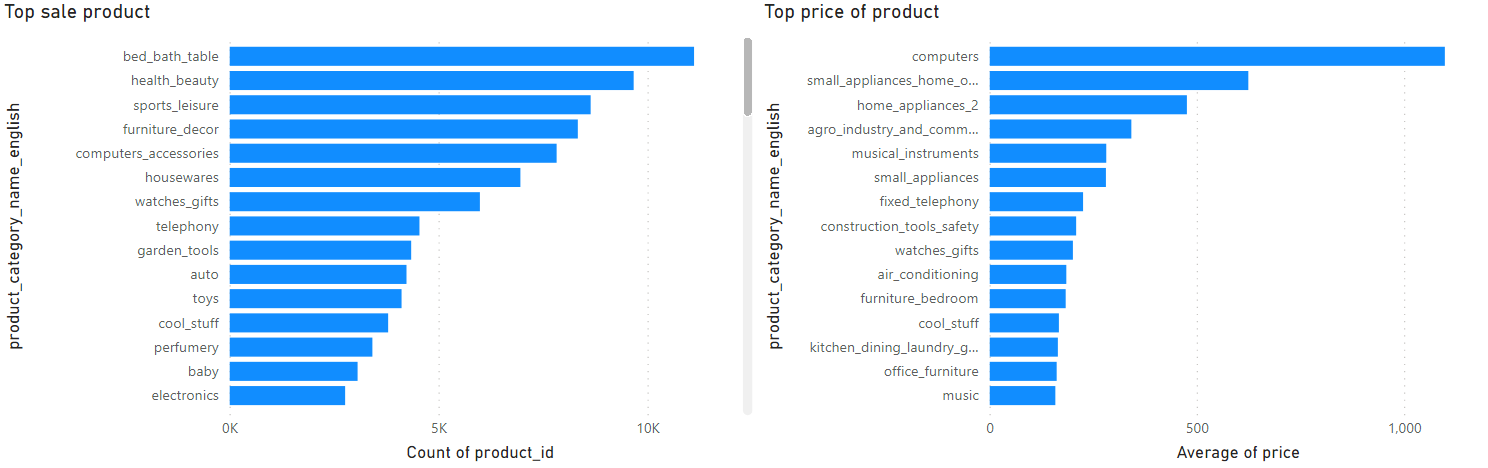
Challenges:

* Data: The dataset was huge and merging 9 datasets to one single master dataset served lots of issues.
* ARIMA Model: Despite having a good amount of data, we had very limited time-series related data points which gave us an Arima model which did not make much importance.
* Only ID: Due to the data security issues, we had the seller/customer related data only in the form of ID. Hence, our market basket analysis didn't have much impact when observed in the form of ID.
* Portuguese Dataset: The whole dataset was in a different language which was hard. We had to download another dataset to translate all of the text data (Headers, Product Categories) to English.

Recommendations:



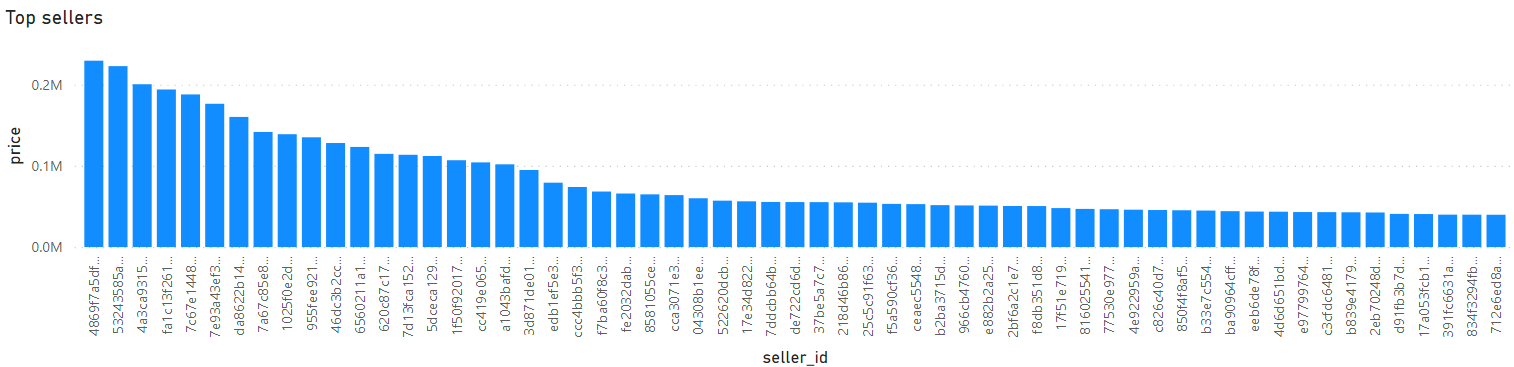
* Sellers: Good market exists in North & North Eastern regions and sellers should make more DC’s or presence there for next further market penetration, South & South Eastern regions, preferably in ZIP starting with 71-92.



* Targeted Marketing: Marketing must penetrate the north eastern market by targeting better coupons to attract new customers.

The highest price product was not the top sell product. So, seller can suggest products accessories.

* Carrier Delays: These must be addressed as delay from 3PL can be handled easily by the company by making the contracts much more stringent. The company can set up cargo warehouse at coastal cities. That reduce the distance to deliver for customer.
* Frequent Item Sets: Olist can use the market basket analysis provided by us for their recommendations to users/customers.
* On-Time Delivery: This is the most important feature which was concluded using both NLP and our classification models and Olist must give extra emphasis on this KIP.



* Sellers rewards: Olist can reward the sellers who can gain the top sales amount, according to the Olist environment.