1_EDA

December 6, 2023

Welcome to Olist dataset

Olist's Business Model

Olist (https://olist.com) is a Brazilian departmental store (marketplace) that operates in e-commerce segment, but is not an e-commerce itself (as she says). It operates as a SaaS (Software as a Service) technology company since 2015. It offers a marketplace solution (of e-commerce segment) to shopkeepers of all sizes (and for most segments) to increase their sales whether they have online presence or not.

Olist says she:

- is a large department store within marketplaces.
- is connected to the main e-commerces of Brazil.
- does not buy products.
- does not keep products in stock.
- does not carry out shipping of any products offered in its store.

All products are sold and shipped by the thousands of shopkeepers (registered on Olist) who sell through Olist. Her strength lies in union of all participating shopkeepers, who are selling physical products. Participant shopkeeper is responsible for separating, packing, and taking products to the logistics operator.

Please note Olist's perspective (a supply chain preview): she prescribes there are many factors that can influence the sales of a shopkeeper e.g. type of product, demand, seasonality, competitive pricing, terms, inventory etc.

1. Overview

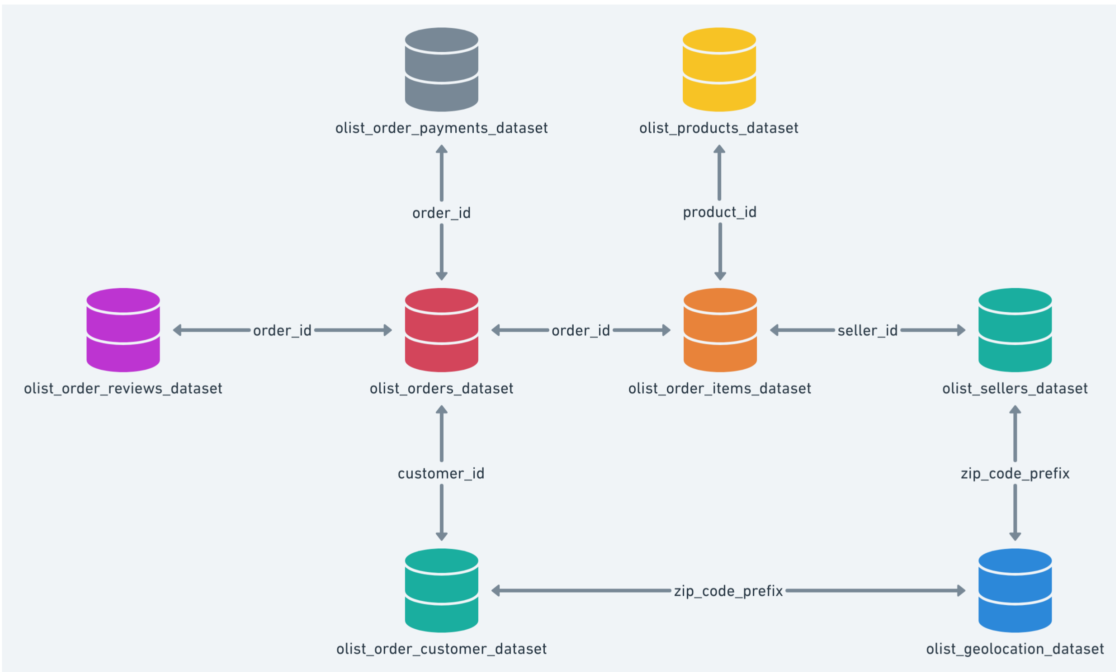
Data source: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data

This is a Brazilian ecommerce public dataset of orders made at Olist. - The dataset has information of 100k orders from 2016 to 2018 made in Brazil.

- Its features allows viewing an order from multiple dimensions:
  - Order status
  - Price
  - Payment
  - Freight value

- Customer location
- Product attributes
- Customer reviews
- Geolocation dataset that relates Brazilian zip codes to lat/long coordinates.

There are some different data sources, each one describing a specific topic related to e-commerce sales. The relationship between these files are described on the schema below.



1.1. Loading data As our dataset is not too big, we could join all datasets (except Geolocation data) to create a single Master data to do our analysis easier. Note that there is a small number of customers who do not make any orders (~0.5%), so that we use `inner join` between these datasets.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 115609 entries, 0 to 115608
Data columns (total 40 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   order_id                       115609 non-null  object
 1   customer_id                    115609 non-null  object
 2   order_status                   115609 non-null  object
 3   order_purchase_timestamp       115609 non-null  object
 4   order_approved_at              115595 non-null  object
 5   order_delivered_carrier_date   114414 non-null  object
 6   order_delivered_customer_date  113209 non-null  object
 7   order_estimated_delivery_date  115609 non-null  object
```

```
 8   customer_unique_id             115609 non-null  object
 9   customer_zip_code_prefix       115609 non-null  int64
 10  customer_city                  115609 non-null  object
 11  customer_state                 115609 non-null  object
 12  order_item_id                  115609 non-null  int64
 13  product_id                     115609 non-null  object
 14  seller_id                      115609 non-null  object
 15  shipping_limit_date            115609 non-null  object
 16  price                          115609 non-null  float64
 17  freight_value                  115609 non-null  float64
 18  payment_sequential             115609 non-null  int64
 19  payment_type                   115609 non-null  object
 20  payment_installments           115609 non-null  int64
 21  payment_value                  115609 non-null  float64
 22  review_id                      115609 non-null  object
 23  review_score                   115609 non-null  int64
 24  review_comment_title           13801 non-null   object
 25  review_comment_message         48906 non-null   object
 26  review_creation_date           115609 non-null  object
 27  review_answer_timestamp        115609 non-null  object
 28  product_category_name          115609 non-null  object
 29  product_name_lenght            115609 non-null  float64
 30  product_description_lenght     115609 non-null  float64
 31  product_photos_qty             115609 non-null  float64
 32  product_weight_g               115608 non-null  float64
 33  product_length_cm              115608 non-null  float64
 34  product_height_cm              115608 non-null  float64
 35  product_width_cm               115608 non-null  float64
 36  seller_zip_code_prefix         115609 non-null  int64
 37  seller_city                    115609 non-null  object
 38  seller_state                   115609 non-null  object
 39  product_category_name_english  115609 non-null  object
dtypes: float64(10), int64(6), object(24)
memory usage: 36.2+ MB
```

1.2. Cleaning data

- There are some datetime columns, which are currently in STRING datatype, it should be converted into DATETIME format for later uses.
- There are some columns contains specific name, these values should be converted into `title` format to look better on the charts
- Adding some more essential columns

2. EDA

2.1. Univariate analysis Exploratory Data Analysis (EDA) could be simple with `pandas.describe()` function. I created a table summarizing the data quality in terms of Null and Unique values. Another option is using pre-built data profiling tool such as `ydata-profiling`.

Profile report structure:

- **Overview** consists of overall statistics. This includes the number of variables (features or columns of the dataframe), Number of observations (rows of dataframe), Missing cells (and percentage), Duplicate rows (and percentage), and Total size in memory. **Alerts** tab is my favorite tab, which contains any type of warnings related to cardinality, correlation with other variables, missing values, zeroes, skewness of the variables, and many others.
- **Variables** gives a detailed information about distribution of all the columns of the dataset. The information presented varies depending upon the data type of variable.
- **Interactions** & **Correlations** represent relationship between each pair of columns (numerical type only) in our dataset.
- Other sections such as **Missing values** or **Sample** are self-explanatory.

Based on univariate analysis above, we could have some general information about olist dataset, which include:

- Number of unique customers: `94.720`
- Number of sellers: `3.090`
- Number of unique orders: `97.916`
- Number of product categories: `73`
- Number of product: `32.789`
- Duration: `09/2016` to `09/2018`

More over, we need to be carefully with alerts:

- `payment_value` is highly overall correlated with `price`: the reason for difference here is that payment is current amount paid by customer, it might be slightly different with the product price (it might be paid totally)
- `order_status` is highly imbalanced (93.4%): most of orders are completed, but there are some orders still incomplete or unsuccessful.
- `delivery_against_estimated` has 2471 (2.1%) missing values and has 1653 (1.4%) zeros
- `review_response_time` is highly skewed: reviews are sent at different times

In the next section, we will focus on **multivariate analysis** to know more about the trend and relationship between multiple columns to understand the Olist's business better.

We will analyse Olist data in terms of Orders, Sales, Rating, Product Category. We will focus on Customer data in a separate notebook.

2.2. Order analysis Looking at the dataset columns, we can see orders with different status and with different timestamp columns like purchase, approved, delivered and estimated delivery, represent different states of order process. I will use `order_purchase_timestamp` for order time.

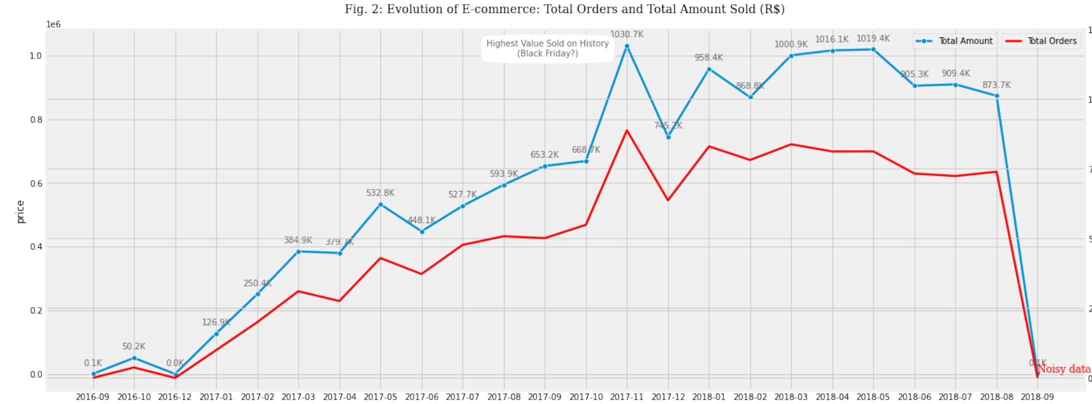
As we mentioned in section 2.1, there is a small difference between `payment_value` and `price`. In this case, I will use `price` for calculating Revenue/Sales.

a. How many orders we have for each status?

Fig. 1: Number Of Orders for each status

b. Orders through time

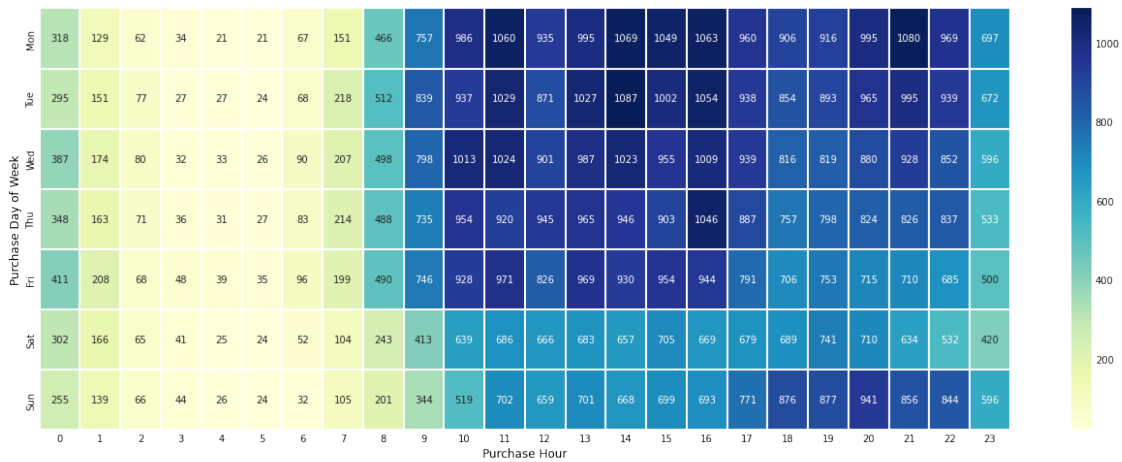
Fig. 2: Evolution of E-commerce: Total Orders and Total Amount Sold (R$)



E-commerce on Brazil really has a growing trend along the time. There is a strong correlation between Revenue and Number of Orders. However, there are some months these metrics change slightly in two different directions such as June-to-August (2018).

c. Orders through days & hours

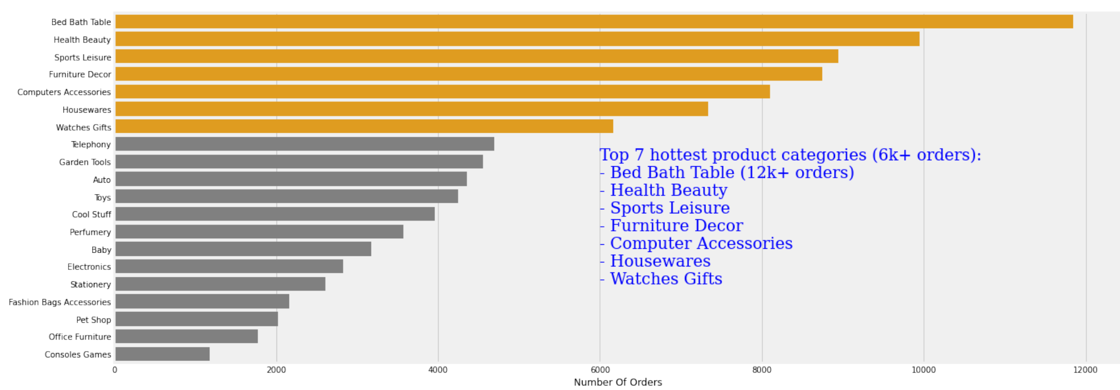
Fig. 3: Orders through days and hours



From the heatmap above, we find that the number of orders decrease gradually from weekdays to weekends. Weekdays, especially on Monday & Tuesday are the prefered days for Brazilian's customers and they tend to buy more at the afternoons. The hottest time slots are Weekdays (10-22h) and Weekends (18-22h).

d. Orders by product categories

Fig. 4: Hottest product categories



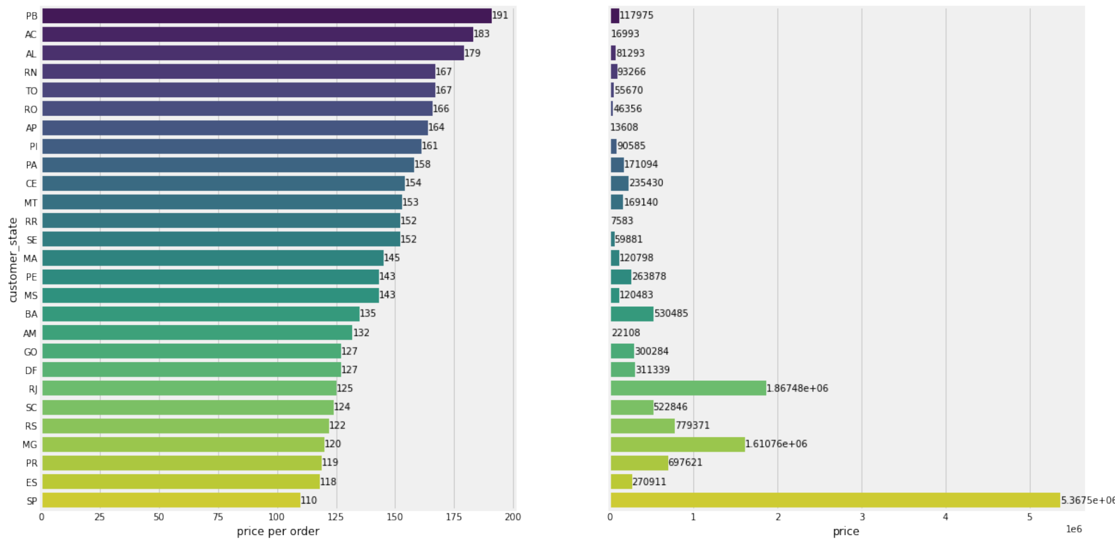
I will dive more deeply into details to know about the trend of product categories through time. Because there are many different product categories, dynamic chart in `plotly` will show better insight in this case.

We could find that `Bed Bath Table` is the best-seller product category of all time. Products in `Furniture Decor` are quite "hot" before 2018, while `Health Beauty` and `Housewares` products are more popular from April to the end of 2018. `Computers Accessories` products are quite seasonal and only hot in the early part of 2018.

2.3. Sales analysis Now, we'll analyze ecommerce cash flow by looking at order prices, shipping rates, and more.

a. Sales by state

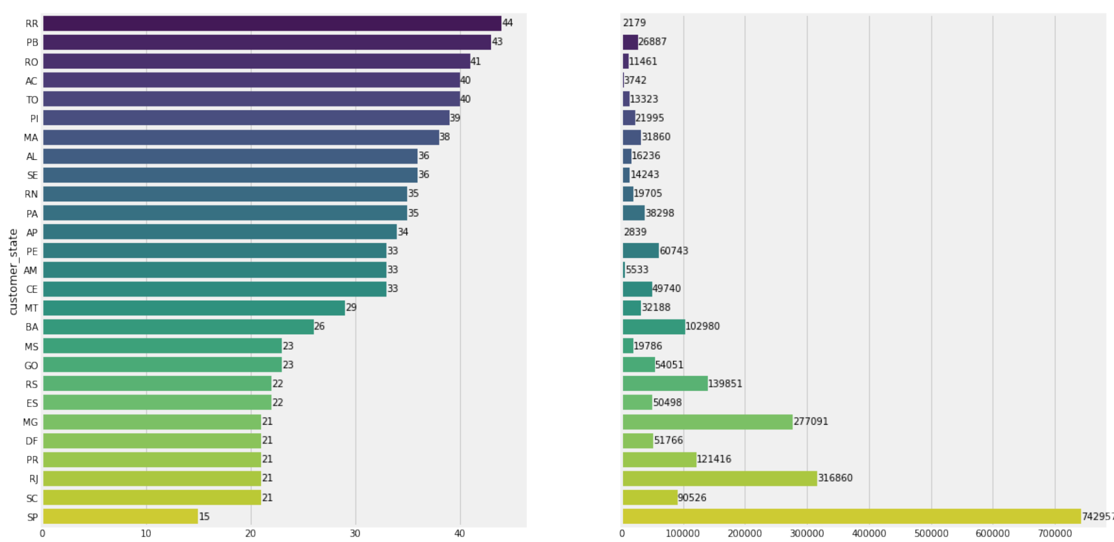
Fig. 6: Sales by state



Some states have a high total amount sold and a low price per order. For example, SP (São Paulo) is the most valuable state for e-commerce (5M+ sold) but also where customers pay less per order (110 per order).

b. Freight value by time

Fig. 7: Freight value by state



We could see that customers in Roraima (RR), Paraíba (PB), Rondônia (RO) and Acre (AC) normaly pays more than anyone on freights.

c. What are the best states to buy in Brazil?

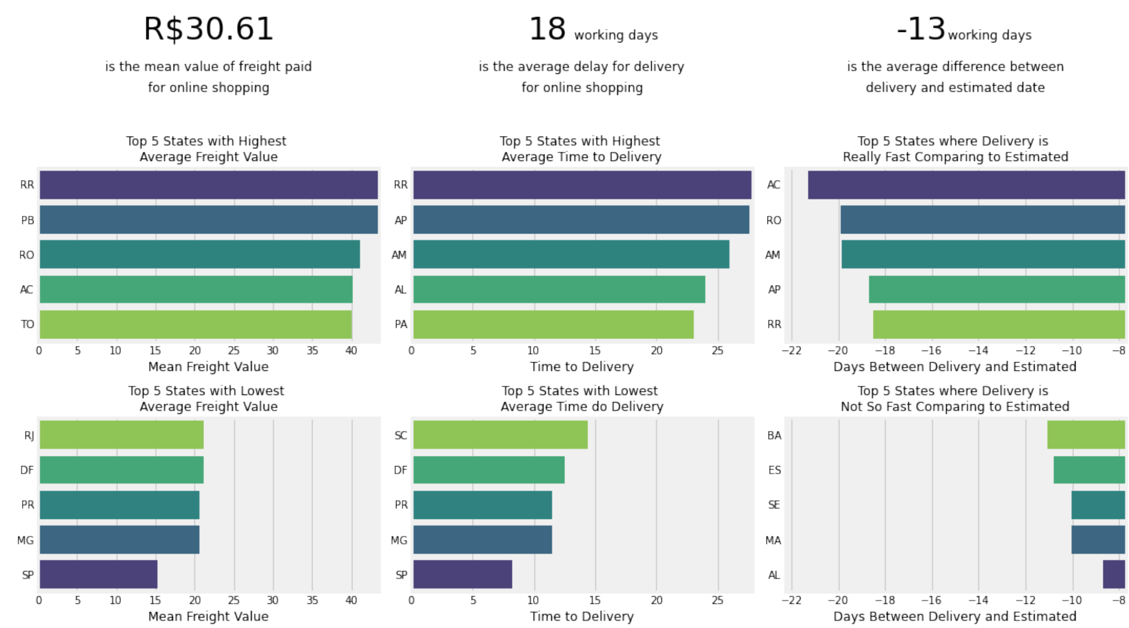


Fig. 8: Comparative Study: E-Commerce on Brazilian States

It looks like delivery system in `SP`, `MG`, `PR` and `DF` states is very good, where customers can place orders at the lowest price shipping rates & receive orders very quickly (within 2 weeks in average).

2.4. Rating Analysis

a. Rating by time

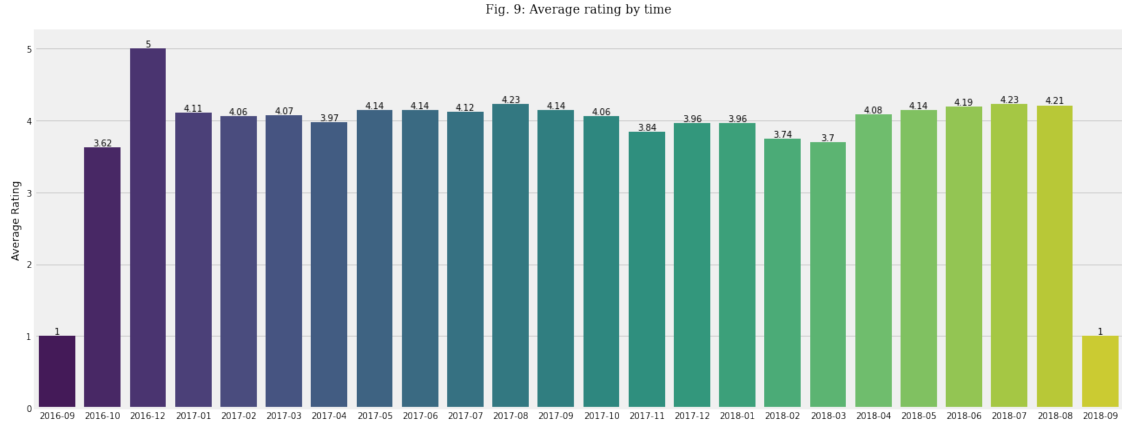


Fig. 9: Average rating by time

If we ignore data before 2017 and 09/2018 because the number of orders is too small (and rating will be biased), we could find that rating for orders are quite stable `between 3.95 and 4.2`. The rating drop in Nov-2017 and Feb, Mar-2018.

b. Rating by categories

Among the top 7 best-selling products (also most reviews), `Bed Bad Table` has the lowest rating (3.89). The highest rating (>4.1) belongs to the 2nd and 3rd best-selling product groups (`Health Beauty` and `Sports Leisure`).

2.5. Payment Type Analysis We can build a mini-dashboard with main concepts: payments type and payments installments, which aims to present enough information to clarify how e-commerce buyers usually prefer to pay orders.

Fig. 11: A Distribution of Payment Installments

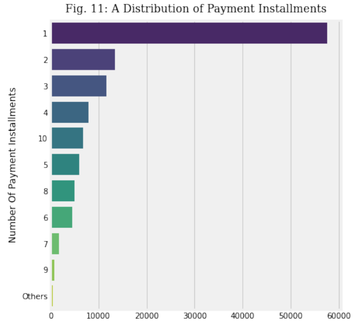
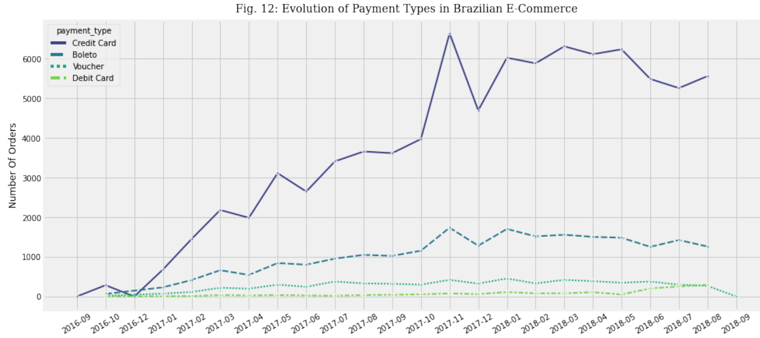


Fig. 12: Evolution of Payment Types in Brazilian E-Commerce



We can see that payments made by `credit card` really took marjority place on Brazilian e-commerce. Since Mar-2018 it's possible to see a little decrease on this type of payment. On the other side, payments made by `debit card` is showing a growing trend since May-2018, which is a good opportunity for investor to improve services for payments like this.

On the bar chart above, we can see how Brazilian customers prefer to pay the orders: mostly of them pay once into 1 installment and it's worth to point out the quantity of payments done by 10 installments.