Olist is an online marketplace where small businesses can display their products and the platform is responsable to show products adds in huge online stores in Brazil, such as: Amazon, Mercado Livre, Americanas and Casas Bahia. It works like ebay and google ads combined:

- This process optimizes the online presence of a store and increases its sales.
- Olist enhances the results faster and effortless with a low cost.
- Offers an app to build the online store.
- They offer a logistic operation for the business to pick up, transport and deliver to the customer.
- They create reports of the e-commerce scenario for each store. This reports can be divided by city, region and trends for the next year.

My objective with this project is to **perform exploratory analysis** and obtain **insights** about customers, company's growth and reviews.

Therefore, I created a interactive dashboard of the business, thinking of Olist's executive board in order to obtain an overview of the business myself.

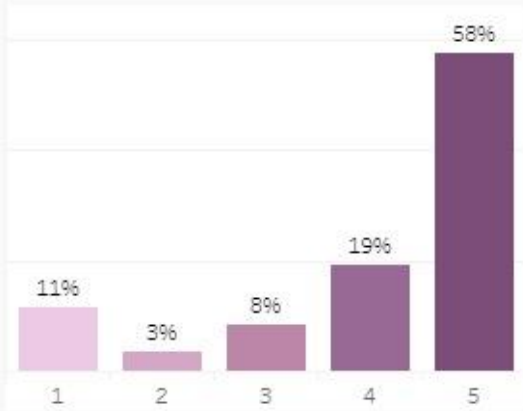
This dashboard was crucial to understand the evolution of Olist's operations and see the concentration of sellers, customers, it's division by categories, regional differences, etc.

Olist - Interactive Dashboard

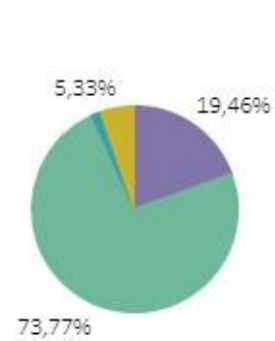
Orders and Revenue per Month

Orders per Year

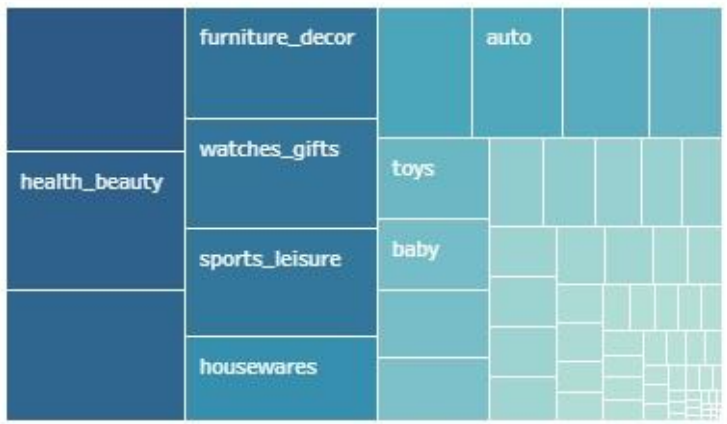
Review Score



Payment type



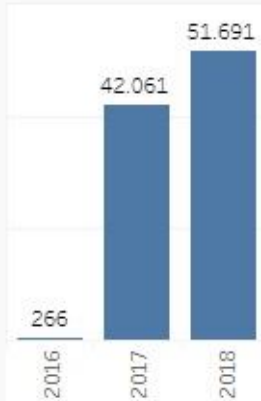
Categories



Sellers

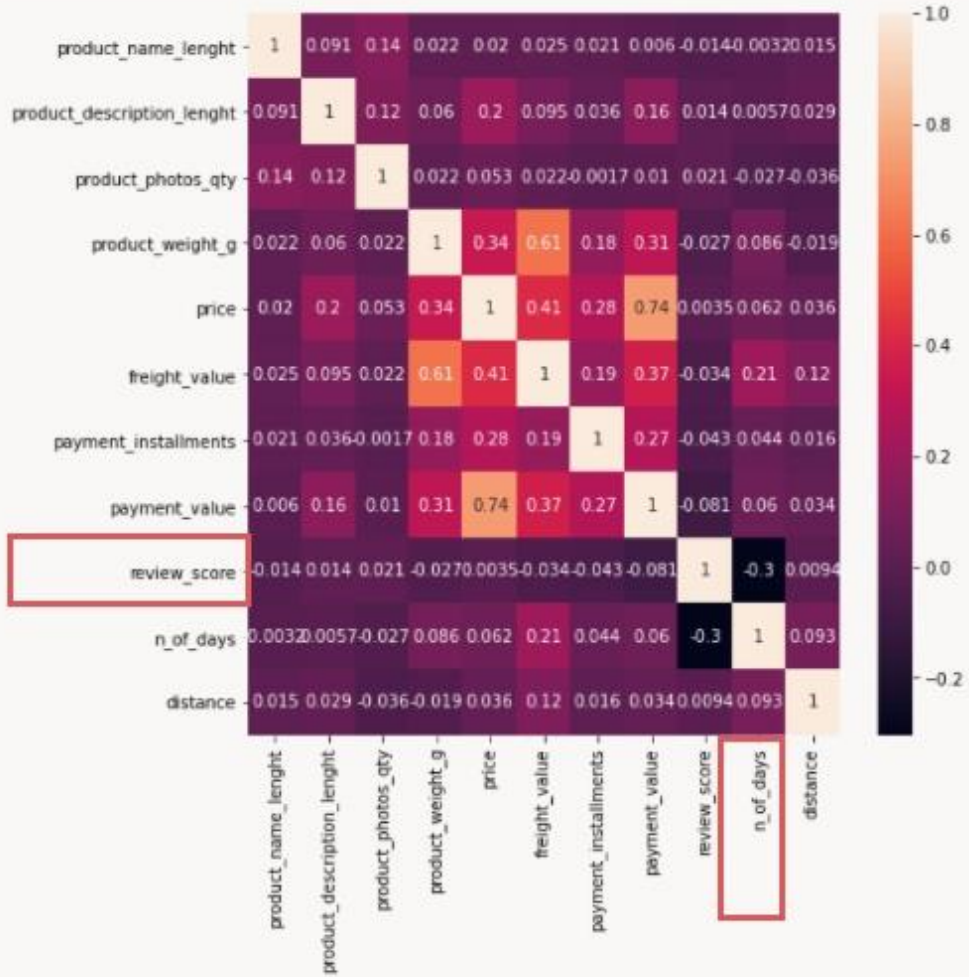
Select a State (Tudo)

Customers



Correlation Matrix

- The most significant correlation is a weak negative correlation between n_of_days to process the order and review_score.
- The other strong correlations are between product weight, price, freight value, payment value (price of product + freight). Which are clear relations and don't need further explanation.



Analysing Linear Regression

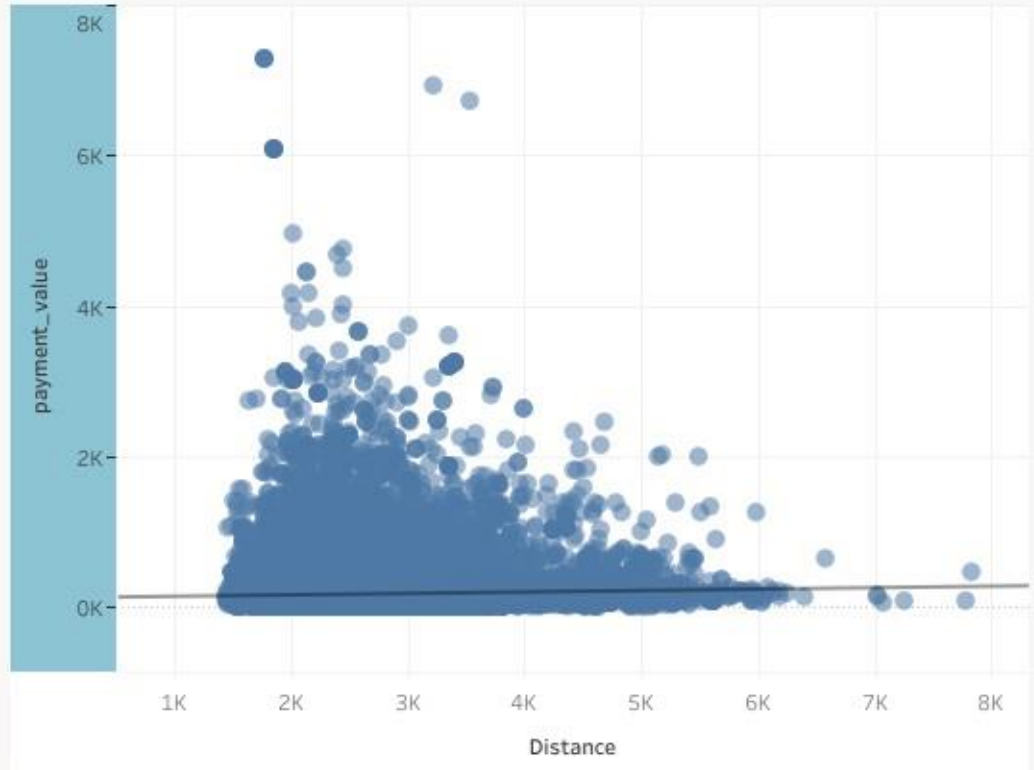
My initial though was to describe deeper if there is a relationship between the variables n_of_days to process an order and the review_score.
Since review score is a categoric variable, I had to create a dummy dataset in python to simulate the data and I got the following results:

Mean squared error: 1.5967
R2 score: 0.1154

The mean squared error is too high, this means that the linear regression may not be the best model to predict the review_score and there are more variables that could influence the results.

Analysing the variables: Distance and Payment value

Let's investigate this variables further:

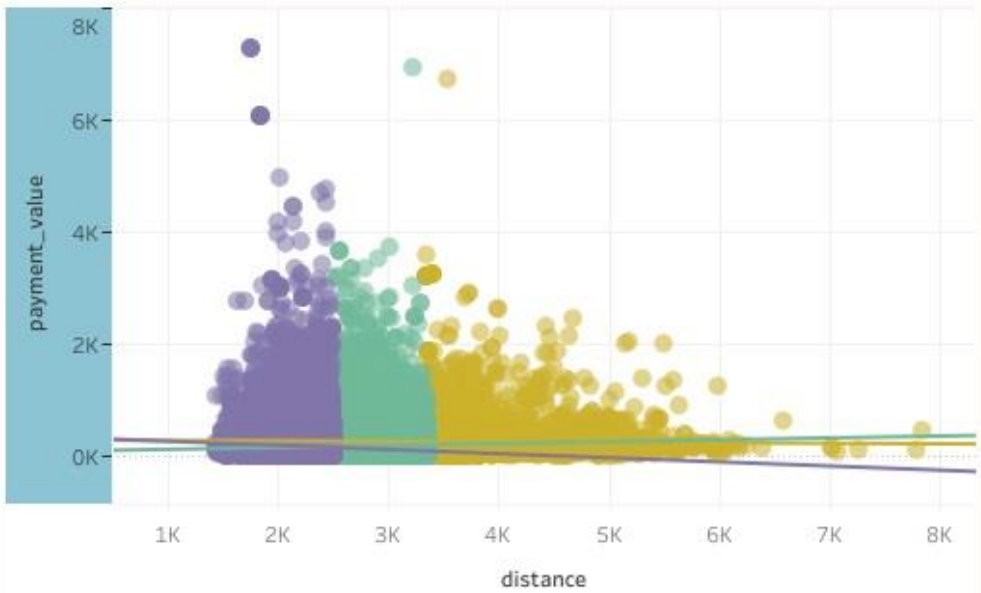


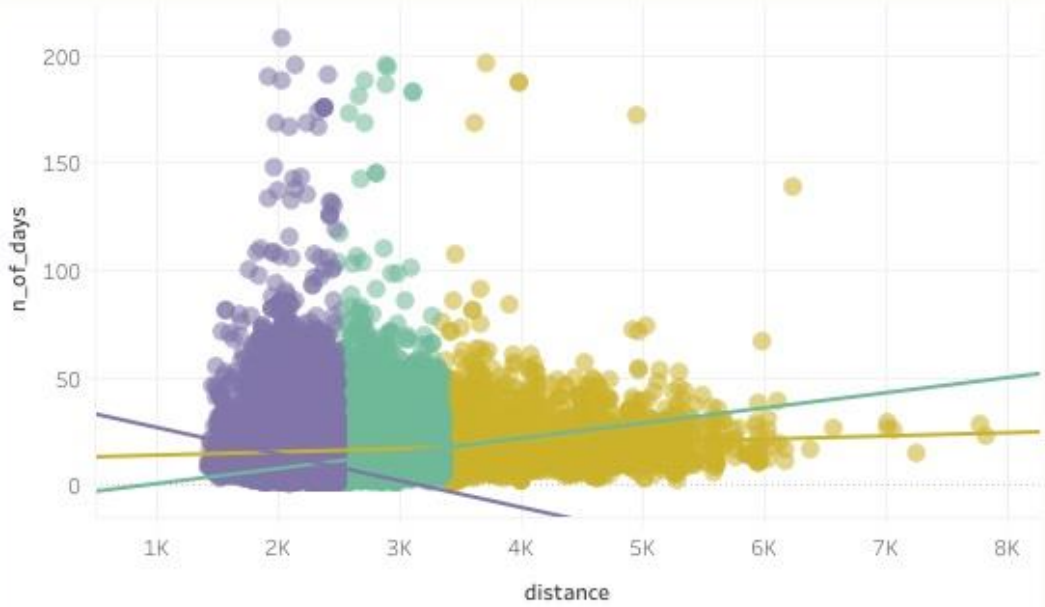
The trend line shows that the higher the distance, the payment value increases slightly.

Grouping in Clusters

Clusters were made to check patterns in the dataset. Using distance and payment_value, I expected that the higher the distance to deliver, more the customer have to pay, but actually, with three clusters the results were different:

For clusters 1 and 3, the higher the distance, lower is the total value.
For cluster 2, the higher distance, higher is the total value.
All the trend lines present a p < 0,05, which makes the results significant.



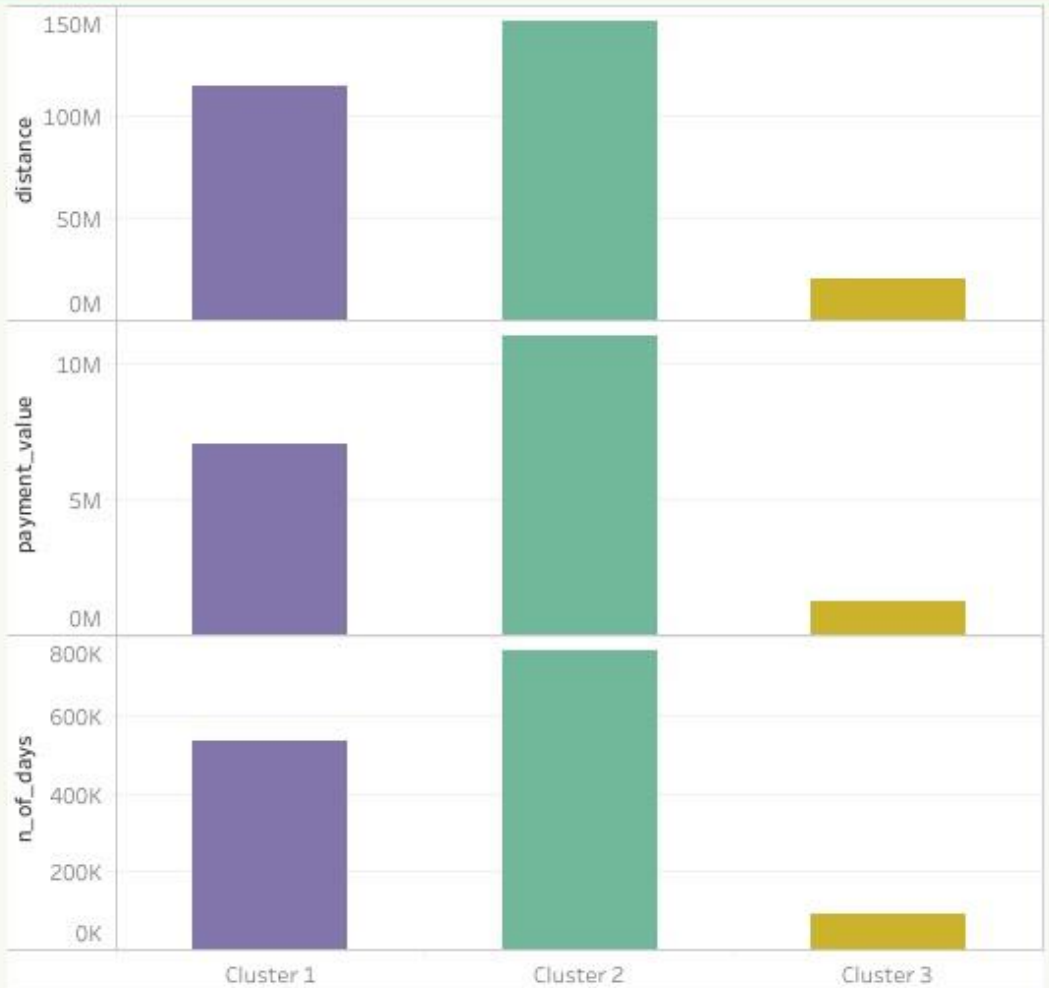


In this cluster analysis, I decided to check the variables n_of_days to process the order and the distance between seller and customer.

For clusters 2 and 3, the higher the distance, the longer it takes to process the order.
For cluster 1, the higher the distance, the lower are the number of days to process, but since the standart deviation is really high, it means that are more variables that influence the number of days to process the order.

All trend lines present a p<0,05, which makes the results significant.

Cluster Analysis



The clusters were interesting to analyse, because of cluster's 2 differences.

Clusters 1: the longer the distance, more expensive is the order.
Cluster 3: the shorter the distance, cheaper is the order.

Cluster 2: longer the distance, the cheaper is the average order's price.

Which makes me wonder: is it a matter of regional logistic costs or lower product's prices?

I checked the mean and median value of review_score is 5 in all clusters, which is the best score a seller can obtain.

Cluster	1	2	3
avg_distance	3652	2788	2282
avg_payment	354	170	166
avg_n_of_days	18	12	9

Final Thoughts & Recommendations

Olist was founded in 2015 and has a focus to help small and medium business to prospect and find clients all over Brazil.

Although their operations are new and the dataset only offered data until July of 2018, it is visible that its presence in brazilian market was not only needed but necessary, so these businesses could thrive, specially in a pandemic scenario.

Recommendations:

- Keep track of reasons for low review score;
- Check the products, categories and regions where the logistic costs are lower;
- Adopt another method for analysing the relationship between review_score and n_of_days to process the order;
- Gather more data to trace seasonality;
- Check how many of the sellers and customers are active and not only registered.