Olist Porject

Data Analysis / Data Science

Project Purpose

Demonstrate applied knowledge of Data Analysis and Data Science

Data source

In this prooject we will use an open database published on the kaggle platform.
data base link

Olist overview

- Founded in 2015
- Operates in Brazil
- 100% digital company
- Online e-commerce service for sellers
- Connects merchants to marketplaces
- Offers logistic and inventory management services to sellers

Olist overview

"What is **Olist** and how does it work?
Olist is a large store made up of more than 25 thousand professional retailers, who act as distribution centers throughout the country.

In practice, we work like this: for each sale made, the partner retailer who has the product in stock receives the order, issues the electronic invoice and prepares the packaging to send it to the consumer safely and quickly.

All **Olist** retailers go through a rigorous evaluation process and must meet a series of requirements. Thus, we guarantee that it will offer our consumers an excellent quality service."

Sorce: olist web site - link

Data Assumptions

Revenues

Variable revenue
Olist charges a 10% cut on each product sold by a seller, if the order was delivered.

Fixed revenue
Sellers have to pay Olist 80 BRL per month to use the platform

Costs

IT cost
IT Cost scale with the square-root of number of orders processed, we assume that IT cost grow relatively smaller and smaller with the number of ordered processed (economies of scale).

Reputation Cost
Bad reviews affect on Olist's reputation and long term profits.

- Annual cost over the average annual score:
  - 1 -> R$ 100
  - 2 -> R$ 50
  - 3 -> R$ 40
  - 4 and 5 -> R$0

Data Assumptions

IT cost

estimated IT cost

IT Cost scale with the square-root of number of orders processed, we assume that IT cost grow relatively smaller and smaller with the number of ordered processed (economies of scale).

So we are gonna use the following formula to calculate the IT cost:

$$\text{Cost per order} = \frac{47.8}{\sqrt{\text{Order Count}}} + 2.2$$

Project Stages

Data Structure

Organize database structure and create new views to help the query prosses

Primary Analysis

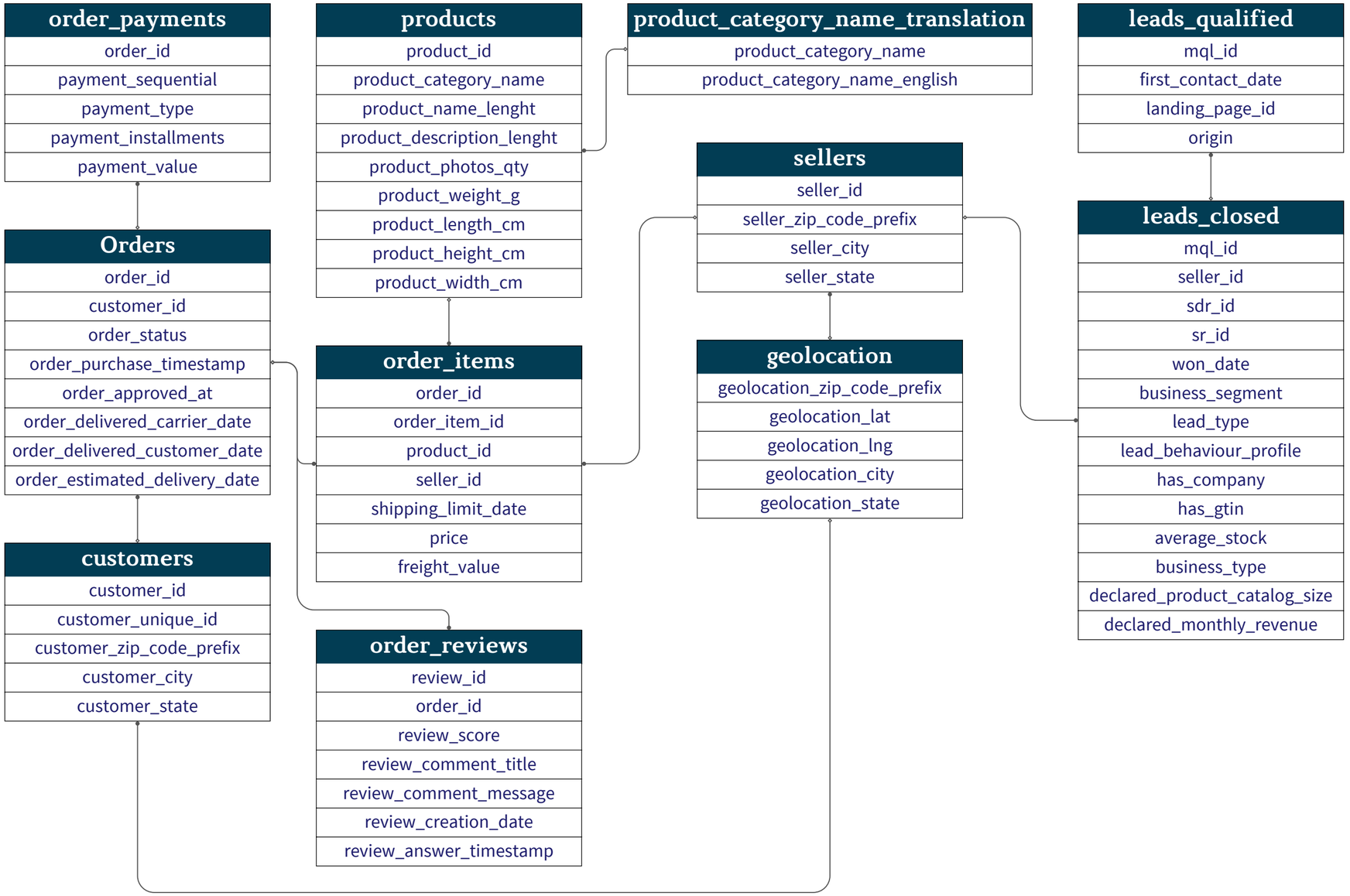
Make a primary analysis of all main KPIs and understand the best round map to follow

Going Further

Take the main opportunity of improvement on customer perspective and understand how we can improve

Data Structure

Database - ER Diagram

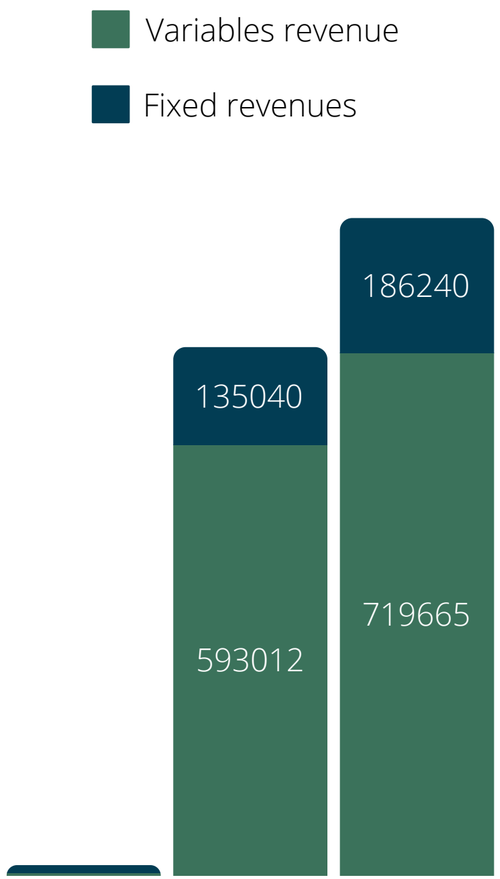


Primary Analysis

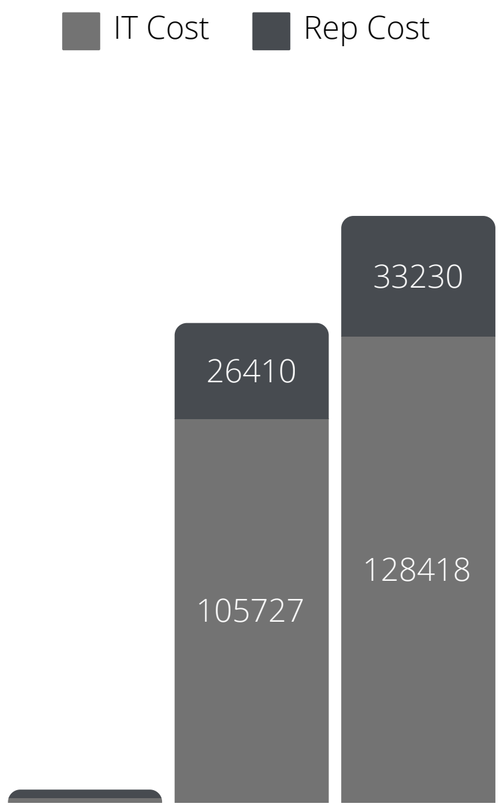
Financial Overview

Financial Overview

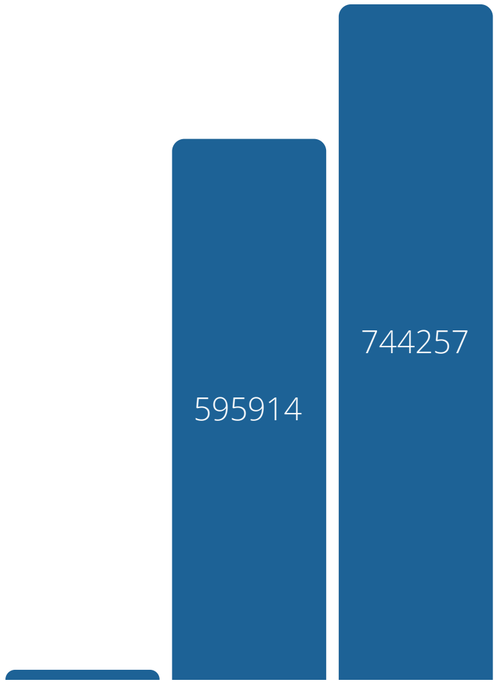
Revenue - 2016 to 2018



Cost - 2016 to 2018



Result - 2016 to 2018

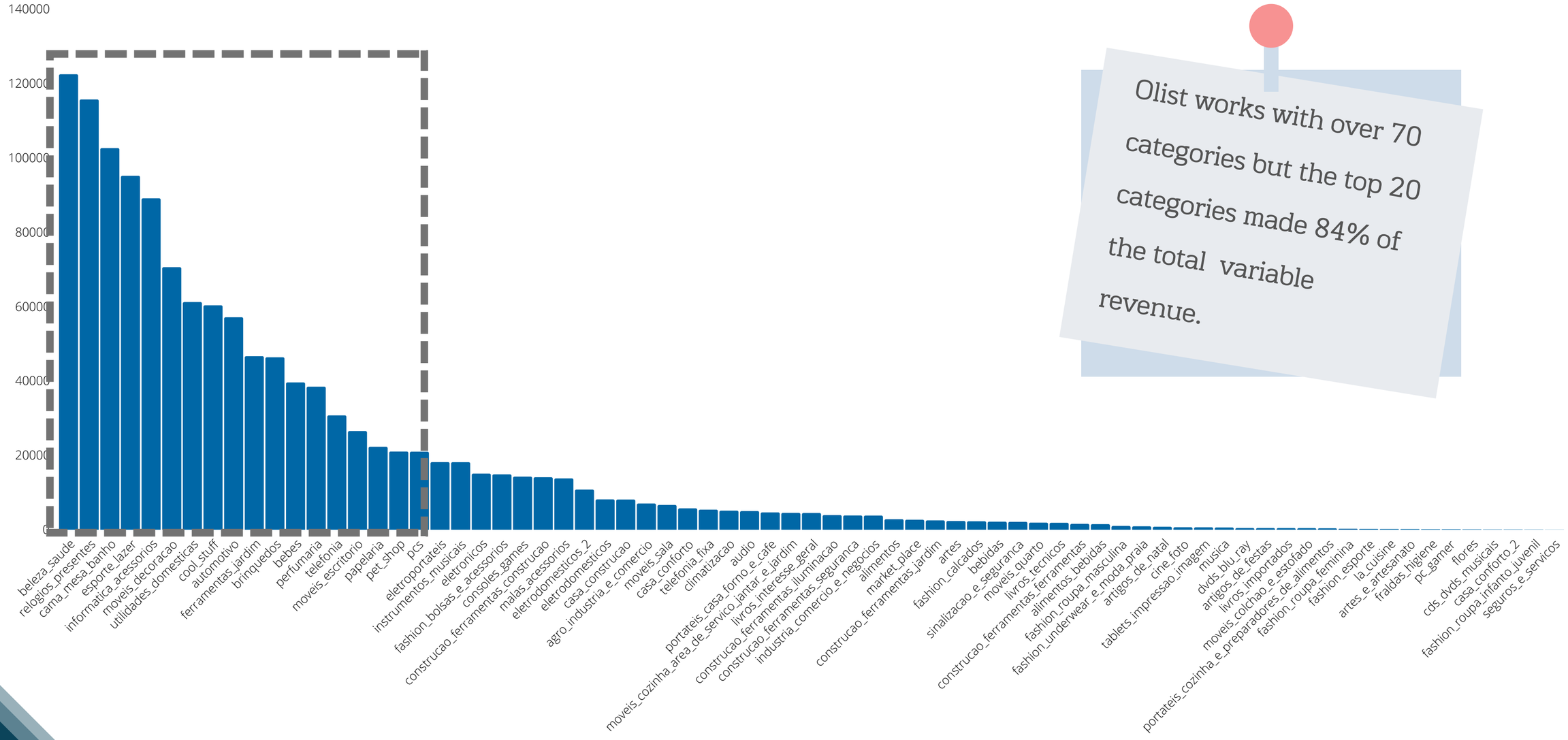


We only have a feel orders from 2016 on this database resulting in a poor analysis for 2016.

Despite of this data break we do have a good data from 2017 and 2018 and as we can see Olist has a good start with a growth trend in there results during this period.

Financial Overview

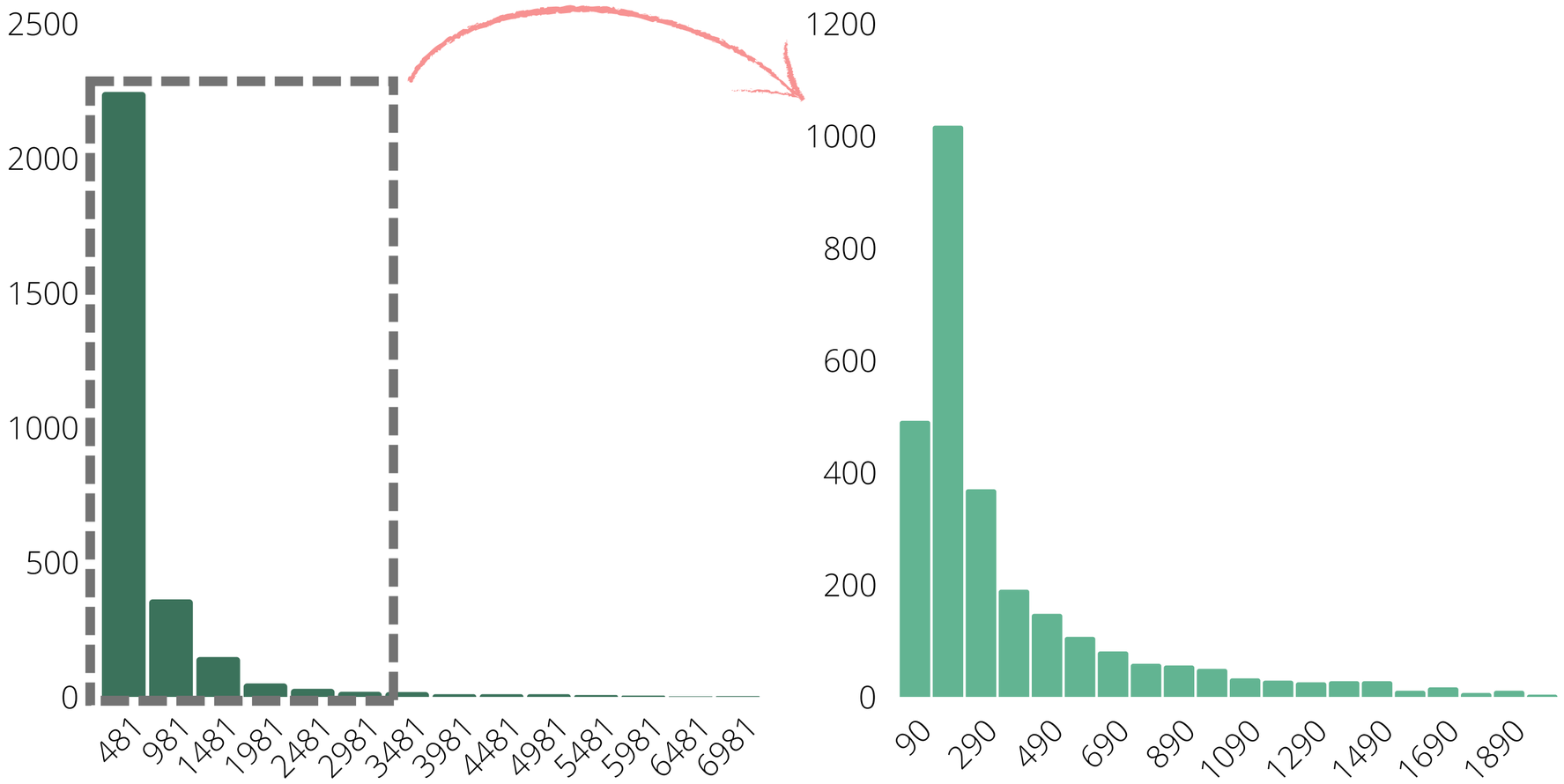
Variable revenues per category (2017 + 2018)



Olist works with over 70 categories but the top 20 categories made 84% of the total variable revenue.

Financial Overview

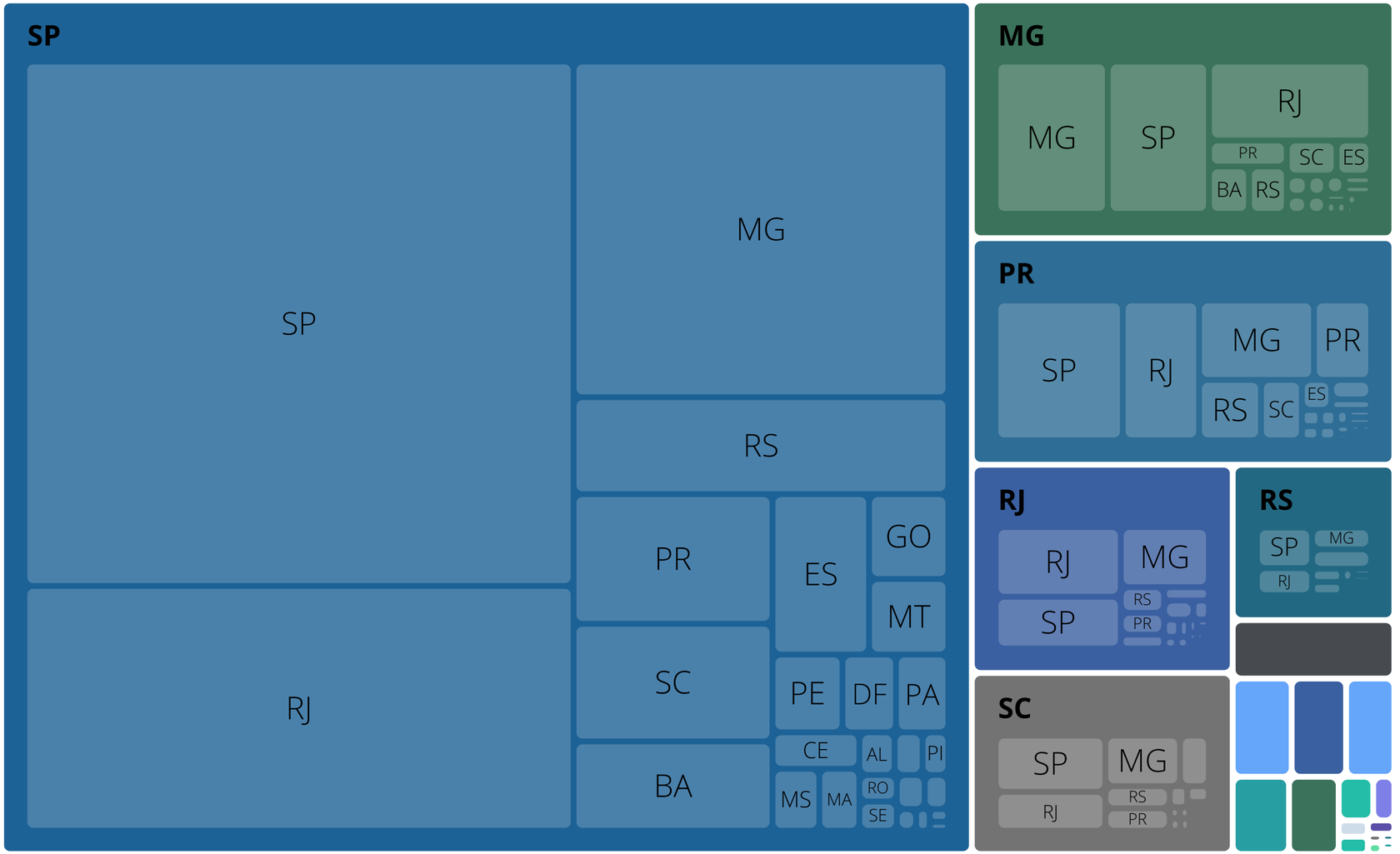
Revenues per Seller (2017 + 2018)



95% of sellers contributed at most R$ 2.000 with emphasis on clusters between 100 and 200.

Financial Overview

Number of orders from state to state (2017 + 2018)

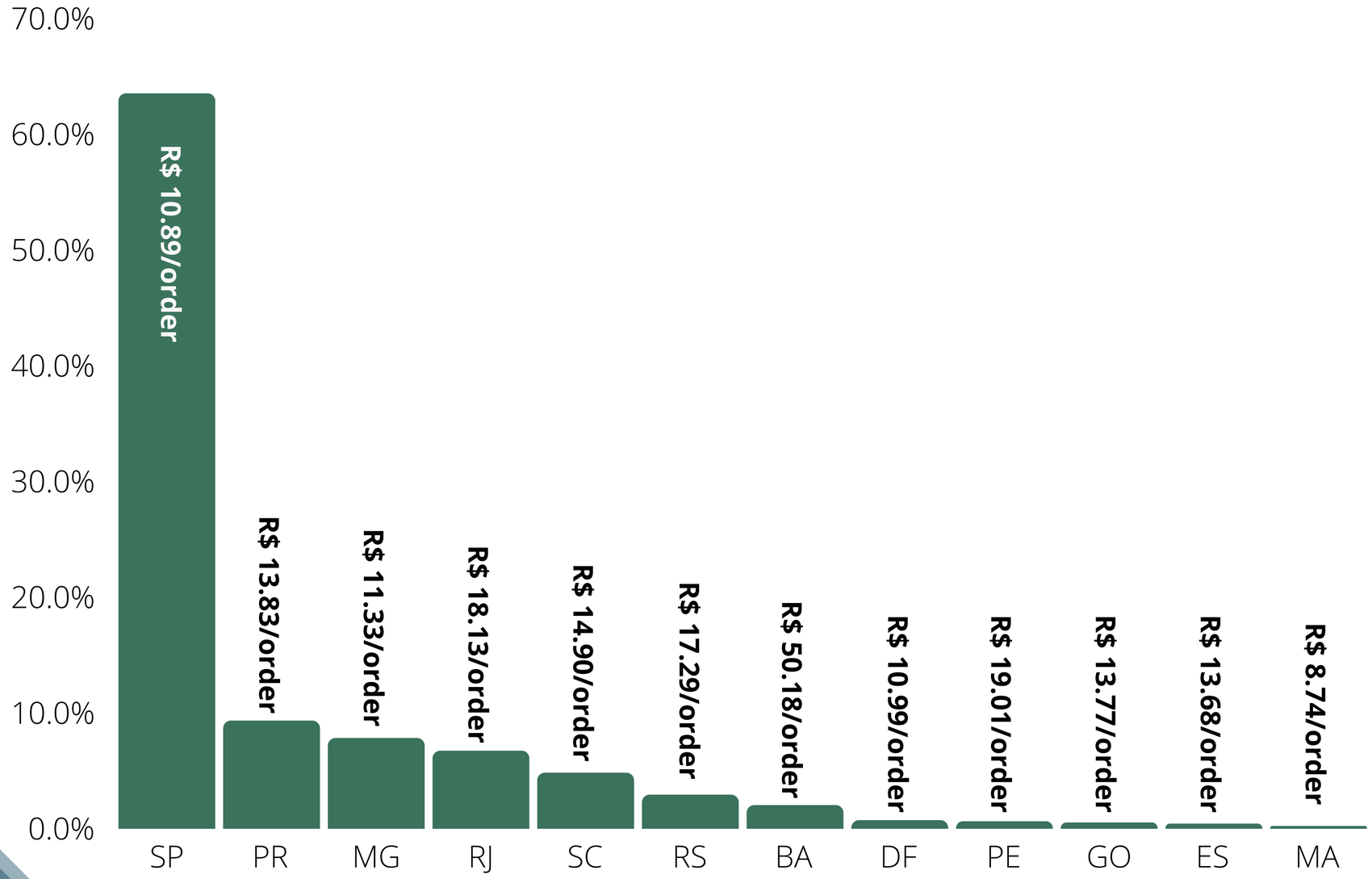


70% of all orders are sold from São Paulo following by Minas Gerais 8.3%, Paraná 8.1% and Rio 4.5%.

We can see a predominance of south-east states.

Financial Overview

Varible Revenue share and value per order (2017 + 2018)



The revenue have the same behavior, São Paulo lead with 64%.

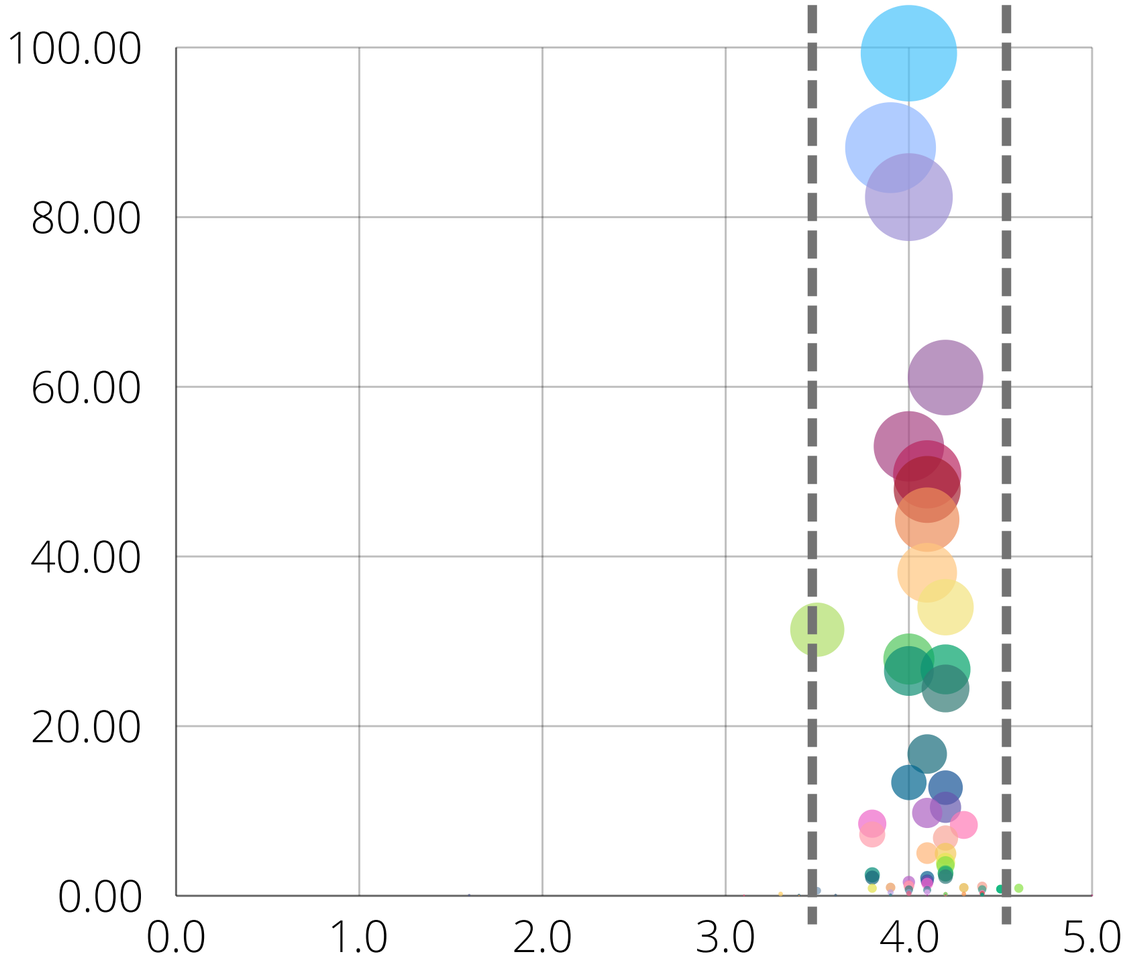
Despite the reavenure share São Paulo have one of the lowest value per order. This scenario can be justified by the state's high competitiveness.

Primary Analysis

Customer Satisfaction

Customer Satisfaction

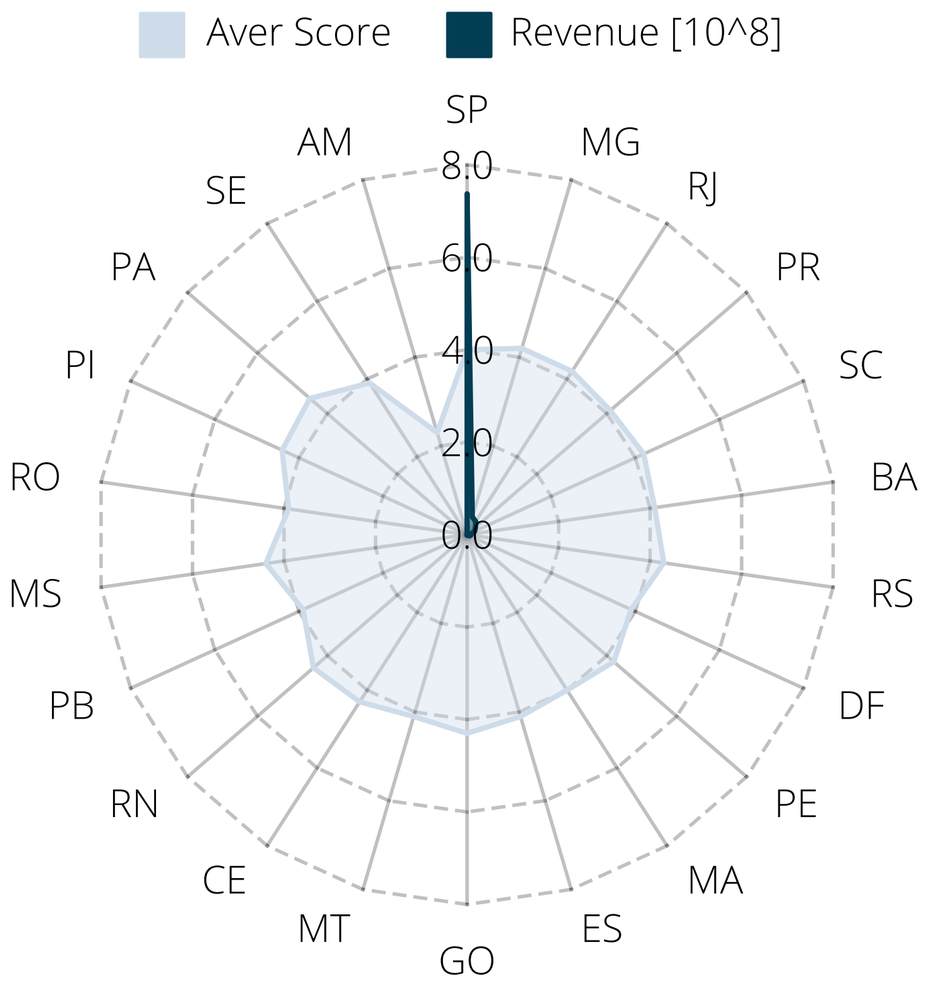
Revenues vs Score vs Category (2017 + 2018)



Considering all relevant categories they have a score between 3.5 and 4.5.

Customer Satisfaction

Revenues vs Score vs State (2017 + 2018)



With the exception of AM, all states have an average Sscore of 4.

Going Further

Customer Decision Tree

Customer Decision Tree

To better understand all opportunity gaps that Olist have fist we need to understand **Olist** customer.

To do so we are goanna apply Machine Learning techniques on Olist database to get the features importances over customer perspective.

We can only analyze features that are in the database, which are:

- Payment installments
- Payment Value
- Freight weight
- Item per order
- Delivery time
- Distance (seller – customer)
- Payment type

Customer Decision Tree

ML - Random Forest Classifier model

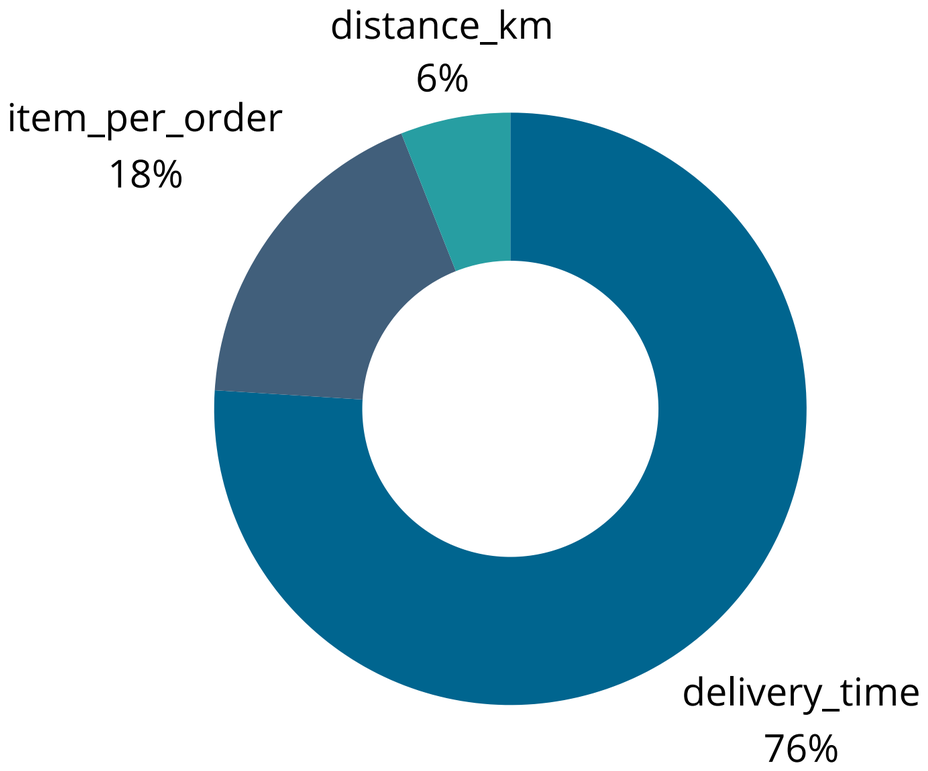
With SKlearn library we created a **Random Forest Classifier** model to get the following features importance proportion:

Feature	Importance
delivery_time	76%
item_per_order	18%
distance_km	6%

Obs: For this modeling we are taking the feedback score of each order and classifying as 5 - good, 4 - neutral and 3 and below - bad. If you want to see all steps you can check on this **link** for jupyter notebook file.

ML - Random Forest Classifier model

Fesatures importance [%]



Customer Decision Tree

ML - Decision Tree Classifier

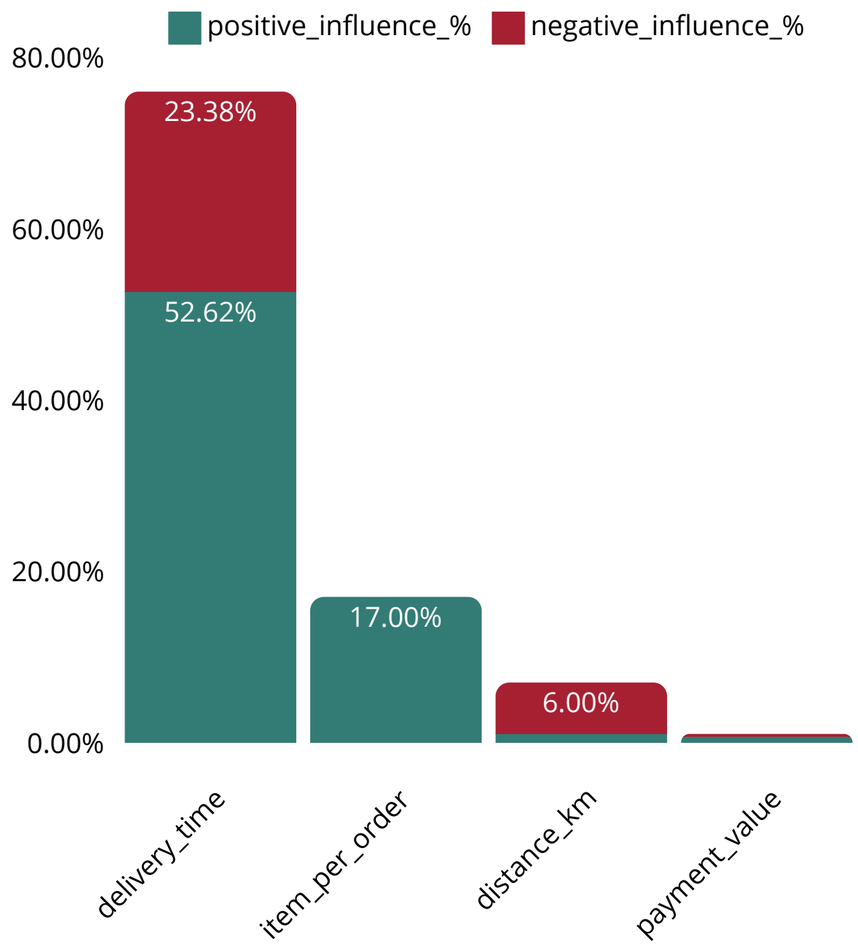
This time we created a **Decision Tree Classifier** model and get the following features importance proportion:

Feature	Importance
delivery_time	76%
item_per_order	17%
distance_km	7%
payment_value	1%

Obs: For this modeling we are taking the feedback score of each order and classifying as 5 - good, 4 - neutral and 3 and below - bad. If you want to see all steps you can check on this **link** for jupyter notebook file.

ML - Decision Tree Classifier

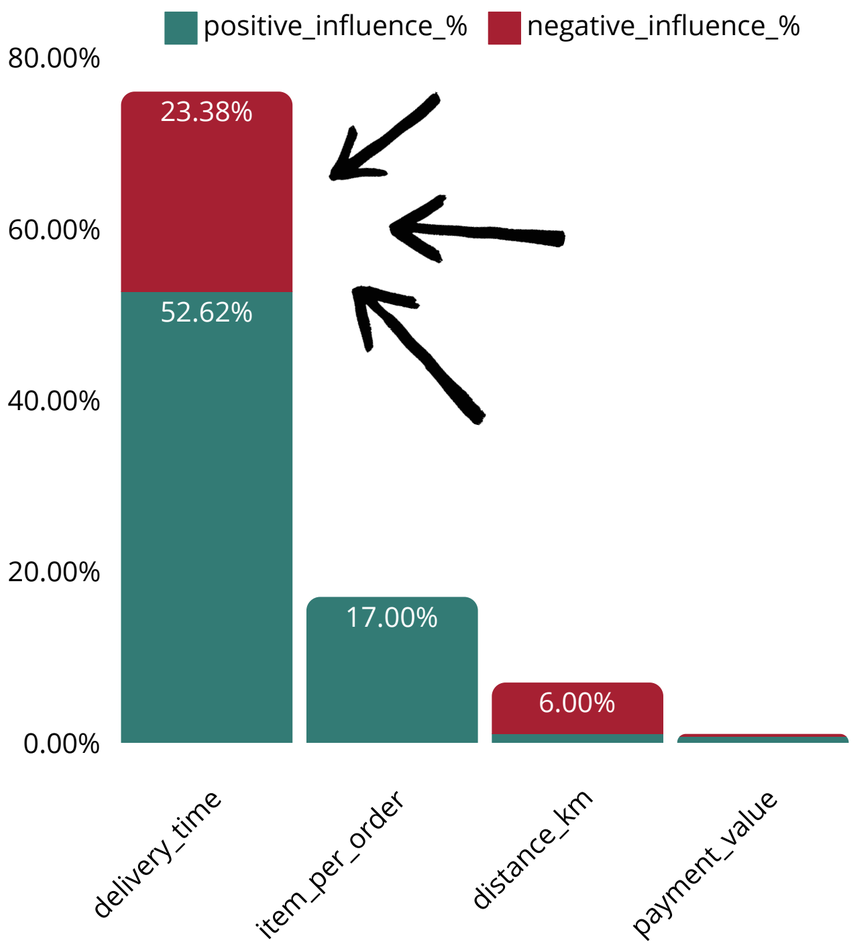
Fesatures importance split by pos e neg [%]



Customer Decision Tree

ML - Decision Tree Classifier

Fesatures importance split by pos e neg [%]



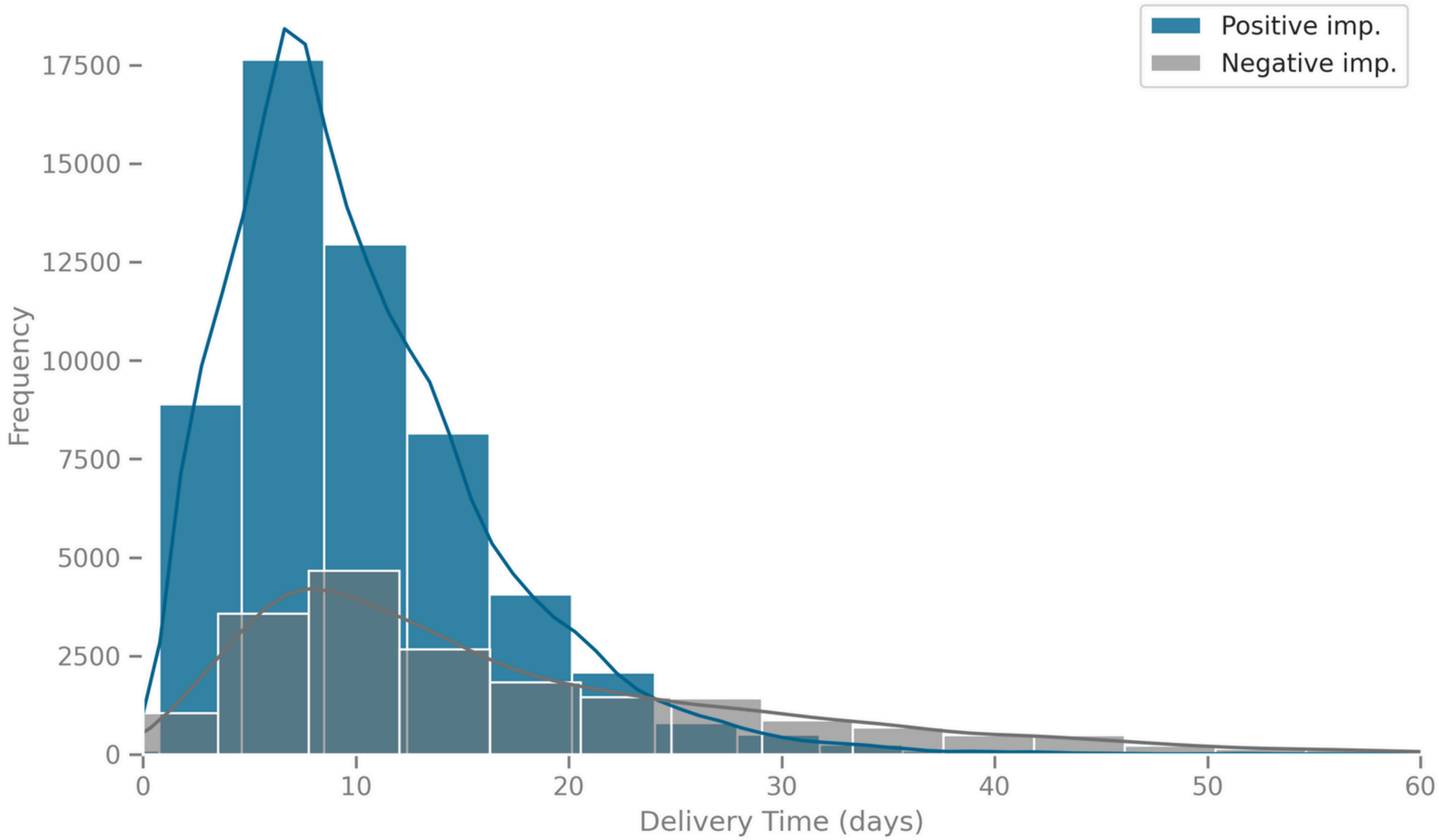
As we can see from our Customer Decision Tree analysis delivery time is the feature how has the greatest impact and the biggest opportunity. So we are going to focus on it.

Going Further

Delivery time

Delivery time

Histogram of impression

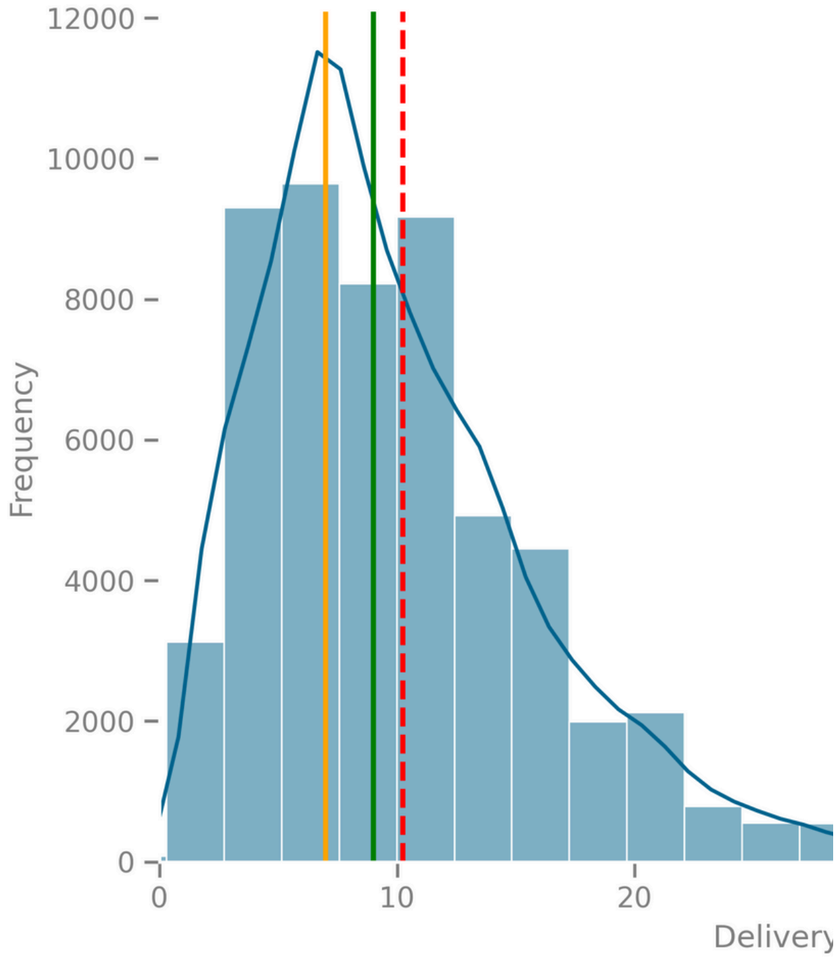


We can observe that the histograms have a positive skew and looking for each variance there is a big difference between them.

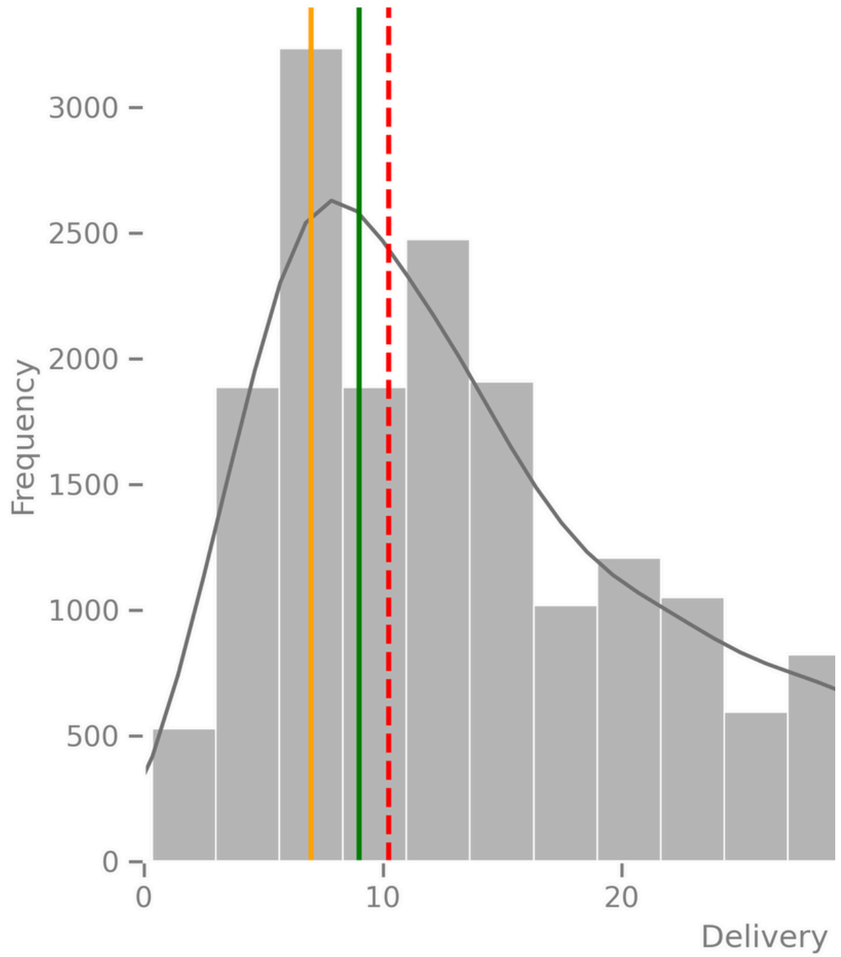
Impressions	Skew	Variance
Positive	3.5	45.9
Negative	2.6	190.9

Delivery time

Histogram - Positive impression



Histogram - Negative impression

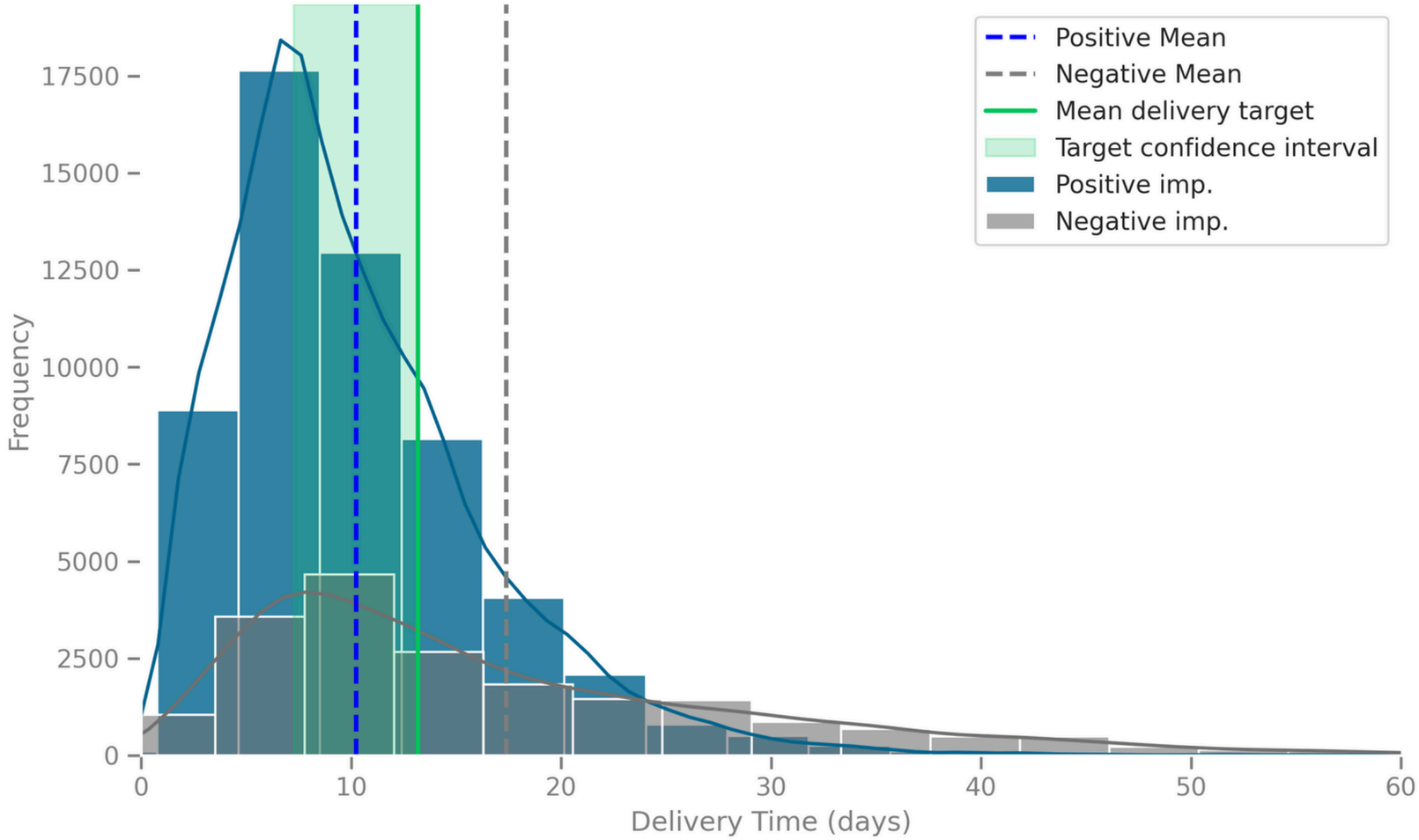


Measures of Central Tendency

	Positive	Negative
Mean	10.2	17.4
Median	9	13
Mode	7	7

Delivery time

Analysis of Mean



To understand what would be the improvement we should have on negative mean delivery time to make those impressions be positive we use statistic to get a minimum **mean of 13 days** with a 95% of confidence interval.

Next steps

1. Create a sistem to track and score the delivery time of each seller to bonus how make there deliverys faster.

2. Monitor the main offenders and, if necessary, unregister them.

Thank you!