OLIST Brazilian E-Commerce Analysis

**Abstract**: This paper presents findings of a public dataset of the infamous Brazilian OLIST e-commerce service connecting marketplaces and merchants. We analyse products, sellers and customers within this marketplace, giving recommendations on how OLIST can increase profits within each of these domains. Using OLS to model relationships between independent variables affecting the review score, we can say the delay between the time the order was purchased and delivery date is the biggest driver in bringing review score down. Using K-Means Algorithm to perform Customer Segmentation, classifying customers into different groups so we can modify marketing campaigns for different clusters of customers and improve customer satisfaction.

# Introduction

This paper explores a series of datasets containing information about sellers, buyers and customers from the OLIST national E-Commerce site. There are a total of 3028 sellers, 96516 customers and 32171 sellers which this database has records for featuring from the 4th of September 2016 to the 3rd of September 2018.

Each seller that has joined OLIST, uploaded a product catalogue, gets notified when the product is sold and then hands over the item to a logistic carrier. These sellers are charged a monthly fee. Customers browse products on the marketplace and purchase products from the OLIST store. The order is placed, and the customer can pay in installments, with different payment types offered.

Chart, pie chart

Description automatically generated

Figure : Payment Type Percentage for all Orders

The seller then gives the product to a logistic carrier delivering it to the customer. The customer is notified of an expected delivery time.

We aim to improve performance and popularity of the OLIST service, since there is a decline in orders being made. To complement our focus, we analyse a range of different questions and then build a customer retention model.

Starting off, we perform a brief analysis of top performing products, customers and sellers so that they can be given rewards/ discounts to incentivize further activity. The option to create leaderboards in the e-commerce service is proposed.

We proceed to look at why some product categories are underperforming compared to others in terms of review score and profit, modelling these via ordinary least squares regression to find the independent variables affecting each of these dependent variables. We also investigate factors affecting seller performance by state.

We finally perform an RFM customer segmentation, which is a data-driven technique enabling marketers to make tactical decisions.

# Analytical Questions

With the size of this dataset, there are a high number of questions we can investigate. This paper’s research focuses on the retail domain, focusing on the products, sellers and customers separately.

We analyse why certain product categories underperform compared to others by focusing on their review score and model the relationships between the independent and dependent variable using (OLS) Ordinary Least Squares Regression. We similarly do the same for sellers, identifying the reasons behind.

We explore the following:

* Product Analysis
  + What are the top 10 products?
  + What are the top 10 most profitable product categories? Top 10 least profitable product categories?
  + Why do certain product categories underperform compared to others?
  + How the product weight/dimensions and photo quantity affect sales/ profits and review score?
* Seller Analysis
  + What are the top 10 sellers? (include the city/ state of these too) What are the bottom 10 sellers? (profits take away freight value)
  + Why are some sellers underperforming compared to others?
* Customer Analysis
  + What are the top 10 customers?
  + Which groups of customers should be targeted? Which customers are the most loyal?

## Approach

This public data base is directly sourced from Kaggle. Featuring nine open-sourced datasets. This is a large database, and all the datasets can be connected via primary and foreign keys, which can be represented as seen in the data-schema below:

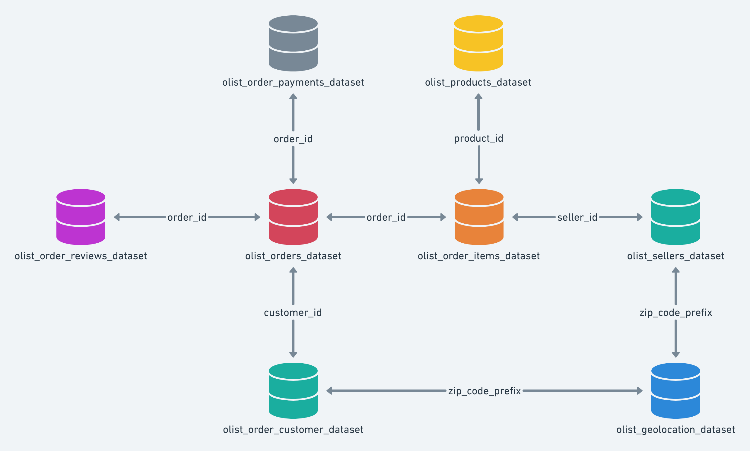


Figure : Data Schema

Merging all these datasets together, we create a master data frame. We answer each analysis question by extracting specific columns needed from the master data frame to answer research questions.

# Data (Materials)

## Key Characteristics

The data-intake report of the Master data frame is shown below:

Table

Description automatically generated

Given the high number of features, null values and cardinality of the dataset, for every individual research question, we select a subset of features from which we can then investigate.

## Suitability and Assumptions

There are a mixture of float, object, and integer data types and after some pre-processing, we converted all dates to the ‘date-time’ format. Further steps involve breaking date-time features into months, years and days (all integer types).

After data cleaning and re-engineering of existing columns, we create new columns, for example multiplying the price and quantity to get revenue. Then we take that away from the freight value which is the cost of transportation of goods from the logistic carrier to the customer to get profit.

The profits for the Olist marketplace, and for sellers is not clear. We assume freight value is the only loss of revenue since cost of production is not given. We assume the profits/losses Olist make are the differences between the final payment values (including no. of instalments) and the profits made by the seller (*(price \* quantity) – freight\_value*).

The limitations of the data include the fact we have customer, seller and product IDs instead of names. These consist of a long string composed of random letters and numbers; hence we cannot truly identify the name of the top sellers/ customers and products. Moreover, the missing values in the review dataset, especially for review comments and titles, so we will not use these features for analysis and model creation.

# Analysis

## Surface-Level Analysis

We have the highest achieving products, sellers and customers in OLIST. OLIST can look to offer rewards to sellers and customers. Also, OLIST can tell sellers to sell more of these products.

Graphical user interface, diagram

Description automatically generated with medium confidence

Figure : Top 10 Products, Sellers and Customers

Customers who buy a certain number of times from the marketplace can also have discounts given to them. We explore this further with customer segmentation. A similar scheme to sellers can also be applied, involving the most frequently retained customers receiving free items and a mention on the website.

Graphical user interface

Description automatically generated with low confidence

Figure :Revenue and Profits for Product Categories

We can see that watches generate the most profit, but health and beauty has generated the most revenue. This is perhaps due to the price of the watches. There is a smaller number of products within the watches category that have been delivered.

## Time-Analysis

#### Orders

Chart, bar chart

Description automatically generatedWe can see that the number of orders per month in Olist is only 377 during 2016, and the number of orders increases significantly during 2017, with there being an all time high of 8881 in November 2017. Majority of growth and usage in the Olist market occurred during 2017, with the number of orders then stagnating in 2018, with a slight decrease during the year.

Figure : Order Time-Series -Number of orders per month from 2016 – 2018

#### Customers

Customers have followed a similar pattern to orders with the maximum number of customers using the service in November 2017. Following on from this, we see the number of customers starts to level of from 6000-7000 each month from 2018 onwards.

Chart, histogram

Description automatically generated

Figure : Customer Time-Series: Number of customers using Olist from 2016-2018

#### Sellers

The number of sellers using the service follows the same pattern as orders and customers, with a peak in November 2017 following a consistent pattern in 2018.

Chart, line chart

Description automatically generated

Figure : Seller Time-Series: Number of sellers from 2016-2018

## In-Depth Analysis

#### Why some product categories underperform compared to others?

Chart, funnel chart

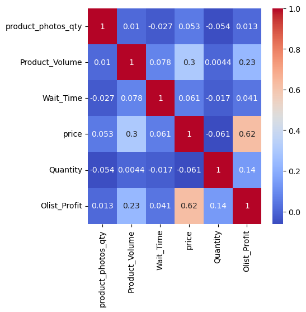
Description automatically generated

Bed-Bath-Tables, Health and Beauty and Sports Leisure product categories record the highest number of orders. OLIST should encourage more sellers from these categories to sell these products, since they are popular among customers.

Table

Description automatically generatedSecurity and Services has made the least number of orders and is recorded to have the smallest profit, with no money being made by Olist.

Sellers for these product categories need to be contacted and failure to improve performance should result in removal.

****We investigate a few interesting features and attempt to model them using OLS (Ordinary Least Squares Regression) which affect the review score of the product category. We can see that a higher product volume has a slight positive correlation with the price of the product, and therefore the profits made.

Formula for review score:



Table

Description automatically generatedEvaluating the model:

Chart, bubble chart

Description automatically generatedAll independent variables bar Product Volume and product photos quantity have p-values less than 5% suggesting these are all statistically significant. Wait Time and Quantity show a reasonable negative correlation with review score, suggesting that Olist sellers need to improve on delivery and quality of multiple-item orders.

We can see that a larger product volume and a higher wait time both contribute to a lower review score. This could suggest that larger products take a longer time to deliver, so there is a longer period between estimated and actual delivery time, reducing customer satisfaction.

It is important to minimise wait time as much as possible, since this is a driving factor in poor reviews of the Olist service, dissuading new customers from joining.

#### Why sellers from some states perform better compared to others?

After some analysis, we investigate a few features which could correspond to state performance for sellers.

Chart

Description automatically generated with medium confidence

Each point in this scatter graph represents a state. We see that neither wait-time nor delay influences the final review score.

## Customer Segmentation Model

#### Building the model (Scaling data, Hyperparameter tuning)

We use the K-Means Algorithm to find out segment customers into different categories based on their recency (when they last purchased from Olist), frequency (how many times they purchased from Olist) and monetary (the total amount of money paid by each customer) values.

K-Means is very simple to implement, the only hyper-parameter we need to tune is the number of clusters. This unsupervised machine learning algorithm is a popular choice for data scientists wishing to implement customer segmentation.

Chart

Description automatically generated

After feature engineering and creation of the RFM dataset, we take the log transformation of our data points due to Frequency and Monetary distributions being heavily skewed. We then use a Standard Scaler to standardise data points before training our model.

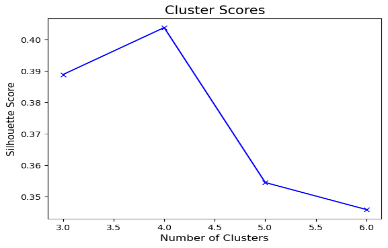
Chart, histogram

Description automatically generated

We can see that Frequency and Monetary histograms are less skewed.

Chart, line chart

Description automatically generatedWe look at the ‘Elbow Method’ for evaluating the number of clusters. We can see there is no clear “elbow mark”, however there is a slight bend when there are 3, 4 and 5 clusters. Hence, we determine the highest silhouette score for these K values.



We clearly can see that K = 4 has the highest silhouette score, so we train a final model breaking down and grouping customers into 4 clusters.

#### Cluster Analysis

Chart, bar chart

Description automatically generatedCluster 1 has the largest number of customers (48393), with these customers having the highest recency value of 300 days on average, suggesting they haven’t made orders through the service for the longest amount of time. There is a small average monetary value and a frequency of one, so we classify this cluster has **low spending** customers.

Cluster 2 (21254) and these customers all having slightly lower recency values than Cluster 1 suggesting they brought from the service more recently. A high monetary value suggests this cluster contains **high-spending** customers.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedCluster 3 (2807) and has the highest average frequency value of 2.1 suggesting these customers are frequent users. A high monetary value also suggests that these are high spenders. Combining all the statistics, we can classify this cluster to having **frequent and higher spending customers.** Considering the monetary value is almost the same as for Cluster 2 for a significantly smaller population, these customers are key for Olists e-commerce growth and profits. It is very important to retain these customers.

Cluster 4 features has the lowest recency value; Frequency values average at 1, and the Monetary value is low. We classify these customers to be **low-spending recent customers.** Marketing campaigns can be tailored to entice these customers to come back and reuse the service.

# Findings, Reflections and Further Work

#### Findings

To conclude, this data has provided insights into Olist Performance of buyers, sellers and customers. To start with we can provide leaderboards for customers, sellers and products and focus on underperforming product categories.

There is no correlation between seller state and review score. There is a correlation however, between seller state and profit. The wait time is the biggest impacting factor on review score of products.

We can see the Top 10 products, customers and sellers below. A seller leaderboard can be created with rewards (a free purchase or a gift voucher) can be given to these sellers to incentivize them further.

In terms of customer retention, the primary focus should be on keeping customers from Cluster 3, which are frequent buyers and high spenders. We can see from our time-series analysis that the OLIST E-Commerce service orders, sellers and customers are all dependent on each other. November 2017 was the peak of OLIST, with there being no further growth in 2018.

#### Conclusions and Reflections

There are lots of outliers within the RFM data. A disadvantage of K-Means is that it does not consider outliers, so our clusters could contain anomalous results. Also, we have only evaluated customer retention on the recency, frequency and monetary values within our model; other factors should be considered such as the waiting time the customers.

Product categories which are underperforming can be removed from the marketplace. Some of these product categories result in a loss in terms of overall profit of Olist.

Analysing which customers are our best customers and the characteristics of them will help OLIST with their marketing strategy as they can improve their service to different types of customers. This will draw more customers into OLIST, motivating sellers to do the same.

Olist should set a threshold of over 7500 orders by introducing new sellers to the company offering a different range of products. An improvement to the user website and making the website more accessible via search engines can be considered. Video adverts and advertising on other websites are two options to be considered.

#### Further Work

Marketers can use RFM to identify and differentiate customers into homogenous groups so they can use different marketing strategies and campaigns on different groups. This is a common technique for companies to keep their most loyal customers, while also attracting less frequent customers more. RFM is used to find out which customers contribute to the churn rate, and linked to this, we can also evaluate models in making a churn prediction, identifying which customers are likely to keep using the marketplace however this goes beyond the scope of this paper.

We can look at ways of enticing new sellers, since there are a higher number of customers in comparison with sellers. A time-series analysis of the number of sellers can be constructed to monitor seller activity. Sellers can be leaving the marketplace. We explore whether sellers who offer customers to pay in installments generate more profit for themselves. Sellers can work to minimize freight value by choosing less expensive logistic carriers. We can use an ARIMA/ SARIMA model to predict the future revenue/ profit/ OLIST Profit in the future.

RFM Model Analysis can be done, analysing customers from cluster 2 to find out why these customers have a higher frequency than other customers.

Another option is creating a Churn Prediction to see if customers are likely to come back. This investigates all the factors as to whether a customer will be retained. We are looking at a classification model question here.

# Word Counts

Abstract : 103

Introduction: 287

Questions: 263

Data: 259

Analysis: 1061 (61 words over)

Findings, Reflections, Further Work: 591