Ecole polytechnique fédérale de Lausanne

May 11, 2019

Data science in practice

Data-driven strategic analysis of Olist e-commerce

Authors:
Augustin Fontugne
Paul Mermod
Kilian Girardet
Thevie Mortiniera
Mathieu Shiva

Supervisor:
Omar Ballester

Teacher:
Christopher Bruffaerts

Contents

1 Introduction & problem statement

1.1 Context

Olist is the largest department store in Brazilian marketplaces. It aims to connect merchants to the largest marketplaces such as Amazon, Carrefour or Walmart. Olist offers a complete solution where the sellers can sell their products through the Olist store and ship to the customers by means of Olist logistic partners. Olist services fill the gap between small businesses and marketplaces, and facilitate the introduction to e-commerce for these merchants. The solution is not only intended to small businesses, but also apply to tenants who seek better placements in marketplace or retailer searching for new sales channels.

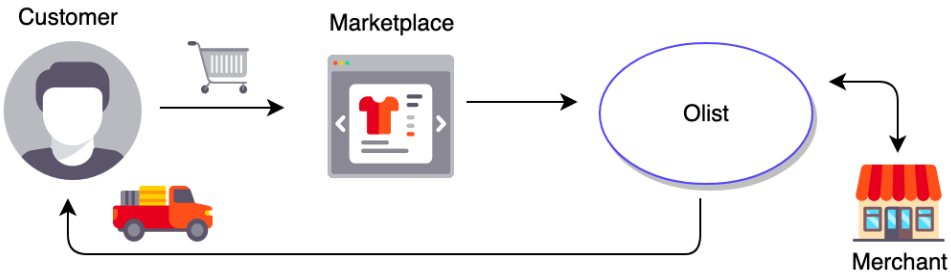


Figure 1: Purchasing flow through Olist

1.2 Business matter

The guiding thread of this study is to add value to Olist's business. E-commerce businesses such as Olist need to maintain a good reputation in order to stay competitive. Firstly, our main concern would be to find which sellers or what factors could affect Olist's credibility and propose solutions to improve the weaknesses. Secondly, we help Olist in their marketing decisions to increase their sales.
The preliminary investigations consist in identifying factors that make up customer satisfaction or dissatisfaction through customer reviews.

Bad reviews are not necessarily Olist's responsibility but are rather related to the seller. The next step is therefore to segment the seller basis because Olist's business plan rely mainly on these stakeholders. Think that each time Olist is chosen during a sale on an online marketplace, Olist receives a commission. On the other hand, if customers on marketplaces are not satisfied with their orders, Olist's brand image is tainted even though it is likely to be the seller's responsibility. From these perspectives, we would reward sellers that are the most valuable and reliable based on indicators such as sales and customer reviews. Proceeding further, we aim to understand the drivers of these valuable sellers through the product they sell, the delivery time and other features. On the other way around, we would like to isolate the worst sellers like the ones that makes the least sales or the ones that are the least reliable. The next step is to identify the drivers of being a unreliable or a worthless seller.

However, we shall be entitled to ask how such an analysis could help Olist in improving their business. Actually, this could aid Olist's actions taken towards their sellers. For instance, if

a merchant is identified to be a big seller and that its associated customer reviews are always positive, Olist would be tempted to promote him on marketplaces and find him the best possible spot. At the opposite, if a merchant always causes bad review because of delays in delivery time or quality problems, Olist would take measures against this merchant and possibly breach his contract. Think about a more subtle situation where a merchant does not bring much sales but shows a good customer satisfaction, Olist can see in him a potential and offer him better positions on marketplaces.

A deep analysis of the seller basis is thus crucial to improve Olist's reputation and sales.

In this logic of adding value to Olist, as a last element of this study we provide a review of the marketing side of the business. We will first examine Olist advertisement policy and give them insights to improve it, then we conduct a sale analysis to help Olist in their choice of products to spotlight on marketplaces.

1.3 Data driven strategy

We will use Olist's public dataset to tackle this problem. One may want to ask why a data driven approach is appropriated for such an analysis. For instance, we are willing to identify the best and the worst sellers according to some metrics, so using a data approach is convenient for such a task. Since we are also interested to understand why a seller has a certain value or reliability, we want to find out the relationships between features that could not be seen naturally. In other words, the data driven approach helps us to identify unseen dependencies and bring a quantitative framework to the analysis.

2 Dataset

2.1 Description

The datasets are provided by Olist. We will use the following datasets :

- ***The orders dataset*** gives informations such as purchase date, delivery date and estimated delivery date, each observation corresponds to an order. This dataset serves a as benchmark for merging the datasets together and contains approx. 100K observations before cleaning.
- ***The payments dataset*** stores informations about payment type and payment value for each order.
- ***The customers dataset*** contains informationd related to customers and their location. Each order is assigned to a unique customerId, while customerUniqueId is used to identify customers that made repurchases at the store.
- ***The reviews dataset*** shows satisfaction survey details, such as review score and review comment, completed by the customer once the product is delivered.
- ***The products dataset*** stores the product ID, the category, the dimensions, and other attributes. ***The order items dataset*** retrieves the items of each order, the product ID, the seller, the freight value and other attributes. Be aware that the order_items dataset

may have several rows with the same ”order ID” because one order can contain several items.

- ***The seller dataset*** gives seller state and city.
- **The geolocation dataset** basically contains spatial coordinates of cities where orders are performed.

The figure (2) shows how the different datasets are linked together.

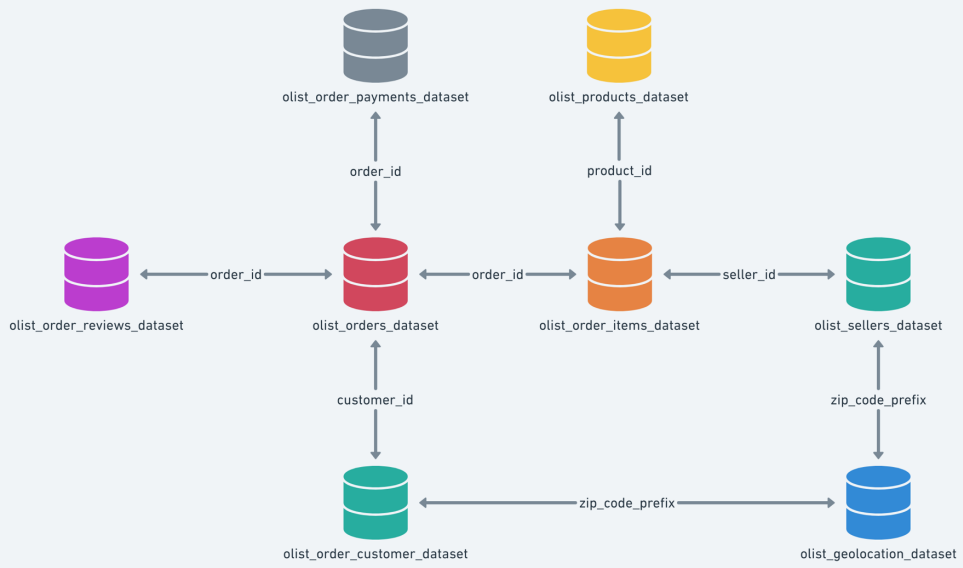


Figure 2: Relational Diagram of Olist's Datasets

For the forthcoming study, we will not use the geographical dataset nor the seller dataset.

2.2 Preprocessing

In the preprocessing, we have to verify mainly if there are missings values and duplicates. The idea is to obtain an exhaustive dataset that gather all the relevant features so that we can use it in the different analysis. We make the choice to base this dataset on an order_id basis, meaning that each row corresponds to an order.
For ***the orders dataset*** we drop the lines where there are missing values.
For ***the payments dataset*** we identify the duplicates with the ”order ID” and drop them.
We repeat this data cleaning process for ***the products dataset*** and ***the order items dataset***. This two can be merged on the product_id key. We call this merged dataset ***products order items*** .
For ***the customers dataset*** we have to verify if the ”customer ID” is unique. This one has to appear only once, as we said before, because each ”order ID” give one ”customer ID” although the purchase is made by the same customer. On the contrary ”customer unique ID” can appear more than one time and allows us to verify if one customer has purchased more than one time on Olist.
For ***the reviews dataset*** we choose to not directly drop null rows for comment title and review message because it is very likely that a customer doesn't write any comment although he

grades it. Furthermore, these rows represent almost half of the data set. We will just fill them with a "The customer did not comment", which in portuguese is translated to "O cliente não comentou". We also have to verify the duplicates for this data set.

Then, we merge ***the orders dataset*** with ***the customers dataset*** with the order_id key. To this, we add the payments dataset on the same key. Because we want to keep the key to be the order_id, we merge it with ***products order items*** on the order_id key. However, when doing so, a problem arises because ***products order items*** may contain several row for an order (i.e. if this order is composed of several items). Therefore, for these multiple items orders we modify the payment value to be the one of the full order, same for the freight value. Concerning product characteristics, we choose to take the ones of the first item in the order. Note that 90% of the orders are composed of only one item, and we will eventually remove rows that consist of orders with multiple items if needed.

3 Customer Satisfaction

The goal of this section is to analyze what makes up satisfaction and disappointment. We will not focus on review scores prediction but rather on discovering where to take actions for increasing customer satisfaction. We first try to understand which features are relevant to determine the review score. Secondly, we conduct a sentiment analysis by examining review comments.

3.1 Exploratory analysis

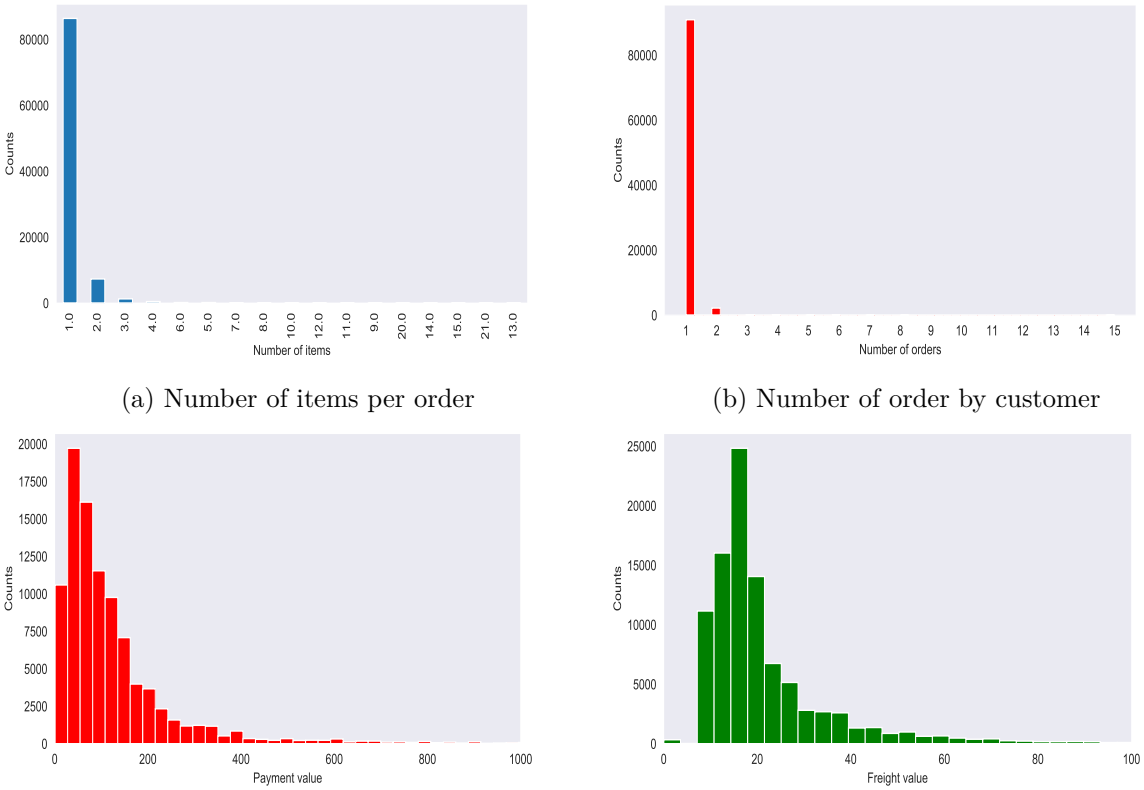
As a preliminary analysis, we present basic statistics that help us understand customers behavior. Note that the price of an order does not include the freight.
From the set of figures (3) we note several observations:

- Figure (3a): 90% of the orders are composed of only one item.
- Figure (3b): more than 95% of the customers have ordered with Olist only once. It would not be so relevant to spotlight the most frequent customers in this case, as the vast majority of them is highly infrequent.
- Figure (3c): in general, people tend to go for cheap orders. The mean price of an order is **137** and the median is **86**, the distribution is rightly skewed.
- Figure (3d): the freight value distribution follows approximately the price distribution

We may also want to know which states are generating most of the sales. From figure (4a) and (4b), it is in the Sao Paolo state (SP) that there are the most orders, but at the same time this is the state with the cheapest orders in average. This is expected as they are the most populated places in Brazil.

3.2 Effect of features on the review score

We put now the emphasis on the categorical and continuous features that may have an effect on the review score.



(a) Number of items per order

(b) Number of order by customer

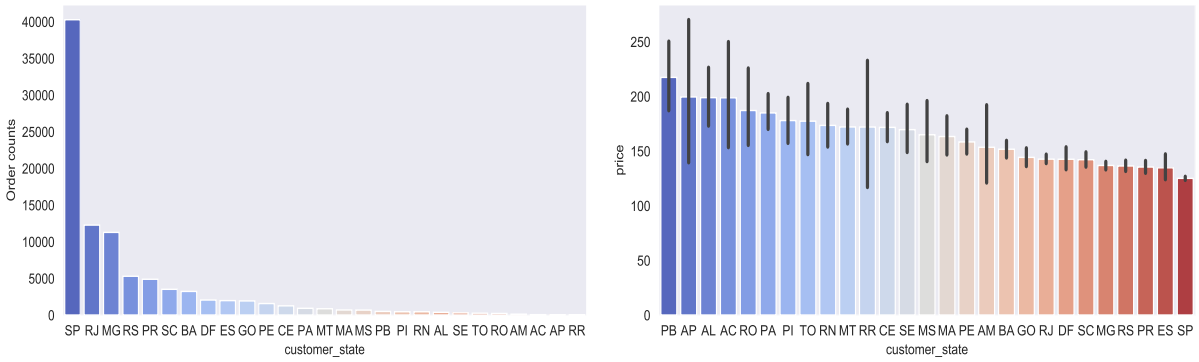
(c) Price of goods per order

(d) Number of order by customer

Figure 3: Univariate statistics on orders

It is clear that the natural proxy who reflects customer satisfaction is the review score. This measure is given on a Likert scale from 1 to 5 where the worst score is 1. From figure (5), we see that the review distribution is highly unbalanced and is mostly composed of the best score with a mean score of **4.14**. It means that most of the customers are happy with their order, but there is still 10% of the orders where customers are totally unhappy. 10% of disappointment is consequent and not negligible, all these disappointed customers will certainly not use Olist facilities again.
We propose to add additional features that can help in explaining the feature score.

- $\textit{Real Delivery Time} = \textit{Order Customer Delivery Date} - \textit{Order Approved At}$
- $\textit{Estimated Delivery Time} = \textit{Order Estimated Delivery Date} - \textit{Order Approved At}$
- $\textit{Delay} = \max(\textit{Real Delivery Time} - \textit{Estimated Delivery Time}, 0)$
- $\textit{Late} = \min(1, \textit{Delay})$ (binary)
- $\textit{Freight Ratio} = \textit{Freight Value} / \textit{Price of the items in the order}$
- *Review Before Delivery* which equals 1 when the review form is fulfilled before the order is delivered. Note that that the review survey is sent



(a) Order counts

(b) Average price of an order

Figure 4: Geographical statistics

- *Simplified Review Score*, equals 1 when *Review Score*= 4 or 5, 0 otherwise.

For now, we keep only relevant features that may have an effect on review score, which are shown in the correlation matrix (c.f. figure (6)). We directly see that there is a correlation between review the real delivery time, the delay and late deliveries. Note that these time related features are correlated with each other, and we will mostly refer to late and delay for the following statistics.

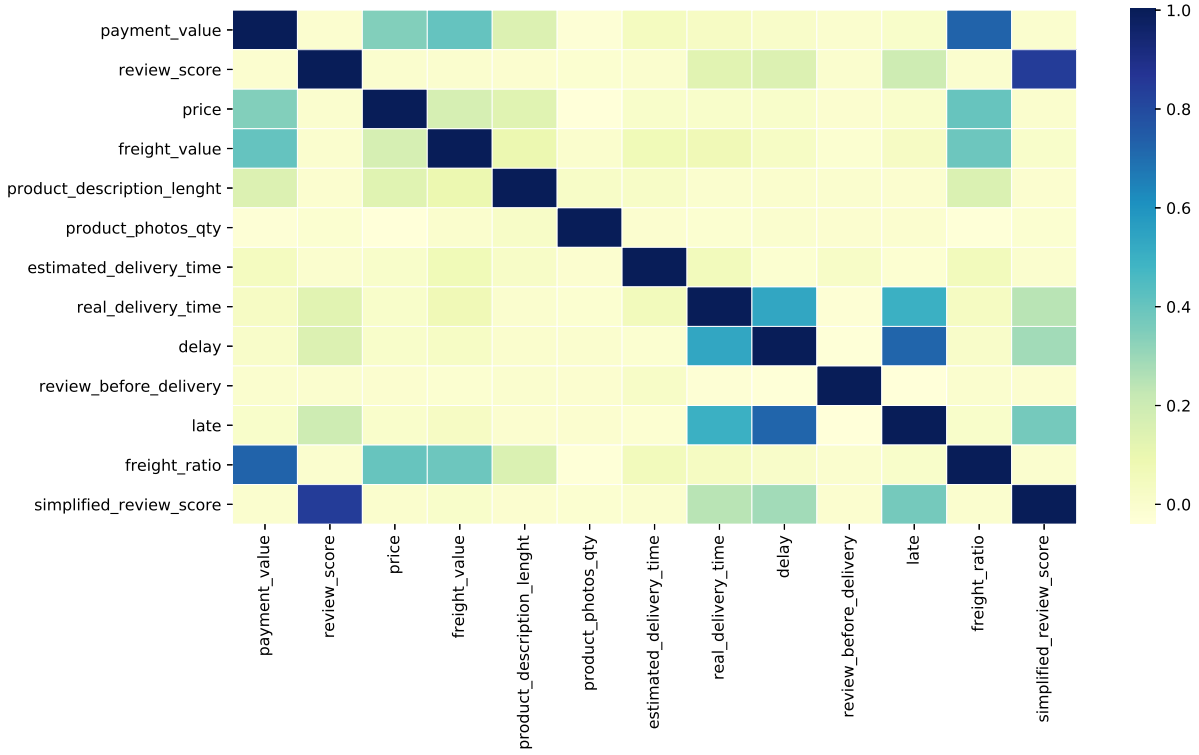


Figure 6: Correlation matrix

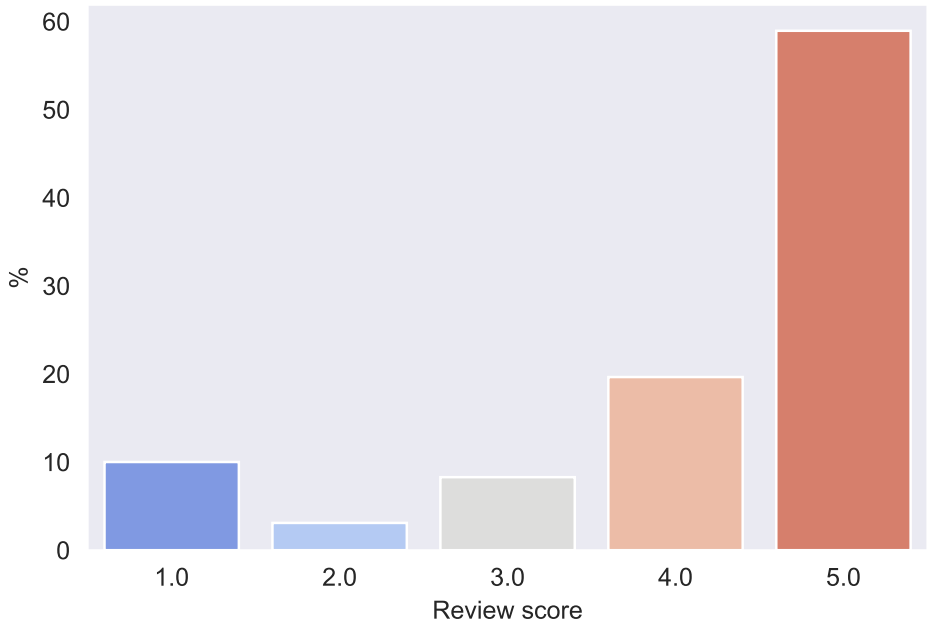
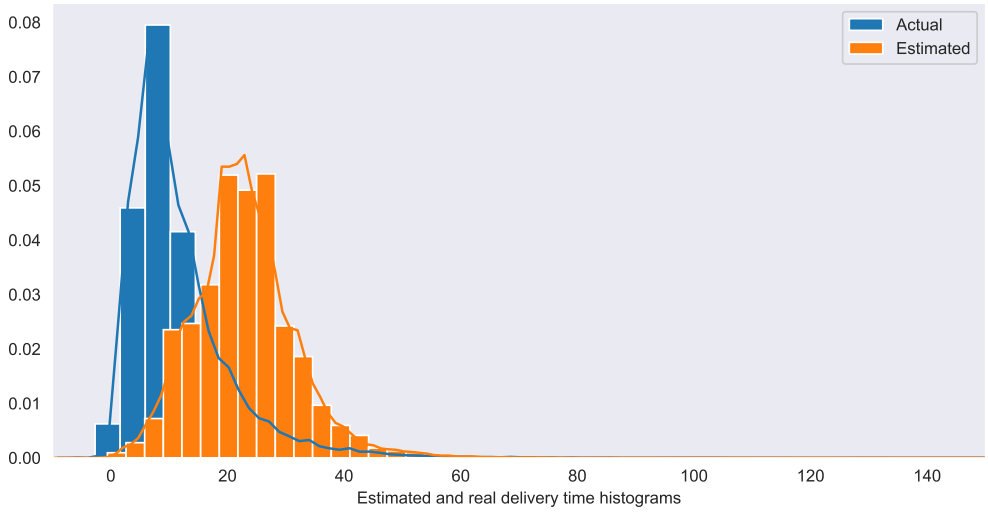
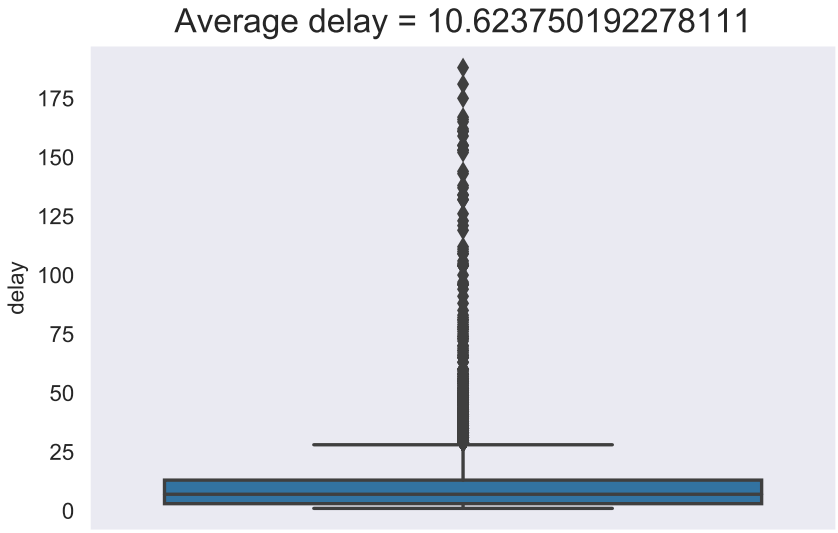


Figure 5: Review score counts

In figure (7a), we see that the estimated delivery time is approximately normally distributed while the real delivery time is rightly skewed, meaning that the right tail corresponds to late deliveries. Considering delayed orders only, the average delay is of 10 days, figure (7b) reveals a bunch of outliers with a delay greater than one month!



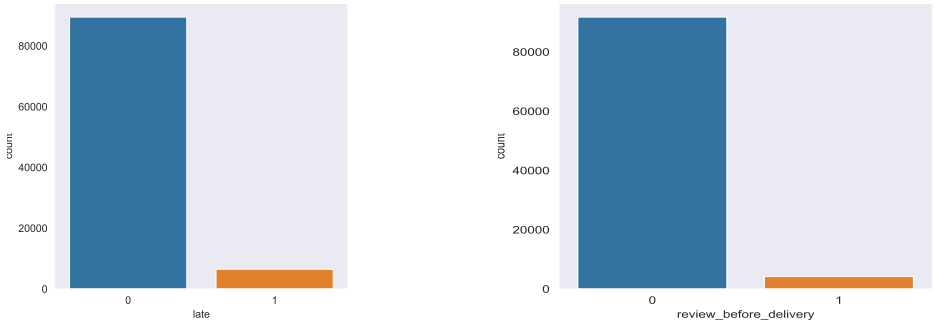
(a) Distribution of real and estimated delivery time.



(b) Boxplot of the delay

Figure 7: Distribution of shipping time features

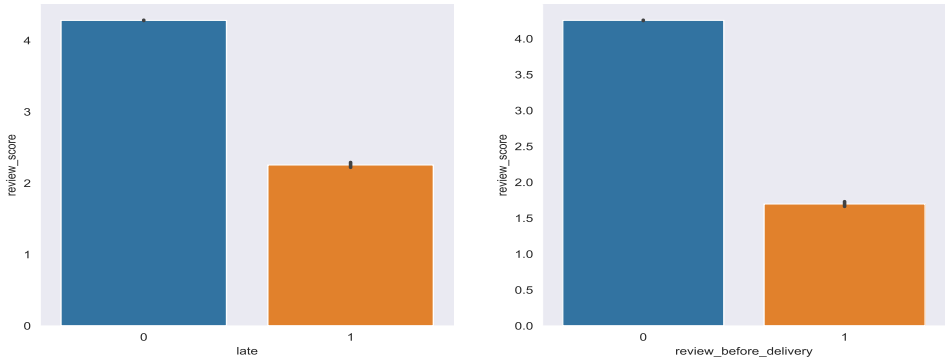
About the late frequency, we see in figure (8a) that the late delivery frequency is very low ≈ 0.067. This is actually a good performance, less than 10% of the of the deliveries are late.



(a) Late delivery = 1 (b) Review fulfilled before delivery = 1

Figure 8: Counts of additional features

We have to point out that the review score are somehow biased because the form is automatically received the day of the estimated delivery. Hence, if the parcel is late the customer will have to fulfill the review survey before having received its order. In figure (9b), we notice that receiving and fulfilling the form before the delivery is impacting consistently the average review score. However, receiving the survey before the delivery is not strictly speaking the real cause. The first cause of this drop in scores is more about lateness, as it is shown in figure(9a). But note that the drop is even greater for people that filled the review form before having their order. Indeed, imagine a customer whose order is delayed. It is no surprise that this customer will be more rude in its review if he has to review the order before receiving it than if he had already received it. Therefore, we recommend to Olist to change their survey mailing policy by sending the review form once the order is really received.

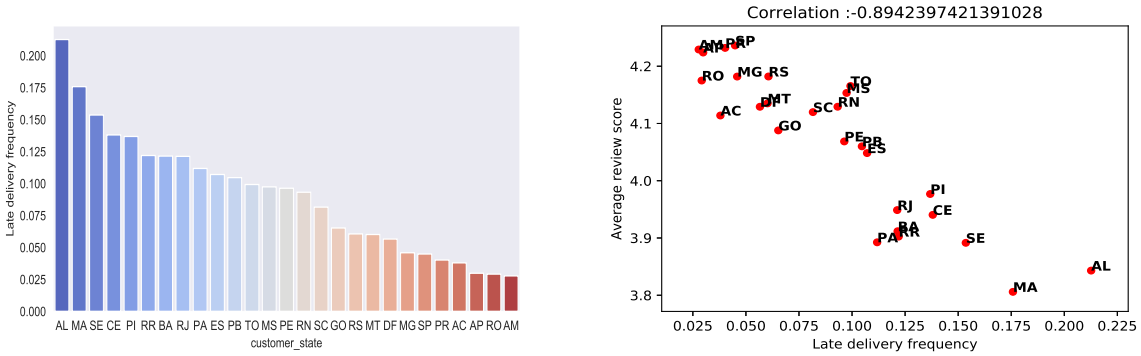


(a) Late delivery = 1 (b) Review fulfilled before delivery = 1

Figure 9: Average review score for different groups of observations

We see in figure (10b) that there is a strongly negative correlation between the average state review score and the late delivery frequency. But what is really remarkable is the variation of the lateness rate between state where in Alagoas (AL), more than 20% of the orders are delayed. Olist should figure out why in some states the lateness frequency is high and find ways to improve their shipping solutions. In parallel, Olist should be less optimistic on their

estimated delivery days.



(a) Late delivery rate

(b) State review score and delay frequency

Figure 10: Per state basis statistics

We have discarded the product category from the dataset, but it is interesting to see its impact on the average review score. We can see in figure (18) that the average review score can varies depending on the product category. Note that the worst rated product category; seguros e servicos, which is basically selling insurances, has only been sold 2 times, so we should avoid early conclusions on these under represented categories. Nevertheless, this figure gives insights to Olist on the categories that are probably related to bad quality products to take action towards sellers who sell these goods.

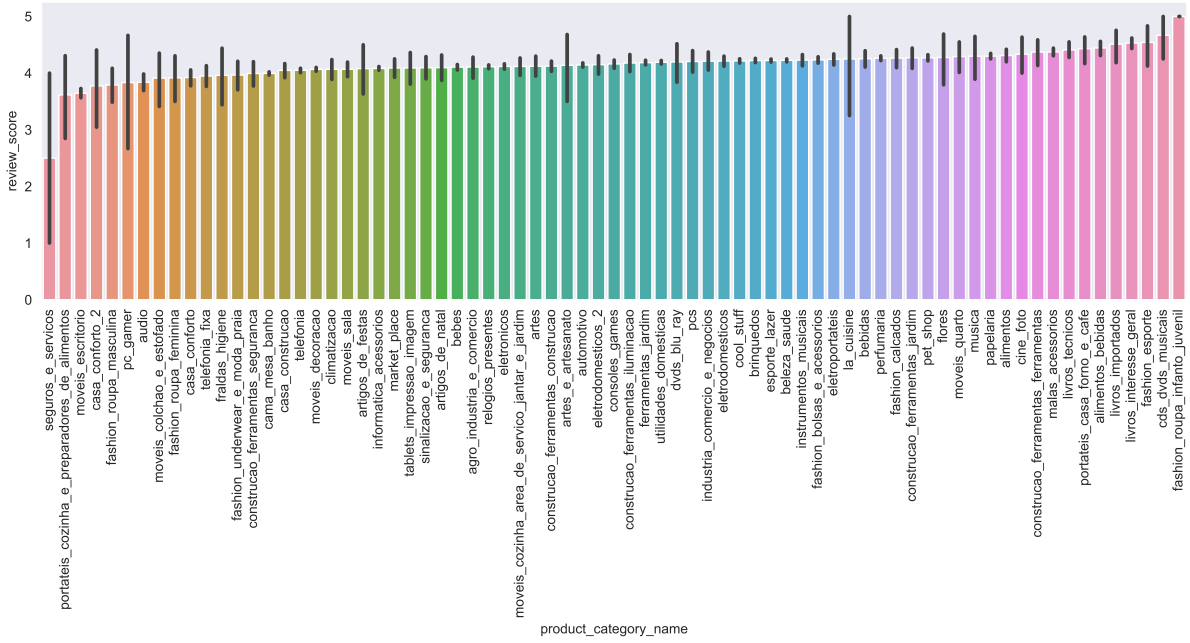


Figure 11: Average review score w.r.t. product category

We have developed an extra trees and a random forest classifier to rank the importance of each feature. To simplify the classifying task, we used as a dependent variable the simplified review score so that the outcome is binary. We used the dataset that has the features of the correlation matrix (c.f. figure (6)) and split it in a train and test set. Moreover, we rebalanced the data using random undersampling. It turned out that the results are not satisfying. In particular, the classifiers both retrieve inaccurate predictions and a very bad recall for the bad score class $\approx 27\%$. Thus, we cannot really interpret the importance of the feature according to classifiers ranking. We attribute the bad performance of the classifiers mainly because of two factors:

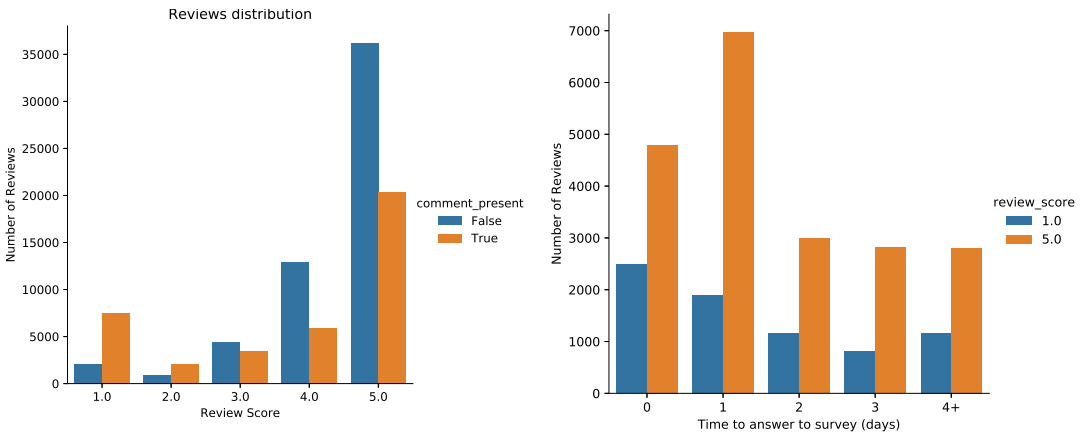
- Late delivery frequency is very low, so it is not seen as a relevant decision rule in the decision trees.
- We are missing features that capture causes of dissatisfaction (e.g. product quality, incomplete order, defective product).

To capture the fact that the unobserved factors play a role into the review score, we propose an approach based on review comments.

3.3 Sentiment analysis

We have seen that the data set features are note sufficient to explain customers satisfaction, the reason is that there are unobserved factors. Therefore, we will try to capture these unobserved drivers trough an approach based on review comments. The core idea is to capture customers sentiments by determining the main issues or positive points that appears in review comments. In prior, we show interesting statistics on customers reviews comments.
In figure (12a), we see that 30% of the customers leave a comment. Moreover, clients are more likely to leave a comment when they are not satisfied than when they are not. However, when someone is happy with his order, we imagine he has other things to do than explaining why he is so satisfied with his order. In the same idea, we see in figure (13) that happy customers write in general short review comments. Similarly, unsatisfied customers also write short comments, where they perhaps just explain briefly reasons of their dissatisfaction. However, there is a peak in the histogram for long comments where the unhappy customer is probably trying to explain all the misadventure of its order dissatisfaction. Interestingly, unsatisfied customers are answering faster to the survey than happy ones as we can see in figure (12b). With this glimpse of data visualization, we have clearly seen behavior differences toward the review comment process between satisfied and unsatisfied customers:

- Frustrated customers are more likely to leave a comment and when they do so, to send the review survey quickly
- Happy customers are more likely to not leave a comment but when they do so, they tend to write very short comments.

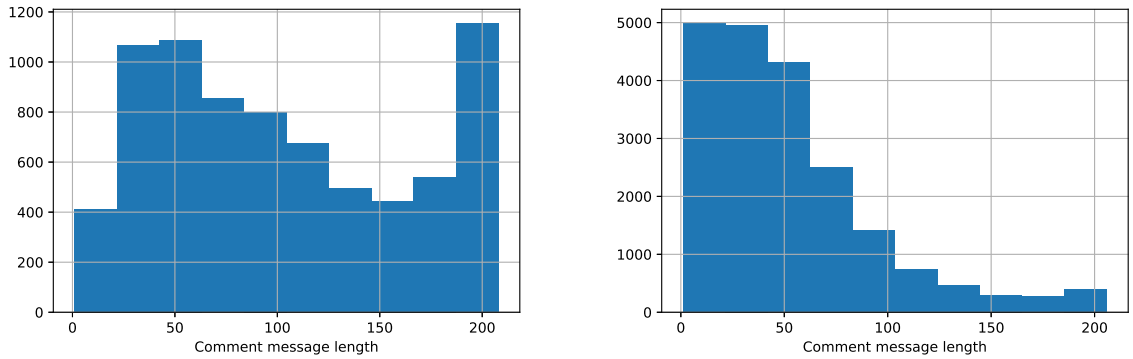


(a) Review comment counts

(b) Time to answer the survey (with comment)

Figure 12: Review comment statistics depending on customer satisfaction

Let's dig now deeper into the content of these comments. For the purpose of this analysis, we discard customers who did not write a review comment and customers whose review score is 2, 3 or 4. Indeed, the most relevant for Olist is to determine the causes of total dissatisfaction. Similarly, understanding full satisfaction is also the second essential point. From now on, we call good reviews the ones whose score is 5 and bad reviews the ones whose score is 1.



(a) review score = 1

(b) review score = 5

Figure 13: Review comment length distribution, maximum 200 characters

The most natural way to determine the causes of bad (or good) reviews is to look at the frequency of repeated sequences of words in the review comments. More precisely, we use Natural language processing to rank the 20 most common sequences of three words (3-gram) in the bad and the good review comments. We choose 3-gram over uni-gram or bi-gram because they can highlight the particular sentiment instead of just pointing out the underlying topic. For the sake of clarity, we translated the Portuguese sequences in English after the processing, this is why some 3-gram in Portuguese may have a different number of words in English.
From figure (12a), good reviews are frequently associated with in advance deliveries. What also causes a good review is the product quality, which appears in the 8th most frequent 3-gram in positive comments.
In the bad comments side, poor reviews are in most cases caused by surveys received and filled before deliveries, which is by far the most frequent matter in negative comments. The remaining causes of bad reviews are bad quality products, defected products, incomplete orders and wrong product shipped.
Hence, Olist should really put an effort in improving their shipping solutions to reduce delayed orders. In addition, they also have to change their survey policy by sending it after the order delivery. Also on the logistic side, they must improve the communication with the merchants since some customers complain for incomplete orders and wrong products. These order content problems are not necessarily Olist's fault, but rather related to sellers' negligence. Moreover, product anomalies or quality issues can be attributed to sellers, this is why we need to conduct a seller analysis to spot who are the sellers that contribute to lowering the review scores.

	positive_comments	negative_comments
0	arrived before deadline	I did not recieve the product
1	well before term	I have not received it yet.
2	delivered before term	product not delivered
3	delivery before term	product has not arrived
4	product arrived before	product not yet
5	arrived well before	it did not arrive yet
6	product delivered before	I did not receive the moment
7	Super fast delivery	not yet delivered
8	great quality product	I have not received it now
9	ahead of schedule	Related searches
10	before term product	I received a product
11	before expected date	I want money back
12	I received before term	I did not receive merchandise
13	before long recommend	product came defect
14	before stipulated deadline	I got no answer
15	fast delivery product	I received product yet
16	delivered within	I did not receive it today.
17	good quality product	I got the wrong product
18	product delivered deadline	I bought two products
19	product arrived well	product came wrong

Figure 14: Most common sequence of three words for good reviews (i.e. review score = 5) and bad reviews (i.e. review score = 1)

4 Sellers Analysis

During the previous section, we identified factors that may conduct an order to be rated positively or negatively. Even though the average satisfaction order review is quite high, i.e, average score 4.1, Olist's managers might be interested in raising those numbers. In this section, we will run an analysis on the sellers to spot which are more valuables than others, what are the main issues, and which actions can be taken.

4.1 RFM Analysis

In order to distiguish between valuables and invaluables sellers, we will use a RFM analysis on the sellers. In our particular case, RFM stands for :

- Recency : How recently did the seller sell an item?
- Frequency : How often does the seller sell items ?
- Monetary Value : How much does the seller bring in revenue ?

To calculate recency, we will look into the invoice dates. We will consider the last invoice as the most recent one and will subtract to it the most recent invoice date for each seller. The observed orders period ranges from 2016-10-04 at 09:43:32 to 2018-08-29 at 15:10:26.
As for frequency, we'll sum the number of invoices for each sellers.

seller_id	recency	monetary	frequency	refs	f_quartile	m_quartile	r_quartile	RFMScore
4869f7a5dfa277a7dca6462dcf3b52b2	4	247634.47	1115	192	5	5	5	555
4a3ca9315b744ce9f8e9374361493884	2	232423.08	1717	115	5	5	5	555
53243585a1d6dc2643021fd1853d8905	7	230797.02	348	431	5	5	5	555
fa1c13f2614d7b5c4749cbc52fecda94	0	200044.11	571	26	5	5	5	555
da8622b14eb17ae2831f4ac5b9dab84a	0	182482.13	1278	25	5	5	5	555
1025f0e2d44d7041d6cf58b6550e0bfa	3	171541.24	892	154	5	5	5	555
7a67c85e85bb2ce8582c35f2203ad736	7	160720.88	1131	426	5	5	5	555
955fee9216a65b617aa5c0531780ce60	0	156609.41	1254	20	5	5	5	555
6560211a19b47992c3666cc44a7e94c0	0	148994.13	1804	19	5	5	5	555
1f50f920176fa81dab994f9023523100	1	142022.61	1379	71	5	5	5	555

Figure 15: Sellers RFM table : top 10 Sellers

seller_id	recency	monetary	frequency	refs	f_quartile	m_quartile	r_quartile	RFMScore
77128dec4bec4878c37ab7d6169d6f26	531	15.22	1	2848	1	1	1	111
702835e4b785b67a084280efca355756	561	18.56	1	2890	1	1	1	111
ad14615bdd492b01b0d97922e87cb87f	564	19.21	1	2896	1	1	1	111
7ab0dd5487bab2dc835337b244f689fb	510	23.46	1	2824	1	1	1	111
c18309219e789960add0b2255ca4b091	267	24.00	1	2383	1	1	1	111
5b92bfa4120daa27c574daa2e386c693	555	24.96	1	2881	1	1	1	111
20d53aad4fe5ee93a64f8839609d3586	276	26.98	1	2417	1	1	1	111
9e25199f6ef7e7c347120ff175652c3b	520	27.02	1	2834	1	1	1	111
0f94588695d71662beec8d883ffacf09	546	27.59	1	2870	1	1	1	111
bee36b4f9a2b9fdcaff6ec05df202ed0	560	27.86	1	2888	1	1	1	111

Figure 16: Sellers RFM table : bottom 10 Sellers

Finally, for the monetary value, we will sum the total order price for each seller.
Our approach to RFM is to assign a score for each dimension on a scale from 1 to 5. To assign scores, we order the data by frequency in descending mode. Then, we divide the number of sellers on the data set (filtered to keep unique orders only) by 5. We assign a number from 5 to 1 to each one of the previously created segments. We apply the same strategy for monetary and recency too. However, for the recency dimension, since the most recent orders are better, we order them in ascending mode.
The maximum score represents the preferred behavior. Finally, we concatenate the three scores, labelling the perfect behaviour being ”555” denoting a top score in each dimension and the undesired behaviour being ”111” denoting the opposite.
Furthermore, we managed to extract valuable statistics for Olist’s managers, using RFM scores. Namely, out of the 2958 sellers observed (filtered out of the 4661), we obtained 229 top sellers, i.e, with a high score on each dimension, and 169 inactive and cheap sellers, i.e, sellers with lowest score on each dimension. We summarize some results into two tables.

4.2 Sellers and Customers Reviews

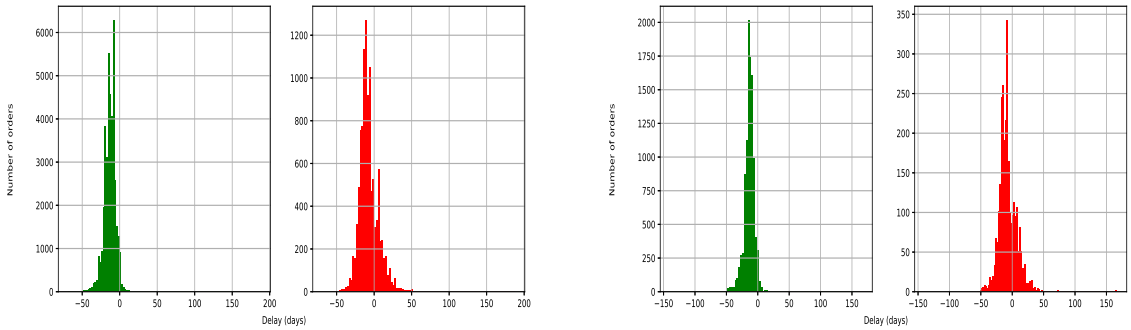
We want to know whether our RFM analysis performed on the sellers do reflect the discussed topics during Customer Sentiment Analysis.
In order to do this, we will consider separately the best (rfm score = 555) and worst sellers (rfm score $\leq$ 333) from the RFM analysis while confronting good reviews (score = 5) and bad reviews (score $\leq$ 3) in order to point out problems discussed in the previous section, namely bad quality products and late deliveries.

4.2.1 Do worst sellers sell lower quality product than best sellers ?

We will use the tri-grams generated during customer sentiment analysis pointing out bad quality products and try to see in which proportion they appear in each subset.
We analyzed 20'503 commented reviews among the best sellers, and only 0.1% of them contained comments mentioning bad products, while out of the 5'478 commented reviews from worst sellers, 0.15% mentioned them. The numbers are quite similar in both cases, hence we might want to conclude that low quality products are not necessarily recurrent for worst sellers. Furthermore, the numbers being quite low, this issue is not significant enough to be a major issue for bad reviews.

4.2.2 Is delivery delay an important element in determining whether a seller is good or not ?

We used the same methodology as the one with the bad products, i.e, we will use the tri-grams generated during customer sentiment analysis pointing out late deliveries issues and try to see in which proportion they appear in each subset. Surprisingly, exactly 0.18% of them were mentioned in each subset out of 20'503 and 5'478 for best and worst sellers respectively. Hence, it seems that delivery delay does not allow us to distinguish between good and bad sellers.
Although we obtained similar results, we wanted to find whether delivery delay is an deciding factor for customer review. To do this, within each subset, we looked at the distribution of delivery delays confronting orders which received good reviews against bad ones.



(a) Best Sellers : good and bad reviews

(b) Worst Sellers : good and bad reviews

Figure 17: Distributions of late deliveries in days for best and worst sellers

Although the distributions are quite similar between best and worst sellers, we notice, when looking at the plots above, that within best and worst sellers, late delivery is actually an important element making the decision for an order review score since good reviews tends to be orders delivered mostly in advance while bad reviews come order issued both in advance and very late.

4.3 Spotting unreliable sellers !

While during the first part we identified best and worst sellers, the rfm analysis did not take into account customer sentiment analysis. Using our previous conclusion about this, one might

be interested to further seller analysis by spotting the sellers who often delivers orders late. This can be used as a driver for stakeholders to either warn best sellers who gets too confident about their position and popularity at the expense of Olist platform popularity when not taking into account customer reviews, or by discarding those bad sellers who do not bring that much value to the company and at the same time tarnish the reputation of the platform.
In order to spot them, we separated best sellers from worst ones, grouped the orders by sellers, and we computed for each of them their average delivery delay. We found that among the best sellers, none of them have a positive average delivery delay while among the worst ones, we found 74 of them.
The table below gives, among the worst sellers, the bottom ten with the highest average delivery delay.

seller_id	late_delivery
df683dfda87bf71ac3fc63063fba369d	167.000000
a154d7316f158bb42e6fa18bbe3afd3a	42.000000
e09887ca8c7bf8a4621ce481820414ef	41.000000
8e670472e453ba34a379331513d6aab1	35.000000
4fb41dff7c50136976d1a5cf004a42e2	33.000000
8629a7efec1aab257e58cda559f03ba7	33.000000
6f1a1263039c76e68f40a8e536b1da6a	31.000000
eebb3372362aa9a46975164bed19a7e7	27.000000
391bbd13b6452244774beff1824006ed	24.000000
9b522ba7eae9e1d04082f267144583cc	23.666667

Figure 18: Bottom 10 sellers with highest average delivery delay

5 Sales Analysis

In this section, we will have a look at the sales in a more general manner: we will not distinguish between sellers but will rather group them all together and analyse the different categories (of bought object) in sales, the different sales period(s) in the year for each category and the different traffic sources that are used in order to attract customers. To do so, we will use 2 datasets: products.csv and a processed dataset: orders.csv.

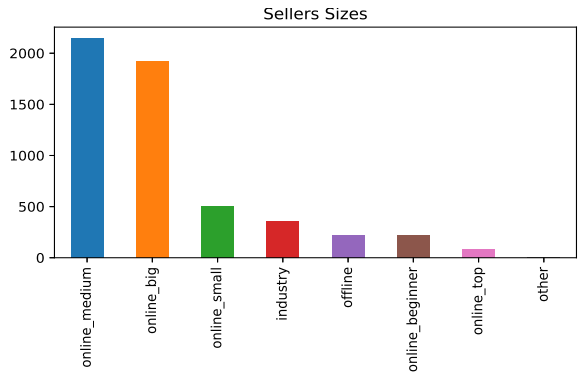
5.1 Sales Decomposition

In this section, we will provide preliminary information about the sellers and the category of merchandises that they sell.
The sellers on Olist are decomposed of 2 different categories: manufacturers and resellers.
There are 73 different business segments (categories of objects sold) within all sellers such as: 'housewares', 'perfumery', 'auto', 'pet_shop', 'stationery','furniture_decor', 'office_furniture', 'garden_tools','computers_accessories', 'bed_bath_table', 'toys', 'construction_tools_construction', 'tele-

phony', 'health_beauty','electronics', 'baby', etc.
The 7 largest categories (with the biggest number of orders) are:
'watches_gifts', 'housewares', 'furniture_decor', 'computers_accessories', 'sports_leisure', 'health_beauty', 'bed_bath_table', Which represent by them alone, about 51% of all the orders.

business_type	manufacturer	other	reseller	All
business_segment				
All	829	3	4661	5493
health_beauty	58	0	810	868
watches	0	0	599	599
household_utilities	132	0	456	588
home_decor	168	0	344	512
construction_tools_house_garden	63	0	292	355
audio_video_electronics	66	0	242	308
pet	29	0	270	299
car_accessories	49	0	162	211
bed_bath_table	40	0	160	200
sports_leisure	15	0	173	188

(a) Top 10 business segments



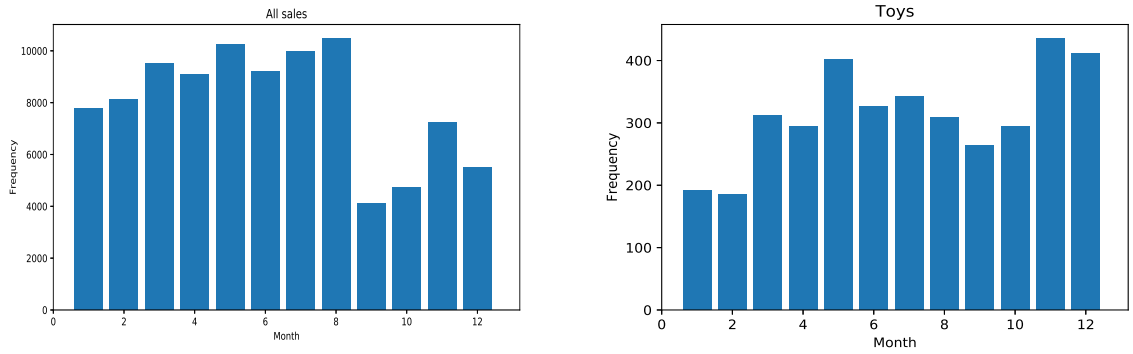
(b) Sizes of manufacturers and resellers combined

Figure 19: Sellers' Decomposition

As it can be seen on figure (19), all segments combined, there are a total of 829 manufacturers and 4661 resellers. We can also see that most of the sellers (approximately 75%) are large and medium sized. However for the further analysis below we will group the manufacturers and reseller together and treat them as a whole unit of sellers.

5.2 Sales Period Analysis

In this section we will analyse the sales periods for each category. As we don't really know the marketing strategy of Olist, we suppose that they have a constant marketing effort for each category of objects during the year. That is for example a user can step upon an advertisement from Olist with uniform probability no matter which date of the year.
The goal here is to see if a category has naturally more sales in some periods of the year. Knowing so Olist could increase their advertisements and marketing effort in general in those periods of the year in order to make even more sales.
To do so, we took the purchase_timestamp in the orders.csv dataset extracted the month when each order was made. We then made a histogram for the 12 months in the year, first, grouping all the sales together and second, making a histogram for each category.
One would naturally think that giant e-commerce businesses such as Olist have most of their sales in the month of December because of Christmas (Christianity being the largest religion in Brazil). This intuition could be right for some categories of objects such as toys but, as it can be seen on the figure (21) this is not necessarily the case in all segments. Note: the number on the x axis corresponds to the month (1 being January and 12 being December).



(a) All segments sales

(b) Toys sales

Figure 20: Sales Histograms

Generally, we can observe a trend in sales: The months with the most sales are May, July and August and there is a drop of sales from August to September. We don't have the causal explanation for this drop as it can be related to several socio-economic factors.

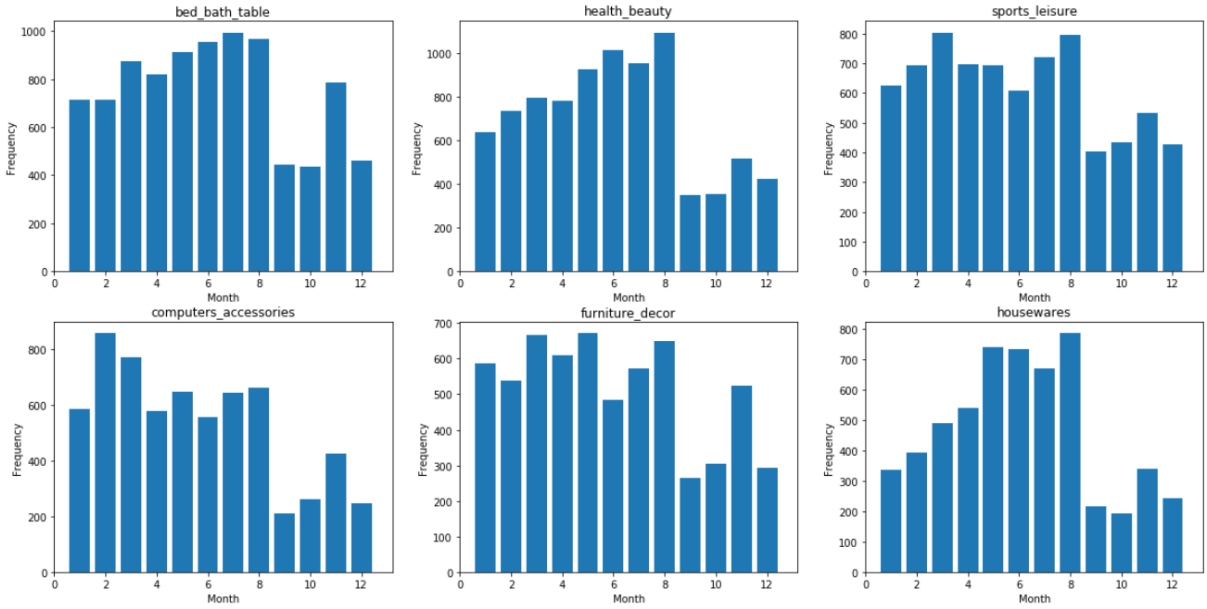


Figure 21: Sales of the 6 biggest segments

Finally we would recommend Olist to implement a personalized advertising/marketing strategy for each segment (category of objects sold) in order to increase their sales. If that is not possible because it would cost too much for the company or if Olist prefers to have a common strategy for all of the items combined, then we would recommend to increase their marketing effort in the first and especially the second and the beginning of the third quarter of the year. They then lay back down until the month of November, when they could again increase their advertisements especially for Christmas gifts.

5.3 Traffic Sources Analysis

The main goal here is to provide Olist with insight on their traffic sources such that they can put more effort in the traffic sources that have a large click-to-conversion rate, meaning that the customer that came up across a product via that traffic source (email for example) would actually buy the product, and maybe discard traffic sources that don't really work and have low click-to-conversion rates.
To do so, we used the products.csv dataset where each line represents a potential customer, the column *origin* represents the traffic source upon which the customer came across the seller's page and the binary column *Closed* specifies whether the customer later bought the object or not. There are 10 different categories of traffic sources in the dataset:

- **Paid search**: Traffic from a search engine results that is the result of paid advertising (via Google AdWords for example)
- **Organic search**: Traffic from a natural search engine search's results list
- **Social**: Traffic from a social network, such as Facebook, LinkedIn, Twitter, etc.
- **Email**: Traffic from email marketing
- **Referral**: Traffic from a site other than a major search engine
- **Direct traffic**: Any traffic where the referrer or source is unknown
- **Display**: Traffic generated from a banner or flash ads
- **Other publicities**
- **Other**
- **Unknown**

There were also 67 NaN values which we dropped as they were only an insignificant portion of all the data and we weren't sure which category they could be included in. We also preferred to keep the *other* and *other publicities* separated as we though that the other category might include factors that aren't direct publicities from Olist.

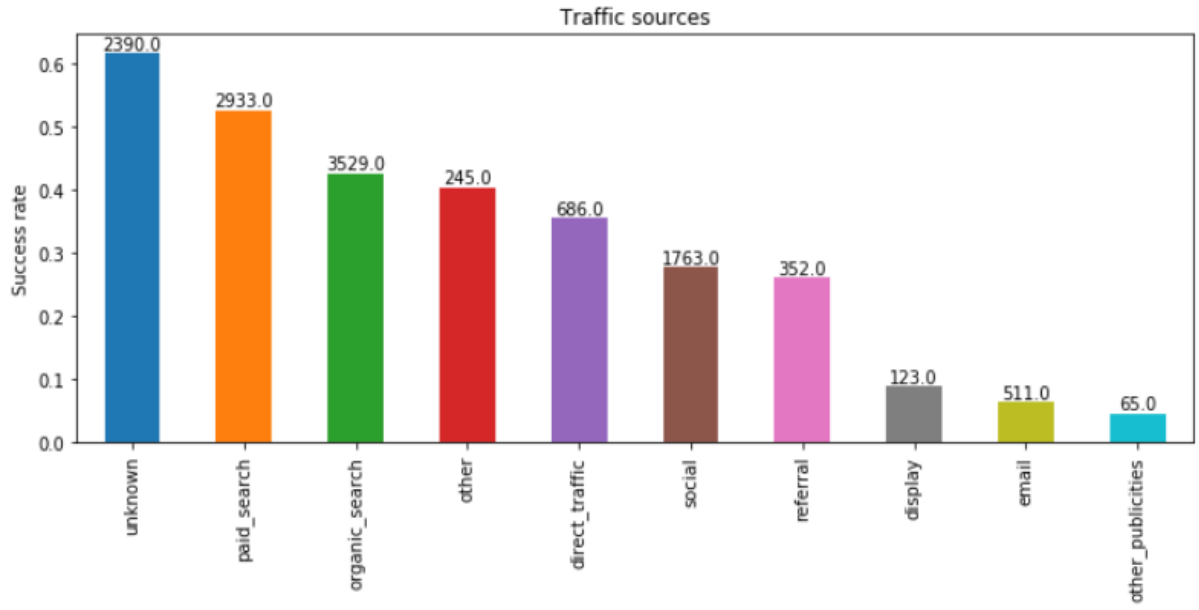


Figure 22: Click-to-conversion rates of the traffic sources

The figure (22) shows the obtained results, the success rate being the click-to-conversion rate calculate as the total number of closed cases divided by the total number of cases for each traffic source. The total number of customers that used each traffic source is displayed on the top of each bar.

We can see that the most successful traffic source is *unknown* which unfortunately doesn't tell us anything useful. The second best traffic source is the paid search with a 0.52 click-to-conversion rate with means that over 1 person out of 2 who steps upon a paid advertisement in his search results, actually ends up buying the product. This also demonstrates the effectiveness of Google and other search engines in proposing relevant and targeted search results to the user.

Finally we would recommend Olist to rather focus their efforts on paid search and direct_traffic which seem to be the most prominent traffic sources. We would also recommend them to stop email marketing as they have a very low click-to-conversion rate with a rather medium traffic flow: compared to direct_traffic for example, the have similar traffic flows (686 people for direct traffic and 511 for email) but the direct_traffic has a click-to-conversion rate that is 4 times higher.

6 Conclusion

6.1 Data driven observations

The goal of this project was to analyse Olist's business using data analytic to find the high-performing sectors to value and those that need to be improved. In order to conduct this research, we dived into three main areas, with the following observations :

- Customer Satisfaction : We mainly used basic Natural Language processing techniques to determine areas of disrepair requiring improvement and those for which we should

keep putting emphasis on. Those were highlighted by customers directly on the surveys through review comments. We came to the conclusion that mostly, customers were happy when the order was shipped fast, i.e, before the estimated delivery date and the received product quality matched their expectations online. Likewise, they did not hesitate to show their discontent when it was not the case.

- Sellers Analysis : We performed a RFM analysis on Olist's sellers to segment their value. Based on their money they bring, the frequency and the recency of their sales, we could identify most valuables and invaluables ones. On top of this analysis, we discovered that delivering bad quality product is equally a concern of both best and worst sellers, while their presence are not significant at all. As for late deliveries, they arise more often but are equally distributed among bad and good sellers. However, we discovered that late deliveries are actually a relevant factor deciding whether or not an order will be rated positively independently on the quality of the sellers. Using this, we managed to spot those sellers which at the same time were labelled as bad ones by RFM analysis and contributed to tarnish Olist's reputation by often delivering late orders.

- Sales and Advertisement Analysis : After proper identifcation of the different categories of items sold through Olist's platform, we noticed that some categories tends to have an increase in sales depending on the period of the year. We found that traffic sources with the largest click-to-conversion-rate are paid searches, with 1 out of 2 person who steps upon a paid advertisement in his search results actually ends up buying the product, while email marketing was found to be quite inefficient.

6.2 Business Strategies Insights

Our observations should serve as basis for Olist's managers to actually make decision to increase the business value. We suggest the following directions :

- Customer Satisfaction : The average review score being quite high already, Olist should work on customer retention to ensure product quality, for example, by having a charter of integrity signed by the sellers, penalizing them and offering customer benefits in the event of a problem that is the responsibility of the seller. Furthermore, given the large size of Brazil, it is not surprising to have issues with late deliveries. One might be interested to further our analysis to try to spot geographical locations of late deliveries. Hence, if this analysis come conclusive, Olist managers could develop relay points to improve the fast delivery service.

- Seller Analysis : One might be interested to have insights from a Marketing Specialist from Olist to understand how are managed Olist's Sellers on the platform. During our analysis, we managed to identify the VIP sellers, hence, in a similar manner as a good spot in a commercial center or a Page rank score for webpages, it might be interesting to give better exposure on Olist's platform to those best sellers since they tend to be the one the customers prefer in term of popularity and the ones who bring the most monetary value to the company. Similarly, Stakeholders might take actions against those bad sellers revealed by our analysis, to either put pressure on them to make a marketing effort to match Olist's expectations or to discard them from the platform to prevent them from tarnishing the company and give the spot to potential sellers who might be interested in

joining the platform. In a future work, one would be interested to develop an automated predictive tool using state-of-the-art Machine Learning techniques to determines what are the drivers that make a good seller.

- Sales and Advertisement Analysis : Similarly, while we mentioned giving better exposure to some sellers on the platform, it might be interesting to target fashionable items, depending on the period of the yea,r and put them on the front page to give them better exposure, set up with the marketing team a promotional strategy on these items, to increase sales, hence profit. Furthermore, the company should stop investing on advertisement strategies which arise low click-to-conversion-rate, like email, and put a financial effort on most effective ones like paid searches to drive customers to the platform and also increase sales and profit.

References

[1] *Olist official website:* https://olist.com/

[2] *Olist public dataset on Kaggle:*https://www.kaggle.com/olistbr/brazilian-ecommerce

[3] *Random Forest algorithm:* https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[4] *Predicting Customer Satisfaction:* https://www.kaggle.com/andresionek/predicting-customer-satisfaction

[5] *RFM Analysis* https://towardsdatascience.com/apply-rfm-principles-to-cluster-customers-with-k-means-fef9bcc9ab16

[6] *Christopher Bruffaerts: slides for the course Data Science In Practice, EPFL, 2019* https://moodle.epfl.ch/pluginfile.php/2644872/mod_resource/content/1/Sales_Marketing_Analytics_2019.html