Sarah Hellmann, Eleanor Pulsipher, Isabel de Obaldia, Ryan Boden
Professor Mallipeddi
Business Analytics Practicum
23 November 2020

E-commerce Dataset Report

1. Background

1.1. About the dataset

The Ecommerce dataset is a public dataset pulled from the Kaggle website (kaggle.com) about orders made from the Brazillian firm: Olist Store. The data contains randomly selected information of 100,000 orders from 2016-2018 made at multiple marketplaces in Brazil. The dataset was provided in a csv and dta format, which we uploaded to R in order to perform our analysis and derive recommendations to the firm. There are 73 variables provided to view and order from multiple dimensions. We focused on the order status, price, payment, product category, customer location, customer reviews, and shipping details variables.

1.2. About the firm

Olist was founded in 2015 and manages an online e-commerce platform that connects merchants to consumers from all over Brazil. Olist is the largest department store in Brazilian marketplaces and operates much like Amazon in the United States. Through Olist, merchants can register their products to be sold on the online Olist Store. Then, when a customer purchases a product from the store, the seller fulfills that order. Orders are then shipped to customers using Olist logistics partners. Lastly, the customer can fill out a satisfaction survey for their order once they receive the product (or once the estimated delivery date is due).

2. Interesting finds

2.1. Sao Paolo State Clear Outlier in Orders Across Product Categories

One of the first things we wanted to explore within the dataset was the orders by product category. When exploring the dataset, we looked at the product categories broken down by State. Analyzing this data through Tableau bar chart visualization, the Sao Paulo state (SP) purchases the most products across all 6 product categories (Auto and Industrial, Books, Electronics, Fashion, Home Goods, and Office). Over 40% of all purchases across all product categories is from the Sao Paulo state alone. Compared to the state driving the second most number of orders, Rio de Janeiro, Sao Paulo purchases roughly 68% more products than Rio de Janeiro. After researching more about the Sao Paulo state, we found that it is the wealthiest state in Brazil and that its capital, the city of Sao Paulo, is the most populous in Brazil too. This information explains why 40%+ of all orders across all product categories come from the Sao Paulo state.

2.2. Disparity Between Satisfaction Score and Product Popularity

As we continued to explore the different products that Olist had to offer, we decided to filter the top 10 with the highest demand and popularity among consumers to see if we could find an interesting insight. Once we had retrieved the top 10 products with the highest order amounts in the dataset, we were interested in finding out their performance among consumers by looking at their average satisfaction scores. As we plotted the relationship between product popularity and overall satisfaction scores, we noticed that the most popular product was also the one with the lowest rating in the group. With a total order amount of 9330, Cama Mesa de Banho received an average satisfaction score of 3.96. Meanwhile, Brinquedos, with an order amount of 3869, almost 60% less than Cama Mesa Banho's demand, received an average rating of 4.17. This can be attributed to the fact that products that are in greater demand and generate more output in the

market are more susceptible to suffering damages via the supply chain or manufacturing phase. For this reason, products like Cama Mesa Banho are more likely to be delivered in unsuitable conditions to the hands of the customer. It is true that there are many other different factors that can affect the overall rating of a product, but this disparity between satisfaction scores and product popularity definitely caught our eye when exploring the dataset.

2.3. Average Review Scores Across Products

To further analyze the performance of Olist's products among the customer base, we decided to divide the orders by their average review score and calculate their percentages. By doing this, we found out that 57.40% of Olist's products received a review score of 5, the highest possible rating, and approximately 20% of the orders received the second highest possible score. It is a good sign to see that around 77% of the company's products satisfied consumers' needs. However, to be well-informed and have a clear picture of the actual state of the company, we must also look at the other side of the spectrum. Our results show that 11.47% of Olist's products received a satisfaction score of 1. Having almost 12% of your products receive the lowest possible rating is a definite concern and a factor the company must be aware of.

2.4. Most Popular Payment Methods Among Customers

To gain more knowledge on Olist's customer preferences and habits, we decided to explore payment information to figure out what was the preferred and most common payment method among consumers. First, we retrieved the amount of orders received from each payment method to figure out the most popular one. Then, we proceeded to calculate the average order revenue by payment method. Our results showed that credit cards were the most popular method accounting for 74479 orders, approximately 75% of all Olist's total order amount. Credit cards were also the payment method which was used to conduct transactions of higher order value. Right behind credit cards in both results were Brazilian boletos. Boletos are tickets regulated by the Brazilian Federation of Banks typically used by people who don't have a bank account nor a credit card. In Brazil, 200 million people do not have credit cards, and 55 million don't have bank accounts, and for many this is the only payment method to use online. The popularity and importance of this payment method is relevant information Olist should keep in mind about its customer base moving forward.

3. Recommendation 1: Reach Out to More Sellers in High Population Areas

Based on our findings, we recommend that Olist should increase its number of sellers in areas with high population to maximize its profits. This recommendation is based on one particular finding that caught our eye when analyzing the top 10 states that generated the highest level of demand. When studying our list of results, we had the idea that the number of orders per state would be directly correlated to the size of its population. This seemed to be the case for most of the states on the list, however, when we got to analyzing the population size of the last three states, we found some discrepancies. Distrito Federal (DF) is the seventh state that generates the highest orders for Olist, followed by the states of Espirito Santo (ES) and Gorias (GO). However, DF has a population size of 3 million people which is less than half of GO's population of 7 million and 1 million less than that of ES. This led us to ask ourselves the reason why DF's order amount exceeded those generated by these two high population states which have a bigger market size. To answer this question, we proceeded to analyze the number of sellers in each area. Our results showed that the number of sellers of DF was two times greater than the number of sellers operating in ES and GO. To maximize its reach and not miss out on potential profit, Olist

should consider increasing the amount of sellers in high population areas to gain more revenues, popularity and exposure. See Appendix 1 and R code for Recommendation Derivation.

4. Recommendation 2: Diversify Sources of Income

To get to understand Olist's business model better, we chose to analyze the different product categories that this ecommerce site offers. Olist offers a wide range of products that fall into the following 6 categories: Auto Industrial, Fashion, Office, Electronics, Books and Home Goods. To get a sense of the individual market performance of each category, we proceeded to calculate an approximation of the total amount of revenue generated by each group. Our results showed that the Home Goods category generated around $6.4 million dollars of total revenue, accounting for over 50% of Olist's total revenues. We also found that less than 5% of revenues come from the Books category. To lower risk and increase profit in the long term, we believe that Olist would greatly benefit from diversifying its sources of income to not be heavily dependent on just one product category and to not waste resources on underperforming categories. Additionally, we also examined the order amounts per category and our results revealed that just like for the revenue, Home Goods represented a significantly high amount of the company's total orders, almost accounting for 90% of Olist's total orders. If the company continues to rely heavily on just one market segment, it may be detrimental in the long term if that market underperforms. See Appendix 2 and R code for Recommendation Derivation.

5. Recommendation 3: Improve Shipping Estimations and Logistical Infrastructure in the Northeast Region

In our analysis of the dataset, we discovered that the shipping estimates Olist provides to their customers are pervasively inaccurate. The average customer receives their order 12 days before the firm's anticipated delivery date, but there is a huge level of variation within this measure; the standard deviation of 10.8 shows a high degree of spread, and the earliest arrival in the dataset was fulfilled 147 days ahead of schedule while the latest was delivered 188 days after the firm's estimate. This level of variation is troubling, and suggests that Olist's methods of estimating transportation time are flawed in some way.

To better understand the factors that influence this disparity in arrival dates, we used statistical testing to explore the relationship between the length of shipping disparity and the other variables in the dataset. We found that the factor that was most negatively correlated with an order arriving ahead of schedule was the customer living in one of the following states: Algoas, Maranhao, Ceara, Piaui, Roraima, and Sergipe. When all other factors were held constant, customers from each of these states received their orders over a week later than customers in the rest of Brazil. Through geographical and contextual research, we found that all of these states share a significant commonality: they are located in the northeast region of the country.

The northeast region is relatively rural; it is geographically isolated by the Brazillian Highlands and the Amazon and Atlantic rainforests, and one of the poorest regions in the country. Because of this, one might expect that customers in this region would have low service expectations and would not be upset by shipping delays. However, we also found that customers in these states have the lowest average satisfaction scores in the country. Thus, it is clear that customers in this region are underserved and less likely to be satisfied with their Olist purchases. These customers currently make up a lower proportion of Olist's client base, but these states have a combined population that is close to 25 million, representing a significant growth opportunity for the firm.

We recommend that the firm improve service in this region, and also improve the accuracy of their shipping estimates in general. See Appendix 3 and R code for Recommendation Derivation.

6. Recommendation 4: Re-evaluate the Distribution of Vouchers

Vouchers are one of the four payment methods that are accepted by Olist, and they are similar to coupons in that they provide a discount to the customer on an item or total purchase. In traditional marketing, this type of promotion is generally expected to positively affect customer satisfaction and incentivize higher spending. However, in our analysis we found that customers who used vouchers as their main payment method had, on average, the lowest order values, ordered the least items per order, and had the lowest average satisfaction scores compared to customers who used any other payment type. Because providing these vouchers presumably comes at a price to either the individual seller or the firm itself, it is not likely that the results of this program are worth the associated costs. Vouchers do not positively impact revenue or customer satisfaction, so Olist should either reduce the total amount distributed or limit their availability to high-value customers who have demonstrated loyalty by making consistent purchases. See Appendix 4 and R code for Recommendation Derivation.

7. Recommendation 5: Focus Business Expansion Efforts to Sao Paulo (SP), Rio de Janeiro (RJ), and Minas Gerais (MG) Regions

We recommend that, in the long term, Olist should focus any business expansion efforts to the São Paulo, Rio de Janeiro, and Minas Gerais regions. As mentioned in Recommendation 2, Olist should consider diversifying their sources of income to rely less heavily on orders from one product category. So, in order for Olist to reduce their income's exposure to overconcentration they should add another product category. When adding another product category Olist should focus on marketing and building that product category into the São Paulo, Rio de Janeiro, and Minas Gerais regions. These regions perform well across all categories, so there is a lot of potential for a new product category to succeed in these regions. However, we recommend this business expansion in the long term because we do not have enough data about the firm's current financial state. Given this lack of financial information, we do not know if Olist is able to currently support new business ventures into other product categories. So, to account for the lack of financial information we recommend that Olist, in the short term, should focus on improving their shipping, selling, and marketing efforts in the regions (São Paulo, Rio de Janeiro, and Minas Gerais) and product category (Home Goods) that drive most of their business. We recommend that Olist should capitalize on their strong sources of income in the short term so they can be in a healthy financial state in the future to support our long term recommendation of business expansion. See Appendix 5 and R code for Recommendation Derivation.

8. Recommendation 6: Place Limits on Number of Payment Installments

Olist currently allows their customers who purchase items on layaway to choose the number of installments in their payment plans. These installments are made monthly, with the average customer choosing to pay off their purchases in 2 to 3 months, with a standard deviation of 2.7. Following the normal distribution, this means that 99.7% of customers select purchase plans that are less than 11 months. However, there are a number of outliers in the dataset, with some individuals taking up to 22 months to pay off their purchases. This lengthy installment option is atypical amongst comparable retailers and not utilized by many customers, yet introduces significant risk to the firm and its sellers. Thus, we wanted to assess if there was any

CRM-related benefit to keeping this option available. Through regression, we noticed that payment installments were significantly negatively correlated with satisfaction scores; customers who chose longer payment plans were actually less likely to be satisfied with their orders. Thus, it seems that offering this option does not improve customer sentiment. We recommend that Olist should limit the number of payment installments to under 12 months, as it would have no impact on the majority of consumers while significantly mitigating risk to the firm's accounts receivable. Installment plans are distinctively popular with Brazillian consumers and considered a crucial aspect of retailing in the country, with 80% of e-commerce orders in Brazil being paid in installments. Thus, optimizing the firm's installment policies could make a significant impact on the firm's revenue and minimize their liabilities. See Appendix 6 and R code for Recommendation Derivation.

9. Recommendation 7: Photo Quantity and Review Rating

Olist has products on their site with 0 to 20 images uploaded on the product page. The majority of products sold and in the dataset (49.4% of the transactions) have only one image. Further, a total of 1,419 products sold do not have any images at all. We recommend that Olist optimize their product pages to include a minimum of two photos. While more photos is better than none, we understand it may be unfeasible to recommend each product have the maximum of 20 photos. Further, many products may be small enough where 20 photos would be repetitive, and users would be seeing duplication in images. It is in Olist's best interest to include as many photos as would be reasonable for each product individually. At the very least, without seeing the exact products sold, it can be assumed two photos would be a reasonable minimum. This way, too, Olist would not be overwhelmed in optimizing all of their pages. In the long run, we would recommend optimizing products which see the highest volume in transactions. This would be those products with 1-6 images on their page (97% of all transactions in the dataset). We would recommend increasing all products to a minimum of 4-6 photos after each product has at least two images. This is a bigger task to accomplish, so a two photo minimum should be prioritized first. See Appendix 7 and R code for Recommendation Derivation.

Appendix

1. Recommendation 1 Derivation

This recommendation came from filtering the top 10 states with the highest order amounts, conducting outside research on the size of their populations and finally retrieving the number of sellers per state. To get the list of the ten states that generate most demand, we made use of the "pipes" in the dplyr package to filter, group by and summarize the count of orders per state and then arrange by descending order. Once we got our results, we proceeded to record the population sizes of each state from outside sources. To get the number of sellers per state, we implemented a similar code to the one described above but this time grouping by sellers. Once we had retrieved all of this data, we analyzed it and used it to drive and support our first recommendation of increasing the number of sellers in high population areas.

2. Recommendation 2 Derivation

In order to figure out the amount of revenue each category generated, we implemented a code using the dplyr package. We assigned each category a variable with its name, applied a filter per category and used the summarize function to get a count of the amount ordered and the total revenue for each. Once we applied this code to each category, we put together our results and examined them to drive our recommendation. After making sense of our final results, we proceeded to illustrate our findings using bar graphs in Tableau. The graphs we produced clearly demonstrate how Olist's revenues and demand depend heavily on just one particular product category. We believe that it is in Olist's best interest to diversify its risk by diversifying its sources of income moving forward.

3. Recommendation 3 Derivation

To quantify the disparity between the orders' estimated and actual dates of arrival, we first created a new variable using the mutate function in DPLYR called "Shipping Disparity," which subtracted an order's actual arrival date from its estimated date. This resulted in a new column which contained the number of days an order arrived ahead or behind of its estimate - with negative values representing orders that arrived behind schedule and positive values representing orders that arrived ahead of schedule. In our initial review of this variable, we used the summarise function to find the mean, median, maximum, minimum, standard deviation, and IQR of the shipping disparities. To better understand the factors that influenced shipping disparity, we first ran a linear regression using radiant with shipping disparity as the response variable and used our domain knowledge to select order price, freight value, customer state, distance, and product height, weight, length, and width as our independent variables. We found that orders coming from the states mentioned in section 5.1 were significantly negatively correlated with shipping disparity, each with p values less than 0.001. After we discovered these states were all located in the Northeast region, we used the if_else function within mutate to create a new variable called "Northeast" as a factor, which returns 1 if an order was placed in any of the states in this region and 0 if the order was placed in another region. To validate the results of this regression and examine its impact on customer satisfaction, we used the "compare means" function in Radiant to run 2 t-tests to find if there was statistically significant variation between the mean review scores and length of shipping disparity between customers in the Northeast and other regions. The t-tests validated the results of our regression, and found that with a 95% confidence level customers in the northeast were more likely to have lower mean satisfaction scores and lower mean shipping disparity (meaning orders arrived fewer days ahead of schedule) than customers in other regions, with p values of >0.001 and 0.023 respectively. This showed that the difference in mean shipping disparity between regions was somewhat significant, but had

a dramatic impact on review scores. To further visualize this relationship, we used Tableau to plot the mean values of shipping disparity and review scores per state, and saw that both values were indeed lowest in the Northeast.

4. **Recommendation 4 Derivation**

In our exploration of the dataset, we decided to analyze the effect of payment type on average order price, number of items in order, and satisfaction scores. Through cursory observation with the summarise function in DPLYR, we found that customers who used vouchers had the lowest mean values in all 3 of the above variables. To validate this initial observation, we ran 3 t-tests using the "compare means" function in radiant using main_payment_method as our explanatory variable and order_value, order_total_items, and average_review_scores as our respective response variables. For each, Voucher was the reference group, and other payment types were compared to it with the null hypothesis being that the average values for each payment type were the same, and the alternative being that the mean values for other payments were greater than the mean values for voucher orders. At a 95% confidence level, we found that voucher payment is significantly statistically correlated with lower mean values for both order value and total items compared to any other payment method, with p values all falling below 0.001. In the t-test comparing average review scores per payment method, we found that voucher orders had a somewhat significantly lower mean score than debit card orders (with a p value of 0.002), but the correlation was not as strong as the relationship between voucher payment and order value and number of items. However, due to the voucher customer group having the lowest mean satisfaction scores numerically, we would still consider this to be relevant information for the firm to explore further.

5. **Recommendation 5 Derivation**

This recommendation came from our cluster analysis of data pulled from Olist's orders grouped by state and product category. To get this data, we focused on the variables of the 6 categories (Auto and Industrial, Books, Electronics, Fashion, Home Goods, and Office) grouped by State. Using the dplyr package, we were able to use the "pipes" to group by, filter, and count the order data per State for the categories D1-D6 columns. Then, we saved this filtered data as a new dataframe to explore visualizations through Tableau. Further, we used the filtered data to conduct a cluster analysis on order data counts per state separated by product category. This cluster analysis divided the States into 5 groups. After analyzing these 5 groups, it becomes clear that most of Olist's orders come from the regions in groups 4 and 5 (São Paulo, Rio de Janeiro, and Mina Gerais) the Home Goods product category. Most of the population and the wealth in Brazil can be found in these three states, which explains the high proportion of orders coming from these regions. Nearly 50% of all orders belong to the Home Goods category. Specifically, groups 4 and 5's Home Goods category product orders account for 65% of the total Home Goods product category orders. So, these regions and the Home Goods product category are of great importance to Olist.

6. **Recommendation 6 Derivation**

To better understand the distribution of payment installments in the dataset, we first used the summarise function in DPLYR to find the mean, median, maximum, and standard deviation for the number of installments chosen by customers. We found that the average customer takes 3 months to pay off their order with a standard deviation of 2.7. Following the rules of the normal distribution, we can assume that 99.7% of customers choose installment plans that are within 3 standard deviations of the mean (0 months to 3+2.7*3 = 11.1 months). However, the maximum installment length was 22 months, which is an atypically long payment schedule for Olist's main

product category offerings - in comparison, Amazon only offers customers the option to pay in up to 4 installments, and 12 installments is customary for many other Brazillian retailers. To analyze the sentimental benefit of maintaining this longer option, we ran a linear regression in Radiant using avg_satisfaction_scores as the response variable and payment_installments as the explanatory variable. In this regression, we found that installments were significantly negatively correlated with satisfaction scores (with a p value of >0.001 and a coefficient of -0.66). We can interpret this result as stating that for each additional month of installments chosen by the customer, their average review score decreases by 0.66 (when all other variables are held constant). To better understand this relationship, we created a new variable called Long_Payment as a factor, which displayed 1 or 0 if a customer paid in more or less than 11 months respectively. We found that customers who paid in over 11 installments had a mean satisfaction score of 3.9, while customers who did not had an average score of 4.1. We also visualized this relationship with Tableau by creating a line graph that plotted mean review scores as installment length increased. Though the difference in mean satisfaction may be relatively minor, offering more installments certainly does not improve scores and the liability to the firm posed by offering this amount of installments is significant in itself, thus the firm should re-evaluate its availability.

7. **Recommendation 7 Derivation**

We ran a logistic regression to see if the quantity of photos had any effect on the product having a greater review score. Through our analysis, we determined that there was a highly statistically significant correlation between photo quantity and review score. The coefficient for product photo quantity was 0.01163, meaning as the photo quantity increased by 1 photo, the odds of the review score rising increases by 1.01%. While this number may seem low, Olist has a large amount of transactions already. Had the products sold had one additional image, our model predicts the average review score would be tremendously higher. We believe it is in Olist's best interest to first optimize all products to at least two images, and then optimize all products to include 4-6 images dependent on the product.