Analysis of Order Ratings, Sales, and Customer Relationship on a Brazilian E-Commerce Website

Authors: Adam Foley, Mann Purohit, Nhat Pham, Sarvagna Shukla
GitHub Repo: https://github.com/adam-foley/olist

I. Summary:

A. Business Question

Brazil has the biggest and fastest growing e-commerce market in Latin America and the Caribbean, concentrating 34 percent of the market in the region in 2020 [1]. Since the market is dominated by numerous large-size marketplaces rather than by a few like in the US (Amazon, Ebay etc.), Brazilian merchants face the operational burden of integrating with those marketplaces [2]. Olist was founded to solve that problem and help merchants connect to a larger customer base. The platform provides tools that support all the stages of an e-commerce operation, from managing inventory to fulfillment, customer services, and payments [3]. As a result, the driver for the company's business is to attract more clients and raise the quality of the process. The motivation in this project is to support this effort through exploring Olist's database.

B. Data Mining Solution

1. Description of the dataset:

The data provided by Olist[4] includes 8 tables (Figure 1, Appx.) of data on all areas of a transaction including customers, sellers, orders, products, payments, etc. Although they provide ways to connect these tables through "keys", in many instances it was not in a tidy format or easy to query. Therefore, we performed data cleaning and tidying for some of the tables. For example, in the payments tables (Figure 2, Appx) we consolidated the table to have a composite key of order_id and payment_type. Therefore, we had to find all instances of a payment type for each order and aggregate their payment installments and payment value.

Once we decided on the schema and keys we would like to use, we leveraged a MySQL instance hosted on AWS in order to store our database as there were over 400k rows. The reason we decided to go with a MySQL instance was to ensure that all of the table tidying we did was captured in each team member's work which ensured consistency amongst the various problems we were trying to solve. In Figure 3 (Appx.), you will see the final schema that we decided on and the fields we needed to answer the questions we had. All of the code for processing and loading the data into MySQL can be found in this notebook.

2. Project goals:

Based on the multi-level datasets available, we provided a comprehensive understanding of Olist's business through four learning objectives. First, we examined customer satisfaction based on analysis of order review scores. Our final results indicate that issues with the delivery process such as long delivery time and late deliveries, high total freight and price value, and large orders negatively impacted customer experience. Second, we performed a sentiment analysis of product reviews, and were able to create a reliable Natural Language Processing model that predicts a customer review's sentiment given a textual message. Third, we would like to target larger segments of customers by knowing their shared values, likes and buying behavior. Our customer segmentation analysis identified different clusters of customers, yet without labels and limited data it was difficult to discern each cluster's traits. Finally, we examined Olist's sales activities and discovered patterns on seasonality and trend from historical sales which might be important for the logistic concept. We also identified important product categories, differences in performance across different geographical markets, and areas that Olist should focus on to improve and expand its business.

II. Data Analysis and Modeling

A. Customer Satisfaction Analysis:

Surveys report that 93% of consumers are influenced by online reviews when making purchase decisions [5]. By leveraging its order review data, Olist can gain insights into its customers' wants and needs, and therefore enhance its services and products to provide better customer experience and attract more merchants to the platform. Through EDA and machine learning models for classification, we are interested in identifying key features that affect a good or bad review score, using information on reviews, orders, customers, sellers, products, and payments.

1. Exploratory Data Analysis

We started by looking at the distribution of review score ratings of all the orders in our dataset. The majority of our orders were given 4 and 5 scores (Figure 4, Appx.). To improve the imbalanced distribution, we categorize 1, 2, 3 scores as negative scores and 4, 5 scores as positive scores. Next, the team examined the attributes that could potentially impact review scores.

We first assessed factors related to the delivery process, including delivery time (number of days it takes from time of purchase to actual delivery date) and late

delivery time (difference in estimated and actual delivery date). We found that orders that were delivered late received significantly lower scores on average than those delivered on time (Figure 5, Appx.), and that the longer it takes for an order to arrive, the less happy the customers are (Figure 6, Appx.). Also, mean delivery time among all orders is 10 days, much higher than Amazon (3.39 days average, 2017 data) [6]. This is a KPI Olist might want to improve to maintain its competitiveness in the long term.

The team also reviewed other order-based characteristics. We saw that large orders (those with multiple items or multiple sellers) have lower scores on average than single-item orders (Figure 7, 8, Appx.). This could be because large orders are more prone to fulfillment errors. Yet, 86% of all orders are single-item-orders so the pattern might not be representative. Meanwhile, we found no difference in the average number of photos, product description length, and product name length per order across different review scores. Proportion of negative and positive reviews is similar across different payment types, indicating that customers' experience do not depend on payment methods chosen. When reviewing sellers' profiles, we saw that sellers with fewer negative reviews are the popular ones (high order volumes) who also have a low proportion of delayed orders (Figure 9, Appx.).

2. *Data Cleaning & Feature Engineering*

As mentioned above, we assigned a binary variable for 'positive' reviews (4, 5 scores) and 'negative' reviews (1,2,3 scores). The number of positive and negative reviews was still imbalanced, and we will talk about the effect it had on our models in the *Modeling & Evaluation* section. In addition, we created some aggregate measures to assess their potential relationship with review scores, including order value, item count per order, seller count per order, delivery time, order counts per product, and order count per seller.

We noticed several numerical features that were heavily positively skewed. These features were not transformed for the first few models. After that, models were rerun using log-transformed data to see whether this improves model accuracy. Next, multicollinearity was checked for by assessing correlations between predictor features, as this can cause issues with some models. The multicollinearity matrix showed that this is not an issue. We also performed one-hot encoding for categorical variables such as payment type and purchase day (of week) to prepare for modeling. We then ended up with 27 attributes to be considered for this analysis. Finally, because independent features were on different scales, they were transformed and normalized using the StandardScaler function in sklearn.

3. *Modeling & Evaluation*

We split the dataset into 80% for the training set and 20% for the testing set, then used Random Forest and Logistic Regression algorithms for this analysis. These models were chosen as appropriate for modeling our binary response variable, "negative" versus "positive" review scores.

a. *Evaluation Metrics*

As the main evaluation metrics, we used the F1 scores of both classes and Area Under the Curve (AUC). The F1 score calculates the harmonic mean between precision and recall and is a suitable measure because there is no preference for false positives or false negatives in this case. AUC is a common, easily interpretable metric for binary classifications. Graphs can be plotted together for comparison across different models to identify the best performer. (See Figure 10, Appx. for a comparison of all models).

b. *Logistic Regression*

Running the Logistic Regression (LR) model with 27 features gave us a fairly good weighted average F1 score of 0.78. However, it is notably worse at predicting negative than positive reviews, with the recall rate being 0.26 for the negative class and 0.96 for the positive class. The AUC has a value of 0.704. Figure 11 (Appx.) shows the top 10 feature importance for the model. As expected, delivery time and whether or not the order is delayed are the most important features, followed by seller count per order and item count per order, and whether the order used boleto as the payment type.

Applying log transformation to right-skewed features and rerunning LR, we achieved an AUC of 0.698, which underperformed the basic LR model. In attempts to improve our Logistic Regression model, we performed random oversampling and undersampling on the imbalanced original data set and rerun the model. LR with oversampled data results in an AUC of 0.707, and while the recall rate improved (0.49), the precision rate for both classes worsened noticeably (0.44 for positive class) compared to the original model. LR with undersampled data gave us similar results.

c. *Random Forest*

We also tried running a basic Random Forest model to see if the accuracy is improved. Random Forest classifier algorithm is a set of decision trees from randomly selected subsets of the training set. It aggregates the votes from different decision trees to decide the final class of the test

object. The “RF Basic” model gave a weighted F1-score of 0.79 and an AUC of 0.711, while the RF model with oversampled data achieved a weighted F1-score of 0.80 and an AUC of 0.715, better than the logistic regression models.

As can be seen from Figure 12 (Appx.), some of the most important features in the Random Forest model are delivery time, order price and freight goal size and some product properties such as name length and description length, which we did not expect from EDA, and delay status of the order

d. Conclusion

Our best model was the Random Forest model with randomly oversampled data (Figure 10, Appx.). Interestingly, each model performed worse at predicting negative than positive reviews, with a lower true negative rate than true positive rate i.e., it classified quite a few negative reviews as positive, but relatively few positive reviews as negative. Possibly the factors that might cause negative reviews are more likely to be beyond the scope of the data used in this analysis, such as quality of the products, poor marketing and customer services.

Figure 10. Review Classification Model Comparison

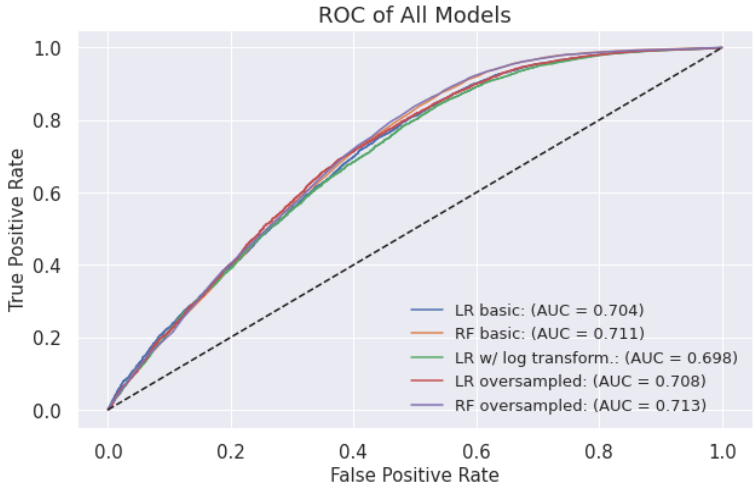


Figure 11. Feature Importance, Logistic Regression

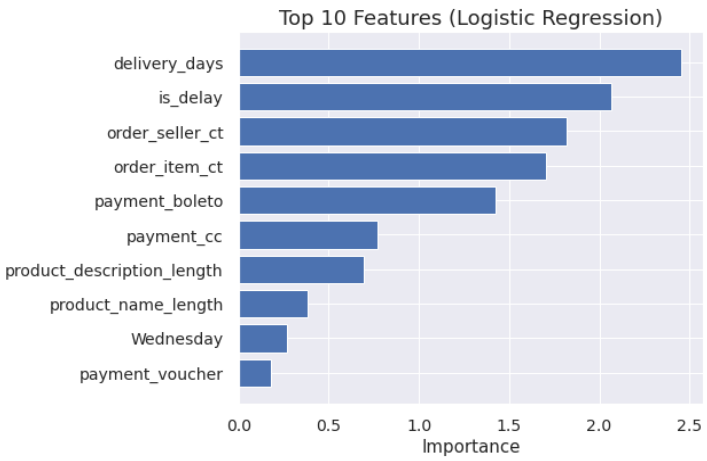
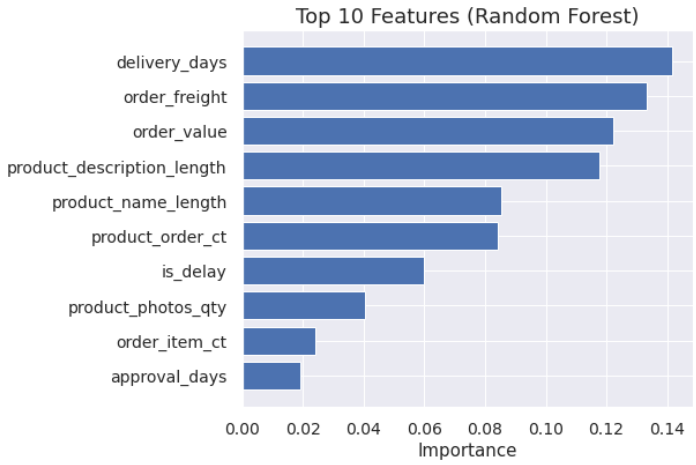


Figure 12. Feature Importance, Random Forest



B. Review Text Analysis:

1) Applications of NLP

Natural Language Processing (NLP) is intriguing because it is a technology through which we can train machines to learn the “Natural Language of humans", like text and audio, and generate predictions on the trained model. Performing sentiment analysis using textual data has exciting real world applications that can be implemented impactfully to leverage user experience through recommendation. Some of the popular applications are Product Analysis, Customer Support, Customer Feedback, and Social Media Monitoring. Social Media posts contribute hugely to any market place and extracting key insights from what the customers think about the product becomes integral. We can extract these insights through finding the most important words for positive and negative reviews for

each category. Using NLP, companies could look into different reviews and extract whether something is wrong with a particular product, and whether they need to update it in order to meet customer demand. This would in turn benefit customers too, as they would get the desired product. For example, consider a review with the message "the color of the black shirt that I ordered faded. Not a good product". OLIST could use NLP to predict sentiment of this review and extract key features from it. They could then let the supplier of this product(black shirt) know about the problem that the above customer faced.

2) Data Cleaning and Preprocessing

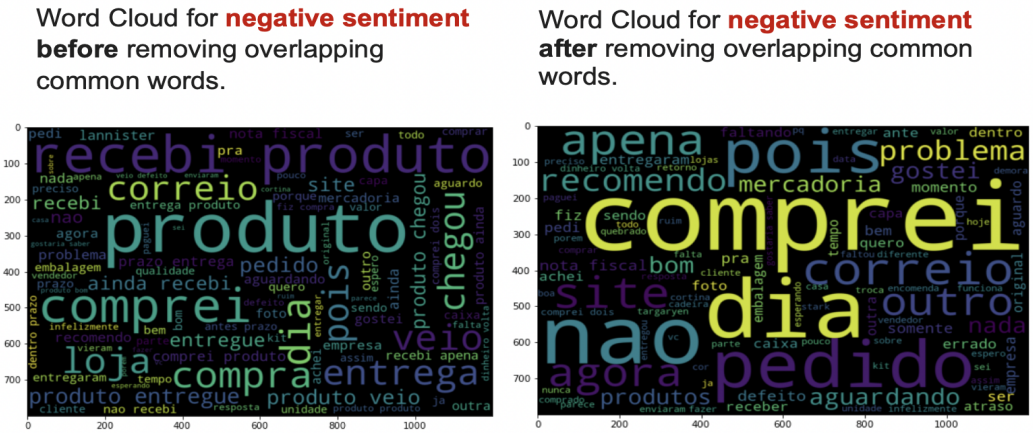
We use and compare different methods like TF-IDF from sklearn, and Tokenizer from keras to transform text data into word vectors. We then use 4 machine learning algorithms - Logistic Regression, Random Forest, Vanilla (Simple/Plain) Neural Networks, and Recurrent Neural Networks to predict a new category for the given input. We use accuracy as the performance metric to evaluate the above models.

The text data in the *message* column of the dataset as input text and the *review_sentiment* column as the target column. We start by cleaning and preprocessing the data. Firstly, we perform data cleaning using *"Regular Expressions",* in which we first remove all the special characters, digits, emojis, and any other text that does not contain alphabets and convert this text to lowercase. We do so because they are insignificant while predicting the output, and keeping them results in increased dimensionality, which results in expensive computations (it should be avoided). We then remove extra spaces and stopwords, they are words which do not add any important information to a sentence, for example "is" and "the". After this step, the corpus (corpus is a combination of all the textual data present in the samples) will have words that contain alphabets only. We move on to data preprocessing, where we have 2 techniques to preprocess text data - *i) Stemming* and *ii) Lemmatization.* The aim of both the techniques is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. Stemming is a process of chopping off the ends of words in order to achieve the above goal and it also often removes derivational affixes. Lemmatization refers to the process of using a vocabulary and morphological analysis to reduce the inflectional forms of words and return a word's base form, known as lemma. Stemming performs the following mapping - {car, cars, cars', car's} => {car}, while lemmatization aims to map {am, are, is} => {be}. We use lemmatization over stemming as chopping off words from the end may change a word's interpretation. This completes cleaning and preprocessing the data.

3) Data Visualization

We plot the most frequently occurring words in both the categories to perform EDA, and find that both categories had overlapping frequently occurring words. To deal with this, we remove these overlapping words from the text data and plot most common words again using Word Clouds, where larger font words have higher frequency.

Figure 13. Word Clouds for Negative Sentiment, Before vs. After Removing Common Words



We fit different models on different datasets and show their accuracy in the table below :

Table 1. NLP Model Comparison

Model	Dataset	TF-IDF	Tokenizer	Embedding
LR	Train	89.01	67.85	NA
RF	Train	98.85	98.89	NA
NN	Train	94.5	NA	95.23
LR	Test	**88.1**	67.85	NA
RF	Test	86.76	77.46	NA
NN	Test	86.49	NA	86.13

Interestingly and surprisingly, Logistic Regression, (the simplest algorithm among the above models) fits the data perfectly with low bias and low variance, and has the best accuracy.

C. Customer Segmentation:

1. *Feature Identification*

In any business understanding your customer base is very important to the strategic decisions you make to improve their experience and in return drive sales for your company. In the case of Olist, this is no different; therefore we tried to understand their customers by using clustering, an unsupervised machine learning technique to group customers based on the similarity between features that best identify them.

For our clustering, we chose features which we believed would allow us to best group together different segments of customers. The features we selected can be split into four separate categories including orders, deliveries, payments, and reviews. First, we thought that order details such as number of orders, days since last order, unique products purchased in an order, and total sales would be the most important. In regards to delivery details we believed that another feature between customers could be the number of orders that were early and late. The rationale behind this being that customers with early orders would be more willing to buy again compared to customers with poor delivery times would be less likely to come back to the site. Also, with the rise in "Buy Now Pay Later" options, we looked at average payment installments as we felt customers that had the choice to pay over more installments would also be willing to spend more on each order. Lastly we felt that how often a customer reviewed their orders and what score they gave would be another way to cluster customers as it would separate those customers with great experiences from those with poor experiences on the site.

2. *Data Preparation*

After we identified what features to select for each customer, we prepared and cleaned the data for clustering. First, we began by visualizing the distribution of each feature to understand where most of our customers fell, as shown in Figure 14 (Appx.). We observed that there were some customers with outliers in total sales and also order counts; and therefore we chose to filter out customers with sales more than $2000 or customers with more than 3 orders.

3. *Dimension Reduction and Modeling*

With our data prepared, we could now move onto modeling. However, before we could cluster, we needed to scale our features using a standard scaler and also reduce the number of features down from 9 to 2. In order to do this we tried two separate techniques for dimensionality reduction - PCA and t-SNE. First we began with PCA, which is a linear technique which can reduce the dimensions to a desired amount while still allowing us to explain most of the importance from the 9 original features; The two components explained 32.4% and 17.4% of the variance respectively for a total or ~50%. We also tested using t-SNE, a non-linear dimension reduction technique, however due to the size of the data we could not tune the t-SNE to provide any real separation which makes it useless for clustering. In Figures 15 and 16 (Appx.) you will see images of K-means clustering with PCA and then t-SNE which provides some more context to why we chose PCA.

After selecting PCA as our technique for dimension reduction, we were left with the 2 components for all customers in our data set that we could use for clustering. Much like dimension reduction, we analyzed several clustering algorithms to compare their results. The two we choose to highlight include K-means and Birch as they were able to be run on our large data sets. Another important aspect when clustering is understanding how many clusters

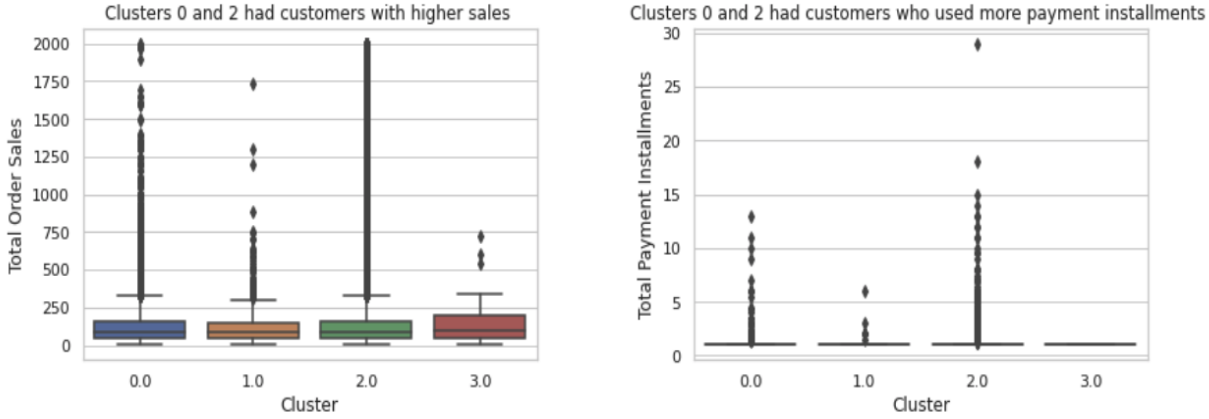
naturally appear in the data, for this we use a technique where we visualize the Within Cluster Sums of Squares in order to find the point where the sum declines at a slower rate as by the nature of clustering, increasing the number of clusters will always improve this and therefore we want to avoid overfitting. In our clustering, we chose to go with 4 clusters which can be shown in Figure 17 (Appx.).

4. Results

Once we selected the cluster techniques and the number of clusters to use, we can now analyze the two clustering methods we ran - K-means and Birch. As shown in Figure 16 in the appendix , we can see that although they are very similar, it does appear that Birch provides possibly a better separation amongst the clusters when compared to K-means. After looking at the features of each cluster from both techniques it appeared our hypothesis was true. Therefore, we decided to focus on drawing conclusions using customer clusters derived through Birch.

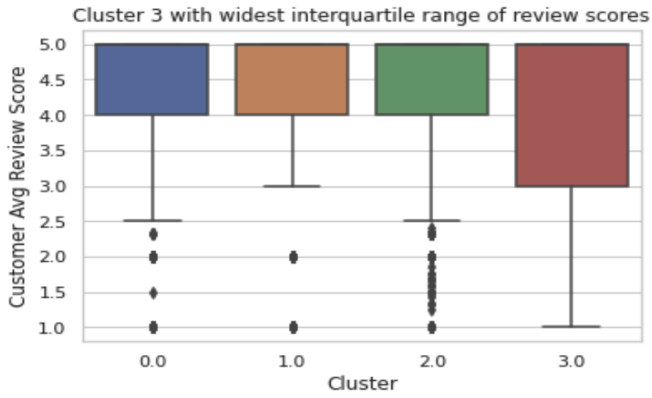
Due to the nature of unsupervised machine learning having no labels or ways to know how accurate our clustering was through accuracy metrics, we will look to EDA to possibly identify differences in each of the 4 clusters. First, we decided to look at sales and payments to identify some differences. As shown in the figure on the left below, clusters 0 and 2 had larger right-skewed distributions with more customers having higher total sales volume compared to clusters 1 and 3. Once we observed this we wanted to understand if payment installments were high in clusters 0 and 2 which would back up our prior thought that customers who paid in higher installments also bought more. As you can see in the figure below on the right, it was the case that clusters 0 and 2 also had right-skewed data and more customers with higher payment installments.

Figure 18. Sales and Payment Installments per Cluster



Along with assessing sales and reviews amongst the clusters, another area we identified some differences between clusters was with review scores. As shown in the figure below, cluster 3 had the widest interquartile range of review scores amongst the clusters from 5 to 3 which tells us that this cluster has a more varying group of customers with poor and bad reviews. Another observation is that cluster 0 and 2 have a right-skewed distribution and cluster 1 has the tightest distribution of reviews.

Figure 19. Review Score by Cluster



D. Sales Analysis:

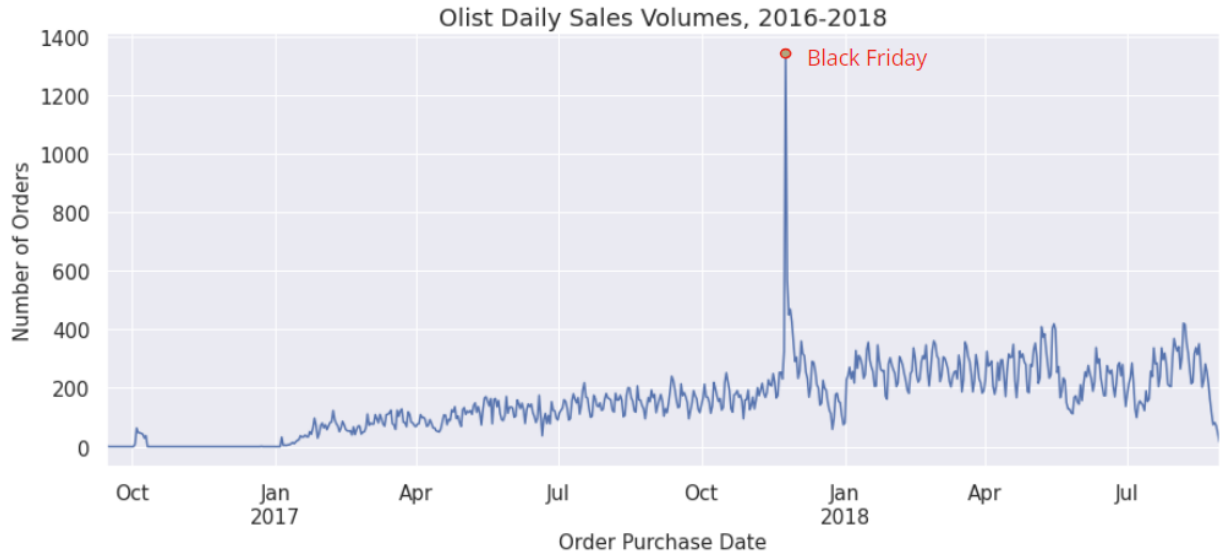
Conventional retailers are often focused on only one specific market segment. This can be fashion, food or general merchandise. Olist on the other hand, serves a very wide market with no targeted segment. Hence, a comprehensive understanding of sales and revenues is crucial for several business decisions such as marketing, inventory, and strategizing about incorporating warehousing services. In this analysis, we will use EDA as a means to identify business trends over time and across sectors, providing a foundation for future forecasting tasks.

1. *Trend and Seasonality*

We examined daily sales, sales by day of week, and within a day. Historical sales data showed a steady upward trend with an extreme spike in November 2017, which was a Black Friday. The trend appears to be linear but has a high variance. We also observed peak effect during some holidays such as Free-shipping Day (last Friday of April), Independence Day (September 7) (Figure 20, Appx.). Because we had limited historical data (2016-2018), very volatile data in 2016, when Olist was first launched, and extreme seasonality, performing sales forecasting can be challenging and unreliable.

Each of the top 3 product categories (based on sales volume) have pretty different seasonality and peak effect, as can be seen in Figure 21 (Appx.), yet all received record sales on Black Friday. Based on this information, in the future when more data is available, we will do a bottom up forecast, creating a separate model for each product category and summing it up for the overall forecast.

Figure 20. Daily Sales Volumes



Next, we looked at the order delivery situation since delivery days affect customer experience. Overall, the business delivers orders as estimated or faster (Figure 22, Appx.). Black Friday impacted order shipment planning, and an occurence in mid June 2018 which made estimated shipping duration suddenly increase. This period fell into the FIFA Soccer World Cup, an important event for Brazillians. Given the big difference between estimated and actual delivery days, Olist should look into improving its logistic systems.

We also zoomed in on the weekly and daily trend, and saw that more orders were made on Monday than any other day of the week: Monday blues seemed to show an effect on purchases (Figure 23, Appx.). Afternoon boredom also seemed to make customers buy more items than any other time of the day (Figure 24, Appx.). Olist could use this information to strategize their business efforts to attract new customers, boost sales, and prepare for increased web traffic during this day/ time.

2. *Product Categories*

Barcharts in Figure 25, 26, 27 (Appx.) visualize a clear picture. Top 15 product categories that generate the highest revenue (accounting for 80% of Olist's total revenue), highest order volumes (77% of Olist's total order volumes) and broadest number of available products (77% of Olist's product counts) have a lot in common, hence are crucial to Olist's business. It would be interesting for Olist to study the correlation between those product categories and see whether any diversification gain can be achieved by growing other product segments.

Since there was a lot of overlap in 73 product categories, we tried grouping them into 13 category types and compared customers' review scores for each.

Interestingly, orders of books had the highest rating (4.4), and furniture had the lowest rating among all categories (3.96) (Figure 28, Appx.). Olist might want to consider expanding its business in the books section, while evaluating customers' feedback for furniture products since it is one of the top 15 categories.

3. *Sales by Region*

When examining sales across different states, we observed that the majority of Olist's customers, sellers, hence sales and revenue, are coming from the Southeast and South regions of Brazil (Table 2). Customers in Northern and Northeastern regions experience longer delivery time, more late deliveries, and consequently, tend to give lower review scores. This can partly be explained by the long distance from the main sellers. Two largest cities, São Paulo and Rio de Janeiro, have, by a big margin, the biggest overall customer count and revenue. On the other hand, the company should plan to improve their performance in Southern regions, through expanding their business to retailers in these areas, improve their logistic processes to reduce waiting time, increase their marketing efforts and conduct product improvement state-wise to improve sales towards Southern states.

III. Discussions

Our EDA and models for classifying positive versus negative review scores indicate that delivery time and whether the order is delayed have a strong impact on customer's experience, followed by freight fee and order values, and number of items in the order. Furthermore, our machine learning models using NLP will help Olist better understand a customer' sentiments and the dynamics that cause poor customer experiences through classifying negative and positive review messages. Incorporating findings from our business trend analysis, our recommendations to Olist include: (1) improving their logistic and fulfillment systems to increase accuracy in estimation of delivery time and efficiency in the delivery process, (2) consider expanding their business to Northern and Northeastern states, where there is a growing demand but few retailers and, as a results, high delivery time (3) study sales correlations between important product categories and consider growing other product segments with high interests and good customer satisfaction, for example to protect them from economic shock.

As Olist has a diverse customer base, it is beneficial to be able to identify target customer groups based on their similar behaviors. Since we did not have target labels, it was difficult to assess the performance of our unsupervised machine learning models for customer clustering. In the future, we can attempt other segmentation methods such as the RFM (Recency, Frequency, Monetary) analysis which calculates R-, F-, and M-scores for each customer based on their buying behavior (how recent, how often, and how big the order is), categorize them into different customer types, and develop suitable marketing strategies. Due to the limited experience in the retail industry, working with retail experts to better understand how to classify customers based on features in the data and in conjunction with other analysis results would improve our results in a preferable and sensible way. Further improvements of our analyses also include performing hyperparameter tuning on our review score classification models, exploring other resampling methods for imbalance data, as well as other classification models such as XG Boost and Gradient Boost to see if model accuracy is improved.

Due to the imbalanced nature of available data, with much higher proportion of positive versus negative review scores, and limited historical data, it was challenging to develop a model that can reliably predict both classes. Most of the orders are single-item-orders and there are few repeated customers in the dataset. With access to more data and in a longer period, possibly for at least 5 years, we will be able to better assess sales seasonality, customer churn rates, and provide a recommendation system based on customers' past purchases.

References:

1. Chevalier, S. (2021, October 25). *Latin America: E-commerce share by country 2020-2021*. Statista. Retrieved December 13, 2021, https://www.statista.com/forecasts/256166/regional-distribution-of-b2c-e-commerce-in-latin-america.
2. Degordian. (2020, July 7). *Olist is empowering small merchants to sell online.* Valor. Retrieved December 13, 2021, https://valorcapitalgroup.com/case-studies/olist-redesigned-the-marketplace-business-model-to-fit-the-realities-of-ecommerce-in-brazil/.
3. Cramos. (2020, November 25). The state of Ecommerce in Latin America: Interview with Olist CEO Tiago Dalvi |. LAVCA The Association for Private Capital Investment in Latin America RSS. Retrieved December 13, 2021, https://lavca.org/2020/11/23/the-state-of-ecommerce-in-latin-america-interview-with-olist-ceo-tiago-dalvi/.
4. Olist. (2021, October 1). Brazilian e-commerce public dataset by Olist. Kaggle. Retrieved December 13, 2021, from https://www.kaggle.com/olistbr/brazilian-ecommerce
5. Diana Kaemingk(2021, August 17). 20 online review stats to know in 2019. Qualtrics. Retrieved December 13, 2021, https://www.qualtrics.com/blog/online-review-stats/#:~:text=91%25%20of%2018%2D34%20year,reviews%20influenced%20their%20purchase%20decisions
6. Mazareanu, E. (2019, January 17). Amazon: Average delivery time 2018. Statista. Retrieved December 13, 2021, from https://www.statista.com/statistics/957782/parcel-carriers-on-time-delivery-rate-peak-season/

Contributions

Each team member contributed to the research, design, and execution of the data mining problem. Team members met together weekly to consult and review together. The contributions below reflect participation in particular topics:

- Adam Foley: Designing, creating, and loading of the AWS MySQL instance and the Customer Segmentation analysis
- Mann Purohit: Review text analysis
- Nhat Pham: Performing the Customer Satisfaction analysis and Sales analysis
- Sarvagna Shukla: Review text analysis

Appendix

I. Supplementary Figures

Figure 1. Olist Data Schema

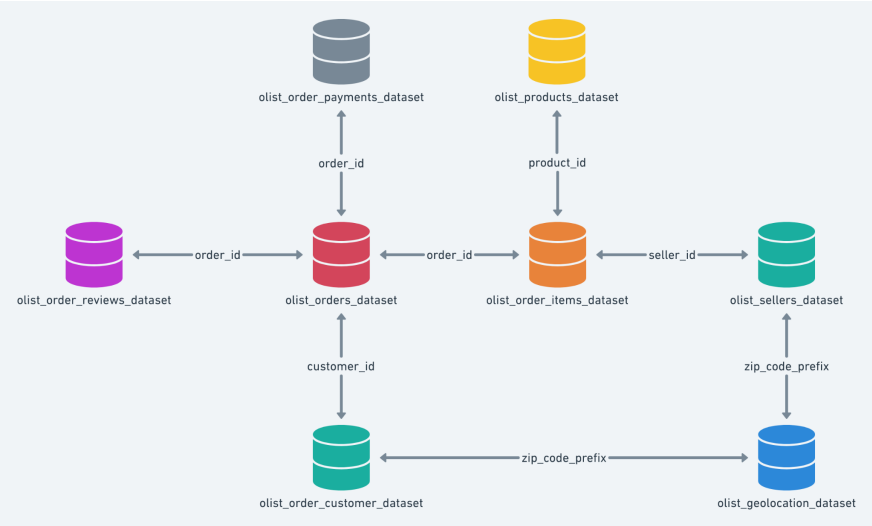


Figure 2. Data Cleaning Example

order_id	payment_sequential	payment_type	payment_installments	payment_value
1	1	voucher	1	89.46
1	2	voucher	1	23.99
1	3	voucher	1	4.78

order_id	payment_type	payment_installments	payment_value
1	voucher	3	118.23

Figure 3. Final AWS MySQL Schema

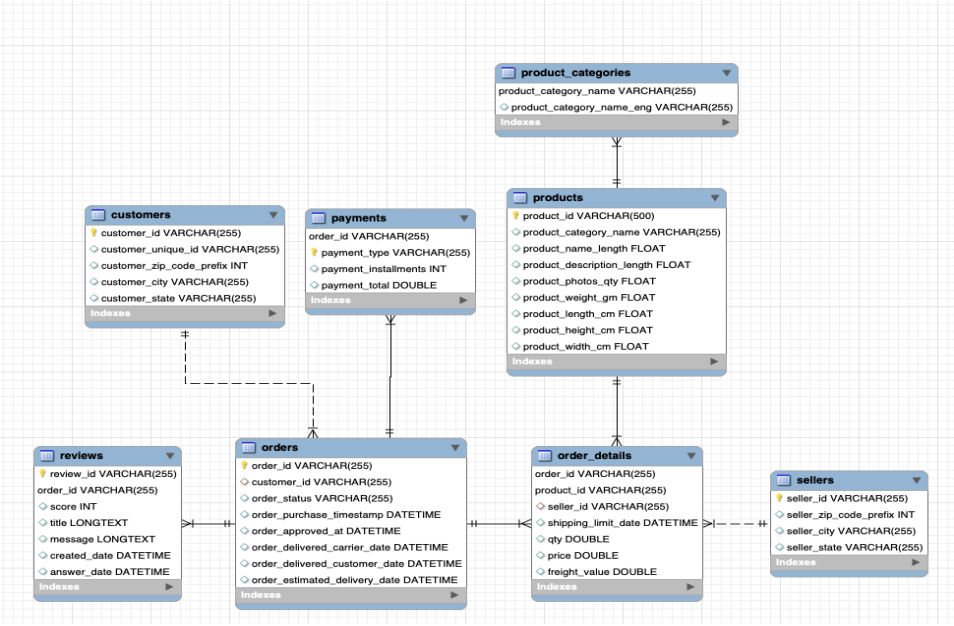


Figure 4. Order Count by Review Score

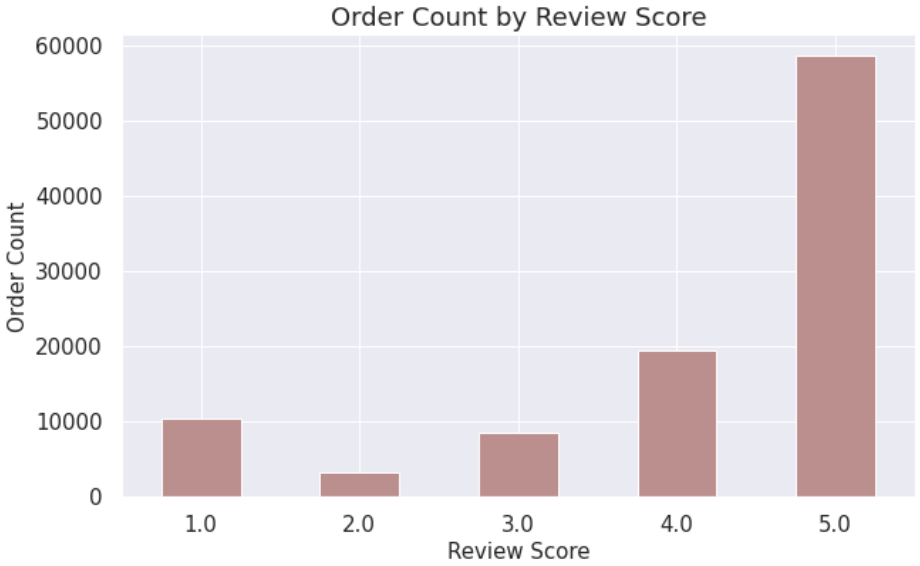


Figure 5. Average Review Score by Late Delivery Days

Figure 6. Average Review Score by Delivery Days

Figure 7. Avg. Score by Item Count per Order

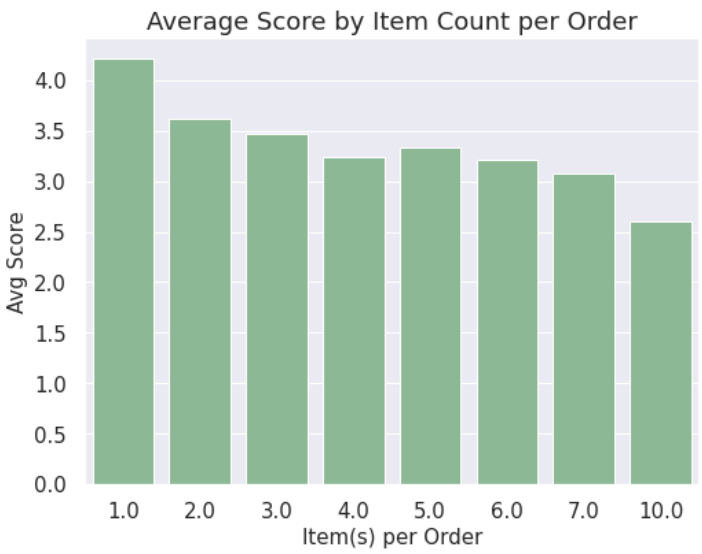


Figure 8. Avg. Score by Seller Count per Order

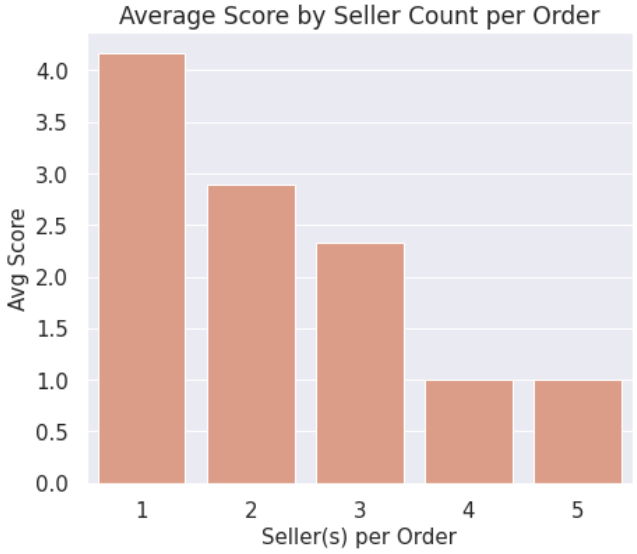


Figure 9. Seller Quality based on Late Delivery and Negative Review

Figure 14. Clustering Outlier Removal

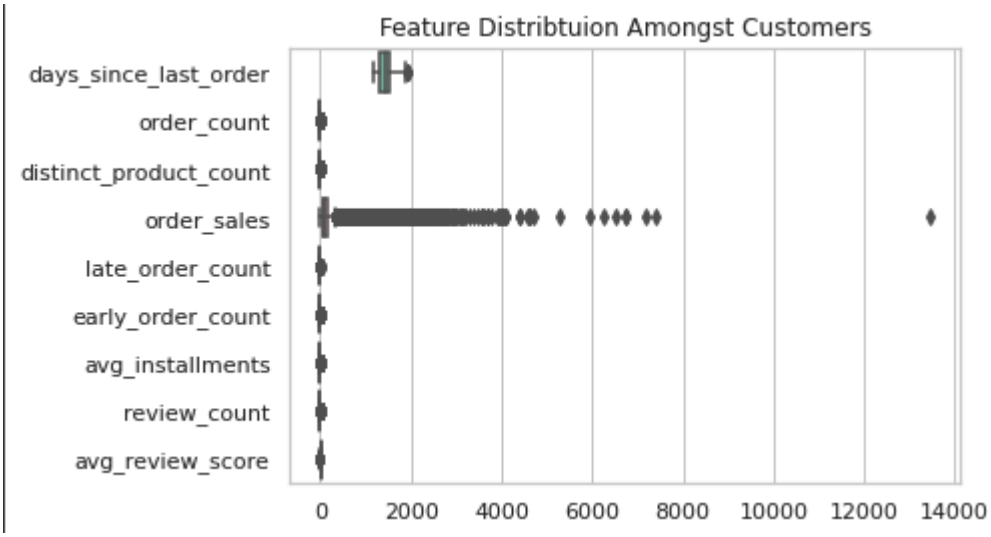


Figure 15. K-means clustering comparison between PCA and t-SNE

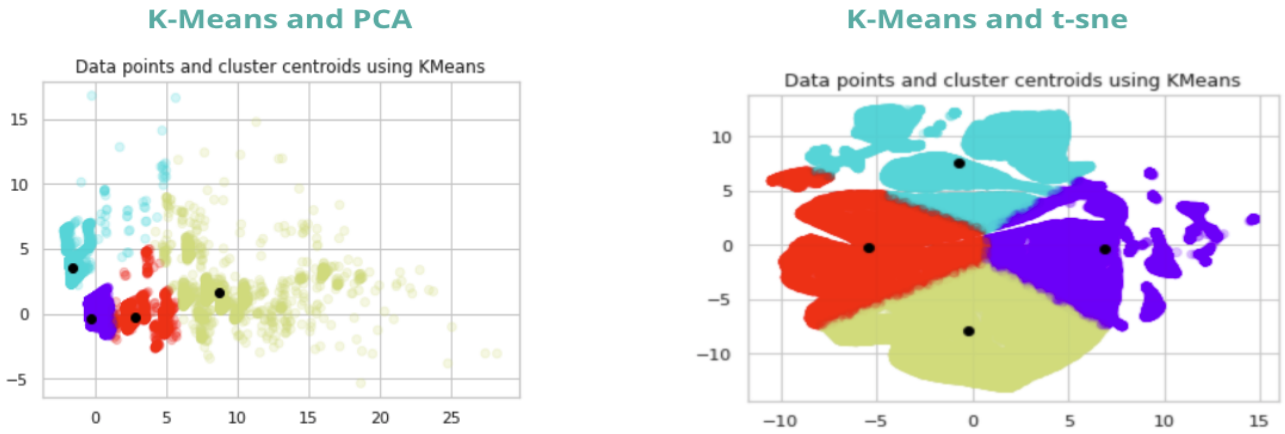


Figure 16. K-means and Birch clustering comparison using PCA

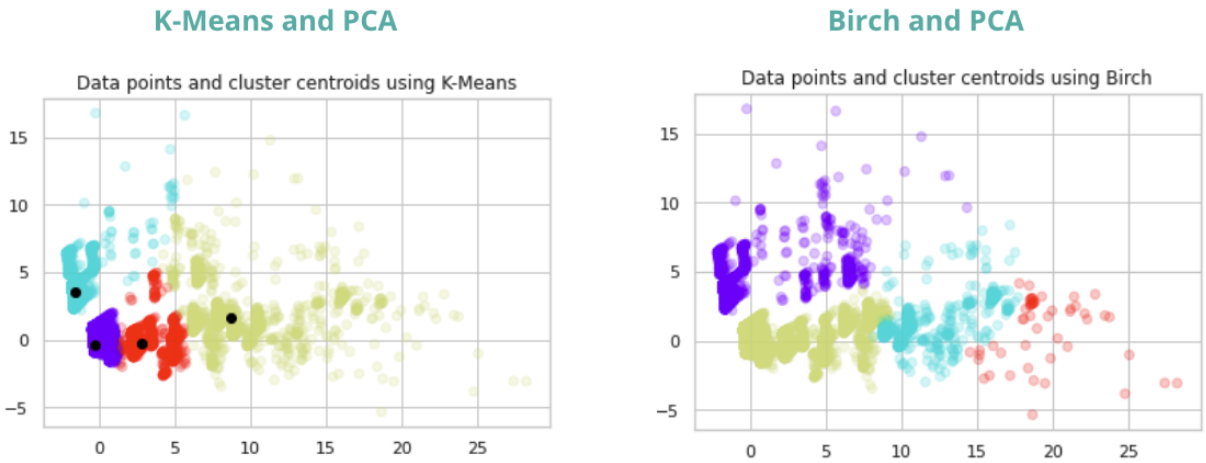


Figure 17. Number of cluster selection

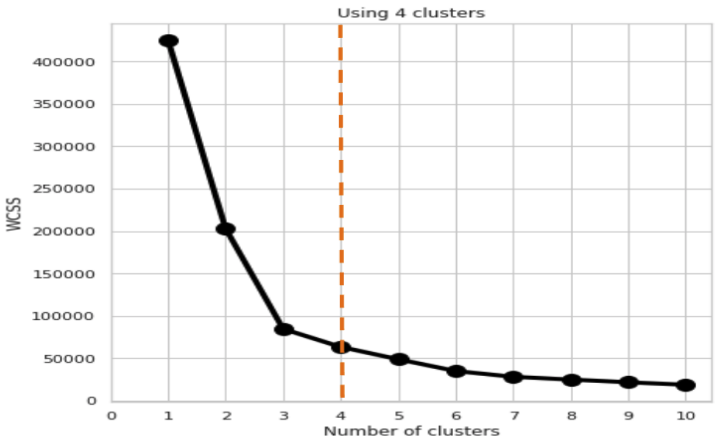


Figure 21. Weekly Order Volumes of Top 3 Product Categories

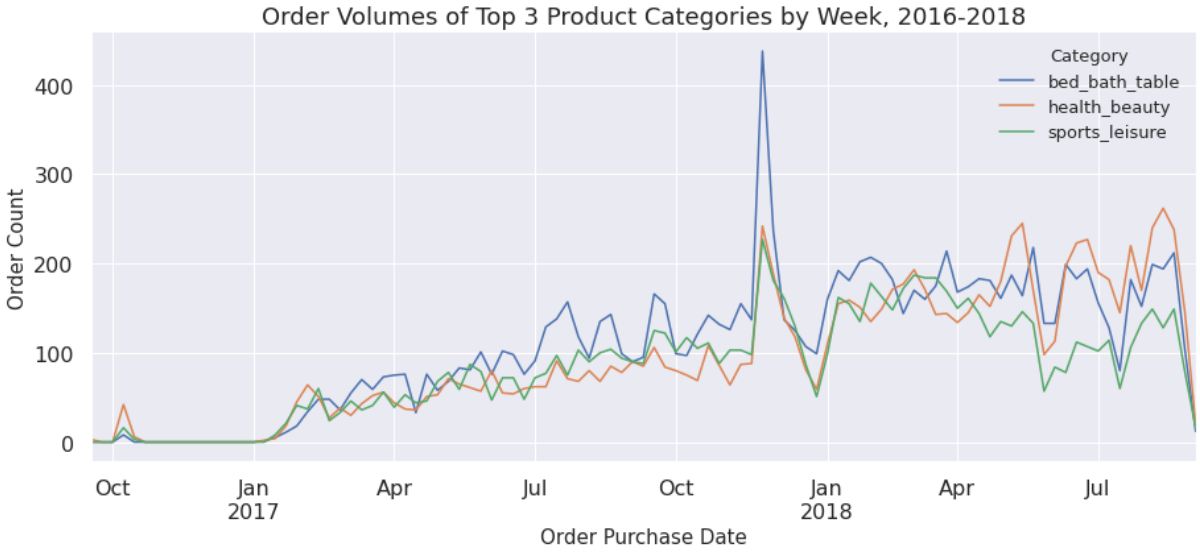


Figure 22. Weekly Order Volumes of Top 3 Product Categories

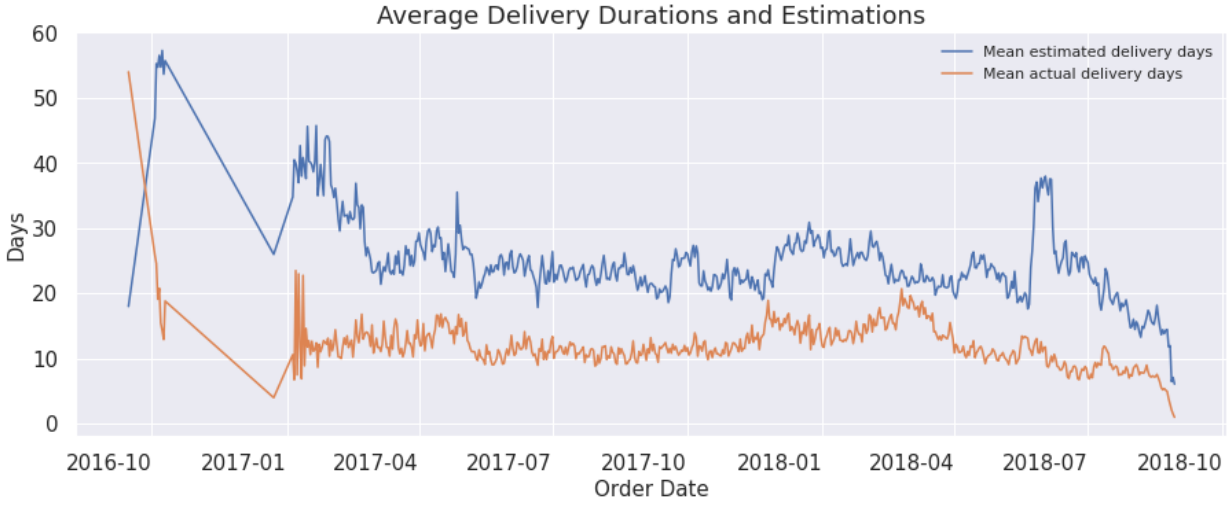


Figure 23. Order Volumes by Day of Week

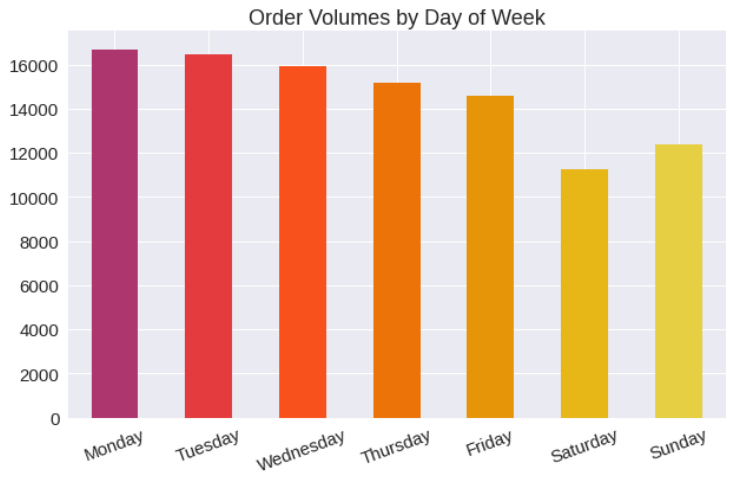


Figure 24. Order Volumes by Time of Day

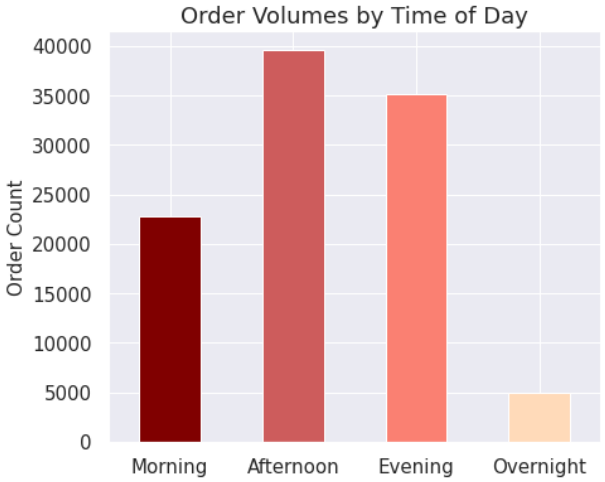


Figure 25. 15 Categories with Highest Revenue

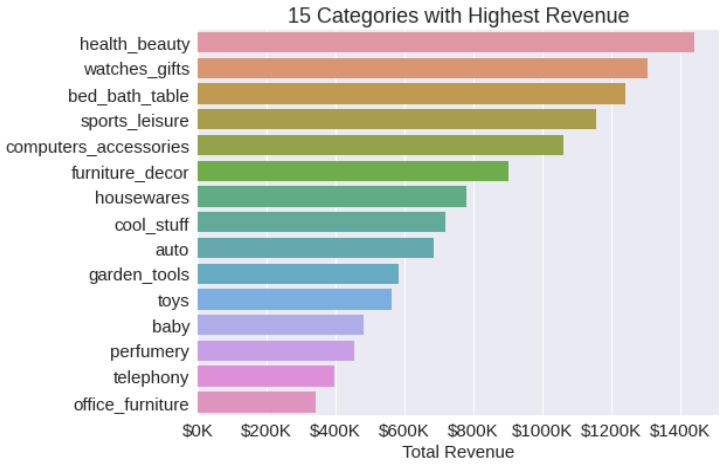


Figure 26. 15 Categories with Highest Order Volumes

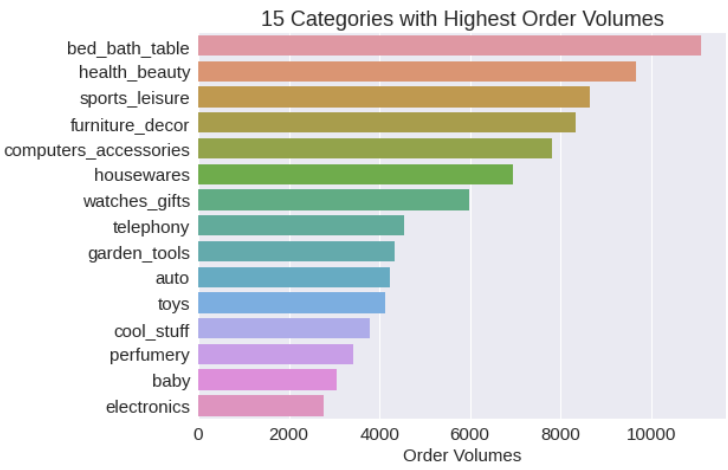


Figure 27. 15 Categories with Highest Number of Products

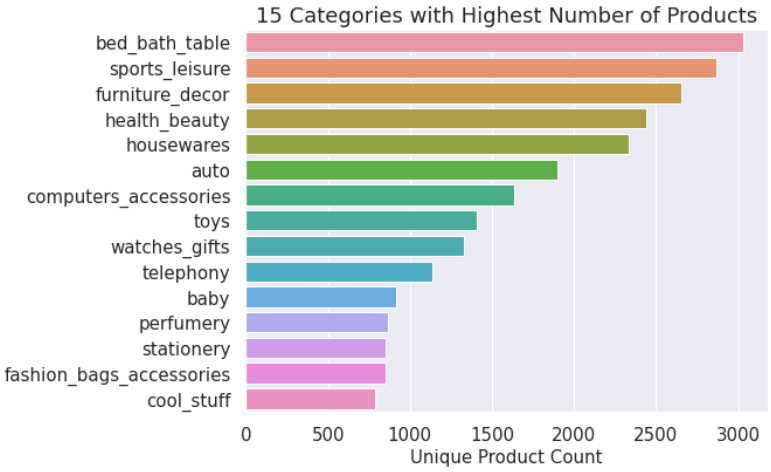


Figure 28. Average Review Score by Product Category

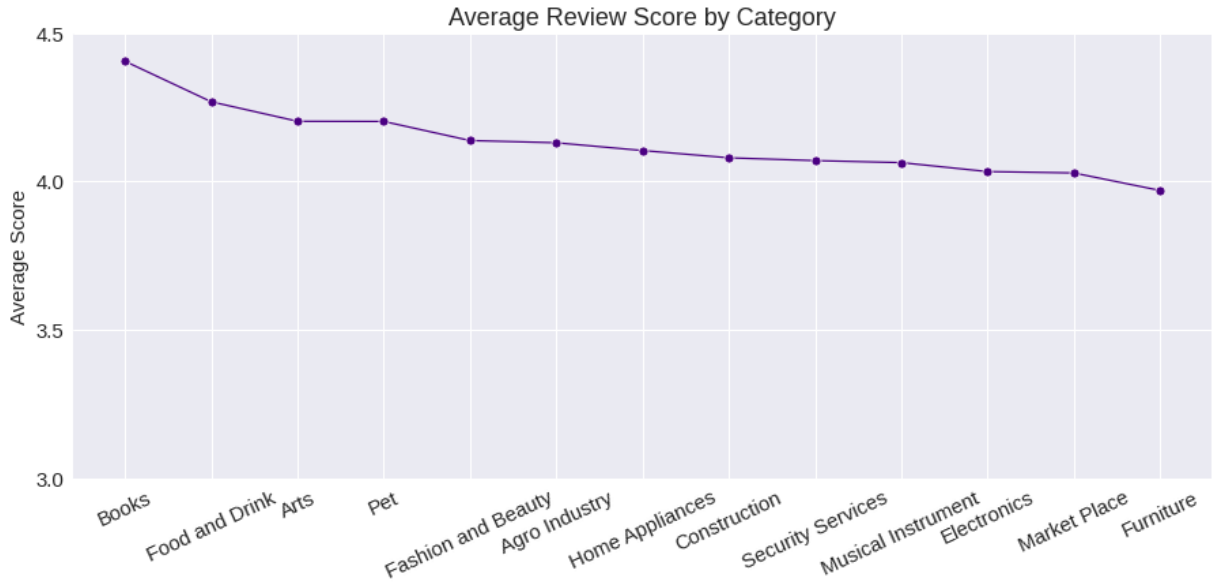


Table 2. Sales by State and Region

State	Region	Avg Score	% Positive Review	Avg Delivery Days	% Late Order	Order Count	Revenue	% Customers	% Sellers
Distrito Federal	Center West	4.10	77.9%	12	6.9%	2154	318,306	2.2%	0.8%
Goiás	Center West	4.08	76.5%	15	7.9%	2031	298,910	2.0%	0.5%
Mato Grosso do Sul	Center West	4.13	78.1%	15	11.8%	722	119,872	0.7%	0.0%
Mato Grosso	Center West	4.08	76.9%	17	6.5%	924	164,954	0.9%	0.1%
Acre	North	4.06	74.1%	21	3.7%	81	16,301	0.1%	
Amazonas	North	4.22	81.8%	26	4.7%	148	22,841	0.1%	0.0%
Amapá	North	4.22	81.2%	29	5.8%	69	13,852	0.1%	
Pará	North	3.89	71.9%	23	11.7%	963	180,737	1.0%	0.0%
Rondônia	North	4.12	75.9%	19	2.8%	253	48,239	0.3%	0.0%
Roraima	North	3.90	68.3%	29	12.2%	41	7,057	0.0%	
Tocantins	North	4.15	79.2%	17	12.3%	284	51,245	0.3%	
Alagoas	Northeast	3.83	71.0%	24	23.1%	403	79,472	0.4%	
Bahia	Northeast	3.90	72.0%	19	13.6%	3347	514,732	3.4%	0.6%
Ceará	Northeast	3.92	72.1%	21	15.0%	1310	230,271	1.3%	0.1%
Maranhão	Northeast	3.78	68.6%	21	19.2%	739	133,045	0.7%	0.4%
Paraíba	Northeast	4.05	75.3%	20	10.5%	526	114,546	0.5%	0.0%
Pernambuco	Northeast	4.06	76.6%	18	10.6%	1618	259,857	1.6%	0.4%
Piauí	Northeast	3.98	74.8%	19	16.0%	481	87,326	0.5%	0.0%
Rio Grande do Norte	Northeast	4.13	78.7%	19	10.2%	488	85,591	0.5%	0.1%
Sergipe	Northeast	3.90	72.1%	21	14.7%	341	57,389	0.3%	0.0%
Paraná	South	4.21	80.6%	11	4.8%	5058	699,716	5.1%	7.7%
Rio Grande do Sul	South	4.15	78.7%	15	7.0%	5524	766,113	5.6%	2.0%
Santa Catarina	South	4.10	77.2%	14	9.5%	3639	527,734	3.7%	3.7%
Espírito Santo	Southeast	4.04	76.2%	15	11.7%	2039	269,546	2.0%	0.3%
Minas Gerais	Southeast	4.17	79.1%	11	5.5%	11711	1,625,518	11.8%	7.9%
Rio de Janeiro	Southeast	3.93	72.5%	15	13.1%	12675	1,836,941	12.7%	4.3%
São Paulo	**Southeast**	**4.22**	**80.5%**	**8**	**5.7%**	**41910**	**5,369,263**	**42.1%**	**71.1%**

II. Brazilian E-Commerce Dataset by Olist Documentation

For more information on the dataset please visit:
https://www.kaggle.com/olistbr/brazilian-ecommerce?select=olist_order_payments_dataset.csv

customer_id: key to the orders dataset. Each order has a unique customer_id
customer_unique_id: unique identifier of a customer
customer_zip_code_prefix: first five digits of customer zip code
customer_state: customer state
order_id: order unique identifier
order_item_id: sequential number identifying number of items included in the same order.
product_id: product unique identifier
seller_id: seller unique identifier
shipping_limit_date: Shows the seller shipping limit date for handling the order over to the logistic partner.
price: item price
freight_value: item freight value item (if an order has more than one item the freight value is splitted between items)
order_id: unique identifier of an order.
payment_sequential: a customer may pay an order with more than one payment method. If he does so, a sequence will be created to accommodate all payments.
payment_type: method of payment chosen by the customer.
payment_installments: number of installments chosen by the customer.
payment_value: transaction value
review_id: unique review identifier
order_id: unique order identifier
review_score: Note ranging from 1 to 5 given by the customer on a satisfaction survey.
review_comment_title: Comment title from the review left by the customer, in Portuguese.
review_comment_message: Comment message from the review left by the customer, in Portuguese.
review_creation_date: Shows the date in which the satisfaction survey was sent to the customer.
review_answer_timestamp: Shows satisfaction survey answer timestamp.
order_id: unique identifier of the order.
customer_id: key to the customer dataset. Each order has a unique customer_id.
order_status: Reference to the order status (delivered, shipped, etc).
order_purchase_timestamp: Shows the purchase timestamp.
order_approved_at: Shows the payment approval timestamp.
order_delivered_carrier_date: Shows the order posting timestamp. When it was handed to the logistic partner.
order_delivered_customer_date: Shows the actual order delivery date to the customer.
order_estimated_delivery_date: Shows the estimated delivery date that was informed to the customer at the purchase moment.
product_id: unique product identifier
product_category_name: root category of product, in Portuguese.
product_name_length: number of characters extracted from the product name.
product_description_length: number of characters extracted from the product description.
product_photos_qty: number of product published photos
seller_id: seller unique identifier
seller_city: seller city name
seller_state: seller state
product_category_name: category name in Portuguese
product_category_name_english: category name in English
product_order_ct: counts of orders that product is in
delivery_days: delivery time (from purchase date to delivery)
delivery_days: delivery time (from purchase date to delivery)

order_seller_ct: number of sellers in an order
order_item_ct: number of items in an order
order_value: total price of the order
order_freight: total freight value of the order
product_order_ct: number of orders a product is in
seller_order_ct: number of orders a seller is in