Title: Analysis of Brazilian E-Commerce Market

Authors: Harshada Sasturkar, Azim Tamboli, Sakshi Kulhari, Mrunal Malekar, Rohit Nair

Summary

The Brazilian E-Commerce dataset, provided by OList Store, consists of data on about 100K orders placed from 2016-2018 at various marketplaces. The data was provided by OList, the largest department store in Brazilian marketplaces.The dataset allows us to view the order details from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers.
The data is divided into multiple datasets including customers dataset, which has information about customers and location. It also contains the order items dataset which has information about the items that were placed in a single order, which can have multiple items. Each item order is fulfilled by a distinct seller. The order payments dataset contains information about payment options and amount.The reviews dataset has information about customer reviews and ratings.The orders dataset consists of information related to order delivery time.The products dataset has information about products sold by OList.
As part of this project, we aim to understand the customer market better by performing customer segmentation on the basis of RFM (Recency, Frequency and Monetary) metrics. Additionally, we conduct market basket analysis in order to better understand the consumer behavior based on the kind of products that are frequently purchased together. We plan to identify the frequency of orders placed and analyze it with day of week, month of year and season. The result is used to reveal the consumer behavior for the season with maximum orders. Along with this, we intend to estimate the products' delivery times based on the region of the buyer and seller. Delivery estimates are then visualized on a map. Text analysis is performed on user comments to gauge public opinion on purchases.

Methods

Market Basket Analysis: With the help of the orders items, products sold by seller and order details dataset we explored which set of items had higher probability to be purchased together by the customers.
Data Tidying and Preprocessing**:** There were 0.01% missing values in the orders data table, which were dropped. Apart from that names of Product categories were cleaned to remove underscores in them.

Analysis and Preliminaries: The data was analyzed to visualize the count of transactions day, hour, month and season wise. We are interested in performing market basket analysis on transactions that took place in the season with the most demand. Association Rule Mining and Apriori Algorithm was used to find a list of items that had the highest associations between them and were most likely to be purchased together by customers. Association rules states If the customer purchases item A, the likelihood of item B being purchased by the customer under the same Transaction ID is determined. There are 3 association metrics to measure association between items are - **Support**:The support metric was used to find the degree of association between two items. **Confidence:** Confidence tells us given the number of times item A occurs, it tells us how frequently A and B occur together. **Lift:** The strength of a rule over the occurrence of A and B at random.
Implementation of Apriori Algorithm and Association Rules**:** To figure out which products customers buy together the most we will primarily use three data sets which are order items dataset.csv, products dataset.csv, product_category_name_translation.csv and orders_dataset.csv.We calculated the support

metrics for product categories.Then we applied the Apriori Algorithm to find the frequent itemsets by setting min support threshold to 0.00005. It was observed that most items list has 1 item in it so pruning was performed by increasing support threshold (0.01) to include itemsets with more than 1 item. Using the frequent itemsets the association rules were found out with metric 'Confidence' and then were sorted using Lift Metric.Finally the association Rules were visualized using heatmaps using confidence Metric which returned the product categories which were most likely to be purchased together.

Geospatial Analysis: We used the geolocation dataset along with the customers and sellers datasets to plot the distribution of customers and sellers across states and cities. As part of this analysis, we also analyzed the peak order purchase times (month and hour of the day) for each state and their respective average delivery times using the orders dataset.

Data Tidying and Preprocessing**:**

1. Checked for the coordinates of the extreme points in the Brazilian territory. These coordinates provided geological constraints. Location data points outside these constraints were considered as outliers and were removed.
2. Removed the entries with missing timestamps in the orders dataset.
3. Converted the order purchased and order delivered timestamps into DateTime objects for calculations.

As there were multiple coordinate entries for each state, we grouped the data according to states and calculated the mean of latitude and longitude entries for each state. Similarly the mean coordinates for each city were computed. These were stored in separate state and city dataframes. We merged the customers and sellers data with above state and city data each to plot the distribution on map. Next, we imported the geojson shape file containing polygon objects representing each state's boundary. This was used to plot the basic map of Brazil onto which the choropleth showing the distribution of customers and sellers was added. For peak time analysis we joined the customers and orders datasets on customer id and grouped them by state to get the month and the hour of the day(averaged over all the months) at which maximum number of orders were purchased. We also calculated the average delivery time (in days) of the orders placed at the peak hours of each state respectively.

Delivery Time Estimation: We used the geolocation data along with orders, orders items, customers and sellers to estimate the amount of time taken to complete an order. The dataset had details on the estimated delivery time provided by the website to the customer. However, when we calculated the RMSE value of the estimated delivery time (provided by the website) and the actual time taken to complete the order, it came out to be around 360 which motivated us to explore this aspect a little more.
Data Tidying and Pre-processing**:**

1. Removed records containing data points outside of the Brazilian coordinates
2. Converted the order purchased and order delivered data into date time data
3. Filtered the orders data only for delivered items.
4. Every zip code contained multiple coordinate data against it. UNique coordinate point was assigned for each zip code by calculating the mean of latitude and longitude.
5. Deleted zip codes mapped to multiple states.

We merged the data together to get the customer and seller zip code for every order that was delivered. Post this, we pulled the zip codes and the corresponding latitude and longitude data for every record. We then calculated the distance between the customer and seller using the Geopy library.We then plotted the

Distance V/S Delivery Time graph and could notice the linear relationship between the two variables. We also calculated the average time taken to complete an order for every state based on two conditions:

a) When the customer and seller state is the same

b) When the customer and seller states are different

Post the exploration, we fit the data in a linear regression model, to predict the delivery time only on the basis of distance.

RFM Analysis and K means clustering:

A model known as recency, frequency, and monetary value (RFM) is used in marketing analysis to divide a customer base into several groups based on their buying tendencies. It analyzes customers' frequency (how frequently they make purchases), recency (how long ago they made a transaction), and monetary value in particular (how much money they spend). On the processed data, we apply the k-means clustering algorithm to group the customers. We analyze, then group, and determine what each cluster stands for. The list of products that loyal customers have purchased is then analyzed after we choose the cluster that best reflects them.

Text Analysis:

Data tidying: The first step in Text analysis is to tidy the data. Since Review data is user written, it consists of a large number of irregularities that need to be cleaned. This may include numeric data, emails, usernames, special symbols etc. These inconsistencies are deleted from the review data. We then tokenize the data to split the text into tokens of words and bigrams. Since textual data usually contains stop words, these are removed to create a structured dataset from unstructured text data.

Sentiment analysis: Using the ratings provided by the users, each review is categorized as positive or negative reviews (sentiments). We use rating 4 and 5 as positive and 1 and 2 as negative. A rating of 3 can be considered as neutral, thus these are ignored. We further analyze the words and bigrams related to both the positive and negative comments.

Modeling: Since there are reviews in the dataset with no rating, we may need to predict user sentiments based on reviews written. For this, we compare 3 classification models to select the one that provides the most accuracy**.**

Results

The results obtained on performing **Market Basket Analysis** are as below:

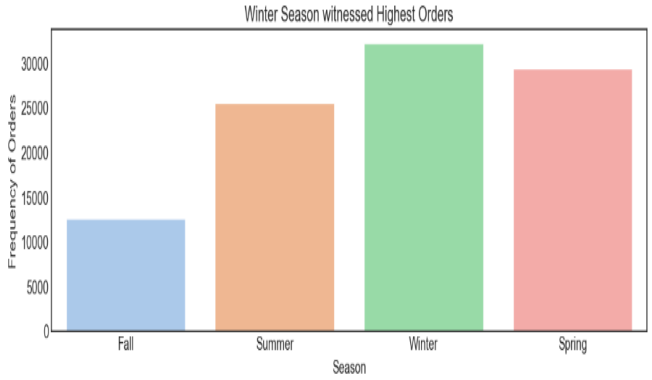


Figure 1: Visualizing frequency of transactions according to Seasons

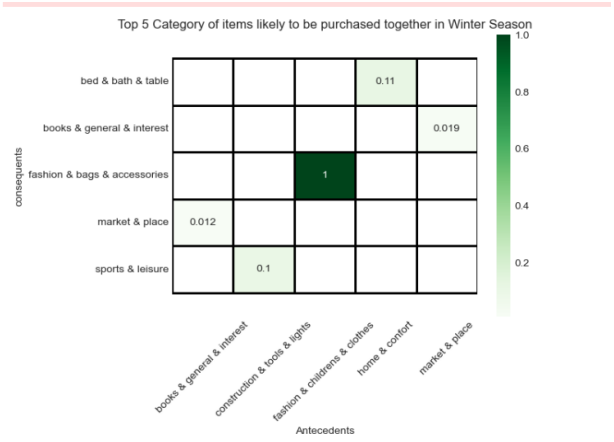


Figure 2: Top 5 results by Association Rule Mining

The results obtained on performing **Geospatial Analysis** are as below:

Figure 3: State wise distribution of customers and sellers

Figure 4: City wise distribution of customers and sellers

Figure 5: Histograms showing distribution of states with peak months and peak hours of the day for order purchase.

state	peak_time_hour	average_delivery_time_days
RR	0	6.708333
SP	16	8.701200
DF	16	12.232331
MG	16	12.318712
PR	11	12.343957
ES	11	14.467014
SC	21	15.124645
MS	17	15.297414
RS	20	15.540625
GO	13	15.574405

Table 1: Average Delivery time (days) for the peak hour of each state.

The results obtained from **Delivery Time Estimation** are as below

Figure 6: Relationship between Delivery Time and Distance

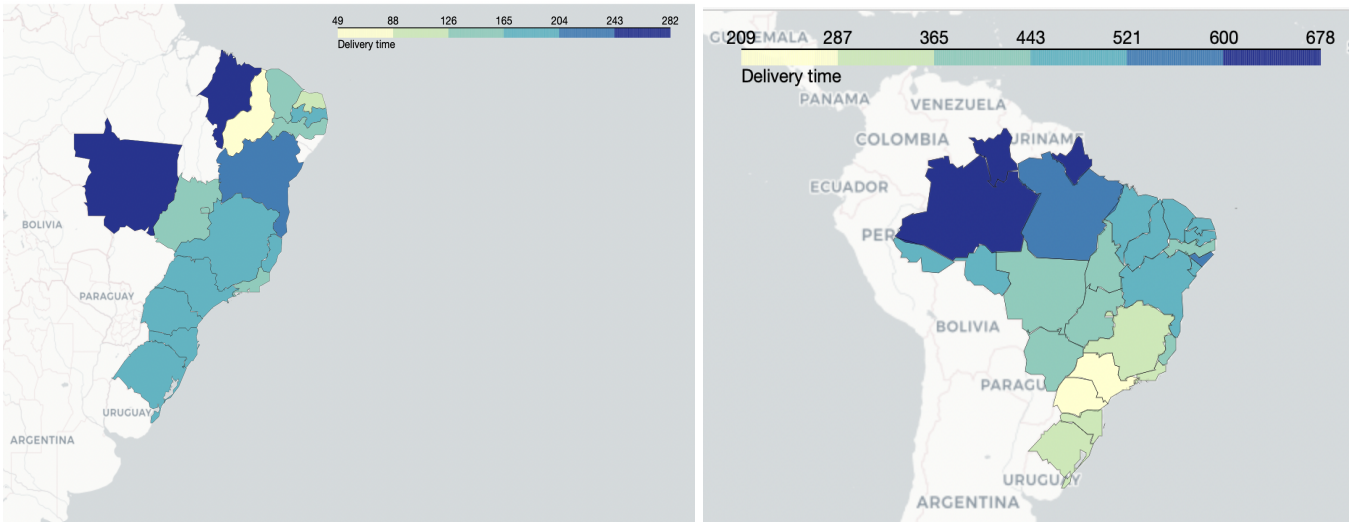
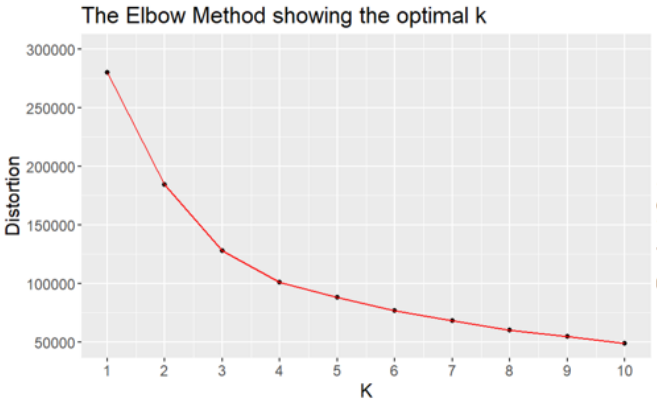


Figure 7: Delivery time when buyer and seller states are same (left) and different (right)

The results obtained using **RFM Analysis and K Means Clustering** are as below:



Cluster <int>	RecencyMean <dbl>	FrequencyMean <dbl>	MonetrayMean <dbl>
1	169.9085	1.000000	292.74522
2	432.4044	1.022223	122.24193
3	236.6726	2.625945	597.94873
4	152.4270	1.012098	65.24922

Figure 8: line plot showing the distortion from the centroid when k clusters are formed and avg value of RFM when k mean applied with 4 clusters.

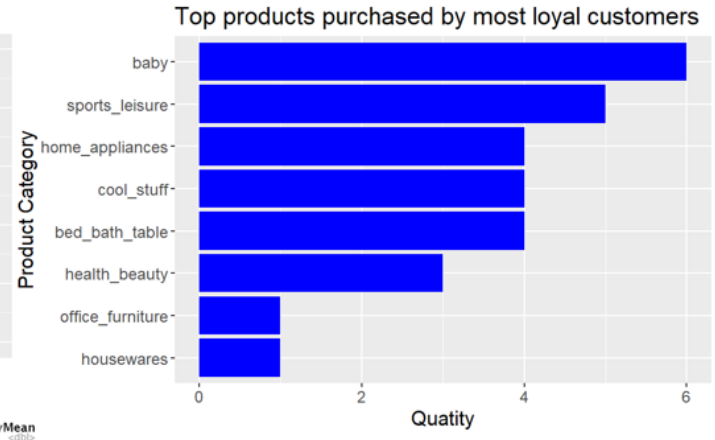


Figure 9: Histogram showing the type and average quantity of products purchased by the most loyal customers

The results obtained on performing **Text Analysis** are as below:

Figure 10: Most common positive words: Considering only the positive bigram comments, we notice that most reviews consist of the words “before deadline”, “arrived before” , “fast delivery”. This shows that delivery of the product before the expected delivery date is an important aspect to most customers. This speaks to the efficiency of the E-commerce website delivery. Related to the actual products itself, we see that customers are most satisfied when the quality of the product is good.

Figure 11: Most common Negative words: In the negative reviews, delivery time has been mentioned a lot again. Customers have given a bad review if the product was delivered late.

Figure 12 : Comparison of classification model accuracy

Logistic Regression : Model score on Testing Data - Test Accuracy = 91.045434%
Naive Bayes : Model score on Testing Data - Test Accuracy = 90.074989%
Support Vector Machines : Model score on Testing Data - Test Accuracy = 92.368769%
Comparing the three models, as expected, the Support Vector Machine model has the most accuracy. We can use SVM for sentiment prediction**.**

Discussion

We have used Market Basket Analysis on Brazilian Ecommerce Data to understand the customer buying pattern using the transaction data, purchase history, products and categories data and to understand which products are likely to be brought together by the customers. In Market Basket Analysis, we have used Association Rules and Apriori Algorithm as well to study associations between product categories. This can be helpful to the retail sector who can use it for upselling and cross selling to sell more items. The results that were obtained showed that Fashion Children Clothes and Fashion Bags were likely to be bought together by customers in the Winter season. Probably because winter season comes with Christmas it can be said that during shopping more customers were buying these two category items together. It showed 1.0 confidence metric which showed that 100% of the customers who bought fashion children's clothes also bought fashion bags with it. This market basket analysis on ecommerce data can be used to boost sales, help understand customer buying patterns.The results we obtained showed us that customers are more likely to shop more in the winter season. These results can be used by the sellers to accordingly increase the stock of these items in winter season, or offering promotional discounts on these items to be purchased together. In stores, they can place these items of these product categories close to each other as they have a high probability to be bought together.As part of future work we can include the time and study patterns with time.

The distribution plots obtained in the **geospatial analysis** can help understand whether the supply is meeting the demand and where the business could be expanded. We can see that the distribution of both customers and sellers is highly skewed with SP being the state with the highest number of both. The states RJ and MG have far more customers as compared to the sellers.The state PR has a lot fewer

customers than sellers. There are a couple states showing no sellers for their customers which is not an ideal case. The city wise distribution shows that there are alot of cities with customers who could potentially have more sellers as currently most of them are concentrated around only a few of them. Thus, more attention could be given to these regions to improve the business. The **peak time analysis** plots show that the majority of states see maximum orders in the month of may and july. Similarly maximum orders were placed at 4 pm (16) and 11 am. Another important observation is that there were no orders purchased between 12 am (0) and 11 am. We've also displayed the top 10 states with lowest average delivery time(days) for their respective peak hour. Analyzing this can help determine the rush hours for the business and find solutions on how to decrease the delivery time by better management of resources. During this analysis we noticed that due to the size of the datasets, we weren't able to render maps showing every single data point and could only plot data grouped together. This can be worked on to get more precise distributions of customers and sellers within each city as well.

The outputs obtained from the **delivery time estimation** pointed towards a linear relationship between the distance and delivery time. Post fitting the data in a linear model, the RMSE for the predicted and actual values came out to be around 101 which is significantly better as compared to the estimated delivery time provided by the website. Hence, the model can be used in place of the existing methods to provide a better delivery time estimation to the customers leading to better customer satisfaction and hence retention. We also noticed that the states like Piaui (PI), Rio Grande do Norte (RN) and Goiás (GO) have the lowest delivery time when the seller and customer states are same while the states like Sao Paulo (SP), Puerto Rico (PR) and Minas Gerais (MG) have the lowest delivery time when the seller and customer states are different. In future we can study the causes leading to delay in the deliveries.

We used RFM analysis to categorize customers into groups, with each group representing a different pattern. We then designated one group as the most loyal customers and another as the least loyal customers. The data set of loyal customers, along with the products purchased by them, will assist shopkeepers to increase profits by focusing on these products. From the results of Text analysis, we notice that most customers are concerned about the quality of the product. By improving the quality of products bought by the most loyal and repeating customers, sellers can expect an increase in sales.

Statement of contributions

Mrunal Malekar: Worked on EDA , Market Basket Analysis
Harshada Sasturkar : EDA, Geospatial Analysis
Sakshi Kulhari : EDA, Delivery Time Analysis, Delivery Time Prediction
Azim Tamboli : EDA, RFM analysis, K-mean clustering
Rohit Nair : EDA, Text analysis , Sentiment Analysis , Sentiment prediction based on user review

References

1. https://towardsdatascience.com/exploratory-data-analysis-in-r-for-beginners-fe031add7072
2. https://www.techtarget.com/searchdatamanagement/definition/RFM-analysis
3. https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce
4. https://medium.com/@24littledino/association-mining-leverage-in-python-86b0e476edeb
5. https://github.com/codeforgermany/click_that_hood/blob/main/public/data/brazil-states.geojson

Appendix

1. Link to program code - https://github.com/ronair212/Analysis-of-Brazilian-E-Commerce-Market