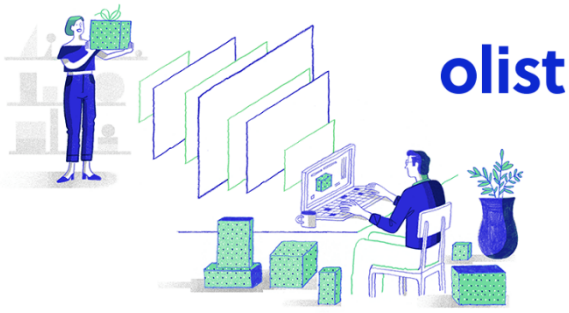
Uncovering Brazilian Shopping Trends: A Statistical and Machine Learning Approach to Enhance Olist's E-commerce Performance

May 30, 2024

Abstract

In this project, we first conduct a review of Olist's business performance and apply the time series analysis for the order volume records to uncover the underlying shopping patterns of Brazilian customers. Secondly, we conduct a thoughtful linear model with review score as the response variable. Furthermore, we implement sentiment analysis using the advanced BERT model to enhance our dataset and employ the Random Forest, XGBoost, and LightGBM models to predict review scores. The four models aim to identify gaps between customer expectations and the real services provided and explore how to boost review scores. Afterward, we explore the monthly sell frequency of different product categories, factors influencing the freight value most and the geographical distribution of sellers and customers. Finally, based on the above exploration, we provide some suggestions to Olist company from the perspective of marketing means, regional publicity and regional construction.

1. Background Settings



Olist is a Brazilian e-commerce marketplace integrator that connects small businesses with large product markets, facilitating entrepreneurs in expanding their customer base. The company provides SaaS licenses to small brick-and-mortar businesses, enabling them to gain market share nationwide.

In this project, We first conduct a review of Olist's business performance and uncover the underlying shopping patterns of Brazilian customers. Obviously, Olist's attempt was a great success. Thus, we wonder whether there exists any gaps between customer expectations and the real services provided based on the analysis of review score and how to boost customers' review scores. And finally, based on our exploration, we provide some suggestions to the Olist company from different aspects.

2. Business Performance

We draw histograms about the order volume, sales and number of active users per month for the Olist company to have an overall sketch of its business performance.

2.1. Order Volume

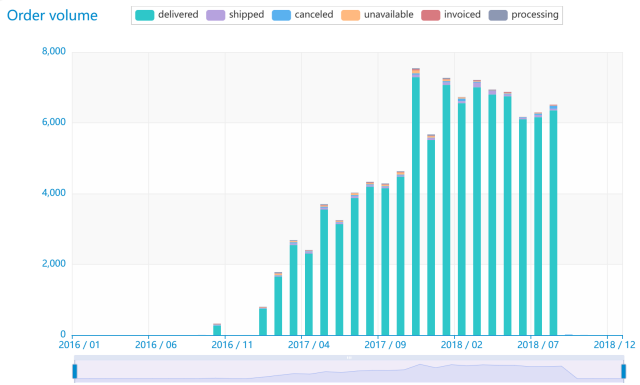


Figure 1. Order Volume

Our dataset collects order records starts from 2016/10 to 2018/08. Apparently, an increasing trend showed in the order volume time plot. What's more, we can see an outlier in the graph. The highest order volume occurred in November

2017, this is due to Black Friday.

Important shopping festivals in Brazil include Christmas, Black Friday, etc. Black Friday is a colloquial term for the Friday after Thanksgiving. It traditionally marks the start of the Christmas shopping season. Many stores offer highly promoted sales at discounted prices.

Pro tip: Brazil spends big on holidays.

As delivered orders account for the majority of orders in all order statuses, we only considered the orders that were delivered in the following analysis.

2.2. Sales

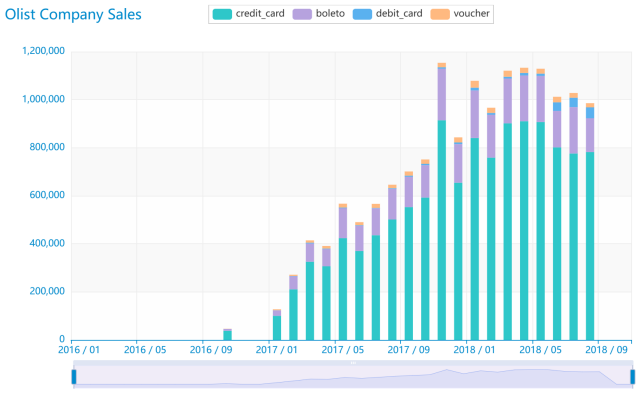


Figure 2. Sales

Similar trend as in order volume time plot. From sales time plot, the most widely used type of payment is Credit Card, followed by local payment boleto. The increase in Credit Card payment is also the most significant compared to other payment methods.

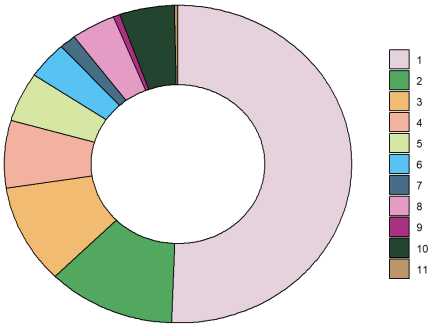


Figure 3. Payment Installments

From the pie plot for payment installments, half of the orders are paid in installments(label '11' in the above pie graph represents payment installment greater than 10). Brazilians are keen to spend in installments, which in turn stimulates an increase in sales.

Pro tip: The most common payment methods in Brazil? Credit and installments.

2.3. Active Users

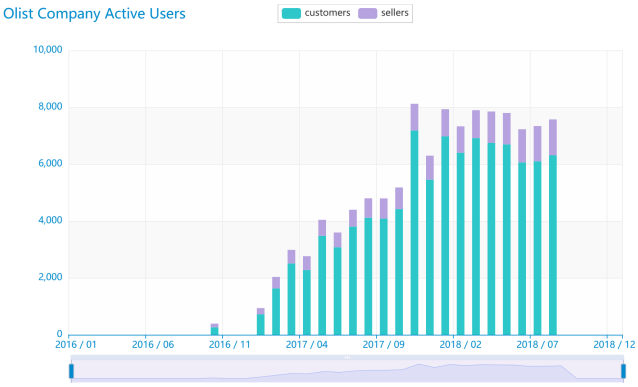


Figure 4. Active Users

We draw histogram of the number of active users including both the customers and sellers per month. Over time, the number of active customers continually increase until it stabilized at 6,500+ per month in 2018 and the number of active sellers stabilized at 900+ per month.

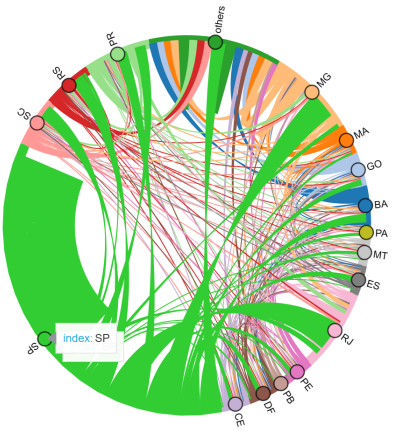


Figure 5. Seller-to-buyer geographic flow

The above chord diagram displays the geographic flow from seller's state to customer's state. It shows that SP(Prefeitura de St. Paul in portuguese and São Paulo in English) is the state with the most active trading flows.

Pro tip: The state of São Paulo is the commercial center of Brazil.

2.4. Time Series Analysis

For the order records, it is a time series. Time-series analysis happens when we consider part or the entire time series to see the "bigger picture." Time series analysis helps organizations understand the underlying causes of trends or systemic patterns over time.

Since the abnormal order volume during Black Friday, we cut the timeline into three pieces: period from 2017/01/09 to 2017/10/31, period from 2017/11/24 to 2017/12/31 and period from 2018/01/01 to 2018/08/26. And apply time series analysis respectively. The first one is used to explore the increasing trend due to the development of Olist company, the second one is used to explore the decreasing trend after the crazy Black Friday, and the last one is used to forecast the order volume for the future four weeks.

To correctly analyze time-series data, we need to look at the three components of a time series:

- Trend: increasing or decreasing long-term movement of the time series.
- Seasonality: regular periodic occurrences within a time interval.
- Noise: random variation in the series.

2.4.1. Time Series from 2017/01/09 to 2017/10/31

Firstly, we construct a time series about the daily order volume from 2017/01/09 to 2017/10/31 to depict patterns of order volume in the early development stage of Olist company. There is no surprise to find an increasing trend in the time series plot6.

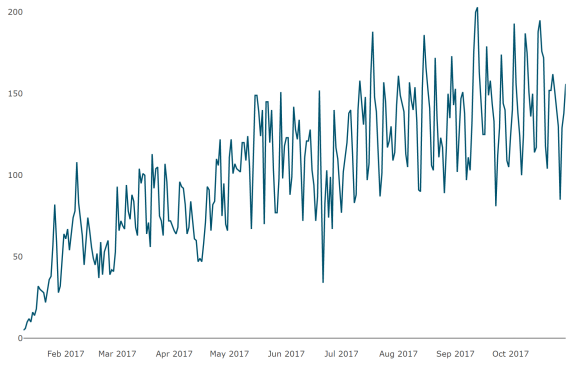


Figure 6. Daily Order Volume from 2017/01/09 to 2017/10/31

We set the frequency of the time series to 7 and draw the acf plot to check whether there exists the seasonality. The following acf plot peaks at lag 1, 2, 3 etc and decrease slowly, suggesting the existence of the seasonality.

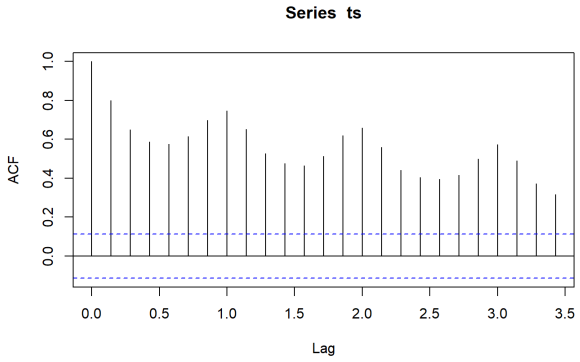


Figure 7. ACF Plot

For a visual look at the seasonality, we draw the following boxplot and heatmap. In the boxplot, the seasonality appears to be a significant drop on Saturday and a rise on weekdays.

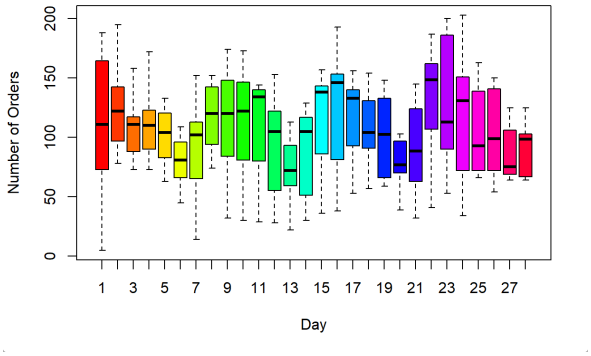


Figure 8. Boxplot for Daily Order Volume

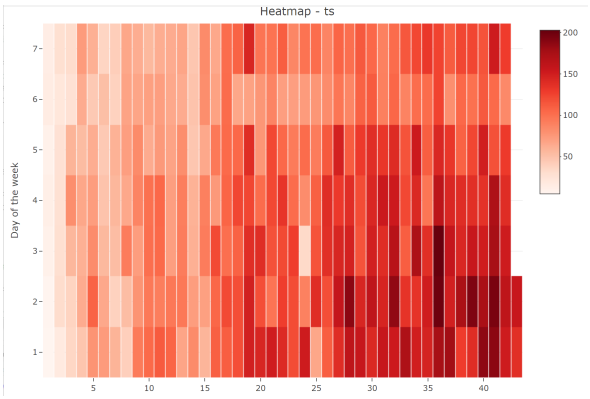


Figure 9. Heatmap for Daily Order Volume

This pattern is more intuitive in the heatmap, where Saturday and Sunday have lighter color than weekdays.

Pro tip: Brazilian consumers shop more on weekdays.

2.4.2. TIME SERIES FROM 2017/11/24 TO 2017/12/31

Next, we construct a time series about the order volume per hour from 2017/01/09 to 2017/10/31 to depict the impact of Black Friday on order volume. Apparently, an decreasing trend showed in the times series plot.

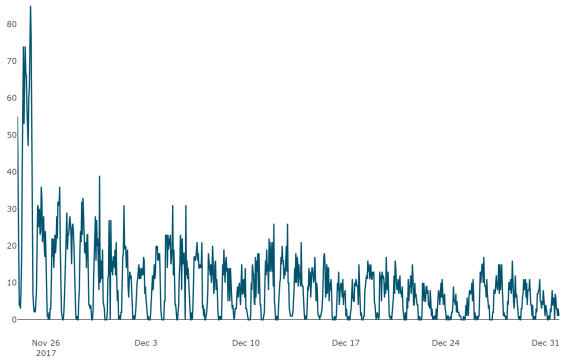


Figure 10. Time Series Plot for Time Period within 2017/11/24–2017/12/31

We set the frequency of the time series to 24 and draw the acf plot to check for the seasonality. The ACF plot peaks at lag 1, 2 etc, which shows there exists the seasonality.

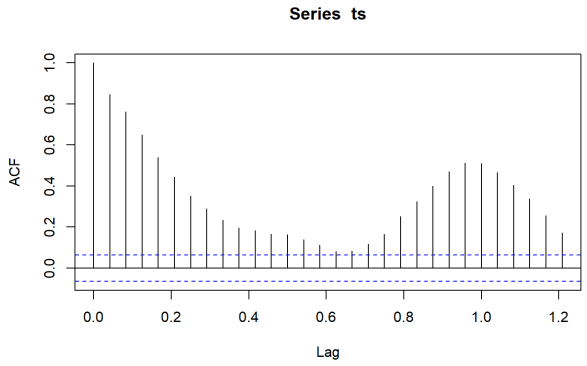


Figure 11. ACF plot

Similarly, we draw the following boxplot to depict the seasonality. In the boxplot, the number of orders peaks during the day and experienced a very significant drop at midnight, and then a significant rise in dawn. In addition, the plot also support the previous conclusion that Brazilian consumers shop more on weekdays. Note that the first day is Friday in Figure12 since our time series start at Black Friday(November 24th).

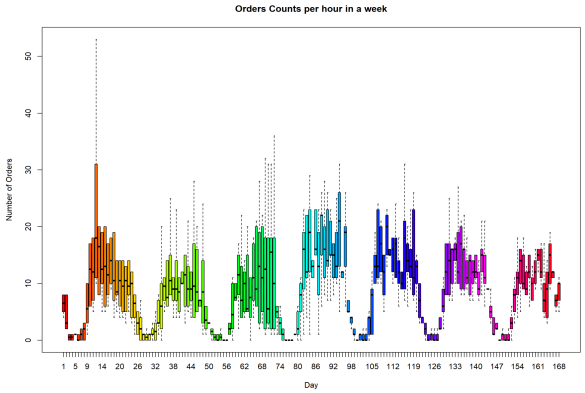


Figure 12. Boxplot for Orders Counts per hour in a week

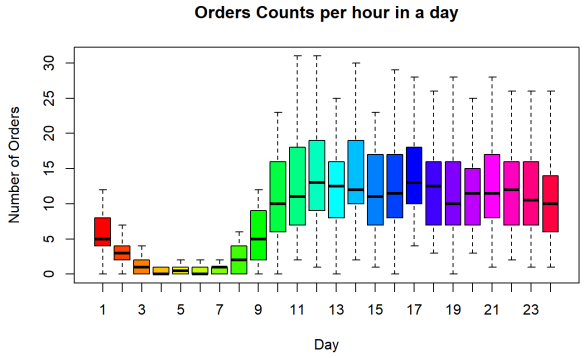


Figure 13. Boxplot for Orders Counts per hour in a day

2.4.3. TIME SERIES FROM 2018/01/01 TO 2018/08/26

Finally, we construct a time series about the order volume per day from 2018/01/01 to 2018/08/26 to forecast the order volume in the future four weeks. Though our forecasting for year 2018 is meaningless since we are now at 2024. Our purpose here is to set up an formal analysis procedure for forecasting order volume.

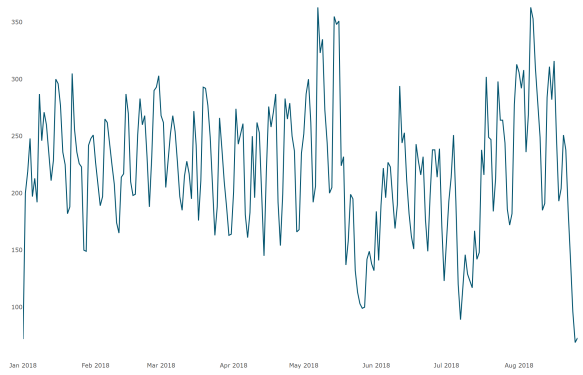


Figure 14. Time Series Plot for Time Period within 2018/01/01 - 2018/08/26

It is evident that there is no trend in the third time series. But the seasonality does exist based on the previous analysis. We set the frequency of the time series to 7 and the acf plot peaks at lag 1, 2 etc.

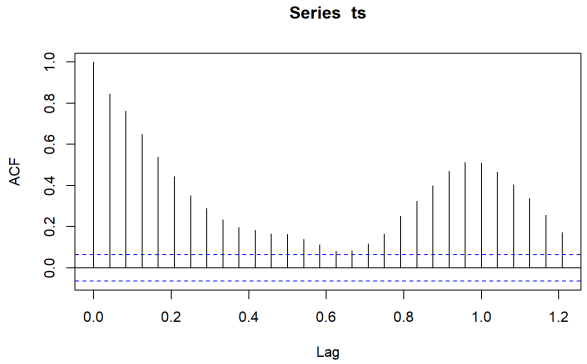


Figure 15. ACF plot

A time series is said to be "stationary" if it has no trend, exhibits constant variance over time, and has a constant auto-correlation structure over time. After differencing the time series to eliminate the seasonality, we test whether the time series is stationary by performing an augmented Dickey-Fuller test availabled in R package 'tseries'.

```
        Augmented Dickey-Fuller Test

data:  dts
Dickey-Fuller = -6.7419, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Figure 16. augmented Dickey-Fuller test

The p-value is less than 0.05, so we reject the null hypothesis and conclude that the time series is stationary after differencing at significant level 0.05.

Then we use the 'auto.arima' function in R package 'forecast' to fit an ARIMA model setting d=0, D=1. The best fitted model is ARIMA(1,0,1)(2,1,0)[7] with drift and its AIC value is 2298.37. The estimated coefficients are listed below:

Table 1. Parameter Estimation

	ar1	ma1	sar1	sar2	drift
$coefs$	0.8777	-0.1531	-0.5442	-0.2143	0.0771
$s.e.$	0.0448	0.0792	0.0699	0.0691	1.2395

We then examine the residual plot and its Q-Q plot to check the model fit and assumptions. Ideally, the residuals should behave like a random scatter of points forming an approximately constant width band around the identity line, satisfying constant variance, normality, and independence.

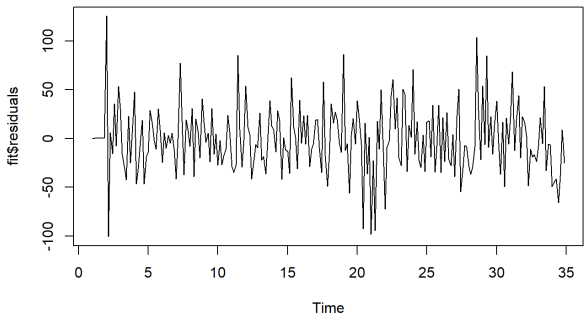


Figure 17. Residual Plot

From the residual plot, except for several residuals, the rest of the residuals behaves like white noise with zero mean and constant variance. Q-Q plot performs quite well except for points at both ends.

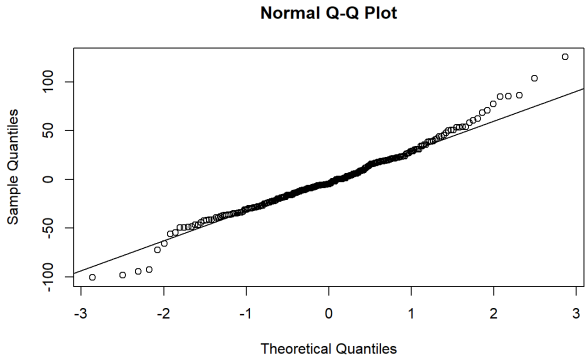


Figure 18. Q-Q Plot

Since Brazil spends big on holidays. We guess these outliers are due to holidays. Overall, our model is already satisfactory. Based on the current model, now we are ready to forecast the order volume for the next four weeks.

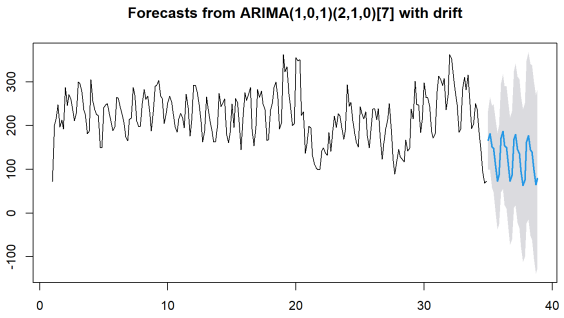
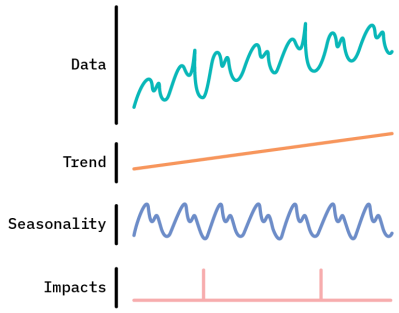


Figure 19. Forecasting the order volume

3. What is a structural time series?

Many time series display common characteristics, such as a general upwards or downwards trend; repeating, and potentially nested, patterns; or sudden spikes or drops.



Popular open source packages like bsts (Bayesian Structural Time Series, in R) and the TensorFlow Probability sts module support the state space model formulation of structural time series.

Impact effects: Some time series have discrete impact effects, active only at specific times. For instance, sales for some consumer products are likely to peak strongly on Black Friday. This isn't part of a weekly recurring pattern; sales don't peak to the same level on every Friday. The date of Black Friday also moves annually. However, whenever it is Black Friday, sales will spike.

This kind of component is especially useful for **modeling holidays**, which occur every year, and often on a different day of the week.

In order to learn such effects, we must have several examples of the event or holiday. Otherwise, we'll introduce a new parameter for a single data point, and the component will also fit any extra noise at the time it is active.