Olist Data Analysis

Ruimin Li, Yihan Zhao, Rui Liu, Sihui Li
Southern University of Science and Technology

1 Abstract

This paper aims to analyze a public e-commerce dataset provided by Olist on the Kaggle platform. Utilizing SAS, Python, and R, the dataset undergoes data processing and is examined through various dimensions and metrics. The analysis includes the development of a user segmentation RFM model and a logistics transportation model, ultimately summarizing the characteristics of Olist user consumption behavior and proposing recommendations.

In the last section, we conducted sentiment analysis based on sentiment polarity, sentiment polarity and review score of the reviews dataset in terms of sentiment distribution, geographical visualization and customer's satisfaction. Generally, customers care more about the transportation time and the product quality. Moreover, geographical analysis shows that the top 5 cities which receive the highest average sentiment polarity are nova lima, irati, balneario camboriu, colorado and artur nogueira.

2 Introduction

In recent years, e-commerce platforms have developed rapidly. Olist Store, the largest department store in the Brazilian market, connects small-scale enterprises from all over Brazil with a barrier-free and unified contract channel. These merchants can sell their products through the Olist store and use Olist's logistics partners to deliver them directly to customers.

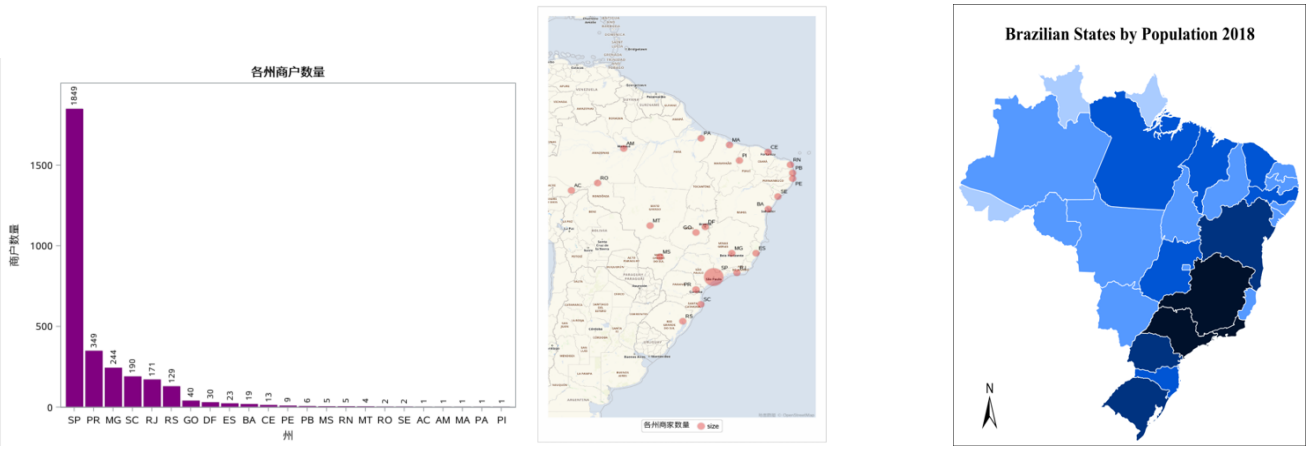
The dataset used in this analysis contains information on over 100,000 orders made between 2016 and 2018. There has been numerous analysis based on this dataset which are posted on kaggle, aiming to shed light on future development of e-commerce. Utilizing tools such as SAS and Python, this paper will analyze the historical sales data of the Olist store from multiple perspectives, including product line configuration, price customization, channel distribution, and sentiment analysis. The aim is to identify potential issues within the customer marketing system and devise targeted and feasible marketing strategies.

3 Sellers and Products

3.1 Sellers

3.1.1 Geographical Distribution of Sellers

Firstly, we draw a bar chart based on the number of sellers in each state. The number of sellers varies greatly from state to state. In St.Paul(SP), there are 1849 seller totally, there are also many sellers in Paraná(PR), Minas Gerais(MG), Santa Catarina(SC) and Rio de Janeiro(RJ), but most states only have few sellers. We also use the size of bubbles to show the geographical distribution of sellers on the map. Comparing it to the population map, we can find that the sellers concentrated in the southeast region, which is in line with the population.

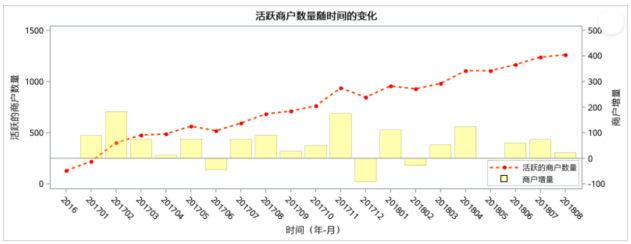


(a) Number of Sellers in Each State (b) Bubble Map (c) Population Map

Figure 1: Geographical Distribution

3.1.2 Number Change of Active Sellers over Time

The number of active sellers from September 2016 to August 2018 was calculated based on the order time. Due to some missing data, the months of September December 2016 were merged into one set of data. It can be seen that the number of sellers is showing a significant increasing trend. Due to the more than 200 merchants at the end of 2016 and the beginning of 2017, they quickly rose to over 1000 within a year, and can reach up to over 1200 at most. However, the increase in merchant growth is gradually stabilizing. Also, we use *PROC UCM* in SAS to built a simple time series prediction with trend and seasonality.



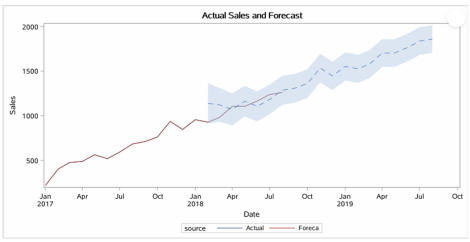
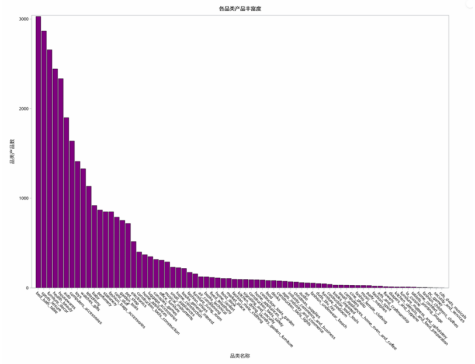


Figure 2: Number Change of Active Sellers over Time

3.2 Products

3.2.1 Variety of Products

We want to learn the variety of products for each category. The "bed bath table", "sports leisure", furniture dector", "health beauty" and "housewares" have variable products. More than half of categories only have few products. On the one hand, some categories themselves are relatively niche. On the other hand, considering the properties of online marketing, some categories such as flowers are not suitable for online sales, and some categories have not opened the online market. By computing, the correlation coefficient between product variety and category sales reached 0.97, and we can figure out that the relationship between variety and sales performance was very consistent on the scatter plot.



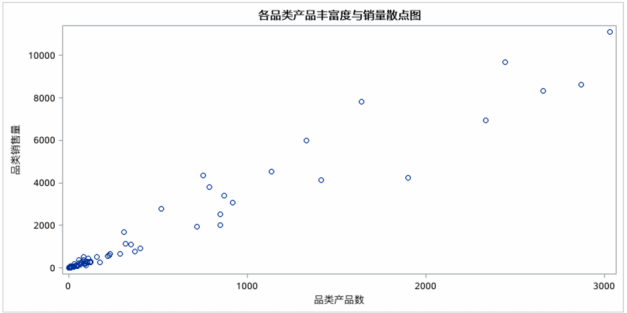


Figure 3: Variety of Products

3.2.2 Revenues and Sale

We calculated the top 15 product categories by revenues and by sales during the data period. It can be seen that the sales figures for "health beauty", "sports leisure", "furniture dector", "computers accessories", "watches gifts" and "bed bath table" are relatively good. We can also notice that the ranking of revenues and sales are not exactly the same.For example, the revenues of the computer category was not high, but it appeared in the 12th place in the sales ranking. Although the sales of watches gifts ranked second, it only ranked 7^{th} in terms of revenues.

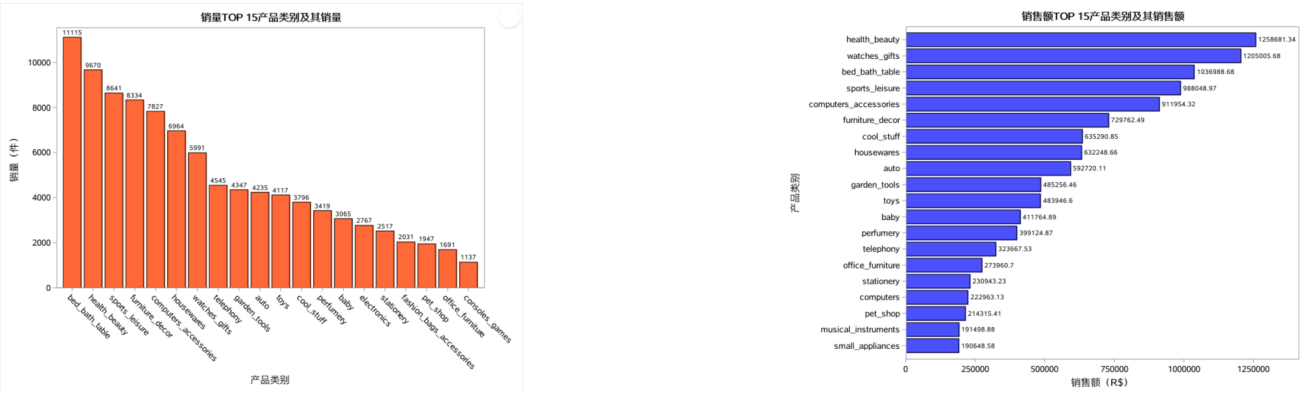


Figure 4: Top 15 Categories

Therefore, we need to consider the price of different categories. We use the following figure to show the prices, revenues and sales. It can be seen that although the sales of some household daily necessities are very high, its unit price is not high. In categories like computers and some small household appliances, although the sales volume is not very high, their unit price is very high. This shows that different types of sales strategies should be adopted for different categories of products, such as computers and small household appliances, which need to be more accurately marketed, and some daily necessities and consumables can be promoted on a large scale to increase their sales.

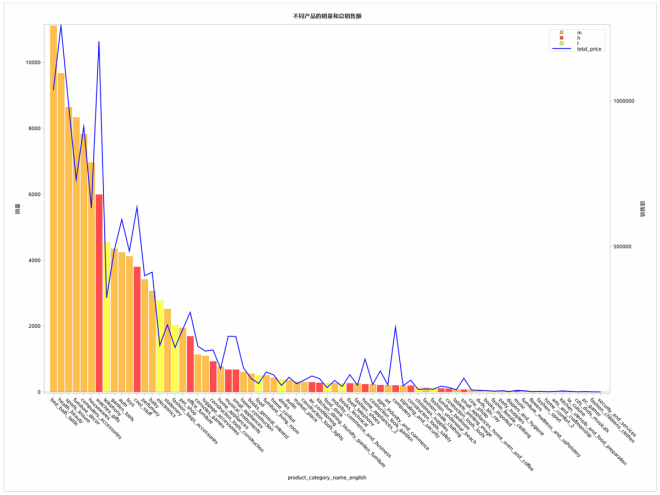


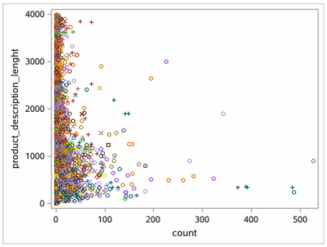
Figure 5: Prices, Revenues and Sales

3.2.3 Selling Frequency of Products

In this part, we want to see the selling frequency of each products. There are more than 18000 items have been sold only once, and more than 5000 items have been sold twice. Only a very small number of items can be sold more than 300 times. The most sold item reached 527 times. This shows that it is on OLIST, the head effect or long tail effect of commodities is very obvious. That is, there are a small number of products that can account for a very large share of sales, while most of the products only account for a small market share.

3.2.4 Prediction and Suggestions

In the next section, we would like to make a sales forecast. First of all, we guess that the sales volume will be related to the description of the product, the score, and other factors. However, the scatter plot does not perform well, and the calculated correlation coefficient is not high. So I thought about splitting the sales back into months and doing some transformations.



销量和几个变量的相关系数

CORR 过程

Pearson 相关系数, N = 32341				
	count	avg_price	product_photos_qty	product_description_lenght
count	1.00000	-0.03281	0.00385	0.00831
avg_price	-0.03281	1.00000	0.06008	0.19265
product_photos_qty	0.00385	0.06008	1.00000	0.10874
product_description_lenght	0.00831	0.19265	0.10874	1.00000

Figure 6: Sales and Description

According to the monthly sales volume changes of some products, it can be seen that the sales status of each product is seriously stratified, although the sales volume of some products fluctuates slightly, but most products have always maintained a certain sales level. By calculating the correlation coefficient matrix, it can also be found that the monthly sales volume has little to do with the variables of the product itself, including ratings, description length, number of pictures and price, but only related to historical sales data.

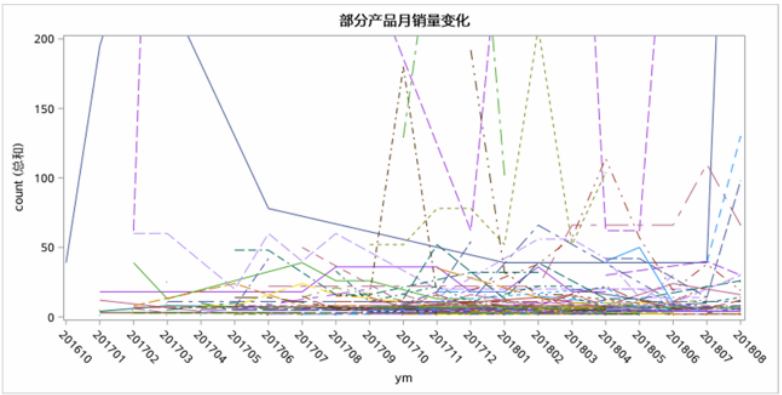


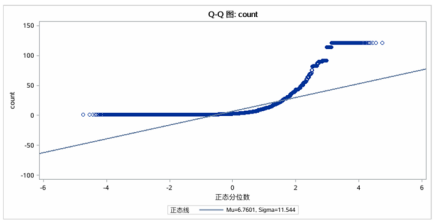
Figure 7: Monthly Sales Volume of Some Products

CORR 过程

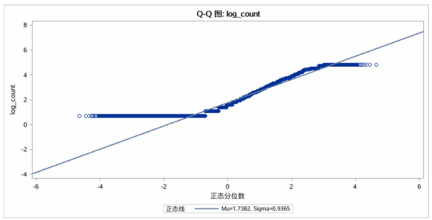
Pearson 相关系数, N = 581904														
	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm	price	review_score	freight_value	count	avg_price	avg_freight	last_ym_count
product_name_lenght	1.00000	0.06873	0.10961	-0.00173	0.05261	-0.03515	0.09904	-0.00368	-0.00875	0.00967	0.05018	-0.00884	0.00605	0.05038
product_description_lenght	0.06873	1.00000	0.10724	0.09520	-0.02367	0.07639	-0.11896	0.20482	0.01796	0.13025	-0.02829	0.20334	0.13589	-0.02847
product_photos_qty	0.10961	0.10724	1.00000	-0.01037	0.01557	-0.09423	-0.03029	0.03066	0.02488	-0.03006	0.02560	0.03188	-0.02416	0.02598
product_weight_g	-0.00173	0.09520	-0.01037	1.00000	0.42798	0.60497	0.47151	0.33573	-0.02172	0.58734	-0.02971	0.33617	0.64543	-0.02906
product_length_cm	0.05261	-0.02367	0.01557	0.42798	1.00000	0.18111	0.56899	0.13438	-0.01797	0.24810	0.01949	0.13064	0.28124	0.02009
product_height_cm	-0.03515	0.07639	-0.09423	0.60497	0.18111	1.00000	0.24064	0.20960	-0.02666	0.37182	-0.01557	0.20878	0.41357	-0.01541
product_width_cm	0.09904	-0.11896	-0.03029	0.47151	0.56899	0.24064	1.00000	0.15817	-0.00662	0.27038	0.05906	0.15403	0.30307	0.05942
price	-0.00368	0.20482	0.03066	0.33573	0.13438	0.20960	0.15817	1.00000	0.00814	0.38266	-0.05474	0.99053	0.41395	-0.05344
review_score	-0.00875	0.01796	0.02488	-0.02172	-0.01797	-0.02666	-0.00662	0.00814	1.00000	-0.03384	-0.01520	0.00699	-0.00813	-0.01507
freight_value	0.00967	0.13025	-0.03006	0.58734	0.24810	0.37182	0.27038	0.38266	-0.03384	1.00000	-0.04024	0.38049	0.66470	-0.03954
count	0.05018	-0.02829	0.02560	-0.02971	0.01949	-0.01557	0.05906	-0.05474	-0.01520	-0.04024	1.00000	-0.05992	-0.05761	0.97760
avg_price	-0.00884	0.20334	0.03188	0.33617	0.13064	0.20878	0.15403	0.99053	0.00699	0.38049	-0.05992	1.00000	0.41381	-0.05804
avg_freight	0.00605	0.13589	-0.02416	0.64543	0.28124	0.41357	0.30307	0.41395	-0.00813	0.66470	-0.05761	0.41381	1.00000	-0.05635
last_ym_count	0.05038	-0.02847	0.02598	-0.02906	0.02009	-0.01541	0.05942	-0.05344	-0.01507	-0.03954	0.97760	-0.05804	-0.05635	1.00000

Figure 8: Correlation Coefficient Table

Through logarithmic transformation, it can be seen that the data is difficult to approach normal due to the influence of heavy tail. Considering that the tail is heavily long, data deletion is meaningless, and only analysis and suggestions are made based on the results.



(a) Before transformation



(b) After Log Transformation

Figure 9: QQ Plots

Using a linear model and stepwise selection, it was found that only the previous month's sales volume and product category were significant in the model.

We can give some suggestions for this phenomenon. For the sellers, they need to provide more variety of product categories. For Olist platform, it should focus on products in the head and middle, through some activities and promotions to provide a better competition and sales environment for these products. It also should improve the function of product description and scoring.

4 Customers

4.1 Order Volume and Transaction Amount

First, we draw a line chart of daily order volume and transaction amount as follows, and we find that there is a sudden increase on November 24^{th}, 2017, which turn out to be Black Friday, a big shopping festival in Western countries.

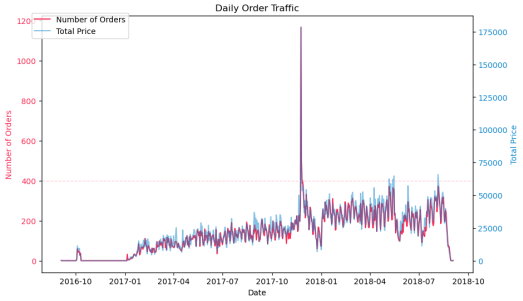
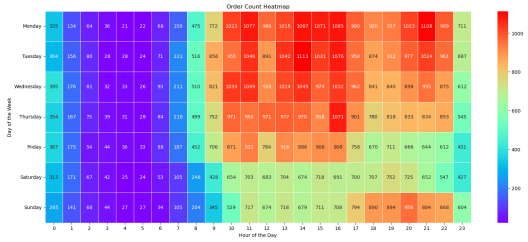
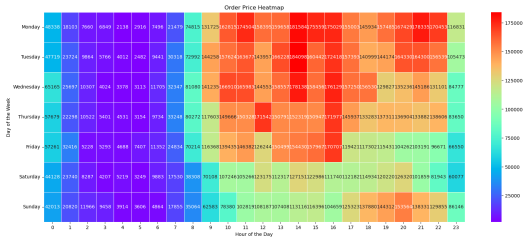


Figure 10: Daily Order Volume and Transaction Amount

After removing the data on November 24^{th}, 2017, we draw a heatmap of 7 days 24h daily and order volume transaction amount respectively. We speculate that people prefer offline shopping at the weekends and from about 18:00 on Sundays to the weekdays, they tend to buy products online.



(a) Daily Order Volume in 7 Days 24 Hours

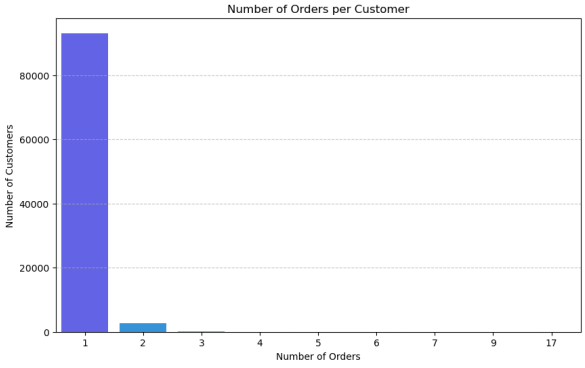


(b) Daily Transaction Amount in 7 Days 24 Hours

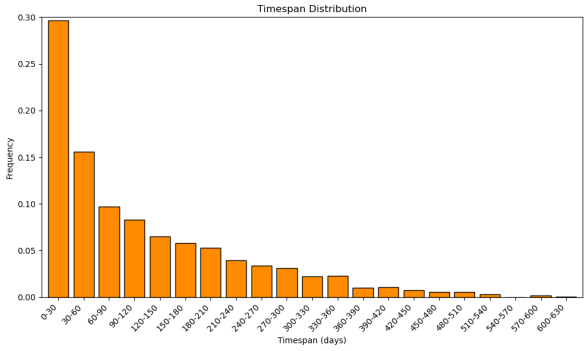
4.2 Repurchase

Then we want to study the way of customer repurchase. We draw the histogram of number of orders as (a) shows. After counting, only 3.12% customers have a second purchase and the proportion of continued purchases after two purchases is 8.41%. So we can conclude that customer loyalty is at a low level.

In the plot (b), we delete the 0 values make a histogram with width of a month of time difference between the first two purchases, which indicates that most of the second purchasing behavior happens in the first 2 months. So the platform should should promote information to users and strive to retain customers in the first 2 months.



(a) Histogram of Number of Orders



(b) Time Difference Between the First Two Purchases

4.3 Payment Method

From the ring diagram, we can discover that the most popular payment method is "credit card", which is convenient. As for the second popular one, "boleto", it's a safe offline payment method and is more trusted by Brazilians than credit card, for it doesn't require customers to fill out card information online. Hence, the unfriendly nature of online payments towards "boleto" may explain the phenomenon of low online shopping intentions.

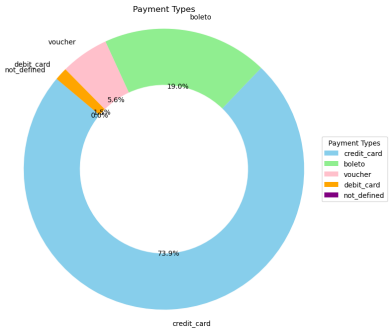


Figure 13: Ratio of Payment Method

4.4 Installment Method

As for the installment method, we use linear model to describe. From the summary we know that the interaction of "credit card" and payment value is significant for the model, which means that the payment type "credit card" affects the impact of the payment amount on the installment method. This may be because the bank may offer different discounts for different installment method. And it's surprising that, despite of other factors like income and habit of customers, only the two explanatory variables and their interaction can take up about 30% of the installment method.

```
> ModelInteract <- lm(payment_installments ~ payment_value + payment_type + payment_value * payment_type, data = payments)
> summary(ModelInteract)

Call:
lm(formula = payment_installments ~ payment_value + payment_type +
    payment_value * payment_type, data = payments)

Residuals:
    Min      1Q  Median      3Q     Max
-67.750  -1.816   0.000   0.246  20.115

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            1.000e+00  1.952e-02   51.24   <2e-16 ***
payment_value                         -4.808e-16  7.560e-05    0.00        1
payment_typecredit_card                1.718e+00  2.201e-02   78.06   <2e-16 ***
payment_typedebit_card                -1.558e-13  6.992e-02    0.00        1
payment_typevoucher                   -1.608e-13  3.953e-02    0.00        1
payment_value:payment_typecredit_card  4.833e-03  8.412e-05   57.45   <2e-16 ***
payment_value:payment_typedebit_card   3.629e-16  2.482e-04    0.00        1
payment_value:payment_typevoucher      3.067e-16  2.695e-04    0.00        1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.271 on 103875 degrees of freedom
Multiple R-squared:  0.2858,    Adjusted R-squared:  0.2857
F-statistic:  5938 on 7 and 103875 DF,  p-value: < 2.2e-16
```

Figure 14: Summary of the Linear Model Established

4.5 Review Scores

About 80% of the customers give 4 or 5 to their online shopping experience, implying that most of the customers have a pleasant online shopping experience. And the heatmap shows that customers like to rate at noon and at midnight, so the platform can provide targeted push notifications to them at this time.

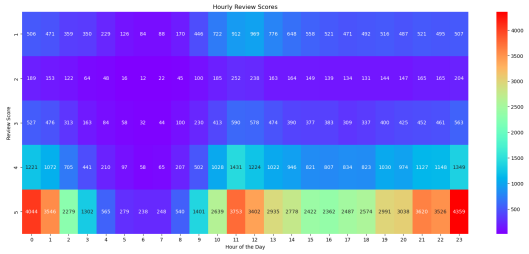
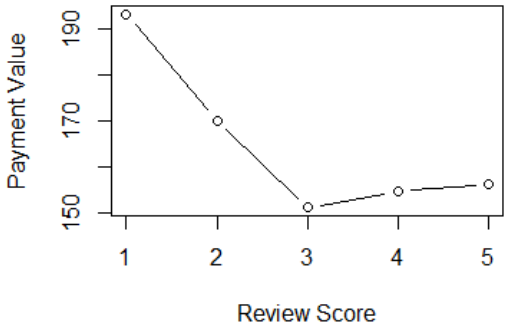


Figure 15: Distribution of Review Score in 24 Hours

Next, we want to know what factors affect the review scores. As for "Review Score" and "Payment Value", we can identify that for lower order amount, customers may rate moderately, when money gets higher, the score gets higher at first and then get much lower, this might because when people spend more money, they may have higher expectation, which ultimately transforms into dissatisfaction. And when we try to fit a linear model, it shows that there's little linear relationship between the factors.

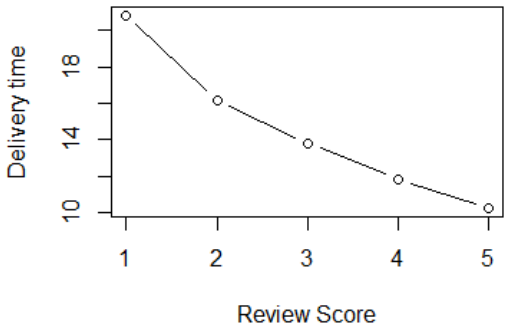


(a) Average Payment Value of Different Review Score

```
> cor(score$review_score, score$payment_value)
[1] -0.04211074
> ModelValue <- lm(review_score ~ payment_value, data = score)
> summary(ModelValue)

Call:
lm(formula = review_score ~ payment_value, data = score)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1920 -0.1820  0.8164  0.8345  2.5314

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.195e+00  5.131e-03  817.59   <2e-16 ***
payment_value -2.492e-04  1.905e-05  -13.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.284 on 96356 degrees of freedom
Multiple R-squared:  0.001773,  Adjusted R-squared:  0.001763
F-statistic: 171.2 on 1 and 96356 DF,  p-value: < 2.2e-16
```

(b) Linear Model of Review Score and Payment Value

As for "Review Score" and "Deliver Time", they show a strong positive linear relationship. In this situation, the platform ought to try to shorten the delivery time in order to gain a higher score.

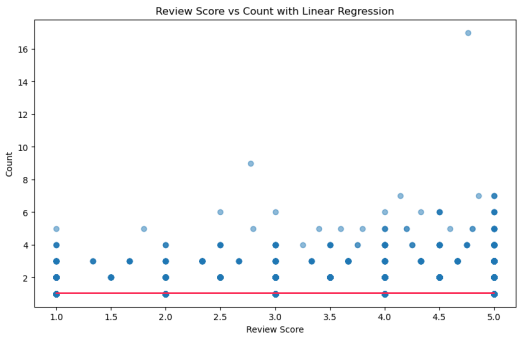


(a) Average Deliver Time of Different Review Score

```
> cor(score$review_score, score$deliver_time)
[1] -0.3336026
> ModelTime <- lm(review_score ~ deliver_time, data = score)
> summary(ModelTime)

Call:
lm(formula = review_score ~ deliver_time, data = score)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7019 -0.4301  0.5246  0.7511  8.7705

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.7019397  0.0063221   743.7   <2e-16 ***
deliver_time -0.0453069  0.0004125  -109.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.211 on 96356 degrees of freedom
Multiple R-squared:  0.1113,    Adjusted R-squared:  0.1113
F-statistic: 1.207e+04 on 1 and 96356 DF,  p-value: < 2.2e-16
```

(b) Linear Model of Review Score and Deliver Time

Now that we know something about the score given by consumers, we have interest in whether the scores have influence on repurchase behavior. Unfortunately, it looks as if no matter how much the customers score, they are not interested in buying from the platform.



(a)

```
> Model <- lm(score_reorder$count ~ score_reorder$review_score)
> summary(Model)

Call:
lm(formula = score_reorder$count ~ score_reorder$review_score)

Residuals:
    Min      1Q  Median      3Q     Max
-0.0411 -0.0411 -0.0411 -0.0394 15.9591

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.0369227  0.0026456 391.939   <2e-16 ***
score_reorder$review_score 0.0008272  0.0006153   1.344    0.179
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2549 on 95378 degrees of freedom
Multiple R-squared:  1.895e-05,  Adjusted R-squared:  8.466e-06
F-statistic: 1.807 on 1 and 95378 DF,  p-value: 0.1788
```

(b)

5 Logistics Status

5.1 Logistics Location

Firstly, I explored the geographical locations of logistics companies and plotted a distribution map of customers and merchants. I found that they are mainly concentrated in the lower right states of Brazil. Therefore, I suggest that logistics companies establish more bases in São Paulo, Rio de Janeiro, and Minas Gerais.

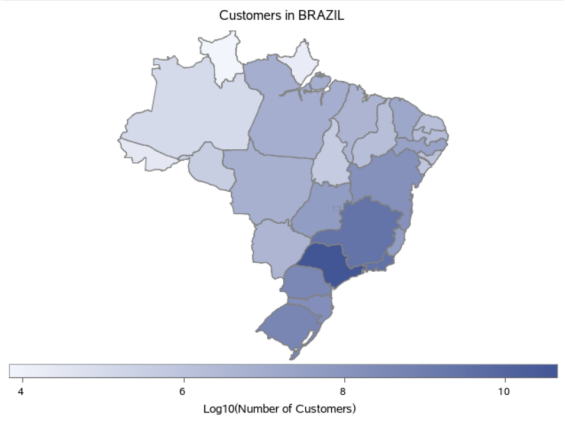


Figure 19: Customer

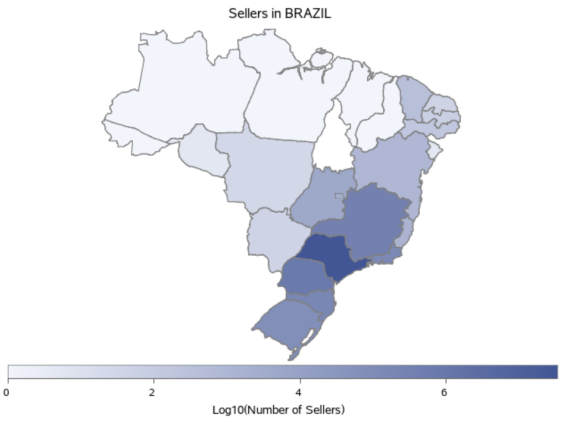


Figure 20: Seller

Then, I found that unlike the merchants, customers are also distributed in significant numbers in several states in the northwest. So, I found a population distribution map of Brazil from 2006 and discovered that many people are also distributed along their riversides. Therefore, it could also be suggested for logistics companies to develop a route along these rivers.

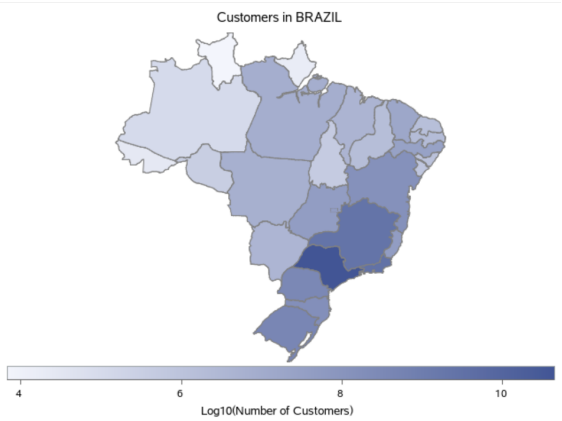


Figure 21: Customer

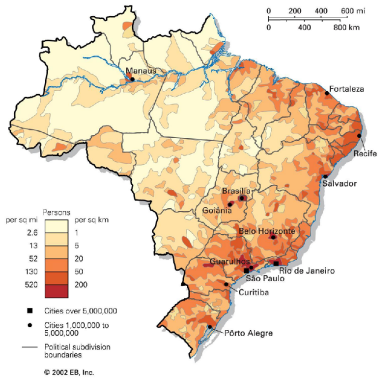


Figure 22: Population distribution from https://zh.brazilmap360.com

5.2 Logistics Time

About the logistics timing, I roughly found that it involves the process from purchase to approval, then to the shipping company, and finally to the customer's home. It's worth noting that the platform itself provides an estimated delivery time. Then, the entire order process from placement to delivery is particularly important for both customers and merchants. So, I attempted to explore and model whether it's possible to utilize the estimated delivery time provided by the platform to establish a linear model to help both parties predict the time it takes for orders to be fulfilled.

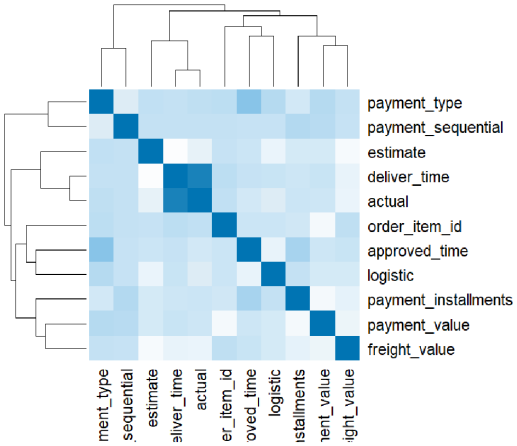
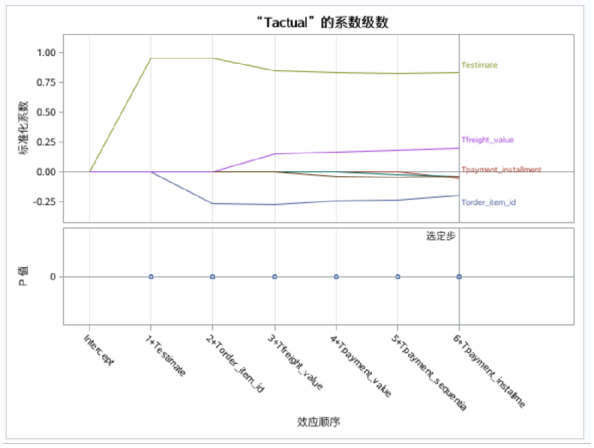


Figure 23: Correlation

In my initial variable screening, I found a very strong correlation between transport time and actual delivery time, which aligns with common sense. Furthermore, since delivery time is only known once the order is delivered, it wouldn't be feasible to use it in predicting order time. Therefore, I didn't include transport time as one of the variables for prediction in my model.

So, I began by first removing outliers and then checked for linearity, homoscedasticity, normality, and multicollinearity in the model. Afterward, I established a model. And finally, I used the stepwise method to select a model.



(a) Stepwise process

逐步选择汇总

步	进入的效应	删除的效应	引入效应数	调整 R 方	AIC	CP	SBC	F 值	Pr > F
0	Intercept		1	0.0000	1226055.38	4164899340	1120185.95	0.00	1.0000
1	Testimate		2	0.9167	962903.46	346783029	857043.60	1165339	<.0001
2	Torder_item_id		3	0.9869	767020.17	54431342	661169.88	567547	<.0001
3	Tfreight_value		4	0.9994	448222.96	2579585	342382.24	2044236	<.0001
4	Tpayment_value		5	0.9995	421838.19	1987196	316007.04	29964.5	<.0001
5	Tpayment_sequential		6	0.9997	382293.41	1334811	276471.83	47942.0	<.0001
6	Tpayment_installment		7	1.0000*	105887.01*	7*	75.00*	1334806	<.0001

* 准则的最佳值

(b) Stepwise process

After removing variables with very high tolerance, I was surprised to find that the R-squared value of my model reached 0.9968. However, through the previous stepwise procedure, it's evident that the main reason for this high fit is likely the accuracy of the predicted time provided by the platform. Therefore, the model ultimately fits very well.

均方根误差	11.29631	R 方	0.9968
因变量均值	3.08442	调整 R 方	0.9968
变异系数	366.23755		

参数估计

变量	标签	自由度	参数估计	标准误差	t 值	Pr > |t|	容差
Intercept	Intercept	1	1.53804	0.00037446	4107.32	<.0001	.
Tpayment_installments	payment_installments Transformation	1	-0.03531	0.00008840	-399.50	<.0001	0.20393
Tpayment_sequential	payment_sequential Transformation	1	-0.10689	0.00032155	-332.42	<.0001	0.42694
Tpayment_value	payment_value Transformation	1	-0.00069478	0.00000188	-369.93	<.0001	0.19057
Tfreight_value	freight_value Transformation	1	0.02522	0.00002971	848.76	<.0001	0.28380
Testimate	estimate Transformation	1	0.05937	0.00002865	2071.93	<.0001	0.21375

Figure 25: Final Model

6 Feature Extraction with the Reviews Comment Dataset

6.1 Sentiment Polarity, Subjectivity and Visualization

Sentiment analysis is a crucial part of text analysis, which assists with exploiting the inner aspects of the data. In this sention, we utilized *LeIA*, an adapted version of the sentiment processor *VADER* to delve into the dtataset *olist_order_reviews_dataset.csv*.

6.1.1 Sentiment Polarity

The reviews dataset has two basic feature. Firstly, both the comment title and comment message are included, but the latter one contains more information. We choose to drop NA values according to **comment message** in order to gain more information, and analyze the sentiment features of both the two columns. Secondly, the review dataset mainly consists of Portuguese and emojis, due to which we can't directly use *VADER*. Therefore, *LeIA* is used for sentiment polarity analysis. This is a tool adapted from the framework of *VADER*, and can handle sentences with emojis.

review_comment_title	review_comment_message	title_compound	title_neg	title_neu	title_pos	message_compound	message_neg	message_neu	message_pos
recomendo	aparelho eficiente. no site a marca do aparelh...	0.3612	0	0	1	0.4215	0	0.896	0.104
Super recomendo	Vendedor confiável, produto ok e entrega antes...	0.7506	0	0	1	0.6705	0	0.522	0.478
Não chegou meu produto	Péssimo	-0.296	0.423	0.577	0	0	0	1	0
Ótimo	Loja nota 10	0.3612	0	0	1	0	0	1	0
Muito bom.	Recebi exatamente o que esperava. As demais en...	0.4215	0	0.263	0.737	0.1901	0.062	0.831	0.107

Table 1: Sentiment polarity of review comment title and message.

6.1.2 Sentiment Subjectivity

For Portuguese text data, there are no mature toolkit like *textblob* to analyze the sentiment subjectivity. Due to the limitation of translation APIs and fares, we choose to use *textblob* to analyze the data without translation, hoping to provide a brief scratch of the sentiment polarity versus subjectivity plot. The sentiment subjectivity is classified as follows:

review_comment_title	review_comment_message	tb_title_Pol	tb_title_Subj	tb_message_Pol	tb_message_Subj
recomendo	aparelho eficiente. no site a marca do aparelh...	0	0	0	0
Super recomendo	Vendedor confiável, produto ok e entrega antes...	0.333333	0.666667	0.5	0.5
Não chegou meu produto	Péssimo	0	0	0	0

Table 2: Sentiment Subjectivity for title and message.

Combining the previous results, we plot the sentiment polarity versus sentiment subjectivity plot for both review comment title and review comment message. These plots indicate that reviews are slightly more positive, with the distribution of sentiment polarity being roughly balanced along the x-axis. More over, The reviews doesn't reveal a clear connection between polarity and subjectivity. This might lead to an assumption that the reviews are justified to some extent.

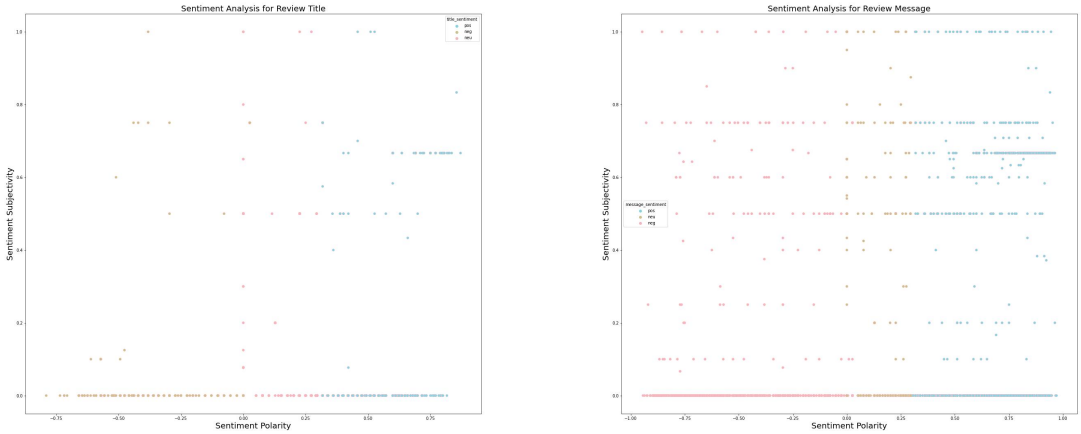


Figure 26: Polarity vs. Subjectivity for title(left) and Polarity vs. Subjectivity for message(right).

In the last part of this subsection, we calculated the correlation coefficient of review comment message, review comment title and review score, to see if our sentiment analysis is meaningful. The correlation coefficient is 0.6, which means that the sentiment analysis is helpful.

6.1.3 Wordcloud for Sentiment Polarity

In this section, we want to explore the word frequency of the review comment dataset. In order to avoid the phenomenon that word and its corresponding phrases which have the same meaning are used multiple times, and avoid using stopwords which are meaningless in data analysis, we used *nltk*, which is a package for natural language processing. Here we apply the method *STOPWORDS* and *PorterStemmer*. Note that before the process of counting word frequency, we did a preprocessing to get rid of emojis and special tokens like commas. Words with top 30 frequency and their wordcloud picture are shown below:

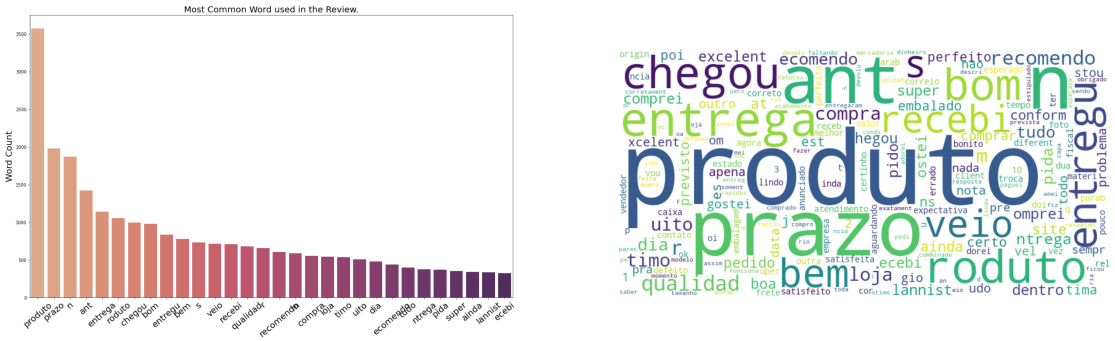


Figure 27: Barplot of counts for the top 30 words(left) and the Wordcloud(right).

It's clear that customers care much about the **transportation** and quality of the product: the top 10 words are mainly concerned with the transportation time of the product. This means that the transportation time is an aspect which must be paid more attention to. And within the top 30 words we can also see the word **recomendo**, which is close to the meaning of recommend. This might indicate that the customers who left comments are generally satisfied with the product they bought.

Next, we want to divide the review dataset into customers who are not satisfied and customers who are satisfied. The rule is that if the review score is between 1-3, then the customer is not satisfied. Otherwise, the customer is satisfied. We plot the wordcloud for customers who are not satisfied as the left picture, and customers who are satisfied as the right picture.

Figure 28: Wordcloud for review of customers who are not satisfied(left) and who are satisfied(right).

This time, it's clearer now that the reviews concerned with transportation time are from customers who are not satisfied. This may shed light on improvements for sellers that to get more positive feedbacks, they can find ways to shorten the transportation time of their products.

6.2 Exploring Which Products Receive More Positive Reviews

Now, with the sentiment analysis results at hand, we add more datasets to see what kind of conclusion can the sentiment result explain. Firstly, we explored the relationship of product categories and customer review sentiment.

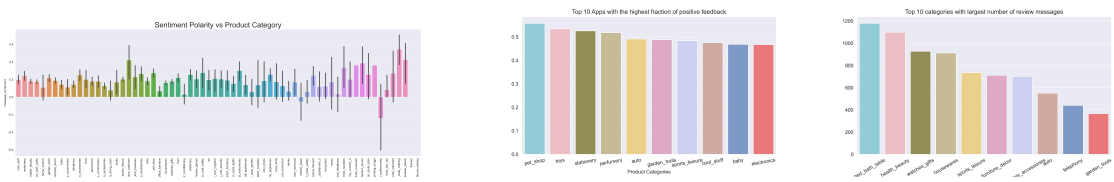


Figure 29: Barplot of categories and their average sentiment(left), categories and their fraction of positive reviews(middle) and Number of reviews according to the product category(right).

From these figure, we can see that the products which averagely receive the highest positive review sentiment

belong to the *fashion_male_clothing*. *fasion_underware_beach* and *furniture_bedroom* rank second and third. This might mean that these products have better quality and their transportation time is shorter. Meanwhile, in the right picture, *bed_bath_table* and *healthy_beauty* receive the top 2 largest number of reviews according to the product category. However, this phenomenon might also be influenced by advertisement, since there might be some spammers inside the reviews. It's also noticeable that *art_and_craftmanship* receive an averagely negative review.

We also explored the relationship between the fraction of positive reviews and the product category. Results are provided as above. Note that products with more reviews in this picture doesn't necessarily indicate that they sell well, since we dropped lots of reviews which only have review comment title.

6.3 Geographical Analysis with Sentiment

In this section, we combine the sentiment of reviews and the geolocation dataset, aiming to provide a geographical visualization of sentiment according to the customer's state/city and the seller's state/city. We used the python package *Plotly* to draw two interactive plot which are saved as html files.

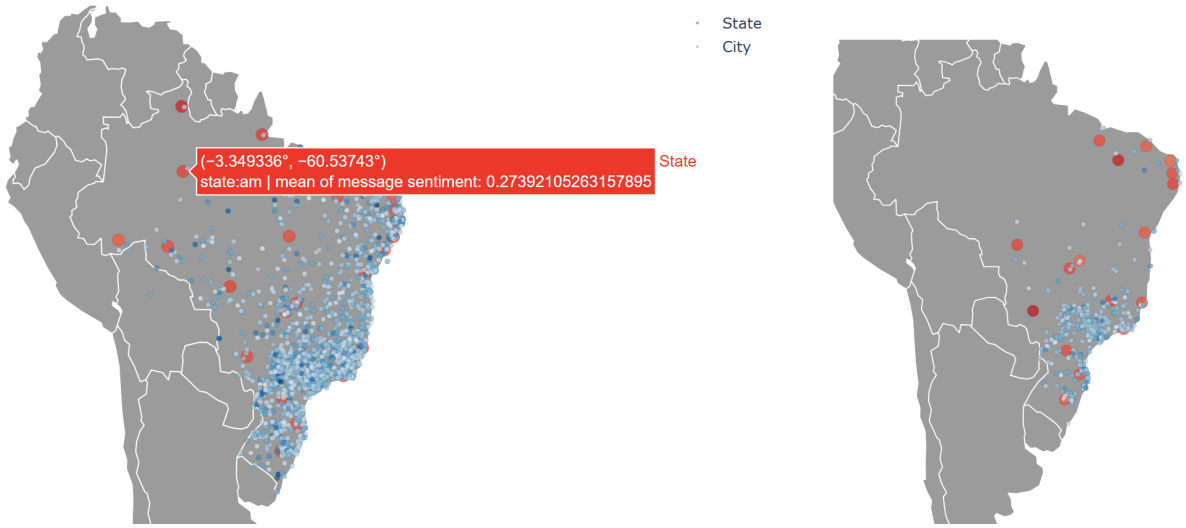


Figure 30: Geographical visualization of sentiment according to the customer's state/city(left) and the seller's state/city(right).

Here in these two plots, the size of the dots depicts the count of reviews, the color depicts the sentiment polarity of the average sentiment polarity of reviews. We provide both city view and state view of the average sentiment. Blue dots indicate city view and red dots indicate state view. You can click on the legend to select whether to show the state view or city view or both of them in the html files.

These two plots show that the dataset mainly consists of transactions within the southeast side of Brazil. Sellers mainly distributed in that area as well. Compared to sellers, customers are distributed into more cities. The abbreviation of top 5 state which receive the highest average sentiment polarity are ms(0.599400), pi(0.522400), pe(0.243585), mg(0.237556) and mt(0.237378). The top 5 state which give the highest average sentiment polarity are rr(0.482938), ap(0.305560), am(0.273921), mt(0.257599) and al(0.232268).

Meanwhile, the top 5 cities which receive the highest average sentiment polarity are nova lima(0.93710), irati(0.89340), balneario camboriu(0.88005), colorado(0.86580) and artur nogueira(0.86410). The top 5 cities which give the highest average sentiment polarity are aparecida(0.9286), sitio novo(0.9214), andre da rocha(0.9081), agua doce do norte(0.9062) and tuparetama(0.9051).

6.4 LightGBM for Customer Satisfaction

We also trained a LightGBM model with the aim of predicting the Customer satisfaction. Customers who give a review score larger than 3 is considered as satisfied. The model is fitted using the variables *freight_value*, *price*, *product_weight_g*, *order_status_delivered*, *product_category_name_english*, *customer_state* and *seller_state*. We encode the categorical variables with one hot encoding, standardized the numerical variables, and deleted the

outliers according to quantile(the upper limit is the third quantile plus 1.5 times the interquantile range). The confusion matrix for the final model is as follows:

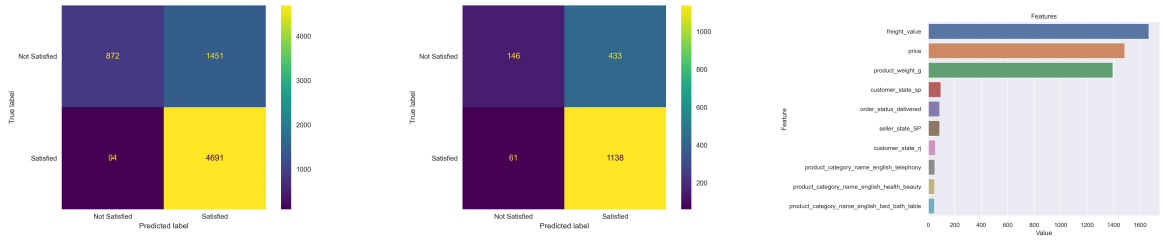


Figure 31: Confusion matrix for the training data(left) and test data(middle), and importance of variables with random forest(right).

It can be seen that the model is not good at classifying customers who are not satisfied. We also used a random forest to see the variable importance. The fact that the numerical variables contribute more may boil down to too many dummy variables for the customer/seller's state.

7 Conclusion

The concentration of Olist e-commerce users is primarily in the eastern coastal regions, with the southeastern coast being the most densely populated area. Future research should delve into the consumption habits of customers in the southeastern coastal region to refine operations for specific user groups. Additionally, an analysis of the reasons for the lack of activity among users in the western and central regions is necessary, along with suggestions for improvement.

The cities and states with the strongest purchasing power are Sao Paulo and SP, respectively. In terms of average sentiment polarity, the top five states are MS (0.599400), PI (0.522400), PE (0.243585), MG (0.237556), and MT (0.237378). Furthermore, the top five cities with the highest average sentiment polarity are Nova Lima (0.93710), Irati (0.89340), Balneario Camboriu (0.88005), Colorado (0.86580), and Artur Nogueira (0.86410).

The majority of users prefer to pay with credit cards, which not only confirms the influence of consumer psychology but also reflects the habits of Olist users and the fact that most customers do not have high deposit balances.

The number of online shopping orders during the week is higher than on weekends. Additionally, users are mostly active between 10:00 a.m. and 10:00 p.m. The following recommendations are proposed: a series of promotional activities can be implemented during the weekend to encourage user activity and purchases.

Popular product categories among the general public include health and beauty, home goods, and sports and leisure. In the popular product categories for main users, health and beauty, home goods, and sports and leisure are all ranked at the top, and the average sentiment polarity of user reviews is relatively high. This indicates that these products are popular among the general public and are suitable for promotion and marketing to various user groups. Watches, furniture, and computer accessories, on the other hand, are niche popular products that are more popular among specific types of users. Therefore, it is recommended to select appropriate user groups for promotion.