A Visual Analytics Approach To Explore E-commerce Data And It's Implication to Olist's Key Stakeholders

Hoi Ting Ethel Cheung
Department of Computer Science
City University of London
London, U.K.
hoi.cheung.2@city.ac.uk

Abstract—An effective analysis of e-commerce sales and customer review data could provide Olist key stakeholders with actionable insights to make decision in their buying and marketing strategies and their products and services upgrade. The report focuses in analysing sales performance and trend, geographical performance, customer review scores and text and visualised the results with the most appropriate plots. Python was used to perform data preparation, text translation and sentiment analysis, and Tableau was used to perform most of the data visualisations. After analysing over 90,000 data, we are able to identify trends in the data and provide actionable insights to Olist store. A sentiment analysis was performed in the end of the paper and built model that is able to classify the reviews to positive or negative.

1 Problem Statement

E-commerce has become increasingly popular among shoppers in recent years. Take the United States as an example, it was predicted that online buyers in the country will be increased to 278.33 million by 2024, which is around 91% of the entire country's population [1]. Therefore, a lot of studies regarding e-commerce have been done to understand the purchase behaviour and preferences of consumers, so retailers could leverage the insights to improve their ways of selling and marketing of their products and services.

With a focus of analyzing e-commerce data in this study, our paper will be using real anonymized commercial data of a Brazilian e-commerce store called Olist to address the following questions:

- *What are the more revenue-driven categories and regions?*
- *What categories do customers from each region prefer to shop?*
- *What feature of the products or services are most important for customers and how is the performance of the feature now?*
- *Are we able to predict the sentiment of the customer based on their reviews?*

As a retailer, it will be crucial to run similar analysis once a quarter to keep track of the trend of customer preferences on their e-commerce store and make necessary adjustments based on profit/loss and customer reviews. We aim to provide actionable insights to Olist key stakeholders based on the analysis and plots.

The Python codes and the graphs set up in Tableau for this report could be replicated to other e-commerce data to create similar report.

2 State of the Art

There are several research papers published in recent years around this topic that dealt with similar data and problems and applied visualisations to solve the problems.

The first paper I studied was a conference paper that discussed big data analytics and its application in E-commerce [2]. It has inspired me that e-commerce data is not only useful for analyzing what categories and products are trending among consumers, but also useful for improving the products and services by using Text Mining technique and Sentiment Analysis. Visually, it has used simple stacked bar charts with trend line to illustrate the trend by categories over the years.

The second paper I studied was about UK e-commerce grocery market sales [3]. It investigated the geography of e-commerce activity and visualized the online grocery penetration by using a UK map with a scale of color to represent the penetration with darker color means higher penetration and vice versa. It also used circles on the map to indicate the number of grocery stores locally, with larger circle means more local stores and vice versa.

The third paper I studied was an analysis of resilience measure for supply networks [4]. In the paper, the author has investigated the relationship between resilience and importance of suppliers and has visualized the result with a quadrant plot. For example, suppliers in upper right quadrant are most resilient and more important while the suppliers in bottom left quadrant are of least importance and less resilient compared to other suppliers.

Considering all the above, these papers have provided us a general guidance on performing analysis with e-commerce data which includes using colored plots to showcase category/product performance and using geographical maps with color scale to showcase sales or traffic effect and use marks e.g. triangles or circles on the maps to present additional data. In addition, it would be interesting to use quadrant plots to present the relationship between 2 variables as we can

quickly spot the set of items that have common traits by separating them in 4 different quadrants.

3 Properties of the Data

The datasets were collected between 2016-2018 and there was a total of 9 files available. It contains data from 99,441 of unique orders. We have used 7 files that are essential for us to solve the research questions, which include: customer orders, product category names, order items, customer location, order reviews, order payments and geolocation that relates Brazilian zip codes to latitude and longitude coordinates.

We used Python to perform data preparation for our datasets. To prepare for the first dataset, we have merged all the files to customer orders data except order reviews with the function DataFrame.merge(right, how = ‘’, on=’’). For example. We have joined the customer orders file with order items file by using the function olist_orders.merge(olist_order_items, on='order_id', how='left'). We then removed columns that are not relevant to our analysis such as ‘order_item_id’, ‘product_length_cm’ and ‘product_id’ and visualized the missing values with a heatmap.

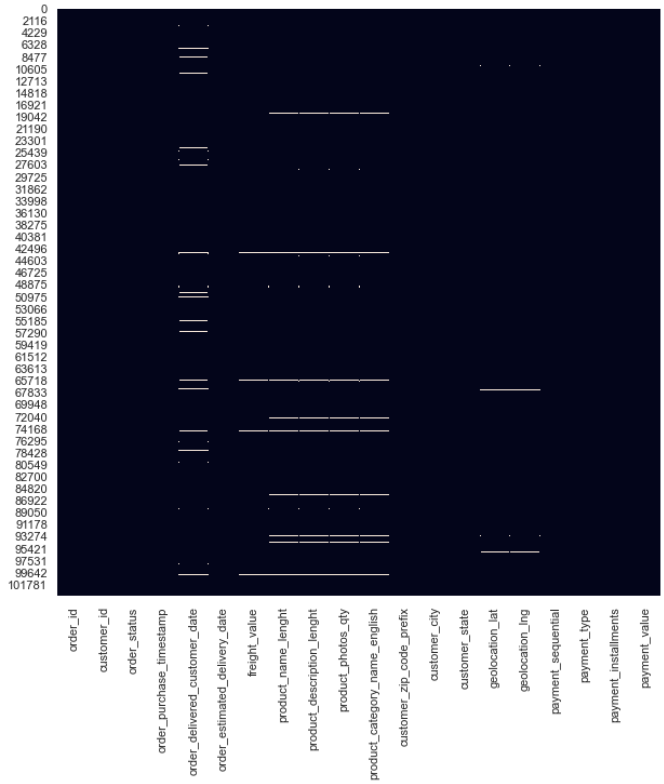


Figure 1: Heatmap of missing values in the first dataset

From the heatmap (Fig.1), we noticed that the missing values are distributed evenly, which means that the values are missing completely at random. Hence, we decided to drop the rows with missing values. As we have a large dataset, removing the missing values would not affect our analysis. Also, this method of removing missing values is more preferred since complete removal of data with missing values could result in robust and highly accurate model.

As a last step of data preparation of this dataset, we have inspected outliers in 7 of the numerical columns by using boxplots. We have identified that all the columns contain outliers. However, we have decided not to remove them after thorough inspection as removing them might lose interesting findings and useful information. For example, the outliers in the column 'payment_value' may imply that the e-commerce store has certain VIPs that have high spending power, the outliers in the column 'payment_sequential' may help us detect fraud payment etc.

As a result, the first dataset has 94,838 rows and 19 columns.

To prepare for the second dataset which contains customer review data, we have first inspected the missing values and visualized them with heatmap. We noticed that only the columns ‘review_comment_title’ and ‘review_comment_message’ have missing values and they are missing completely at random because the values are distributed evenly across all orders on the plot. Hence, we removed all the rows with no comment because one of the goals of this research is to analyse customer reviews. We have also removed unnecessary columns ‘review_creation_date’ and ‘review_answer_timestamp’ and then imputed the missing values of 'review_comment_title' as empty string.

As a result, the second dataset has 41,753 rows and 5 columns.

Note that we did not merge the second dataset of customer reviews with the first dataset because the former contains a large amount of missing values and removing them from a merged dataset would result in a much smaller dataset and loss in information.

Fig. 2 is a summary of the fields after data preparation, their meaning, and types of the values.

Variable	Meaning	Type of Values
order_id	unique identifier of the order	object
customer_id	key to the customer dataset. Each o rder has a unique customer id.	object
order_status	Order status (delivered, shipped)	object
order_purchase_ti mestamp	Purchase timestamp	object
order_delivered_c ustomer date	Order posting timestamp. When it was handled to the logistic partner	object
order_estimated_d elivery date	Actual order delivery date to the cu stomer	object
freight_value	Item freight value item	float64
product_name_len ght	Number of characters extracted fro m the product name	float64
product_descriptio n lenght	Number of characters extracted fro m the product description	float64
product_photos_qt y	Number of product published photo s	float64
product_category_ name english	Category name in English	object
customer_city	Customer city name	object
customer state	Customer state	object
geolocation_lat	Latitude	float64
geolocation_lng	Longitude	float64
payment_sequenti al	A customer may pay an order with more than one payment method. If he does so, a sequence will be creat ed to accommodate all payments.	float64
payment_type	Method of payment	object
payment_installme nts	Number of installments	float64
payment value	Transaction value	float64
review_id	Unique review identifier	object
review_score	Ranging from 1 to 5 given by the c ustomer on a satisfaction survey	int64
review_comment_ title	Comment title from the review left by the customer, in Portuguese	object
review_comment_ message	Comment message from the review left by the customer, in Portuguese	object

Figure 2: variable summary

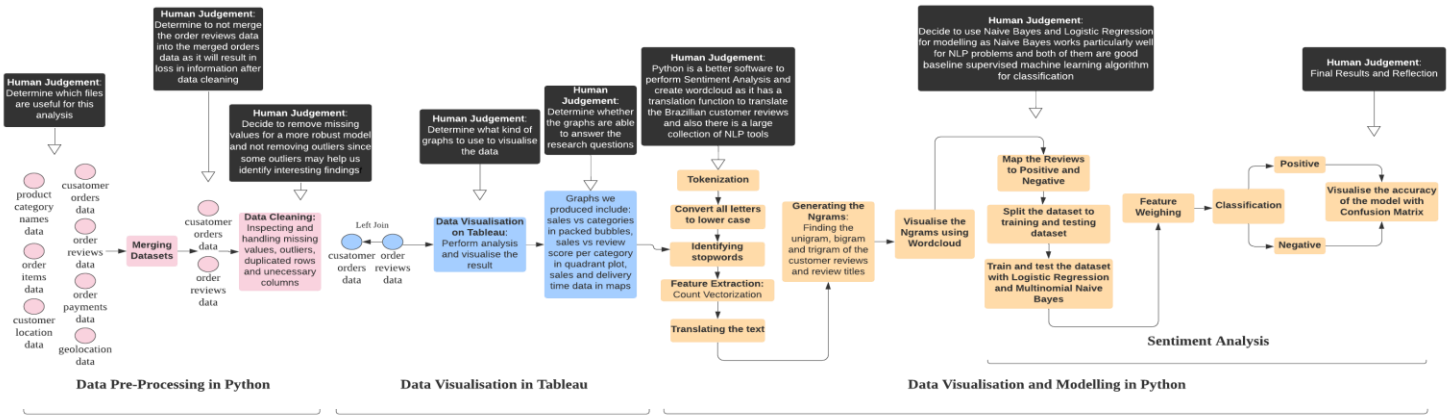


Figure 3: Schematic diagram of data visualization and analysis workflow

4 Analysis

4.1 Approach

In this research, we will be using both Tableau and Python for data visualization. Each of them has its own strengths: Tableau has simple user interface and is able to create beautiful visualizations easily, whereas Python has extensive libraries suitable for different analytical purposes.

As mentioned in Section 3 and as shown in Fig. 3, we have done data preparation in Python. Human judgement was involved to decide to remove missing values for a more robust model and not removing outliers since some outliers may help us identify interesting findings.

After we loaded the cleaned customer orders data into Tableau, we used a left join to combine order reviews data with it. Orders with no reviews will have 'null' appeared in the cell. This will preserve all the customer review data for later analysis.

To find out the most appropriate visual displays to answer each question, we will first brainstorm some possible options to visualize the data and then create and compare multiple plots when we are analyzing each question and use human judgement to decide which plots are the most suitable ones.

We will use packed bubble plots and heatmap to visualize the revenue and trend of the product categories. To understand how customers feel about the different categories, we will also use a quadrant plot to visualize the relationship between customer review score and total payment value. We will then use human judgement to summarize the key similarity between the categories of each quadrant.

To create a more effective graph to represent the geographic performance of product categories, we have decided to use a hex-tile map instead of the usual map to eliminate the effect of the different sizes of states on a map.

After performing data visualization in Tableau, we will use Python to visualize customer reviews data. We have made a human judgement to use Python because it has a translation function available for us to translate the reviews from Portugese to English. Also, it has a large collection of NLP tools for us to perform Sentiment Analysis.

We will perform the standard procedure of NLP which includes word tokenization, identifying stopwords, extracting features, translating texts, and identifying Ngrams. We will then visualise the Ngrams in wordclouds.

As a last analytical step of the research paper, we will perform Sentiment Analysis with the goal to build a model that is able to classify customer reviews as negative or positive. We have made human judgement to use Naïve Bayes and Logistic Regression for modelling because both of them are good baseline supervised learning algorithm for classification [5] [6].

After training and testing the dataset with the two regression models, we will classify the most important words with positive and negative sentiments and translate the text to English. As a last step, we will use confusion matrix to visualise the accuracy of the models. Human judgement will be involved in this step to reflect on all the analytical steps and summarize the findings.

4.2 Process

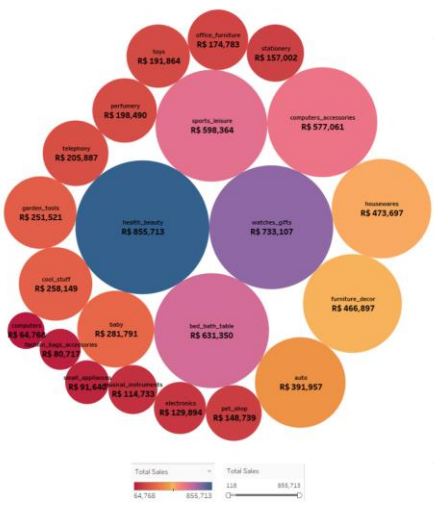
Most Revenue-Driven Categories and Trend

We first investigated the top product categories in terms of sales by using a packed bubbles plot (Fig. 4). The color ranges from blue (most sales) to red (least sales), the size of the bubbles also indicates sales value, where largest bubble means highest sales and vice versa. We can see that daily essentials and gifts eg. health_beauty, watches_gifts, bed_bath_table and sports_leaisure are the top categories while items that people usually buy only when they are worn out or needed an upgrade eg. musical_instruments, small_appliances, fashion_bags_accessories and computers generated the least sales values.

However, it will be more useful to find out whether there is any trend in terms of categories sales as we might be able to identify some up-and-coming and going out of fashion categories so Olist's buyers could adjust their buying strategies based on the findings.

We have created a heatmap with individual color legends for the two columns: Total Sales and % of Difference in Total Sales in 2017 in the heatmap and assigned a different color palette to each column for more accurate comparison (Fig. 4).

We noticed that small_appliances_home_oven_and_coffee has 5,924% growth compared to 2017 and some of the other high growth categories are diapers_and_hygiene, arts_and_craftmanship and construction_tools_lights. This might be due to more people staying at home nowadays and they want to decorate their home.



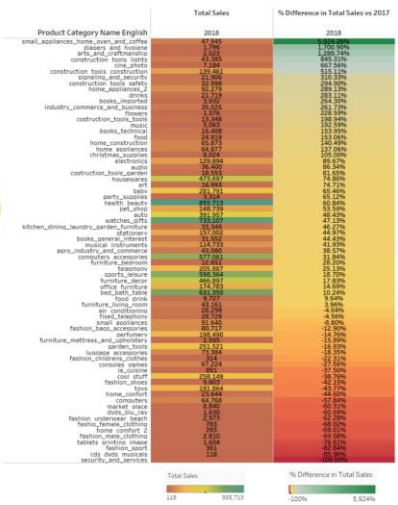


Figure 4: Packed Bubbles Plot (Left): Total Sales by product categories, Heatmap (Right): Total sales (left column) and % Difference in Total Sales in 2017 (right column) by product categories

Figure 5: Barplots of number of order (leftmost column), Total Sales (middle column) and Sales per order (rightmost column) by product categories

We also noticed from the heatmap that most of the categories with negative growth vs 2017 are related to fashion eg. fashion_sport, fashion_male_clothing, fashion_female_clothing. People are either not interested in shopping for clothes or they prefer to shop from other destinations. Hence, the buyers should consider stock less of these categories and focus on the homegoods category.

We have also engineered a new feature 'Sales per Order' by dividing the Total Sales by Number of Orders. From the bar chart (Fig. 5), we can see that some categories like computers, home_appliances_2, agro_industry_and_commerce and small_appliances_home_oven_and_coffee have low number of sales and total sales but the Sales Per Order generated are much higher compared to other categories. This is due to the products from these categories have higher price point. Perhaps markerters in Olist could spend more effort in promoting categories with higher sales per order as these categories can easily generate more revenue if they marketed it properly.

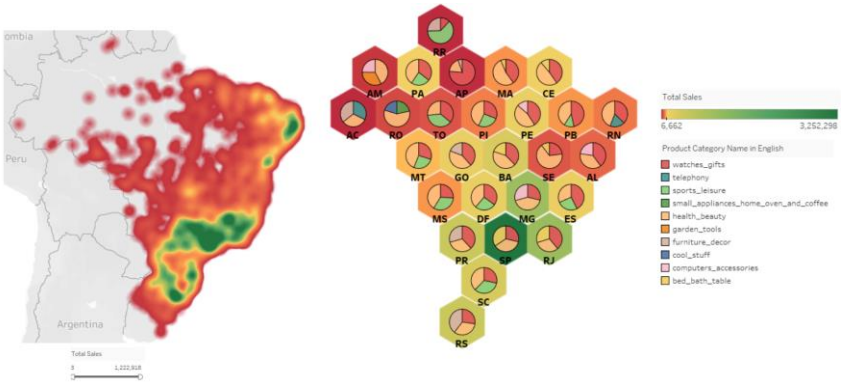


Figure 6: Density map (left): Total Sales per city (Green: highest sales, Red: lowest sales), Hex-tile map (right): Total Sales per city (Green: highest sales, Red: lowest sales) and pie charts of product categories performance by total sales

Geographical Performance

To investigate the geographical performance in terms of total revenue and that of product categories, we have used a density map and a hex-tile map with pie charts to visualise the data Fig 6).

From the density map, we noticed that coastal areas generated more revenue than inland, which is aligned with the findings of a research about Spatial income inequality of Brazil that richest areas were located around Rio de Janeiro [7].

As mentioned in Section 4.1, we decided to use hex-tile map for geographical performance of product categories because we would not be able to display pie charts on smaller states.

To create a hex-tile map, we first downloaded a file from a source which contains Brazilian states and the respective numbers in the form of their location in columns and rows [8]. We then added hexagon shape to our Tableau shapes repository. Next, we added the Rows and Columns data of the file to the Tableau's rows and columns. Then we changed the mark type to shapes and added state code to shapes and changed the shape to hexagon which we have added to our repository.

From the hex-tile map, we can see that SP (Sau Paulo), RJ (Rio de Janeiro) and MG (Minas Gerais) are the most revenue-driven states which is confirmed by our findings in the map. We also noticed that top categories are different across states. For instance, health_beauty, watches_gifts and bed_bath_table are top categories in Sao Paulo whereas sports_leisure is one of the top categories in its neighbouring state SC (Santa Catarina). It is also interesting to note that health_beauty is a top category

across all the states, telephony is the top category only in RN (Rio Grande do Norte) and garden_tools is the top category in AM (Amazonas). This indicates that the marketing effort of Olist should be more regional and category specific.

As we have investigated the top product categories, the trend and their geographical performance in terms of sales, we have proceeded to investigate customer reviews data. It will be interesting and useful to analyse the relationship between customer reviews and sales and whether there are any similarities between certain product categories.

Customer Reviews Analysis

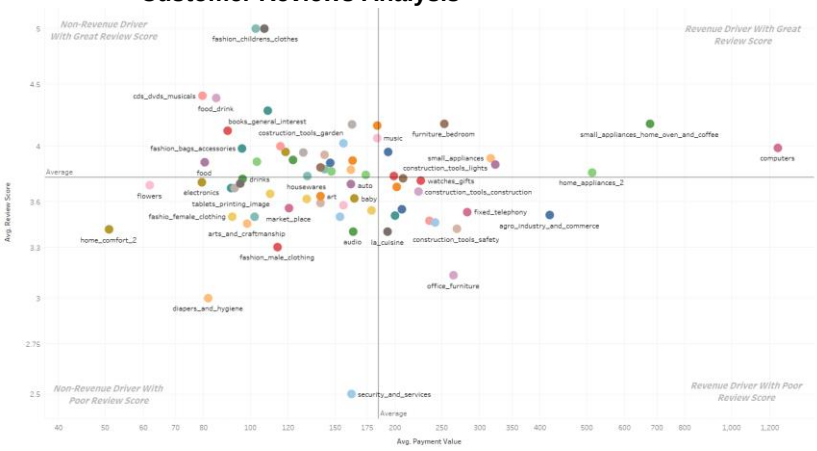


Figure 7: Quadrant plot of customer review score vs average payment value by product categories

Quadrant charts are very useful for plotting data with three measures. In this paper, we visualized the relationship between review score (Y-axis), payment value (X-axis) and product categories (color) with a quadrant plot (Fig. 7). We divided the plot into four quadrants, the top right is 'Revenue Driver With Great Review Score', top left is 'Non-Revenue Driver With Great Revenue Driver', bottom right is 'Revenue Driver With Poor Review Score' and bottom left is 'Non-Revenue Driver With Poor Review Score'.

We noticed that small_appliances_home_oven_and_coffee and computers are categories that are located in the most top-right area, which means they are the most revenue-driven and received the highest review score. Furniture_bedroom, small_appliances and home_appliances_2 are also performing well. This aligns with our findings in earlier section which suggested that homegoods is a category that Olist should focus on.

Categories that Olist should be aware of are categories in the bottom left quadrant as they are neither generating revenue nor getting high review scores. The categories in this quadrant should be either improved or removed completely from sale.

Similarly, categories such as office_furniture, fixed_telephony and watches_gifts that are in the bottom right quadrant should be relooked at and received product upgrades as they are generating revenue but received poor review score.

Figure 8: Wordclouds of customer reviews (top row) and customer reviews titles (bottom row) in the sequence of unigram (leftmost), bigram (middle) and trigram (rightmost)

To analyse the customer reviews in Python, we first defined stopwords, performed tokenization to divide the sentences into smaller parts and converted all letters to lower case for consistency. Following that, we translated the text from Portuguese to English by using the function google_translator() after imported the google_translator library from google_trans_new python API. As translating all the customer reviews data is very time-consuming, we decided to use CountVectorizer to get the most important 50 words in the form of Unigrams, Bigrams and Trigrams from the reviews and review titles. We have also combined the phrases using underscore to visualize it as wordclouds.

From the customer wordclouds (Fig. 8), we can see that the top keywords are related to delivery time, product quality and positive words such as good, great and super. This implies that people are satisfied with the products and services of Olist in general and would recommend to others. It also indicates that delivery time is one of the things customers most care about after purchased something from an e-commerce. However, there are also some negative comments in the wordcloud such as product_not_delivered, wrong_product, different_product_photo.

As delivery time is a key metric in e-commerce, we proceeded to investigate the delivery performance of Olist.

Delivery Time Analysis

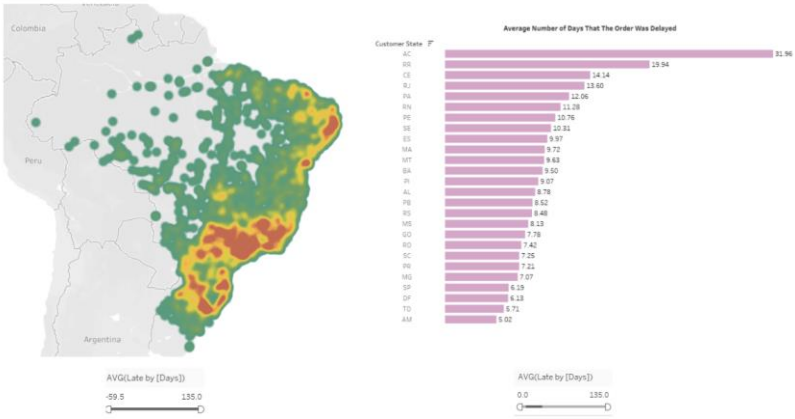


Figure 9: Density map (left): Average number of days that the order delivery was delayed (Red: highest delay, Green: least delay), Bar plots (right): Average number of days that the order delivery was delayed by customer states

We engineered a new variable "Late by [Days]" by subtracting 'order_delivered_customer_date' from

'order_estimated_delivered_date' and used map and bar plots to visualise the data (Fig. 9). We found that cities in coastal area have the most serious delay. Note that it is also the area that we identified as revenue-driven in the earlier Geographical Performance section. Meanwhile, there are also some inland cities received their packages earlier than estimated day as they have negative average (late by [Days]) values.

From the bar plots, we noticed that AC (Acre) has the most serious delay, followed by RR (Roraima) and CE (Ceará). Olist's logistic team should focus on improving these areas in order to reduce the average number of days delayed, which will be reflected in future customer reviews wordclouds.

4.3 Results

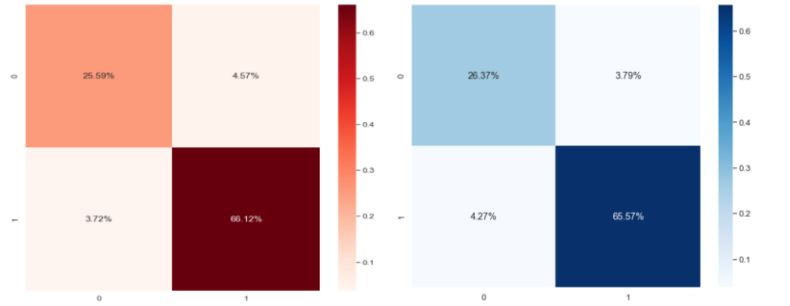


Figure 10: Confusion Matrix: Logistic Regression model (left), Multinomial Naïve Bayes (right)

As a final step of the analysis, we performed Sentiment Analysis with the goal to build a model that can classify customer reviews as negative or positive.

We split the dataset into training and testing dataset, defined, trained and tested the Logistic Regression and Multinomial Naïve Bayes models. We then visualized the accuracy of the models by plotting confusion matrix (Fig. 10). From the confusion matrix, we can conclude that the models are quite accurate in classifying the reviews because we got 25%+ as True Negative and 65%+ as True Positive.

The result also implies that customers are satisfied with Olist's products and services as more than 65% of the reviews are with positive sentiment, which aligns with our findings from the wordclouds.

To further improve the customer reviews or to get more positive reviews, Olist should relook at their logistics service to minimize delivery delays. Also, they should be aware of some categories that are trending down and reallocating resources to potential product categories and market them to the right regions.

5 Critical Reflection

We have translated the Portuguese customer reviews to English in this paper using google_trans_new python API. This might result in inaccurate analysis because auto translation tool might not be able to translate the full text accurately and to preserve the sentiment of the customer. In fact, research has shown that output of online translation tools of Arabic UGC can either fail to transfer the sentiment by producing a neutral target text, or completely flips the sentiment polarity of the target word [9]. Unless there is further research done on the accuracy of the output of translation tool of Portuguese UGC, we are not able to estimate the accuracy of our models and wordclouds. Alternatively, we could have created wordclouds and performed Sentiment Analysis in Portuguese first then translate the output.

Research has shown that when performing sentiment analysis, using Sentiment-aware tokenizing and negation marking together would yield the highest mean accuracy if the data contains at least 6000 training texts [10]. Further analysis could be done to compare the accuracy between the basic text tokenizer we used in this paper and sentiment-aware tokenizing with negation marking and visualise the result with a model accuracy plot.

Also, we have only used part of the reviews and review titles in this analysis because it takes long time to translate all the text in the customer reviews. If we are able to figure out a faster way to analyse all the text data, we would get a more accurate result and might be able to identify a more diverse set of ngram keywords in the wordclouds.

Overall, the graphs we produced in Tableau are able to answer the research questions, however, apart from using both Tableau and Python for this paper, we could use TabPy in the future. TabPy is a Tableau Python Server which allows users to execute Python scripts in Tableau environment. We could then save time on navigating between both software and the graphs would be more consistent in terms of layout and design if we conducted in the entire research on Tableau.

Table of word counts

Problem statement	247/250
State of the art	344/500
Properties of the data	495/500
Analysis: Approach	497/500
Analysis: Process	1464/1500
Analysis: Results	186/200
Critical reflection	346/500

Python Code: https://github.com/ethel121/Coursework-Task/blob/main/Final/INM433%20Visual%20Analytics%20Coursework%20-%20Code%20.ipynb

REFERENCES

The list below provides examples of formatting references.

[1] Tugba Sabanoglu (2020, November) United States: number of digital buyes 2017-2024 [Online] Available: https://www.statista.com/statistics/273957/number-of-digital-buyers-in-the-united-states/

[2] Uyoyo Zino Edosio (2014, April) Big Data Analytics and its Application in E-Commerce [Online] Available: https://www.researchgate.net/publication/264129339_Big_Data_Analytics_and_its_Application_in_E-Commerce

[3] Elena Kirby-Hawkins, Mark Birkin, Graham Clarke (2018, February) An investigation into the geography of corporate e-commerce sales I the UK grocery market [Online] Available: https://journals.sagepub.com/doi/full/10.1177/2399808318755147

[4] Dmitry Ivanov (2019, July) A new resilience measure for supply networks with the ripple effect considerations: a Bayesian network approach [Online] Available: https://www.researchgate.net/publication/334810177_A_new_resilience_measure_for_supply_networks_with_the_ripple_effect_considerations_a_Bayesian_network_approach

[5] Standard University, Text Classification and Naïve Bayes, The Task of Text Classification [Online] Available: https://web.stanford.edu/class/cs124/lec/naivebayes.pdf

[6] Daniel Jurafsky & James H. Martin (2020, December) Logistic Regression [Online] Available: https://web.stanford.edu/~jurafsky/slp3/5.pdf

[7] Eustaquio Reis (2014, May-August) Spatial income inequality in Brazil 1872-2000, Volume 15 Issue 2 [Online] Available: https://www.sciencedirect.com/science/article/pii/S1517758014000198

[8] Tableau Workbook - DistribuiodeMdicosnoBrasil [Online] Available: https://public.tableau.com/views/DistribuiodeMdicosnoBrasil/DistribuilcaoMedicos?:embed=y&:toolbar=yes&:embed_code_version=3&:loadOrderID=0&:display_count=yes&:tabs=no&:showVizHome=no 000198

[9] Hadeel Saadany, Constantin Orasan (2020, October) Is it Great or Terrible? Preserving Sentiment in Neural Machine Translation of Arabic Reviews [Online] Available: https://www.aclweb.org/anthology/2020.wanlp-1.3.pdf

[10] Christopher Potts (2014, May) Sentiment Analysis CS 244U: Natural language understanding [Online] Available: https://web.stanford.edu/class/cs224u/2014/slides/cs224u-2014-lec15-sentiment.pdf

Xing Fang, Justin Zhan (2015, June) Sentiment analysis using product review data [Online] Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2

Kishan Maladkar (2018, February) 5 Ways to Handle Missing Values in Machine Learning Datasets [Online] Available: https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/