Big Data Analysis and Project

Assignment 1: Part A (Question Formation and Exploratory Analysis)

Possakorn A1873765

June 12, 2023

Problem description

The E-commerce industry has experienced significant growth and transformation in recent years, fueled by technological advancements and changing consumer preferences. With the abundance of digital data generated by online platforms, companies operating in the E-commerce sector have a unique opportunity to leverage this wealth of information to gain insights, make data-driven decisions, and enhance their competitive edge. In this industry, data science applies techniques to analyse and interpret various types of data, such as transaction records, customer demographic, product descriptions, and customer reviews, including various kinds of information such as complaints, direct comments, or emoticons expression to automate and optimise the processes to perform the maximum profit (Bayhaqy *et al.*, 2018). Examining E-commerce data enables online retailers to gain insights into customer expectations, identify potential challenges, and improve overall customer experiences(Vanaja and Belwal, 2018).

One of the crucial factors for achieving success involves identifying potential customer segments through a deep understanding of their unique needs and preferences. This knowledge allows for personalized marketing efforts, strategic optimization, and improved customer relationships. Additionally, analyzing sentiments expressed in written content, such as customer reviews and social media interactions, provides valuable insights into customer satisfaction and product preferences. By leveraging this information, E-commerce companies can identify relevant issues, respond effectively to customer feedback, and develop suitable marketing strategies. Ultimately, this leads to the first initial question: **"How does sentiment impact business profitability in the E-commerce industry on each segment"**

Dataset description

Regarding the potential data sources available on the **Kaggle dataset**, some datasets pertain to **an e-commerce platform known as Olist Store in Brazil**(Olist and Sionek, 2018). These datasets encompass approximately 100,000 records from 2016 to 2018 and originate from various online marketplaces. They embody the key characteristics of Big Data, namely volume, velocity, variety, and veracity.(Erl, Khattak and Buhler, 2016, p. 19 - 35)The volume characteristic is evident due to the substantial data from three primary sources over the specified period. The rapid data collection process demonstrates the velocity characteristic, ensuring real-time capture of transaction records.
Additionally, the data sources exhibit variety, encompassing multiple formats, including quantitative values and text formats. However, it should be aware that in this project involving customer reviews data, there may be missing values resulting from incomplete comments, affecting the veracity of the dataset. Lastly, the dataset offers intriguing information regarding transaction records, encompassing details such as order items, status, pricing, customer attributes, product details, and customer feedback, as depicted in the diagram below:

Therefore, there are potential data sources, including with

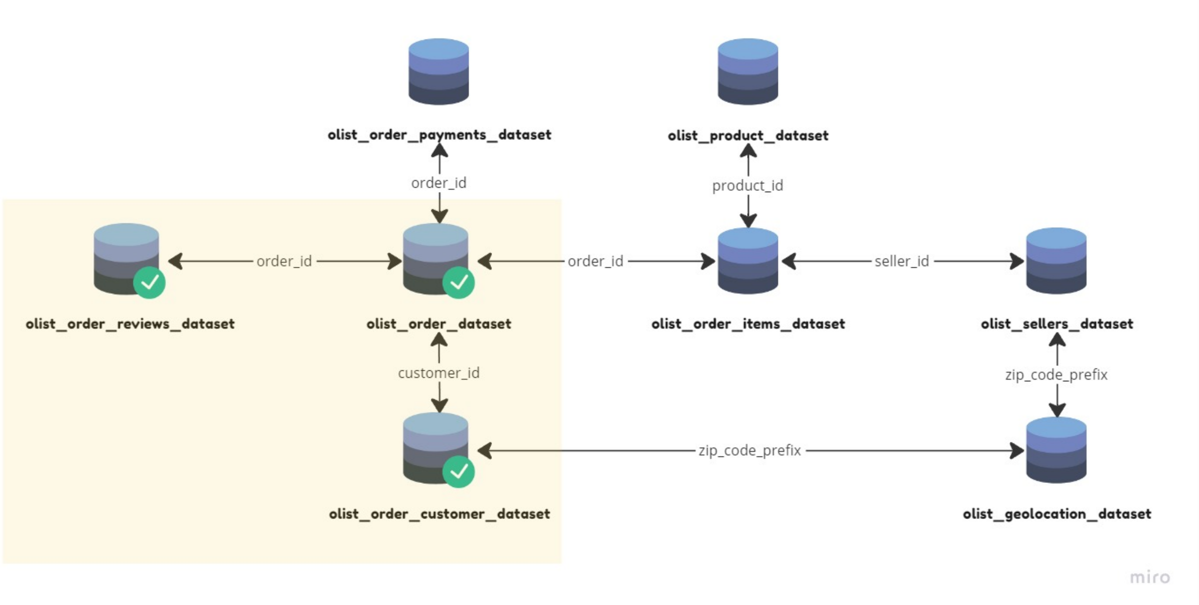

Figure 1: Potential data schema - diagram

01 olist_orders_dataset (core table);

- Background: this is the core dataset to provide the information related to transaction records from 2016 to 2018.
- Detail: 99,441 transaction records with order status and all related transaction dates.
- Formats: CSV formats with size 17.65 MB

02 olist_order_customer_dataset;

- Background: This dataset has information about the customer and its location
- Detail: almost 96k unique customers with zip code, city, and state.
- Formats: CSV formats with size 9.03 MB
- join Key: customer_id

03 olist_order_reviews_dataset;

- Background: This dataset includes data about the reviews made by the customers.
- Detail: there are 98,410 customer reviews with review_Score, comment title, comment detail, and related_review data
- Formats: CSV formats with size 14.45 MB
- join Key: order_id

Initial Data processing

01 Exploratory Data Analysis - Overview

Before delving deeper into the analysis, obtaining a foundation understanding of each dataset is essential. The datasets relevant to this inquiry consist of customer, order, and review data. Hence, assessing the quality, gathering basic information, and developing a comprehensive understanding of each data source is imperative.

Dataset01 - olist_orders_dataset
The Olist order dataset comprises 99,441 rows and exhibits near-complete data coverage, with a completeness

rate of nearly 100%. However, there are minor instances of missing data in fields such as order_approved_at, order_delivered_carrier_date, and order_delivered_customer_date. Furthermore, the dataset encompasses an order period from September 4, 2016, to October 17, 2018.

Table 1: Data summary

Name	Piped data
Number of rows	99441
Number of columns	8
Column type frequency:	
character	3
Date	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
order_id	0	1	32	32	0	99441	0
customer_id	0	1	32	32	0	99441	0
order_status	0	1	7	11	0	8	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
order_purchase_timestamp	0	1.00	2016-09-04	2018-10-17	2018-01-18	634
order_approved_at	160	1.00	2016-09-15	2018-09-03	2018-01-19	611
order_delivered_carrier_date	1783	0.98	2016-10-08	2018-09-11	2018-01-24	547
order_delivered_customer_date	2965	0.97	2016-10-11	2018-10-17	2018-02-02	645
order_estimated_delivery_date	0	1.00	2016-09-30	2018-11-12	2018-02-15	459

After examining the customer order data, an intriguing observation emerges: most customers on the platform make purchases on Olist only once, accounting for approximately 3.12 percent of the total. Consequently, the platform's behaviour aligns with the preferences and patterns of most customers who place a single order.

Table 4: A summary of percent of total customer who had order more than once

be_more_than_one_order	percent_user
0	96.88
1	3.12

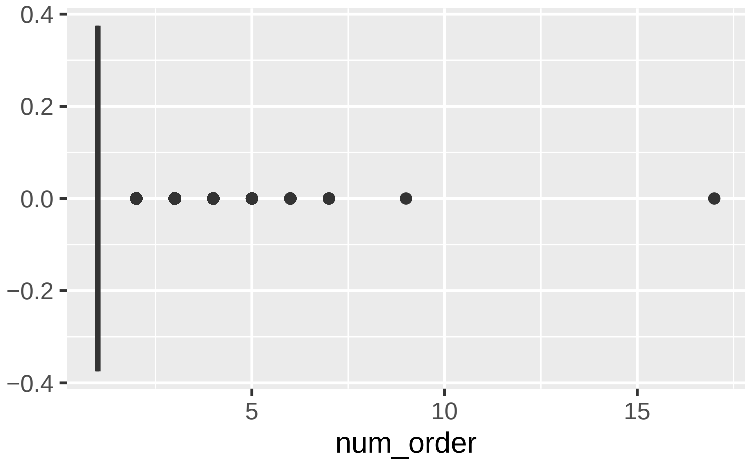

Figure 2: Distribution of amount of order for each customer

Dataset02 - olist_order_customer_dataset
The Olist order customer dataset consists of 99,441 rows, encompassing 96,096 distinct customer IDs, and exhibits complete data coverage in quantity. Considering the number of states, the customers of the Olist platform are distributed across all states in Brazil, spanning over 4,119 cities.

Table 5: Data summary

Name	Piped data
Number of rows	99441
Number of columns	5
Column type frequency:	
character	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
customer_id	0	1	32	32	0	99441	0
customer_unique_id	0	1	32	32	0	96096	0
customer_zip_code_prefix	0	1	5	5	0	14994	0
customer_city	0	1	3	32	0	4119	0
customer_state	0	1	2	2	0	27	0

When examining the geographical distribution of users across all states, São Paulo (SP) emerges as the most prevalent, accounting for 41.98% of the total, followed by Rio de Janeiro (RJ) and Belo Horizonte (MG), with percentages of 12.92% and 11.70%, respectively.

Table 7: Top 5 states following percent of total user in each state of Brazil

customer_state_name	customer_state	percent_user
São Paulo	SP	41.98
Rio de Janeiro	RJ	12.92
Minas Gerais	MG	11.70
Rio Grande do Sul	RS	5.50
Paraná	PR	5.07

Dataset03 - olist_order_reviews_dataset
The Olist order review dataset comprises 99,224 rows, providing complete coverage of review scores. However, there are significant occurrences of missing data in the review comment title and comment fields, accounting for 88.34% and 58.70%, respectively. It is essential to mention that these columns contain Portuguese text. Finally, the review dataset demonstrates a median review score of 5.00 and a mean review score of 4.08.

Table 8: Data summary

Name	Piped data
Number of rows	99224
Number of columns	7
Column type frequency:	
character	4
Date	2
numeric	1
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
review_id	0	1	32	32	0	98410	0
order_id	0	1	32	32	0	98673	0
review_comment_title	0	1	0	26	87656	4528	2
review_comment_message	0	1	0	208	58247	36160	27

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
review_creation_date	0	1	2016-10-02	2018-08-31	2018-02-02	636
review_answer_timestamp	0	1	2016-10-07	2018-10-29	2018-02-04	715

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
review_score	0	1	4.09	1.35	1	4	5	5	5

02 Data preparation

After performing null value checks and eliminating duplicate values using statistical methods in R Studio, these datasets are combined using a join key to integrate all the relevant information for further analysis.

03 Deficiencies and pitfalls

After conducting a comprehensive exploratory data analysis, certain deficiencies have been identified. These include incompleteness regarding review titles and comments and the dataset needing to be updated since 2018. Furthermore, significant challenges and pitfalls arise from concentration risks, specifically concerning the uneven geographical distribution that heavily focuses on São Paulo (SP). Additionally, there are challenges associated with customer behaviour, particularly for single-order customers. Furthermore, significant challenges arise when dealing with NLP-based translation due to the inherent complexity of human language.(Ayata, 2018) Languages inherently possess ambiguity and context-dependence, posing difficulties in developing algorithms that accurately translate Portuguese to English.

Refined problem and Plan

Refined the core questions

After thoroughly reviewing the data quality and gaining a basic understanding of each dataset, the question can be further refined to focus on profitability by considering customer lifetime value and its specific impact on individual customer behaviour. Therefore, the refined question becomes: “How does the sentiment expressed in customer feedback impact the customer lifetime value in the E-commerce industry, particularly when analysed within each specific customer segment?”

Implementation Plan

Step 01: Define the research question and identify reliable data sources.
Step 02: Assess the feasibility of the project, which includes data cleaning, removing duplicates, and merging the datasets into a unified table. Refine the problem statement and identify any potential deficiencies or pitfalls.
Step 03: Conduct a detailed exploratory data analysis to describe the relevant key features of the dataset, identify clusters and patterns, and visualise the data to determine relationships. Identify additional data sources if necessary.
Step 04: Pre-process the data to prepare it for modelling, which may involve dimension reduction and feature extraction techniques.
Step 05: Implement machine learning models, including Natural Language Processing techniques and Classification Algorithms, to analyse the data and extract insights.
Step 06: Evaluate the performance and accuracy of the models using appropriate metrics to assess the approach’s effectiveness.

Alternative plan and questions

Another crucial factor that contributes to a company’s success is profitability. Therefore, it is undeniable that employing methods to increase profits in e-commerce companies and utilising large volumes of data for implementing more efficient pricing strategies are crucial. Machine learning and artificial intelligence, such as predictive analytics, dynamic pricing or demand forecasting, play significant roles across various business units, enabling the development of practical and cost-effective pricing methods.(amazinum, 2022) Therefore, this opportunity of machine implementation leads me to the alternative question: “How does the dynamic pricing approached affect maximising the profit on eCommerce?”

Based on the related analysis fields, the examination will involve selecting the same primary data sources and additional datasets for addressing alternative questions. The accompanying diagram illustrates the supplementary datasets:

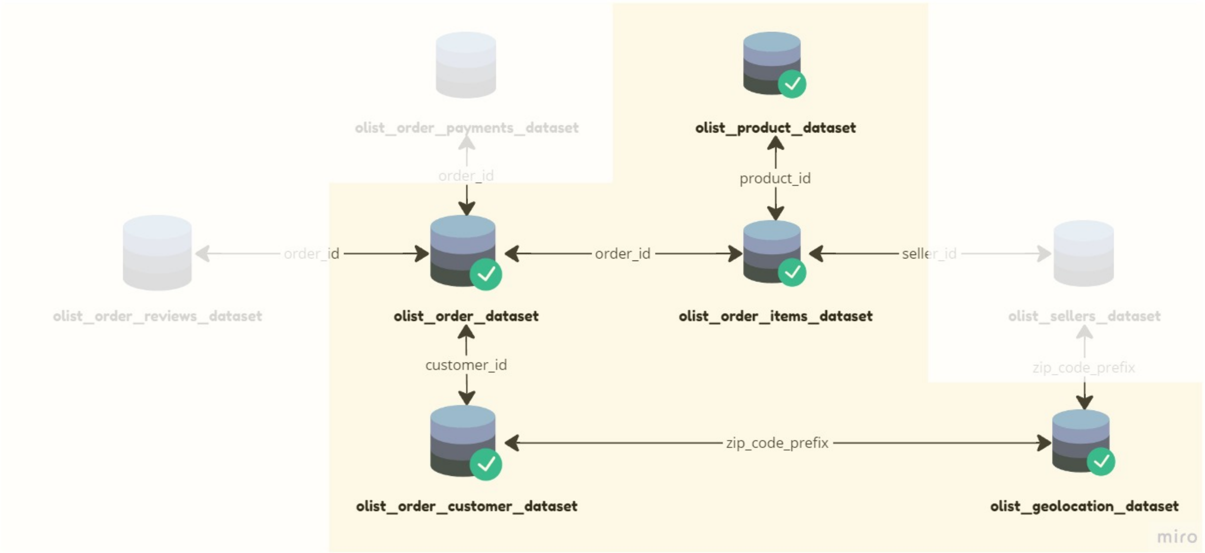

Figure 3: Potential data schema - diagram

04 olist_products_dataset;

- Background: this is the core dataset to provide the information related to products sold by Olist.
- Detail: 32,951 product lists with 73 unique categories and all related product attributes.
- Formats: CSV formats with size 2.38 MB
- join Key: product_id

05 olist_order_items_dataset;

- Background: This dataset has information about the customer and its location
- Detail: 98,666 transactions in item-level records with price, time stamp, and freight value.
- Formats: CSV formats with size 15.44 MB
- join Key: product_id

06 olist_geolocation_dataset;

- Background: This dataset includes Brazilian zip codes and lat/long coordinates. Use it to plot maps and find distances for both customer and seller.
- Detail: 1 M zip code, city, state, and geolocation coordinates.
- Formats: CSV formats with size 61.27 MB
- join Key: geolocation_zip_code_prefix

Reference

amazinum (2022) “Machine learning for e-commerce: Price optimization.” amazinum. Available at: https://amazinum.com/insights/machine-learning-for-e-commerce-price-optimization/.

Ayata, D. (2018) *Applying machine learning and natural language processing techniques to twitter sentiment classification for turkish and english.* PhD thesis.

Bayhaqy, A. *et al.* (2018) “Sentiment analysis about e-commerce from tweets using decision tree, k-nearest neighbor, and naïve bayes,” in *2018 international conference on orange technologies (ICOT)*, pp. 1–6. Available at: https://doi.org/10.1109/ICOT.2018.8705796.

Erl, T., Khattak, W. and Buhler, P. (2016) *Big data fundamentals: Concepts, drivers & techniques.* Prentice Hall (Always learning). Available at: https://books.google.com.au/books?id=GUkYswEACAAJ.

Olist and Sionek, A. (2018) “Brazilian e-commerce public dataset by olist.” Kaggle. Available at: https://doi.org/10.34740/KAGGLE/DSV/195341.

Vanaja, S. and Belwal, M. (2018) “Aspect-level sentiment analysis on e-commerce data,” in *2018 international conference on inventive research in computing applications (ICIRCA)*, pp. 1275–1279. Available at: https://doi.org/10.1109/ICIRCA.2018.8597286.