Table of Content

1.0 Abstract

This paper leverages Business Intelligence tools and methodologies to conduct a comprehensive examination of an e-commerce dataset of an emerging Brazilian e-commerce platform, Olist. Using a varied range of methodologies, this paper uncovers valuable insights while providing strategic optimization recommendations to aid in well informed decision making and potential future strategic planning processes.

The dataset contains a well rounded view of both quantitative and qualitative data that enables a deep and detailed understanding and analysis of both operational efficiency, market dynamics and provides a detailed view of the company's business landscape and performance.

Several methodological techniques are employed to analyse and interpret the data to provide actionable insights. Exploratory Data Analysis, through descriptive statistical calculations and visualizations, reveal significant trends, cyclical patterns, correlations and fluctuations in performance across various variables to convert them into usable intelligence. The paper also probes into sentiment analysis of customer reviews, unearthing their overall satisfaction, opinions and commonalities that could provide insights on customer retention and loyalty.

Employment of clustering techniques, geographical segmentation in this case specifically, categorises customers based on their locale into 5 distinct clusters to examine the logistical optimization potential while the application of predictive modeling using Weka aimed at training different models to identify the highest performing model that is able to capture and learn non linear relationships, returning an r2 value of 0.81 for the Sept 2018 Sales Forecast. Consolidation and visualization of trends and performance metrics for quick decision making and strategic findings via an interactive dashboard was later developed using Power BI, showing steady ~12% YOY growth and identification of a seasonal retail calendar.

The study demonstrates the objective integration of Business Intelligence techniques to enable data driven strategies aiming to boost sales, enhance operations and strengthen competitive positioning while identifying potential optimization strategies in the e-commerce industry.

2.0 Introduction

Olist, a Brazilian e-commerce platform company, offers a SAAS marketplace platform for small and medium retailers. Founded in 2015, at the pivoting point between brick and mortar towards e-commerce, Olist offers a platform and tools that allow merchants to manage everything from inventory oversight to payment handling to logistical support. This widespread access to e-commerce tools allows retailers to connect with a larger group of Brazilian consumers, simultaneously simplifying their operations complexity and overhead expenses.

In addition to marketplace integration, Olist offers marketing tools, performance analytics and partnerships with external logistics providers. This allows them higher order efficiency, delivery tracking and cost management, giving sellers operational insights and market reach.

The publicly available Olist e-commerce dataset comprises about 100,000 orders from 2017 - 2018, made available by Olist, spans customer demographics, product information, payment records and customer feedback and satisfaction, amongst others. This rich dataset enables a range of analytical applications that will be used for analysis through 5 main objectives, Exploratory Data Analysis (EDA) to discover patterns and trends, Sentiment Analysis to evaluate customer feedback to understand customer satisfaction, Clustering to segment customers to strategize potential optimization, Sales Forecasting to model and predict demand, Dashboard Development in Power BI to visualize business performance and identify potential future strategic planning

This paper applies these objectives to discover actionable insights that could enhance customer satisfaction, increase sales efficiency and optimize logistics in Brazil’s competitive yet vast untapped e-commerce sector.

3.0 Literature Review

Electronic commerce(E-commerce) is an activity involving selling and buying of products or services on digital platforms. This process usually involves transactions through online websites, and mobile apps. Singh, S. (2016) mentioned that the rise of e-commerce has transformed the traditional business models into more convenient transactions, by ignoring the constraints of geography or operating hours in the traditional business models. Unlike the traditional brick-and-mortar retail, e-commerce allows business to be connected with customers on a larger scale, expanding their product or service to serve customers globally. As mentioned by Vadwala and Vadwala (2017), expanding the market reach with e-commerce will greatly improve the transaction rate, and reduce the operational cost. The operational cost reductions are achieved by eliminating rental expenses and utilities expenses that are associated with physical retail spaces. Moreover, the digital nature of e-commerce ensures that every transaction can be recorded and saved in a database. These historical transaction data allow data analytics to perform analysis and gain insight on the sales performance, to better design marketing strategy, tailored promotion, and product recommendation. Bilgic and Duan (2019) emphasize that valuable insight obtained from e-commerce data allows better strategic decision-making, ultimately improving customer satisfaction and loyalty. Bachir et al. (2024) demonstrate the importance of necessary operational elements of e-commerce in shaping customer experiences such as shipping speed, payment convenience, and customer demographic adaptability. This supports the company's in strategically service-based and trading-facilitation model, which declare that marketplace efficiency develops the retailer results and buyer trust.

Pandey et al. (2023) used the Brazilian E-Commerce dataset prepared by the company Olist, which has over 100,000 order and customer related data across different Brazilian market regions. The dataset contains different attributes such as order status, price, payment method, freight performance, customer location, product information, and customer feedback. Its well-designed structure is supported in the natural language processing, clustering, sales prediction, delivery performance analysis, and feature engineering of authors work. Resulted in making it a valuable resource for analyzing customer behavior and retailer performance in Brazil’s e-commerce sector.

Moussas et al. (2023) has defined the business process flow and the key functions of Business Intelligence of the Olist company as a strategic cycle that starts with the process of the data collection, integration, analysis, and the communicating of insights to support decision-making. They highlighted the BI’s role in developing the operational efficiency and strategy planning by components such as predictive analytics, reporting, and KPI tracking. The authors gave importance that successful BI adoption requires strong leadership, well structured implementation, ongoing training, and alignment with organizational objectives, making sure that BI becomes an integral part of the organization’s operational and strategic processes. Widjaja et al. (2023) has reported the business process flow and key functions of effective Business Intelligence use through a classification of 53 antecedent factors grouped into 10 categories. These also include BI application features, system capabilities,

data-related factors, and technology, also includes the organizational aspects such as leadership, human resource management, and environmental factors, as well as individual capabilities and behaviors. The authors highlighted that effective BI use proceeds further beyond system access to how it is applied to achieve organizational goals, requiring alignment between technological, organizational, and human elements for optimal performance, and improving sales being one of the key objectives in most companies.

Chate et al. (2022) has described the dependent or targeted variable as the review score for the next order in the Olist e-commerce dataset. The author used to classify customer feedback into positive (score > 3) or negative (score ≤ 3). This binary classification target was central to the study's aim of predicting customer satisfaction levels using machine learning models.

Anitha and Sherly (2024 received and published in 2025) has reported that the dependent or targeted variable as customer churn, showing whether a customer is likely to leave the platform. Using the Brazilian Olist and E-commerce Sales datasets, the author's study predicted the binary outcome through a GraphSAGE model built on features derived from RFM scoring and aspect-oriented sentiment analysis.

Widjaja et al. (2023) has conducted his study using a systematic literature review, following with the Ghapanchi and Aurum's four-step approach which is used for the identifying of research resources, selecting studies, integrating and arranging the data, and presenting results. They also performed a search to determine keywords and focused on three databases such as the Business Source Complete, Scopus, and ProQuest Central—applying the PRISMA framework to screen and filter articles published between 2000 and 2023. Whereas, only the quantitative, peer-reviewed journal articles providing statistical proof of the factors that influence BI use were included, resulting in 36 eligible studies for analysis.

Xu, Tan, Wang, and Li's study in 2023 gives the similarities with other Olist-focused research in using the sales and customer feedback data to test the platform performance. They did that by identifying service gaps, and proposing improvements for customer satisfaction and the operational effectiveness. However, it varies more on concentrating the seasonal sales trends, supplier cooperation issues, and delivery efficiency, while also allowing the limitations such as a single-year dataset and a lack of broader, more innovative analytical approaches.

Alzami et al. along with Sambasri, F. D., Nabila, M., Megantara, R. A., Akrom, A., Pramunendar, R. A., and Sulistiyawati, P. in the year 2023 demonstrated strong domain knowledge in e-commerce customer segmentation. They focused specifically within the Customer-to-Customer (C2C) business model. They have applied the RFM method to assess customer loyalty and purchasing behavior. Their process involved merging multiple Olist e-commerce datasets, performing EDA, and using K-Means clustering to segment customers, with insights displayed via a Streamlit dashboard that has the supporting targeted marketing strategies and improved customer relationship management.

4.0 Methodology

4.1 Olist E-Commerce Dataset

The e-commerce sales dataset was obtained from a database called Kaggle. This dataset is the e-commerce dataset provided by the company called Olist. There are 7 different dataset files provided, with each containing different information data across different departments. These 6 different files were able to be merged into one main dataset using the common identifier. Figure 1 below shows the relationship between 7 different dataset files and the common identifier that links between the dataset.

Figure 1: Relationship Between Raw Dataset

After merging all the datasets into one single dataset, the resulting dataset has 38 columns and 94,842rows. From the total of 38 columns, 23 of the variables are categorical and 15 of the variables are numerical.

Data cleaning was performed using python in the google colab platform. Key steps involved in data cleaning are:

- Handling missing values:
  All the numerical missing values were replaced by the mean of their respective column, preserving the size of the dataset.
- Handling inconsistent formatting:
  All the categorical variables were reformatted into a uniform format especially for the date variable to ensure the compatibility and consistency of the dataset.

After data cleaning, feature engineering was performed to further enhance the usefulness of the dataset for analytical usage. In this step, additional features were created from the existing features to represent a more informative pattern in the dataset. Nine variables were created from the feature engineering step including five categorical and four numerical. The five

categorical variables are: general product categories, customer full state name, delivery status, binned product volume, and binned product weight. The four numerical variables are: order year, order month, product volume, and actual delivery days.

The final dataset contains a total 47 variables where 28 are categorical variables and 19 are numerical variables and is fully prepared for subsequent analysis and modelling to extract insights and support decision-making.

Numerical Features	Elaboration
geolocation_lat	Latitude of customer location
geolocation_lng	Longitude of customer location
price	item price
freight_value	item freight value item (if an order has more than one item the freight value is splitted between items)
product_name_length	number of characters extracted from the product name.
product_description_length	number of characters extracted from the product description.
product_photos_qty	number of product published photos
product_weight_g	product weight measured in grams.
product_length_cm	product length measured in centimeters.
product_height_cm	product height measured in centimeters.
product_width_cm	product width measured in centimeters.
payment_sequential	a customer may pay an order with more than one payment method. If he does so, a sequence will be created to
payment_installments	number of installments chosen by the customer.
payment_value	transaction value.
order_year	Year of order
order_month	Month of order
product_volume_cm3	Product volume measured in cm^3
actual_delivery_days	Actual delivery days
review_score	Review score from 0 to 5

Table 1: Elaboration of Numerical Features

Categorical Features	Elaboration
customer_zip_code_prefix	first five digits of customer zip code
customer_id	key to the orders dataset. Each order has a unique customer_id.
customer_unique_id	unique identifier of a customer.
customer_city	customer city name
customer_state	customer state displayed in prefix
order_id	unique identifier of the order.
order_status	Reference to the order status (delivered, shipped, etc).
order_item_id	order unique identifier
product_id	product unique identifier
seller_id	seller unique identifier
shipping_limit_date	Shows the seller shipping limit date for handling the order over to the logistic partner.
product category	root category of product
general categories	Grouped product category
customer_state_name	customer state
delivery_status	Status of delivery
payment_type	method of payment chosen by the customer
review_id	unique review identifier
review_comment_title	Comment title from the review left by the customer, in Portuguese.
review_comment_message	Comment message from the review left by the customer, in Portuguese.
review_creation_date	Shows the date in which the satisfaction survey was sent to the customer.
review_answer_timestamp	Shows satisfaction survey answer timestamp.
order_purchase_timestamp	Shows the purchase timestamp.
order_approved_at	Shows the payment approval timestamp.
order_delivered_carrier_date	Shows the order posting timestamp. When it was handled to the logistic partner.
order_delivered_customer_date	Shows the actual order delivery date to the customer.

order_estimated_delivery_date	Shows the estimated delivery date that was informed to customer at the purchase moment.
binned_product_volume	Binned/ Categorized product volume
binned_product_weight	Binned/ Categorized product weight

Table 2: Elaboration of Categorical Features

4.2 Methodology for Business Objectives

4.2.1 Exploratory Data Analysis

The methodology flow for this objective consists of four main stages: Numerical Univariate Analysis, Categorical Univariate Analysis, Correlation Analysis, Bivariate Categorical Analysis

Numerical Univariate Analysis

The dataset was loaded from "main_dataset (2).csv" and a set of numerical variables, such as price, freight_value, and actual_delivery_days, was selected for univariate analysis. For each variable, missing values were excluded before calculating key descriptive statistics, including count, number of missing values, sum, mean, median, minimum, maximum, range, variance, standard deviation, quartiles (Q1 and Q3), interquartile range (IQR), skewness, and excess kurtosis. Outliers were identified using the IQR method, where values below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ were flagged and counted. These measures provided a detailed profile of the distribution, spread, and shape of each variable, as well as insights into potential outliers. The results were compiled into a single table and exported as "numerical_univariate_analysis.csv" for further examination in Google Sheets, enabling a clearer understanding of data patterns, variability, and quality issues before proceeding with deeper statistical analysis.

Categorical Univariate Analysis

We conducted a univariate analysis of all categorical variables in the dataset to understand their distribution and identify the most common values. First, we specified a list of relevant categorical columns, including customer details, order information, product classifications, and delivery status. For each column, we calculated the total number of entries, the number of unique values, and the count of missing values. We then generated a frequency table for the top 10 most common categories in each variable, along with their counts and percentage share. To support visual inspection, we prepared plots for each variable and stored them in a dedicated folder. The results were compiled into a summary table, showing for each column the number of unique values, the most frequent category, its count, and percentage, enabling quick identification of dominant trends and potential data quality issues.

Correlation Analysis

To identify relationships between numerical features, a Pearson correlation matrix was computed. Pearson’s method was chosen because it measures the linear relationship between continuous variables. The resulting correlation matrix was visualized as a heatmap using the Seaborn library, allowing for quick identification of strong positive or negative correlations.

This step focused exclusively on numerical variables, with all categorical features excluded from the analysis.

Bivariate Categorical Analysis
Categorical variables were identified from the dataset, excluding high-cardinality identifiers and timestamps. For each pair of remaining categorical variables, the five most frequent categories from each variable were selected to reduce noise. The dataset was filtered to include only these top categories, and contingency tables (crosstabs) were generated to show the frequency distribution of category combinations. Results were stored for all variable pairs, with sample outputs reviewed to identify notable associations.

4.2.2 Clustering

The methodology flow for this objective consists of six main stages; Data Cleaning, Feature Standardization, Model Evaulation, Model Selection, Clustering and Geographic Visualization.

The geolocation dataset was imported and latitude and longitude data along with customer zip code was used as the primary features for clustering, with rows containing missing values removed. The data was standardised using StandardScaler to ensure zero mean and unit variance, preventing scale related biases in clustering.

Three clustering algorithms were evaluated, KMeans, Agglomerative Hierarchical Clustering (HAC), and Gaussian Mixture Model (GMM). KMeans was applied with fixed clusters and random seed for reproducibility. HAC was run using Ward's linkage criterion to minimize cluster variance. GMM clustering enabled soft probabilistic cluster assignments by assuming data originated from a Gaussian mixture distribution.The results showed that KMeans had the highest silhouette score of 0.4968 and was then chosen as the strongest model for clustering.

KMeans clustering was applied with 5 clusters and a fixed random state of 42 to ensure reproducibility. Cluster visualization was performed using a scatter plot with longitude on the x-axis and latitude on the y-axis.

KMeans was later used to introduce temporary centroids by assigning each customer to the nearest centroid and recalculating them as the average location of all customers in the cluster until it converges. This allowed determination of individual centroids for each cluster.

With the centroids identified, geographic visualization was implemented using the Folium mapping library to overlay the map behind the clusters and centroids to give a better representation of the clusters. A base map of Brazil was generated and each cluster specific data point was plotted in varying colors for differentiation between clusters. Centroids were displayed as larger icons for clear identification.

4.2.3 Sentiment Analysis

The methodology flow for this objective consists of six main stages: Data Cleaning, Sentiment Prediction, Score Integration, Sentiment Categorization, Descriptive Analysis, and Relationship Analysis.

The raw dataset (olist_order_reviews_dataset.csv) was first pre-processed by removing records with empty review messages. Sentiment analysis was then conducted using a pre-trained Portuguese BERT-based model (lipaoMai/bert-sentiment-model-portuguese). The model was executed on a GPU where available to optimize computational efficiency. Reviews were processed in batches to predict both the categorical sentiment label (Positive, Neutral, or Negative) and an associated probability score representing model confidence.

The output were then used to integrate both the subjective rating and the linguistic sentiment, the review_score and associated probability score representing model confidence variables were converted to numeric form. Review scores (ranging from 1 to 5) were normalized to a 0–1 scale to ensure comparability with sentiment probability scores. An overall sentiment score was then computed using a weighted formula, assigning 60% weight to the normalized review score and 40% weight to the sentiment probability.

And finally The overall sentiment scores were subsequently categorized into discrete sentiment classes based on predefined thresholds:

Value	Rating
≥ 0.66	Positive
0.33 to < 0.66	Neutral
< 0.33	Negative

Table 3. Overall Sentiment Scores

This categorization facilitated qualitative interpretation of the results and supported downstream analysis. Following the computation and categorization of overall sentiment scores, a set of descriptive and visual analyses was conducted to summarize the sentiment distribution within the dataset.

The dataset was analyzed to assess sentiment polarity and score distribution. Frequency counts and proportions were calculated for each sentiment category (Positive, Neutral, Negative), and results were presented in a combined table. A bar chart, color-coded by sentiment polarity, was used to visualize category distribution. For the continuous variable overall_sentiment_score, descriptive statistics were computed, and its distribution was examined using a histogram with a kernel density overlay to assess central tendency, variability, and skewness.

The relationship between sentiment outcomes and other variables was analyzed separately for numerical and categorical data. Numerical variables were assessed against overall_sentiment_score using Pearson and Spearman correlations, with results ranked by absolute Pearson correlation. Categorical variables were tested against overall_sentiment_label using the Chi-square test of independence, with Cramér's V measuring association strength and results ranked accordingly. This process identified the variables most strongly linked to sentiment for deeper interpretation and modeling.

4.2.4 Sales Forecasting Modelling

The methodology flow for this objective consists of four main stages: Dataset Development, Feature Selection, Model Selection, and Modelling & Evaluation.

The sales forecasting objectives were set to be forecasting the monthly sales. From the original merged dataset, each row entries were representing a unique customer with their product purchased.

A dataset was purposely created using python in google colab to obtain the monthly sales figure from the original dataset. By performing feature engineering, the unique combination of year and month were extracted from the customer purchase date variable in the original dataset. Then, each unique year and month combination was used as a grouping key to get aggregated features including total sales, product rating, delivery rating, customer satisfaction, delivery days, product weight, product volume, number of product photos, product description length, product name length, and advertising spending. All these newly engineered features provide a monthly level view of sales amount and product/customer-related attributes.

After creating a dataset that is suitable to predict monthly sales, the next step was to perform feature selection to obtain a strong predictor for monthly sales. The feature selection was done by correlation analysis, conducted using Microsoft Excel, with monthly total sales as the dependent variable. From the correlation matrix, high correlated features will be selected as the final predictor variable for forecasting models.

In the next step, models that are used for sales forecasting modelling will be selected. This model selection was based on the related work and previous research in sales forecasting. Data mining models will be selected based on their proven performance and popularity in previous research.

In the final modeling and Evaluation step, all selected models will be trained and tested in Weka. The Weka Explorer interface was used for training and evaluation, with 10-fold cross-validation applied to ensure that model performance was assessed reliably across different subsets of the data.

4.2.5 Power BI Dashboard

There are mainly 7 stages to this, KPI Overview, Categorical Analysis, Temporal Analysis, Operational Performance, Customer Satisfaction, Geographical Analysis, Actionable Insights

Several methodologies were adopted in creating the Power BI dashboard. The overall layout is a top-down KPI-driven approach, starting from overall key performance indicators that later drill down into the details. Categorical analysis was conducted to show sales distribution by product category and payment type allowing for identification of top performing segments while temporal analysis, to examine monthly sales performance and potential identification of seasonal purchase behavior to identify high growth and low periods. Operational performance metrics were included to highlight delivery time distribution and delivery patterns to analyze logistical efficiency. Customer satisfaction was determined through review score analysis, allowing association between operational factors and customer satisfaction. Visualized geographical mapping of customer purchase behavior gives insights to customer demand concentrations, and allows for logistics planning and inventory management. Combined, these methodologies allow actionable, data-driven analysis of business performance and future strategic planning.

5.0 Result and Analysis

5.1 Relationship between Main dataset and AI dataset

The main dataset itself is a cleaned and merged version of several datasets that was done for easier analysis. The generated AI dataset contains further customer data that relates to the main dataset through “Customer ID” and the full relationships between datasets can be seen in Power BI.

5.2 Exploratory Data Analysis

5.2.1 Univariate Analysis

5.2.1a Univariate Numerical Feature Analysis

The results are collected and shown below. Table of Numerical Features count, sum, mean, median, first quartile, third quartile, max, range, variance, standard deviation, IQR, skewness ,excess kurtosis and outlier count.

Column	Count	Sum	Mean	Median	Min	Q1	Q3	Max	Range	Variance	Std Dev	IQR	Skewness	Excess Kurtosis	Outlier Count
geolocation_lat	94842.00	-2010904.00	-21.20	-22.93	-36.61	-23.59	-20.15	42.18	78.79	31.48	5.61	3.44	1.66	3.54	15491.00
geolocation_lng	94842.00	-4380423.00	-46.19	-46.63	-72.67	-48.11	-43.63	-8.58	64.09	16.52	4.06	4.49	0.03	2.34	4014.00
price	94842.00	11890954.00	125.38	79.00	0.85	41.80	139.90	6735.00	6734.15	36045.35	189.86	98.10	7.83	118.54	7268.00
freight_value	94842.00	1914863.00	20.19	16.39	0.00	13.30	21.25	409.68	409.68	250.86	15.84	7.95	5.60	58.98	10362.00
product_name_length	94842.00	4635358.00	48.87	52.00	5.00	43.00	57.00	76.00	71.00	99.69	9.98	14.00	-0.92	0.18	1025.00
product_	94842.00	75260954	793.54	607.00	4.00	348.00	996.00	3992.00	3988.00	427846.1	654.10	648.00	1.99	4.83	5835.00

description_length		.00								0					
product_photos_qty	94842.00	213542.00	2.25	2.00	1.00	1.00	3.00	20.00	19.00	3.05	1.75	2.00	1.85	4.49	2798.00
product_weight_g	94842.00	20000000.00	2104.49	700.00	0.00	300.00	1813.00	40425.00	40425.00	14123465.00	3758.12	1513.00	3.61	16.40	13543.00
product_length_cm	94842.00	2859857.00	30.15	25.00	7.00	18.00	38.00	105.00	98.00	260.27	16.13	20.00	1.76	3.75	3059.00
product_height_cm	94842.00	1563008.00	16.48	13.00	2.00	8.00	20.00	105.00	103.00	177.10	13.31	12.00	2.27	7.53	6354.00
product_width_cm	94842.00	2187230.00	23.06	20.00	6.00	15.00	30.00	118.00	112.00	137.67	11.73	15.00	1.72	4.63	2167.00
payment_sequential	94842.00	96962.00	1.02	1.00	1.00	1.00	1.00	19.00	18.00	0.05	0.23	0.00	25.43	1163.08	1545.00
payment_installments	94842.00	276715.00	2.92	2.00	0.00	1.00	4.00	24.00	24.00	7.35	2.71	3.00	1.61	2.43	5961.00
pay	9484	1492	157.	103.	0.01	60.0	175.	1366	1366	4682	216.	115.	9.48	260.	7356

ment_value	2.00	8386.00	40	12		1	04	4.08	4.07	9.64	40	03		34	.00
order_year	94842.00	19100000.00	2017.55	2018.00	2016.00	2017.00	2018.00	2018.00	2.00	0.25	0.50	1.00	-0.25	-1.77	0.00
order_month	94842.00	57251 0.00	6.04	6.00	1.00	3.00	8.00	12.00	11.00	10.38	3.22	5.00	0.21	-0.97	0.00
product_volume_cm3	94842.00	14400000 00.00	15210.65	6450.00	168.00	2816.00	18375.00	296208.00	296040.00	5440000 0.00	23329.44	15559.00	4.09	25.79	83778.00
actual_delivery_days	94842.00	1145835.00	12.08	10.00	0.00	6.00	15.00	209.00	209.00	91.06	9.54	9.00	3.83	39.44	4931.00

Table 4. Result of Univariate Numerical Feature Analysis

The numerical feature analysis provided a comprehensive overview of the dataset's continuous variables, including central tendency, dispersion, and distribution characteristics. For geolocation coordinates, the latitude values ranged from –36.61 to 42.18 with a mean of –21.20, while longitude ranged from –72.67 to –8.58, indicating the geographical spread of customers and sellers across Brazil. The price variable displayed high variability (range of 6,734.15), with a positively skewed distribution (skewness = 7.83) due to the presence of extreme high-value products. Similarly, freight_value showed a moderate positive skew (5.60) and outliers in high delivery fees.

Product-related dimensions such as product_name_length, product_description_length, product_photos_qty, and product_weight_g also exhibited noticeable variance. In particular, product_description_length was heavily right-skewed (skewness = 1.99) with a wide range up to 3,992 characters, while product_weight_g displayed extreme kurtosis (16.40), highlighting the presence of very heavy items in the catalog. Dimensional attributes like product_length_cm, product_height_cm, and product_width_cm were moderately skewed, with some unusually large products contributing to the observed outliers.

In payment-related metrics, payment_value had a long-tailed distribution (skewness = 9.48) influenced by a small proportion of high-value transactions. The order_year and order_month fields showed minimal variability, consistent with the dataset's temporal coverage. The actual_delivery_days variable, while generally clustered around the median of 10 days,

revealed extreme values (up to 209 days) that represent delayed deliveries. These findings indicate the dataset contains both typical e-commerce transactions and a small fraction of extreme cases that could significantly affect model training if not treated.

5.2.1b Univariate Categorical Feature Analysis

The results are collected and shown below. Table of Categorical Features unique values, most frequent category, most frequent count, and most frequent %

Column	Unique Values	Most Frequent Category	Most Frequent Count	Most Frequent %
customer_zip_code_prefix	14695	22790	134	0.141287615
customer_id	94842	274fa6071e5e17fe303b9748641082c8	1	0.001054385
customer_unique_id	91807	8d50f5eadf50201ccdcedfb9e2ac8455	14	0.014761393
customer_city	4029	sao paulo	14803	15.60806394
customer_state	27	SP	39926	42.09738302
order_id	94842	28db69209a75e59f20ccbb5c36a20b90	1	0.001054385
order_status	2	delivered	94836	99.99367369
order_item_id	1	1	94842	100
product_id	30546	99a4788cb24856965c36a24e339b6058	427	0.450222475
seller_id	2901	6560211a19b47992c3666cc44a7e94c0	1804	1.902110879
shipping_limit_date	89564	11/6/2018 3:31	6	0.006326311
product category	73	bed table bath	9149	9.646570085
general categories	7	Clothing and Fashion	30648	32.31479724

customer_state_name	27	Sao Paulo	39926	42.09738302
delivery_status	2	On Time	87168	91.90864807
payment_type	4	credit_card	71986	75.90097214
order_purchase_timestamp	94342	2/8/2018 12:06	3	0.003163156
order_approved_at	86879	27/2/2018 4:31	9	0.009489467
order_delivered_carrier_date	78714	9/5/2018 15:48	47	0.049556104
order_delivered_customer_date	94058	2/12/2017 0:26	3	0.003163156
order_estimated_delivery_date	444	20/12/2017 0:00	502	0.529301364
binned_product_volume	5	Very Small	18998	20.0312098
binned_product_weight	5	Very Light	22090	23.2913688

Table 5.Result of Univariate Categorical Feature Analysis

Categorical feature analysis revealed varying degrees of diversity and concentration among categories. Customer location identifiers (e.g., customer_zip_code_prefix, customer_city, customer_state) showed high cardinality, with São Paulo (SP) emerging as the dominant state, accounting for over 42% of all orders. Similarly, in customer_city, “sao paulo” represented 15.6% of transactions, confirming the city’s role as a major e-commerce hub.

The order_status field was overwhelmingly dominated by “delivered” orders (99.99%), reflecting the dataset’s post-fulfillment nature. Product-related identifiers (product_id, seller_id) showed a highly fragmented distribution, though some products and sellers appeared disproportionately often, with the top product accounting for 0.45% of orders. The product category field included 73 categories, with “bed table bath” being the most frequent (9.65%), while the higher-level general categories grouped these into 7 broad segments, the largest being “Clothing and Fashion” (32.31%).

Shipping-related categorical variables indicated that 91.9% of orders were delivered “On Time.” Payment methods were concentrated on credit card usage (75.9%) .Temporal identifiers (order_purchase_timestamp, order_approved_at, etc.) had high uniqueness,

consistent with timestamp-level granularity. The engineered binned variables (binned_product_volume and binned_product_weight) revealed that the majority of products were classified as “Very Small” (20.03%) and “Very Light” (23.29%), aligning with typical consumer goods profiles in online retail.

5.2.2 Bivariate Analysis

5.2.2a Bivariate Analysis Numerical

The results are collected and shown below. Are a table of numerical feature correlation between all numerical features.

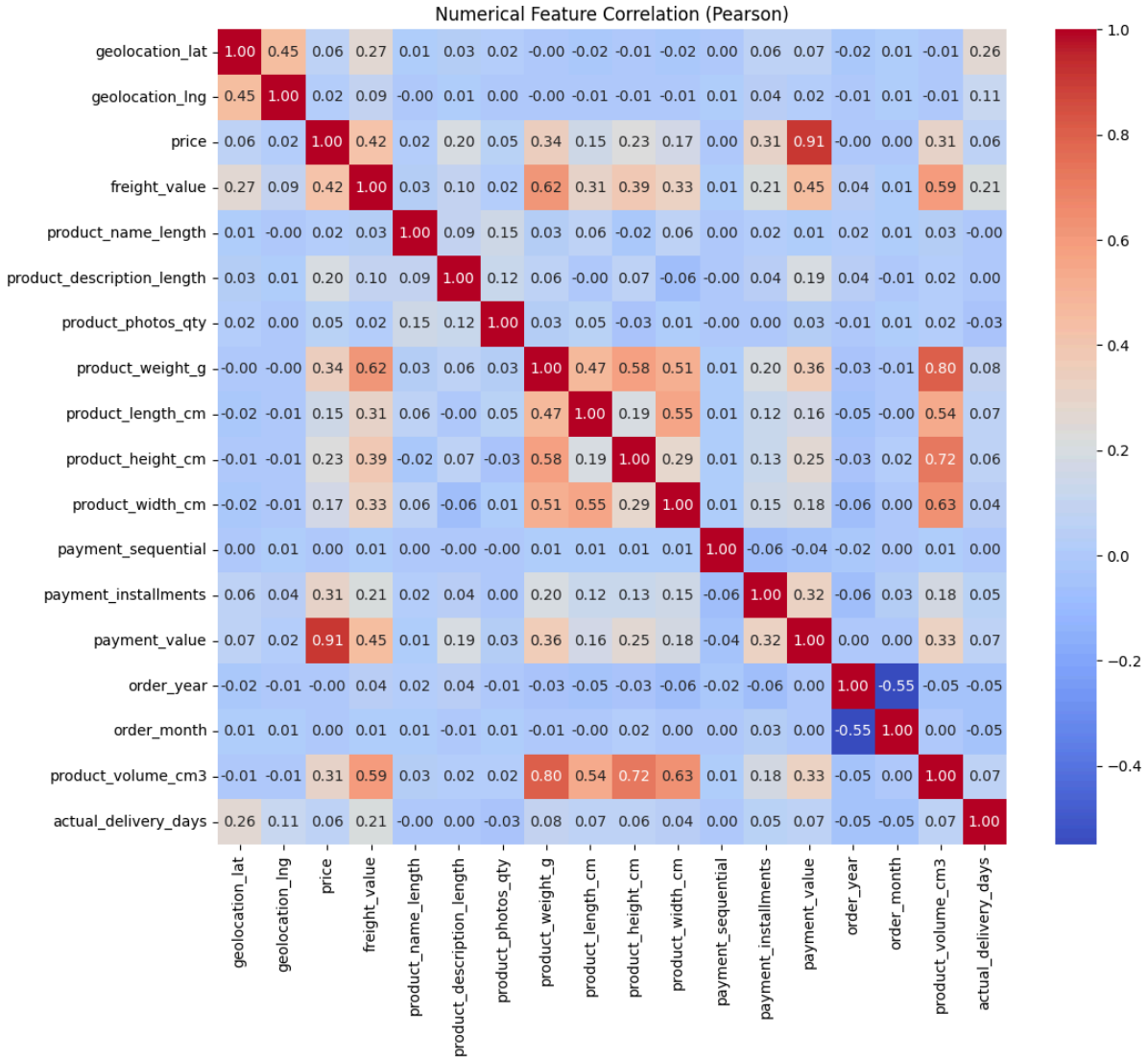


Figure 2. Bivariate Analysis for Numerical

The Pearson correlation analysis among the numerical features reveals several meaningful relationships in the dataset. Notably, delivery-related variables such as actual_delivery_days show a moderate positive correlation with geolocation_lat (0.26) and freight_value (0.21), suggesting that delivery times tend to be longer in certain locations and for higher shipping costs. The financial variables price and payment_value are highly correlated (0.91),

indicating that most purchases are fully paid and that payment closely reflects product price. Similarly, physical product dimensions—product_weight_g, product_length_cm, product_height_cm, and product_width_cm—show strong inter-correlations (ranging from 0.47 to 0.72), which is expected as larger products tend to weigh more and have larger dimensions. Moderate correlations are also observed between product_volume_cm3 and weight/dimensions, reflecting the natural geometric relationships among these features.

Other variables exhibit relatively weak correlations, indicating they are largely independent of each other. For example, order_year and order_month show minimal correlation with most numerical features, except a slight negative relationship with order_month and order_year themselves (-0.55), which may reflect seasonal trends across different years. Payment-related variables such as payment_installments show modest correlations with price (0.31) and payment_value (0.32), suggesting that higher-value purchases are more likely to be split into installments. Overall, these correlations provide insight into the structural relationships in the dataset, highlighting clusters of interrelated variables (such as product attributes and financial/payment features) while also identifying features that are largely independent, which could be particularly informative for predictive modeling or feature selection.

5.2.2b Bivariate Analysis Categorical

Below shows 76 contingency tables of top counted categories against top counted categories.

```
Bivariate Crosstab: customer_state_name vs delivery_status
delivery_status      Delayed  On Time
customer_state_name
Minas Gerais             272     4754
Parana                   116     2026
Rio Grande do Sul        176     2199
Rio de Janeiro           735     4842
Sao Paulo               1088    16710

Bivariate Crosstab: customer_state_name vs payment_type
payment_type          UPI  credit_card  debit_card  voucher
customer_state_name
Minas Gerais         1005         3835          63      123
Parana                486         1560          29       67
Rio Grande do Sul     584         1708          30       53
Rio de Janeiro       1009         4318          79      171
Sao Paulo            3449        13588         332      429

Bivariate Crosstab: customer_state_name vs binned_product_volume
binned_product_volume  Extra Large  Large  Medium  Small  Very Small
customer_state_name
Minas Gerais                  1001    973    1059    981        1012
Parana                         429    398     436    429         450
Rio Grande do Sul              516    522     474    424         439
Rio de Janeiro                1176   1147    1106   1167         981
Sao Paulo                     3611   3404    3558   3699        3526

Bivariate Crosstab: customer_state_name vs binned_product_weight
binned_product_weight  Heavy  Light  Medium  Very Heavy  Very Light
customer_state_name
Minas Gerais            1005    921     899        1049        1152
Parana                   461    401     358         400         522
Rio Grande do Sul        518    407     446         476         528
Rio de Janeiro          1211    996    1004        1195        1171
Sao Paulo               3587   3362    3186        3347        4316

Bivariate Crosstab: delivery_status vs payment_type
payment_type      UPI  credit_card  debit_card  voucher
delivery_status
Delayed           771         2587          49       86
On Time          7710        29672         603     1006

Bivariate Crosstab: delivery_status vs binned_product_volume
binned_product_volume  Extra Large  Large  Medium  Small  Very Small
delivery_status
Delayed                        764    670     672    776         611
On Time                       7791   7687    7874   7770        7869

Bivariate Crosstab: delivery_status vs binned_product_weight
binned_product_weight  Heavy  Light  Medium  Very Heavy  Very Light
delivery_status
Delayed                  735    647     586         786         739
On Time                 7863   7328    7026        7583        9191

Bivariate Crosstab: payment_type vs binned_product_volume
binned_product_volume  Extra Large  Large  Medium  Small  Very Small
payment_type
UPI                           1535   1647    1690   1713        1896
credit_card                   6743   6344    6502   6469        6201
debit_card                      91    119     147    144         151
voucher                        186    247     207    220         232

Bivariate Crosstab: payment_type vs binned_product_weight
binned_product_weight  Heavy  Light  Medium  Very Heavy  Very Light
payment_type
UPI                     1643   1509    1479        1559        2291
credit_card             6605   6137    5796        6555        7166
debit_card               113    122     117          91         209
voucher                  237    207     220         164         264

Bivariate Crosstab: binned_product_volume vs binned_product_weight
binned_product_weight  Heavy  Light  Medium  Very Heavy  Very Light
binned_product_volume
Extra Large             1850    224     714        5714          53
Large                   3898    677    1902        1632         248
Medium                  1909   1989    2746         745        1157
Small                    766   2974    1428         223        3155
Very Small               175   2111     822          55        5317
```

Figure 3.Bivariate Analysis Categorical 1

```
Bivariate Crosstab: customer_state vs binned_product_volume
binned_product_volume  Extra Large  Large  Medium  Small  Very Small
customer_state
MG                            1001    973    1059    981        1012
PR                             429    398     436    429         450
RJ                            1176   1147    1106   1167         981
RS                             516    522     474    424         439
SP                            3611   3404    3558   3699        3526

Bivariate Crosstab: customer_state vs binned_product_weight
binned_product_weight  Heavy  Light  Medium  Very Heavy  Very Light
customer_state
MG                      1005    921     899        1049        1152
PR                       461    401     358         400         522
RJ                      1211    996    1004        1195        1171
RS                       518    407     446         476         528
SP                      3587   3362    3186        3347        4316

Bivariate Crosstab: order_status vs product category
product category  Furniture Decoration  HEALTH BEAUTY  bed table bath  \
order_status
delivered                         2796           3871            4078

product category  computer accessories  sport leisure
order_status
delivered                         2860           3339

Bivariate Crosstab: order_status vs general categories
general categories  Clothing and Fashion  Consumer Electronics  \
order_status
canceled                               1                     0
delivered                          13732                  8174

general categories  DIY and Hardware  Home and Furniture  \
order_status
canceled                         0                   0
delivered                     2362               11139

general categories  Media and Entertainment
order_status
canceled                                0
delivered                            5793

Bivariate Crosstab: order_status vs customer_state_name
customer_state_name  Minas Gerais  Parana  Rio Grande do Sul  Rio de Janeiro  \
order_status
canceled                       0       0                  0               1
delivered                   5026    2142               2375            5576

customer_state_name  Sao Paulo
order_status
canceled                     0
delivered                17798

Bivariate Crosstab: order_status vs delivery_status
delivery_status  Delayed  On Time
order_status
canceled               0        1
delivered           3493    38990

Bivariate Crosstab: order_status vs payment_type
payment_type   UPI  credit_card  debit_card  voucher
order_status
canceled         0            1           0        0
delivered     8481        32259         652     1092

Bivariate Crosstab: order_status vs binned_product_volume
binned_product_volume  Extra Large  Large  Medium  Small  Very Small
order_status
canceled                         0      0       1      0           0
delivered                     8555   8357    8545   8546        8480

Bivariate Crosstab: order_status vs binned_product_weight
binned_product_weight  Heavy  Light  Medium  Very Heavy  Very Light
order_status
canceled                   0      0       1           0           0
delivered               8598   7975    7611        8369        9930
```

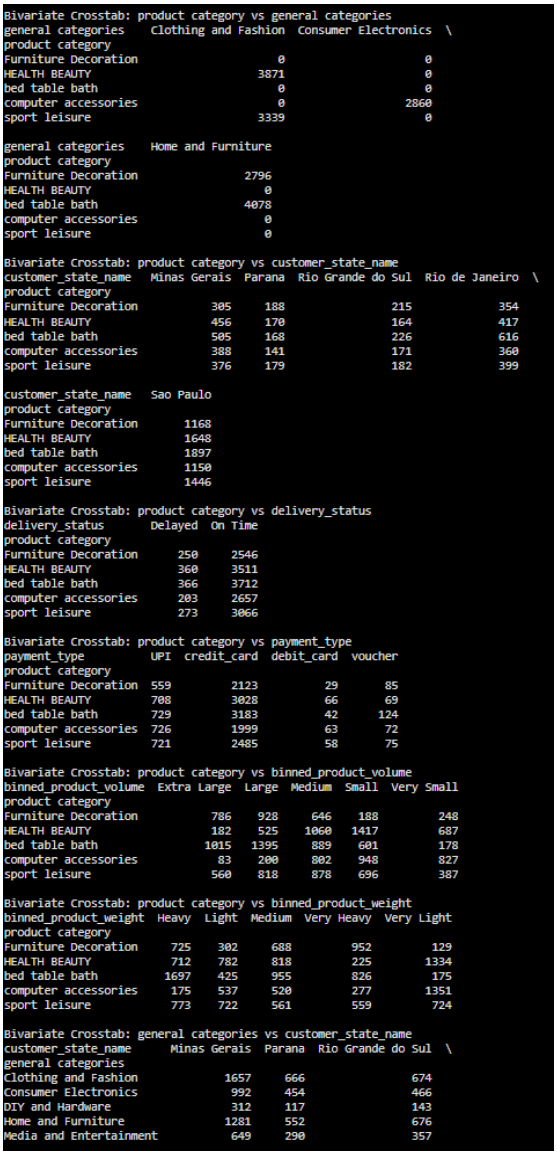
Figure 4.Bivariate Analysis Categorical 2

```
Bivariate Crosstab: customer_city vs general categories
general categories  Clothing and Fashion  Consumer Electronics  \
customer_city
belo horizonte                       379                   222
brasilia                             309                   161
curitiba                             204                   131
rio de janeiro                       866                   530
sao paulo                           2168                  1173

general categories  DIY and Hardware  Home and Furniture  \
customer_city
belo horizonte                    77                 313
brasilia                          33                 202
curitiba                          30                 161
rio de janeiro                   158                 883
sao paulo                        303                1907

general categories  Media and Entertainment
customer_city
belo horizonte                           158
brasilia                                 120
curitiba                                  88
rio de janeiro                           454
sao paulo                                838

Bivariate Crosstab: customer_city vs customer_state_name
customer_state_name  Minas Gerais  Parana  Rio de Janeiro  Sao Paulo
customer_city
belo horizonte              1195       0               0          0
curitiba                       0     635               0          0
rio de janeiro                 0       0            2992          0
sao paulo                      0       0               0       6617

Bivariate Crosstab: customer_city vs delivery_status
delivery_status  Delayed  On Time
customer_city
belo horizonte        71     1124
brasilia              68      778
curitiba              35      600
rio de janeiro       342     2650
sao paulo            431     6186

Bivariate Crosstab: customer_city vs payment_type
payment_type     UPI  credit_card  debit_card  voucher
customer_city
belo horizonte   208          948          18       21
brasilia         154          661           6       25
curitiba         149          454          10       22
rio de janeiro   530         2316          54       92
sao paulo       1176         5148         131      162

Bivariate Crosstab: customer_city vs binned_product_volume
binned_product_volume  Extra Large  Large  Medium  Small  Very Small
customer_city
belo horizonte                 222    222     258    245         248
brasilia                       149    158     202    177         160
curitiba                       123    114     125    136         137
rio de janeiro                 628    614     601    627         522
sao paulo                     1228   1214    1378   1406        1391

Bivariate Crosstab: customer_city vs binned_product_weight
binned_product_weight  Heavy  Light  Medium  Very Heavy  Very Light
customer_city
belo horizonte           207    233     228         238         289
brasilia                 153    174     182         136         201
curitiba                 117    119     102         121         176
rio de janeiro           645    528     524         658         637
sao paulo               1298   1325    1221        1104        1669

Bivariate Crosstab: customer_state vs order_status
order_status    canceled  delivered
customer_state
MG                     0       5026
PR                     0       2142
RJ                     1       5576
RS                     0       2375
SP                     0      17799
```

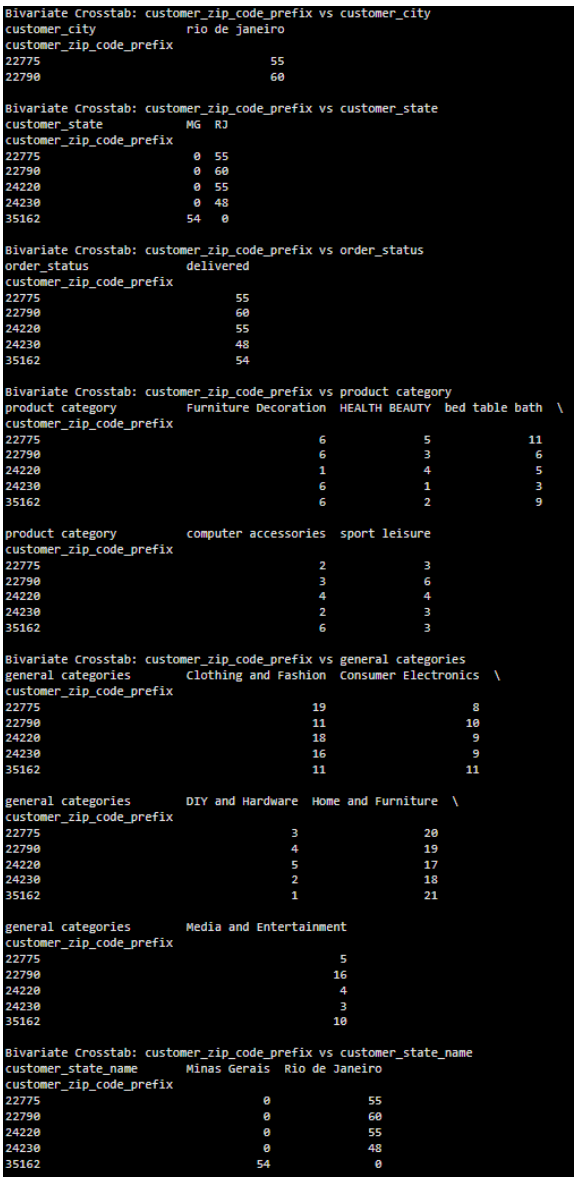
Figure 5. Bivariate Analysis Categorical 3

```
customer_state_name      Rio de Janeiro  Sao Paulo
general categories
Clothing and Fashion               1684       5733
Consumer Electronics               1008       3194
DIY and Hardware                    330        925
Home and Furniture                 1578       5000
Media and Entertainment             810       2345

Bivariate Crosstab: general categories vs delivery_status
delivery_status          Delayed  On Time
general categories
Clothing and Fashion        1138    12595
Consumer Electronics         671     7502
DIY and Hardware             208     2154
Home and Furniture           953    10186
Media and Entertainment      435     5358

Bivariate Crosstab: general categories vs payment_type
payment_type              UPI  credit_card  debit_card  voucher
general categories
Clothing and Fashion     2575        10615         209      334
Consumer Electronics     1883         5944         153      194
DIY and Hardware          543         1730          39       50
Home and Furniture       2092         8583         136      328
Media and Entertainment  1149         4406          84      154

Bivariate Crosstab: general categories vs binned_product_volume
binned_product_volume    Extra Large  Large  Medium  Small  Very Small
general categories
Clothing and Fashion            1457   1928    3079   4058        3211
Consumer Electronics             822    723    1519   2066        3043
DIY and Hardware                 548    732     640    198         244
Home and Furniture              3900   3163    2102   1218         756
Media and Entertainment         1557   1561     885    701        1089

Bivariate Crosstab: general categories vs binned_product_weight
binned_product_weight    Heavy  Light  Medium  Very Heavy  Very Light
general categories
Clothing and Fashion      1913   3724    2680        1423        3993
Consumer Electronics       754   1612     950        1092        3765
DIY and Hardware           951    225     330         667         189
Home and Furniture        3240   1215    2232        3760         692
Media and Entertainment   1465   1005    1145        1212         966

Bivariate Crosstab: customer_state_name vs delivery_status
delivery_status      Delayed  On Time
customer_state_name
Minas Gerais             272     4754
Parana                   116     2026
Rio Grande do Sul        176     2199
Rio de Janeiro           735     4842
Sao Paulo               1088    16710

Bivariate Crosstab: customer_state_name vs payment_type
payment_type          UPI  credit_card  debit_card  voucher
customer_state_name
Minas Gerais         1005         3835          63      123
Parana                486         1560          29       67
Rio Grande do Sul     584         1708          30       53
Rio de Janeiro       1009         4318          79      171
Sao Paulo            3449        13588         332      429

Bivariate Crosstab: customer_state_name vs binned_product_volume
binned_product_volume  Extra Large  Large  Medium  Small  Very Small
customer_state_name
Minas Gerais                  1001    973    1059    981        1012
Parana                         429    398     436    429         450
Rio Grande do Sul              516    522     474    424         439
Rio de Janeiro                1176   1147    1106   1167         981
Sao Paulo                     3611   3404    3558   3699        3526

Bivariate Crosstab: customer_state_name vs binned_product_weight
binned_product_weight  Heavy  Light  Medium  Very Heavy  Very Light
customer_state_name
Minas Gerais            1005    921     899        1049        1152
Parana                   461    401     358         400         522
Rio Grande do Sul        518    407     446         476         528
Rio de Janeiro          1211    996    1004        1195        1171
Sao Paulo               3587   3362    3186        3347        4316
```

Figure 6.Bivariate Analysis Categorical 4



```
Bivariate Crosstab: product category vs general categories
general categories    Clothing and Fashion  Consumer Electronics  \
product category
Furniture Decoration                     0                     0
HEALTH BEAUTY                         3871                     0
bed table bath                           0                     0
computer accessories                     0                  2860
sport leisure                         3339                     0

general categories    Home and Furniture
product category
Furniture Decoration                2796
HEALTH BEAUTY                          0
bed table bath                      4078
computer accessories                   0
sport leisure                          0

Bivariate Crosstab: product category vs customer_state_name
customer_state_name   Minas Gerais  Parana  Rio Grande do Sul  Rio de Janeiro  \
product category
Furniture Decoration           305     188                215             354
HEALTH BEAUTY                  456     170                164             417
bed table bath                 505     168                226             616
computer accessories           388     141                171             360
sport leisure                  376     179                182             399

customer_state_name   Sao Paulo
product category
Furniture Decoration       1168
HEALTH BEAUTY              1648
bed table bath             1897
computer accessories       1150
sport leisure              1446

Bivariate Crosstab: product category vs delivery_status
delivery_status       Delayed  On Time
product category
Furniture Decoration      250     2546
HEALTH BEAUTY             360     3511
bed table bath            366     3712
computer accessories      203     2657
sport leisure             273     3066

Bivariate Crosstab: product category vs payment_type
payment_type          UPI  credit_card  debit_card  voucher
product category
Furniture Decoration  559         2123          29       85
HEALTH BEAUTY         708         3028          66       69
bed table bath        729         3183          42      124
computer accessories  726         1999          63       72
sport leisure         721         2485          58       75

Bivariate Crosstab: product category vs binned_product_volume
binned_product_volume  Extra Large  Large  Medium  Small  Very Small
product category
Furniture Decoration           786    928     646    188         248
HEALTH BEAUTY                  182    525    1060   1417         687
bed table bath                1015   1395     889    601         178
computer accessories            83    200     802    948         827
sport leisure                  560    818     878    696         387

Bivariate Crosstab: product category vs binned_product_weight
binned_product_weight  Heavy  Light  Medium  Very Heavy  Very Light
product category
Furniture Decoration     725    302     688         952         129
HEALTH BEAUTY            712    782     818         225        1334
bed table bath          1697    425     955         826         175
computer accessories     175    537     520         277        1351
sport leisure            773    722     561         559         724

Bivariate Crosstab: general categories vs customer_state_name
customer_state_name      Minas Gerais  Parana  Rio Grande do Sul  \
general categories
Clothing and Fashion             1657     666                674
Consumer Electronics              992     454                466
DIY and Hardware                  312     117                143
Home and Furniture               1281     552                676
Media and Entertainment           649     290                357
```



```
Bivariate Crosstab: customer_zip_code_prefix vs customer_city
customer_city            rio de janeiro
customer_zip_code_prefix
22775                                55
22790                                60

Bivariate Crosstab: customer_zip_code_prefix vs customer_state
customer_state           MG  RJ
customer_zip_code_prefix
22775                     0  55
22790                     0  60
24220                     0  55
24230                     0  48
35162                    54   0

Bivariate Crosstab: customer_zip_code_prefix vs order_status
order_status             delivered
customer_zip_code_prefix
22775                           55
22790                           60
24220                           55
24230                           48
35162                           54

Bivariate Crosstab: customer_zip_code_prefix vs product category
product category          Furniture Decoration  HEALTH BEAUTY  bed table bath  \
customer_zip_code_prefix
22775                                        6              5              11
22790                                        6              3               6
24220                                        1              4               5
24230                                        6              1               3
35162                                        6              2               9

product category          computer accessories  sport leisure
customer_zip_code_prefix
22775                                        2              3
22790                                        3              6
24220                                        4              4
24230                                        2              3
35162                                        6              3

Bivariate Crosstab: customer_zip_code_prefix vs general categories
general categories        Clothing and Fashion  Consumer Electronics  \
customer_zip_code_prefix
22775                                       19                     8
22790                                       11                    10
24220                                       18                     9
24230                                       16                     9
35162                                       11                    11

general categories        DIY and Hardware  Home and Furniture  \
customer_zip_code_prefix
22775                                    3                  20
22790                                    4                  19
24220                                    5                  17
24230                                    2                  18
35162                                    1                  21

general categories        Media and Entertainment
customer_zip_code_prefix
22775                                           5
22790                                          16
24220                                           4
24230                                           3
35162                                          10

Bivariate Crosstab: customer_zip_code_prefix vs customer_state_name
customer_state_name       Minas Gerais  Rio de Janeiro
customer_zip_code_prefix
22775                                0              55
22790                                0              60
24220                                0              55
24230                                0              48
35162                               54               0
```

Figure 7. Bivariate Analysis Categorical 5

Figure 8.Bivariate Analysis Categorical 6

The bivariate analysis reveals clear links between customer location, order status, and product choices. Zip codes map directly to cities and states, and most orders are delivered on time, though some areas, especially larger cities like São Paulo and Rio de Janeiro, show slightly higher delays. Customer location also influences payment methods, with credit cards being the dominant option. Product volume and weight vary by region, with larger, heavier items mostly in furniture and home-related categories, and smaller, lighter items in health, beauty, and electronics.

Product and general categories vary across cities and states, reflecting regional preferences. “Clothing and Fashion” and “Home and Furniture” dominate in São Paulo, while Rio de Janeiro and Minas Gerais show higher demand for “Consumer Electronics” and “Media and Entertainment.” Certain product categories, like “bed table bath” and “HEALTH BEAUTY,” experience slightly more delays, but most deliveries are on time. Payment type, product size,

and weight correlate with location, highlighting how customer city and state strongly influence purchasing patterns, delivery outcomes, and order composition.

5.3 Clustering

The objective was to identify data in the dataset that could be clustered to discover structure in the data and for data exploration and visualization. This segmentation intends to cluster customers based on their geographical locations to identify potential groupings to both discover patterns and for potential optimization.

The data was prepared by focusing on customer's zip code location its associated geolocation data, then standardising it so both latitude and longitude are on the same scale for use with Euclidean distance)

The methodology of clustering was determined by running three separate models, KMeans, GMM and Agglomerative to identify the most accurate model to be used, with KMeans giving the highest silhouette score of 0.4968, and being a model that minimises cluster variances, it was thus chosen. KMeans was also chosen due to its efficiency with large datasets and its ability to interpret centroids easily.

Model	Silhouette Score
KMeans	0.4968
GMM	0.3558
Agglomerative	0.4624

Table 6. Model Comparisons

Using KMeans, a Matplotlib scatter plot was created to determine cluster spread and boundaries.

Cluster	Avg Latitude	Avg Longitude	Latitude Range	Longitude Range
0	-19.02	-49.40	-23.43 to -10.83	-59.96 to -44.19
1	-9.22	-38.16	-18.07 to 42.18	-45.17 to -8.58
2	-22.61	-45.30	-25.01 to -14.27	-48.59 to -39.74
3	-27.24	-50.97	-36.61 to -21.13	-64.28 to -48.33
4	-5.04	-51.59	-14.54 to 3.84	-72.67 to -44.17

Table 7. KMeans Clustering Results

These clusters give us some insights on potential segmentation of the customer base, it is inferred that;

Cluster 0 (Blue) - Covers the central to western region of Brazil, likely spanning metropolitan areas while also extending into inland regions, potentially with slightly longer delivery times due to distances from major ports and urban postal hubs.

Cluster 1 (Purple) - Covers the north eastern coastal strip of Brazil, likely including more densely populated urban hubs and coastal regions which show medium to high density. This cluster does show extreme far north to north east points which could signal data anomalies and potential outliers.

Cluster 2 (Orange) - South of cluster 1, it seems to have a high density, possibly representing customers with mid to high purchasing power with better logistical connections, and inferred to be the main cities of Sao Paulo and Rio De Janeiro.

Cluster 3 (Green) - Further south of Brazil, possibly interpreted to be more price sensitive with more seasonal shopping rather than consistent purchases. This could be inferred to a more distinct southern customer base with different buying habits.

Cluster 4 (Red) - The most scattered of the clusters going westward inland. These customer segmentation is inferred to be in more remote or semi-urban areas, potentially facing significantly longer delivery times and higher freight costs due to poorer infrastructure and lower amounts of logistical partners.

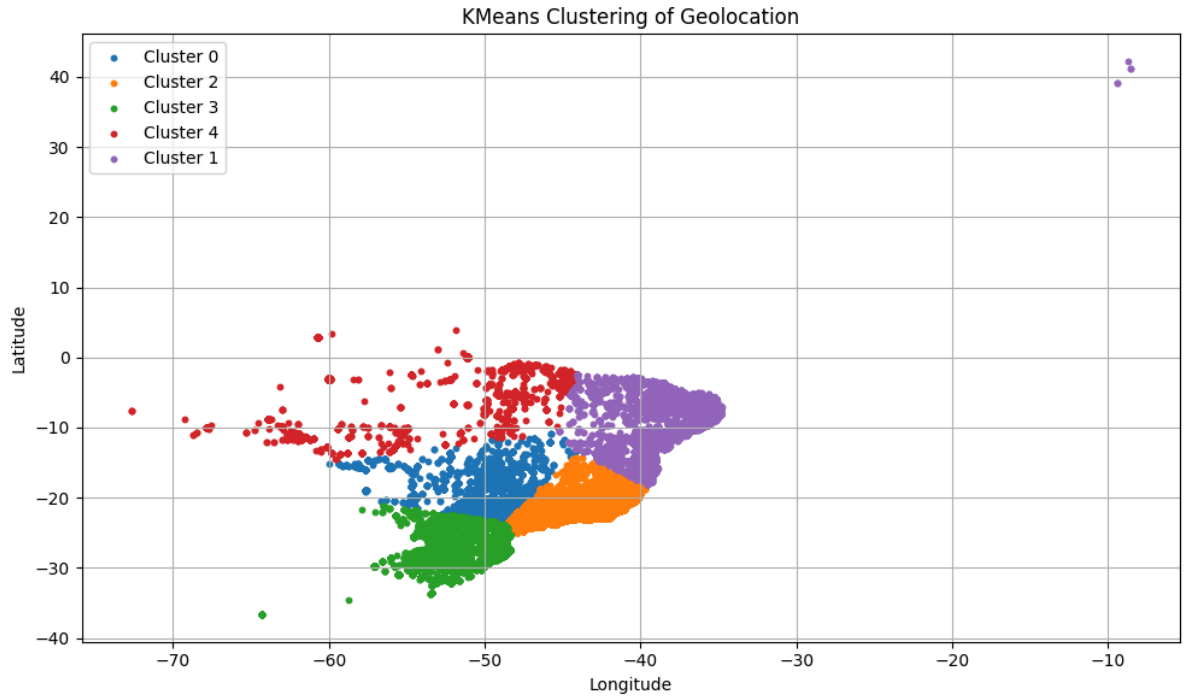


Figure 9. KMeans Scatterplot Clustering

With the customer geolocations segmented into 5 clusters and the high customer concentration determined, mostly in cluster 1 and 2, while more sparse concentrations in clusters 0, 3 and 4, some potential optimization could be proposed.

Logistical Optimization

Given that currently, Olist as a platform not just provides e-commerce services, but also manages shipping through their integration with various carriers, a potential avenue for optimization could be the identification of strategic areas within each cluster for a regional hub that could increase logistical efficiency and reduce transit and delivery times.

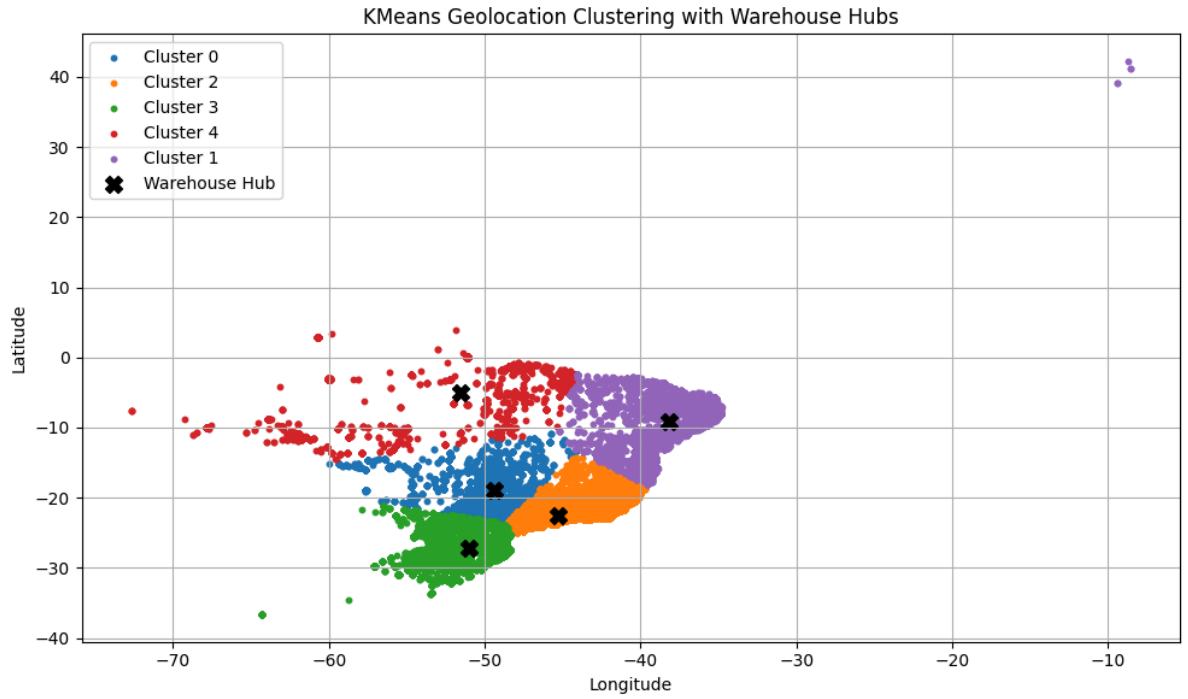


Figure 10. KMeans Scatterplot Clustering with Centroid Overlay

The previous scatter plot of customer locations, showed only their locations, giving rise to the potential for logistical optimisation. By calculating the centroid of each cluster (represented by the ‘x’), it represents potential warehouse and delivery hub locations that could increase logistical efficiency for that cluster.

Applying the KMeans model works by introducing temporary centroids and assigning each customer to the nearest centroid and recalculating them as the average location of all customers in the cluster until it converges. The central idea by placing these hubs near to each centroid will minimize average travel distance to customers within the region, increasing logistical efficiency and reducing freight time.

The final image shows an overlay of the cluster and suggested regional cluster hubs overlaid onto the map of Brazil. Although these centroids are not placed in the areas of highest customer density, this strategy is intended to minimize the total travel distances for the entire cluster, not just the most concentrated parts.

This strategy aims to balance service across the entire cluster, where in instances of high density cities like Sao Paulo or Rio De Janeiro, which would concentrate customers in one direction, using the KMeans model finds the centroids that reduce average distance to all customers. This may shift the hub away closer to the cluster's geometric center, away from high density cities but still closer to all customers.

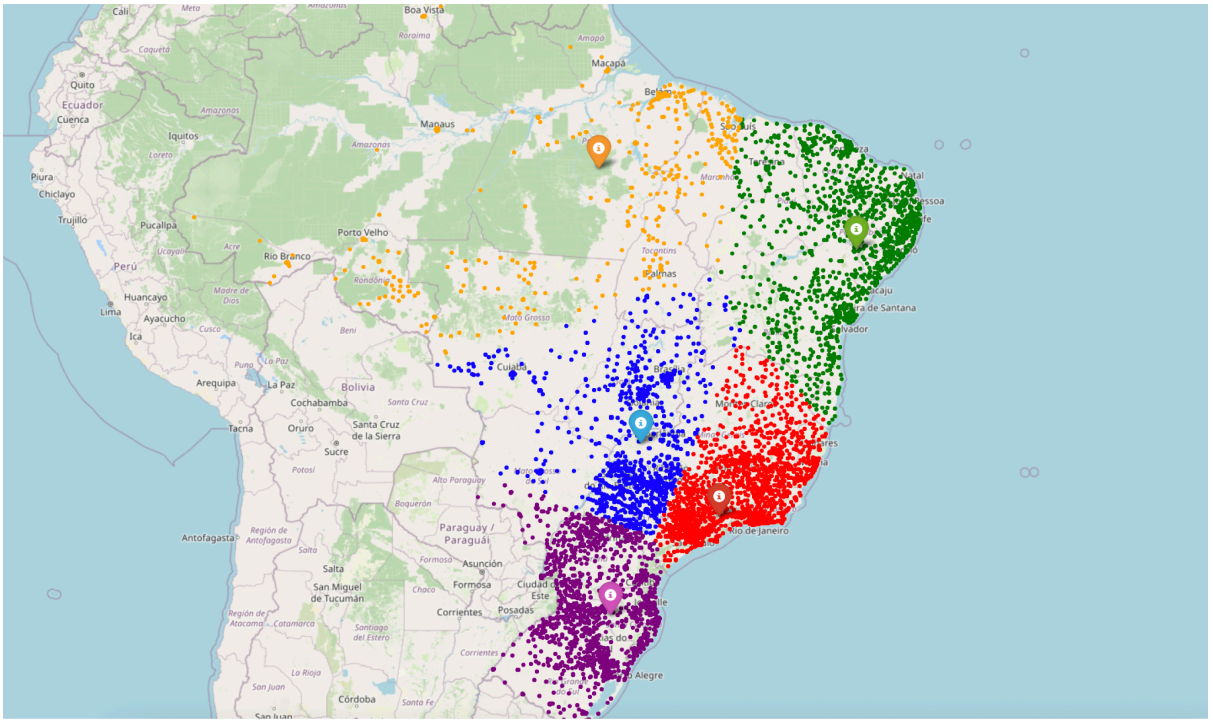


Figure 11. Cluster Distribution with Logistic Hub Optimization with Map Overlay

https://imailsunwayedu-my.sharepoint.com/:u:/r/personal/25083247_imail_sunway_edu_my/Documents/Sunway%20University/Block%201/BAA5043%20Business%20Intelligence/Assignment/Olist/Resources/geolocation_clusters_overlay_map.html?csf=1&web=1&e=jYnbzd

This strategy works especially well in cluster 0 and 4, where it improves connectivity to customers in more remote areas, also reducing long distance deliveries. Further potential logistical benefits of this strategy is the balanced workload across all delivery routes, lowering fuel costs from the overall reduced travel mileage while concurrently improving regional coverage and improved delivery reliability.

5.4 Sentiment Analysis

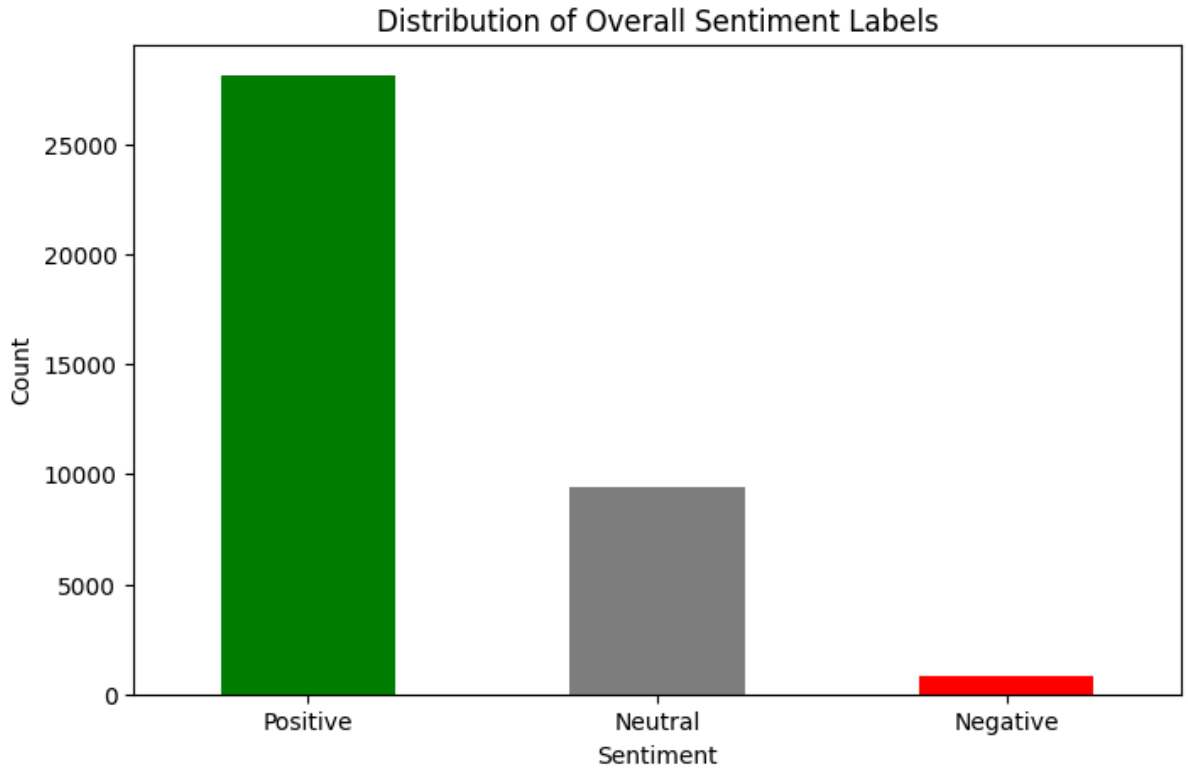


Figure 12. Bar Graph of Distribution of Sentiment Labels

Out of 94,842 orders, 38,391 customers left reviews. Among these, 73.26% were positive, 24.59% neutral, and 2.15% negative. This shows most customers were satisfied, but understanding what affects their feelings is important.

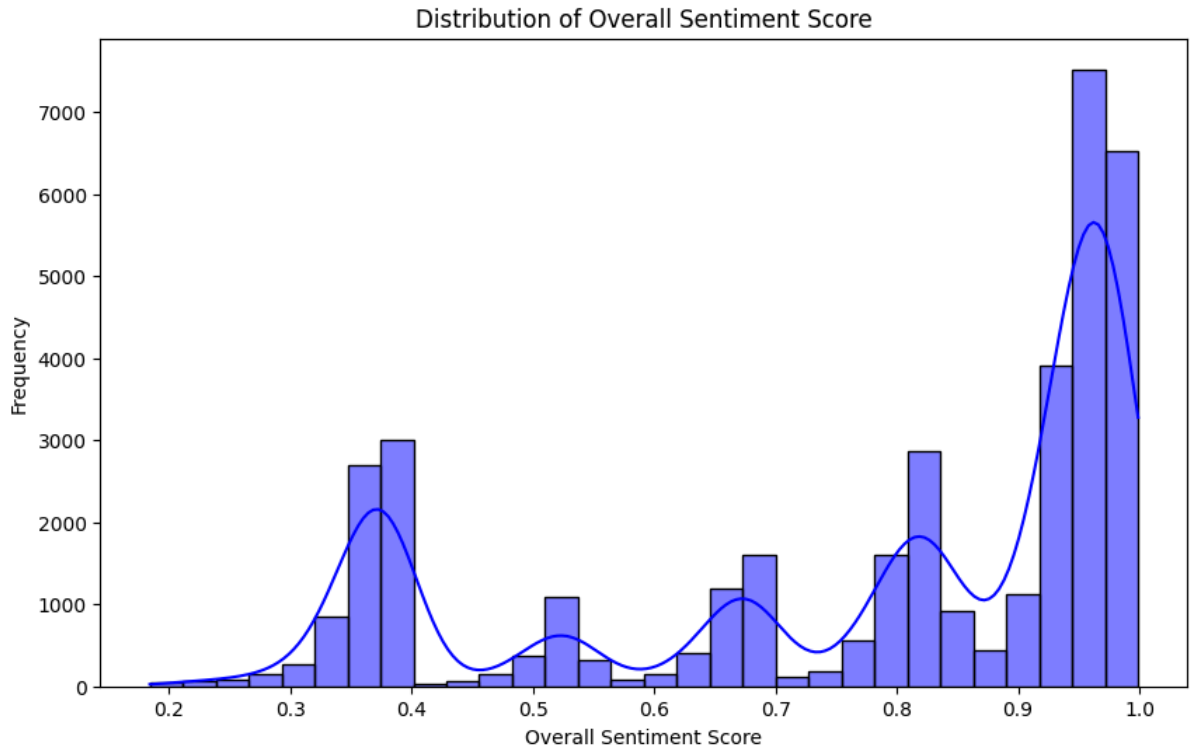


Figure 13.Distribution of Sentiment Score

The overall sentiment score has an average (mean) value of approximately 0.77 with a standard deviation of about 0.23, indicating moderate variability around the mean. The scores range from a minimum of roughly 0.18 to a maximum close to 1.00. The 25th percentile is around 0.64, the median (50th percentile) is about 0.88, and the 75th percentile reaches approximately 0.96, showing that most scores are skewed towards the higher end of the scale. Each spike of the overall sentiment score reflects a review score of 1,2,3,4 and 5.

This analysis assessed relationships between predictors and sentiment outcomes.
Numerical features were correlated with overall_sentiment_score using Pearson and Spearman correlations, with coefficients and p-values reported and ranked by absolute Pearson correlation.

Categorical features were compared with overall_sentiment_label using Chi-square tests and Cramér's V, then ranked by association strength.
Results highlight variables most strongly related to sentiment for further modeling or interpretation.

Numerical Variable	**Pearson Corr**	**Pearson p-value**	**Spearman Corr**	**Spearman p-value**
actual_delivery_days	-0.329	0.00E+00	-0.231	0.00E+00

payment_value	-0.056	7.80E-28	-0.064	9.77E-36
geolocation_lng	-0.042	3.06E-16	-0.042	1.99E-16
order_year	-0.039	1.23E-14	-0.036	1.10E-12
order_month	0.039	1.30E-14	0.043	2.97E-17
freight_value	-0.033	1.64E-10	-0.04	5.64E-15
product_weight_g	-0.032	2.61E-10	-0.006	2.63E-01
product_volume_cm3	-0.03	6.74E-09	-0.011	2.43E-02
product_length_cm	-0.024	2.31E-06	-0.01	6.21E-02
product_photos_qty	0.024	3.76E-06	0.024	3.41E-06
product_height_cm	-0.023	4.47E-06	-0.009	7.26E-02
payment_installments	-0.023	4.65E-06	-0.01	5.98E-02
geolocation_lat	-0.018	3.13E-04	-0.014	6.48E-03
product_width_cm	-0.01	5.23E-02	-0.002	6.42E-01
price	-0.007	1.85E-01	0	9.35E-01
product_description_length	0.007	1.95E-01	0.004	4.44E-01
product_name_length	-0.006	2.18E-01	-0.005	3.06E-01
payment_sequential	0	9.23E-01	-0.004	4.89E-01

Table 8.Pearson Corr and Spearman Corr for Numerical

Categorical Variable	**Chi-square**	**p-value**	**Cramér's V**
delivery_status	4498.871	0.00E+00	0.342
product category	506.444	2.16E-42	0.069
customer_zip_code_prefix	23730.327	7.37E-02	0.064
customer_state	305.212	1.58E-37	0.057
customer_state_name	305.212	1.58E-37	0.057

general categories	126.58	2.99E-21	0.039
binned_product_weight	39.916	3.32E-06	0.02
binned_product_volume	27.703	5.34E-04	0.016
order_status	8.776	1.24E-02	0.013
payment_type	7.789	2.54E-01	0.005
customer_city	6197.892	7.45E-01	0

Table 9.Chi-square and Cramér's V for Categorical

From this top 10 factors that were affecting the sentiments were:

Rank	**Factor**	**Type**	**Measure**	**Comment**
1	actual_delivery_days	Numerical	Pearson corr = -0.329, $p < 0.001$	Delivery timing is crucial
2	delivery_status	Categorical	Cramér's V = 0.342	Delivery status strongly affects sentiment
3	payment_value	Numerical	Pearson corr = -0.056, $p < 0.001$	Amount paid impacts sentiment
4	product category	Categorical	Cramér's V = 0.069	Product type moderately affects sentiment
5	geolocation_lng	Numerical	Pearson corr = -0.042, $p < 0.001$	Location longitude influences sentiment
6	order_month	Numerical	Pearson corr = 0.039, $p < 0.001$	Month of order influences sentiment
7	freight_value	Numerical	Pearson corr = -0.033, $p < 0.001$	Shipping cost influences sentiment
8	product_weight	Numerical	Pearson corr = -0.032, $p < 0.001$	Heavier products show slightly worse sentiment scores
9	product_volume_cm3	Numerical	Pearson corr = -0.030, $p < 0.001$	Larger products tend to face more delivery challenges.
10	customer_state	Categorical	Cramér’s V = 0.057	Sentiment varies slightly across regions.

Table 10. Rank of Factors Affecting the Sentiments

Delivery speed has the biggest impact on how customers feel, with actual delivery days being the most important factor. How much a customer pays also affects their satisfaction. Where a customer is located, especially their longitude, plays a role in their experience. The year and month of the order show that timing and season can change how customers feel, possibly

because of service changes or demand shifts. The size and weight of products matter too, as bigger items might face more delivery problems. Shipping costs also influence customer opinions.

Among categories, delivery status is the strongest factor. Whether an order arrives on time or is delayed greatly affects sentiment. The type of product also matters since some categories get better reviews. Customer location, such as zip code and state, shows regional differences in satisfaction. Other factors like product weight and volume have smaller but clear effects on how customers feel.

5.5 Sales Forecasting Modelling

In e-commerce, sales performance is the most important information that sellers need to adjust their sales strategy and consistently optimise their efforts. In this study, monthly sales was selected as the target variable to be forecast using different data mining models.

To get the monthly sales from the original dataset, a new data set was purposely developed for sales forecasting modelling. Feature engineering was required to develop this new dataset by extracting year and month from the user order purchase date in the original dataset. The unique year and month combination were then used to create new features, including monthly sales, average product rating, average delivery rating, average customer satisfaction, average delivery days, average product weight, average product volume, average product photos quantity, average product description length, average product name length, and average advertising spending. These newly derived features represent monthly values corresponding to each distinct year and month combination.

	totalsales
totalsales	1
order_year	0.847355833
order_month	-0.07996489
product_rating	-0.39164778
delivery_rating	-0.61861855
customer_satisfaction	-0.51318295
actual_delivery_days	0.031366041
product_weight_g	0.195436921
product_volume_cm3	0.072860441
product_photos_qty	0.307491227
product_description_length	-0.42736951
product_name_length	-0.31410685
ad_spend	0.525263335

Figure 14: Correlation with sales as target variable

Correlations were performed using excel with monthly total sales as the target variable. Figure 14 above shows 5 variables out of 12 were having high correlation with the monthly total sales and will be selected as the good predictor variable to use in the modelling.
From the related work, a few models were selected based on their popularity in past research. These models are Random Forest, Multi-Layer Perceptron(MLP) , k-nearest neighbors(KNN), and Stacking which are frequently recommended in the literature for sales forecasting purposes.

Weka was selected to perform modelling of sales forecasting due to its simplicity. All 3 models were trained and evaluated in the Weka explorer with 10 fold cross validation.

Model	RRSE	MAE	R-squared
MLP	0.43	0.10	0.81
Random Forest	0.46	0.12	0.79
KNN	0.47	0.11	0.78
Stacking	0.45	0.12	0.80

Table 11: Result of Models Performance Metrics

From the Table 11 above, an illustration of performance comparison between 4 models was shown based on three evaluation metrics which are Root Relative Squared Error (RRSE), Mean Absolute Error (MAE), and R-squared. The high R-square value indicates that the model is able to explain a larger variance in the target variable, ensuring strong prediction. From the result, MLP stands out as the best model for e-commerce sales forecasting as it has the highest R-square value while maintaining low prediction error. For Random Forest and Stacking they both perform competitively, but having higher RRSE and MAE compared to MLP, suggesting marginally less accurate predictions. KNN has the worst performance with the lowest R-squared value among the models, indicating inability of capturing underlying patterns of the data.

Order Year	Delivery Rating	Customer Satisfaction	Description Length	Advertising Spend	Forecasted Sales
Sept 2018	3.8	3.5	900	600	852535.3

Table 12: Forecasted Sales in September 2018

With the use of trained MLP model, the sales in September 2018 were forecasted to be R$ 852,535.30. This result is based on the features that represent monthly performance and investment which are: delivery rating, customer satisfaction, product description length and advertising spending. By experimenting with different values of predictor features, customer satisfaction and advertising spend shows high influence with forecasting of monthly sales. Both customer satisfaction and advertising spend are having strong directly proportional relationships with the forecasted sales. Thus sellers should focus more on customer satisfaction and advertising spend if they wish to improve monthly sales.

Overall, MLP achieved high performance in predicting sales due to its ability to capture and learn non linear relationships. Capturing non linear patterns is important when it comes to sales data, due to the sales dataset having a lot of features that do not have purely linear relationship with sales. In addition, MLP is able to handle high dimension and correlated data well due to its internal weight adjustment. With the iterative weight adjustment during the model training, MLP learns to assign appropriate weight for each predictor feature reducing the negative impact of multicollinearity. This weight adjusting feature in MLP allows the model to joint effect of correlated features, making the model more generalised.

In closing, the model has demonstrated reliable short-term predictive capabilities, however, with data only available up to August 2018, forecasting further into Q4 could risk introducing uncertainty and reduce reliability of the results, especially given Brazil's year end peak sales season. Therefore, it would be more prudent to limit forecasting only into September 2018 until which time more complete datasets are available for longer term projections.

5.6 E-commerce Dashboard

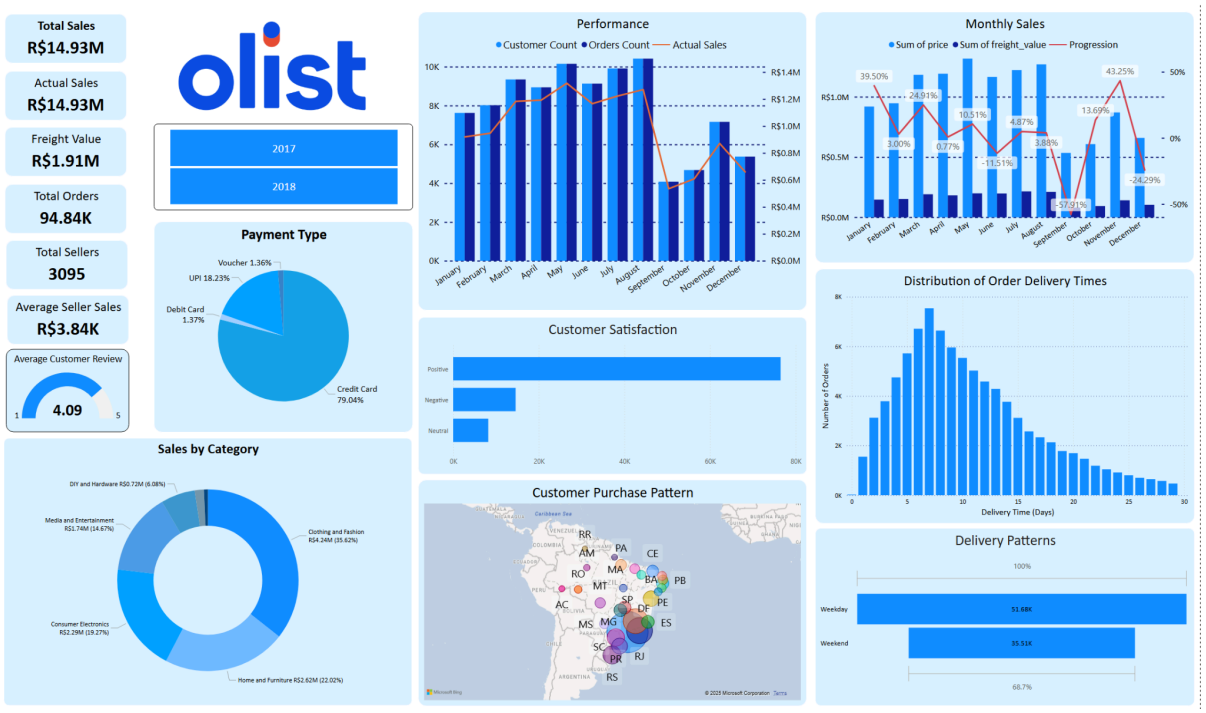


Figure 15. Dashboard Overview for 2017 and 2018

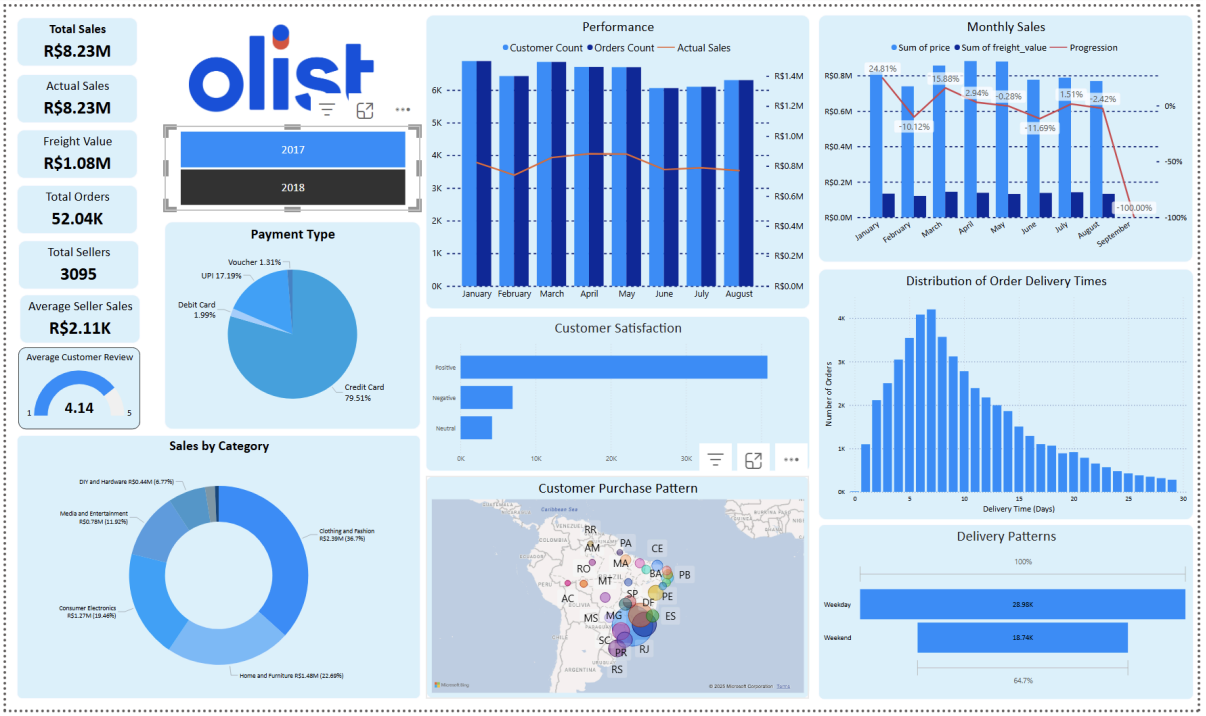


Figure 16. Dashboard Overview for 2017

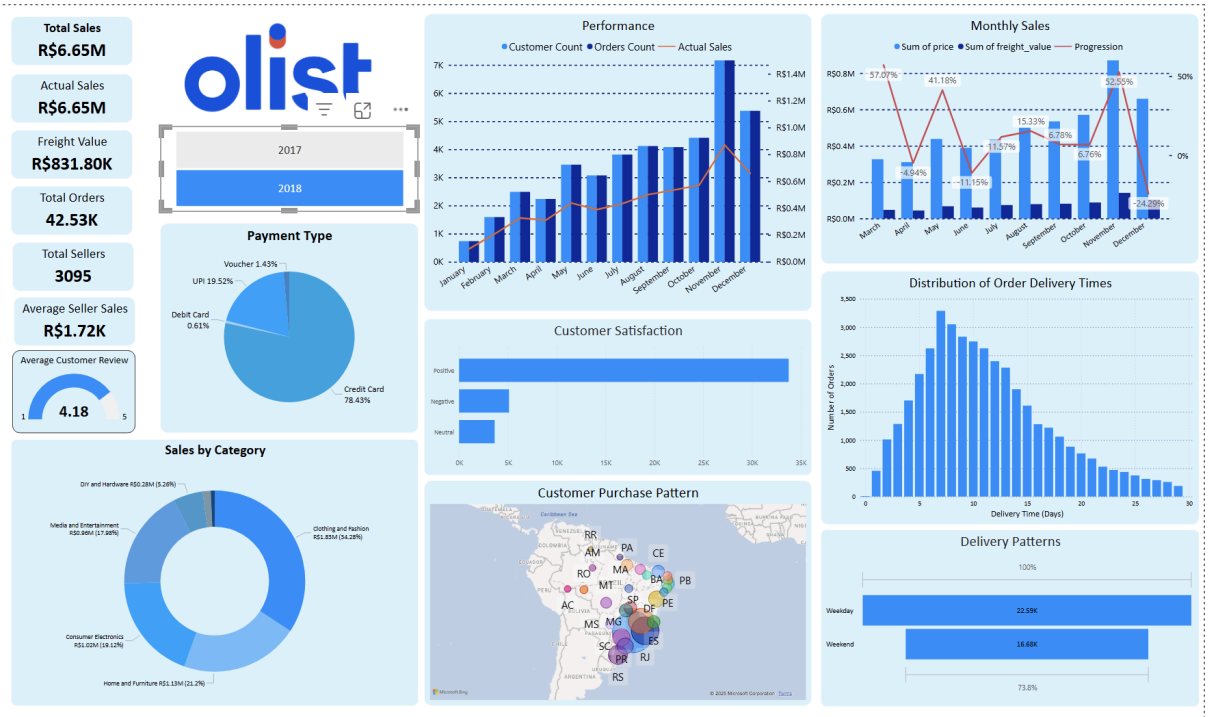


Figure 17. Dashboard Overview for 2018

The developed Power BI dashboard provides a consolidated view of Olist's business performance, allowing for data driven analysis and tracking of business performance and an exploration of future strategic planning.

Overall business performance metrics show that for the period of 2017 - 2018, Olist generated R$ 14.93 million in revenue, R$ 11.89 million of which from sales and an additional R$ 1.91 million from shipping, this across 94.84K orders, with 3095 total sellers bringing in an average of R$ 3.84K. The average customer review score of 4.09 indicates an overall high level of customer satisfaction, but does leave room for improvement with still 15% of negative reviews.

The dashboard shows an average overview of the 2017 and partial 2018 period (ending August 2018). Given that these years are Olist's yearly years, with the company only being founded in 2015, performance trends reflect early stage growth and establishment of market presence and refinement of their commercial strategy. An important point to note is that as mentioned, data is only available until August 2018, meaning incomplete Q3 and Q4. This limits direct year on year comparisons on trends, particularly sales and order volumes. Additionally, the insufficient data also prevents tracking peak retail periods in Q4 for Brazil as present in 2017 data.

What remains consistent across both years however is the domination of credit cards as the preferred payment method (~78%), potentially indicating preference for deferred payments. Clothing & Fashion, Home & Furniture and Consumer Electronics remained top performing categories while Media & Entertainment saw an increase in performance throughout 2018.

The average customer review score saw a slight decrease YOY, from 4.18 to 4.14, possibly indicating market share growth and the need for optimization of their logistical processes, while delivery time distributions remained similar with deliveries completed within 5 - 10 days, with 2018 seeing a 9.1% shift from weekday to weekend deliveries. The southeast region showed concentrated sales throughout, reflecting Brazil's national capital and highly dense urban areas and the country's economic centers.

This dashboard provides an overview of Olist's startup stage and initial growth phase. The observed performance patterns seem to indicate early scaling trajectory, although data limitations make it hard for us to accurately determine market share growth but it is highly inferred based on a ~12% growth in both total sales and order volume for both years between Jan - Aug. Limitations in available data also do not allow us to track seller base growth.

In summary the dashboard shows overall growth in Olist's business performance, with peaks correlated to Brazil's retail calendar. However, incomplete Q4 2018 data limits a full year on year comparison but overall trends point strongly towards business expansion and steady growth.

6.0 Conclusion and Discussions

This paper used Olist's e-commerce dataset to show how Business Intelligence can transform operational and sales data into objective and actionable insights. With initial fragmented datasets and vague performance visibility, the analysis combined exploratory data analysis, clustering, sentiment analysis, sales forecasting and a Power BI dashboard to create an overview of business performance and metrics.

This revealed several key indicators, from consistent seasonal sales peaks in line with Brazil's retail calendar, strong reliance on consumer preferences for deferred payments and top performing product categories with strongest and consistent consumer demand. It also highlighted potential avenues for improvement and optimisation, such as delivery speed, more accurate delivery estimations and logistical optimisation.

While data limitations existed, with data only spanning up to August 2018, leaving out Q4, Brazil's busiest retail season, the Jan - Aug data shows promising growth YOY of about 12%, suggesting Olist's steady market share and sales growth. Furthermore, the MLP model was able to forecast sales for September at $852,535.30 with an r2 value of 0.81, but it was not deemed reliable to forecast for Q4, due to insufficient overall data. The analysis helps turn data and trends into actionable insights that help frame strategic decisions, operation optimization and inventory planning, while adding discovered opportunities in regional clustering.

Moving forward, with complete multi-year data, inclusion of greater and broader market information, as well as real time performance updates would allow for a stronger and more reliable model. Overall, despite the data gaps and its limitations, this paper shows that Olist is going through its growth phase with steady expansion and market penetration and how Business Intelligence can be a key driver in guiding their strategic planning and ongoing expansion.

7.0 References

Singh, S. N. (2016). E-commerce : Role of e-commerce in today's business. Computing Trendz - the Journal of Emerging Trends in Information Technology, 6(1). https://doi.org/10.21844/cttjetit.v6i1.6699

Vadwala, A. Y., & Vadwala, M. S. (2017). E-Commerce: Merits and Demerits A Review Paper. International Journal of Trend in Scientific Research and Development, Volume-1(Issue-4). https://doi.org/10.31142/ijtsrd106

Bilgic, E., & Duan, Y. (2019). E-commerce and Business Analytics: A Literature review. Lecture Notes in Business Information Processing, 173–182. https://doi.org/10.1007/978-3-030-30874-2_13

Pandey, T. N., Vasudev, A., Sagayanathan, D., Anjan, G., Arshad, D., & Patra, S. S. (2023, November). Predicting customer satisfaction in brazil e-commerce: A comparative study of machine learning techniques. In 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 505-510). IEEE. https://ieeexplore.ieee.org/abstract/document/10425505

Moussas, K., Hafiane, J., & Achaba, A. (2023). Business intelligence and its pivotal role in organizational performance: An exhaustive literature review. *Management Sciences, LARTI2D Laboratory, ENCG, Ibn Zohr University*. https://www.researchgate.net/publication/382250264_Business_intelligence_and_its_pivotal_role_in_organizational_performance_An_exhaustive_literature_review

Chate, P. A. (2022). *Behavioral modelling of customer marketing patterns and review prediction using machine learning techniques* (MSc Research Project). National College of Ireland. NORMA Repository. https://norma.ncirl.ie/6100/

Anitha, M. A., & Sherly, K. K. (2025). Churn prediction with GraphSAGE model based on the derived features using RFM and sentiment analysis. *Journal of the Chinese Institute of Engineers*, 1–13. https://www.tandfonline.com/doi/abs/10.1080/02533839.2025.2478185

Widjaja, A. A., Ghapanchi, A. H., & Bingley, S. (2025). Exploring the Antecedents to the Effective use of Business intelligence: A Systematic Review approach. *Information Systems*

Management, 1–21.
https://www.tandfonline.com/doi/full/10.1080/10580530.2025.2479737?af=R#infos-holder

Xu, Z., Wang, X., Tan, Y., & Li, X. (2023). Data-Driven Analysis for the Operation Status of the E-Commerce Platform based on Offlist.
https://www.scitepress.org/Papers/2022/118366/118366.pdf

Alzami, F., Sambasri, F. D., Nabila, M., Megantara, R. A., Akrom, A., Pramunendar, R. A., Prabowo, D. P., & Sulistiyawati, P. (2023). Implementation of RFM Method and K-Means Algorithm for Customer Segmentation in E-Commerce with Streamlit. *ILKOM Jurnal Ilmiah*, *15*(1), 32–44.
https://www.academia.edu/download/109391985/pdf.pdf