**Technical Report on Olist Stores**

**A report on our insights from: Feedback Sentiment Analysis,**

**Clustering, Sales Prediction, Sale Stores vs Location,**

**Delivery Time Performance**

**1. Introduction**

This report presents an analysis of the Olist Store, Brazil's largest online retailer, using a dataset of 100,000 orders from various marketplaces. The dataset provides insights into order status, pricing, payments, freight, customer locations, product attributes, and reviews, along with geolocation data mapping Brazilian zip codes. Our analysis focuses on four key areas: feedback sentiment analysis to gauge customer review polarity, sales prediction based on historical data, location-based sales performance to identify high-sales regions, and delivery time performance to assess logistical efficiency.

**2. Methodology**

**Data Source**

The primary data source that we used was the dataset from the “Brazilian E-Commerce Public Dataset by Olist.” This included 9 datasets displaying information about various parts of the Olist Store.

**Data Preparation**

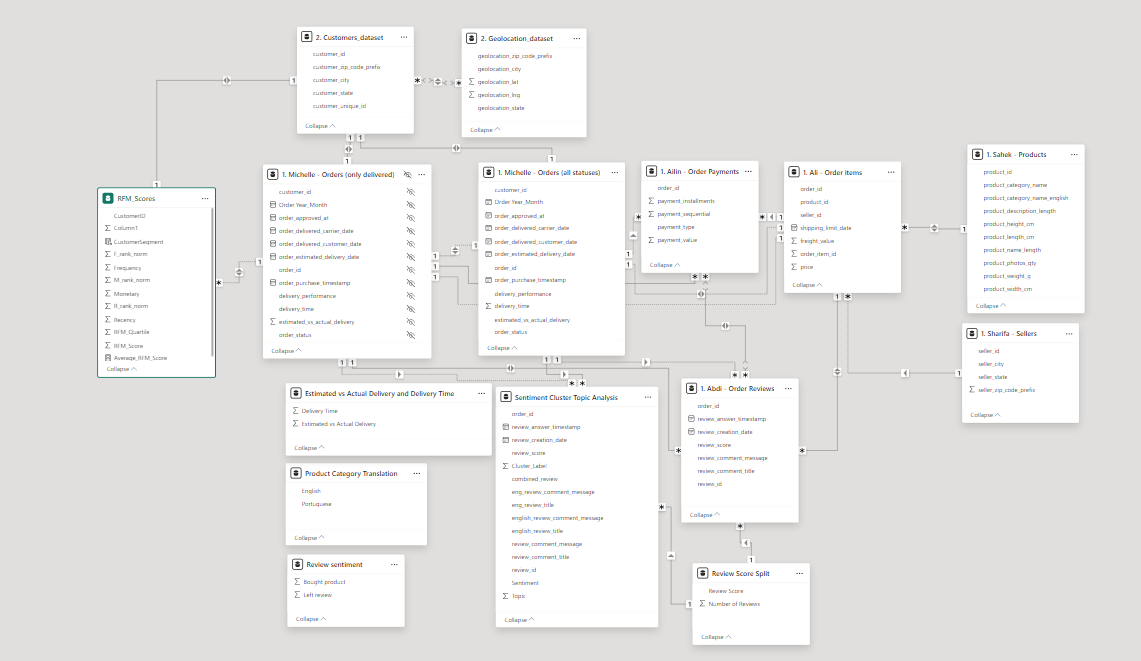
To prepare the data for analysis, we began by examining and establishing relationships between the various datasets. This involved performing joins to understand how different tables could be connected, ensuring the data was accurately linked and ready for comprehensive analysis.

**Data Cleaning**

The data preparation process also included a thorough cleaning phase. We translated text from Portuguese to English, corrected spelling errors, removed unnecessary information and addressed missing values by cleaning blank fields. These steps ensured that the dataset was accurate, consistent, and ready for further analysis.

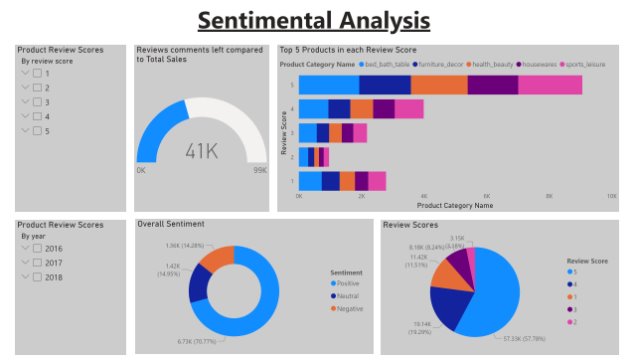
For the ***order\_payment***dataset, we noticed that one type of payment was labelled in Portuguese. To clean this data, we used Excel to replace all Portuguese terms with their English equivalents. Other than that, there are no issues with this dataset.

**Data Model Layout**



**3. Analysis and Visual Aids**

**Sentiment Analysis**: The sentiment of each customer review message was analysed using a sentiment analysis model. Details are mentioned on Section 4. Results – 4.2 Clustering.



The donut chart highlighted where the overall sentiment is positive.

A clustered bar chart is used the illustrate the top 5 products in each review score.

The bed bath is represented by the blue colour, and we can see it has the highest positive and negative reviews.

**Additional Insight: Sample Analysis of Customer Reviews**

1. **Overview of Analysed Reviews**
   1. **Total Reviews Analysed:** 184
   2. **Language:** Portuguese and English
2. **Sentiment Breakdown**

|  |  |  |
| --- | --- | --- |
| **Sentiment** | **Count** | **Percentage** |
| Positive | 81 | 44% |
| Neutral | 42 | 23% |
| Negative | 61 | 33% |
|  |  |  |

1. **Example Reviews**
   1. **Positive Reviews:**
      1. *Portuguese:* "Amei o produto, super fácil de manusear e chegou bem antes do prazo!"

*English:* "Loved the product, super easy to handle and arrived well before the deadline!"

* + 1. *Portuguese:* "Entrega perfeita, até antes do prazo. Produto em perfeito estado e de ótima qualidade."

*English:* "Perfect delivery, even ahead of schedule. Product in perfect condition and of great quality."

* 1. **Neutral Reviews:**
     1. *Portuguese:* "Produto chegou conforme anunciado. Gostei."

*English:* "Product arrived as advertised. I liked it."

* + 1. *Portuguese:* "Demorou quase trinta dias. Tempo demais."

*English:* "It took almost thirty days. Way too long."

* 1. **Negative Reviews:**
     1. *Portuguese:* "Recebi apenas 1 detector de metal e a compra foi de 2 unidades. Como podemos resolver isso?"

*English:* "I received only 1 metal detector, but I purchased 2 units. How can we resolve this?"

* + 1. *Portuguese:* "O produto veio fora da especificação. Não gostei."

*English:* "The product came out of specification. I didn't like it."

This analysis provides a granular view of customer feedback within a small sample, emphasizing diverse sentiment distribution and specific customer experiences.

**4. Results**

The sentiment analysis reveals that 70% of customer feedback demonstrates positivity. Both the Bed Bath Table and Sports Leisure categories exhibit the highest average ratings. However, there are notable occurrences of negative and neutral reviews, indicating opportunities for improvement. By leveraging insights from this analysis, we can formulate recommendations aimed at enhancing customer satisfaction. Addressing these concerns has the potential not only to improve customer experience but also to drive increased sales.

This structured approach ensures that actionable insights are derived from the data, enabling strategic improvements that align with customer expectations and business objectives.

**4.1 Feedback Sentiment Analysis**

The Bed Bath Table and Sports Leisure categories have the highest average ratings. However, the Bed Bath Table exhibits significant polarity, characterized by a bimodal distribution of ratings, with a notable prevalence of both 1-star and 5-star reviews, indicating divergent customer experiences.

The objective of this analysis is to identify patterns in customer reviews to understand the factors contributing to customer satisfaction and dissatisfaction. Specifically, we aim to:

* Analyse customer sentiments (positive, neutral, negative) expressed in reviews.
* Correlate customer sentiments with review scores to uncover potential discrepancies or insights.
* Highlight themes and topics within customer feedback to identify key drivers of satisfaction or dissatisfaction.

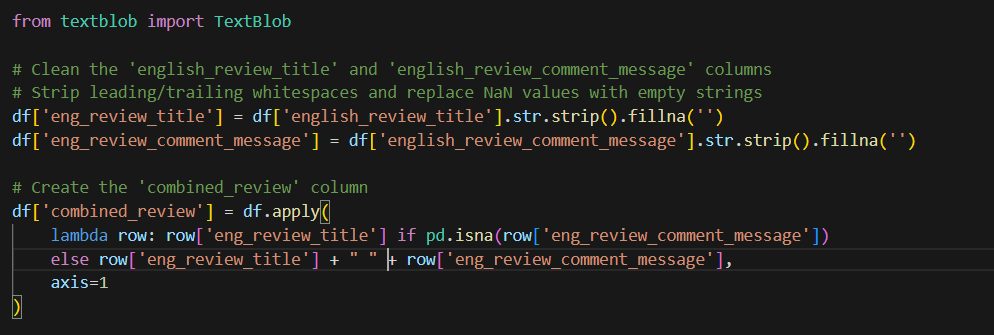
**Preprocessing and Dataset Transformation**

The analysis began with the **olist\_order\_reviews\_dataset.csv**, which contained raw customer reviews. The dataset underwent several preprocessing and transformation steps to ensure it was clean, consistent, and ready for analysis:

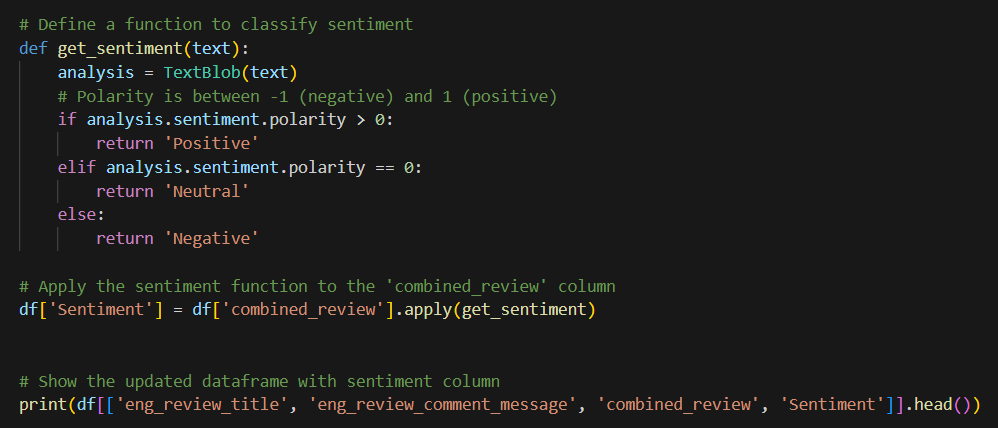
1. **Filtering Missing and Short Data**:

* Rows with empty values in the columns ‘review\_comment\_message’ and ‘review\_comment\_title’ were removed to focus only on reviews with substantive content.
* Additionally, reviews with a combined character length of fewer than three were excluded, as such entries often lack meaningful information.

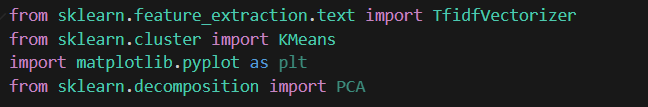
1. **Translation**: The columns ‘review\_comment\_title’ and ‘review\_comment\_message’, originally in Portuguese, were translated into English using Google Translate function within Google Sheets. Two new columns, english\_review\_title and english\_review\_comment\_message, were created to store the translated content.
2. **Text Combination and Sentiment Analysis**: After the translation, the next step was to preprocess the review text. Using Python in Visual Studio Code, we combined the translated english\_review\_title and english\_review\_comment\_message columns into a new column called combined\_review. This combined review text was then used for sentiment analysis.

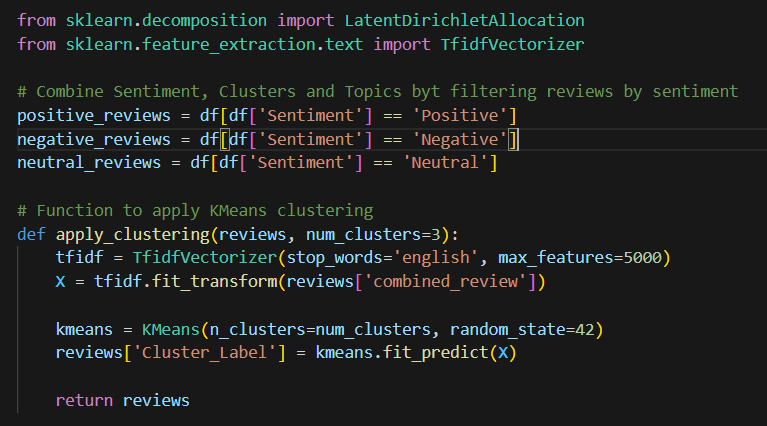


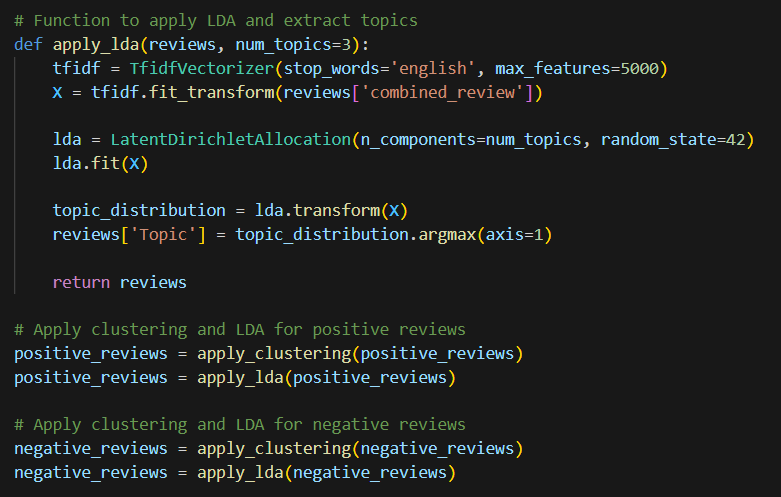
1. **Sentiment Analysis**: The sentiment of each review was analysed using a sentiment analysis model. The TextBlob library in Python was employed to classify the sentiment of the combined review as either **Positive**, **Neutral**, or **Negative** based on its polarity score.

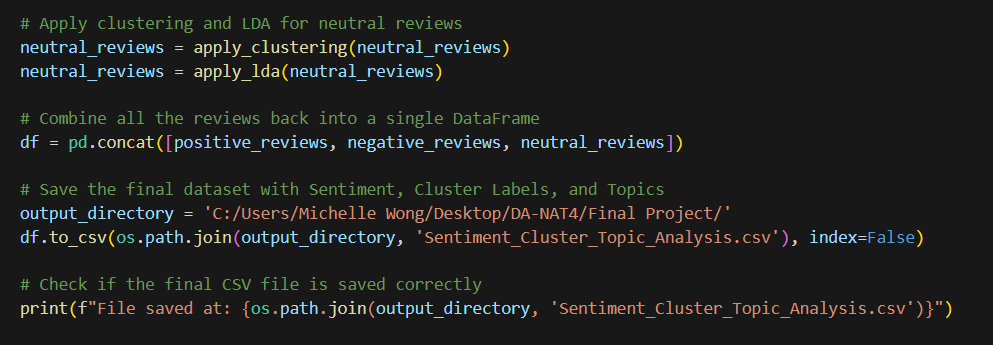


1. **Clustering and Topic Modelling**: In addition to sentiment analysis, clustering was applied to group similar reviews. K-means clustering was used to create clusters based on the content of the reviews, and a new column, Cluster\_Label, was generated to store the cluster labels (0, 1, 2, etc.). To further enhance the analysis, **Topic Modelling** was performed using the **Latent Dirichlet Allocation (LDA)** model. The model grouped reviews into topics, and a new column, Topic, was created to represent the topic assigned to each review.









1. The results of these steps were saved into a new CSV file, Sentiment\_Cluster\_Topic\_Analysis.csv, which included the original review text, sentiment, cluster labels, and topic information for further analysis and visualization in Power BI.

**Remarks:**

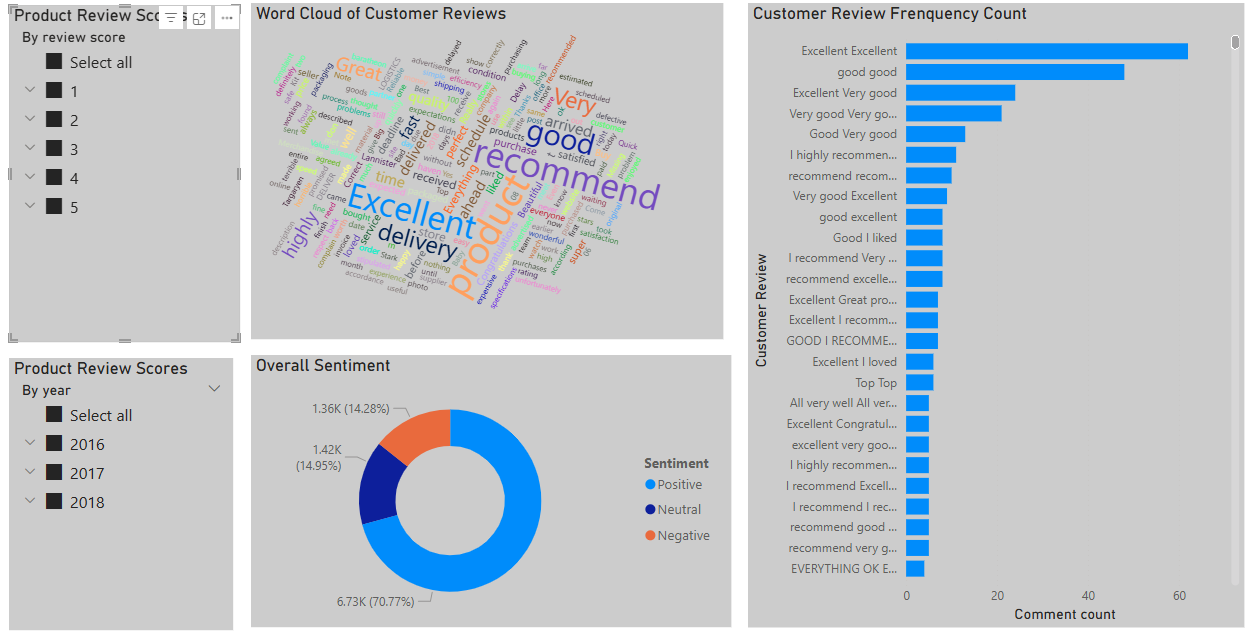
Although we initially applied **clustering** and **topic modelling** (using K-means and Latent Dirichlet Allocation) to explore patterns in the customer reviews, the results did not provide sufficiently meaningful or actionable insights in relation to the objectives of this analysis. The clusters and topics identified did not align well with the customer sentiment or the specific reasons behind customer satisfaction or dissatisfaction.

As a result, while these techniques were implemented in the dataset, we chose not to focus on them finally. Instead, we concentrated on **sentiment analysis** and the **review scores** to provide more direct insights into customer feedback.

**Visualisation and Insights**

In this analysis, sentiment and review score were analysed together can provide even deeper insights and make the analysis more comprehensive. Word Cloud, Bar Chart, Donut Chart and Slicers were used in Power BI.

**Overall Customer Review and Scores**

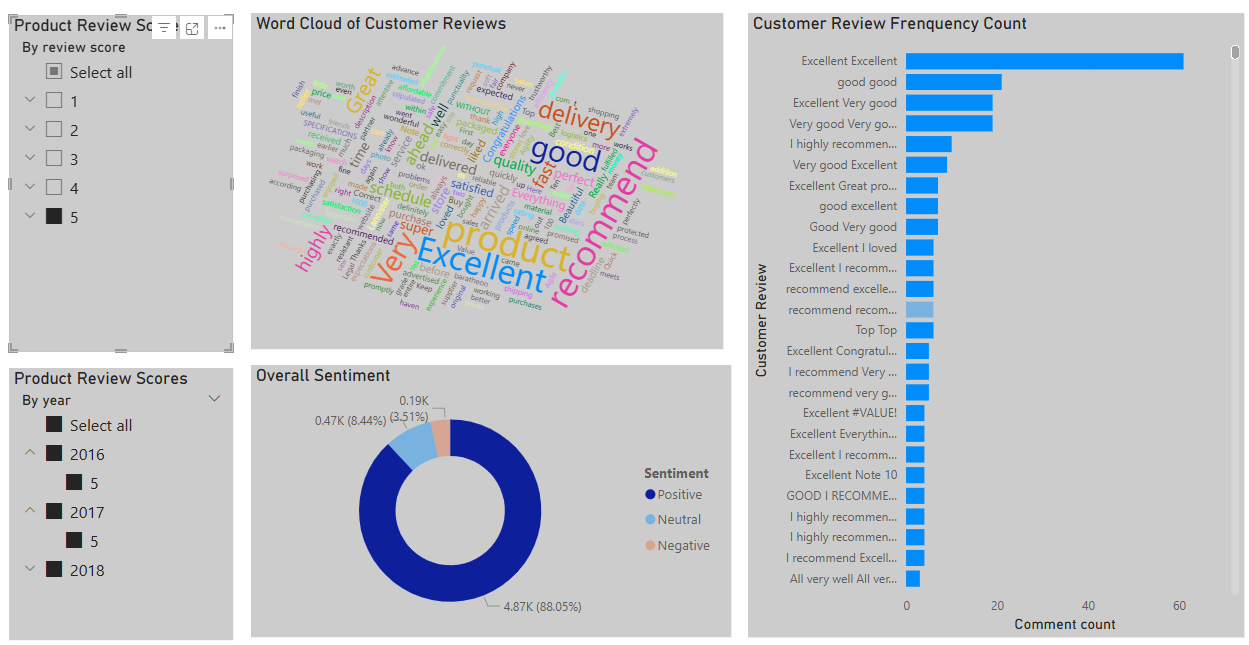


The analysis of customer reviews and scores reveals that majority of sentiments are positive, accounting for approximately 70% of all reviews. Both neutral and negative sentiments hold similar percentages, at 14.95% and 14.28%, respectively.

The most frequently mentioned words in customer reviews are **‘Excellent,’** **‘Recommend,’** **‘Product,’** and **‘Good,’** which consistently highlight customer satisfaction.

Common phrases that appeared in the reviews include **‘Excellent,’** **‘Good,’** **‘Very Good,’** and **‘I highly recommend excellent,’** indicating strong endorsement and positive customer experiences.

**Reasons Behind Customer Satisfaction**



To identify the factors contributing to customer satisfaction, we filtered the data to include only reviews with a score of ‘5’ (the highest rating) and a positive sentiment.

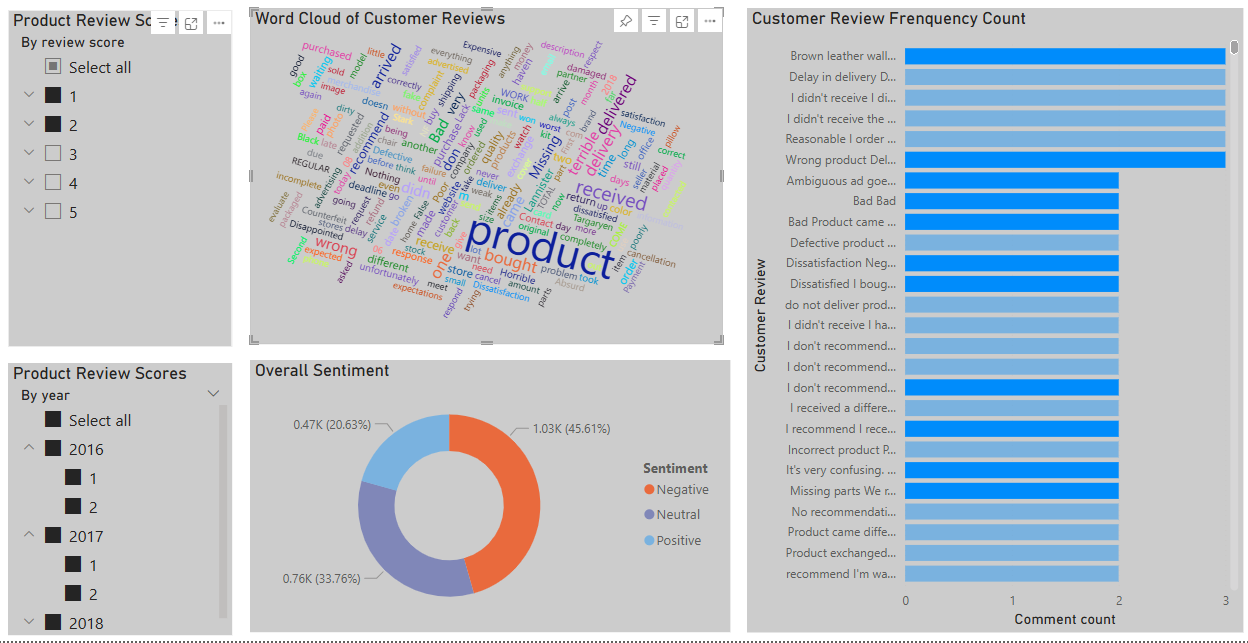
The Donut Chart highlights that 88% of customers with a review score of 5 expressed positive sentiment. This indicates a strong correlation between high scores and satisfaction.

Key themes from the analysis include:

The most frequently mentioned words in these comments are ‘Product,’ ‘Excellent,’ ‘Recommend,’ ‘Delivery’ and ‘Great,’ reflecting customer appreciation for product quality, reliability, and service efficiency.

Detailed review comments are presented in the Bar Chart, the most frequently mentioned review is ‘Excellent,’ with 61 customers explicitly leaving this comment in their message, further emphasising the high level of satisfaction among these reviewers.

**Reasons Behind Customer Dissatisfaction**



To analysing the reasons, the customers are dissatisfied can help us address the critical pain points and take immediate corrective actions to improved customer retention and brand reputation.

To identify the factors contributing to customer dissatisfaction, we filtered the data to include reviews with scores of '1’ and ‘2’ and with negative sentiment.

The Donut Chart highlights that 45.61% of reviews fall into this category, indicating a strong relationship between low scores and negative sentiment. This suggests that customers who rate their experience poorly are more likely to express dissatisfaction in their comments.

Key themes from the analysis include:

The most frequently mentioned words in these comments are ‘Product,’ ‘terrible,’ ‘delivery,’ ‘Missing,’ ‘Delivered,’ ‘received’ and ‘wrong.’ These keywords suggest that customer complaints primarily revolve around issues such as:

**Product quality:** Words like ‘terrible’ and ‘wrong’ indicate dissatisfaction with the product itself. In the Bar Chart, the specific comments are: ‘Missing parts,’ ‘Dissatisfied. I bought it in yellow, they delivered it in pink.’ and ‘Bad product came wrong.’ etc.

**Delivery problems:** Frequent mentions of ‘delivery,’ ‘missing’ and ‘delivered’ point to logistical challenges, such as delays, incorrect items, or undelivered orders. In the Bar Chart, the specific comments are: ‘Wrong Product Delivery of the product different from that requested,’ ‘Only one multimedia centre arrived, and I bought 2’ etc.

**4.2 Clustering**

The clustering analysis conducted on customer data using Recency, Frequency, and Monetary (RFM) scores, combined with sentiment analysis of customer reviews. The primary goal was to segment customers into meaningful groups based on their purchase behaviour and sentiment toward the company.

**Methodology**

* **RFM Analysis**: Calculated RFM scores (Recency, Frequency, Monetary) and how these metrics were used to segment customers. Details as shown below:

import pandas as pd

import datetime as dt

import numpy as np

# Load CSV file

dfprice = pd.read\_csv(r'C:\Users\Michelle Wong\Desktop\olist\_order\_payment\_dataset.csv')

dfprice.head()

# Repeat this process for the other dataset we need as well

dforder = pd.read\_csv(r'C:\Users\Michelle Wong\Desktop\cleaned\_olist\_orders\_dataset.csv')

dforder.head()

# Merge dfprice and dforder on 'order\_id' to get the 'order\_purchase\_timestamp' and 'payment\_value' together

dfmerged = dforder.merge(dfprice[['order\_id', 'payment\_value']], on='order\_id', how='left')

# Ensure 'order\_purchase\_timestamp' is in the merged DataFrame

dfmerged['order\_purchase\_timestamp'] = pd.to\_datetime(dfmerged['order\_purchase\_timestamp'], format="%d/%m/%Y %H:%M")

# Show the merged DataFrame

print(dfmerged.head())

# Now we can begin to calculate the recency

# we first group all the customer IDs, and grab the latest order purchase timestamp

df\_recency = dfmerged.groupby(by='customer\_id', as\_index=False)['order\_purchase\_timestamp'].max()

# then we rename the columns to be something more appropriate

df\_recency.columns = ['CustomerID', 'LastPurchaseDate']

# We grab the most recent data from the LastPurchaseDate in order to help calculate our recency score (rather than using the current date)

recent\_date = df\_recency['LastPurchaseDate'].max()

# finally, we calculate the amount of days between the most recent date and the date of the purchase

df\_recency['Recency'] = df\_recency['LastPurchaseDate'].apply(lambda x: (recent\_date - x).days)

# we can print the head of the dataframe to make sure all is well

df\_recency.head()

# here we calculate the frequency of the orders

# we calculate it by grabbing the customer ID, and counting the number of order purchases made by that customer

frequency\_df = dfmerged.groupby(by=['customer\_id'], as\_index=False)['order\_purchase\_timestamp'].count()

# then we rename the column titles as usual

frequency\_df.columns = ['CustomerID', 'Frequency']

frequency\_df.head()

# finally, we have to calculate how much we are spending

# here we're going to simply calculate the sum of the payment values per customer

monetary\_df = dfmerged.groupby(by='customer\_id', as\_index=False)['payment\_value'].sum()

# Once again, change the column titles

monetary\_df.columns = ['CustomerID', 'Monetary']

monetary\_df.head()

# Now we can begin to calculate our RFM!

# first we merge recency and frequency

rf\_df = df\_recency.merge(frequency\_df, on='CustomerID')

# then we merge this table to monetary to get our RFM table!

rfm\_df = rf\_df.merge(monetary\_df, on='CustomerID').drop(columns='LastPurchaseDate')

rfm\_df.head()

# Now we normalise the data

# first, we grant each customer a rank within their columns

rfm\_df['R\_rank'] = rfm\_df['Recency'].rank(ascending=False)

rfm\_df['F\_rank'] = rfm\_df['Frequency'].rank(ascending=True)

rfm\_df['M\_rank'] = rfm\_df['Monetary'].rank(ascending=True)

# normalizing the rank of the customers

rfm\_df['R\_rank\_norm'] = (rfm\_df['R\_rank']/rfm\_df['R\_rank'].max())\*100

rfm\_df['F\_rank\_norm'] = (rfm\_df['F\_rank']/rfm\_df['F\_rank'].max())\*100

rfm\_df['M\_rank\_norm'] = (rfm\_df['F\_rank']/rfm\_df['M\_rank'].max())\*100

# finally, we drop the original rank columns, to avoid confusion!

rfm\_df.drop(columns=['R\_rank', 'F\_rank', 'M\_rank'], inplace=True)

rfm\_df.head()

# now we calculate rfm

# we use a preset formula for this, using the normalised rankings

rfm\_df['RFM\_Score'] = 0.15\*rfm\_df['R\_rank\_norm']+0.28 \* \

rfm\_df['F\_rank\_norm']+0.57\*rfm\_df['M\_rank\_norm']

rfm\_df['RFM\_Score'] \*= 0.05

# once this is calculated, we round it to two decimal places

rfm\_df = rfm\_df.round(2)

# and print the head to check all is well!

rfm\_df[['CustomerID', 'RFM\_Score']].head(10)

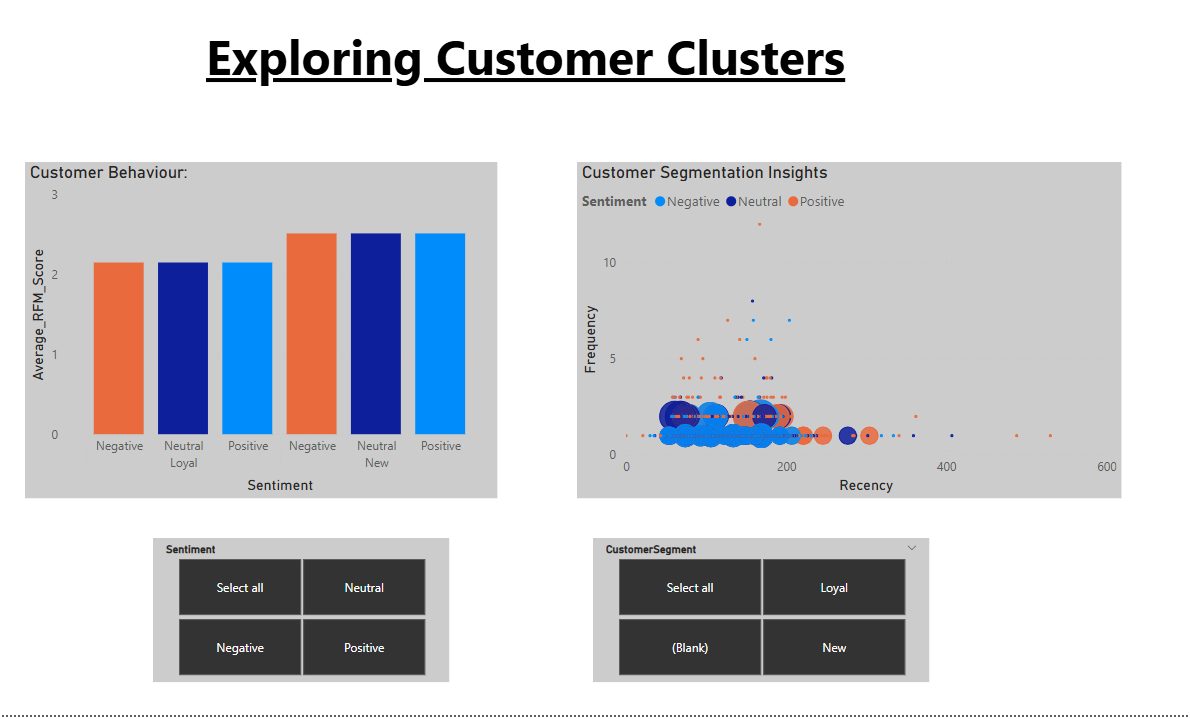
# Finally, we are going to split it into quartiles, in order to assign a 'grade' to each of our customers

# RFM\_Quartile 1 customers are our most valuable, whilst RFM 4 tend to be our least valuable!

rfm\_df['RFM\_Quartile'] = pd.qcut(rfm\_df['RFM\_Score'], 4, ['4','3','2','1'])

rfm\_df.head(10)

* **Sentiment Analysis**: it was already analysis in **4.1 Feedback Sentiment Analysis**



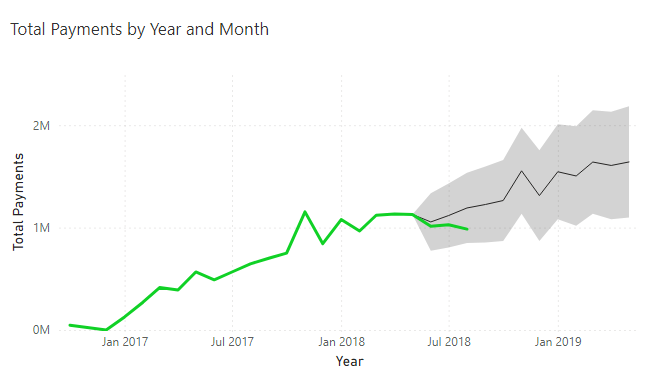
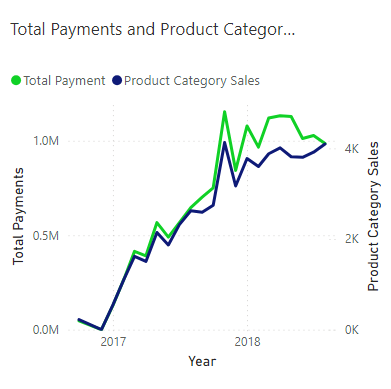
**Bar Chart** shows the average RFM scores for each customer segment (Loyal, At-Risk, and New) against their sentiment. There is not a significant difference between loyal and new customers in terms of their average RFM scores. Despite differences in engagement behaviour, new and loyal customers exhibit similar overall satisfaction and spending patterns."

#### **Scatter Plot Overview** which visualizes the relationship between Recency, Frequency, and RFM scores, with the bubble size representing the standard deviation of the RFM score across different sentiment types (positive, negative, and neutral). The larger the bubble, the more variation there is in the RFM score for that sentiment type. Smaller bubbles suggest less variation or more uniform behaviour. As the points are aligned horizontally, it indicates that Recency or Frequency doesn’t vary much for many customers. This suggests that most customers fall into predictable purchasing patterns."

* Loyal Customers (Small Bubbles): Loyal customers tend to show stable RFM scores, indicating consistent engagement. They often show high frequency but moderate recency, suggesting they engage regularly but not necessarily recently.
* At-Risk Customers (Large Bubbles): At-risk customers display more variance in their RFM scores. Some show high frequency but low recency, while others have high recency but low frequency. This inconsistency makes them harder to retain.

**4.3 Sales Prediction**

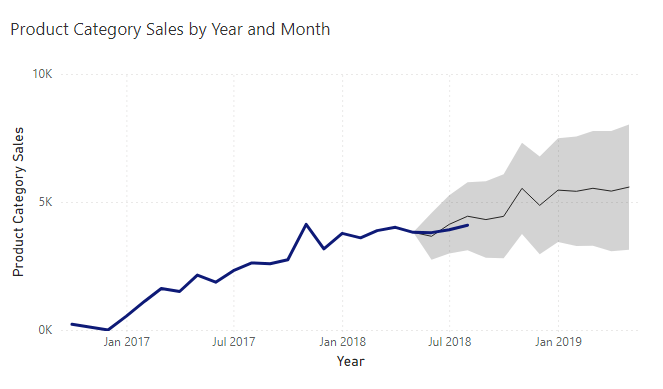
We aim to predict sales for Olist Stores in two aspects: the number of product sales (dark blue line) and the total sales value (green line) over time, as shown in the Power BI line charts. Both trends show an increase, reaching their peak in October 2017. Following this, there was a decline until December, after which the trends increased again until August 2018. Based on this observation, we predicted that in the next nine months, the number of products sold would continue to increase, reaching a new peak in November 2018, followed by a drop in December and a subsequent increase in 2019.

In the chart with the blue line, the black line represents the prediction for product sales, while the grey area indicates the confidence interval. This means that while the sales may not exactly follow the black line, they are expected to fall within the grey area. For the prediction, we started forecasting from May 2018 instead of August 2018. This approach allowed us to compare the predictions with actual data to assess their accuracy. From the chart, we can observe that the predictions for the last three months closely align with actual sales. Therefore, we believe that our prediction is reliable. 

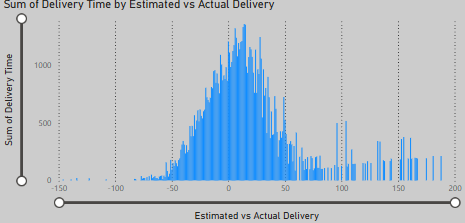
We used a similar method to predict the sales value. In the chart with the green line, we can see that the trend of sales value was also increasing, reaching its peak in November 2017. After that, it declined until December, followed by a gradual increase until August 2018. We also used the prediction function in Power BI to forecast the sales value for the next nine months. The black line represents the predicted sales value up to May 2019, while the grey area indicates the upper and lower boundaries of our prediction. According to the chart, we predict that the sales value will continue to increase over the next nine months and reach a new peak in November 2018. However, the increase appears to be more gradual and smoother compared to product sales.

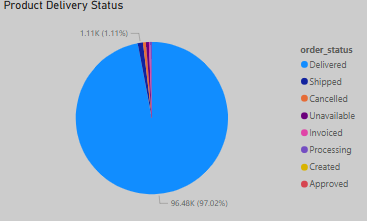
Furthermore, we observed a significant difference between the predicted sales value and the actual sales value from May 2018 to August 2018. While our prediction showed an increasing trend, the actual sales value dropped during this period. This discrepancy highlights that the prediction may not be as accurate as expected.

As a next step, we could analyse the reasons behind the increase in product sales alongside the decrease in total sales value. Additionally, we could extend the prediction period to two years to see if the trend changes over a longer timeframe.



**4.4 Delivery Time Performance**



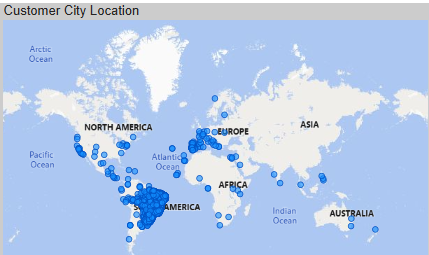


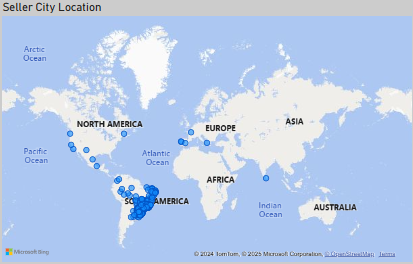
Delivery performance overall is strong with 97% of all orders shipped; 1.2% of orders were cancelled or unavailable. However, the histogram shows that delivery time performance is poor, as the majority of products were delivered with delays. This includes a significant number of orders having 100+ day delays, matching up with one of the primary concerns of the negative reviews from the customers stating that they did not receive their shipment at the pre-ordained date.

**4.5 Delivery Time Performance Additional Insight**

The store location may pose a significant effect on the delivery time. Although most orders are placed within East of South America itself, where most sellers are located, there are a considerable number of orders within Europe and Africa, which do not have as many stores. This may contribute to the late delivery times, as the product will have to be shipped overseas, which is a much longer process than shipping it locally. It may be worth expanding and opening more stores in these continents, particularly the South of Europe, where there are only 3 stores but a considerable number of orders. Opening more stores would increase in both customer satisfaction and could potentially encourage more sales in the area, by increasing the companies reach.

In the maps below, we can see the density of the seller's city, as well as the customers city. Most sellers are in South America, as well as most customers. We can draw a fair correlation between this. North America also has a considerable number of customers, around the same place that the sellers are in North America. The most beneficial next step would be to have more sellers in the south of Europe, as that has a considerable number of customers, with very few sellers. This could be leveraged by building more stores in the south of Europe, which would also speed the delivery time up for customers in Africa. Since, the customers in Africa are quite speed out, it might be worth opening a store in southern Africa too, to increase the reach. In a similar manner, we can see that there are a few customers in both Asia and Oceania but only one store in both. We can assume that the delivery time is much longer in these areas due to the lack of stores. Thus, it would be worth opening a few stores in both south Asia and southeast Oceania to increase the reach and speeding up the delivery process in both. A speedy delivery would also increase customer satisfaction and result in higher ratings.





**5. Conclusion**

Sports products are popular with Olist customers; expanding on this product range could prove profitable for the company. Issues with delivery delays can be solved by acquiring more warehouses abroad to speed up the process; incorrect shipments can be returned to these warehouses for replacement where necessary.

The analysis revealed that customer satisfaction and dissatisfaction consistently revolve around three core themes: product quality, reliability, and service efficiency. Positive reviews highlight strengths in these areas, while negative feedback points to opportunities for improvement, underscoring their critical role in shaping the overall customer experience.

**6. Limitations**

Product data is limited, listing them under vague categories such as ‘furniture decor’ and ‘health beauty.’ This does not allow us to analyse the sales using relative data such as price changes over time, prices in other countries and overall profit margins per product.

The sales data in these datasets ranges from 2016 October – 2018 August, meaning the data is now several years out of date. The patterns and analysis made using these datasets may no longer be relevant in isolation, though they can be used as supplementary analysis should more relevant sales datasets be provided in the future.

Most customer reviews were made in 2018, with none in 2016. This makes it difficult to gauge customer sentiment in 2016 and leaves the overall data skewed towards later months within the dataset.

**7. Recommendations**

**Conduct Product-Specific Analysis**: Focus on the product categories with the highest levels of dissatisfaction to identify potential issues in quality control. This will help pinpoint specific areas for improvement and ensure product reliability aligns with customer expectations.

**Link Reviews with Delivery Metrics**: Perform further analysis of customer reviews in conjunction with delivery times and warehouse or seller locations. This will provide insights into how logistics influence customer satisfaction and identify opportunities to enhance service efficiency through optimized delivery processes.

**Expand on Sports Leisure product Catalogue:** These products received the most 5\* reviews and positive comments, implying a strong appetite from the customers for them. This will increase total product sales and overall profit for Olist.

**Increase number of warehouse locations globally:** Spreading out product storage locations globally will allow Olist to reach more customers as well as provide returns locations to correct customer deliveries, increasing customer satisfaction. The African continent has zero seller locations and could provide Olist an opportunity to enter the African market. Whilst Portugal currently has many seller locations, there are still a considerable number of customers in this location and so expanding logistics here could also be of great benefit.