Academic Year 2021/2022

Term 2

IS428 – Visual Analytics for Business Intelligence

E-commerce Delivery Fleet Deployment in Brazil

Prepared for Instructor Benedict The

G1 - Group 7

Submitted by:

Name	Campus ID
Vania NG Tsin Ee	01370995
YAP Lee Sim	01389699
MA Li	01450610
Charlie ANGRIAWAN Oei	01326239

Table of Contents

1. Background

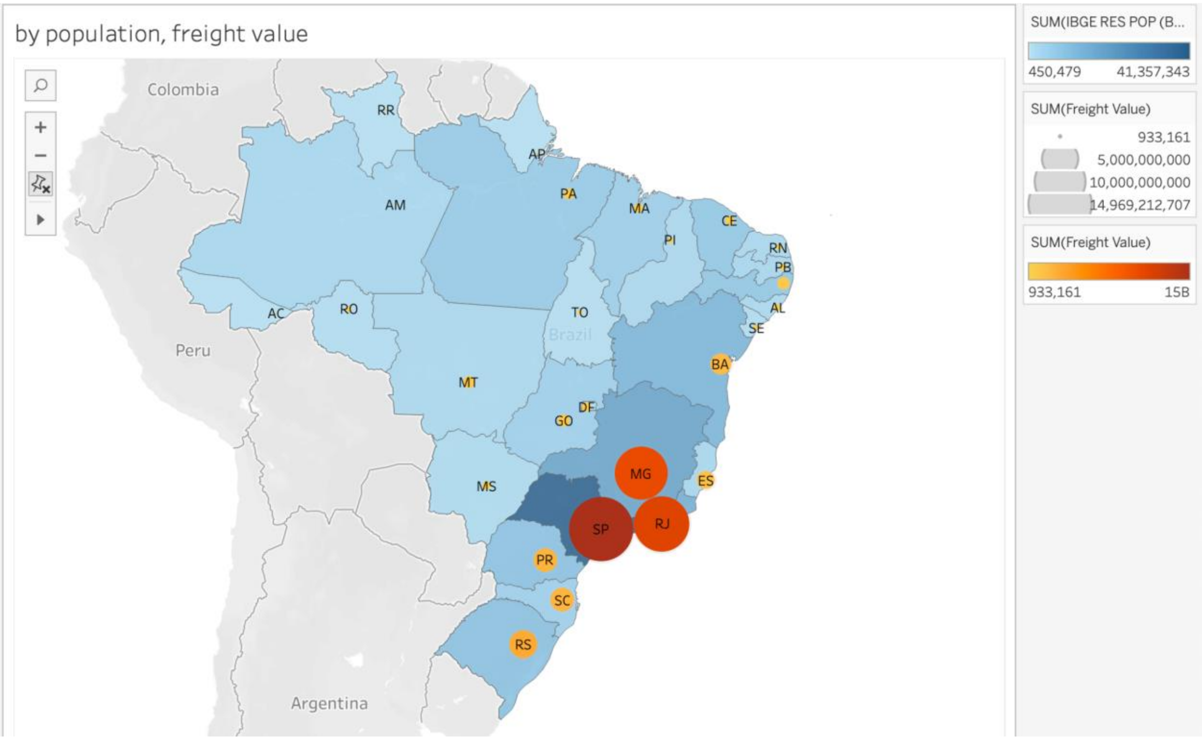
We are Flash Logistics, a Brazilian logistics company, who has recently secured a contract with Olist to take care of their deliveries. Olist is a Brazilian online marketplace that provides a platform for sellers to connect with customers. We aim to increase our delivery productivity and optimise our delivery team deployment to better cater to Olist's delivery needs in the future so that we can increase our business revenue.

Our dataset is sourced from Kaggle and spans across years 2016 to 2018. It is a public dataset of orders made at Olist Store with information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. The data contain information such as order status, price, payment and freight performance to customer locations, product attributes and customer reviews. However, we found that the data for 2016 was largely incomplete and hence, our analysis will only be based on years 2017 to 2018. Since our focus is primarily on the duration and cost of delivery, we filtered the dataset to only include orders with the order_status indicated "delivered".

2. Data Analysis

2.1 Freight value against Brazil's population

To get an understanding of how Olist's deliveries are distributed throughout the country, we first plotted the freight values on a map based on the states in Brazil [Figure 1]. We also wanted to find out if there was any correlation of the Freight Values and Population of the Brazilian states. [Figure 1] shows that the states with the biggest freight values (indicated by bubbles) are Sao Paolo (SP), Rio De Janeiro (RJ) and Minas Gerais (MG), which incidentally are the states with the largest populations (indicated by the filled map). The total freight value of Brazil from 2017 to 2018 is estimated at $2 million and the sum of freight values for the three aforementioned states make up around $1.15 million or close to 60% of the total freight value. As mentioned, these are also the three most populated states in Brazil. Of Brazil's total population of 212 million people, SP has 41 million, MG has 19 million and RJ has 16 million. Seeing how most of Olist's deliveries are clustered in these Brazilian states, we hence want to channel our company's resources and efforts to SP, RJ and MG.



[Figure 1] Population and Freight Value Graph

2.2 Freight value Against Fuel Prices

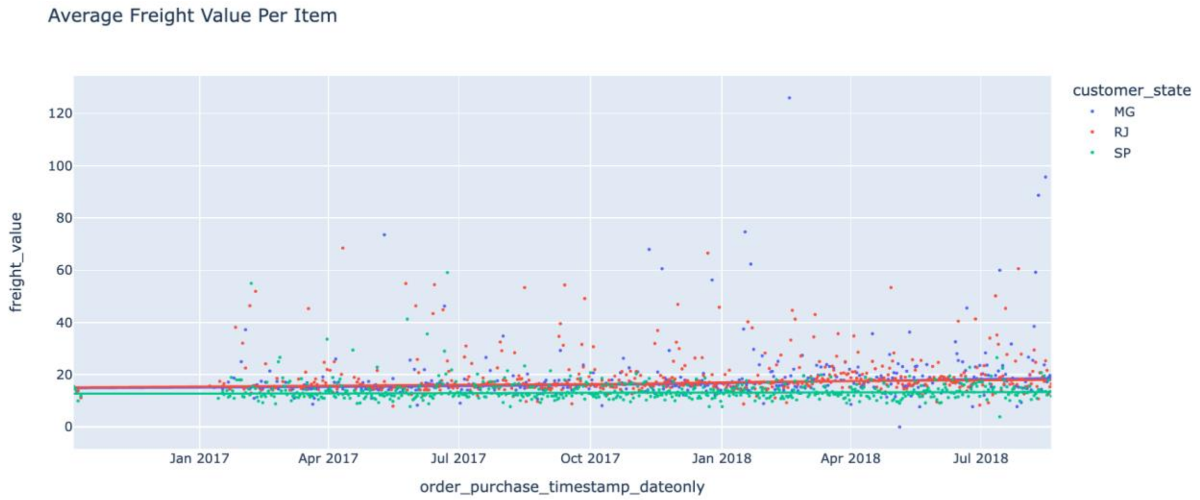
Diesel and gasoline are the most commonly used fuel for heavy-duty vehicles and light-duty vehicles respectively in Brazil. Diesel is used to power our trucks travelling long distances, while gasoline is used to power our vans, cars and motorcycles for distributing goods within cities and towns (Diesel Net, 2020; Petrobas, n.d.). To find out how fuel costs affect freight values and to what extent, which may affect our operating costs. This is done by comparing mean market prices of gasoline and diesel and average freight value in SP, RJ and MG between 5th October 2016 and 19th August 2018.

mean_market_price refers to the average gas price surveyed across the number of gas stations specified in *num_gas_station/stations_consulted*, while *freight_value* refers to the mean of all freight values on a given day.

As seen in [Figure 2, 3, 4], there is a correlation between gasoline and diesel prices and freight values because there is an upward trend for all.

[Figure 2] Average Gasoline Price in SP, RJ and MG from 5th October 2016 too 19th August 2018

[Figure 3] Average Diesel Price in SP, RJ and MG from 5th October 2016 to 19th August 2018

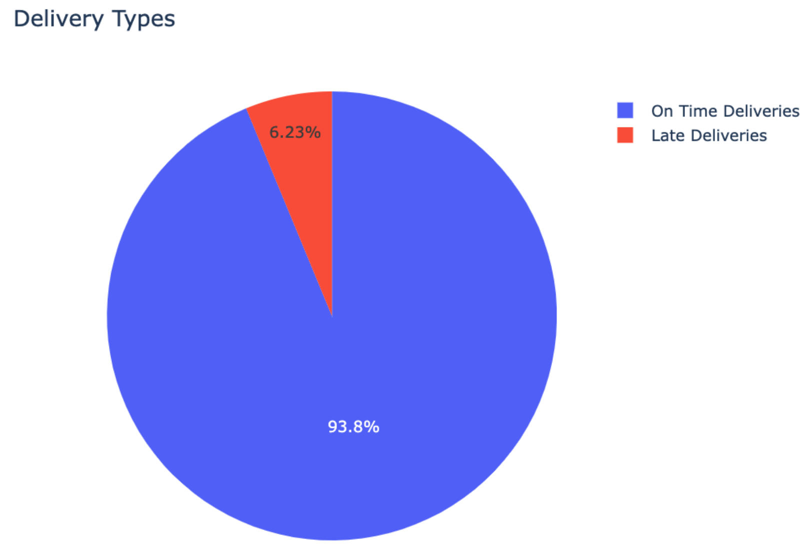


[Figure 4] Average Freight Value in SP, RJ and MG from 5th October 2016 to 19th August 2018

2.3 Delivery Data Analysis

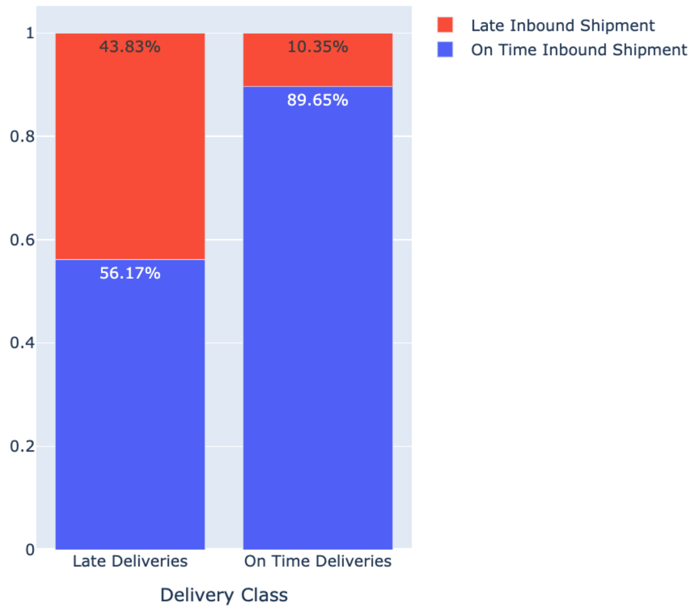
The delivery process is as follows: after a customer's order is approved, the seller hands over the product to the logistics partner (inbound shipment), who delivers the product to the customer (outbound shipment) (Lopienski, 2020). Knowing that spike in fuel costs could potentially increase our operating costs, it is critical to optimise our logistics resources to remain competitive.

We start by looking at the current delivery performance. Based on [Figure 5], 6.23% of the deliveries were late deliveries. A delivery is considered late if *order_delivered_customer_date* > *completed_orders.order_estimated_delivery_date*, meaning the product was delivered to the customer after the estimated delivery date. A delivery is considered on time if *order_delivered_customer_date* < *completed_orders.order_estimated_delivery_date*, meaning the product was delivered to the customer before the estimated delivery date.



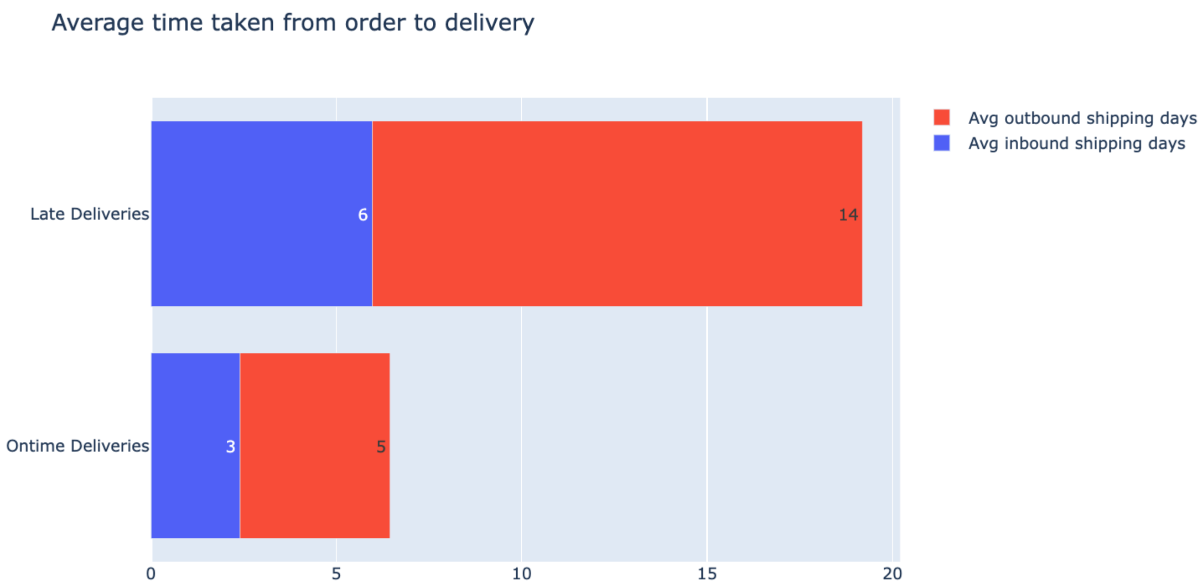
[Figure 5] Percentage of On Time Deliveries and Late Deliveries

Within the late deliveries, 44% of late deliveries were linked to late inbound shipping while 10% of the deliveries still arrived on time even though there was late inbound shipping [Figure 6]. It can be concluded there are other factors significantly contributing to the late deliveries, apart from late inbound shipment. The method to discover these other factors will be discussed later.



[Figure 6] Percentage of On Time and Late Deliveries Involving Late Inbound Shipment

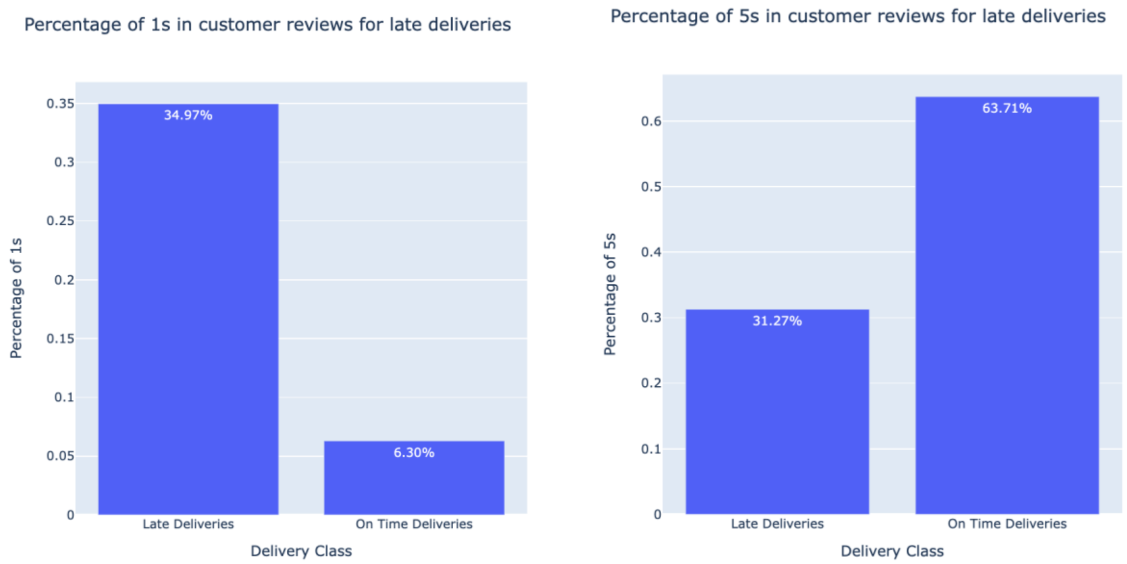
Outbound shipping (what we are responsible for) is the biggest bottleneck in the overall delivery chain [Figure 7], as it takes 14 days on average when the delivery is late, 5 days when the delivery is on time, in contrast to inbound shipping. This suggests that we should look into faster ways for outbound shipping.



[Figure 7] Delivery Bottleneck

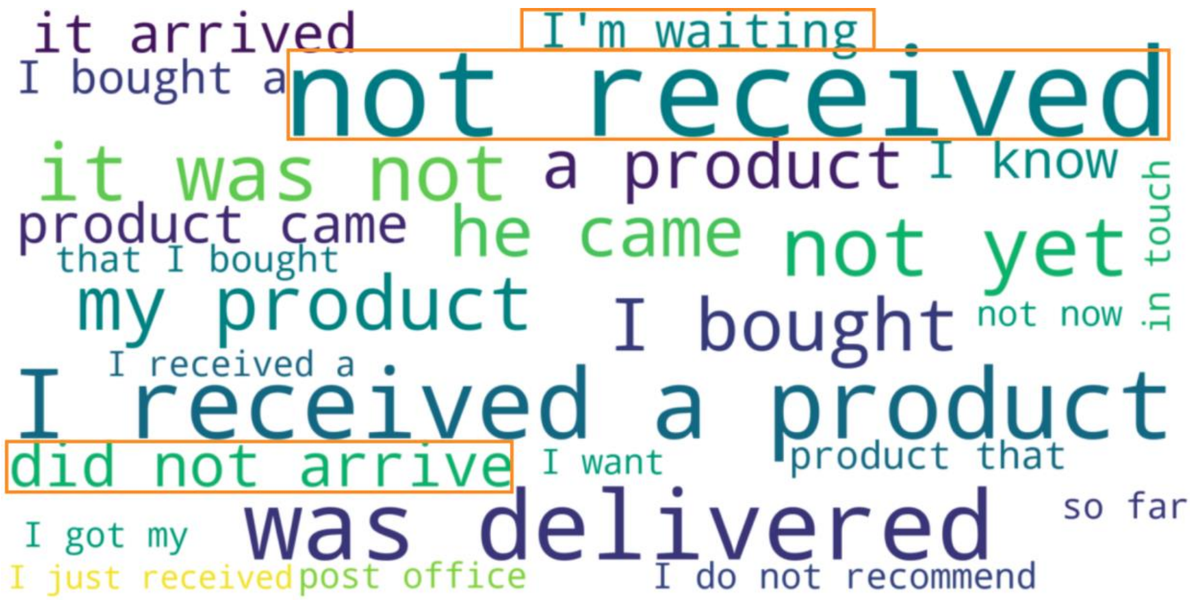
Customer satisfaction is important to our business continuity because it affects Olist's customer retention rate. If the customer retention rate is high, there is a likelihood that the amount of products outsourced to us to deliver will increase.

We found that customers are more likely to give a review score of 1 instead of review score of 5 to sellers in late deliveries [Figure 8].



[Figure 8] Percentage of 1-star Reviews in Delivery Classes vs. Percentage of 5-star Reviews in Delivery Classes

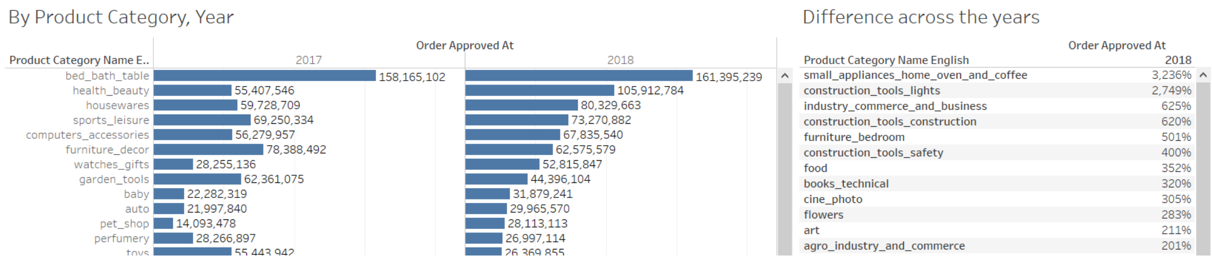
To confirm if poor review scores are caused by delivery-related issues, we generated a word cloud from reviews with a review score of 1. Indeed, words related to delivery such as 'not received', 'I'm waiting', 'did not arrive' were mentioned most frequently [Figure 9].



[Figure 9] 1-star Reviews Word Cloud

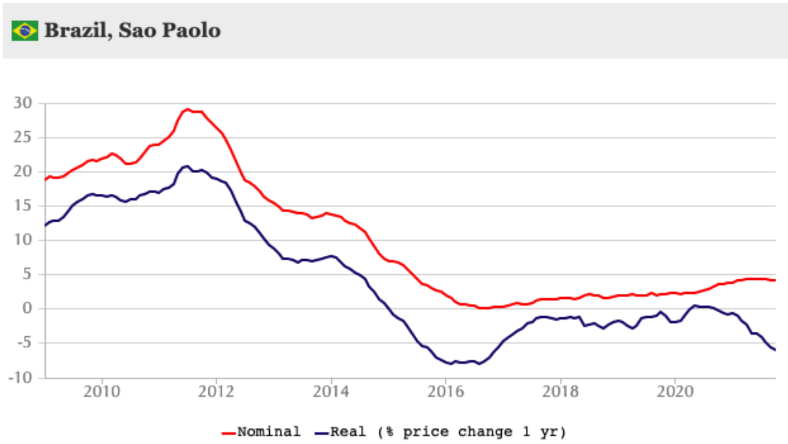
2.4 Most Popular and Up-and-coming Product Categories

Moving on, we also want to know if there are particular categories of products that are selling well or up and coming. By looking at the amount of sales and the difference in sales between 2018 and 2017 [Figure 10], we can see that in 2018, the top 5 categories are Bed_Bath_Table, Health_Beauty, Housewares, Sports_Leisure and Computers_Accessories. The categories that have seen a surge in sales from 2017 to 2018 are small_appliances_home_oven_and_coffee, construction_tools_lights, industry_commerce_and_business, construction_tools_construction and furniture_bedroom.



[Figure 10] Most Popular and Up-and-coming Product Categories

2.5 Brazil's Housing Market



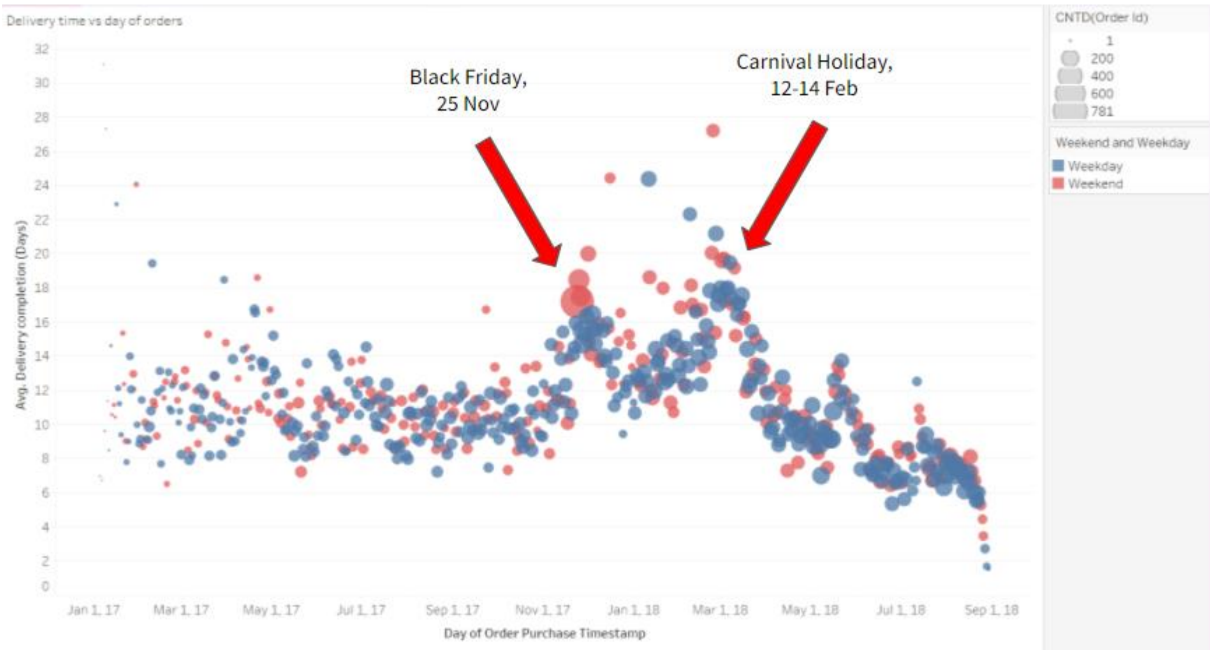
[Figure 11] House Prices in Brazil from 2017 to 2018 (Global Property Guide, 2022)

We used Sao Paolo as a representative of Brazil as it is the most populous state. The graph above [Figure 11] retrieved from GlobalPropertyGuide shows the housing prices in Brazil, Sao Paolo from 2010 to 2020 (Global Property Guide, 2022). Using the site's interactive features, we gathered that there was an increase of 2.82% from -4.71% in 2017 to -1.89% in 2018. It can be inferred from this that there was a demand in housing, which could potentially explain the surge in sales for home appliances and construction materials and tools.

2.6 Forecasting Future Order Quantity

This gives us an indication of what might be hot selling in 2019 and we want to also forecast the potential overall demand in the next 6 months [Figure 12]. From the figure, it shows that if we stick to a linear forecast, the demand in the next 6 months should range from 4750 to 5750 sales per month which means we will need to prepare our fleet to cater for this potential increase in demand. We were not able to find the seasonality because of insufficient data points.

[Figure 12] Forecast of demand for next 6 month

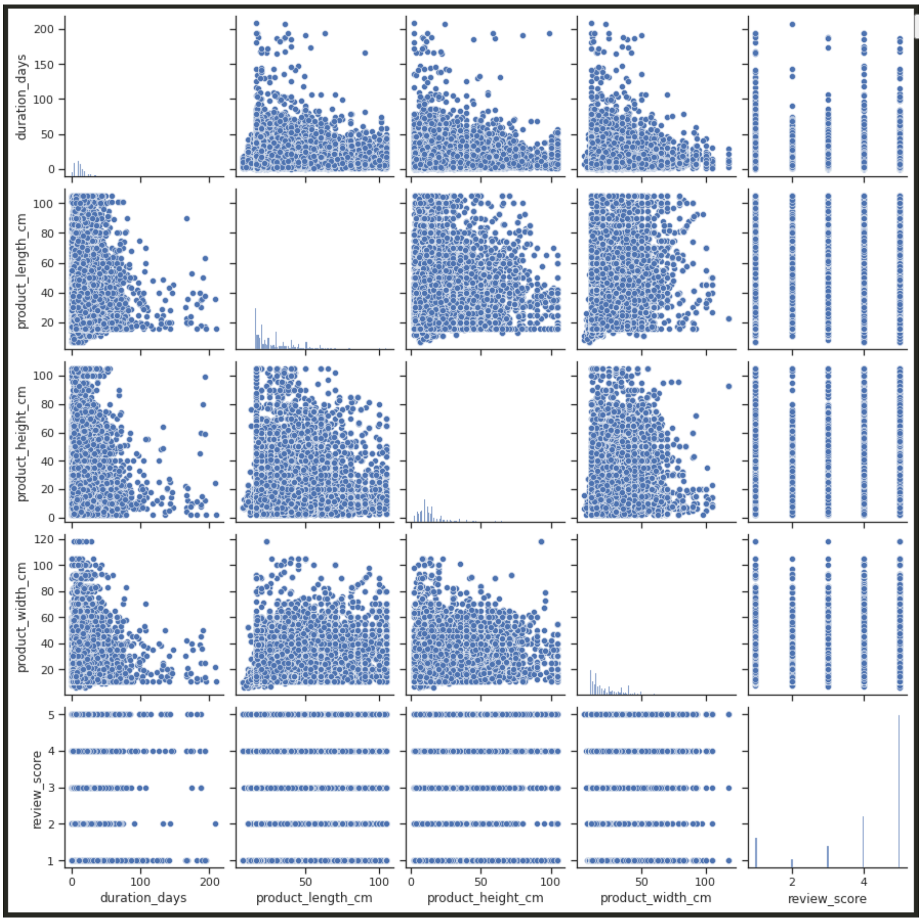


[Figure 13] Order Demand from January 2017 to September 2018

We also wanted to know if there were periods in the year where delivery took especially long compared to other periods so that we can plan and prepare our resources early. As we can see from [Figure 13], there were no obvious differences in delivery time between purchases made during the weekends (Friday, Saturday and Sunday) vs weekdays. However, there were 2 distinct peaks around November to December 2017 and February to March 2018. We believe that the December peak could be due to Black Friday sales held on 25 November and the March peak could be due to the Carnival Holidays from 12 to 14 February.

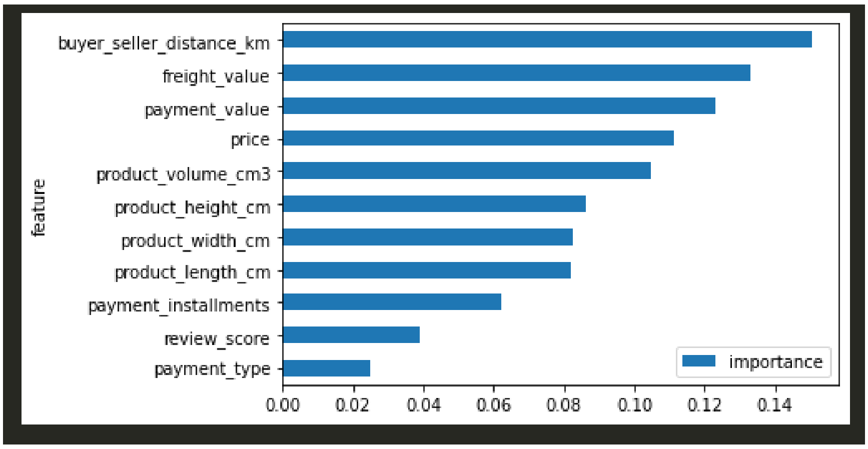
2.7 Factors Affecting Delivery Time

We plotted a correlation matrix [Figure 14] to find out if there were any clear correlations between the variables as there were no clear or strong correlations found between the variables found in the dataset.



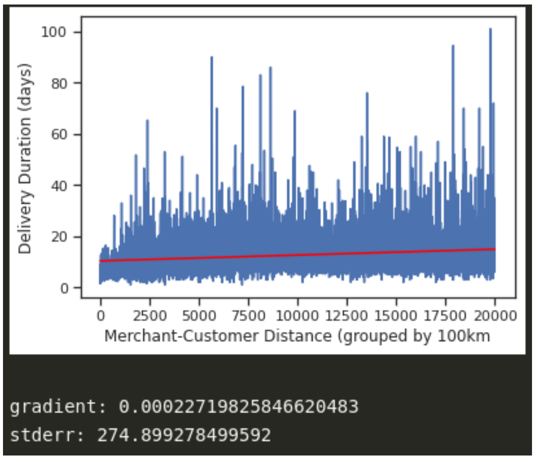
[Figure 14] Duration Days Against Other Variables Correlation Matrix

We also tried to find out what are some of the possible variables that might affect the delivery time e.g. distance between the merchant (seller) with the customer and the product volume. From this analysis, these variables did not seem to correlate with delivery time hence we decided to find out the biggest factors affecting the delivery duration using feature importance. Feature Importance provides a score for all the input features, where the scores represent the importance of each feature. It is commonly used to understand the relationship between the features and the target variable (Shin, 2022). Using a Random Forest classifier, we noticed from [Figure 15] that product volume, which includes the height, width and length, were a few of the factors that affected the delivery time of products.

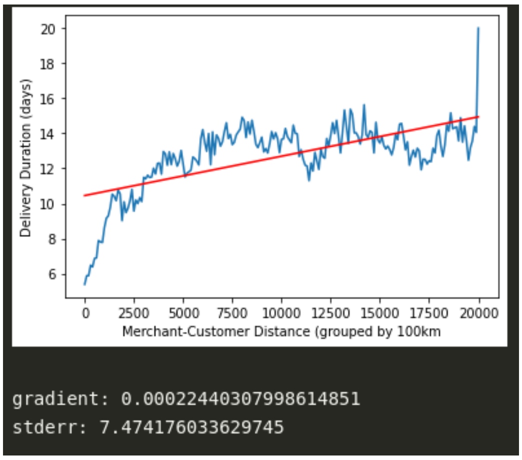


[Figure 15] Derived Feature Importance Using Random Forest

Hence we then dived deeper into how much each of these variables affected the delivery time plotted below:



[Figure 16] Correlation Between Delivery Duration and Buyer Seller Distance

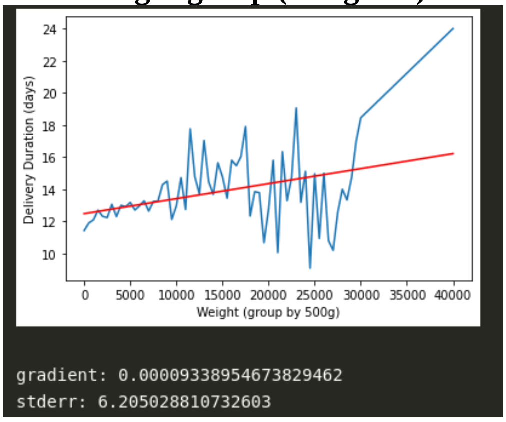


[Figure 17] Correlation Between Delivery Duration and Buyer Seller Distance (after smoothing)

As merchant-customer distance is a continuous variable, the resulting plot is very noisy [Figure 16]. However, we can see from the regression line (highlighted in red) that the gradient is positive, indicating that there might be underlying patterns. To understand the patterns, we implemented a grouping of the x-axis. For this example, we grouped the distance to per-100km and calculated the rolling average for the grouped category. After smoothing, [Figure 17] shows there is a correlation between delivery duration and buyer-seller distance.

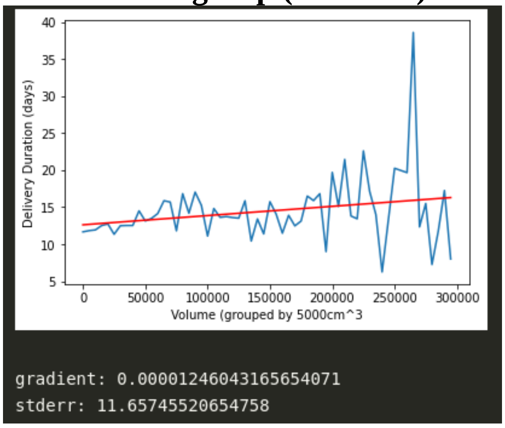
The visualisations below are plotted following the same concept, using linear regression and applying a grouped rolling average on the x-axis.

Weight group (500 gram)



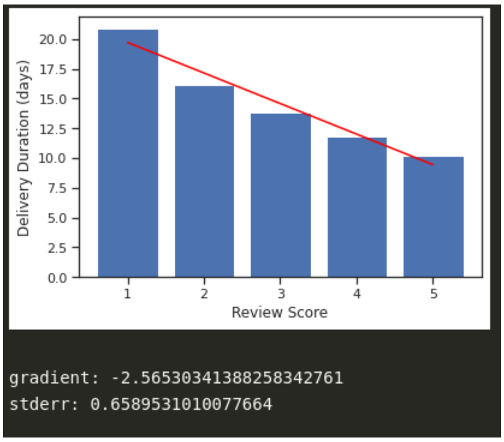
[Figure 18] Correlation Between Delivery Duration and Weight

Volume group (5000 cm^3)



[Figure 19] Correlation Between Delivery Duration and Volume

Review Score

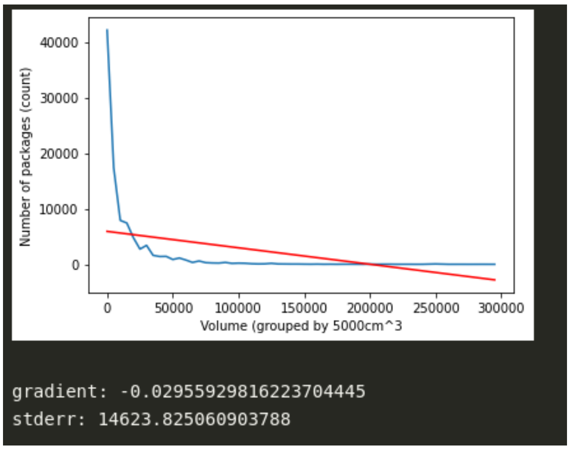


[Figure 20] Correlation Between Delivery Duration and Review Score

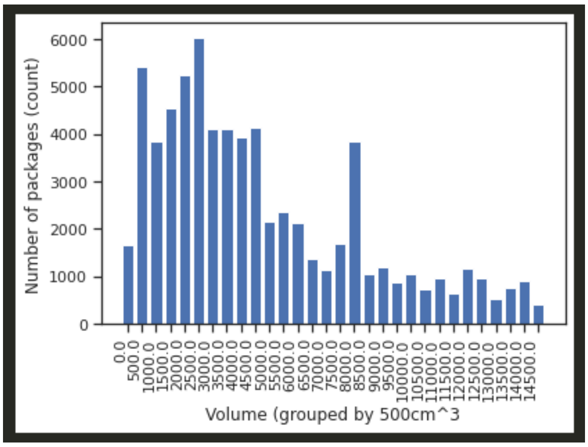
Here, we can see the trends of all the correlation of variables and infer that heavier, bulkier products coupled with a longer distance between the merchant and customer will lengthen the delivery time of the product. We can also see that orders with a long delivery time are likely to receive a low review.

2.8 Popular Box Dimensions

We intend to optimise the packaging of delivery parcels to fit more parcels into our delivery vehicles. This could improve delivery time and save resources on fuel, vehicles and manpower. We then dug deeper into understanding the product volume for the items and noticed that most of them are small sized items as visualised below. [Figure 21] shows the average product volume frequency across the dataset and [Figure 22] shows those with product volume below 15,000 cm^3.

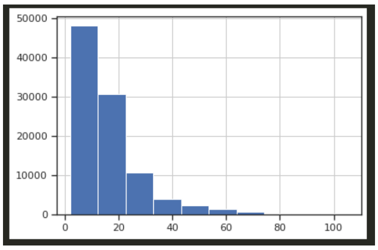


[Figure 21] Frequency of Orders By Product Volume

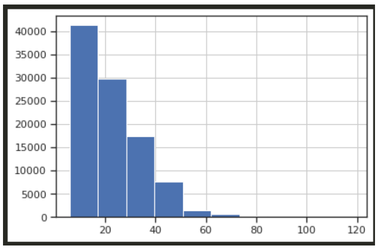


[Figure 22] Frequency of Orders By Product Volume Smaller Than 15,000 □□^{3}

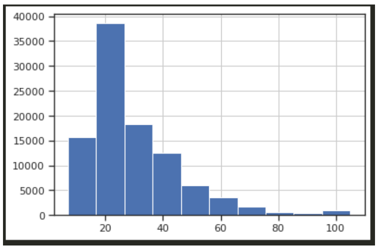
This is followed by finding out what are the most common product heights, widths and lengths [Figure 23, 24 and 25].



[Figure 23]
Product Height Frequency

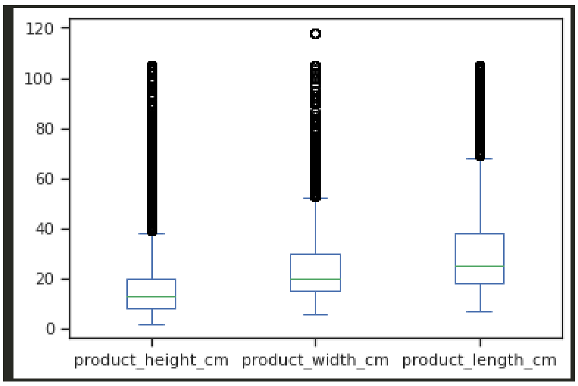


[Figure 24]
Product Width Frequency

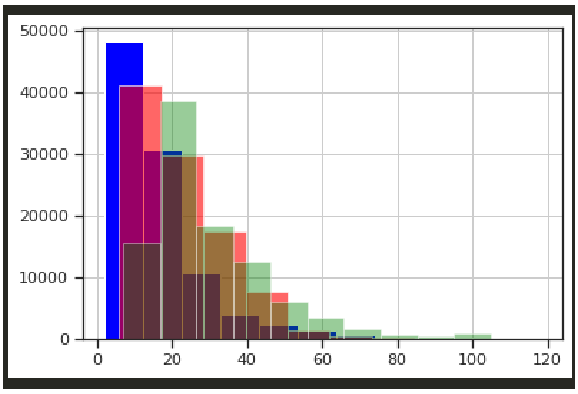


[Figure 25]
Product Length Frequency

There are many outliers in package sizes as visualised in [Figure 26] below. Since our objective is to optimise box sizes, having a vast distribution of product size is inefficient. Therefore we overlapped the bars to check if we could find zones where the product height, width and length overlap. The overlap of the product sizes can be seen in [Figure 27] below.



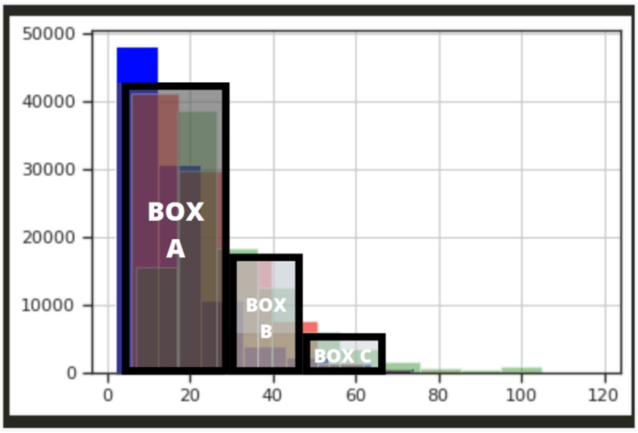
[Figure 26] Product Dimensions Frequency In A Box Plot



[Figure 27] Product Dimension Frequency In An Overlapped Histogram

Based on [Figure 27], we came out with three box sizes that would cater to over 95% of the cases as visualised below on [Figure 28].

- Box A: 30 x 30cm
- Box B: 45 x 45cm
- Box C: 70 x 70cm



[Figure 28] Box Sizes Based on Most Common Product Dimension

3. Recommendations

1. Since SP, RJ and MG make up the biggest freight values, we will focus our resources on them because they can bring in the most revenue for the company.
2. To avoid missing and/or late deliveries, we plan to improve our workflow through the following approaches:
    a. Cluster delivery points based on distance
    b. Work with Olist to remind sellers to send items to the warehouse on time
    c. Incorporate IoT to track goods on every stage of delivery
3. Given that the top 5 up-and-coming categories– Small Appliances Home Oven and Coffee, Lights, Industry Commerce and Business Items, Construction Tools and Bedroom Furniture mostly consist of fragile items, we plan to increase staff training on handling fragile items.
4. To ensure we can handle the spike in delivery demands during sales periods such as Black Friday and Carnival Holiday, we plan to increase our manpower and fleet deployment during these periods.
5. Mass-produce standardised-size boxes, keeping our operation costs low.

4. Future Works

1. Incorporate weather data into our analysis for delivery route planning in order to avoid bad weather and also understand the effects the weather has on each delivery.
2. To keep fuel costs low, more research should be done to discover how to reduce fuel consumption e.g. use of electric vehicles (EVs)

References and Data Sources

Brazilian E-Commerce Public Dataset by Olist. (2021, October 1). [Dataset]. https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_sellers_dataset.csv

Gas Prices in Brazil. (2022, March 11). [Dataset]. https://www.kaggle.com/datasets/matheusfreitag/gas-prices-in-brazil?resource=download

DieselNet. (2022, February). *Emission Standards: Brazil: Light-Duty Vehicles*. DieselNet. Retrieved April 2, 2022, from https://dieselnet.com/standards/br/ld.php

Petrobas. (n.d.). *Gasoline*. Petrobras. Retrieved April 2, 2022, from https://petrobras.com.br/en/our-activities/products/automotive/gasoline/

Enahoro, P. (2021, December 15). Improving Olist's delivery performance - precious enahoro. Medium. Retrieved April 6, 2022, from https://medium.com/@enahoroprecious/improving-olists-delivery-performance-52e0b27d0570

Global Property Guide. (2022, March 7). Brazilian Home Prices, Graphed as Time-Series. Retrieved April 6, 2022, from https://www.globalpropertyguide.com/Latin-America/Brazil/Home-Price-Trends

Shin, T. (2022, January 1). Understanding Feature Importance and How to Implement it in Python. Medium. Retrieved April 7, 2022, from https://towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285#:%7E:text=Feature%20Importance%20refers%20to%20techniques,to%20predict%20a%20certain%20variable.