DATS 6401 – Project Part 3

1. Introduction and background:

The COVID-19 epidemic has accelerated the rise of e-commerce as people shifted their shopping habits online due to physical store lockdowns. As a result, the number of online consumers and purchases increased, and businesses responded by focusing on e-commerce initiatives. As a result, the sector has seen extraordinary growth. The continuous rise in 2021 shows that customer behavior will continue to shift toward online buying. The dataset consists of several tables, including orders, payments, customers, reviews, and products, covering data from September 2016 to September 2018. The dataset contains information on more than 96096 customers, 98666 orders, 32,951 products, and 98410 reviews. The dataset's variables include customer demographics, product categories, payment methods, order dates, and review scores.

2. About the datasets:

The dataset we are intended to use is the Brazilian E-Commerce Public dataset.

Source:

The dataset is available on Kaggle, a platform for data science and machine learning competitions and resources.

Who collected the data? Who funded the project that the data came from? Other important information?

The data was created by Olist, a Brazilian e-commerce company, for the purpose of analyzing customer behavior and improving their business operations. The data was collected from January 2017 to August 2018 and includes information on orders, products, customers, sellers, and reviews. The project was funded by Olist.

Why was the data created and for what purpose? Who collected the data? Who funded the project that the data came from?

In order to manage and enhance their e-commerce operations, the Brazilian e-commerce portal Olist developed the data. Olist gathers and saves information on every step of the online purchasing process, including client orders, product shipments, and payment handling.

By gathering this information, Olist will be able to analyze their e-commerce operations, spot patterns and trends in customer behavior, and make data-driven choices that will help them grow their company. For instance, Olist can spot popular goods and categories and modify their inventory and marketing plans by analyzing the data on client orders and product sales.

What is the timeline or lineage of the data?

The data covers the period from January 2017 to August 2018 and was collected by Olist as part of their ongoing business operations.

Define and describe the variables included in the dataset.

Here is a detailed description of the key variables in the main tables:

order_id: A unique identifier for each order.

customer_id: A unique identifier for each customer.

order_status: The status of the order.

order_purchase_timestamp: The date and time the order was placed.

order_delivered_carrier_date: The date and time when the carrier received the order from the seller to deliver to the customer.

order_delivered_customer_date: The date and time when the order was delivered to the customer.

order_estimated_delivery_date: The estimated date when the order was expected to be delivered to the customer.

order_item_id: A unique identifier for each item included in an order.

product_id: A unique identifier for each product.

seller_id: A unique identifier for each seller.

shipping_limit_date: The date and time when the seller was supposed to ship the order to the carrier.

price: The price of the products in the order.

freight_value: The cost of shipping the order to the customer.

payment_type: The type of payment used by the customer for the order.

payment_installments: The number of installments chosen by the customer for payment.

payment_value: The total amount paid by the customer for the order.

review_id: A unique identifier for each customer review of a product in the order.

review_score: The rating given by the customer in their review of the product.

product_weight_g: The weight of the products in grams.

product_length_cm: The length of the products in centimeters.

product_height_cm: The height of the products in centimeters.

product_width_cm: The width of the products in centimeters.

customer_unique_id: A unique identifier for each customer who placed an order.

customer_zip_code_prefix: The first 5 digits of the customer's postal code.

customer_city: The city where the customer is located.

customer_state: The state where the customer is located.

seller_zip_code_prefix: The first 5 digits of the seller's postal code.

seller_city: The city where the seller is located.

seller_state: The state where the seller is located.

product_category_name_english: The name of the product category in English.

Customer_latitude: The latitude of the customer's location.

Customer_longitude: The longitude of the customer's location.

Seller_latitude: The latitude of the seller's location.

Seller_longitude: The longitude of the seller's location.

How large is the dataset (cases, variables)?

The dataset consists of multiple tables containing information about orders, customers, products, and sellers with a total of 100,000 orders, 100,000 customers, 32,000 sellers, and 100,000 product listings. There are also tables for geolocation, product categories, and customer reviews. The total number of variables is 36.

What locations are included in the dataset?

The dataset includes data from all over Brazil, including information on the location of customers and sellers. Therefore, it is possible to create graphs at the city/county, state, and national/world levels. However, it should be noted that the dataset is not representative of the entire Brazilian population and may have regional biases.

Describe how the datasets will help you achieve your goals/questions that posed in your introduction, including any limitations.

The dataset will help achieve the goals by providing rich and detailed information about customer behavior, product categories, payment methods, order dates, and review scores. By analyzing this data, businesses can identify patterns and trends, optimize their e-commerce initiatives, and improve customer experiences. The dataset can also help stakeholders gain a clear and actionable understanding of the Brazilian e-commerce market and the behavior of online shoppers in Brazil.

3. Data Story:

1) In this bar graph, we will explore the purchase behavior of customers in terms of the every month in which they made their purchases. The bar graph depicts the number of orders purchased monthly in 2017 and 2018. Throughout the specified time, the number of orders purchased grew. We can see that the number of orders purchased was lowest in January 2017 (about 1035) and increased to 8655 as the highest count in 2018. We also identified popular product categories and their seasonality, as well as analyzed the customer behavior in terms of repeat purchases. These insights can be used to improve the marketing and sales strategies of e-commerce businesses and optimize their operations.

Variables used: Order Purchase Timestamp

Dataset used:
olist_orders_dataset: Order Purchase Timestamp
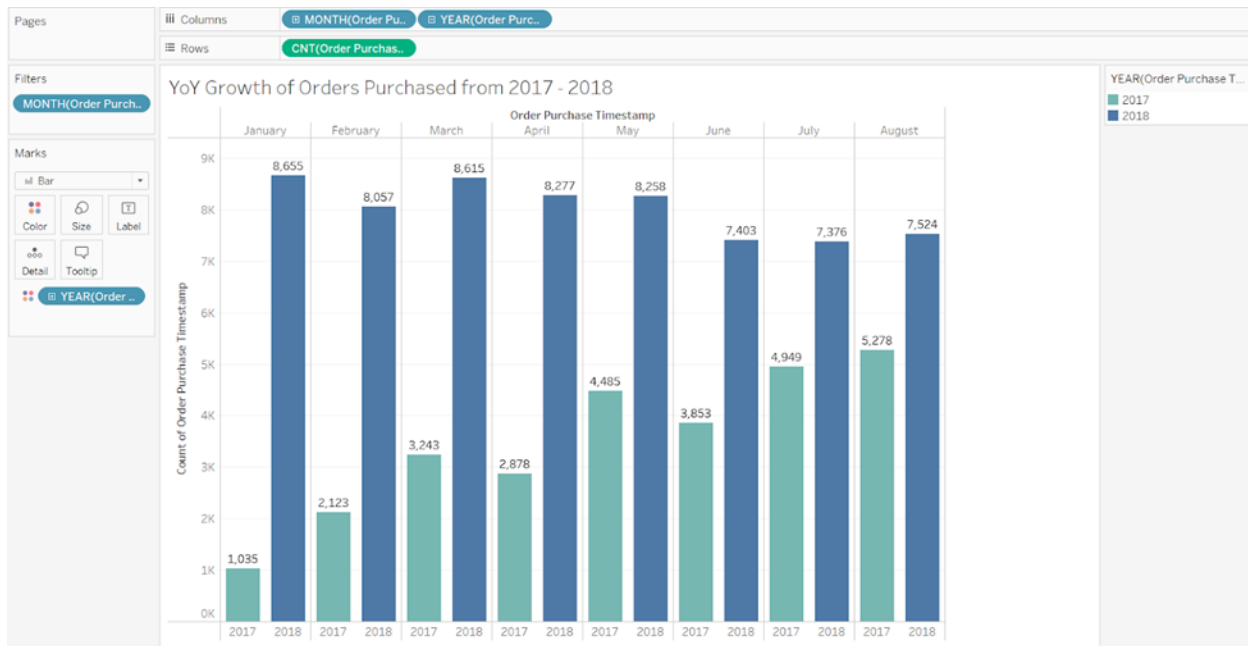
2) In this table, we will observe the purchase behavior of customers on the Olist marketplace in Brazil varied across different product categories and over the years. We observed that some product categories were consistently popular throughout the years, while others were more popular in certain years. Furnishings, Health and Beauty, and Sports and Leisure are the most

popular product categories in terms of order volume. We also identified trends in the overall purchase behavior of customers, as well as fluctuations in the number of customers across different product categories and over the years. These insights can be used by e-commerce businesses to optimize their operations and marketing strategies, and improve customer retention.

Variables used: Order Purchase Timestamp, Product Category

Dataset used:
olist_orders_dataset: Order Purchase Timestamp
olist_products_dataset: Product Category
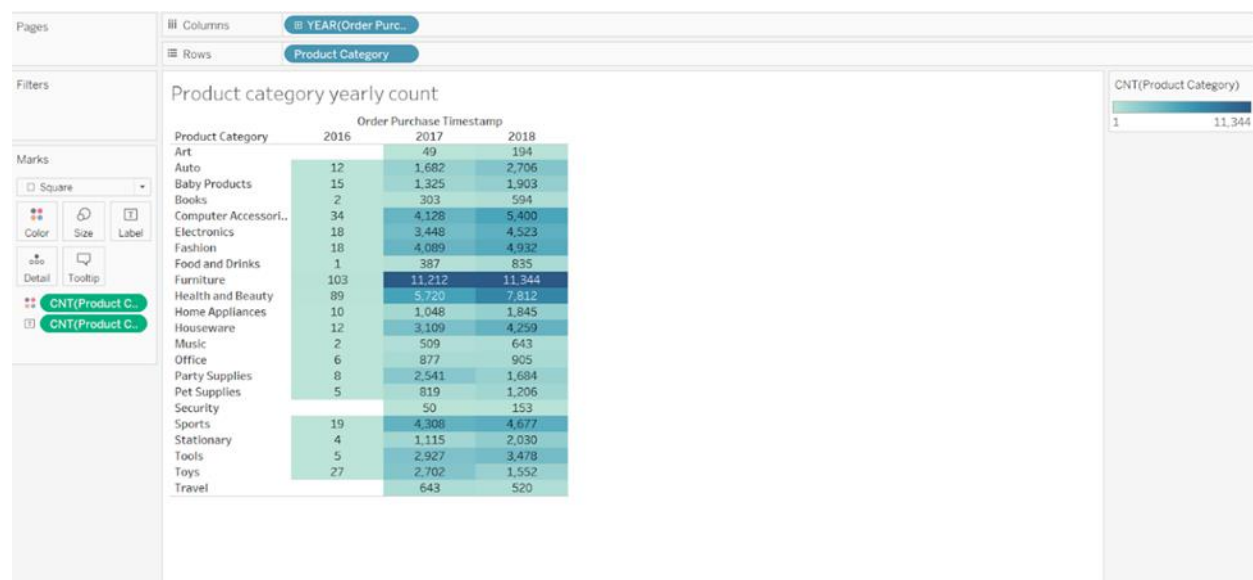
3) In the below map we will explore the geographical distribution (density) of customers on the Olist marketplace in Brazil, based on their longitude and latitude coordinates. we will explore the geographical distribution of customers on the Olist marketplace in Brazil, based on their longitude and latitude coordinates. The choropleth map shows that southeast states with the highest average order value when compared to the remaining regions in Brazil. The purchasing behavior of customers is also related to their geographical location, with certain states having a higher number of orders and a higher average order value. These insights can be used by e-commerce businesses to optimize their operations and marketing strategies, and improve customer acquisition and retention in different regions of Brazil.

Variables used: Customer latitude, Customer longitude

Dataset used:
olist_geolocation_dataset: Customer latitude, Customer longitude
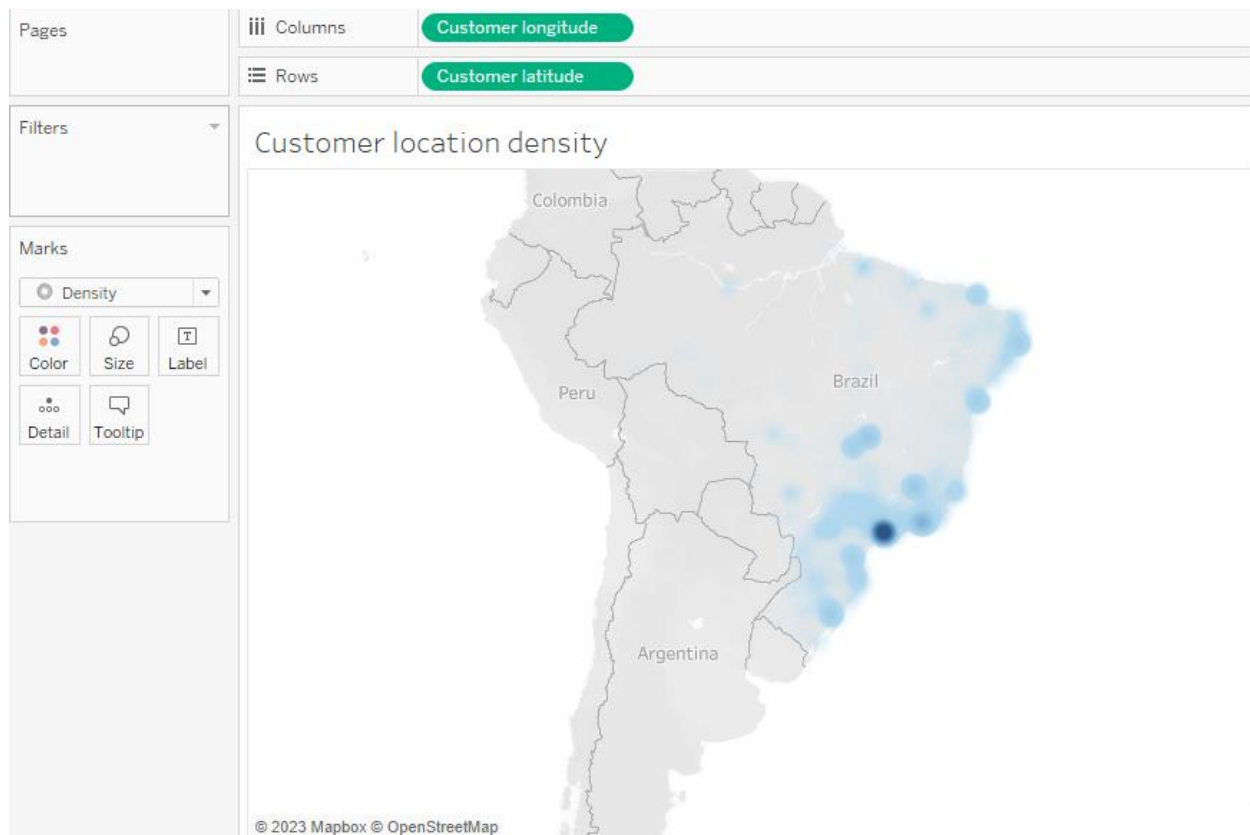
4) In the below line graph, we will explore the purchasing patterns of customers on the Olist marketplace in Brazil, based on their order purchased timestamp (month) and the count of orders and payment values. It depicts the number of orders purchased and payment value on a monthly basis in 2017 and 2018 for the given products. Customers also tend to spend more when they place more orders, indicating that they are willing to pay for multiple items at once. These insights can be used by e-commerce businesses to optimize their operations and marketing strategies, and improve customer acquisition and retention in different regions of Brazil.
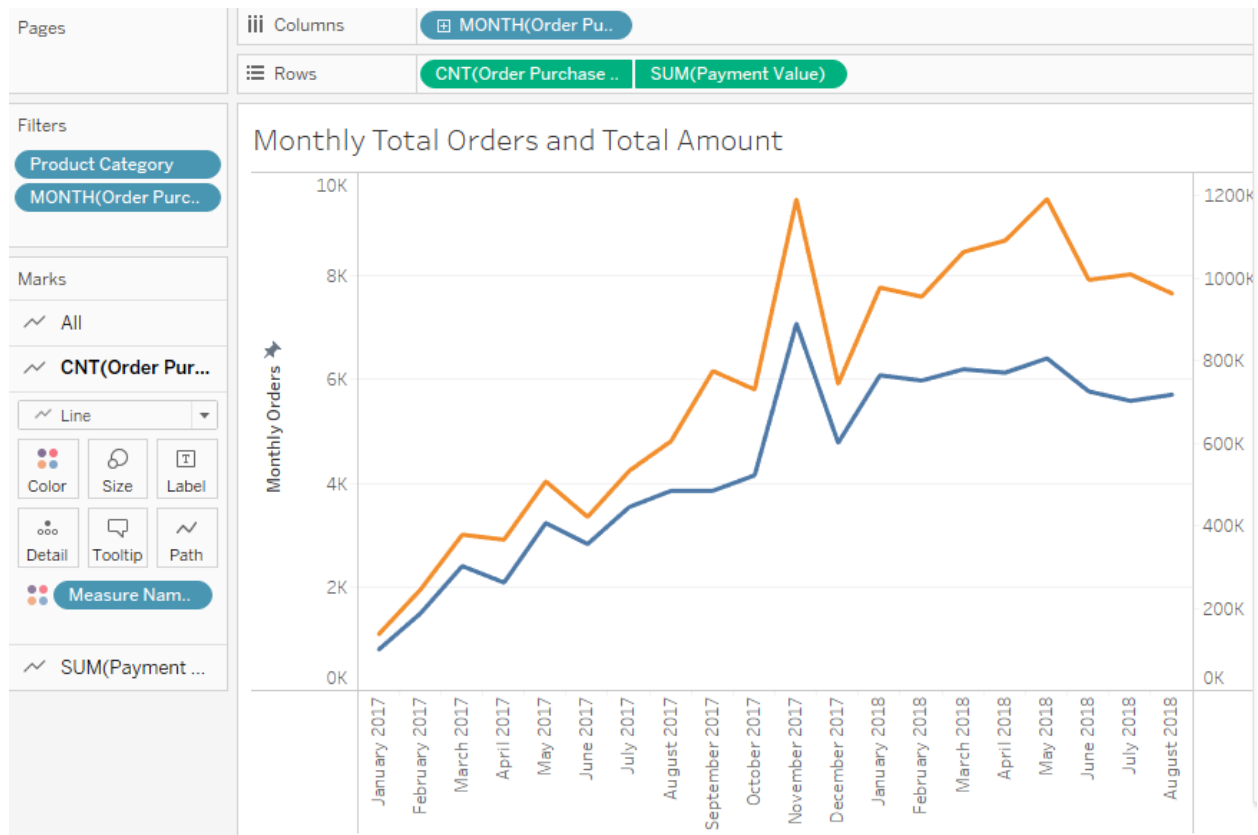
Variables used: Order Purchase Timestamp, Payment Value, Product Category

Dataset used:
olist_orders_dataset: Order Purchase Timestamp
olist_orders_payments_dataset: Payment Value
product_category_name_translation: Product Category

Monthly Total Orders and Total Amount

5) We'll look at the Brazilian Olist marketplace's consumer payment options in the pie chart below. The pie graphic demonstrates that credit cards are the most common form of payment used by consumers, which makes up the majority of purchases placed on the Olist marketplace. The second most common payment method in Brazil is called Boleto Bancário, which enables customers to pay for their purchases using a bank slip. Customers use other payment options like debit cards, vouchers, and bank accounts less frequently. E-commerce companies can use this data to streamline their payment handling procedures and provide the payment options that their clients are most likely to use.

Variables used: Payment type.

Dataset used:
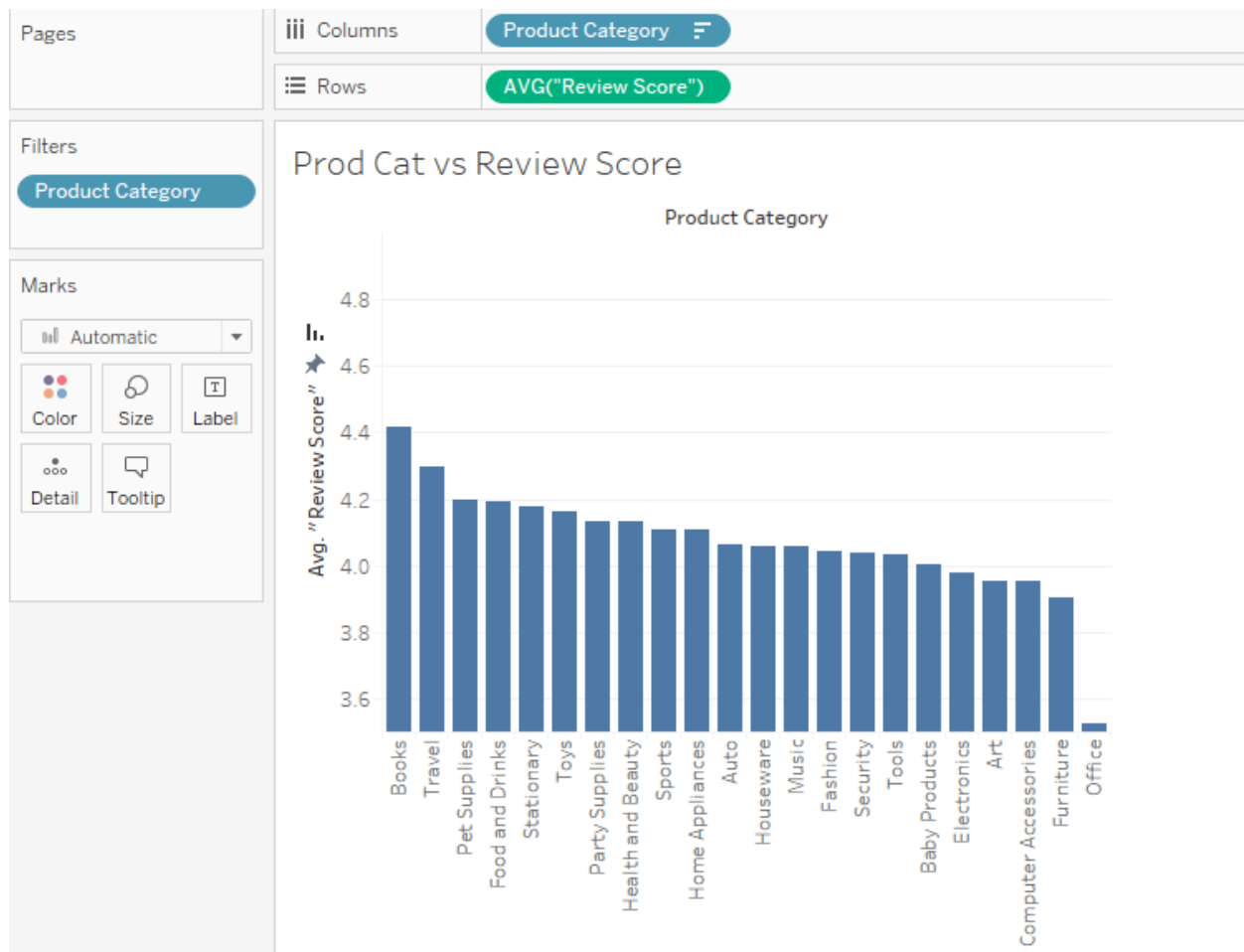olist_order_payments_dataset: Payment type

Payment type
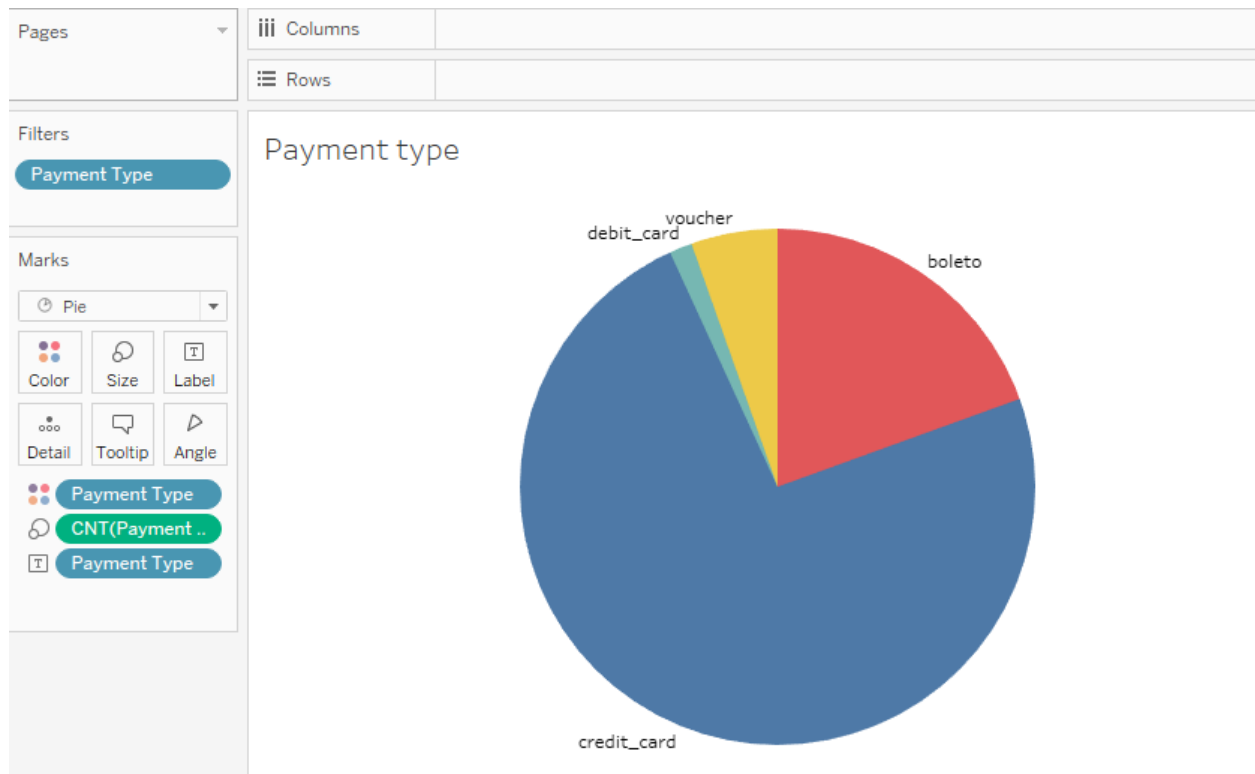
6) Below bar chart we will explore the review scores for different product categories on the Olist marketplace in Brazil. This information can be used by e-commerce businesses to identify areas for improvement and to focus on product categories that are popular with customers and have high review scores. Additionally, businesses can use this information to address issues that may be affecting the review scores for specific product categories.

Variables used: Product Category, Review Score

Dataset used:
olist_products_dataset: Product Category
olist_order_reviews_dataset: Review Score

Pages

Columns Product Category

Rows AVG("Review Score")

Filters
Product Category

Prod Cat vs Review Score

Product Category

Marks

Automatic

Color Size Label

Detail Tooltip

7) We will examine the purchase frequency on Brazil's Olist marketplace in the image below. The bar graph demonstrates that on the Olist marketplace, purchases are made most frequently between the hours of 10 a.m. and 4 p.m., peaking at 2 p.m. Early morning (between 12 am and 6 am) and late nighttime hours are when orders are comparatively less frequent. (Between 8 pm and 12 am). E-commerce companies can use this data to plan their operations and distribute resources based on when purchases are most likely to be placed. Businesses can also use this data to spot patterns and trends in the regularity of purchases over time and make necessary adjustments.

Variables used: Order Purchase Timestamp

Dataset used:
olist_orders_dataset: Order Purchase Timestamp

4. Summary and Conclusions:

The provided study examines consumer habits and shopping trends on the Brazilian e-commerce site Olist. The data used in the project was produced by Olist between January 2017 and August 2018 and contains details on orders, goods, customers, vendors, and evaluations. In order to expand the company, the initiative seeks to better e-commerce operations, analyze customer behavior, and make data-driven choices. In order to understand customer behavior and improve e-commerce operations and marketing strategies, the project investigates a number of factors, including order buy timestamp, product categorization, customer latitude, customer longitude, payment value, and payment method. To provide a thorough grasp of consumer behavior and patterns in the e-commerce sector, the analysis uses bar graphs, tables, maps, line graphs, and pie charts.

5. References:

Kaggle. (n.d.). Brazilian E-Commerce Public Dataset by Olist (Version 7). Retrieved March 4, 2023, from **https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/versions/7**

Morgan Stanley. (2022, February 9). Global E-Commerce Growth Forecast 2022. Retrieved March 4, 2023, from **https://www.morganstanley.com/ideas/global-ecommerce-growth-forecast-2022#:~:text=The%20COVID%2D19%20global%20e,2019%20to%2021%25%20in%202021**.

Global Newswire. (2022, October 10). Outlook on the Brazil B2C E-commerce Market to 2026 - Global E-Commerce Behemoths are Seeking to Launch Their Payment Services in Brazil Amid the Rapid Digital Payment Adoption. Retrieved March 4, 2023, from

https://www.globenewswire.com/en/news-release/2022/10/10/2530907/28124/en/Outlook-on-the-Brazil-B2C-Ecommerce-Market-to-2026-Global-E-Commerce-Behemoths-are-Seeking-to-Launch-Their-Payment-Services-in-Brazil-Amid-the-Rapid-Digital-Payment-Adoption.html#:~:text=Like%20the%20B2C%20e%2Dcommerce,rapid%20adoption%20of%20social%20commerce.

Tableau. (n.d.). 10 Dashboards Every Retailer Should Use. Retrieved March 4, 2023, from https://www.tableau.com/10-dashboards-every-retailer-should-use

Datapine. (n.d.). Sales Dashboard Examples & Templates for Sales Managers. Retrieved March 4, 2023, from https://www.datapine.com/dashboard-examples-and-templates/sales

Statista. (2022). E-commerce worldwide - Statistics & Facts. Retrieved March 4, 2023, from https://www.statista.com/topics/871/online-shopping/

6. Contributions:

We all contributed equally to the project and worked well together. To contribute to the project, each of us performed research and selected a dataset idea. After finalizing the dataset, we collaborated to join the eight datasets to ensure that all information was accurately merged. To develop the data visualizations, we all made two charts and shared them with the team. We then discussed the visualizations and decided on the top seven. During the project, we made certain that the burden was well allocated and that everyone understood their roles.