Analyzing eCommerce Business Performance with SQL

Created by:
Nur Imam Masri
Email : nurimammasri.01@gmail.com
LinkedIn : linkedin.com/in/nurimammasri
Github : github.com/nurimammasri
Portfolio : bit.ly/ImamProjectPortfolio

I am innovative and able to work together in a team orientation. Have an understanding of the fields of **Data Science, Data Analyst, Machine Learning, and Artificial Intelligence** with 5 years of learning experience and a specialist certificate (**TensorFlow Developer Certificate**). Often performs the data mining process by analyzing data, representing data, preprocessing data, making predictions. Excellent in understanding business operations and various data analytics tools (**Python, R, SQL, Pandas, Numpy, Matplotlib, Seaborn, TensorFlow, Sklearn, BigQuery, etc.**) and dashboarding using **Google Data Studio, Power BI, and Tableau.**

Rakamin Academy

Project Introduction

Overview :

In a company measuring business performance is very important to track, monitor, and assess the success or failure of various business processes. This can help us to see current market conditions, growth analysis, and product analysis, and to develop opportunities for new, more effective business methods. Therefore, this project will analyze the business performance of an eCommerce company, taking into account several business metrics, namely customer growth, product quality, and payment types.

In this project, an analysis will be carried out using **PostgreSQL**, and visualizing the results using **Tableau.**

Datasets :

The dataset used is provided by Rakamin Academy, <u>Brazilian E-Commerce Public Dataset by Olist</u>. Or it can be accessed on the <u>kaggle dataset</u>. This is a Brazilian e-commerce public dataset of orders made at the Olist Store. The dataset has information on 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing orders from various dimensions: from order status, price, payment, and freight performance to customer location, product attributes, and finally reviews written by customers. This also released a geolocation dataset that relates Brazilian zip codes to lat/long coordinates.

1. Creating Database & Table :

- **Create Database**

  We can create an *ecommerce* database by using the CREATE DATABASE query or by using the pgAdmin 4 GUI

  ```
  CREATE DATABASE ecommerce;
  ```

- **Create Tables**

  Create several tables in the *ecommerce* database by using the CREATE TABLE query or using the pgAdmin 4 GUI, along with column names, data types, and primary keys.

  ```
  -- ======= customers =======
  CREATE TABLE customers (
      customer_id varchar(50) NOT NULL,
      customer_unique_id varchar(50) NULL,
      customer_zip_code_prefix varchar(50) NULL,
      customer_city varchar(50) NULL,
      customer_state varchar(50) NULL,
      CONSTRAINT customers_pk PRIMARY KEY (customer_id)
  );
  ```

List of Tables Created

customers	order_reviews
geolocation	orders
order_items	products
order_payments	sellers

2. Importing Data :

- **Import Data into Table**

  After make sure column names and data types match, then we enter the .csv data into each table by using the COPY query or using the pgAdmin GUI.

  ```
  -- ======= order_items =======
  COPY order_items(
      order_id,
      order_item_id,
      product_id,
      seller_id,
      shipping_limit_date,
      price,
      freight_value
  )
  FROM
  '\Dataset\order_items_dataset.csv'
  DELIMITER ','
  CSV HEADER;
  ```

- **Pre-Cleaning Data**

  Specifically for geolocation data, cleaning will be carried out first including :

  1. Drop Duplicate Rows

  2. Change Special Character in City

  3. Input new geolocations from customers and sellers

  ```
  CREATE TABLE geolocation AS
  WITH geolocation AS (
      SELECT geolocation_zip_code_prefix,
          geolocation_lat,
          geolocation_lng,
          geolocation_city,
          geolocation_state FROM (
          SELECT *,
              ROW_NUMBER() OVER (
                  PARTITION BY geolocation_zip_code_prefix
              ) AS ROW_NUMBER
          FROM geolocation_dirty2
      ) TEMP
      WHERE ROW_NUMBER = 1
  ),
  custgeo AS (
      SELECT customer_zip_code_prefix, geolocation_lat,
      geolocation_lng, customer_city, customer_state
      FROM (
          SELECT *,
              ROW_NUMBER() OVER (
                  PARTITION BY customer_zip_code_prefix
              ) AS ROW_NUMBER
          FROM (
              SELECT customer_zip_code_prefix, geolocation_lat, geoloca
              FROM customers cd
  ```
  ....

The full query can be seen here

Table Geolocation before cleaning

Result Cleaning
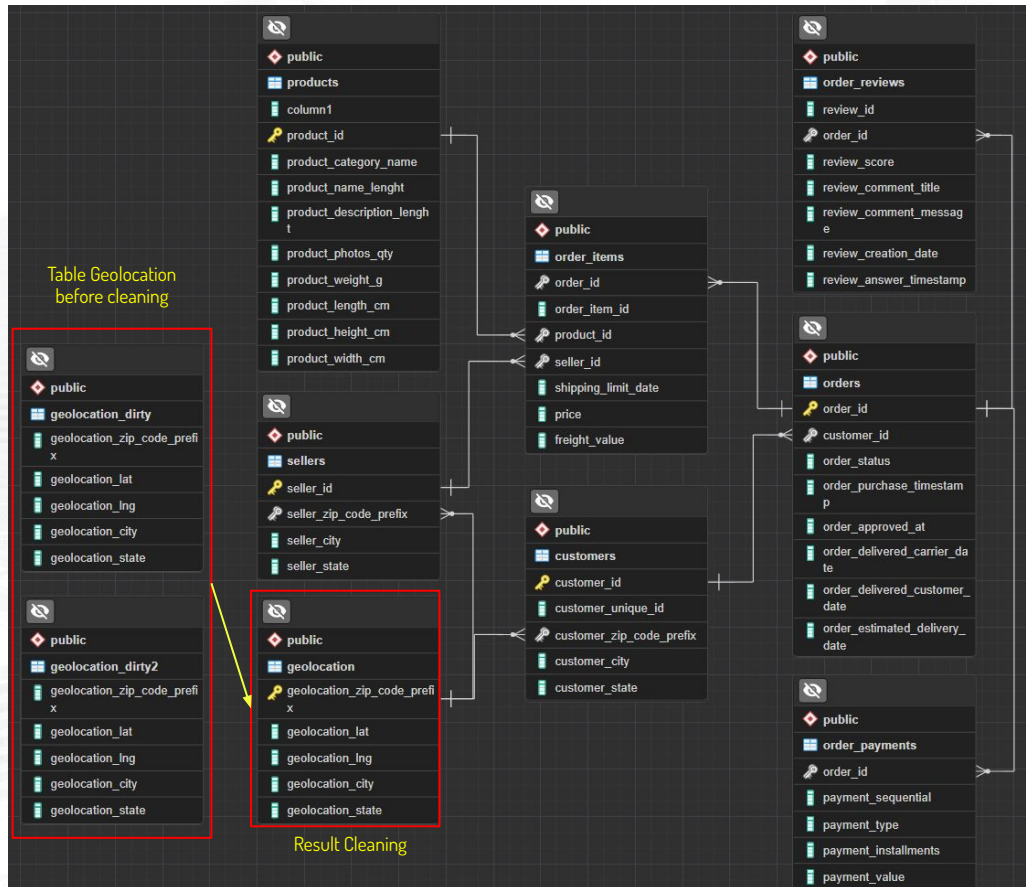
3. Entity Relationship :

Next, create an ERD by specifying the Primary key and Foreign key for each table. Then the relationship between the keys is connected by specifying the Constraint name. So that an Entity Relationship Diagram (ERD) is formed as shown in the picture.

```
-- products -> order_items

ALTER TABLE order_items
ADD CONSTRAINT order_items_fk_product
FOREIGN KEY (product_id) REFERENCES products(product_id)
ON DELETE CASCADE ON UPDATE CASCADE;


-- sellers -> order_items

ALTER TABLE order_items
ADD CONSTRAINT order_items_fk_seller
FOREIGN KEY (seller_id) REFERENCES sellers(seller_id)
ON DELETE CASCADE ON UPDATE CASCADE;


-- orders -> order_items

ALTER TABLE order_items
ADD CONSTRAINT order_items_fk_order
FOREIGN KEY (order_id) REFERENCES orders(order_id)
ON DELETE CASCADE ON UPDATE CASCADE;
```

View or Download ERD here
The full query can be seen here

Annual Customer Activity Growth Analysis

Overview :

The increase in growth customers is a reflection of the effectiveness in sales. That's why customer growth is one of the main goals of an e-commerce company. It is not only essential to grow our customers but also to retain them. One of the metrics used to measure eCommerce business performance is customer activity that interacts on the eCommerce platform. In this section, we will analyze several metrics related to customer activity such as the number of active customers, the number of new customers, the number of customers who make repeat orders and also the average transactions made by customers every year. Analyze whether the eCommerce business performance in terms of customer activity has grown, stagnated or even decreased in the past year.

The following is a yearly table, the combined results of the Monthly Active Users (MAU) table, Number of New Customers, Number of customers who make repeat orders, and the average order each year.

year double precision	avg_monthly_active_user numeric	new_customers bigint	repeat_customers bigint	avg_num_orders numeric
2016	108	326	3	1.01
2017	3694	43708	1256	1.03
2018	5338	52062	1167	1.02

The full query can be seen here

Annual Customer Activity Growth Analysis

1. Average Monthly Active User (MAU) per year

Displays the average number of monthly active users (monthly active users) for each year

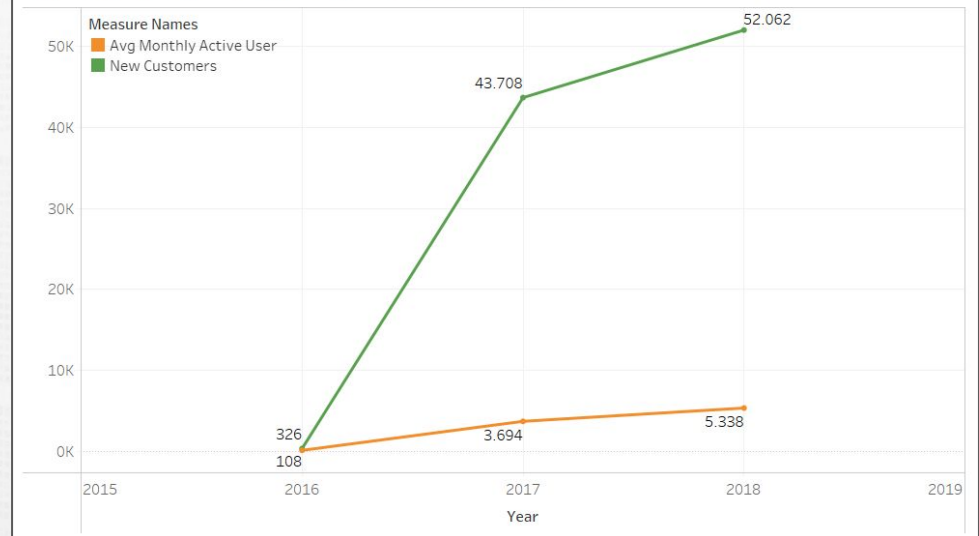| year double precision 🔒	avg_monthly_active_user numeric 🔒
2016	108
2017	3694
2018	5338

2. Total new customers per year

Displays the number of new customers in each year

year double precision 🔒	new_customers bigint 🔒
2016	326
2017	43708
2018	52062

Monthly Average Users (MAU) & New Customers
Monthly active user & new customers increased in last 3 years

Observation :
- Monthly Active Users have increased every year, adding up to thousands of customers
- There was a significant increase in 2017, but it could be because the data is only available for 4 months starting in September. Then it increased again in 2018
- This shows that e-commerce performance is related to having good engagement to attract customer interest, which is indicated by an increase in the average active customer. Also promotion of branding awareness
- good ecommerce so new customers are interested in starting orders, besides that it is possible, there are promos or vouchers for new customers which are also interesting.

The full query can be seen here

Annual Customer Activity Growth Analysis

Monthly Active User

Active user keeps on hitting its peak on November as the holiday season starts.
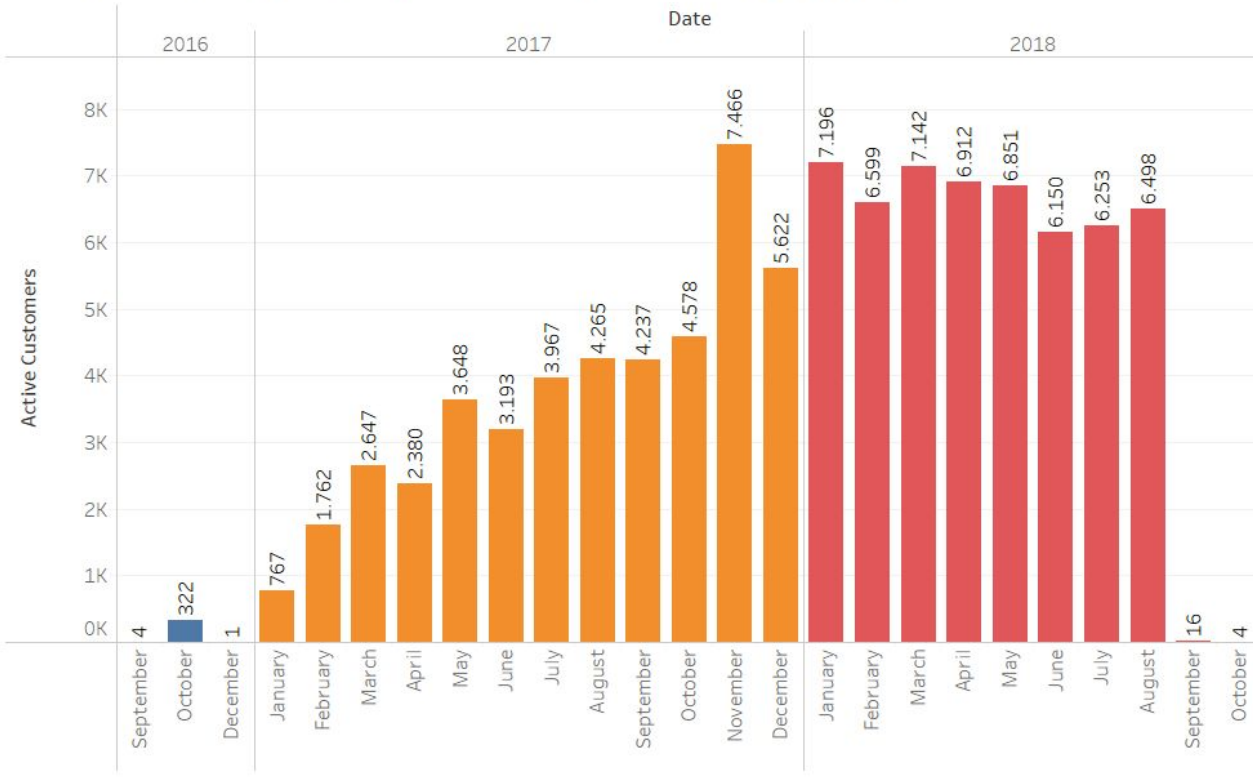
Observation :

- The fact that the data for 2016 starts from 4 months, starting from September
- In 2017, every month experienced an increase, most in November.
- This increase in November is likely due to Black Friday. Black Friday is the colloquial term for the Friday after Thanksgiving. It traditionally marks the start of the Christmas shopping season. Many stores offer highly promoted sales at discounted prices and are often open early, sometimes as late as midnight.
- But 2018 looks a bit stagnant for the increase in active users

The full query can be seen here

Annual Customer Activity Growth Analysis

3. The number of customers who make repeat orders per year

Displays the number of customers who make purchases more than once (repeat orders) in each year

| year double precision 🔒	repeat_customers 🔒 bigint
2016	3
2017	1256
2018	1167

4. Average order frequency for each year

Displays the average number of orders made by customers for each year

year double precision 🔒	avg_num_orders 🔒 numeric
2016	1.01
2017	1.03
2018	1.02

Increase in Repeat Customers

There was an increasing trend from 2017 to 2018, However, in 2018 there was a slight decrease

Measure Names
- Repeat Customers
- Avg Num Orders
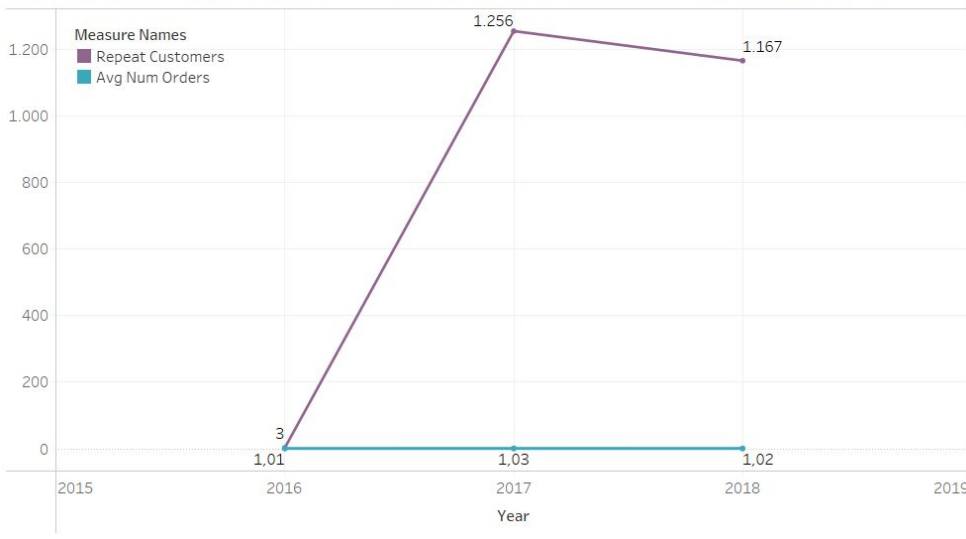
Observation :
- There was a significant increase in repeat orders in 2017. However, there was a decrease in 2018 as many as 100 customers
- From the average value of the number of orders it can be seen that most people order only once in the last 3 years

The full query can be seen here

Additional (continue...)

- It is known that the data only starts at 4 months in 2016 in September. So many new customers have just bought or tried the products in this e-commerce. This can be seen from the increase in new customers and MAU (Monthly Active Users) from 2016 - 2018. However, there is a difference in the number of repeat orders, that in 2016 - 2017 there has been a significant increase, but in 2017 - 2018 it has decreased. So that further analysis can be carried out on the causes of the decline in repeat orders in the last year and also in terms of existing business methods.

- The small average number of orders for each customer may be due to the lack of programs or maintenance from business marketing. This can be improved by holding a loyalty program, introducing new products and attracting potential customers for each new product launched.

Overview :

E-Commerce business performance is of course very closely related to the products available in it. Analyzing product quality in eCommerce can provide decisions for developing a more effective business. By analyzing product quality in e commerce, we can gain insight on how to better develop our business. Product quality is important because it influences a company's success and helps build its reputation in the customer market. When a company can create high-quality products that continuously meet customer demands, it can result in a higher return on investment and increased revenue. In this section, we will analyze product categories that have a positive impact (in terms of revenue metrics) as well as product categories that have a negative impact (in terms of the number of order cancellation metrics) for the company each year.

The following is a yearly table, the combined results of the revenue, number of canceled orders, top categories that generate the most revenue and categories that experience the most canceled orders per year

year double precision	top_product_category_revenue character varying (50)	top_category_revenue double precision	total_revenue_year double precision	top_product_category_canceled character varying (50)	top_category_num_canceled bigint	total_canceled_orders_year bigint
2016	furniture_decor	6899.35	46653.74000000004	toys	3	26
2017	bed_bath_table	580949.200000002	6921535.239999719	sports_leisure	25	265
2018	health_beauty	866810.3399999987	8451584.769999959	health_beauty	27	334

The full query can be seen here

Annual Product Category Quality Analysis

1. Table of revenue per year

Create a table that contains total company revenue information for each year

| year double precision	revenue double precision
2016	46653.74000000004
2017	6921535.239999719
2018	8451584.769999959

2. Table of the number of canceled orders per year

Create a table that contains information on the total number of cancel orders for each year

year double precision	total_cancel bigint
2016	26
2017	265
2018	334

Total Revenue and Total Canceled Orders
In line with each year's increase in revenue and canceled orders
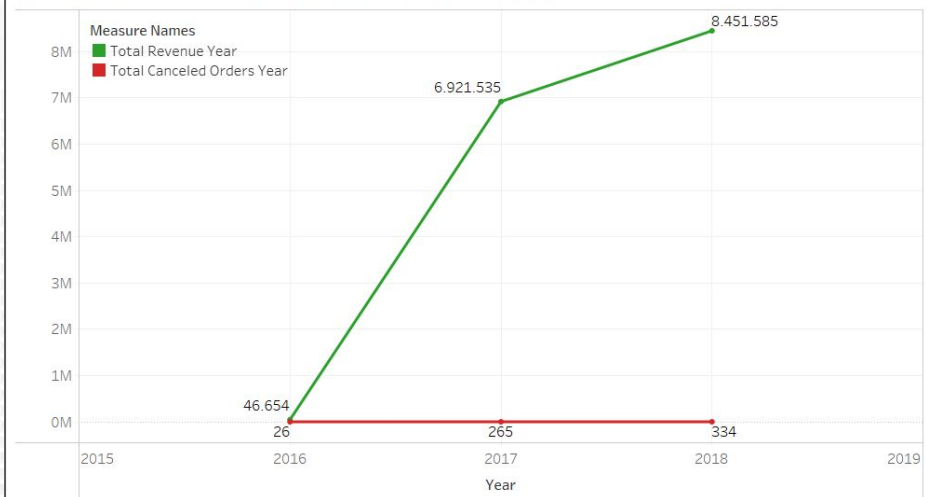
Observation :
- There is an increase in revenue every year. although the previous data, namely repeat orders, showed a decrease from 2017 to 2018. This could be due to a sharp increase in the number of new customers each year
- There was an increase in canceled orders from 2016 to 2018. However, the value is not on a large scale, still in the hundreds compared to the thousands. It is better if additional data is needed regarding the review of reasons for canceling orders to avoid an increase in the number of cancels in the following year

The full query can be seen here

Annual Product Category Quality Analysis

3. Table of the top categories that generate the largest revenue per year

Create a table containing the product category names that provide the highest total revenue for each year

year double precision 🔒	product_category_name character varying (50) 🔒	revenue double precision 🔒
2016	furniture_decor	6899.349999999999
2017	bed_bath_table	580949.2000000025
2018	health_beauty	866810.3399999987

4. Table of categories that experienced the most canceled orders per year

Create a table containing the names of the product categories that have the highest number of canceled orders for each year

year double precision 🔒	product_category_name character varying (50) 🔒	total_cancel bigint 🔒
2016	toys	3
2017	sports_leisure	25
2018	health_beauty	27

Top Product Revenue & Canceled by Year
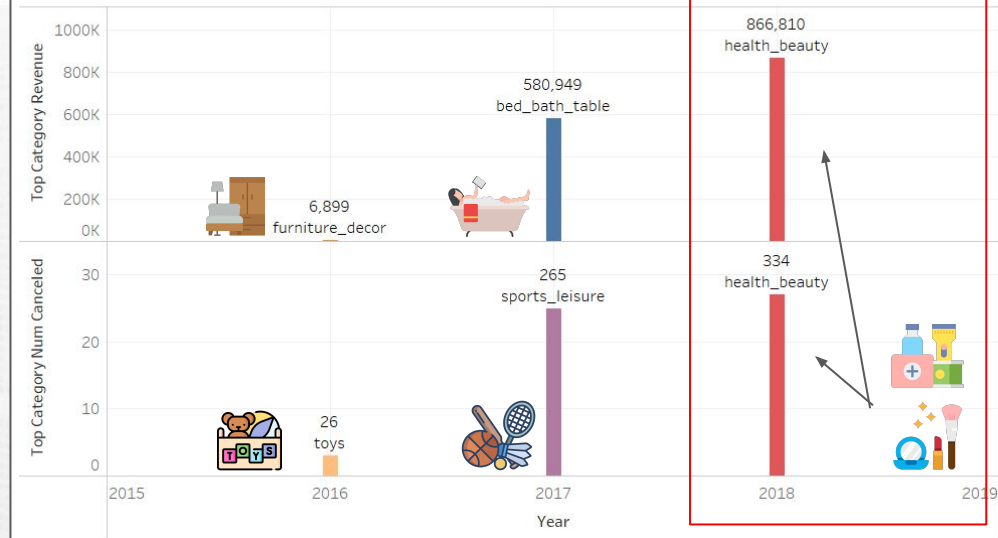
Observation :

- It can be seen that the types of product categories, both top revenue and top canceled, all have different types each year
- However, there was an interesting thing in 2018, it was found that the Health Beauty category was in the same position as the top highest revenue and canceled. This could be because the most orders were in the health and beauty categories

The full query can be seen here

Analysis of Annual Payment Type Usage

Overview :

Effective income analysis requires consideration of the payment method and its impact on customer behavior. Research indicates that customers are likely to abandon their purchase if their preferred payment option is unavailable. In e-commerce, offering an open payment system with diverse options is essential. Analyzing the performance of each payment type will provide valuable insights, enabling strategic partnerships with payment providers to enhance customer satisfaction and optimize revenue streams.

The table below presents the annual data, combining the total number of transactions for each payment type and the corresponding number of users for each year.

payment_type character varying (50)	num_of_usage bigint
credit_card	76795
boleto	19784
voucher	5775
debit_card	1529
not_defined	3

payment_type character varying (50)	year_2016 bigint	year_2017 bigint	year_2018 bigint
credit_card	258	34568	41969
boleto	63	9508	10213
voucher	23	3027	2725
debit_card	2	422	1105
not_defined	0	0	3

The full query can be seen here

Analysis of Annual Payment Type Usage

1. Total usage of each type of payment all time

Displays the total usage of each type of payment at all time, sorted from the favorite

payment_type character varying (50) 🔒	num_of_usage bigint 🔒
credit_card	76795
boleto	19784
voucher	5775
debit_card	1529
not_defined	3

Payment Type Used for All Time
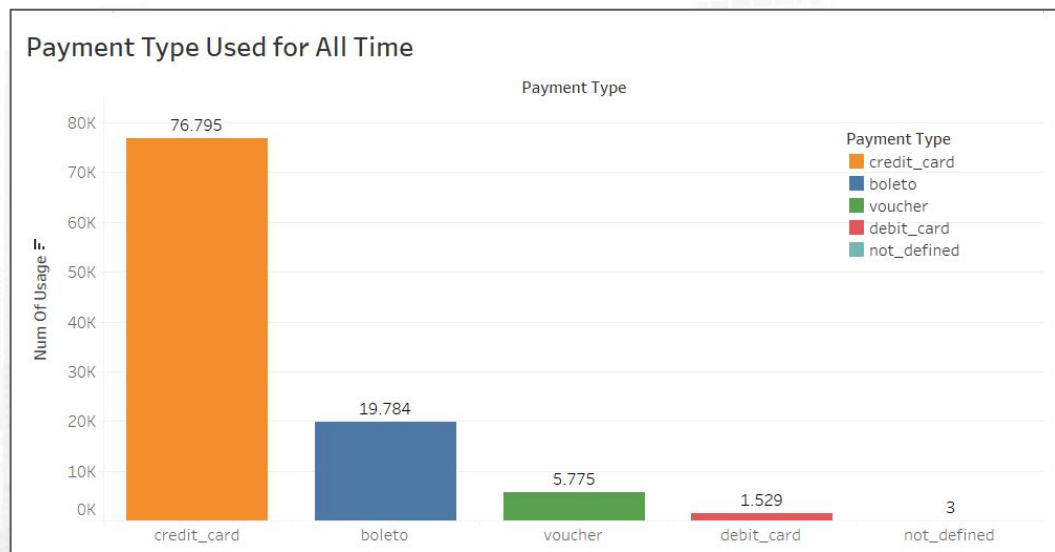
Observation :
- There are 4 common payment methods, Credit Card, Boleto, Voucher, Debit Card
- The use of credit cards as a means of payment is the most compared to other types of payments. By looking at this, companies can work with credit card provider banks to provide more attractive promotions. Further analysis can be carried out on customer behavior in using credit cards, such as the payment credit type chosen, what product categories are usually purchased with credit cards, etc.
- Because many customers use credit cards, boleto and vouchers, bonuses or promotions can be given that target the most favorite payment types.

The full query can be seen here

Analysis of Annual Payment Type Usage

2. Details of the amount used for each type of payment for each year

Displays detailed information on the amount of usage for each type of payment for each year

payment_type character varying (50) 🔒	year_2016 bigint 🔒	year_2017 bigint 🔒	year_2018 bigint 🔒
credit_card	258	34568	41969
boleto	63	9508	10213
voucher	23	3027	2725
debit_card	2	422	1105
not_defined	0	0	3

Observation :
- Furthermore, for each type of payment tends to increase significantly every year.
- However, payments using vouchers have decreased in 2018. This could be due to a reduction in promotions using vouchers.
- Further analysis can be carried out by confirming with other departments, such as Marketing, Analyst or Business Development.

Payment Type Usage each Year

Most used type of payment is Credit Card, All have increased usage every year

The full query can be seen here
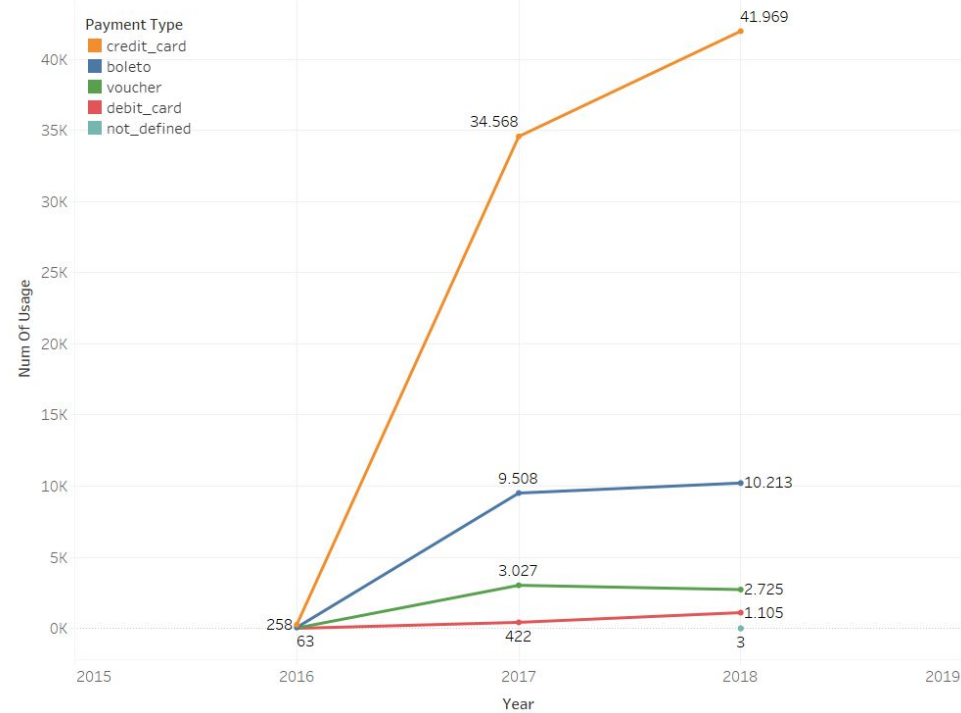
Summary

- Customer activity experienced growth in every metric from 2016 to 2018, including an increase in the number of new customers and Monthly Active Users (MAU). However, repeat customer orders remained stagnant during this period. Moreover, the average number of orders made by customers was only once.
- The analysis of Product Category Quality revealed a consistent growth in the company's total revenue each year. Interestingly, the most canceled product categories and best selling orders changed annually. Notably, the health beauty category emerged as both the most sold and the most canceled product category.
- Additionally, each type of payment showed a significant increase year over year. Credit cards were the most commonly used payment method from 2016 to 2018. However, payments made using vouchers experienced a decline in 2018.