Analyzing eCommerce Business Performance with SQL

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

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.**

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 :

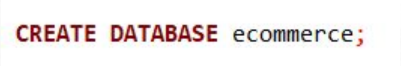
The dataset used is provided by Rakamin Academy, Brazilian E-Commerce Public Dataset by Olist. Or it can be accessed on the kaggle dataset. This is a Brazilian e-commerce public dataset of orders made at the Olist Store. The dataset has information on 100k orders from 2016 to 2018 made at multiple marketplaces 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.

Data Preparation

1. Creating Database & Table :

- **Create Database**

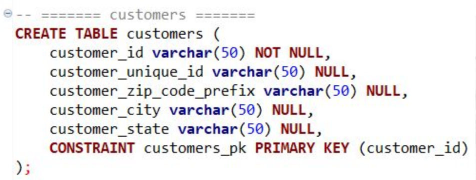
  We can create an *ecommerce* database by using the CREATE DATABASE query or by using the pgAdmin 4 GUI



```
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)
);
```

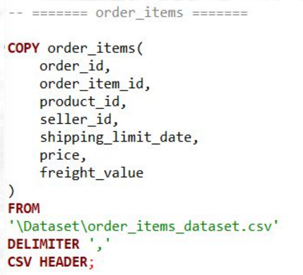
List of Tables Created

customers
geolocation
order_items
order_payments
order_reviews
orders
products
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;
```

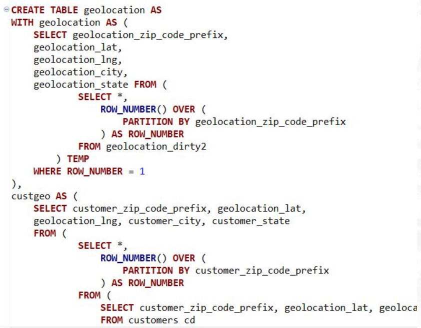
- **Pre-Cleaning Data**

  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

Data Preparation

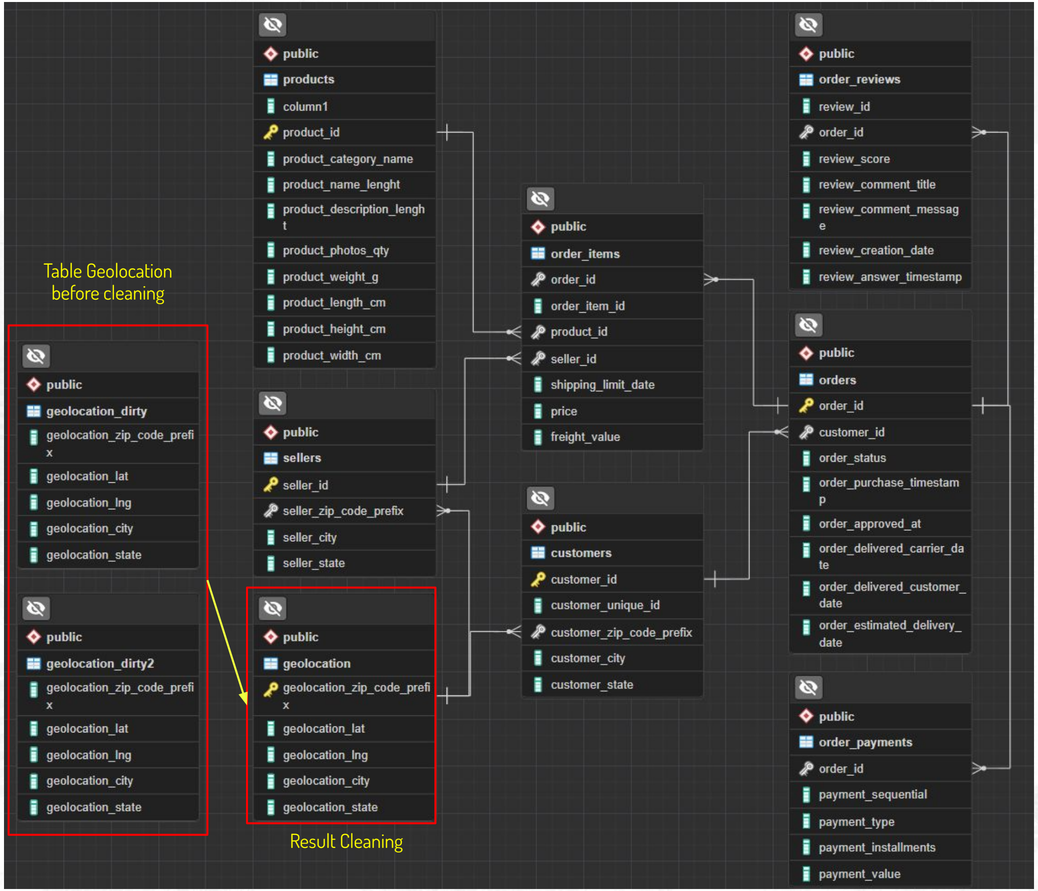
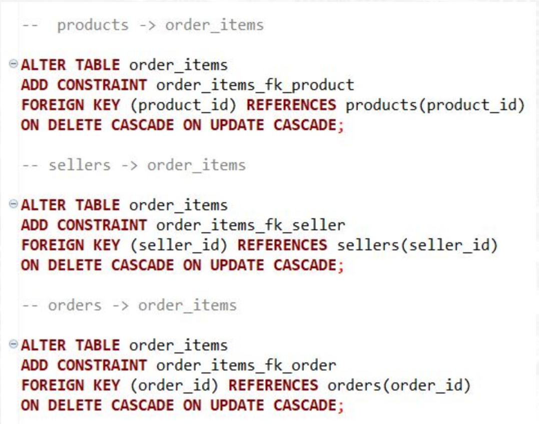


Table Geolocation before cleaning

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.



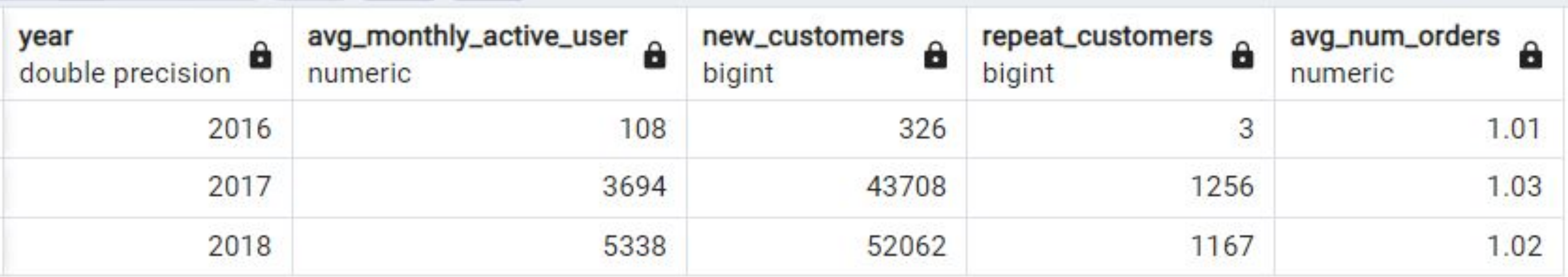
```
--  products -> order_items

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

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.



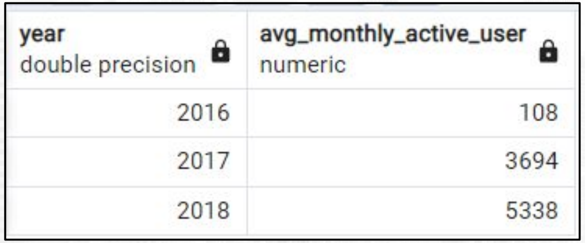
year double precision	avg_monthly_active_user numeric	new_customers bigint	repeat_customers bigint	avg_num_orders numeric
2016	108	326	3	1.01
2017	3694	43708	1256	1.03
2018	5338	52062	1167	1.02

The full query can be seen here

Annual Customer Activity Growth Analysis

1. Average Monthly Active User (MAU) per year

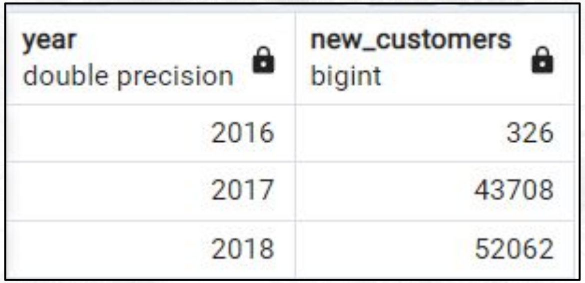
Displays the average number of monthly active users (monthly active users) for each year



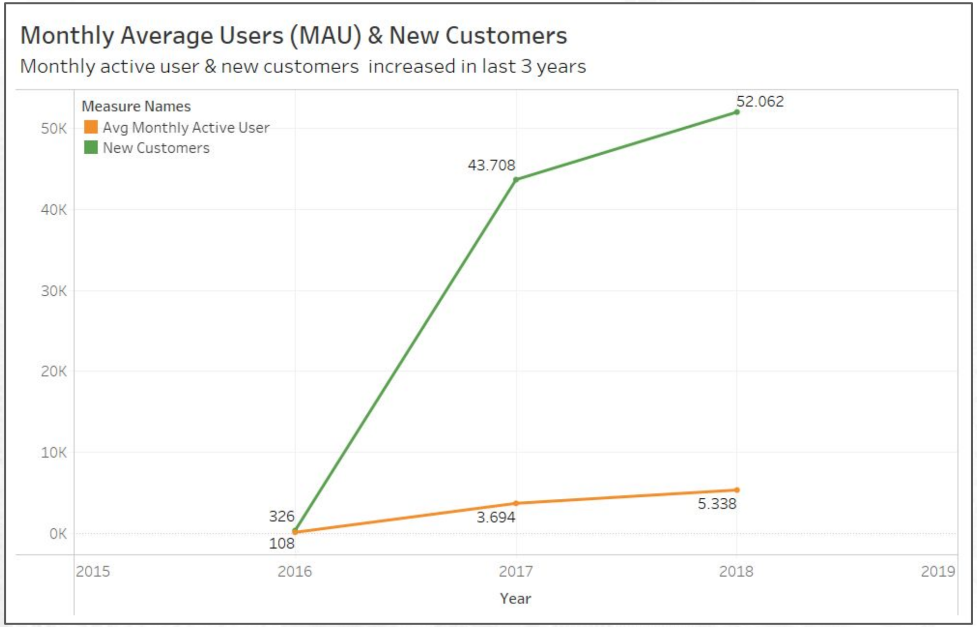
year double precision	avg_monthly_active_user numeric
2016	108
2017	3694
2018	5338

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 :

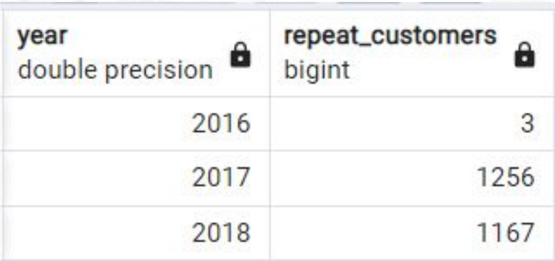
- Monthly Active User
- Ada signifikan kenaikan pada tahun 2017, namun bisa jadi dikarenakan data yang dimiliki hanya pada 4 bulan di mulai September. Kemudian mengalami kenaikan kembali di tahun 2018
-

The full query can be seen here

Annual Customer Activity Growth Analysis

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

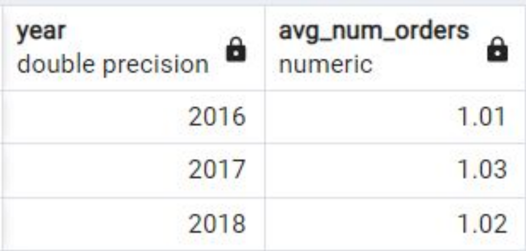
Displays the number of customers who make purchases more than once (repeat orders) in each year



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

The full query can be seen here

Annual Product Category Quality Analysis

- Masukkan master tabel yang berisi informasi :
  - revenue per tahun
  - jumlah cancel order per tahun
  - top kategori yang menghasilkan revenue terbesar per tahun
  - kategori yang mengalami cancel order terbanyak per tahun

  (Hasil dari langkah 5)

- Silakan tambahkan visualisasi lainnya bisa berupa gambar, grafik, dsb untuk informasi yang menurutmu penting mengenai **kualitas produk** berdasarkan tabel di atas kemudian tulislah interpretasi-mu.

- Tambahkan query yang kamu gunakan pada tugas ini, ke dalam file doc yang sudah dibuat pada tugas sebelumnya, dan jangan lupa sematkan link-nya di pojok kanan bawah

- Maximal 4 slide.

Query selengkapnya dapat dilihat disini

Analysis of Annual Payment Type Usage

- Masukkan tabel yang berisi informasi jumlah penggunaan masing-masing tipe pembayaran untuk setiap tahun

- Silakan tambahkan visualisasi lainnya bisa berupa gambar, grafik, dsb untuk informasi yang menurutmu penting mengenai **tipe pembayaran** berdasarkan tabel di atas kemudian tulislah interpretasi-mu.

- Tambahkan query yang kamu gunakan pada tugas ini, ke dalam file doc yang sudah dibuat pada tugas sebelumnya, dan jangan lupa sematkan link-nya di pojok kanan bawah

- Maximal 4 slide.

Query selengkapnya dapat dilihat disini