Analyzing eCommerce Business Performance with SQL

Created by:
Kenneth Wahyudi, S.Si.

kennethwahyudi48@gmail.com
https://www.linkedin.com/in/kenneth-wahyudi-80b886209/

I am an aspiring data scientist with interest in machine learning and paddling through the data lake. I have experiences as a research intern in BATAN, and several data science and machine learning projects under my belt. Specifically intermediate to advanced knowledge in Exploratory Data Analysis, Data Pre-processing, Supervised & Unsupervised Learning, as well as Data Visualization and Storytelling. And I am always looking forward to exploring new and uncharted territories that might allow myself to grow and improve.

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

Overview

Overview

In a company, measuring business performance is of utmost importance to track, monitor, and assess the success or failure of various business processes. Therefore, this paper will analyze the business performance of an eCommerce company, taking into consideration several business metrics, namely customer growth, product quality, and payment methods.

All results are outputs from queries executed in PostgreSQL's management tool Pg Admin 4.

Link to the datasets used in this project can be found here :
https://drive.google.com/drive/folders/1PCdSDyv1dYNZYDmznSKVUgfI0jhP1Zyq?usp=sharing

Data Preparation

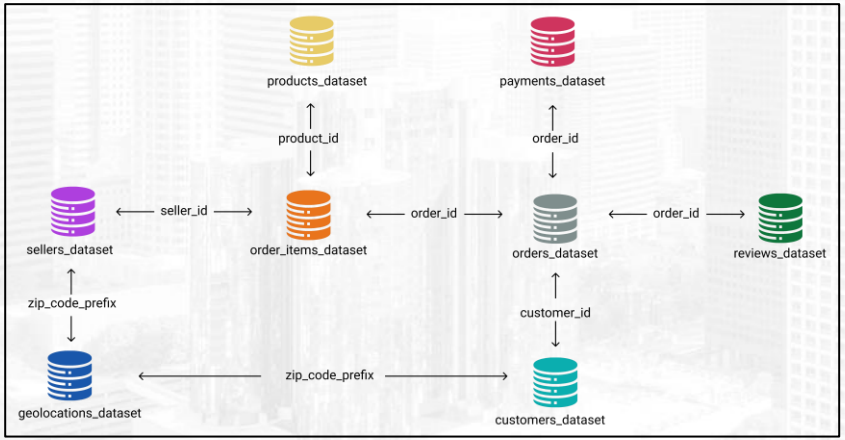
Data Preparation

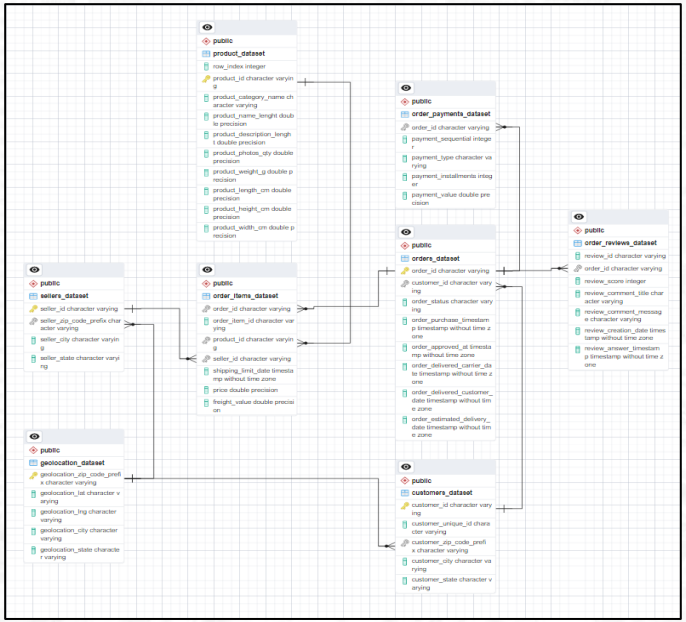
I first created a database in the Pg Admin 4 named ‘ecommerce’. Then, create 8 tables from the datasets : customers_dataset.csv, geolocation_dataset.csv, order_items_dataset.csv, order_payments_dataset.csv, order_reviews_dataset.csv, orders_dataset.csv, product_dataset.csv, sellers_dataset.csv. With the product_id column from the product_dataset table as a primary key, the order_id column from the orders_dataset table as a primary key, the seller_id column from the sellers_dataset table as a primary key, and the customer_id column from the customers_dataset table as a primary key.

Looking at the relationship requirement between the tables, the zip_code_prefix column in the geolocation_dataset table needs to be unique in order to establish the proper relationships. However, the column contains duplicate values. And not only that, some zip_code_prefix values from the sellers_dataset and customers_dataset tables do not exist in the zip_code_prefix column within the geolocation_dataset table. Therefore, data cleaning must be performed that inputs the non-existing zip_code_prefix values into the geolocation_dataset table, and then dropping the duplicate values so that each and every value within the zip_code_prefix column in the geolocation_dataset table is unique. And finally setting that column as a primary key.

Data Preparation

After all that, only then can I create the relationships by establishing a foreign key constraint into the corresponding tables. By obeying the relationship requirement on the left, I created the Entity Relationship Diagram (ERD) using Pg Admin 4 on the right :





The complete query, ERD in png format, and ERD in pgerd format can be accessed in the following link :
https://drive.google.com/drive/folders/1jT952KR6TyS2r56I1UTezjyxUOfo9dCq?usp=sharing

Annual Customer Activity Growth Analysis

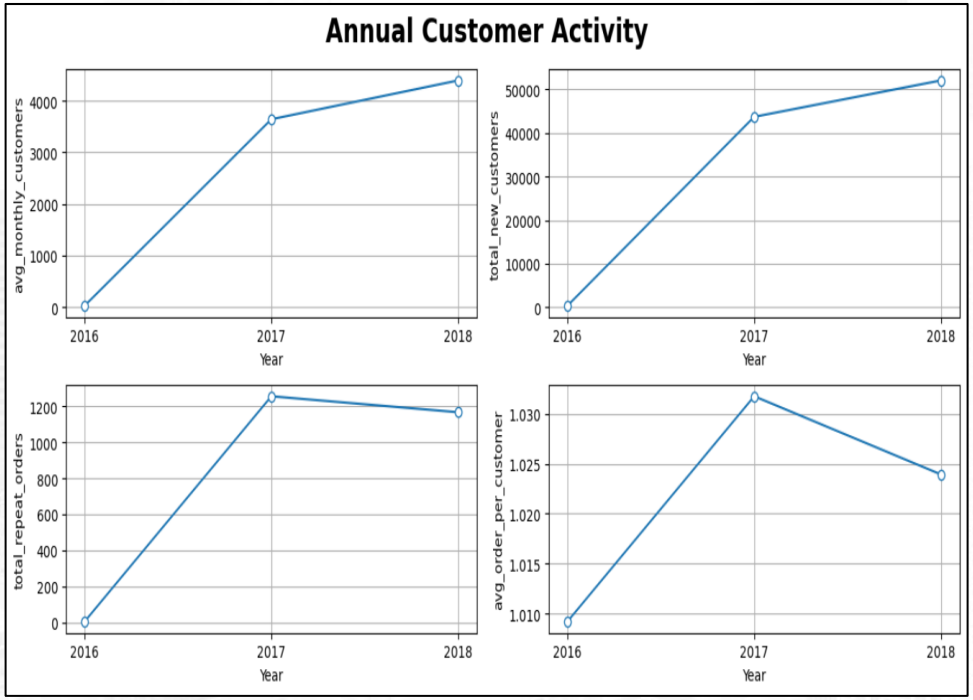
Annual Customer Activity Growth Analysis

year	avg_monthly_customers	total_new_customers	total_repeat_orders	avg_order_per_customer
2016	27.16666667	326	3	1.009202454
2017	3642.75	43708	1256	1.031752568
2018	4395.75	52062	1167	1.023924624

Above is the table that is as a result of 5 steps of queries :

1. Displaying the average monthly customers per year
2. Displaying the number of new customers per year
3. Displaying the amount of customers that do repeat orders per year
4. Displaying the average orders per customer each year
5. Joining to create a master table that houses all the information

Annual Customer Activity Growth Analysis



The complete query in doc format, output file in csv format, and chart in png format can be accessed in the following link : https://drive.google.com/drive/folders/1d0O4lsOrMRaIpWaPhYPMYf6A04qK9YHY?usp=sharing

As you can see from the chart on the left, all of the variables follow a similar trend. With average monthly customers & total new customers are on the rise, but their ascent is slowing down. However, for the total repeat orders & average order per customer, everything seems to be the same until the year 2017. When the numbers actually decline, with average order per customer actually declining quite steeply.

Further analysis should be done on the latter two, as to why they are on the decline. And prescriptive efforts may have to be implemented in order to increase the values in those two aspects such as a better recommender system, discounts & sales, targeted marketing, etc.

Annual Product Category Quality Analysis

Annual Product Category Quality Analysis

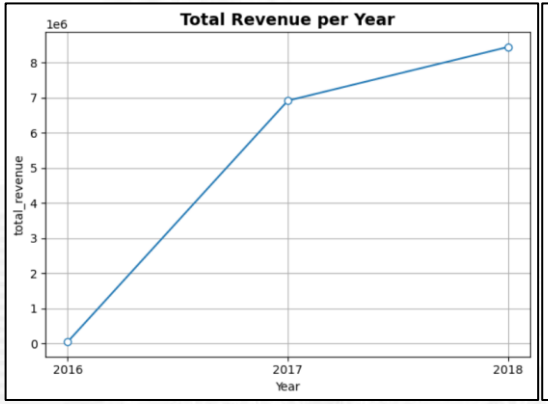
year	total_revenue	total_canceled_orders	top_revenue_category	total_category_revenue	top_canceled_category	total_category_canceled_orders
2016	46653.74	26	furniture_decor	6899.35	toys	3
2017	6921535.24	265	bed_bath_table	580949.2	sports_leisure	25
2018	8451584.77	334	health_beauty	866810.34	health_beauty	27

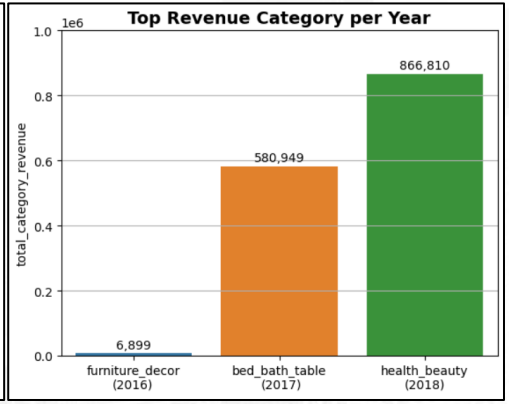
Above is the table that is as a result of 5 steps of queries :

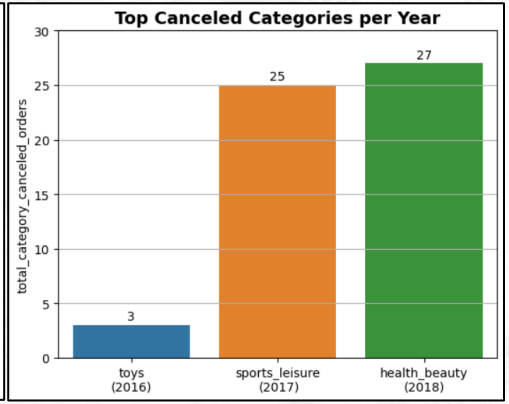
1. Displaying the total revenue each year
2. Displaying the total number of canceled orders each year
3. Displaying the categories that brings the most revenue each year
4. Displaying the categories that has the largest number of canceled orders each year
5. Creating a master table (the table above) that houses all of the information from the above steps of queries.

The complete query in doc format, charts in png format, and table in csv format can be accessed in the following link :
https://drive.google.com/drive/folders/1WyUMh8CJ1qyQxotyuSGY657Dl6KSa2PA?usp=sharing

Annual Product Category Quality Analysis







From the total revenue chart, we can see that the yearly revenue sky-rocketed in 2017. However, that growth slows down significantly in 2018. The same goes for the amount of canceled orders. With more orders resulting in more canceled orders, visible in the spike in 2017. Top revenue category & Top canceled category follow the same trend. However, in 2016 & 2017, the top revenue category is different than the top canceled category. Which may suggest that the top revenue category in those years were performing relatively well compared to the top revenue category in 2018.

Analysis of Annual Payment Type Usage

Analysis of Annual Payment Type Usage

Rakamin Academy

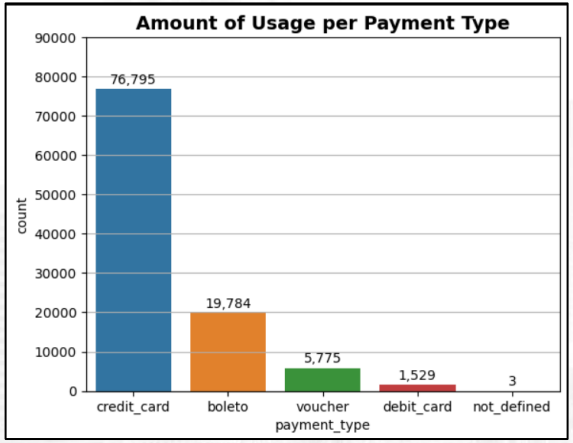
Below is the table that displays the amount of usage per each payment_type :

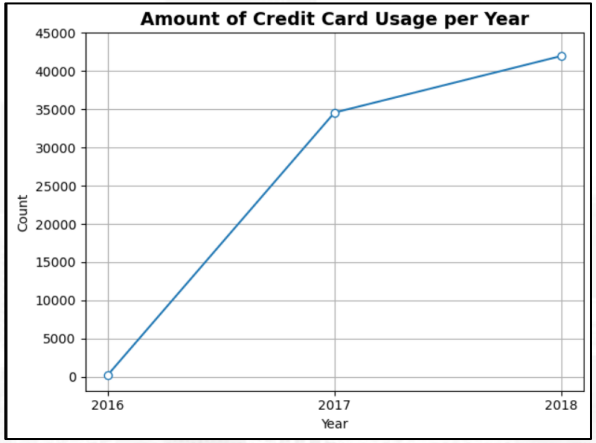
credit_card	boleto	voucher	debit_card	not_defined
76795	19784	5775	1529	3

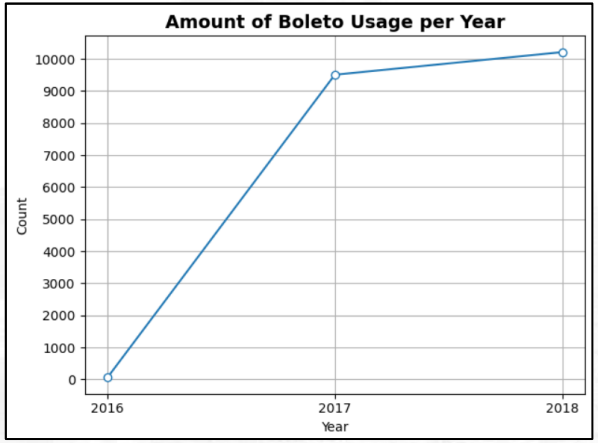
Below is the table that displays the amount of usage by each payment_type per year :

year	credit_card	boleto	voucher	debit_card	not_defined
2016	258	63	23	2	0
2017	34568	9508	3027	422	0
2018	41969	10213	2725	1105	3

Analysis of Annual Payment Type Usage

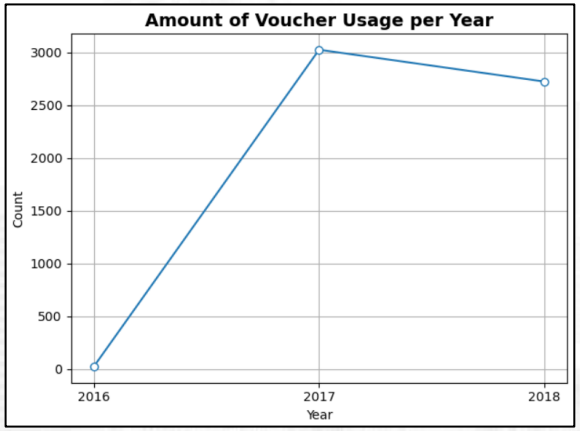


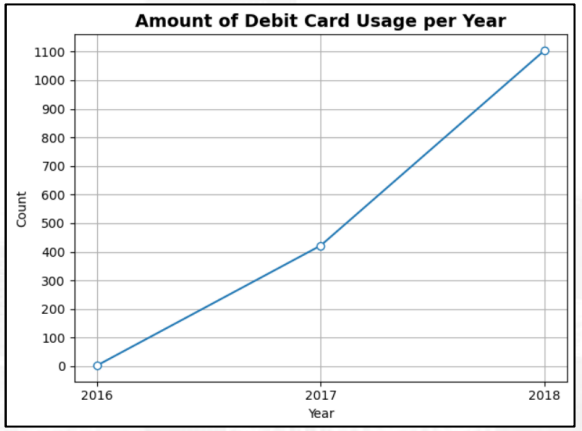


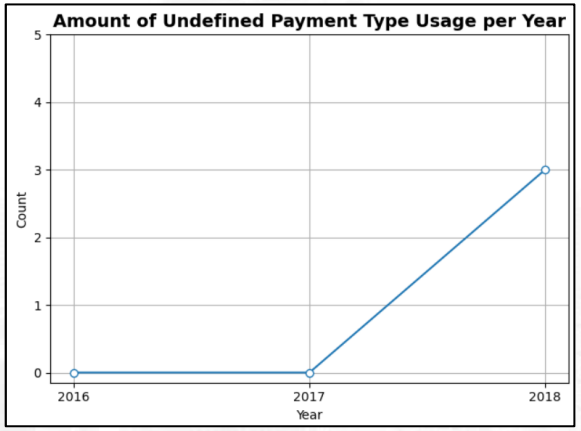


From the chart above, we can see that credit card is by far the most popular method of payment. And in the other charts, we can see that not only is credit card the most popular method, but its growth is also the quite promising. With its growth from 2017 to 2018 being a lot better than boleto that slows down after 2017.

Analysis of Annual Payment Type Usage







From the charts above, it is visible that the voucher payment type is in decline after the 2017 growth boom. Whilst the growth of the usage of debit cards are the best compared to all the other payment types. One interesting note is that in 2018, there are 3 orders with undefined payment types. It is possible to investigate further what kind of payment type it is, especially if it grows exponentially in the following years, being a significant player in the game.

The complete query in doc format, charts in png format, and output tables in csv format can be accessed in the following link : https://drive.google.com/drive/folders/1ycToT7_-QCo9QQ11w75aaYNHHABRm4sG?usp=sharing