Cohort Analysis with SQL

STUDY CASE: THELOOK ECOMMERCE (GOOGLE BIGQUERY DATASETS)

by Raden Ghifary Nurrachman

about theLook Ecommerce

TheLook is an e-commerce clothing store developed by the Google Looker team. This E-commerce data is hosted on Google Big Query and contains information about distribution centers, customers or user, products, orders, logistics, and web events. The contents of this dataset are provided to industry practitioners for the purpose of product discovery, testing, and evaluation.
At the time of this analysis, the e-commerce store recorded sales transaction data from January 2019 to October 2023. But for this analysis i used the e-commerce store recorded sales transaction data from January 2022 to December 2022

Side note: Google BigQuery has several public datasets that are updated periodically and can be used to build projects for your portfolio

How many users coming back to reorder for the following months in 2022?

Scope of Problem

SQL QUERY

Cohort Orders Analysis RUN SAVE QUERY SHARE SCHEDULE

```sql
WITH cohort_customer AS(
    SELECT
     user_id AS customer_id,
     MIN(DATE(DATE_TRUNC(created_at, month))) as First_Date
    FROM `bigquery-public-data.thelook_ecommerce.orders`
    GROUP BY 1
),
user_activities AS (
    SELECT
        ord.user_id AS customer_id,
        DATE_DIFF(DATE(DATE_TRUNC(created_at, Month)), crt.First_Date,Month) AS month_number
    FROM `bigquery-public-data.thelook_ecommerce.orders` ord
    LEFT JOIN `cohort_customer` AS crt
    ON ord.user_id = crt.customer_id
    WHERE EXTRACT(YEAR from crt.First_Date) = 2022 AND EXTRACT(YEAR FROM ord.created_at) = 2022
    GROUP BY 1,2
),
cohort_size AS (
    SELECT
        First_Date,
        COUNT(First_Date) AS number_user
    FROM cohort_customer
    GROUP BY 1
),
retention_table AS (
    SELECT
        c.First_Date,
        a.month_number,
        COUNT(First_Date) AS number_user
    FROM user_activities AS a
    LEFT JOIN cohort_customer AS c
    ON a.customer_id = c.customer_id
    GROUP BY 1,2
)
SELECT
  b.First_Date,
  s.number_user AS cohort_size,
  b.month_number,
  b.number_user AS total_users,
  CAST(b.number_user AS decimal)/s.number_user AS percentage
FROM retention_table AS b
LEFT JOIN cohort_size AS s
ON b.First_Date = s.First_Date
WHERE b.First_Date IS NOT NULL
ORDER BY 1,3;
```

Cohort Table

SUM of percentage		month_number											
First_Date	cohort_size	0	1	2	3	4	5	6	7	8	9	10	11
2022-01-01	1545	100.00%	3.56%	2.72%	2.52%	3.24%	3.17%	3.17%	3.24%	3.37%	3.69%	2.59%	3.04%
2022-02-01	1428	100.00%	3.85%	3.64%	3.36%	2.87%	2.59%	3.50%	3.22%	2.52%	3.08%	3.36%	
2022-03-01	1571	100.00%	2.61%	3.88%	2.48%	2.67%	3.69%	3.69%	3.18%	2.86%	3.37%		
2022-04-01	1698	100.00%	2.77%	3.77%	3.00%	2.77%	2.71%	3.36%	3.18%	3.00%			
2022-05-01	1727	100.00%	3.13%	2.90%	3.30%	3.88%	3.53%	3.65%	3.94%				
2022-06-01	1731	100.00%	3.00%	3.87%	2.95%	3.76%	2.95%	3.58%					
2022-07-01	1814	100.00%	3.31%	2.92%	4.30%	3.75%	3.97%						
2022-08-01	1946	100.00%	2.77%	4.68%	3.96%	3.91%							
2022-09-01	1966	100.00%	4.83%	4.02%	3.87%								
2022-10-01	2080	100.00%	3.65%	4.47%									
2022-11-01	2123	100.00%	4.19%										
2022-12-01	2272	100.00%											
Grand Total	1825.08	100.00%	3.43%	3.69%	3.30%	3.35%	3.23%	3.49%	3.35%	2.94%	3.38%	2.98%	3.04%

In this table we can analyze and conclude, there is a **massive drop in retention rates** entering the first month since they first placed an order. It can be seen in the cohort table above that the average user retention rate **drops drastically from 100% to ±3% in the next month**, it can be concluded that users are returning less and less since they first placed an order. However, each cohort's retention rate has **increased and decreased steadily at 2.94% – 3.69% every month from the beginning of the user's order**. We can see that the **August to November cohort has a higher retention rate** than the previous months, where the **retention rate value is at 2.77% – 4.83% in the month after the first user places an order**. Then we can find out that the **September cohort has the largest retention rate** after its first order in the first month after its first order **at 4.83%**, and the **March cohort has the largest decrease in user retention rate** with **2.48% in its third month**. **Although this retention value is a very small value that requires improvement and evaluation of a more effective business strategy**.

Hypothesis

What could possibly the main reason we have so low retention rate?

I would check the **status of order_items** data and check the **total** number of orders that exist in each status, at the time is the reason why a lot of cohort size have drop retention on the first month after their order.

```
SELECT
ord.status,
COUNT(ord.status) total_status
FROM `bigquery-public-data.thelook_ecommerce.order_items` ord
WHERE EXTRACT (YEAR FROM created_at) = 2022
GROUP BY 1;
```

status	total_status	
Cancelled	7292	15.09%
Processing	9475	19.61%
Complete	12011	24.85%
Shipped	14694	30.41%
Returned	4854	10.04%
	48326	

Hypothesis

The hypothesis that I get is by analyzing the data obtained from various tables on thelook ecomerce dataset, where a conclusion is obtained from a large hole that can be seen in the **item order's status** column. That there are many customer item order statuses that are **canceled** and **returned**, with a **canceled percentage of 15.09%** and a **returned percentage of 10.04%**, that means **25.14% of the items seems failed to be sold**, this percentage value is very huge seeing from **the orders completed in 2022 only 24.85%** of which the rest are still in process and shipping.
This indicates that the seller canceled or sent items that were not supposed to be, which can be a red flag for potential customers to return to the store or even for the reputation of thelook ecommerce it self.

status	total_status	
Cancelled	7292	15.09%
Processing	9475	19.61%
Complete	12011	24.85%
Shipped	14694	30.41%
Returned	4854	10.04%
	48326	

Recomendation

For seller:

- It is important for the seller to monitor its cancellation rate and better understand how to reduce it if it exceeds the required threshold.
- A high cancellation rate can be caused by many factors such as poor inventory management, product listing errors, or communication issues with customers,etc. By improving these areas, sellers can reduce the cancellation and return rates thus providing a better experience for their customers.

For thelook ecommerce:

- Thelook has the right to make regulations that benefit both buyers and sellers. So it is necessary to make rules that set limits on stores canceling orders unilaterally and conduct further analysis if a store often experiences returns from buyers.

Thank you

0857 1000 9556

ghifarynurrachman@gmail.com

Raden Ghifary