“UNDERSTANDING CUSTOMER RETENTION USING COHORT ANALYSIS”

A SQL CASE STUDY ON OLIST

BY: SHRADDHA SAHA

Objective

Customer retention is a key metric for e-commerce businesses, as acquiring new customers is often more expensive than retaining existing ones.

Here, I performed a **cohort-based customer retention analysis** on the **Olist Brazilian E-Commerce dataset** using **SQL** to understand how customer engagement evolved over time after their first purchase.

Dataset Description

This analysis is based on the **Olist Brazilian E-Commerce Public Dataset**, a real-world dataset that captures the operations of a large online marketplace in Brazil. Olist acts as an intermediary between sellers and customers, managing orders, payments, logistics, and reviews across multiple categories.

The dataset contains transactional data from 2016 to 2018 and is structured across multiple relational tables, making it well-suited for SQL-based analysis.

For this cohort analysis, the focus was primarily on the **customers** and **orders** tables. Only orders with a status of **"Delivered"** were considered to ensure that customer activity reflects completed and successful transactions.

What is Cohort Analysis?

Cohort analysis is an analytical technique where users or entities are **grouped into cohorts based on a shared characteristic or event**, and their behaviour is tracked **over time** rather than in aggregate.

In **e-commerce**, cohort analysis helps answer a simple but critical question: *"How does customer behaviour change as they age with the business?"*

Instead of looking at overall metrics like total revenue, average retention, or monthly active users—which mix customers at different lifecycle stages—cohort analysis **controls for time** and allows us to compare **like with like**.

In short: Traditional analysis answers **"What is happening?"** and Cohort analysis answers **"Who is behaving how, and when?"**

Why is cohort analysis especially important in e-commerce?

E-commerce businesses are highly dynamic:

- Customers are acquired continuously
- Product mix, pricing, discounts, logistics, and UX keep changing
- New users behave very differently from repeat or loyal users

Because of this, **aggregate metrics often hide real trends**. For example, Overall retention looks stable. But cohort analysis may reveal that **newer customers churn faster**, while older cohorts are propping up the metric.

This makes cohort analysis a **core tool for understanding customer quality, lifecycle value, and sustainability of growth**.

Different varieties of cohort analysis in e-commerce

Cohort analysis can be classified based on **how cohorts are defined**.

- **Acquisition Cohort Analysis:** Customers are grouped by **when they were acquired** (usually first purchase date/ month).
- **Behavioural Cohort Analysis:** Customers are grouped by **actions or behaviours** they performed (say, Customers who used a coupon vs didn't)
- **Campaign / Channel Cohort Analysis:** Customers are grouped by **source of acquisition** (say, Referral vs paid users).
- **Product / Category-based Cohort Analysis:** Customers are grouped by **what they bought first**.

Acquisition Cohort Analysis (needed for retention analysis)

In most e-commerce retention studies, we work with **acquisition cohorts**.

An **acquisition cohort** groups customers based on the **time-period of their first transaction**.

Customers who placed their **first order in the same month** belong to the same cohort.

Example cohorts: Jan 2017 cohort, Feb 2017 cohort

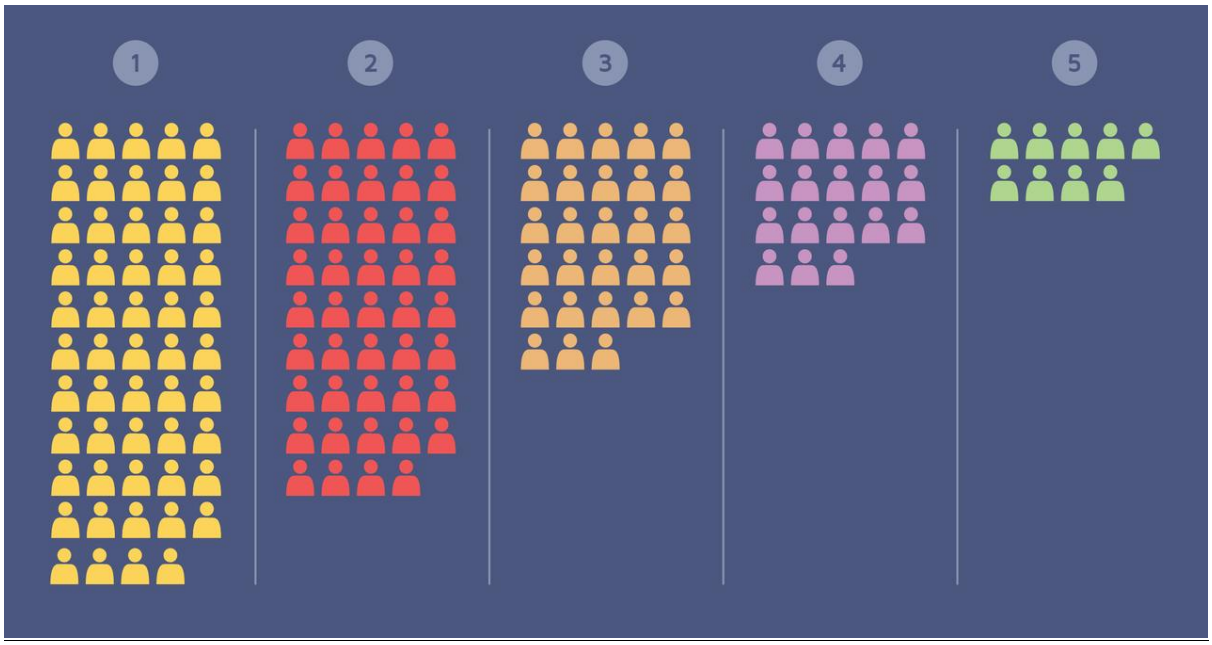
This allows us to track how customers behave **relative to their starting point**, not calendar time.

Retention rate

Retention rate measures the percentage of customers from a cohort who **return and make another purchase** after their first order.

Retention is calculated across **cohort age**, not calendar months:

- Month 0 means first purchase month
- Month 1 means 1 month after first purchase
- Month 2 means 2 months after first purchase



Why does acquisition cohort-based retention matter to businesses?

Retention-based cohort analysis helps answer:

- Are we growing **sustainably** or just acquiring churn-prone users?
- Which acquisition periods produced the **highest quality customers**?
- Is a spike in sales driven by **long-term users or one-time buyers**?

For e-commerce companies, **retention is often a stronger indicator of business health than revenue growth alone**.

Let's start the SQL work!

Step 1 - Identifying Customer Cohorts

The first step in cohort analysis is identifying the **cohort month** for each customer, defined as the month when they made their first purchase.

```sql
use olist_dump;

-- COHORT ANALYSIS --

create or replace view customer_cohorts as
select customer_unique_id,
min(date_format(o.order_purchase_timestamp, '%Y-%m')) as cohort_month
from customers c join orders o on c.customer_id = o.customer_id
where o.order_status = 'Delivered'
group by customer_unique_id;

select * from customer_cohorts;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

customer_unique_id	cohort_month
871766c5855e863f6eccc05f988b23cb	2017-09
eb28e67c4c0b83846050ddfb8a35d051	2017-04
3818d81c6709e39d06b2738a8d3a2474	2018-01
af861d436cfc08b2c2ddefd0ba074622	2018-08
64b576fb70d441e8f1b2d7d446e483c5	2017-02

- min() shows first purchase
- '%y-%m' means month-level granularity
- One row per customer

This defines when each customer entered the platform.

Step 2 – Tracking Monthly Customer Activity

After identifying cohorts, the next step is tracking **customer activity by month**.

```sql
create or replace view orders_with_month as
select o.customer_id, c.customer_unique_id, o.order_id,
date_format(o.order_purchase_timestamp, '%Y-%m') as order_month
from orders o join customers c on o.customer_id = c.customer_id
where o.order_status = 'Delivered';

select * from orders_with_month;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

customer_id	customer_unique_id	order_id	order_month
3ce436f183e68e07877b285a838db11a	871766c5855e863f6eccc05f988b23cb	00010242fe8c5a6d1ba2dd792cb16214	2017-09
f6dd3ec061db4e3987629fe6b26e5cce	eb28e67c4c0b83846050ddfb8a35d051	00018f77f2f0320c557190d7a144bdd3	2017-04
6489ae5e4333f3693df5ad4372dab6d3	3818d81c6709e39d06b2738a8d3a2474	000229ec398224ef6ca0657da4fc703e	2018-01
d4eb9395c8c0431ee92fce09860c5a06	af861d436cfc08b2c2ddefd0ba074622	00024acbcdf0a6daa1e931b038114c75	2018-08
58dbd0b2d70206bf40e62cd34e84d795	64b576fb70d441e8f1b2d7d446e483c5	00042b26cf59d7ce69dfabb4e55b4fd9	2017-02

orders_with_month 2 | Read Only

- Converted every order into a **monthly activity record**

- Kept customer granularity intact

This captures every month in which a customer was active.

Step 3 – Calculating Cohort Index

To measure retention, we need to calculate how many months after their first purchase a customer returns. This is done using a **cohort index**.

```
25 •  create or replace view cohort_analysis as
26    select cc.cohort_month, owm.order_month, owm.customer_unique_id,
27    timestampdiff(month, str_to_date(concat(cc.cohort_month, '-01'), '%Y-%m-%d'),
28    str_to_date(concat(owm.order_month, '-01'), '%Y-%m-%d')) as cohort_index
29    from customer_cohorts cc join orders_with_month owm
30    on cc.customer_unique_id = owm.customer_unique_id;
31
32 •  select * from cohort_analysis;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

cohort_month	order_month	customer_unique_id	cohort_index
2017-09	2017-09	871766c5855e863f6eccc05f988b23cb	0
2017-04	2017-04	eb28e67c4c0b83846050ddfb8a35d051	0
2018-01	2018-01	3818d81c6709e39d06b2738a8d3a2474	0
2018-08	2018-08	af861d436cfc08b2c2ddefd0ba074622	0
2017-02	2017-02	64b576fb70d441e8f1b2d7d446e483c5	0

- By concat with '01' and str-to-date(), converted month strings to real dates
- Timestampdiff calculating month differences between two dates
- cohort_index = 0 means first purchase month
- cohort_index = 1 means one month later and so on

This allows us to see how customer engagement decays over time.

Step 4 – Counting Active Customers

Once the cohort index is calculated, we count how many unique customers are active in each cohort for each month.

```sql
-- Count active customers per cohort per month:

select cohort_month, cohort_index,
count(distinct customer_unique_id) as active_customers
from cohort_analysis
group by cohort_month, cohort_index
order by cohort_month, cohort_index;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

cohort_month	cohort_index	active_customers
2016-09	0	1
2016-10	0	262
2016-10	6	1
2016-10	9	1
2016-10	11	1

- DISTINCT avoids duplicate orders inflating counts

This shows how many customers from each cohort return every month.

Step 5 – Calculating Retention Rate

Retention rate is calculated by dividing the number of active customers in each month by the total number of customers in the cohort.

```sql
-- calculate retention rate:

select cohort_month, cohort_index, count(distinct customer_unique_id) as active_customers,
round(count(distinct customer_unique_id)*100.0/
first_value(count(distinct customer_unique_id))
over( partition by cohort_month order by cohort_index),2) as retention_rate_pct
from cohort_analysis
group by cohort_month, cohort_index
order by cohort_month, cohort_index;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

cohort_month	cohort_index	active_customers	retention_rate_pct
2016-09	0	1	100.00
2016-10	0	262	100.00
2016-10	6	1	0.38
2016-10	9	1	0.38
2016-10	11	1	0.38

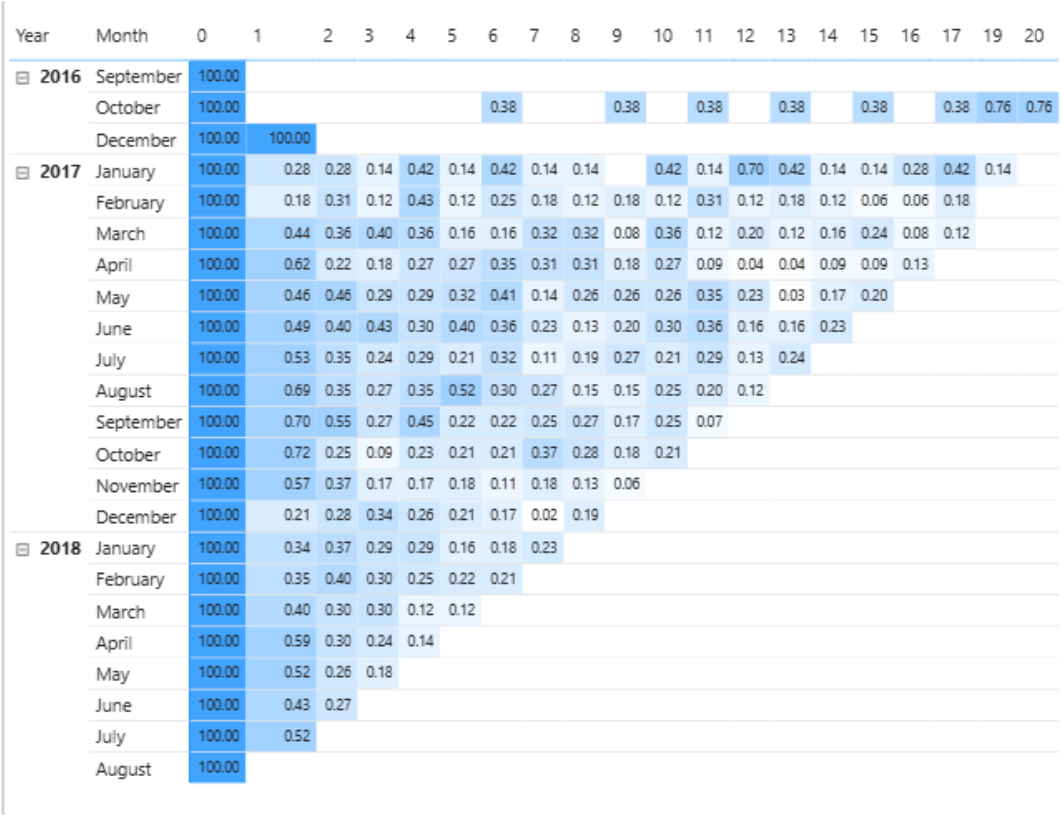
Result 5 | Read Only

- **FIRST_VALUE()** is a window function; it partitions data by cohort_month; orders rows by cohort_index**; extracts Month 0 active users (cohort size)** and uses that value as the denominator for all months.

Retention is measured against the original customer base, ensuring reliable comparisons across months.

Data Visualisation

The cohort heatmap below tracks customer retention over time. Each row represents a cohort of customers grouped by their **first purchase month**, while each column shows the **number of months since that first purchase (cohort index)**. The values indicate the **percentage of customers who returned and placed another order** in subsequent months.



Year	Month	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19	20
2016	September	100.00																			
	October	100.00						0.38			0.38		0.38		0.38		0.38		0.38	0.76	0.76
	December	100.00	100.00																		
2017	January	100.00	0.28	0.28	0.14	0.42	0.14	0.42	0.14	0.14		0.42	0.14	0.70	0.42	0.14	0.14	0.28	0.42	0.14	
	February	100.00	0.18	0.31	0.12	0.43	0.12	0.25	0.18	0.12	0.18	0.12	0.31	0.12	0.18	0.12	0.06	0.06	0.18		
	March	100.00	0.44	0.36	0.40	0.36	0.16	0.16	0.32	0.32	0.08	0.36	0.12	0.20	0.12	0.16	0.24	0.08	0.12		
	April	100.00	0.62	0.22	0.18	0.27	0.27	0.35	0.31	0.31	0.18	0.27	0.09	0.04	0.04	0.09	0.09	0.13			
	May	100.00	0.46	0.46	0.29	0.29	0.32	0.41	0.14	0.26	0.26	0.26	0.35	0.23	0.03	0.17	0.20				
	June	100.00	0.49	0.40	0.43	0.30	0.40	0.36	0.23	0.13	0.20	0.30	0.36	0.16	0.16	0.23					
	July	100.00	0.53	0.35	0.24	0.29	0.21	0.32	0.11	0.19	0.27	0.21	0.29	0.13	0.24						
	August	100.00	0.69	0.35	0.27	0.35	0.52	0.30	0.27	0.15	0.15	0.25	0.20	0.12							
	September	100.00	0.70	0.55	0.27	0.45	0.22	0.22	0.25	0.27	0.17	0.25	0.07								
	October	100.00	0.72	0.25	0.09	0.23	0.21	0.21	0.37	0.28	0.18	0.21									
	November	100.00	0.57	0.37	0.17	0.17	0.18	0.11	0.18	0.13	0.06										
	December	100.00	0.21	0.28	0.34	0.26	0.21	0.17	0.02	0.19											
2018	January	100.00	0.34	0.37	0.29	0.29	0.16	0.18	0.23												
	February	100.00	0.35	0.40	0.30	0.25	0.22	0.21													
	March	100.00	0.40	0.30	0.30	0.12	0.12														
	April	100.00	0.59	0.30	0.24	0.14															
	May	100.00	0.52	0.26	0.18																
	June	100.00	0.43	0.27																	
	July	100.00	0.52																		
	August	100.00																			

- The visual is done in Power BI with the output generated from retention rate calculation done in MYSQL. This heatmap look alike is formed using Matrix visual with conditional formatting.

Insights

- All cohorts start with **100% retention at Month 0**, which validates that the cohort size is correctly defined using customers' first delivered order.
- Across nearly all cohorts, retention drops sharply from **100% in Month 0 to 20–40% in Month 1**. Do not take (2016, December) into account for 100% retention as only 1 customer was active during the phase. A large portion of customers do not return after their first purchase, indicating **high early churn**.
- By **Month 3–4**, most cohorts fall below **15–25% retention**, and by later months, retention stabilizes at **single-digit percentages**. Olist customers tend to be **transactional rather than habitual buyers**.

- Cohorts from **late 2017 and early 2018** demonstrate marginally better retention in the first 1–2 months compared to earlier cohorts. This suggests potential improvements in product experience, logistics, or seller quality over time.
- Despite overall low retention, a small fraction of customers continues to purchase even after **6–10 months**. There exists a **core loyal segment**, though it is relatively small.

Business Recommendations

for the Marketing Team

- Focus campaigns on first-to-second purchase conversion
- Launch post-purchase email nudges within say 7–30 days
- Offer time-bound incentives such as discounts or free shipping for repeat orders

for the Product & Platform Team

- Improve the post-purchase experience with better delivery updates and seller communication
- Highlight highly rated and trusted sellers to increase buyer confidence

for Business & Growth Teams

- Identify high-performing cohorts and analyse what drove better retention
- Allocate marketing budgets toward retention initiatives rather than only new user acquisition.

Conclusion

This cohort analysis on the Olist e-commerce dataset demonstrates how looking beyond aggregate metrics can uncover meaningful customer behaviour patterns. While overall order volumes may appear healthy, cohort-based retention reveals a different story — a significant drop in customer activity immediately after the first purchase.

By tracking customers based on their acquisition month and analysing their repeat purchase behaviour over time, we were able to identify that **early churn is the primary retention challenge**. Most customers do not return after their initial order, and only a small fraction remain active in the long term. This insight would remain hidden without a cohort-based approach.

From a business perspective, the findings highlight the importance of shifting focus from pure customer acquisition to **early lifecycle engagement**. Improving the experience between the first and second purchase can have a disproportionate impact on long-term revenue and customer lifetime value.

Thank you for taking the time to read this analysis. I hope it provided useful insights into how cohort analysis can uncover meaningful customer behaviour patterns. I welcome any feedback, questions, or discussions on *shraddhaasaha@gmail.com*