Cohort analysis

Вступление

В этом проекте мы произведем когортный анализ датасета с покупками клиентов и вычислим retention rate с использованием SQL.

Источник данных

Данные были сгенерированы искусственно. Датасет доступен в той же папке, что и данный файл, в формате .sql и .xlsx.

Описание данных

user_id - ID клиента

order_date - дата покупки

amount - сумма покупки

Анализ

Когорта - это группа пользователей, пришедших в один период. В рамках данного анализа мы возьмем за период первый месяц покупки пользователя.

```
with
--определяем дату первой покупки для каждого клиента
first_purchase as
    (select user_id,
        min(order_date) as first_purchase
    from orders
    group by 1),
--формируем когорты
cohorts as
    (select o.user_id,
        o.order_date,
        date_trunc('month', o.order_date::date) as purchase_mnth,  --месяц покупки
        date_trunc('month', f.first_purchase::date) as cohort_mnth,  --когорта, к которой относится клиент
        extract(month from age(o.order_date, f.first_purchase)) as mnths --месяц относительно когорты
    from orders as o
    join first_purchase as f
        on o.user_id = f.user_id),
--рассчитываем абсолютное кол-во активных юзеров в каждый месяц когорты
absolutes as
    (select cohort_mnth,
```

```
        mnths,
        count(distinct user_id) as active_users
    from cohorts
    group by 1, 2
    order by 1, 2)
--теперь к абсолютным значениям добавляем долю - какой процент активных клиентов когорты до
шел до конкретного месяца   (retention rate)
select *,
    max(active_users) over (partition by cohort_mnth order by mnths) as max_users,
    round(active_users/max(active_users) over (partition by cohort_mnth order by mnths)::numeric * 100,
2) as retention_rate
from absolutes
```

В результате мы получим следующую таблицу:

	cohort_mnth timestamp with time zone	mnths numeric	active_users bigint	max_users bigint	retention_rate numeric
1	2025-01-01 00:00:00+05	0	73	73	100.00
2	2025-01-01 00:00:00+05	1	51	73	69.86
3	2025-01-01 00:00:00+05	2	48	73	65.75
4	2025-01-01 00:00:00+05	3	37	73	50.68
5	2025-02-01 00:00:00+05	0	24	24	100.00
6	2025-02-01 00:00:00+05	1	17	24	70.83
7	2025-02-01 00:00:00+05	2	11	24	45.83
8	2025-03-01 00:00:00+05	0	2	2	100.00
9	2025-04-01 00:00:00+05	0	1	1	100.00