Data Engineering Technical Assessment
Part II: Cohort Analysis

Mohamed Amin Soumih

May 27, 2025

Objective

Analyze user retention behavior over time by grouping users into cohorts based on their signup month, and calculating how many users from each cohort returned weekly for 8 weeks.

Data Pipeline

- A synthetic e-commerce dataset was generated and loaded into a PostgreSQL database.
- Users were grouped into **monthly cohorts** using the expression `DATE_TRUNC('month', signup_date)`.
- Events were joined with users to track activity over time.
- The number of weeks since each user's signup was computed using the formula `(event_week - signup_month) // 7`.
- Records were filtered to include only weeks 0 to 7, ensuring an 8-week analysis window.

Calculating Weekly Retention

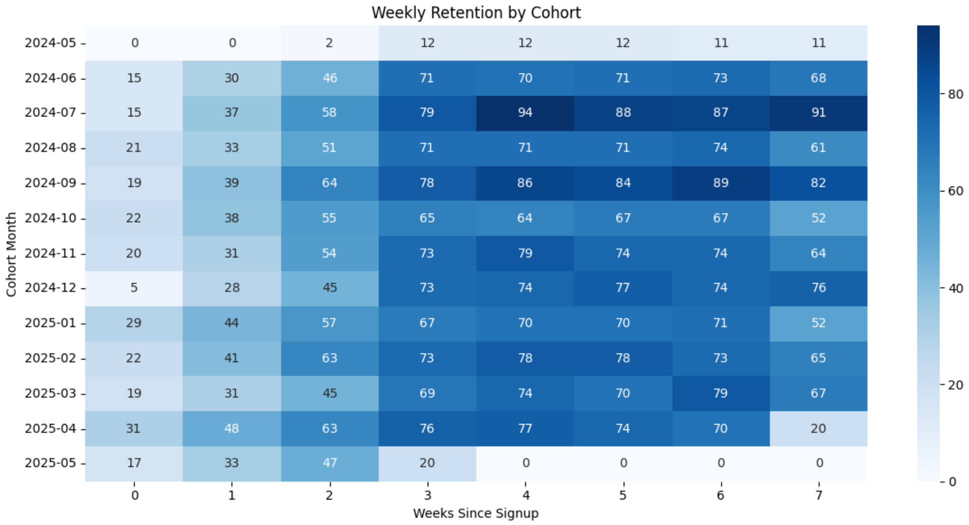
For each user, we determined whether they were active in the week of signup (week 0) and in each of the 7 subsequent weeks. We then calculated the number of distinct users who were active in each week, grouped by their signup cohort.

This produced a **retention matrix**, where:

- Rows = Cohort month
- Columns = Week since signup (0 to 7)
- Cell value = Number of users from that cohort active during that week

Visualization

The matrix was visualized using `seaborn` as a heatmap, saved under `figures/retention_heatmap.png`.



Conclusion

This analysis highlights user retention trends per cohort. It clearly shows drop-off rates over time and identifies which signup periods had higher long-term engagement.