Customer Retention

A marketing analytics solution

TABLE OF CONTENTS

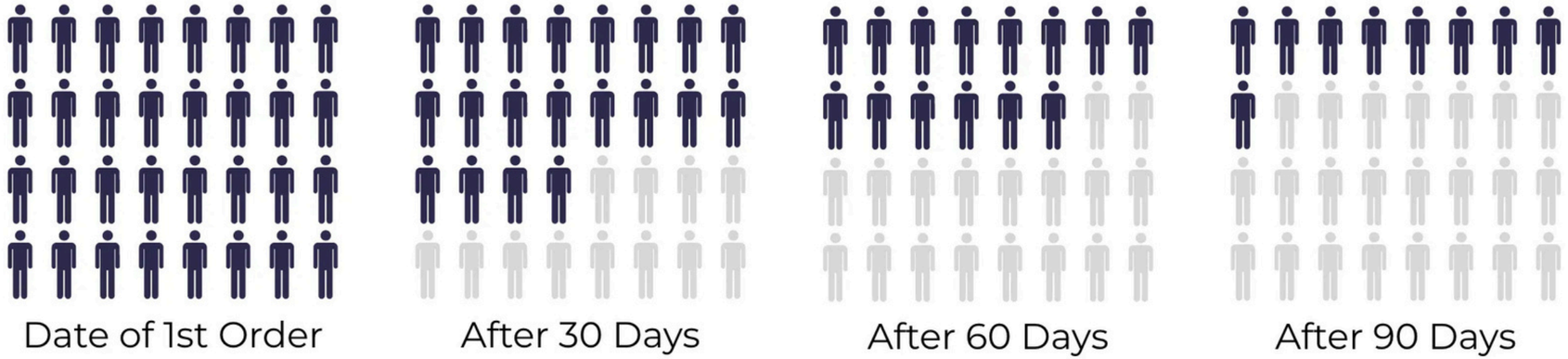
Introduction

Acquiring customers is costly, so understanding whether they return is crucial.

In this analysis, customers are grouped into cohorts based on their first purchase and we explore how long it takes customers to make a second purchase.

This helps us understand how repeat behavior changes across different cohorts.

SCOPE OF THE PROJECT



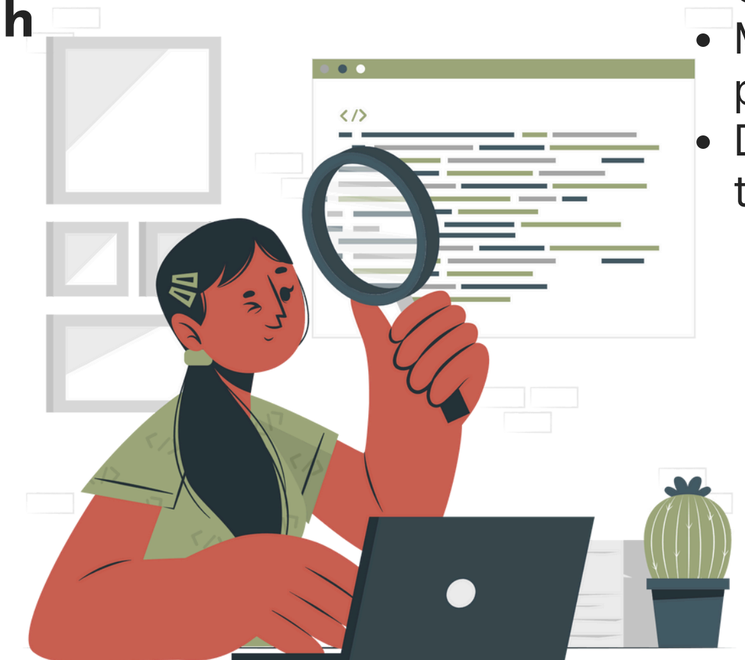
How long does it take for customers to make their second purchase, and how does this vary across different cohorts?

Dataset Description

The dataset contains order-level transactional data with five columns: a technical row identifier (row_id), customer identifier (customer_id), order date (order_date), order identifier (order_id), and sales value (sales). The data covers customer activity from 2024, with an extended version including 2025 data for comparison.

Data Architecture & Approach

The project follows a Bronze-Silver-Gold structure to progressively improve data quality. Each layer has a clear responsibility and is rebuilt manually when needed using SQL notebooks.



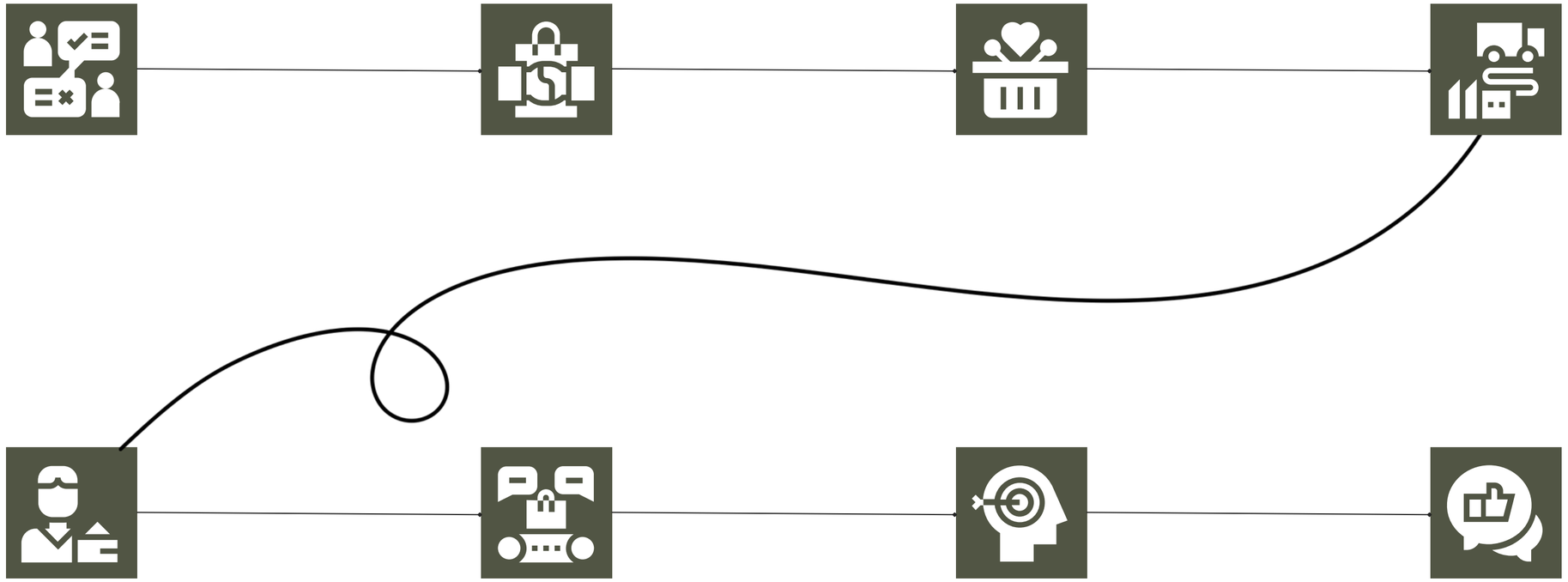
Tools & Technologies

- Databricks SQL (notebooks)
- Delta Lake tables
- Window functions and CTEs

Key Outputs

- Clean, analytics-ready order data
- Customer-level cohort table
- Metrics for retention and repeat purchase behavior
- Dashboards comparing different time periods (2024 vs 2024–2025)

High-Level Flow



Data Loading (CSV - Delta in Databricks) 2024 source

Raw ingestion & Standardization

Bronze Layer

Silver Layer (Clean & Deduplicated)

Gold Layer (Analysis-ready cohort metrics)

Build Reporting Tables (cohort size, retention, repeat depth) **and Create Dashboard**

Data Loading (Append New Data - 2025 source)

Rebuild the reporting tables & Create Dashboard (2024-2025)

Data Transformation Overview

Purpose

Store the original transactional data in a structured and reliable format, without changing its meaning.
This layer acts as the single source of truth for all downstream transformations.

Input

Source tables containing raw order data (src_orders_2024, src_orders_2025 appended later on)

What happened in this layer?

- Raw data was loaded into a Delta table
- Column data types were defined
- Minor formatting was applied (text fields were trimmed to remove extra spaces, sales values were cast to numeric format)

no rows were removed
no bussiness logic was applied

BRONZE

Output raw_orders

Data remains raw but structured and consistent

SILVER

Purpose

Improve data quality and reliability by removing invalid records and resolving duplicates.
This layer prepares the data for accurate analysis.

Input

raw_orders (Bronze layer)

What happened in this layer?

- **Invalid rows were removed**: records with missing customer IDs, missing order IDs, missing order dates, and sales values that were missing, zero, or negative.
- **Duplicate orders were resolved**: orders were grouped by order_id, the most recent valid record was kept, and ties were broken using a technical identifier to ensure deterministic results.
- **Data was standardized**: the final dataset contains one clean, consistent, and trustworthy record per order, ready for analysis.

Output clean_orders

One valid row per oder
Clean, analysis-ready transactional data

Purpose

Transform clean transactional data into business-level insights focused on customer behavior and retention.

Input

clean_orders (Silver layer)

What happened in this layer?

- Orders were analyzed at the customer level
- Each customer's purchase history was ordered chronologically
- Key customer milestones were identified: First purchase date, Second purchase date (if applicable)
- Customers were assigned to cohorts based on their first purchase month
- Time between first and second purchase was calculated to measure early retention

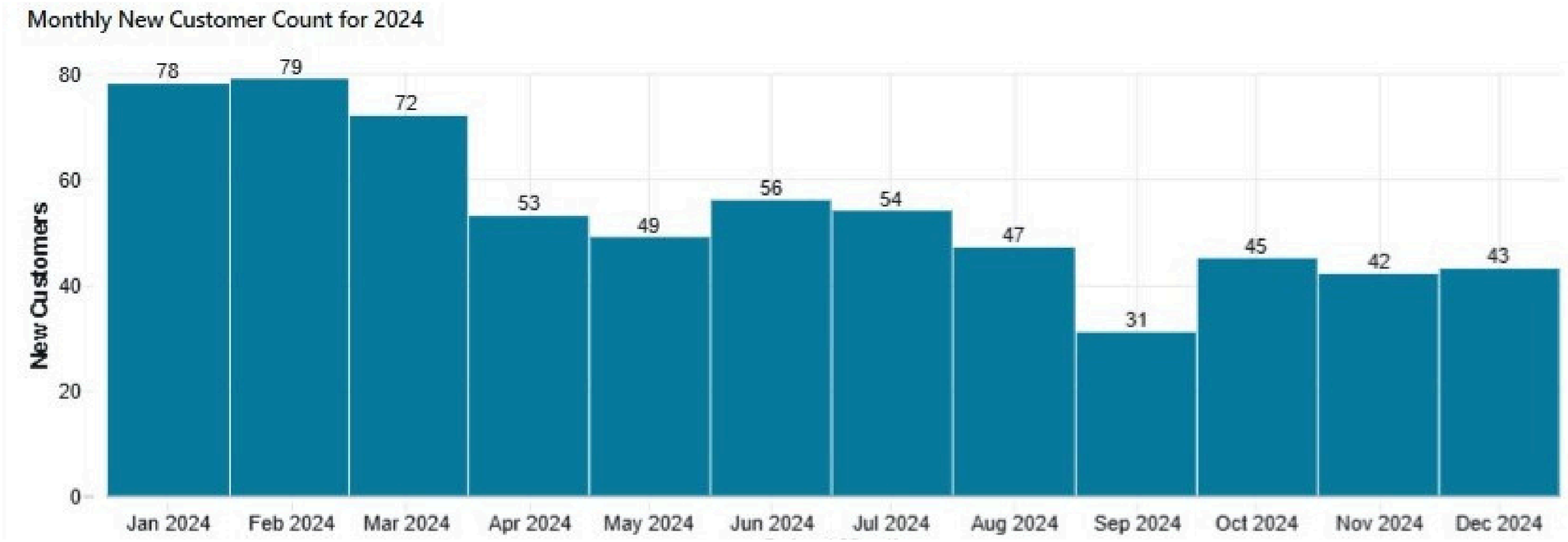
GOLD

Output cohort_customers

One row per customer, including: cohort month, first and second purchase dates, days to repeat purchase

Dataset ready for cohort retention dashboard and analysis

Customer Acquisition by Cohort Month (2024)

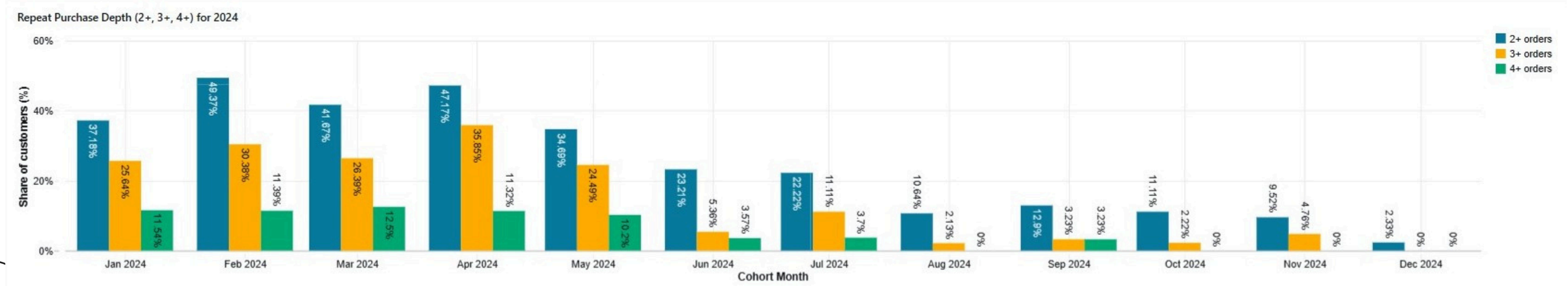


- Shows the number of new customers acquired each month in 2024
- Each bar represents a cohort (customer's first purchase month)
- Higher bars indicate stronger acquisition
- Early 2024 (Jan–Mar) has the strongest acquisition
- Mid-year shows a decline, with a slight recovery toward the end of the year
- Provides context for retention metrics (smaller cohorts naturally lead to lower retention later)

Short-Term Customer Retention by Cohort (2024)

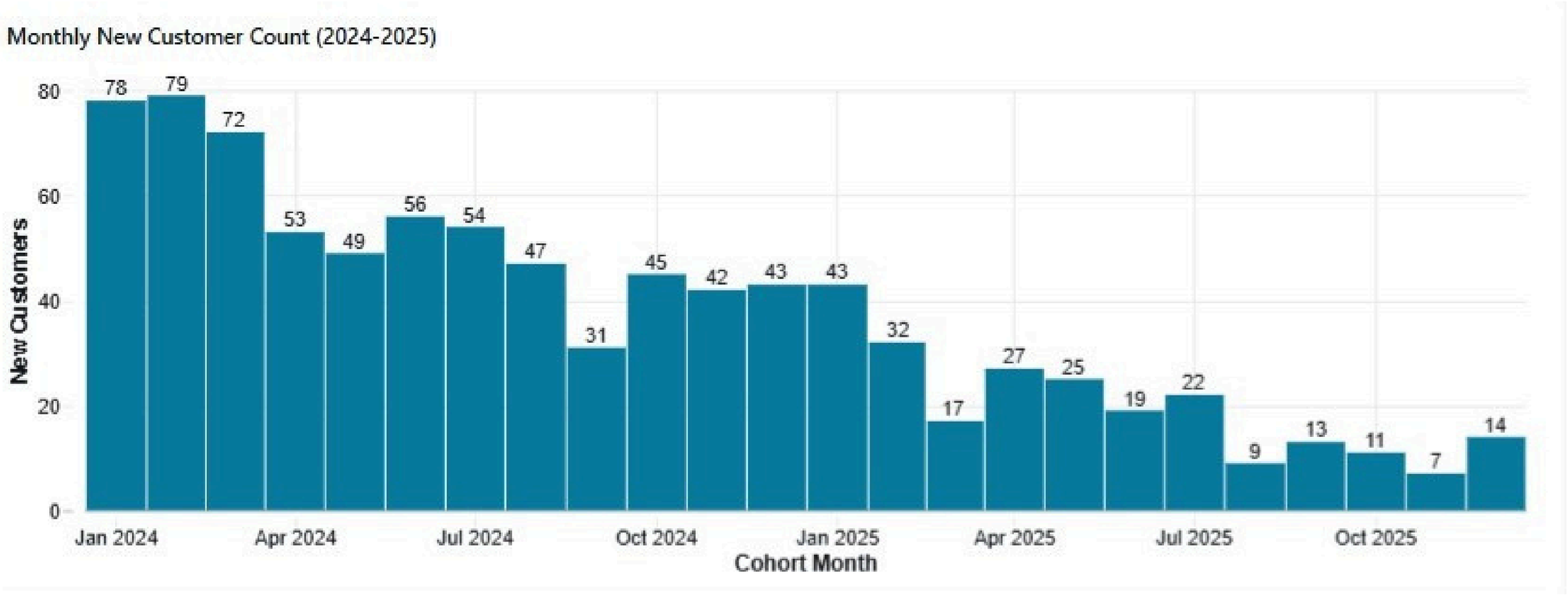
- Shows 1-month, 2-month, and 3-month retention rates per cohort
- Measures how many customers return after their first purchase
- Higher lines indicate better short-term retention
- Earlier cohorts (Feb-Apr) show stronger retention across all time windows
- Later cohorts appear weaker mainly due to limited time to return, not worse performance
- Helps distinguish real performance changes from cohort maturity effects

Repeat Purchase Depth by Cohort (2024)



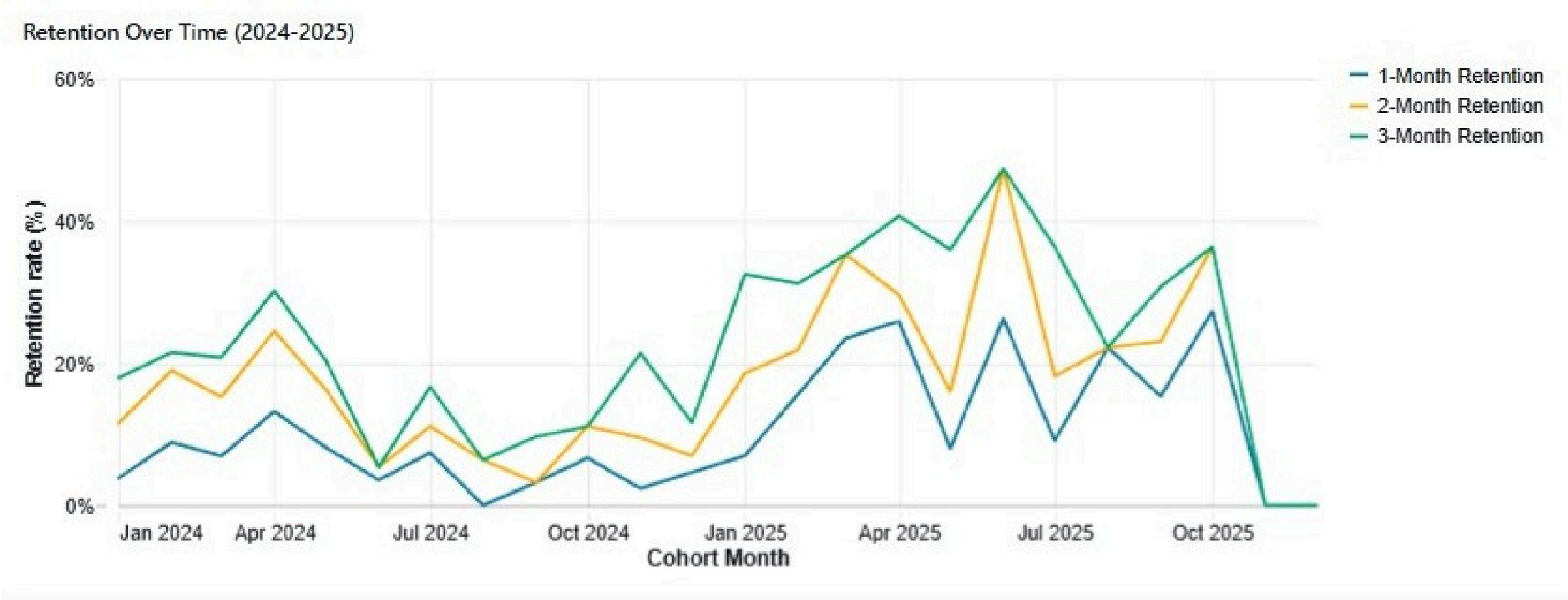
- Shows how deeply customers repeat purchases within each cohort
- 2+ orders: customers who returned at least once
- 3+ orders: customers who returned at least twice
- 4+ orders: highly engaged repeat customers
- Values are shown as a percentage of each cohort
- Older cohorts show higher repeat depth
- Recent cohorts show lower repeat depth due to shorter observation periods

Impact of Adding 2025 Cohorts (vs 2024)



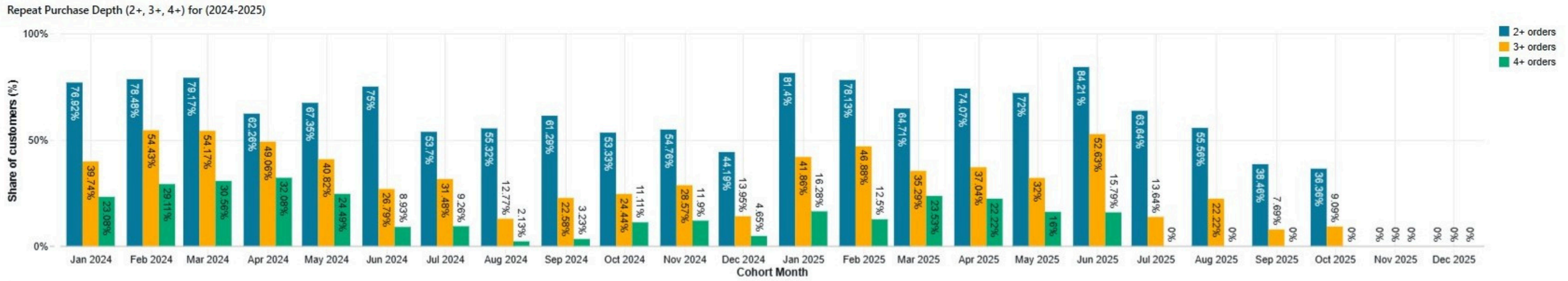
- Shows the number of new customers acquired each month from 2024 through 2025
- Each bar represents a cohort, based on the customer's first purchase month
- Customer acquisition is strongest in early 2024 and gradually declines into 2025
- 2025 cohorts are visibly smaller, indicating slower acquisition or incomplete months
- Provides important context for retention and repeat metrics in later charts

Short-Term Retention by Cohort (2024–2025)



- Retention lines now show more fluctuation, especially in 2025 cohorts
- Several 2025 cohorts display sharp spikes, driven by very small cohort sizes
- Drops in the most recent months are expected, as customers have not yet had time to return
- The longer time range makes cohort maturity effects more obvious

Repeat Purchase Depth by Cohort (2024-2025)



- Repeat purchase depth now includes 2025 cohorts, which mostly show lower values
- High repeat depth is concentrated in older cohorts, especially early 2024
- Many 2025 cohorts show little or no 3+ or 4+ purchase behavior due to limited time
- The contrast between mature (2024) and immature (2025) cohorts is clearer

CONCLUSIONS

- Customer acquisition was strongest in early 2024, creating larger and more stable cohorts.
- Retention and repeat purchase depth are consistently higher in older cohorts, mainly because they have had more time to mature.
- Newer cohorts, especially in 2025, appear weaker across retention and repeat metrics, but this is driven by limited observation time rather than a decline in customer quality.
- Extending the analysis into 2025 introduces more variation, which is expected given the smaller cohort sizes.
- Overall, customer behavior remains consistent: once customers return for a second purchase, repeat engagement deepens over time.

The analysis suggests that improving early re-engagement, rather than acquisition volume alone, is the most effective way to strengthen long-term customer value.

RECOMMENDATIONS

KEEP IT SIMPLE

Send simple follow-up messages to stay in touch with new customers.

DRIVE SECOND PURCHASE

Encourage customers to make a second purchase soon after their first.

MAKE IT MEMORABLE

Use targeted offers or reminders to reinforce repeat purchasing.

TEST & ITERATE

Continue tracking recent cohorts to see how their behaviour develops over time.

THANKS FOR WATCHING!