Project Report

A Retention-Focused Revenue Analysis for an E-commerce Business

Chirag Somashekar

Date: 20 November 2025

Table of Contents

1. Executive Summary

This report analyzes the key revenue and customer behavior drivers of an e-commerce business (thelook_ecommerce: bigquery public dataset), with a focus on understanding why long-term revenue growth is limited despite strong acquisition performance. Using cohort analysis, RFM segmentation, retention analysis and ARPU/LTV modeling, the analysis reveals a single dominant bottleneck: customers rarely return after their first purchase.

Although new customer acquisition is healthy, only 12% of customers ever repurchase and 70-80% of total customer value is captured in Year 0-1. Most cohorts rapidly plateau and long-term customer value remains low because customers stop engaging after their first order. As a result, nearly 44% of all revenue sits in dormant segments such as Needs Attention and Hibernating, representing a large and untapped reactivation opportunity.

At the same time, high value segments like Loyal Customers and Champions contribute 40% of long-term revenue despite being much smaller in size. This imbalance indicates that the business already has a strong base of profitable buyers, but does not effectively move first time purchasers into these high value groups and then push the high value groups further to increase retention.

The findings point to a clear growth path: retention initiatives deliver far higher revenue potential than additional acquisition spending. Improving early-lifecycle retention (particularly the second purchase window) and reactivating lapsed customers would unlock significant LTV uplift without increasing customer acquisition costs.

This report concludes with data-driven recommendations for improving Year-1 retention, reactivating dormant segments and enhancing the post-purchase experience through personalized engagement and reduced friction. Together, these initiatives form a sustainable, compounding path to higher revenue and stronger customer lifetime value.

2. Introduction

This report analyzes customer behavior and revenue performance using the thelook_ecommerce public dataset from Google BigQuery. The dataset represents a mid-size e-commerce business and includes customer, order and product information across multiple years. It provides a realistic environment to evaluate how acquisition, retention and long-term value contribute to business growth.

The analysis is written from the perspective of a Marketing Analyst. The goal is to understand why long-term revenue growth slows down despite healthy customer acquisition and to identify which customer groups offer the highest revenue potential. The targeted audience includes stakeholders in growth, marketing, CRM, product and data teams who rely on data-driven insights to guide decisions around customer lifecycle, retention initiatives and revenue strategy.

To answer the core business questions, the analysis applies industry-standard techniques including cohort analysis, RFM segmentation, retention curve evaluation and ARPU/LTV modeling. These methods help reveal how frequently customers return, how valuable they are over time and where the largest opportunities for revenue uplift lie.

Overall, the introduction outlines the key business context and sets up the analytical direction for the rest of the report.

3. Data Description

3.1 Data Overview

- Public e-commerce dataset from Google BigQuery (thelook_ecommerce).
- Includes customer, order, item, event and marketing touchpoint tables.
- Time period analyzed: January 2019 – November 2025.

3.2 Dataset Scale (Raw)

- 2.43M user events
- 125K orders

- 181K ordered items

3.3 Major Tables used

- events: user interactions, used for attribution and user behavior understanding.
- orders: transaction-level info (order date, customer ID, order value)
- order_items: line-item details used for revenue and quantity.
- users: customer-level attributes.

4. Data Preparation & Cleaning

To prepare the dataset for analysis, a series of clean SQL views was created in BigQuery to provide a consistent and reliable source of truth for each metric used in the project. Only completed orders were retained in the dataset; transactions that were cancelled, returned, or partially processed were excluded to avoid inflating revenue. Timestamps were standardized across tables and several derived fields such as order week, cohort year and year offset were generated to support the time based components of the analysis.

An analysis-ready dataset was then formed by joining the events, orders, order_items and user tables, ensuring that behavioural, transactional and user-level information aligned correctly. Based on this unified dataset, RFM variables (Recency, Frequency and Monetary value) were constructed and each customer was assigned to a registering cohort defined by the date of their first purchase.

For the cohort retention and lifetime value sections, cumulative ARPU was modelled at the cohort level, with predicted future values included where appropriate. A simple last-touch attribution approach was applied to map events to orders and identify the channels associated with conversions.

Through the process, the dataset was checked for completeness and internal consistency, including assessments for missing values, negative quantities, inconsistent timestamps and NULL user identifiers. This data preparation ensured that all subsequent insights including RFM segmentation, funnel performance, cohort retention patterns and ARPU/LTV calculations were derived from clean, dependable data.

5. Business Questions

This analysis was guided by a set of core business questions focused on understanding customer behavior, revenue contribution and long-term growth potential. Specifically, the project aimed to answer the following:

Which acquisition channels bring customers who convert and generate meaningful revenue?

- What are the key customer segments and how do their behaviors differ?
- How does user behavior evolve from first purchase to later purchses?
- Where does revenue leakage occur in the customer journey (funnel + lifecycle)?
- How do customer cohorts retain over time?
- How much value does each customer generate across their lifecycle (ARPU/LTV), and what portion is realistic vs. long-range projection?
- Which customer groups represent the largest opportunity for revenue lift?

6. Methodology

This section outlines the analytical approach used to investigate customer behavior, revenue contribution and long-term value drivers within the dataset. The methodology follows a structured sequence aligned with industry-standard marketing analytics practices.

6.1 Tools & Technologies

- SQL (BigQuery): Used for data cleaning, feature engineering, cohort creation and modelling tables.
- Tableau: Used for dashboard creation and interactive visual exploration.
- Google Sheets: Used for lightweight modelling checks and sanity validation.
- Keynote & Microsoft Word: Used for Presentation and Reporting.

6.2 Exploratory Analysis & Core KPIs

This analysis began with a broad exploration of transaction-level and customer-level behaviour using standard KPIs including:

- Total customers, orders and revenue
- Average order value (AOV)
- Channel-level acquisition and conversion performance
- Order frequency and distribution

This stage helped identify overall demand patterns, customer behavior and anomalies that required deeper investigation

6.3 Customer Segmentation (RFM Analysis)

To identify meaningful customer groups, the analysis applied the RFM (Recency, Frequency, Monetary) segmentation framework. Customers were evaluated across three behavioral dimensions:

- Recency: How long since the customer’s last purchase
- Frequency: How often the customer purchases
- Monetary Value: Total revenue generated

Each metric was transformed into a relative score from 1 to 4 using the NTILE() function, which assigns customers into ranked buckets.

- A score of 1 represents the top-performing group (e.g., most recent purchase, highest spend)
- A score of 4 reflects lower engagement or value.

These three scores were then combined using a rule-based classification system (implemented in SQL) to assign each customer to a behavioral segment. Examples include:

- Champions: high scores across all RFM dimensions
- Loyal Customers: strong spend and frequency, recent purchasers
- Potential Loyal Customers: strong recent activity but fewer orders
- Needs Attention: declining recency and moderate activity.
- Hibernating: long inactive periods and low engagement

This segmentation allowed the analysis to map each customer into a lifecycle state and quantify revenue distribution, retention issues and reactivation opportunities.

6.4 Funnel Analysis

To evaluate how users progress through the purchase journey, an event-level funnel analysis was conducted using three key interaction points: product view, add to cart, and purchase. Because the dataset records user interactions at the event level rather than as aggregated sessions, the funnel measures transitions based on event counts rather than distinct users. This makes the analysis effective for understanding relative drop-off patterns, even though it does not represent a strict user-level conversion funnel.

Traffic-source filters were applied to compare funnel behavior across marketing channels. Email generated the largest volume of top-funnel activity, while all the channels were around 36% purchase attempt rate. This comparison helped identify which channels attract high-intent users versus high-volume users, informing both acquisition and retention strategies.

The funnel results were then used to quantify where the largest leakage occurs and to contextualize subsequent analyses on cohort retention and customer lifetime value.

6.5 Cohort Analysis

To evaluate customer retention and long-term value, registration cohorts were created based on the customer’s first completed purchase. Several cohort granularities were tested, weekly, monthly and yearly. Yearly cohorts were selected because the dataset exhibited low short-term repeat behavior: most customers did not return within the first several months after their initial purchase.

Using yearly cohorts therefore provided a clearer and more stable view of long-term retention patterns, reduced noise created by low repeat volumes in early months and aligned more naturally with the business questions around long term contribution, revenue trajectory and lifetime value modelling.

For each cohort, retention percentages were calculated across year offsets (Year 0 to Year 6) and cumulative ARPU was modelled using historical spend combined with cohort-level growth trends to forecast future value.

6.6 ARPU and Lifetime Value Modeling

To understand customer value over time, cumulative ARPU (Average Revenue Per User) was calculated at the cohort level. For each cohort, total revenue generated in a given year offset divided by the number of customers in that cohort. These yearly ARPU values were then accumulated to model the cumulative growth % which helped model how customer value builds over time.

The prediction model applied a continuation of the cohorts existing trend but was used cautiously, given that long-range projection can become unreliable in business with weak repeat activity.

This approach allowed the analysis to compare value creation across cohorts and understand how much of lifetime value is realized early versus in later years.

A basic channel attribution check was also performed to understand the final touch points leading to conversions, but the main focus of the analysis remained on customer behavior, retention patterns and lifetime value.

These steps ensured that all later findings were built on clean, consistent and well structured data.

7. Results & Analysis

This section presents the key findings derived from the cleaned and modelled dataset. The analysis follows the same structure as the dashboard and presentation: starting from overall business performance, then exploring customer segments, funnel behavior, retention patterns and lifetime value. Each component contributes to a clearer understanding of how users interact with the platform, where value is created and where growth opportunities exist.

7.1 Overall Business Performance (2025 YTD)

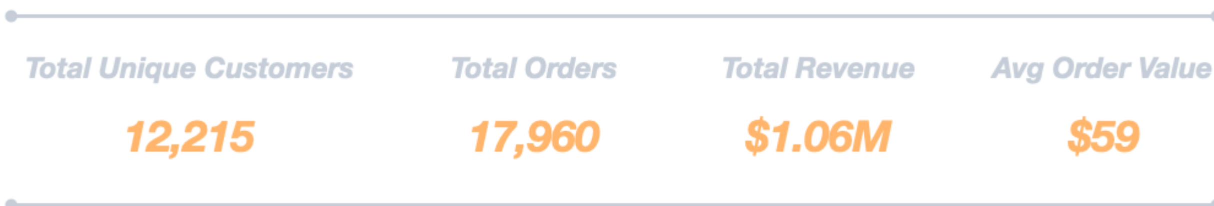

Figure 1 shows the key performance indicator for Jan-Nov 2025, highlighting strong YoY growth in customers (+67%), orders (+69%) and revenue (+66%), while AOV slightly declined (-1.6%).

The business shows strong year to date performance across all major sales KPIs. From January to November 2025, the company generated $1.06M in revenue, driven by 17,960 total orders from 12,215 unique customers. This reflects significant growth compared with the previous year, with revenue up 66%, orders up 69% and customer count increased by 67%.

Despite strong volume growth, average order value (AOV) decreased slightly to $59 (-1.6% YoY). This suggests that revenue expansion in 2025 was primarily driven by higher customer acquisition and increased purchase frequency.

Beyond overall sales performance, understanding which channels contributed most to this growth provides additional context.

Performance across marketing channels shows that Email remains the dominant revenue source, contributing nearly half of total revenue in 2025. Paid channels such as AdWords and Facebook also delivered meaningful volume but were comparatively less efficient. Organic search, although smaller in size, achieved the fastest year-over-year growth (+100%), suggesting improving brand strength and long-term acquisition potential.

These top level KPIs establish the overall performance landscape and set the context for deeper customer-level, funnel, segmentation and retention analysis presented in the following sections.

7.2 Customer Value Distribution (RFM Insights)

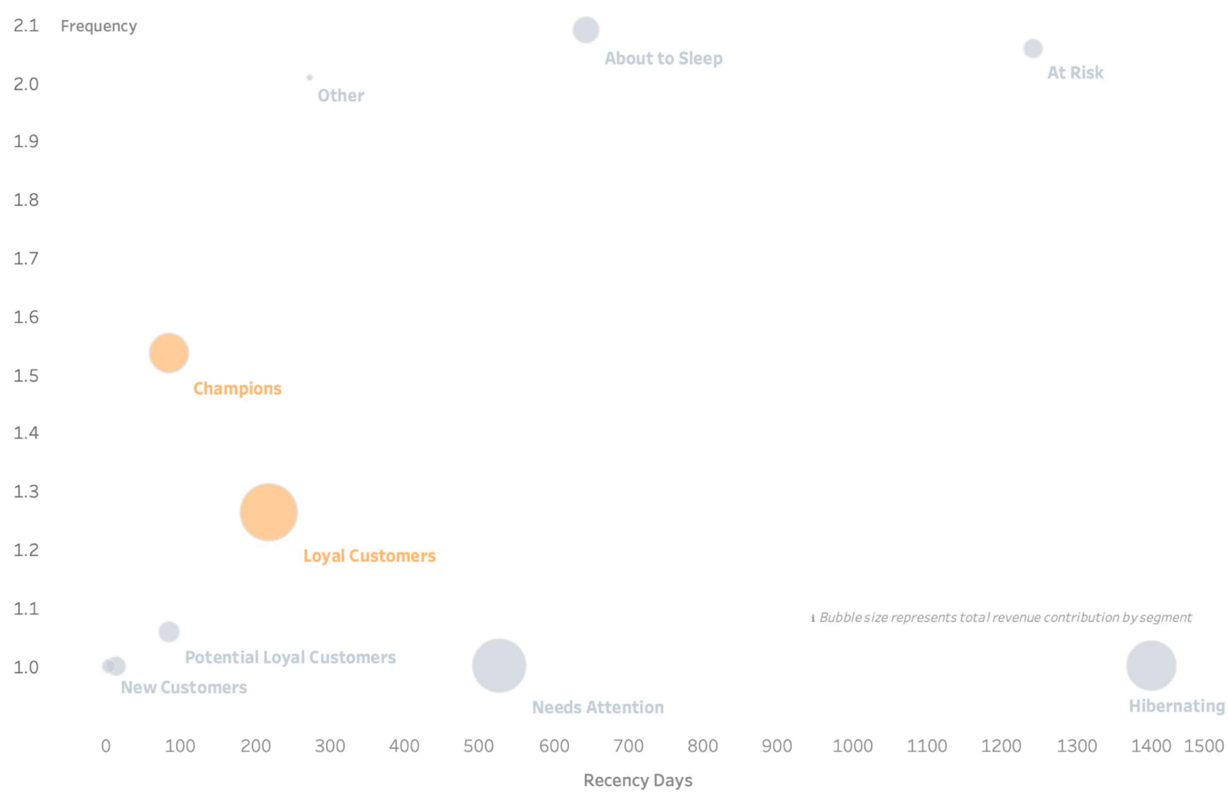

Figure 2 presents the RFM bubble chart showing customer distribution across Recency, Frequency and Monetary value (bubble size).

The analysis reveals a highly uneven distribution of customer value. Although most customers make only one purchase and do not return, two segments, Loyal Customers and Champions generate approximately 40% of all long-term revenue. These groups purchase more frequently, exhibit low recency and form the core of the business's profitable user base.

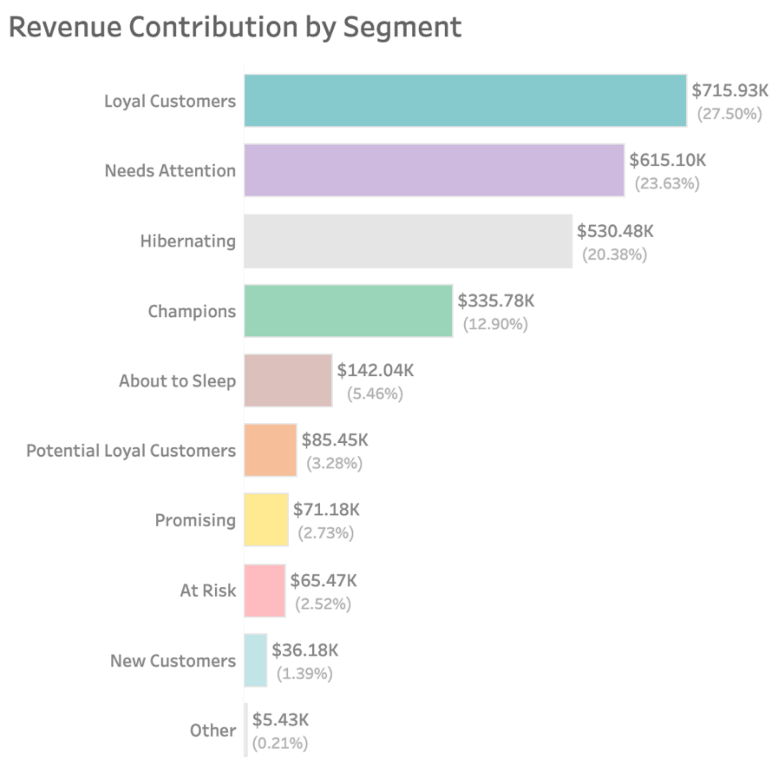

Figure 3 highlights revenue contribution of each RFM segment.

In contrast, a disproportionately large share of revenue, around 44% across all years comes from customers classified as Needs Attention or Hibernating. These users made at least one meaningful purchase but have not returned for an extended period. Their high recency values and low frequency suggests missed opportunities for re-engagement.

The presence of these large dormant segments, combined with a repeat purchase rate of only 12%, indicated that growth is currently constrained by retention rather than acquisition. Even small improvements in bringing Year 1 buyers back into the funnel would shift a significant portion of dormant customers into higher value segments, driving substantial long-term revenue uplift.

Overall, the RFM model shows a clear pattern: acquisition is strong, but retention is weak. The business already attracts high-value customers, the challenge lies in converting first-time buyers into repeat purchasers and reactivating the sizeable dormant base.

7.3 Funnel Performance by Traffic Source

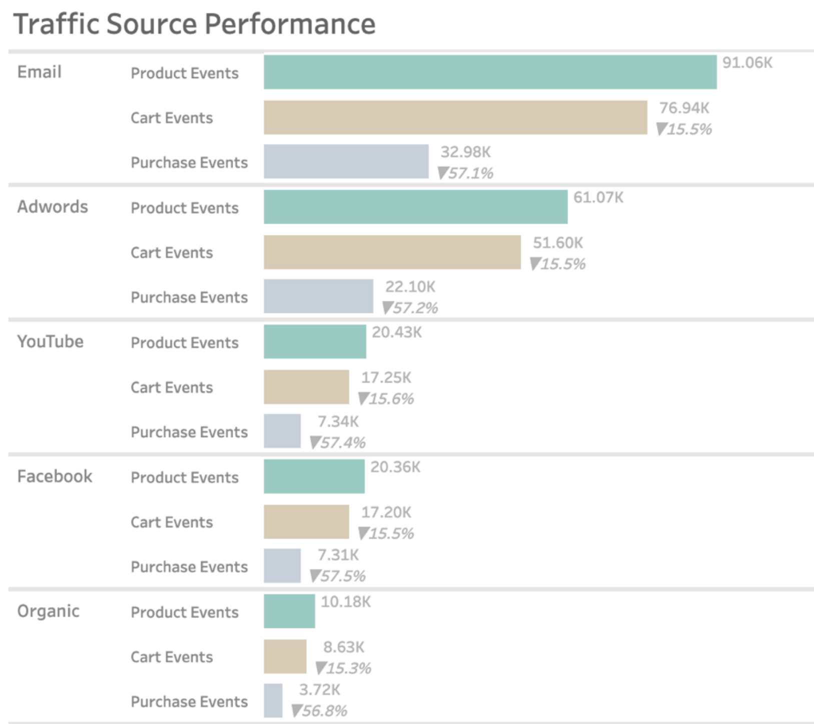

Figure 4 presents the event-level funnel (Product → Cart → Purchase) broken down by traffic source. These metrics do not represent completed orders but measure how users progress through the onsite journey.

Key Insights:

- Email is the strongest performer across every stage of the funnel. It brings the highest volume of product views (91K), cart events (77K), and purchases attempts (33K), demonstrating both high intent and strong conversion efficiency.
- Paid channels generate meaningful volume but have similar drop off patterns. All three lose ~57% of users between Cart to Purchase, suggesting consistently weak final stage conversion across paid traffic.

- Organic traffic has the smallest volume but the most efficient audience. Despite ~10K product events, it maintains the same drop off rate as paid channels and contributes a stable share of purchases.
- Across all channels, the largest leakage point is Cart to Purchase, where 56-57% users drop off. This indicates a system wide issue (checkout friction, low perceived value, or lack of urgency), not channel-specific weakness.

The funnel confirms a core theme in this report:

Acquisition is strong but conversion and retention are weak.

Email is already delivering high-quality demand, while Paid channels bring volume but not efficiency. Improving the final purchase step would unlock immediate revenue across all traffic sources.

Overall, the funnel analysis reveals that while acquisition brings volume, the business loses most uses at the final conversion step, reinforcing the need for checkout optimization and early retention improvements.

7.4 Cohort Retention (Yearly Cohorts)

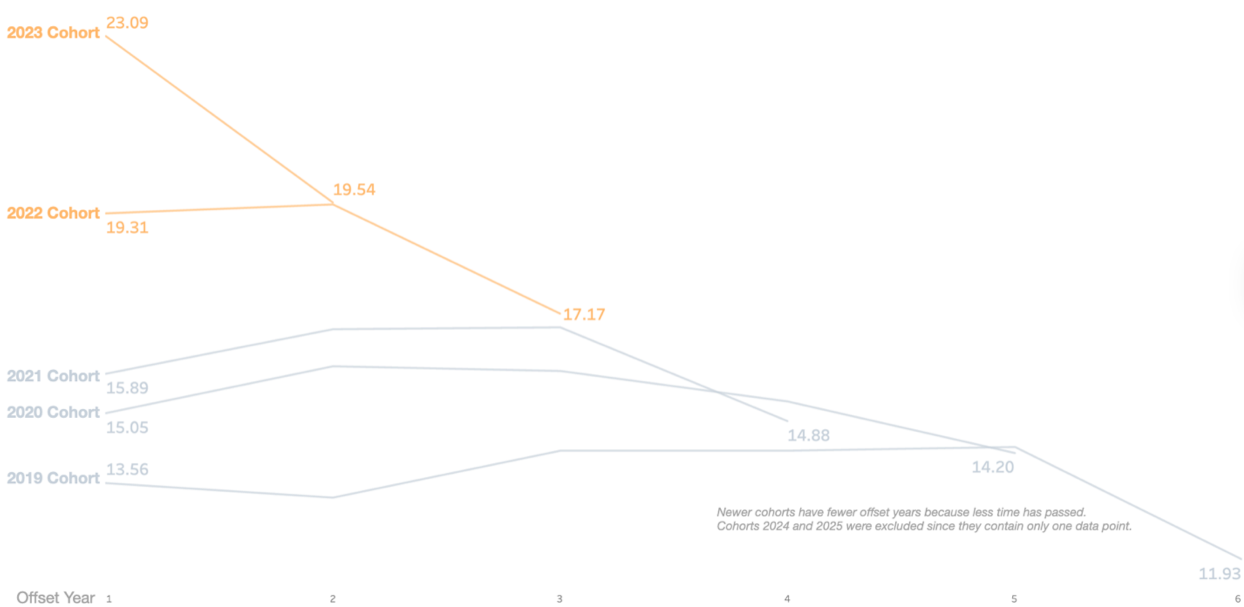

Figure 5 illustrates cohort-level retention from 2019 to 2025, showing how customers repurchase behavior evolves after the first purchase.

Retention patterns reveal two consistent themes:

- Newer cohorts start strong but decline faster. The 2023 cohort begins with a strong repurchase rate of 23% at year 1, but this falls to 19% within 2nd year. The 2022 cohort shows a similar pattern, starting near 19% and declining to 17% in the following period. These steep drops indicate that although recent acquisition efforts bring in high-intent buyers, the business struggles to convert them into long-term customers.
- Older cohorts stabilize at a predictable baseline. Cohorts from 2019-2021 show flatter, more stable curves, clustering around 14-16% retention after Year 1.

Overall insight:

Retention decays rapidly after the first year for all cohorts, with newer groups declining faster. While acquisition drives strong initial activity, long-term value is limited without stronger Year-1 retention actions. Improving the experience, particularly nudging the second purchase remains the most impactful lever for increasing lifetime value.

7.5 Lifetime Value (LTV) & Cumulative ARPU Analysis

Cumulative ARPU (Historical + Predicted)

Year of Coh..	0	1	2	3	4	5	6
2019	$93	$105	$117	$128	$141	$154	$164
2020	$95	$111	$125	$140	$154	$167	$177 (pr)
2021	$94	$111	$127	$143	$158	$170 (pr)	$180 (pr)
2022	$95	$114	$132	$148	$162 (pr)	$175 (pr)	$185 (pr)
2023	$96	$119	$140	$155 (pr)	$170 (pr)	$183 (pr)	$193 (pr)
2024	$100	$129	$147 (pr)	$163 (pr)	$178 (pr)	$192 (pr)	$210 (pr)
2025	$118	$141 (pr)	$161 (pr)	$178 (pr)	$195 (pr)	$210 (pr)	$222 (pr)

Figure 6. Cumulative ARPU by Year Offset and Cohort. (pr) = predicted using cumulative growth %

Customer value was assessed using Average Revenue per User (ARPU) at the cohort level. For each cohort, ARPU was calculated across Year 0 (first purchase) and subsequent offset years to understand how much additional value customers generate beyond their first transaction.

As retention shows that repeat activity after Year 1 declines sharply, and this pattern is confirmed in the ARPU results.

For this project ARPU was modelled at the cohort level using both actual revenue and predicted values. The projects follow the observed historical trend: minimal incremental values is added beyond the first year. The 2024 and 2025 cohorts have slightly higher first-purchase ARPU, but the pattern of rapid decline remains consistent.

This retention behavior has a direct impact on lifetime value. While the 6-year projected LTV reaches approximately an average of $190, most of this value is theoretical because the company has a limited multi-year repeat-purchase history. In reality, nearly 70% of customer value is generated in Year 0-1, and the next two years add only a small incremental lift. For this reason, the 3-year LTV (~$130) provides a more realistic measure of customer value and is better aligned with actual observed behavior. The extended 6 year window should therefore be interpreted as a long-term potential scenario rather than a guaranteed forecast.

8. Recommendations

The analysis indicates that the business has strong acquisition channels, high first-purchase value and clear opportunities to increase long-term customer value. While Year-0 performance is consistently strong, retention and repeat purchase remain the primary constraints on sustainable revenue growth. Based on the insights from RFM segmentation, funnel diagnostics, cohort retention and ARPU/LTV modelling, the following actions are recommended:

8.1 Strengthen Early Lifecycle Retention

- Prioritize customer engagement during the first 90 days, where drop-off is the steepest.
- Introduce onboarding nudges (e.g., personalized email flows, reminders, first-month offers, loyalty programs)
- Optimize post-purchase touchpoints to drive a second order soon.
- A small improvement in Year 1 repeat rate would materially increase LTV and retention curves.

8.2 Reactivate Dormant Customer Segments

- Needs Attention and Hibernating segments represent ~44% of all historical revenue.
- Target these customers with reactivation email sequences and personalized promotions.
- These users are cheaper to re-engage than acquiring new users, improving ROI.
- This supports long-term ARPU growth more efficiently than scaling paid ads.

8.3 Improve Funnel Efficiency on Critical Stages

- The drop between view_item to add_to_cart is consistently the weakest stage.
- Test UX refinements, clearer product information, reviews, returns and price/offer visibility.
- For AdWords and Facebook traffic, focus on audience with higher engagement quality.
- Improving just one or two low-efficiency funnel steps increases conversion at scale.

8.4 Optimize Paid Acquisition with Channel Insights

- Emails remains the most profitable channel; strengthen CRM driven growth.
- Organic search shows the highest YoY growth and should be nurtured for long term advantage.
- Paid channels drive volume but at lower ROI, improve targeting or reduce low-quality spend.
- Use last-touch attribution insights to refine channel budgets quarterly.

8.5 Continue Developing LTV & Cohort Tracking

- Maintain cohort tracking on a quarterly or yearly basis to monitor long-term health.
- Replace 6-year forecasts with shorter 2-3 year LTV models for more realistic planning.
- Build ongoing experiments (discount tests, email flows, onboarding steps) and monitor their effect on retention behavior.
- Integrate retention KPIs into monthly reporting.

9. Conclusion & Key Takeaways

This analysis shows that the business is growing steadily and acquiring customers effectively, but long-term value is constrained by low repeat engagement. Across all cohorts, retention declines rapidly after the first purchase, resulting in roughly 70% of lifetime value being generated in the first year. While projected 3-6 year LTV offers useful directional insight, the most meaningful growth opportunity lies within improving early stage retention, not expanding acquisition.

Customer segmentation reinforces this pattern: the majority of revenue potential sits within Needs Attention and Hibernating customers, indicating a large pool of users who have engaged previously but have not been reactivated. Funnel analysis reveals multiple friction points that contribute to drop-off before conversion.

Overall, the business can achieve stronger, more predictable growth by prioritizing customer lifecycle improvements, especially reactivation, early engagement and post purchase communication. These strategies are most cost effective than acquisition and they directly increase LTV through improved Year-1 retention.

10. Answers to Core Business Questions

1. **Which channels bring customers who actually convert and generate meaningful revenue?**

Email is the strongest traffic source across all funnel stages and contributes nearly half of total revenue. Paid channels deliver high volume but suffer from weak final-stage conversion, losing 57% between Cart to Purchase.

2. **What are the key customer segments and how do their behaviours differ?**

   RFM segmentation shows that Loyal Customers and Champions contribute to ~40% of all revenue with higher frequency and low recency. Needs Attention + Hibernating (44% of revenue) are dormant but high-value, representing the largest reactivation opportunity.

3. **How does user behavior evolve from first purchase to later purchases?**

   Most customers make only one purchase. Repurchases rate drops to ~12%, and Year 1 ARPU declines sharply from $100 to ~29%. Long term behavior stabilizes at very low activity.

4. **Where is the biggest revenue leakage?**

   The main drop-off occurs in the Cart to Purchase stage (56% loss across all channels). This indicates a system wide conversion problem rather than channel specific weakness.

5. **How do customer cohorts retain over time?**

   Retention collapses after the first year. Newer cohorts start higher but decline faster. Older cohorts stabilize around 14-16% by Year 2+.

6. **How much value does each customer generate across their lifecycle?**

   70% of LTV is created in Year 0-1. The projected 6-year LTV is $190 reflects long term potential. A more realistic operational benchmark is 3 year LTV ($135).

7. **Which customer groups represent the largest opportunity for revenue life?**

   Needs Attention + Hibernating segments (44% of revenue) hold the biggest growth potential. Reactivating these users or improving Year-1 retention yields the fastest, most cost-effective LTV uplift.