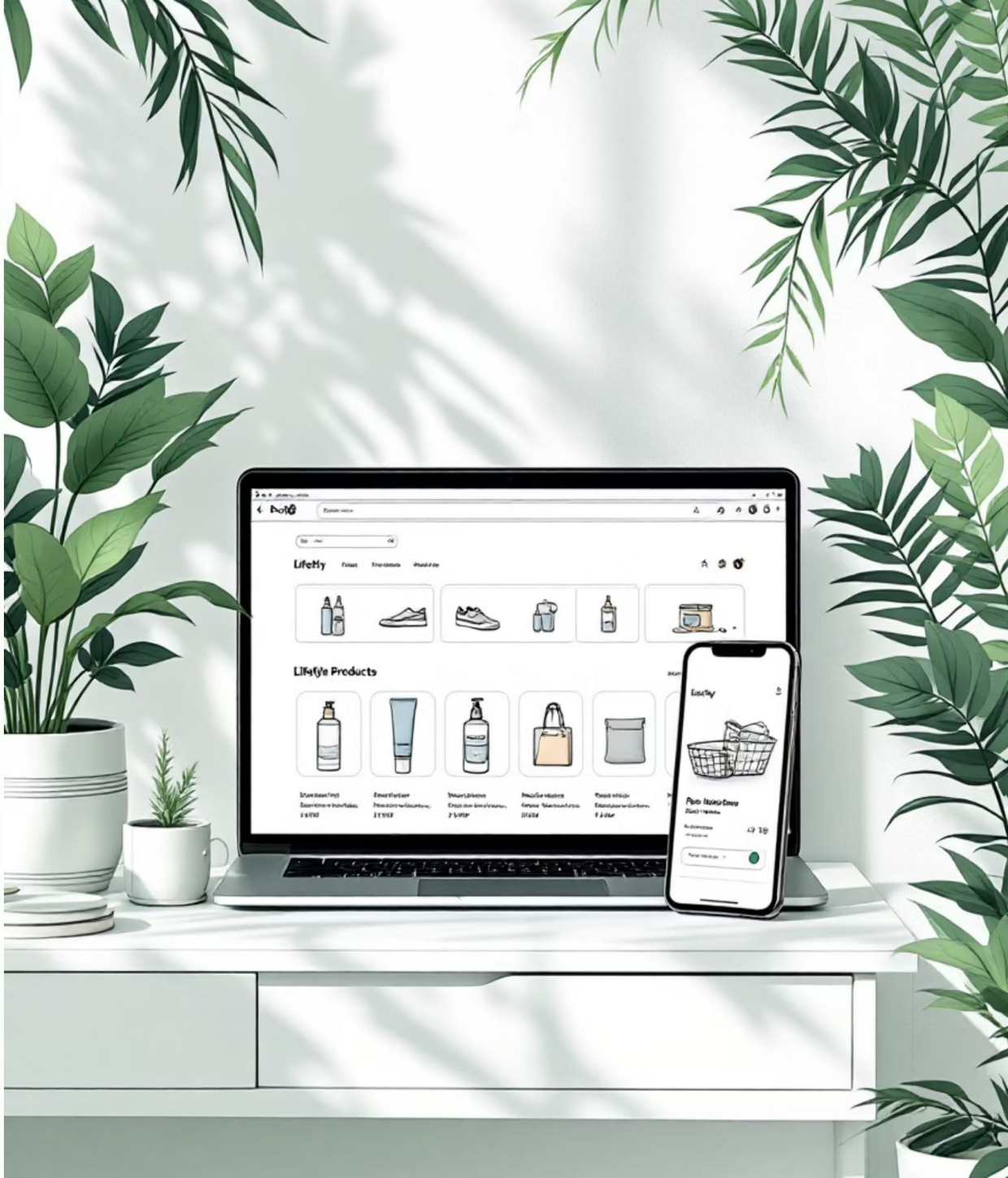
Cohort Analysis for Assessing Customer Retention in the E-commerce Industry



Company Overview

ShopSphere Inc.

Location: United States

Founded in 2012, ShopSphere Inc. started as a niche online marketplace specializing in lifestyle products and home essentials. Over the past decade, it has transformed into a full-scale e-commerce platform with a diverse catalog spanning fashion, electronics, home goods, beauty, and wellness products.

Growth Journey and Milestones:

- **2014-2016:** Expanded beyond lifestyle into electronics and fashion, driving a 250% surge in traffic.
- **2017:** Launched a mobile application, which became a key driver of daily transactions.
- **2019:** Introduced a same-day delivery service across major U.S. metropolitan areas.
- **2021:** Crossed 5 million registered customers and expanded to the U.K. market.
- **2023:** Implemented AI-powered recommendation engines to improve personalization.

Market Presence and Customer Base

Market Reach

Active in both the U.S. and U.K. markets, serving over 8 million customers.

Demographics

A customer demographic ranging from Gen Z online shoppers to middle-aged families.

Positioning

Competitive positioning built around customer-centric policies, quick delivery, and personalized shopping experiences.

Differentiators:

Personalization Engine

Leveraging AI for product recommendations.

Fast Logistics Network

Same-day and next-day deliveries.

Customer Support

Flexible return policies and strong customer support services.

Despite these differentiators, ShopSphere faces one major concern: while customer acquisition remains strong, retention rates have shown a steady decline in the past 18 months.

Business Challenge

The Retention Problem

While ShopSphere enjoys high customer acquisition rates due to aggressive marketing and referral programs, the core challenge lies in customer retention. Specifically:

Declining repeat purchase rate

Many customers make a first purchase but fail to return for subsequent ones.

Customer churn

A growing percentage of new customers disengage after their initial transaction, raising acquisition costs.

Low engagement

Platform analytics show that customers browse less frequently after their first two visits.

Contextual Factors Driving This Challenge:

- Rising competition in the e-commerce sector with platforms offering competitive discounts.
- Shift in consumer expectations toward more personalized experiences.
- Marketing spend focused heavily on acquisition rather than loyalty.
- Operational inefficiencies in converting first-time buyers into long-term customers.

Impact:

- Reduced Customer Lifetime Value (CLV).
- Increased Customer Acquisition Cost (CAC) with diminishing returns.
- Strain on revenue growth and profitability.
- Weakening brand loyalty compared to industry peers.

This scenario makes it crucial for ShopSphere to adopt a data-driven approach to understand customer retention patterns and strategically improve engagement.

Rationale for the Project

Why Cohort Analysis?

> **Cohort Analysis Defined:**
> Cohort Analysis is a data science technique that groups customers based on shared characteristics – typically the time of their first purchase – and tracks their behavior over time. In e-commerce, this allows businesses to measure how well they are retaining customers across different acquisition cohorts.

Relevance in E-commerce:
E-commerce platforms operate in highly competitive spaces where acquiring new customers is expensive. Cohort analysis helps companies understand whether marketing spend translates into long-term engagement or just one-time conversions. For instance, Amazon and ASOS regularly analyze cohorts to fine-tune customer retention strategies.

Top Five Strategic Reasons to Initiate This Project:

Measure Retention Trends
Quantify retention rates across customer cohorts and identify at-risk groups.

Optimize Marketing Spend
Balance investment between acquisition and retention efforts.

Enhance Customer Experience
Identify pain points in the early customer journey and improve onboarding experiences.

Improve CLV and Profitability
Increase the proportion of repeat buyers, boosting lifetime revenue.

Data-Driven Decisions
Enable leadership to make strategic decisions rooted in empirical behavioral patterns rather than assumptions.

Project Objectives

What We Aim to Achieve

The primary objectives of this project are:

Customer Retention Measurement

Track retention patterns across acquisition cohorts to identify how customer behavior evolves over time.

Segmentation of Customers

Group customers into meaningful cohorts (e.g., based on acquisition month, first purchase category, or geography).

Identify Drop-Off Points

Pinpoint the stages where customers disengage (after the first purchase, post-second visit, etc.).

Support CLV Estimation

Enable financial teams to accurately forecast lifetime value across different cohorts.

Actionable Insights for Strategy

Deliver insights to inform targeted marketing campaigns, loyalty programs, and personalized recommendations.

Comprehensive Dataset

Data Structure and Model

The analysis will be based on a synthetic dataset modeled after a real e-commerce transaction log.

Table Name: ShopSphere_records.csv

Column Name	Data Type	Description
InvoiceNo	String	A unique 6-digit integral number assigned to each transaction.
StockCode	String	A 5-digit integral number uniquely identifying each product.
Description	String	Product name/description.
Quantity	Integer	The quantities of each product per transaction.
InvoiceDate	DateTime	The date and time when each transaction was generated.
UnitPrice	Float	The price per unit of the product in GBP.
CustomerID	String	A unique 5-digit integral number identifying each customer.
Country	String	The name of the country where the customer resides.

Data Model:

This is a single-table dataset for simplicity in this initial analysis. The primary key is a composite of InvoiceNo and StockCode, as one invoice can contain multiple items. CustomerID is the foreign key that would link to a hypothetical customer table (containing demographics, signup date, etc.) in a more complex database.

Technology Stack

Tools and Technologies

The technology stack will be beginner-friendly but industry-relevant:

Programming Language

Python

Data Processing

NumPy, Pandas

Visualization

Matplotlib, Seaborn

Statistical Analysis / ML

Scikit-learn (for RFM-based clustering later)

Version Control

Git, GitHub

Collaboration & Documentation

Jupyter Notebooks, Markdown

Data Science Project Scope

CRISP-DM Methodology

The workflow for the project will follow the CRISP-DM methodology adapted to e-commerce analysis:

Business Understanding

Define customer retention issues and align with business KPIs (CLV, churn, repeat purchase rate).

Data Understanding and Ingestion

Explore dataset, check for missing values, and validate structure.

Exploratory Data Analysis (EDA)

- Cohort retention heatmaps.
- Trend analysis across time windows.
- Customer engagement patterns.

Project Workflow Continued

Data Preparation

- Create acquisition cohorts based on InvoiceDate.
- Calculate retention rates month over month.
- Feature engineering (e.g., RFM metrics).

Modeling

- Apply RFM segmentation using clustering techniques (e.g., K-Means or hierarchical clustering).
- Predict propensity to churn using classification models (optional extension).

Version Control

Enable tracking version changes and commits to github.

Reporting and Strategy Recommendations

Recommendations for loyalty programs, personalized marketing, and customer re-engagement strategies.