SEGMENTING SUBSCRIPTION BOX CUSTOMERS USING R FOR CHURN REDUCTION

1. Background

A U.S.-based monthly subscription box company offering curated wellness and lifestyle products experienced a rising churn rate of 31% over six months. While they had demographic and transactional data, the internal marketing team lacked the capability to identify churn signals or build actionable segments. Our task was to use R programming to develop a robust segmentation model based on customer behavior and engagement patterns.

2. Objective

- To segment customers based on behavioral and transactional data using unsupervised machine learning in R
- To identify high-risk churn segments and recommend targeted interventions

3. Data Summary

- **Source:** Customer database (exported via HubSpot and Stripe)
- **Volume:** 14,852 active and past users (Jan–Dec 2023)
- **Variables:**
    - Demographics: Age, Gender, Zip Code
    - Subscription Details: Box Type, Start Date, Renewal History
    - Transactions: Monthly Spend, Failed Payments, Discounts Used
    - Engagement: Email Open Rate, Website Activity, Customer Support Tickets

4. Methodology

4.1 Data Preprocessing

- Removed duplicates and non-U.S. records using dplyr
- Imputed missing numeric values with median and categorical with mode
- Normalized numerical features using the scale() function in R
- One-hot encoded categorical fields such as box type and region

4.2 Feature Engineering

- Created derived features:
  - Average Monthly Spend
  - Tenure in Months
  - Days Since Last Renewal
  - Number of Discounts Applied
  - Total Support Interactions

4.3 Clustering

- PCA reduced 11 features to 4 principal components
- Used Elbow Method and Silhouette Score to identify 5 as optimal cluster count
- Applied kmeans() algorithm and validated with cluster::clusplot() and internal metrics

4.4 Segment Profiling

Cluster	Size	Churn Rate	Characteristics	Retention Strategy
C1	28%	45%	Low spend, frequent discounts, low engagement	Auto-applied loyalty rewards
C2	19%	11%	High spend, active in feedback, low support	Exclusive content + referral bonus
C3	21%	29%	Average spend, new users (<3 months)	Onboarding campaigns + welcome gift
C4	17%	6%	Long tenure, zero complaints, stable spend	Ambassador program
C5	15%	38%	Sporadic purchases, inactive since 2 months	Win-back discount + survey

5. Results

- **Retention pilot targeting C1 and C5** reduced churn by **7.2%** in 2 months
- Custom messaging increased email open rates by **19%** for dormant users
- Lifetime value prediction accuracy improved by 22% after segment tagging

- The segmentation report was adopted into their quarterly analytics pipeline

6. Deliverables

- Fully commented R script (.R file) for EDA, PCA, clustering, and segment profiling
- Visualizations: Scree Plot, Cluster Scatter Plots, Segment Heatmaps (ggplot2, factoextra)
- Report (R Markdown + PDF) summarizing methodology, results, and marketing recommendations
- Final dataset with segment labels and churn probability scores
- A handoff call with the internal data and CRM team for integration

7. Tools and Packages Used

- dplyr, tidyr, ggplot2, cluster, factoextra, caret, readr, psych, stats

8. Future Scope

- Implement real-time churn monitoring by integrating the model into R Shiny dashboard
- Incorporate customer sentiment from reviews and emails using text mining (tm, tidytext)
- Expand segmentation to include social media activity from Facebook and Instagram APIs