BEHAVIORAL SEGMENTATION AND CHURN RISK ANALYSIS USING R FOR A U.S.-BASED SUBSCRIPTION APP

1. Background

A U.S.-based lifestyle and wellness app operating on a freemium subscription model faced high churn rates among paid users. Despite engagement through content and reminders, the team lacked clarity on which behaviors indicated early churn signals. We were brought in to analyze behavioral data in R and produce a detailed churn profiling report for data-driven retention interventions.

2. Objective

- To segment users based on behavioral engagement metrics
- To identify attributes strongly associated with subscription churn
- To guide the design of personalized retention strategies and early warning alerts

3. Data Used

Sources:

- App user interaction logs + subscription status database

Dataset Structure:

- Weekly engagement data for 18,500 premium users (active during Jan–Jun 2023)
- Final structure after cleaning: 240,000 user-week records

Key Variables:

- User_ID, Subscription_Start_Date, Subscription_End_Date, Is_Churned
- Avg_Session_Duration, Workout_Completed, Content_Viewed, Push_Opened, Days_Active_Per_Week, In_App_Purchase_Count, Referral_Code_Used, Region

4. Analysis Methodology

4.1 Data Cleaning and Feature Engineering

- Removed inactive/free users

- Created derived metrics: Total_Engagement_Score, Recency_Score, Weekly_Churn_Flag
- Generated binary flags for churned vs. retained using dplyr and lubridate

4.2 Exploratory Analysis

- Descriptive summaries by churn status
- Plots of session duration and engagement score by week using ggplot2
- Group-wise averages using group_by() and summarise() on usage cohorts

4.3 Segmentation and Statistical Testing

- Performed **k-means clustering** using scaled engagement variables
- Applied **logistic regression** on churn status (glm(family = "binomial"))
- Conducted Chi-squared and t-tests to find significant differences in user traits

5. Key Results

Area	**Insight**
Session Duration	Users with <5 min avg session churned at 2.3× higher rate
Notification Response	Push-open rate <15% was the top churn predictor in logistic model
Recency	65% of churners were inactive in the 2 weeks before subscription expiry
Region	Churn was 40% higher in Midwest users compared to West Coast
Segment Profiles	4 clusters identified – high-engaged, trial-heavy, content-only, inactive

6. Interpretation and Recommendations

- Introduce **renewal reminder nudges** at D-14 for low-recency users
- Set up automated campaigns for users **with push-open rate below 20%**
- Incentivize engagement from content-only users by adding **habit-forming routines**
- Refine onboarding funnel to **flag "under 3 sessions/week" users by week 2**
- Run **regional A/B tests** with revised push copy in Midwest ZIP codes

7. Reporting Output

- **R Markdown Report (PDF, 28 pages)**:
  - Cluster visuals and engagement heatmaps
  - Coefficient tables from churn prediction model
  - Segment-wise churn distribution with explanation
- **Excel Summary Dashboard**:
  - Cluster membership table
  - Churn probability tracker by cohort
  - Segmented user lists for CRM use
- **Code Artifacts**:
  - K-means clustering pipeline
  - Logistic regression model with ROC/AUC plots
  - Pre-built R functions for user flagging and churn scoring

8. Business Impact

- Average churn rate dropped by **7.4% over the next quarter** after applying insights
- Internal retention team adopted **weekly churn probability flagging using R script**
- ROI from campaign optimization increased by **11%** through refined targeting
- Playbook shared with **Series A investors** as part of growth and LTV plan