CUSTOMER SEGMENTATION AND LIFETIME VALUE MODELING USING R: A CASE STUDY FOR A U.S. SUBSCRIPTION BOX BRAND

1. Background

A U.S.-based subscription box company delivering monthly curated self-care products to consumers faced high churn and lacked actionable customer insights. Their marketing efforts were uniform and inefficient due to the absence of segmented audience profiles. The goal was to build a data-driven segmentation model and predict Customer Lifetime Value (CLV) using R.

2. Objective

- To segment customers into distinct groups based on behavioral and transactional variables
- To predict each segment's average lifetime value (LTV) using regression modeling
- To provide actionable insights for segment-specific retention strategies and upselling

3. Data Used

Source: Customer CRM, Shopify transaction logs, support ticket logs

Structure:

- 12,542 customer records
- Timeframe: Jan 2021 to Dec 2023
- Fields: Customer_ID, Signup_Date, Avg_Order_Value, Total_Orders, Last_Purchase_Date, Support_Tickets, Subscription_Length_Months, Referral_Flag

4. Modeling Methodology

4.1 Data Preparation in R

- Used dplyr, lubridate, and janitor for cleaning and formatting
- Created derived fields: Recency, Frequency, Monetary for RFM analysis
- Normalized features using scale()

4.2 Clustering Using K-Means

- Applied elbow method to determine optimal clusters (k = 4)
- Used kmeans() function on scaled RFM and engagement variables

```
kmeans_model <- kmeans(scaled_data, centers = 4, nstart = 25)
```

- Labeled clusters manually:
  - **Segment A**: Loyal, High-Value
  - **Segment B**: New, Potential
  - **Segment C**: Price-Sensitive
  - **Segment D**: At-Risk

4.3 Lifetime Value Modeling

- Modeled LTV as:

$$LTV = Avg\_Order\_Value \text{ x } Subscription\_Length\_Months$$

- Applied multiple linear regression with predictors: Referral_Flag, Support_Tickets, Subscription_Length, Cluster
- Used lm() function and tested assumptions using car and olsrr packages

```
model_ltv <- lm(LTV ~ Referral_Flag + Support_Tickets + Subscription_Length +
as.factor(Cluster), data = customer_data)
```

5. Results

Segment	**Size**	**Avg LTV (\$)**	**Churn Risk**	**Action Strategy**
Segment A	28%	$312	Low	Loyalty rewards & referrals
Segment B	22%	$197	Medium	Welcome offers + upselling
Segment C	34%	$141	High	Discounts + email nudging
Segment D	16%	$105	Very High	Reactivation SMS + feedback

- Regression model adjusted R^2: **0.72**
- Referral_flag and cluster membership were significant predictors ($p < 0.01$)

6. Interpretation and Recommendations

- High LTV correlated with longer subscription lengths and referral behavior

- Support ticket count had a **negative effect** on LTV, suggesting poor experience
- Recommended personalized retention campaigns by segment
- Flagged Segment D customers for urgent win-back strategies
- Suggested testing pricing elasticity for Segment C before campaign deployment

7. Reporting Output

- **R Markdown Report** (25 pages)
  - Cluster profiles
  - LTV regression diagnostics
  - Visuals: boxplots, scatter plots, cluster silhouette plots
- **Executive Dashboard (Excel)**
  - Tab 1: Segment-wise LTV
  - Tab 2: Churn risk triggers
  - Tab 3: Weekly cohort behavior
- **Reusable R Codebase**
  - cluster_customers.R, ltv_model.R, report_generator.Rmd

8. Business Outcome

- Segmentation informed targeted campaigns that lifted retention by **18%**
- Projected revenue uplift of **$94,000** from segment-specific promotions
- Enhanced boardroom-level visibility into **unit economics by segment**
- Framework adopted for quarterly updates in marketing strategy reviews