Table of Content

I. Introduction & Business Context 2
II. Tools & Libraries 2
III. Data Quality & Preprocessing 3
IV. Exploratory Data Analysis Techniques 3
V. Analytical Methods & Validation 4
    1. Statistical Analysis 4
    2. Machine Learning Models 4
    3. Customer Segmentation (Unsupervised Learning) 4
VI. Key Findings from Business Questions 5
    1. Which membership tiers churn more? 5
    2. Do churn rates differ by gender, and how is each churn group composed by gender? 6
    3. Do customers with higher average transaction value churn more frequently? 7
    4. Do customers with higher wallet balances churn less? 8
    5. Is there a churn difference between customers with zero vs. non-zero wallet points? 9
    6. Does higher login frequency correlate with lower churn risk? 10
    7. Which preferred offer type is best for customer satisfaction? 11
    8. Do customers with past complaints churn more? 11
    9. Which features (customer behaviors) have the most impact on churn vs not-churn? (RandomForest + SHAP value) 13
    10. How many separate customer groups based on their behaviors? (Kmean Clustering) 14
VII. Business Recommendations 19
VIII. Conclusion 20
IX. References 21

I. Introduction & Business Context

Our client is a B2B SaaS company operating on a subscription-based revenue model. In such businesses, recurring revenue is the main growth engine, and customer **churn** (customers discontinuing their subscription) poses a direct threat to profitability.

- **Business Problem:** High churn reduces customer lifetime value and increases customer acquisition costs.
- **Objective:** To identify churn drivers, detect high-risk customer groups, and recommend targeted retention strategies.
- **Dataset:** The company provided a dataset (saas_customer_churn.csv) with ~37,000 customer records, including demographic information (age, gender, region), behavioral data (login frequency, time spent), financial activity (transaction value, wallet balance), service interactions (complaints, resolutions), and churn risk scores.

This project applies an **explanatory data analysis pipeline** to uncover insights and guide data-driven decision-making.

II. Tools & Libraries

The analysis was conducted in **Python (Jupyter Notebook)**, using the following key libraries:

- Data Handling & Preprocessing
  - pandas, numpy: data manipulation, missing values, transformations.
  - datetime, re: date/time parsing and cleaning irregular strings.
- Exploratory Data Analysis (EDA)
  - matplotlib, seaborn: visualizations such as bar charts, box plots, histograms, KDE density plots.
  - PercentFormatter: formatting proportions on charts.
- Machine Learning Models
  - scikit-learn: train/test splitting, RandomForest classifier, pipelines, clustering with KMeans, PCA dimensionality reduction.
  - shap: interpretability and feature importance analysis.
- Clustering & Segmentation
  - KMeans: customer segmentation.

    - silhouette_score: validation of clustering quality.

These libraries allow a **full pipeline**: from cleaning raw data → exploratory analysis → predictive modeling → interpretability → customer segmentation.

III. Data Quality & Preprocessing

Initial Audit Findings

- Missing values in key features: region (~15%), wallet points (~9%), preferred offers (~1%).
- Inconsistent categorical values (e.g., “Yes/No/True/Unknown” variations).
- Numeric errors: negative wallet balances, invalid login days, strings like “Error” in numeric fields.
- Outliers: very high transaction values (up to ~100,000).
- Mixed date formats (dd-mm-yyyy, yyyy-mm-dd, dd/Mon/yyyy).

Preprocessing Steps

- **Drop high-cardinality IDs:** security_no and referral_id removed as not useful.
- **Standardize categorical variables:** Normalized Yes/No values, replaced ?, XXXXXXX, Q with “Unknown.”
- **Dates:** Converted joining_date to ISO format (YYYY-MM-DD).
- Numeric corrections:
    - Removed negative balances.
    - Converted “Error” values to zero in login frequency.
    - Removed $ signs from transaction values.
- **Missing value imputation:** Median used for numeric fields; mode/“Unknown” for categorical.
- **Outlier handling:** Values capped at the 99th percentile.
- **Final dataset:** 36,992 records × 22 cleaned columns. No missing values remained.

IV. Exploratory Data Analysis Techniques

A combination of **statistical summaries, visualization techniques, and cross-tabulation** was applied.

- Univariate Analysis

    - Distribution of age, transaction values, wallet balances.
    - Outlier detection using describe() and histograms.
- Bivariate Analysis
    - Churn vs membership category (stacked bar plots).
    - Churn vs gender (nested donut chart).
    - Churn vs spending (KDE density plot).
    - Churn vs wallet balance (boxplots).
- Multivariate Analysis
    - Correlation analysis using Spearman's rank correlation.
    - Feature importance using RandomForest + SHAP.
    - KMeans clustering with PCA projection for visualization.

V. Analytical Methods & Validation

1. Statistical Analysis

    - Descriptive statistics: mean, median, variance, percentiles for numeric fields.
    - Outlier analysis: values capped at 99th percentile.
    - Frequency tables for categorical features.

2. Machine Learning Models

- **RandomForest Classifier:** Built on numerical + encoded categorical features.
- Validation:
    - Accuracy and interpretability prioritized over complex models.
    - **SHAP values** used to interpret feature importance.

    Key Insights from SHAP:

    - Wallet balance → strongest predictor (higher = lower churn).
    - Membership tier → Premium/Platinum reduce churn, No/Basic increase churn.
    - Transaction value and login frequency → moderate importance.
    - Demographics (age, gender) → minimal effect.

3. Customer Segmentation (Unsupervised Learning)

    - **KMeans clustering** with standardized features.
    - Validation methods:

- **Elbow method** → optimal clusters between k=4 and k=5.
- **PCA visualization** → clusters displayed clearly in 2D space.

VI. Key Findings from Business Questions

1. Which membership tiers churn more?

- **Visualization:**

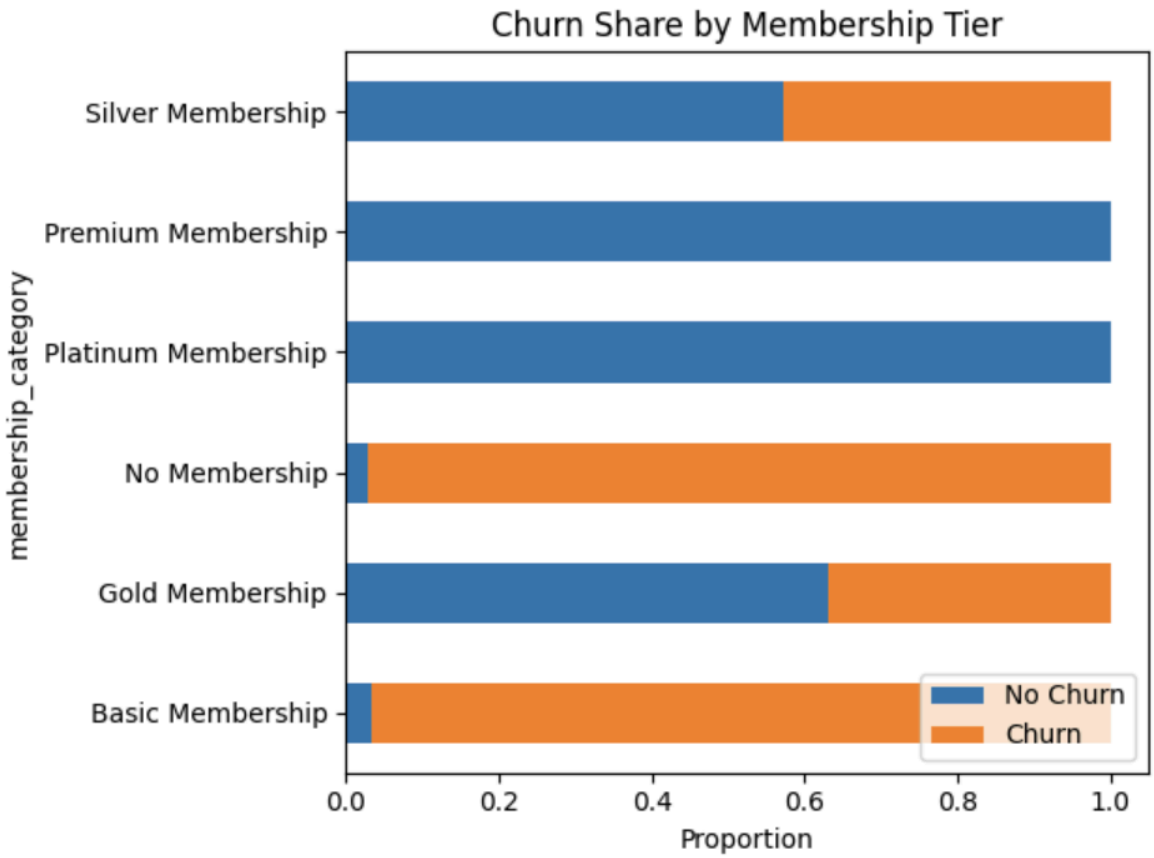

Figure 1: Membership and Churn chart

- **Result:**
  - **Highest churn:** No Membership, Basic → weak perceived value; target upgrades/onboarding.
  - **Mid churn:** Silver, Gold → better but still meaningful; reinforce value.
  - **Lowest churn:** Platinum, Premium → near-zero churn; maintain benefits.
  - **Action:** Prioritize retention offers for entry tiers; protect premium experience.

2. Do churn rates differ by gender, and how is each churn group composed by gender?

- **Visualization:**

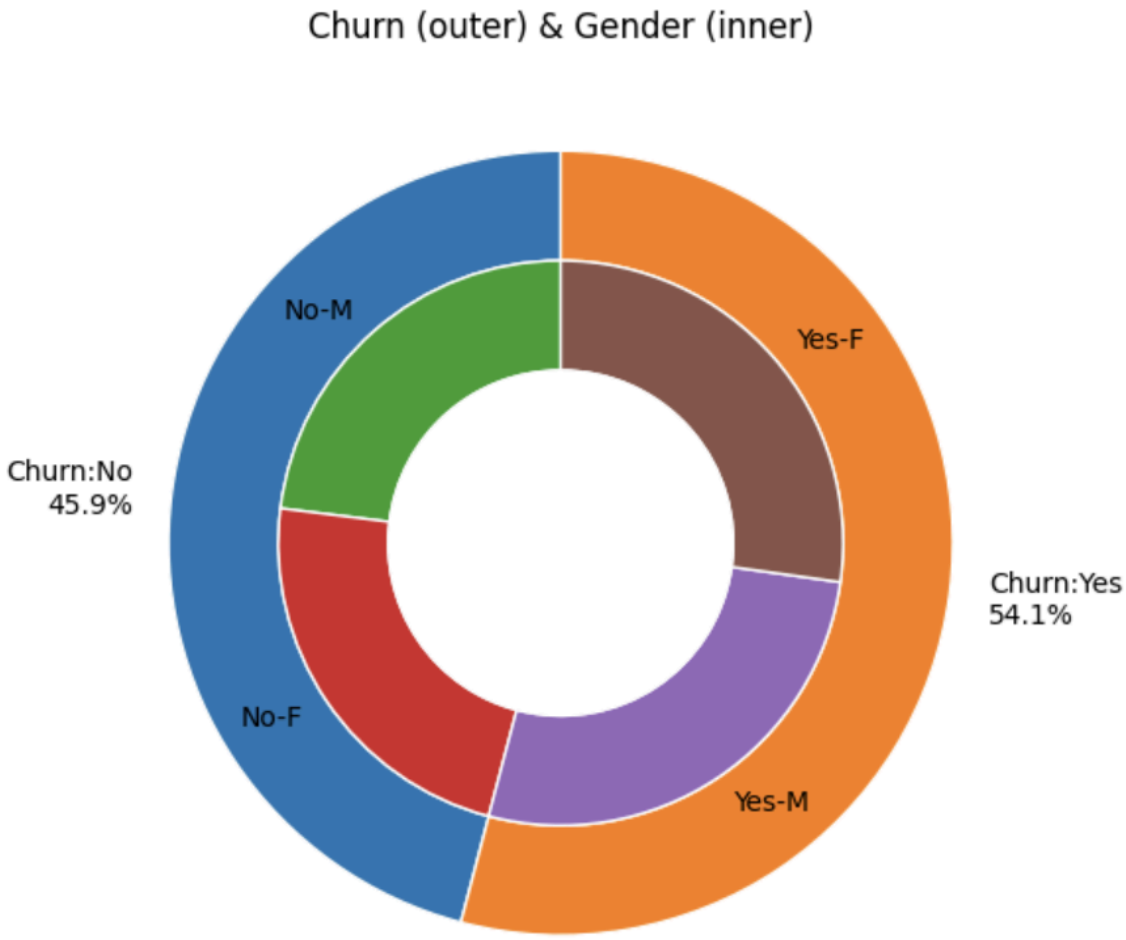

Figure 2: Gender and Churn chart

- **Result:**
  - Both **male and female customers churn significantly**, with no major gender gap.
  - Gender alone does **not appear to be a strong differentiator of churn**.

Recommendation:

Focus churn analysis on **behavioral features** (usage frequency, spending, complaints) rather than gender. Gender can still be used for **marketing segmentation**, but not as a primary churn predictor.

3. Do customers with higher average transaction value churn more frequently?

- **Visualization:**

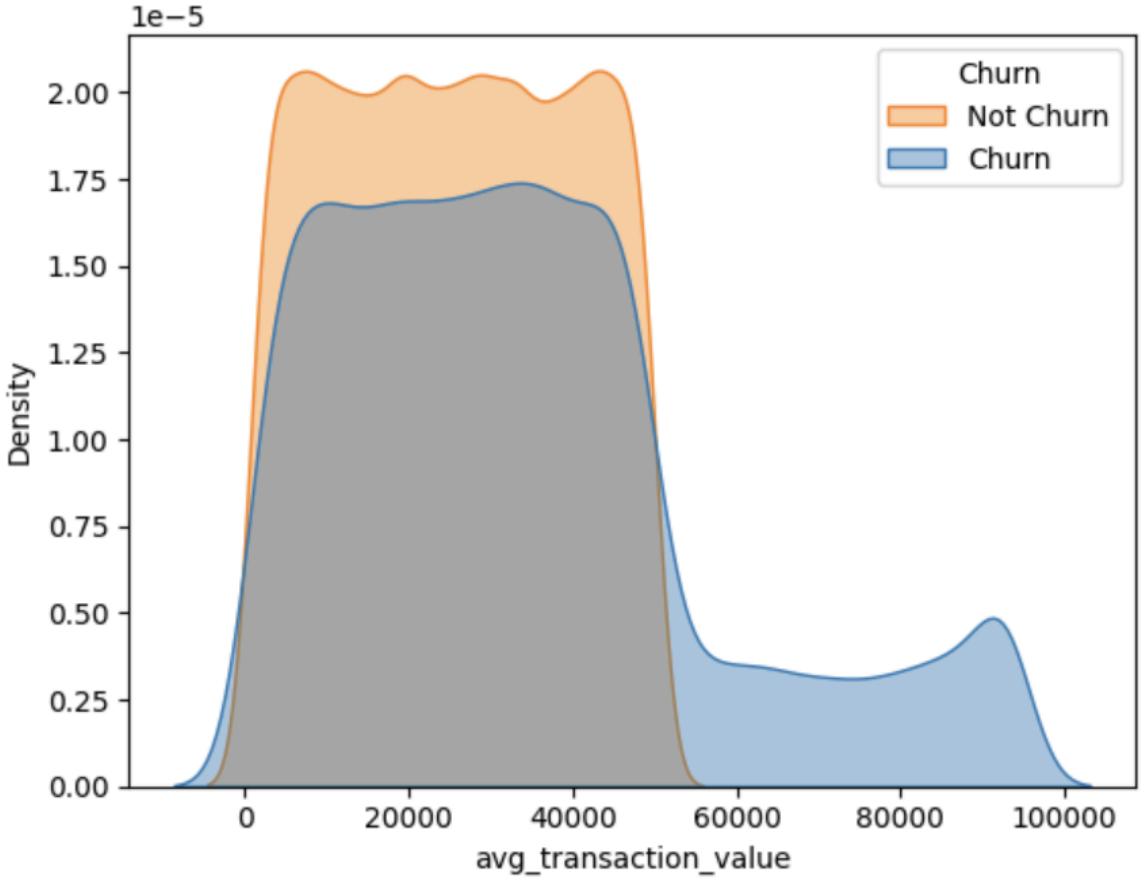

Figure 3: Average Transaction Value and Churn chart

- **Result:**
  - Customers with **lower avg transaction values (0–40K)** are more often **not churn** → low spenders are more stable.
  - Customers with **higher transaction values (60K–100K)** show higher **churn density** → big spenders are more at risk.
  - Suggests **price sensitivity**: heavy spenders may churn if they feel costs outweigh benefits.

Recommendation:

Target high-transaction-value customers with **loyalty rewards, premium service, or discounts** to reduce churn.

4. Do customers with higher wallet balances churn less?

- **Visualization:**

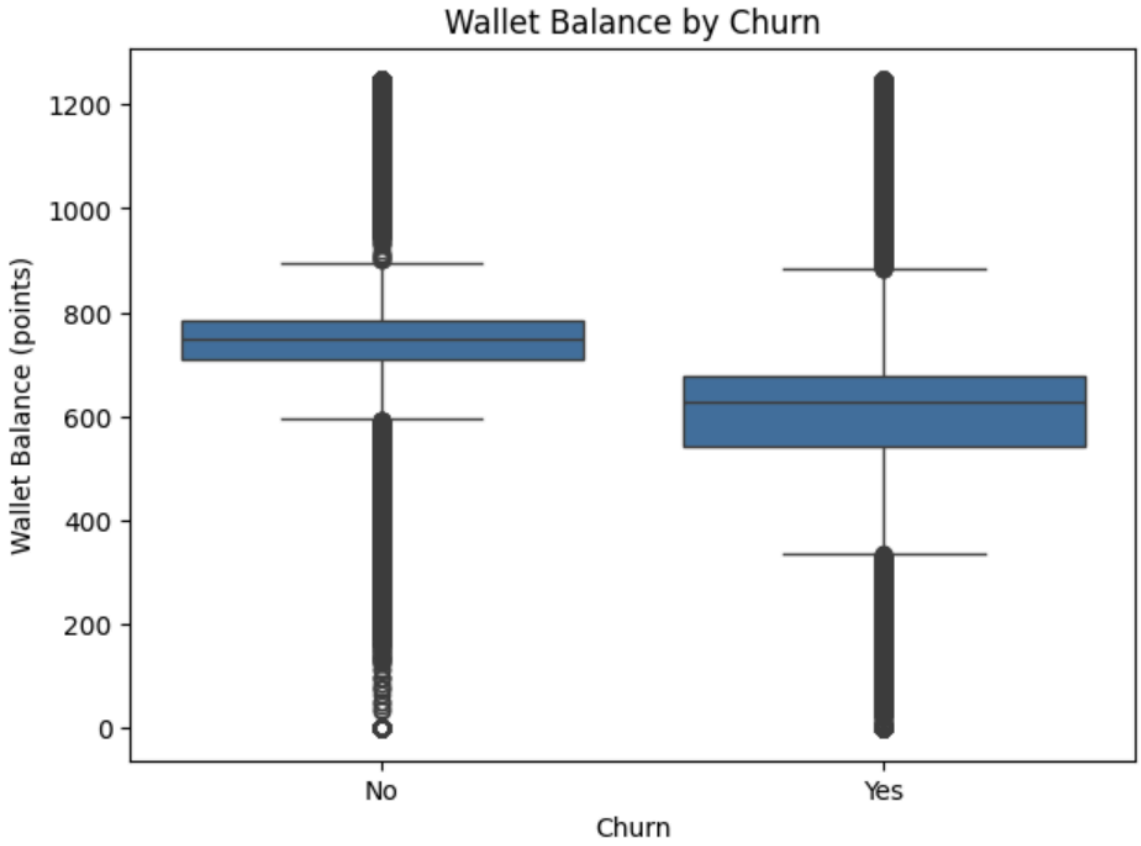

Figure 4: Wallet Balance and Churn chart

- **Box Plot Interpretation: Wallet Balance by Churn**
  Median Balance
    - Non-churned customers (No) have a higher median wallet balance (~750).
    - Churned customers (Yes) have a lower median (~620).
      => Suggests customers with more wallet balance are less likely to churn.

  Spread (IQR – interquartile range)
    - The Non-churn group has a tighter spread (most balances between ~600–800).
    - The Churn group shows a wider spread, meaning more variation in balances.

Outliers

- Both groups have extreme values (very low and very high balances).
- Churned customers show more low-balance cases.

Overall Trend

- Higher wallet balances are linked to lower churn.
- Customers with low wallet balances appear more at risk of churn.

5. Is there a churn difference between customers with zero vs. non-zero wallet points?

- **Visualization:**

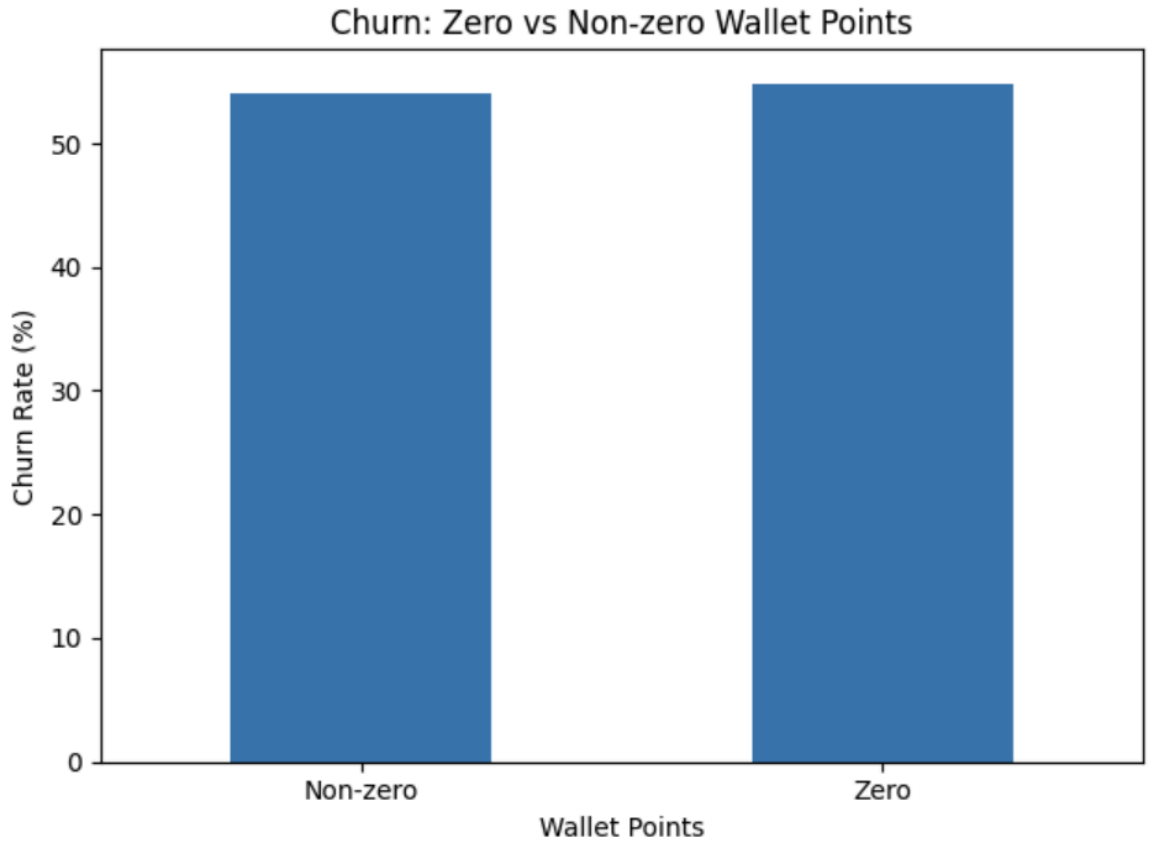

Figure 5: Zero and Non-Zero Wallet Point chart

- **Result:**
  - Churn is ~54–55% for both groups (difference < 1%).
  - Simply having wallet points does **not** meaningfully reduce churn.
  - Focus retention on other drivers (usage, tier, offers).

6. Does higher login frequency correlate with lower churn risk?

- **Visualization:**

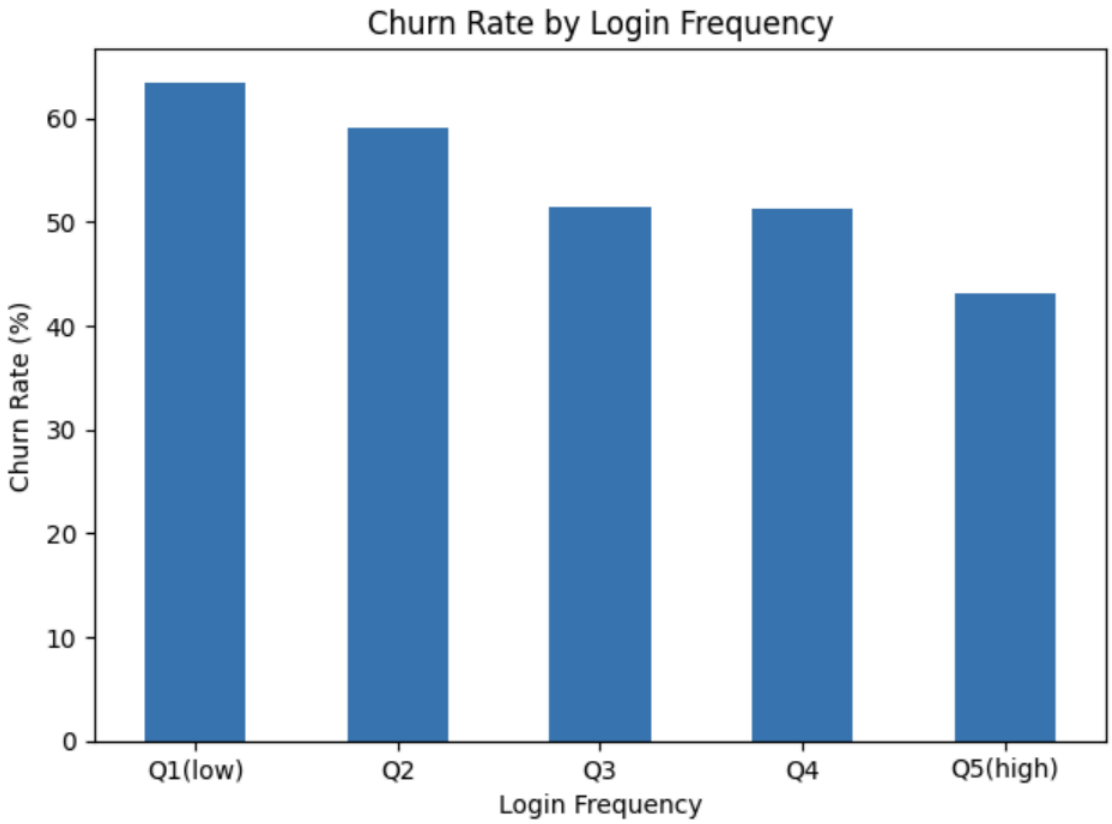

Figure 6: Login Frequency and Churn chart

- **Result:**
    - **Correlation**:
      The Spearman correlation is **-0.151**, showing a **negative relationship** between login frequency and churn risk.
      → Customers who log in more frequently are slightly less likely to churn.
    - **Churn by Frequency Buckets**:
        - **Q1 (lowest frequency)** → Churn rate ~63.5% (highest risk).
        - **Q5 (highest frequency)** → Churn rate ~43.1% (lowest risk).
        - Churn rate steadily declines from **low-frequency to high-frequency users**.
    - **Overall Insight**:
      Customers who log in more often tend to stay engaged and are **less**

likely to churn, while infrequent logins are a strong indicator of churn risk.

Encouraging **more frequent customer logins** (e.g., engagement campaigns, gamification, or reminders) could help reduce churn.

7. Which preferred offer type is best for customer satisfaction?

- **Visualization:**

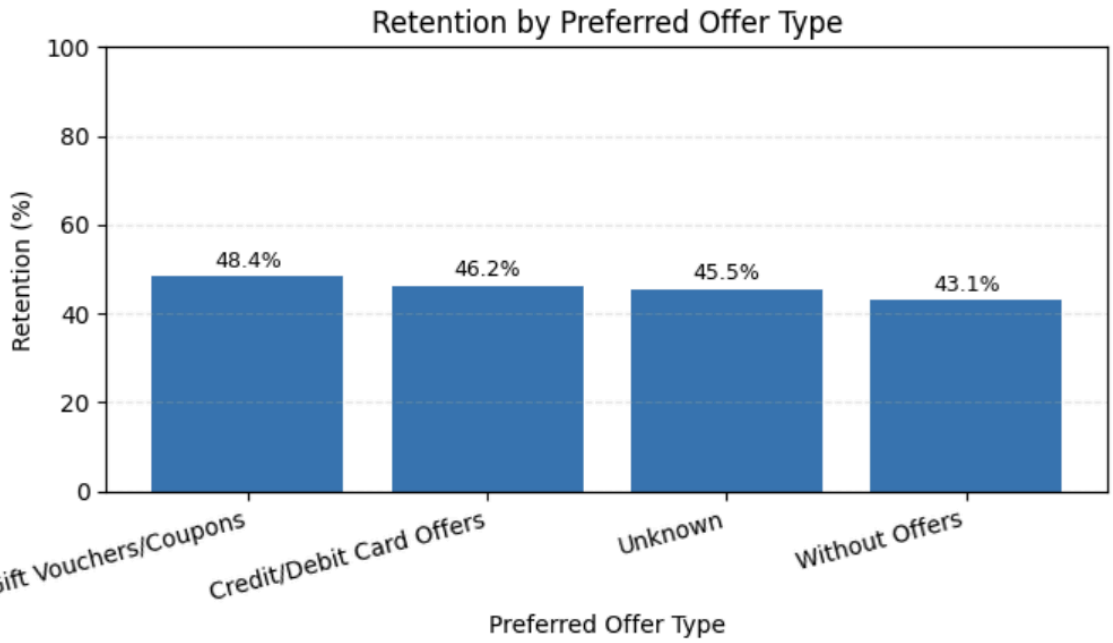

Figure 7: Preferred Offer Type and Churn chart

- **Result:**
  - The gap is modest (~5 pp) from best (Vouchers) to lowest (No offers).
  - This may be statistically small and operationally marginal.

8. Do customers with past complaints churn more?

- **Visualization:**

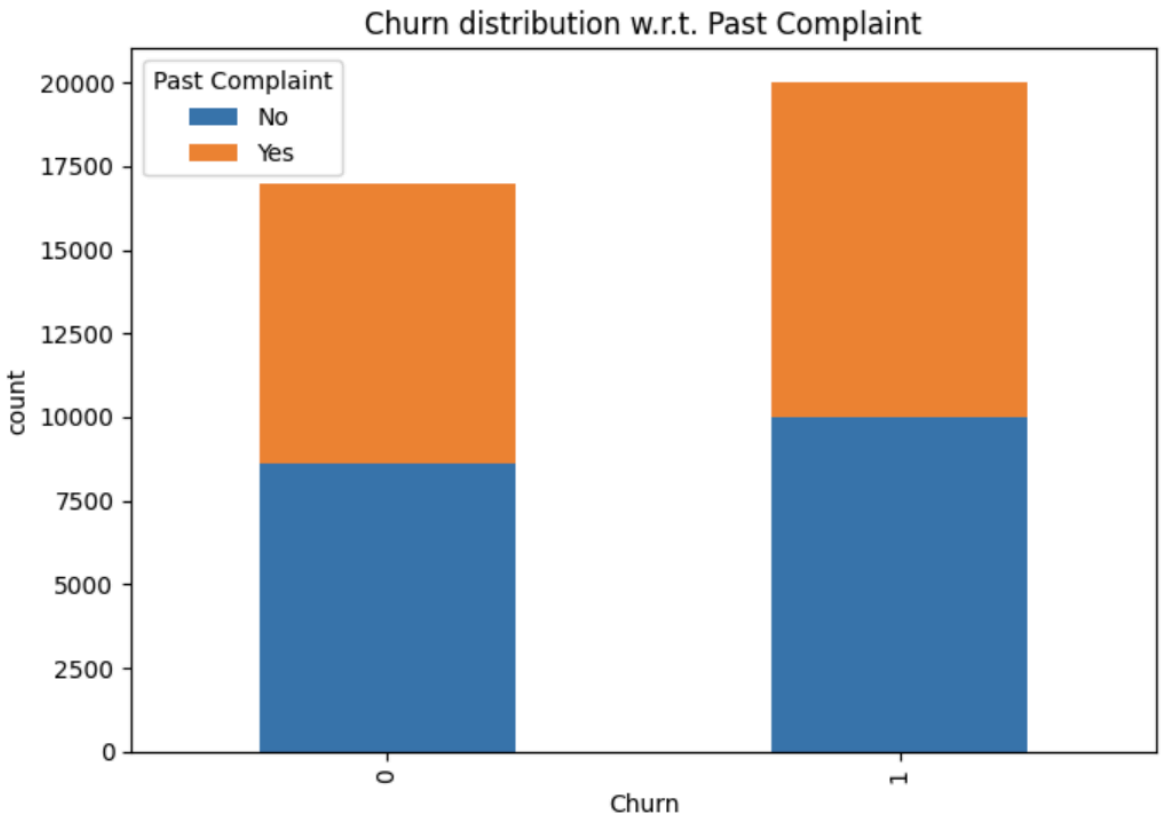

Figure 8.1: Past Complaints Numbers and Churn chart

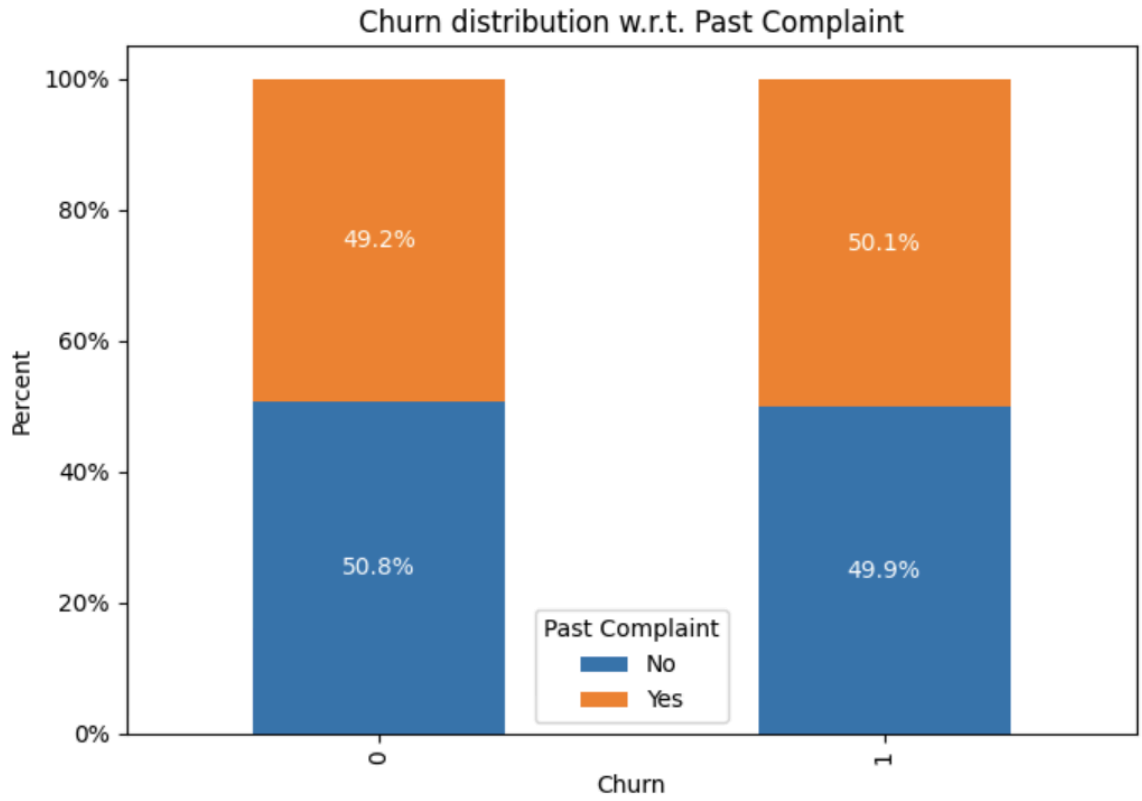

Figure 8.2: Past Complaints Percentages and Churn chart

- **Result:**
    - Churn shares are ~50/50 for both groups → **past complaints alone don't drive churn**.
    - Likely explanation: **post-complaint support is effective**, neutralizing risk.

  Recommendation: Maintain/scale the current service-recovery playbook (fast SLA, follow-ups). Use complaints as a **trigger** to offer loyalty perks or onboarding tips, but prioritize other risk signals (low usage, entry tiers, non-referral) for retention targeting.

9. Which features (customer behaviors) have the most impact on churn vs not-churn? (RandomForest + SHAP value)

- **Visualization:**

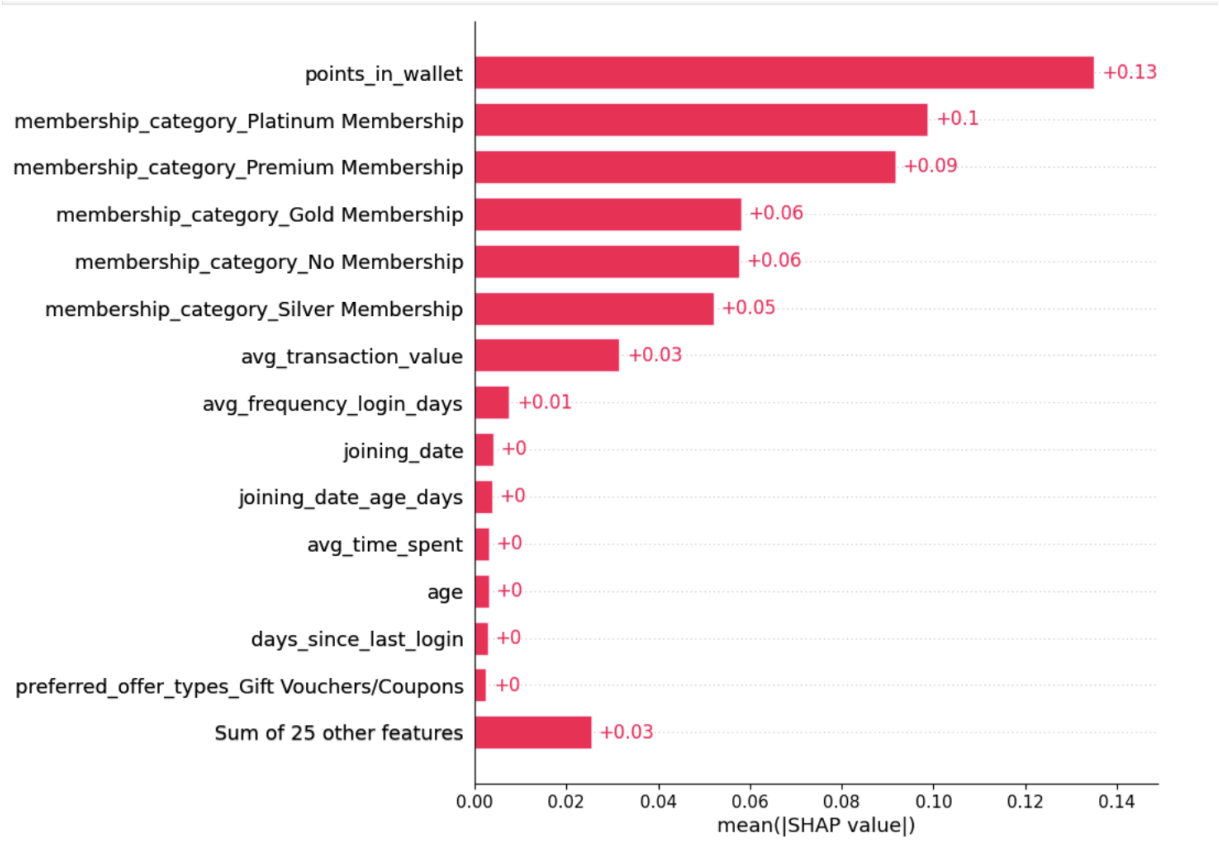

Figure 9: Features (Customer Behaviors) have the most impact on Churn vs Not-Churn

- **Result:**
    - **Wallet balance** → stronge factor; higher balance = lower churn.
    - **Membership tier** → Platinum/Premium/Gold reduce churn; No/Silver increase it.
    - **Spending** (avg_transaction_value) → moderate impact; higher spenders churn less.
    - **Engagement** (login frequency) → small but consistent; frequent users churn less.
    - **Tenure & demographics** → minimal effect.

  Recommendation: Focus retention on **wallet incentives** and **membership upgrades**, with re-engagement for low-activity users.

10. How many separate customer groups based on their behaviors? (Kmean Clustering)

- **Visualization:**

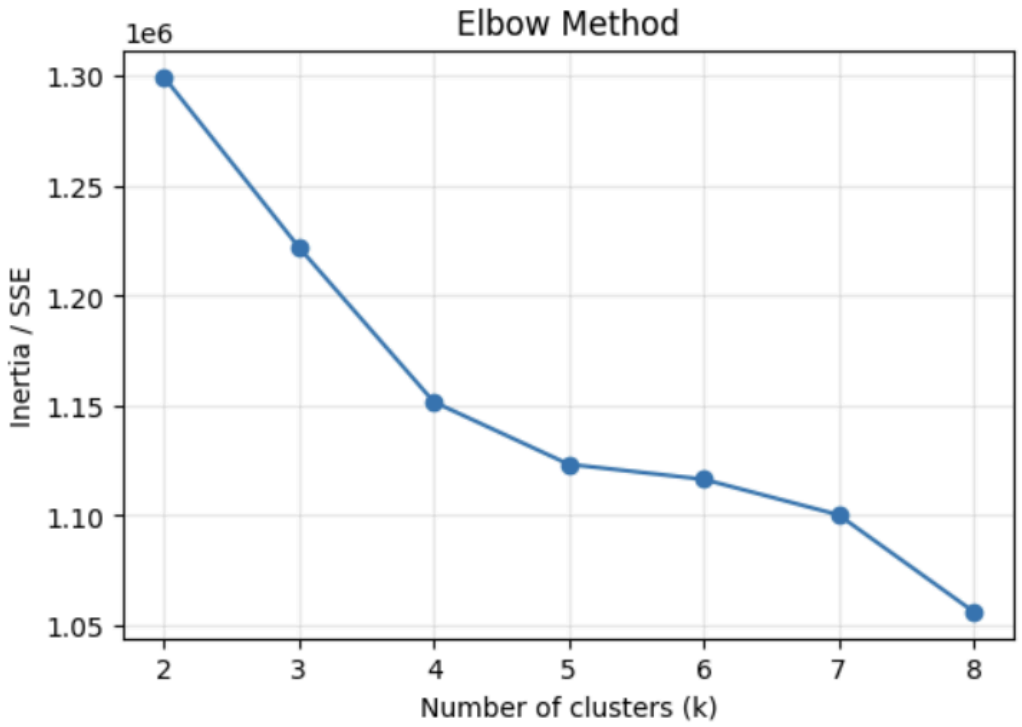

Figure 10.1: Elbow Chart for optimum-k

- **Conclusion: Best K for Customer Segmentation**
    - From the Elbow method, both **k=4** and **k=5** looked plausible.
    - However, the PCA scatter plots show that with **k=4**, clusters are overlapping and not well separated.
    - With **k=5**, the clusters are **clearer and more distinct**, suggesting that the optimal number of customer groups is **5**.

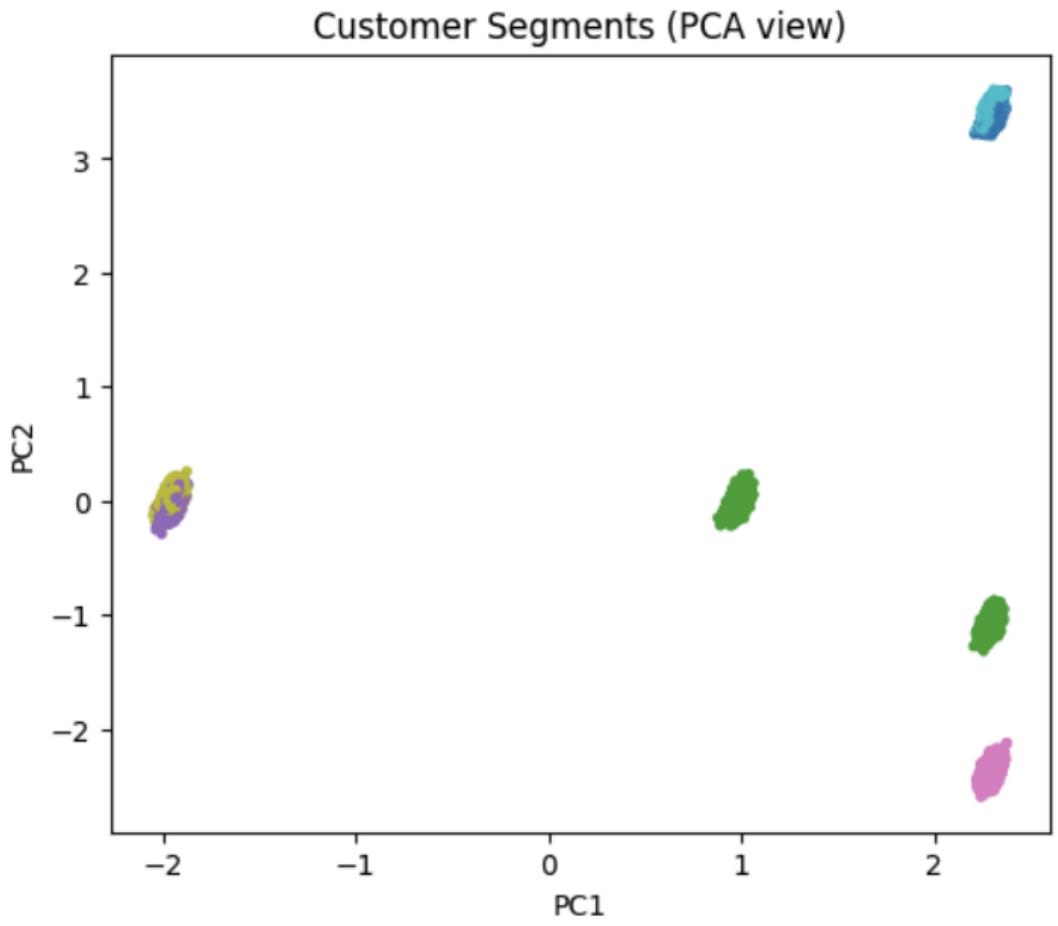

Figure 10.2: Customer Clustering with k=5

- **Customer Groups' Main Features:**

  Cluster 0: top behaviors (z-score sign shows direction)

  - internet_option_Mobile_Data: much higher than the average of all clusters (2.02 SD)
  - avg_frequency_login_days: much higher than the average of all clusters (2.02 SD)
  - preferred_offer_types_Without Offers: much lower than the average of all clusters (1.98 SD)
  - - avg_time_spent: much lower than the average of all clusters (1.91 SD)
  - - region_category_Village: much higher than the average of all clusters (1.87 SD)

  Cluster 1: top behaviors (z-score sign shows direction)

- complaint_status_Solved in Follow-up: much higher than the average of all clusters (2.24 SD)
- feedback_Solved in Follow-up: much higher than the average of all clusters (2.24 SD)
- joined_through_referral_Unknown: much higher than the average of all clusters (2.16 SD)
- gender_M: much lower than the average of all clusters (1.09 SD)
- preferred_offer_types_Gift Vouchers/Coupons: much lower than the average of all clusters (1.07 SD)

Cluster 2: top behaviors (z-score sign shows direction)

- - age: much higher than the average of all clusters (1.52 SD)
- - membership_category_No Membership: much lower than the average of all clusters (1.50 SD)
- - feedback_Not Applicable: much higher than the average of all clusters (1.41 SD)
- - complaint_status_Not Applicable: much higher than the average of all clusters (1.41 SD)
- - past_complaint_Yes: much lower than the average of all clusters (1.41 SD)

Cluster 3: top behaviors (z-score sign shows direction)

- - feedback_Solved: much higher than the average of all clusters (2.24 SD)
- - complaint_status_Solved: much higher than the average of all clusters (2.24 SD)
- - joined_through_referral_Yes: much lower than the average of all clusters (2.01 SD)
- - membership_category_Platinum Membership: much lower than the average of all clusters (1.51 SD)
- - offer_application_preference_Yes: much higher than the average of all clusters (1.25 SD)

Cluster 4: top behaviors (z-score sign shows direction)

- - age: much lower than the average of all clusters (1.78 SD)

- - used_special_discount_Yes: much lower than the average of all clusters (1.65 SD)
- - offer_application_preference_Yes: much higher than the average of all clusters (1.50 SD)
- - feedback_Not Applicable: much higher than the average of all clusters (1.41 SD)
- - complaint_status_Not Applicable: much higher than the average of all clusters (1.41 SD)

- **Conclusion:**

  How many groups?

  Using the elbow method and PCA separation, the data are best segmented into 5 distinct customer groups.

  High-level profiles (top behaviors per cluster)

  - Cluster 0 – Mobile & frequent users (Village-leaning)
    - Much higher mobile-data internet usage and higher login frequency.
    - Lower time spent per session; more Village region.
    - Dislike “Without Offers” (prefer to have some kind of offer available).
  - Cluster 1 – Issues solved via follow-up (female-leaning)
    - Complaints and feedback solved in follow-up far higher than average.
    - Referral source unknown is higher; male share is lower.
    - Lower interest in Gift Vouchers/Coupons.
  - Cluster 2 – Older, established members with few complaints
    - Older age; fewer “No Membership” (i.e., more are members).
    - Complaint/feedback often “Not Applicable”; past complaints lower.
  - Cluster 3 – Quick resolutions & offer-active
    - Solved complaints/feedback much higher.
    - Less “joined by referral”; fewer Platinum members.
    - Higher willingness to apply for offers.
  - Cluster 4 – Younger, offer-applicants but discount-light
    - Younger age.

    - Less use of special discounts, yet more likely to apply for offers.
    - Complaint/feedback often “Not Applicable”.

Takeaway:

Five segments provide clear, behaviorally distinct groups useful for targeted offers, complaint-handling strategies, and channel personalization.

VII. Business Recommendations

Summary of Findings

- **Membership tier is the strongest churn driver**: Low tiers (No/Silver/Basic) have high churn; Premium/Platinum show strong loyalty.
- **Financial behavior matters**: Higher wallet balances reduce churn; high-spending customers are more churn-sensitive.
- **Engagement reduces churn**: Frequent logins lower churn risk; infrequent logins are a warning signal.
- **Complaints are not churn drivers** if resolved quickly—service recovery is effective.
- **Demographics (gender, offers)** play a minor role, but offer type still influences behavior.
- **Best segmentation = 5 clusters** with distinct profiles (mobile-heavy users, follow-up resolution seekers, older loyal members, quick-resolution/offer users, younger discount-light users).

Business Decision-Making

1. **Retention & Loyalty Programs**
    - Push **wallet incentives and tier upgrades** for No/Silver/Basic members.
    - Protect high-value Premium/Platinum members with **exclusive perks**.
2. **Customer Engagement**
    - Increase logins via **gamification, reminders, app campaigns**.
    - For low-frequency users, design **reactivation offers**.
3. **Complaint Management**

- Keep **fast SLA and follow-up processes**, as they neutralize churn risk.
- Use complaints as a **trigger for loyalty outreach**.

4. **Pricing & Offers**
   - Manage high-spender churn with **targeted discounts, premium services, or loyalty rewards**.
   - Segment offers: vouchers for price-sensitive clusters, personalized deals for offer-active clusters.
5. **Segment-Specific Strategies**
   - **Cluster 0** (Mobile & frequent users) → Push app-based offers and regional campaigns.
   - **Cluster 1** (Follow-up resolution seekers) → Emphasize customer care quality.
   - **Cluster 2** (Older loyal members) → Build long-term value programs.
   - **Cluster 3** (Quick-resolution, offer-active) → Upsell through time-limited offers.
   - **Cluster 4** (Young, discount-light) → Promote innovative bundles instead of heavy discounts.

Takeaway

- Churn is best reduced by **upgrading entry-tier members, incentivizing wallet balance, and boosting engagement**.
- The **5 behavioral clusters** enable **personalized marketing and service strategies**, turning insights into **actionable business decisions**.

VIII. Conclusion

This project demonstrates the value of combining **data quality checks, exploratory analysis, machine learning, interpretability, and clustering** to understand customer churn.

- Strengths of analysis:
  - Rigorous cleaning and preprocessing.
  - Multiple techniques (visual EDA, statistical analysis, ML, clustering).
  - Interpretability through SHAP values.
  - Actionable segmentation into 5 customer groups.
- Limitations:

- Dataset provides churn risk scores rather than historical churn events.
- External factors (market competition, pricing strategy) are not included.
- Advanced model tuning was not the main focus.

Final Takeaway:
Customer churn can be reduced by focusing on:

- Upgrading entry-level members.
- Boosting wallet balances with incentives.
- Increasing engagement via logins.
- Maintaining strong service recovery.
- Deploying cluster-based personalized marketing.

These insights translate directly into **data-driven strategies** for marketing, product, and customer success teams, enabling the SaaS company to improve retention and long-term revenue.

IX. References

1. Drive Link:

   https://drive.google.com/drive/folders/1uCo6YthRRWhnZnL-gSexny5WuyBnvwUP?usp=sharing

2. Github:

   https://github.com/tuanTaAnh/SaaS-Customer-Churn-Consulting

3. Video Link:

   https://drive.google.com/file/d/1yNLPKje8U1GLvebs7fPb5hQJMwDPIC5D/view?usp=sharing

(*) You only need to access one of 2 links above for code file (Drive Link or Github)
(**) Because of my OS (MacOS), some installation instructions in Readme could cause problems for you if you are using Windows or Ubuntu.