SmartBank Customer Churn Analysis: Data Preparation & Exploratory Report

Executive Summary

This report presents the first phase of a churn analysis project for **SmartBank**, aimed at understanding customer behavior and laying the groundwork for future predictive modeling.

Key findings show that **churn is not strongly associated** with internal data such as demographics, transaction history, or service interactions, suggesting that **external factors** may influence churn decisions.

Despite this, clustering revealed **three distinct behavioral personas**, each with unique characteristics and levels of engagement. These segments were used to craft **targeted strategic recommendations**, including loyalty incentives, reactivation campaigns, and product cross-selling.

The deliverables include a **cleaned and engineered dataset**, detailed exploratory analysis, and **customer personas** that SmartBank can leverage to improve retention, even in the absence of strong churn predictors.

Future steps involve predictive modeling, integration of external data sources, and experimentation with retention strategies tailored to each segment.

1. Introduction

Customer churn — the rate at which customers stop doing business with a company — is a critical metric in any business sector. For SmartBank, reducing churn is directly tied to long-term profitability, as acquiring new customers often costs significantly more than retaining existing ones. Understanding the behavioural and service-related factors that lead to churn can enable the bank to proactively engage at-risk customers, improve service delivery, and enhance loyalty programs.

This project focuses on laying the groundwork for a churn prediction system by conducting a thorough exploratory data analysis (EDA) and preparing a clean, feature-rich dataset that captures customer behaviour and service interactions over time.

The objectives of this phase are as follows:

- **Identify and Gather Data**: Review the provided data sources and select those most relevant for predicting customer churn, with a focus on customer demographics, transaction history, service interactions, and online activity. Each dataset is selected based on its potential to uncover meaningful behavioural and service-related insights.

- **Perform Exploratory Data Analysis (EDA)**: Use visualisations (e.g., histograms, box plots, scatter plots) and statistical summaries to explore the data, identify trends and relationships, and uncover potential drivers of churn such as service complaints, low engagement, or transaction frequency.

- **Clean and Preprocess the Data**: Handle missing values using appropriate methods, address outliers, normalise or standardise numerical features, and encode categorical data. These steps ensure the dataset is consistent, interpretable, and ready for machine learning.

- **Segment Customers**: Identify distinct behavioural clusters using unsupervised learning techniques to uncover actionable customer personas that can support targeted strategies and business decisions.

- **Extract Churn Insights**: Assess the relationship between churn and key variables — including demographic, behavioural, and service-related features — to support early detection of churn risks, even when direct associations are weak or non-significant.

This foundational work will guide the development of predictive models and data-driven strategies aimed at improving customer retention and overall satisfaction at SmartBank.

2. Data Collection & Integration

This section describes the data sources selected for the churn analysis, the rationale for their inclusion, and the key feature-engineering steps applied before merging. These consolidated data form the foundation for all subsequent exploratory analysis and model development.

2.1 Data Sources Overview

We leveraged five core datasets, each contributing a distinct perspective on customer behaviour:

Dataset	**Description**	**Key Raw Fields**

Customer Demographics	Static attributes describing the customer profile.	CustomerID, Age, Gender, MaritalStatus, IncomeLevel
Transaction History	Records of every purchase, including amount, date, and product category.	CustomerID, TransactionDate, AmountSpent, ProductCategory
Customer Service	Logs of customer support interactions, including type and resolution status.	CustomerID, InteractionDate, InteractionType, ResolutionStatus
Online Activity	Digital engagement metrics such as login frequency and service channel used.	CustomerID, LastLoginDate, LoginFrequency, ServiceUsage
Churn Status	Binary label indicating whether a customer ultimately churned (1) or was retained (0).	CustomerID, ChurnStatus

2.2 Rationale for Inclusion

1. **Customer Demographics**
   Demographic attributes (age, gender, marital status, income level) can influence banking needs and risk of churn. We included these to test for any socio-economic correlations with churn behaviour.

2. **Transaction History**
   Purchase frequency, monetary value, and recency are classical predictors of customer value and loyalty. Aggregating transaction data yields key RFM (Recency, Frequency, Monetary) metrics.

3. **Customer Service Records**
   The volume, type, and resolution of support interactions are strong proxies for customer satisfaction. Both resolved and unresolved complaint counts were extracted to capture friction points.

4. **Online Activity Logs**
   Digital engagement — especially login frequency and platform preference (web vs mobile) — reflects ongoing customer interest and ease of use of SmartBank's digital

channels.

5. **Churn Status**
   The target variable for modeling. Ensuring this label is accurately linked to each customer is critical for all supervised learning and evaluation.

2.3 Feature Engineering & Aggregation

To prevent data duplication and ensure one record per customer, we performed the following aggregations and transformations **before** merging:

1. **Transaction-Based Features**
   - **TotalSpent**: Sum of AmountSpent per customer.
   - **AvgSpent**: Mean spend per transaction.
   - **NumTransactions**: Count of transactions.
   - **FirstPurchase / LastPurchase**: Dates of earliest and most recent transactions.
2. **Service-Based Features**
   - **NumInteractions**: Total support interactions.
   - **UniqueInteractionTypes**: Distinct interaction categories (Inquiry, Complaint, Feedback).
   - **NumComplaints**, **ResolvedComplaints**, **UnresolvedComplaints**: Complaint volumes by resolution status.
   - **HadComplaint / HasUnresolvedComplaint**: Binary flags indicating any complaint or any unresolved complaint.
   - **MultipleInteractionTypes**: Flag for >1 interaction category.
   - **FirstInteraction / LastInteraction**: Dates of earliest and latest support contacts.
   - **InteractionSpanDays**: Days elapsed between first and last interaction.
   - **InteractionFrequency**: Interactions per day over the interaction span.

  - **DaysSinceLastInteraction**: Days from last interaction to data-pull date.

2.4 Merging Strategy

- **One-to-Many Aggregation**: Transaction and service tables were aggregated to single-row summaries per CustomerID to avoid Cartesian product inflation.

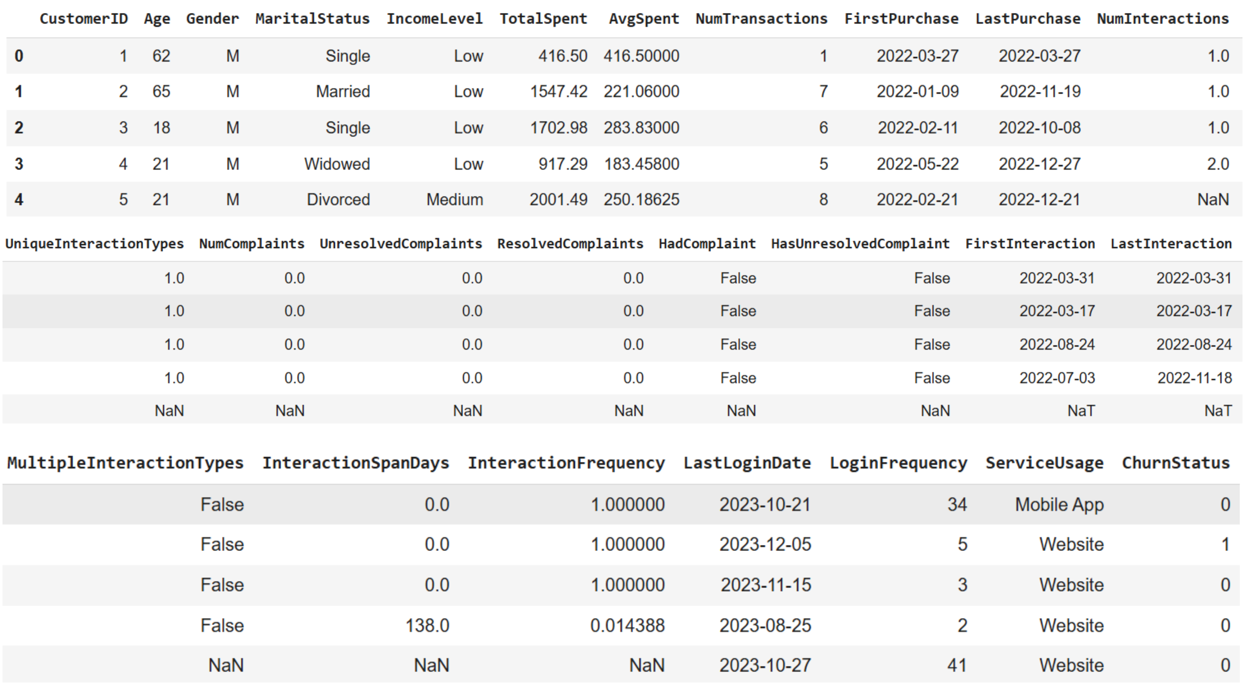
- **Left Joins on CustomerID**: Ensured that all customers in the demographic table were retained, with missing values appropriately imputed (zeros for numerical summaries, "1970-01-01" for timestamps).

- **Data Consistency Checks**: Validated that each CustomerID appeared exactly once in the final merged set.

- **Final Output**: A single "wide" table containing ~1,000 customers and 25+ engineered features, all aligned for exploratory analysis and model training.

2.5 Snapshot of Merged Dataset

Below is a sample of the first five rows of the merged, feature-engineered dataset.



	CustomerID	Age	Gender	MaritalStatus	IncomeLevel	TotalSpent	AvgSpent	NumTransactions	FirstPurchase	LastPurchase	NumInteractions
0	1	62	M	Single	Low	416.50	416.50000	1	2022-03-27	2022-03-27	1.0
1	2	65	M	Married	Low	1547.42	221.06000	7	2022-01-09	2022-11-19	1.0
2	3	18	M	Single	Low	1702.98	283.83000	6	2022-02-11	2022-10-08	1.0
3	4	21	M	Widowed	Low	917.29	183.45800	5	2022-05-22	2022-12-27	2.0
4	5	21	M	Divorced	Medium	2001.49	250.18625	8	2022-02-21	2022-12-21	NaN

UniqueInteractionTypes	NumComplaints	UnresolvedComplaints	ResolvedComplaints	HadComplaint	HasUnresolvedComplaint	FirstInteraction	LastInteraction
1.0	0.0	0.0	0.0	False	False	2022-03-31	2022-03-31
1.0	0.0	0.0	0.0	False	False	2022-03-17	2022-03-17
1.0	0.0	0.0	0.0	False	False	2022-08-24	2022-08-24
1.0	0.0	0.0	0.0	False	False	2022-07-03	2022-11-18
NaN	NaN	NaN	NaN	NaN	NaN	NaT	NaT

MultipleInteractionTypes	InteractionSpanDays	InteractionFrequency	LastLoginDate	LoginFrequency	ServiceUsage	ChurnStatus
False	0.0	1.000000	2023-10-21	34	Mobile App	0
False	0.0	1.000000	2023-12-05	5	Website	1
False	0.0	1.000000	2023-11-15	3	Website	0
False	138.0	0.014388	2023-08-25	2	Website	0
NaN	NaN	NaN	2023-10-27	41	Website	0

3. Exploratory Data Analysis (EDA)

This section presents key findings from the exploratory analysis of the merged dataset. It includes univariate distributions, bivariate and interaction relationships (with churn), correlation structure, and results of statistical tests.

3.1 Data Overview and Summary Statistics

This subsection presents a high-level summary of the dataset's key numeric and categorical features, providing context for subsequent exploratory analyses.

3.1 (a) Numeric Features

Feature	**Mean**	**Std Dev**	**Median**	**Min**	**Max**
Age	43.27	15.24	43.00	18.00	69.00
TotalSpent (USD)	1,267.07	738.59	1,232.88	9.80	3,386.04
NumTransactions	5.05	2.60	5.00	1.00	9.00
LoginFrequency	25.91	14.06	27.00	1.00	49.00
NumInteractions	1.00	0.82	1.00	0.00	2.00

- **Age** is centred around the early forties, with customers ranging from 18 to 69 years.
- **TotalSpent** shows wide variability, indicating a mix of low- and high-value customers.
- On average, customers complete around **five transactions** and initiate **one service interaction**.
- **LoginFrequency** (median 27 logins) suggests a generally engaged user base, with some low-activity outliers.

3.1 (b) Categorical Features

Feature	**Unique Values**	**Most Frequent Value**	**Frequency**
Gender	2	F	513
MaritalStatus	4	Widowed	276
IncomeLevel	3	High	349
HadComplaint	2	False	703
HasUnresolvedComplaint	2	False	832
MultipleInteractionTypes	2	False	785
ServiceUsage	3	Online Banking	349

- **Gender** is balanced, with a slight majority of female customers (51.3%).
- **MaritalStatus** categories are varied; "Widowed" is the largest group (27.6%).
- **IncomeLevel** skews toward "High" (34.9%).
- Fewer than one-third of customers have ever filed a complaint, and only 16.8% have an unresolved complaint.
- Most customers engage via **Online Banking** (34.9%), with the remainder split between Mobile App and Branch.

These summary statistics confirm that the dataset captures a diverse set of customer profiles and behaviours, setting the stage for deeper bivariate and multivariate analyses in the following sections.

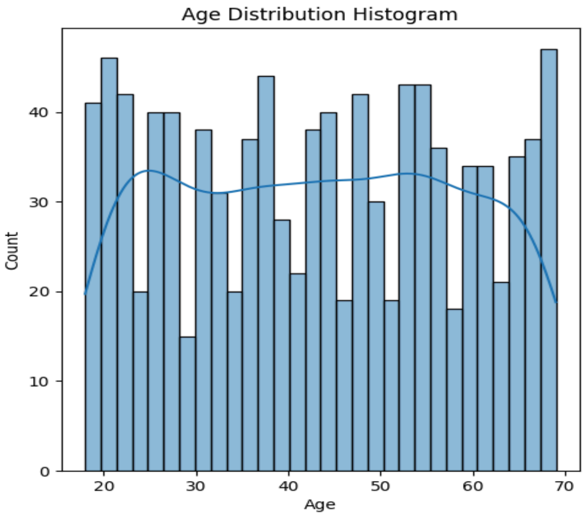
3.2 Univariate Distributions

To better understand SmartBank's customers, we analysed the dataset's individual distributions of key features. This helps uncover general trends and behaviours across the customer base, such as how much they typically spend, how frequently they engage with the bank, and their general demographic characteristics.

Numerical Features

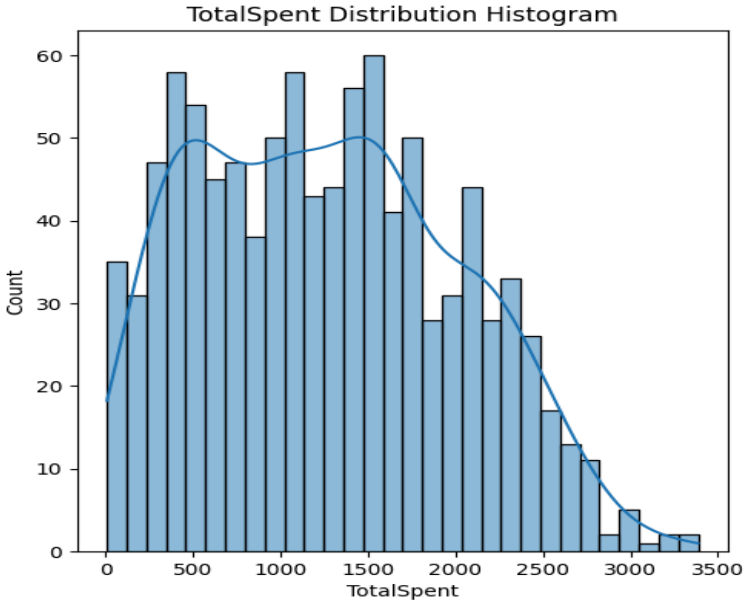
- **Age**
  The age distribution is relatively uniform, with most customers falling between the ages of 18 and 60. The average age is approximately 43 years, with no noticeable skewness, suggesting that SmartBank serves a broad age range.

  Figure 1: Histogram of Age



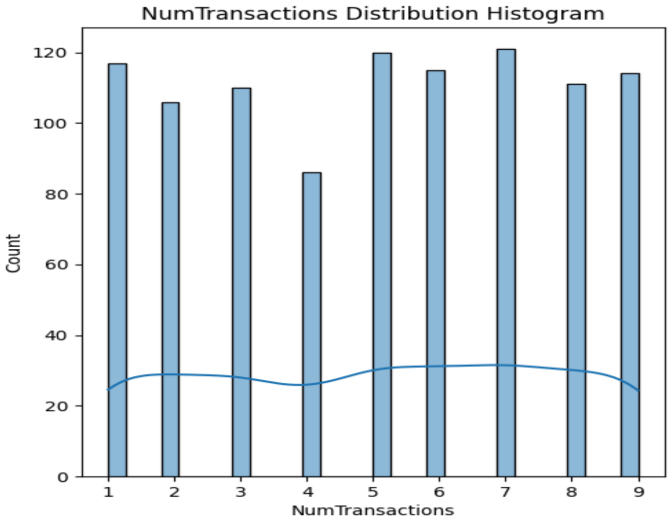
- **Total Amount Spent**
  The total amount spent by customers varies widely. While the average customer has spent around $1,267, some customers have spent as little as $9.80, and others up to $3,386. This right-skewed distribution indicates that a smaller group of customers are responsible for a large portion of spending.

  Figure 2: Histogram of TotalSpent



- **Number of Transactions**
  Most customers have completed between 3 to 7 transactions, with a median of 5. The spread is fairly even, showing that the majority are consistently active.

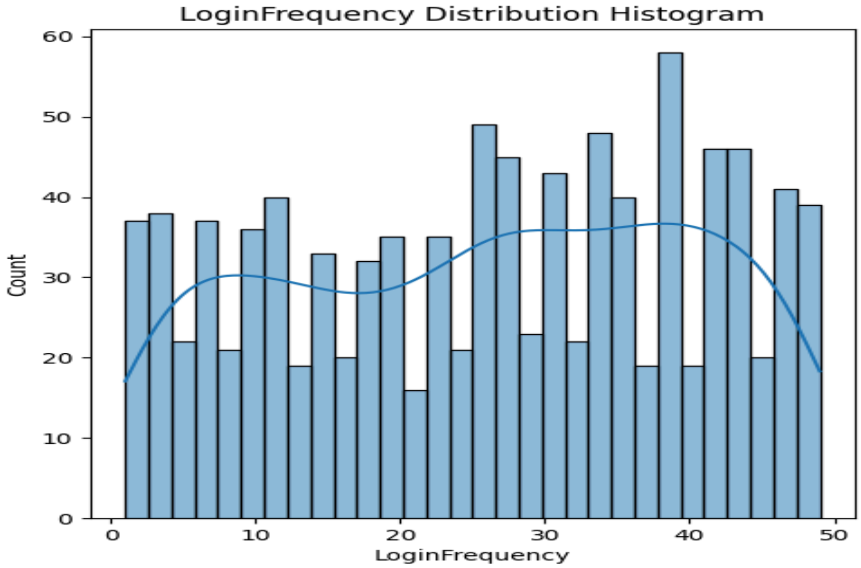
  Figure 3: Histogram of NumTransactions



- **Login Frequency**
  On average, customers logged in 26 times over the observed period. Some logged in

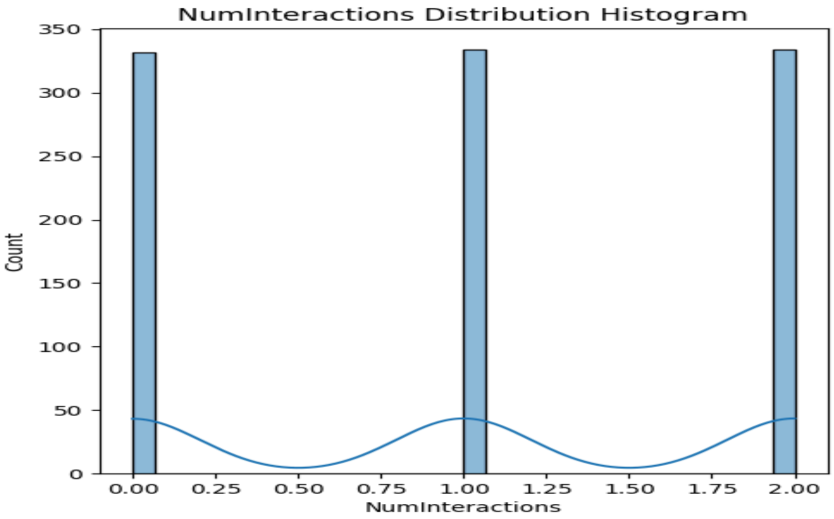
nearly every day, while a few were much less active. The variation in login frequency could reflect different levels of digital engagement, which may be linked to loyalty or satisfaction.

Figure 4: Histogram of LoginFrequency



- **Number of Interactions with Customer Service**
  Most customers had either no contact or only one interaction with customer service. A very small subset reached out more than once. This suggests that the majority of customers did not face major service issues or chose not to report them.

  Figure 5: Bar plot of NumInteractions



Categorical Features

- **Gender**
  The customer base is balanced, with 51% identifying as female and 49% as male.

- **Marital Status**
  SmartBank's customers span across different marital statuses, with the largest segment being widowed (27.6%), followed by married and single customers. This could reflect a strong user base among older customers.

- **Income Level**
  The majority of customers fall into the "High" income bracket (34.9%), followed by "Medium" and "Low" income levels. This highlights SmartBank's reach into financially capable segments, which may influence product preferences and churn behaviour.

- **Had a Complaint**
  Around 30% of customers reported having had at least one complaint during the period. While this suggests some dissatisfaction, it also highlights opportunities for service recovery.

- **Unresolved Complaints**
  Among those who had complaints, a significant portion had their issues resolved. Only 16.8% of the full customer base had unresolved complaints, indicating a generally effective resolution process, though there's still room for improvement.

- **Interaction Types**
  Most customers interacted with the bank through a single service channel. Only a minority used multiple channels, such as both online and in-person services. This

could be a point of opportunity for SmartBank to promote more integrated, omnichannel experiences.

- **Service Usage**
  Online Banking is the most common service type, followed by Mobile App and In-Branch services. This shows a clear shift toward digital banking, which may correlate with other behaviour patterns like login frequency and churn.

3.3 Bivariate Analysis

This section explores the relationships between key features and the churn status of customers. The goal is to identify patterns or associations that might help explain why customers leave and which features are more influential in predicting this behavior.

3.3a. Correlation Between Numerical Features and Churn

We computed the correlation between numerical features and churn status to assess how closely they are related.

- Most features had very weak correlations with churn status (values close to 0), suggesting limited direct linear relationships.

- Notably:

  ○ LoginFrequency had a **negative correlation** with churn status (-0.08), suggesting that **customers who log in more often are slightly less likely to churn.**

  ○ InteractionFrequency had a weak positive correlation (0.06), indicating **higher interaction rates may be slightly associated with churn**, possibly reflecting frustration or repeated contact attempts.

Feature	**Correlation with ChurnStatus**
Age	0.03
TotalSpent	0.00

NumTransactions	-0.01
LoginFrequency	-0.08
NumComplaints	0.02
NumInteractions	0.00
InteractionFrequency	0.06

3.3b. Statistical Testing for Significance (Numerical Variables)

To move beyond correlation and assess **whether differences in features are statistically significant** between churned and retained customers, we performed the **Mann–Whitney U test** for each numerical feature.

Feature	**p-value**	**Result**
Age	0.3466	❌ Not significant
TotalSpent	0.9464	❌ Not significant
AvgSpent	0.2392	❌ Not significant
NumTransactions	0.7805	❌ Not significant

NumInteractions	0.8799	❌ Not significant
UniqueInteractionTypes	0.6923	❌ Not significant
NumComplaints	0.5674	❌ Not significant
UnresolvedComplaints	0.4449	❌ Not significant
ResolvedComplaints	0.8760	❌ Not significant
InteractionSpanDays	0.1910	❌ Not significant
InteractionFrequency	**0.0303**	✅ **Significant**
LoginFrequency	**0.0107**	✅ **Significant**

- **Interpretation**:
  - Only **LoginFrequency** and **InteractionFrequency** showed statistically significant differences between churned and retained customers.
  - These findings support the idea that **ongoing engagement with the platform is a key factor in retention**.

3.3c. Statistical Testing for Categorical Variables

We also conducted **Chi-square tests** to determine whether there were significant associations between categorical features and churn status.

Feature	**p-value**	**Result**
Gender	0.6207	❌ Not significant
MaritalStatus	0.6326	❌ Not significant
IncomeLevel	0.6160	❌ Not significant
HadComplaint	0.6170	❌ Not significant
HasUnresolvedComplaint	0.4981	❌ Not significant
MultipleInteractionTypes	0.5210	❌ Not significant
ServiceUsage	0.2409	❌ Not significant

- **Interpretation**:
  - None of the categorical features showed statistically significant associations with churn.
  - This suggests that **demographic or complaint status alone may not explain churn** without considering the context of customer behavior.

3.4 Multivariate Analysis

In this section, we explore how combinations of features—especially behavioral and complaint-related attributes—interact to influence customer churn. Our aim is to understand how the effects of one factor may vary depending on another.

3.4.1 Feature Binning

To better understand patterns in customer behavior, continuous variables were grouped into three categories using equal-sized bins:

- **Age** was grouped into: *Young*, *Mid-age*, and *Senior*
- **LoginFrequency** into: *Low Freq*, *Medium Freq*, and *High Freq*
- **TotalSpent** into: *Low Spend*, *Medium Spend*, and *High Spend*
- **InteractionFrequency** into: *Low Interaction*, *Medium Interaction*, and *High Interaction*

This binning helped to make comparison easier and to reveal more intuitive insights in the context of churn rates.

3.4.2 Churn Rates by Login Frequency and Complaint Status

We examined how churn rates differ across login frequency categories and whether a customer had unresolved complaints. Here's a summary:

LoginBin	**HasUnresolvedComplaint**	**Retained**	**Churned**
Low Freq	False	77.4%	22.6%
Low Freq	True	68.8%	31.2%
Medium Freq	False	80.1%	19.9%
Medium Freq	True	80.0%	20.0%
High Freq	False	82.8%	17.2%

High Freq	True	81.8%	18.2%

- **Observation**: Customers with **low login frequency and unresolved complaints** had the **highest churn rate (31.2%)**.
- In contrast, churn rates were consistently lower among **high-frequency users**, regardless of complaint status.

3.4.3 Statistical Testing: Interaction Between Login Behavior and Complaints

To investigate whether these differences were statistically meaningful, we performed a **logistic regression** using the binned login frequency and unresolved complaint status as predictors.

Predictor	**Coefficient**	**p-value**
Intercept	-1.2283	0.000
LoginBin: Medium Freq	-0.1672	0.420
LoginBin: High Freq	-0.3421	0.109
HasUnresolvedComplaint (True)	0.4398	0.198
Medium Freq × UnresolvedComplaint (Interaction)	-0.4306	0.376
High Freq × UnresolvedComplaint (Interaction)	-0.3735	0.468

- **Result**: Although the trends suggest that frequent login may reduce churn, especially when complaints are unresolved, the model did not find these effects to be statistically significant (*p > 0.05*).

- **Model Fit**: The overall model had a low explanatory power (Pseudo R² = 0.006), indicating limited predictive strength from these combinations.

3.4.4 Follow-up Test: Using Raw Login Frequency

To further verify whether the number of logins directly impacts churn, we conducted a second logistic regression using the raw (continuous) values of LoginFrequency instead of bins.

Predictor	**Coefficient**	**p-value**
Intercept	-1.0733	0.000
HasUnresolvedComplaint (True)	0.4394	0.289
LoginFrequency	-0.0126	0.040
LoginFreq × UnresolvedComplaint	-0.0111	0.459

- **Result**: In this model, **LoginFrequency was a statistically significant predictor** of churn (*p = 0.040*), suggesting that **higher login frequency is associated with lower churn risk**.

- However, **the interaction with unresolved complaints was not significant**, indicating that unresolved issues don’t significantly alter the effect of login behavior on churn.

3.5 Conclusion

Our exploratory analysis offered a comprehensive look into customer behavior and service interactions at SmartBank, focusing on how they relate to churn.

- From the **univariate analysis**, we observed that churned customers generally exhibit **lower login frequency**, **fewer interactions**, and a slightly higher likelihood of **unresolved complaints**.

- The **bivariate analysis** highlighted that while features like **login frequency** and **interaction frequency** are associated with churn, many other variables — including age, spending, and complaints — showed no significant difference between churned and retained groups.

- In the **multivariate analysis**, although **lower login frequency** appeared to reduce the likelihood of churn, the overall model explained only a small portion of the variation in churn behavior. Many of the variables we tested, including interaction with complaint resolution, were not statistically significant.

These findings suggest that **churn is not entirely predictable from the available internal features**. While behavior and service experiences play a role, **churn decisions may be influenced by other external or unmeasured factors**, such as changes in customer life circumstances, competition, or broader market dynamics.

In other words, while some signals are helpful, **churn still appears partially random**, and relying solely on behavioral data may not fully explain or predict it. This insight is essential as we move toward customer segmentation and strategy design in the next section.

4. Feature Engineering and Preprocessing

This section outlines the transformations and feature engineering steps applied to the raw dataset in preparation for modeling. The goal was to convert the data into a structured format that enhances the ability of machine learning algorithms to identify meaningful patterns.

4.1 Handling Missing Values

While most columns had complete records, a few features—particularly those derived from customer interactions—had missing values, typically indicating the absence of such interactions.

- **Numerical fields** (e.g., NumComplaints, NumInteractions) were filled with **0**, reflecting no recorded activity.

- **Date fields** related to interactions were assigned a default placeholder of **"1970-01-01"**, signifying that the customer never engaged with that channel.

This approach ensured the integrity of the dataset while preserving valuable information about customer inactivity.

4.2 Creating New Features

To extract deeper insights and support predictive modeling, several new features were engineered from existing columns:

- **Interaction Frequency:**
  This was calculated by dividing the total number of interactions by the span of days between the customer's first and last interaction. This provides a measure of how frequently a customer engages with the bank's services over time.

- **Resolved Complaints:**
  Derived by subtracting unresolved complaints from total complaints to highlight customers who had their issues resolved.

- **Unique Interaction Types:**
  Indicates the number of distinct customer service channels used, providing a sense of channel diversity.

- **Average Spent per Transaction:**
  Created by dividing TotalSpent by NumTransactions to capture transaction efficiency or spending behavior.

- **RFM Metrics (Recency, Frequency, Monetary):**
  The **RFM** model was used to segment customers based on three key factors:

  - **Recency (R)**: How recently the customer made a transaction. Higher values indicate more recent transactions.

  - **Frequency (F)**: How frequently the customer transacts. Higher values indicate more frequent transactions.

  - **Monetary (M)**: How much the customer has spent. Higher values indicate higher spending.

- Each metric was scored using quantiles:

  - **Recency**: Scored from 3 (most recent) to 1 (least recent).

  - **Frequency**: Scored from 1 (infrequent) to 3 (frequent).

  - **Monetary**: Scored from 1 (low spender) to 3 (high spender).

- These scores were combined into an overall **RFM score** for each customer (RFM_score_trans), which helped identify high-value and at-risk customers.

4.3 Binning Continuous Variables

To aid interpretation and facilitate advanced analysis (e.g., segment-level comparisons), continuous variables were grouped into meaningful bins:

- **Age** → AgeBin: Young, Mid-age, Senior
- **LoginFrequency** → LoginBin: Low Freq, Medium Freq, High Freq
- **TotalSpent** → SpendBin: Low Spend, Medium Spend, High Spend
- **InteractionFrequency** → InteractionFreqBin: Low Interaction, Medium Interaction, High Interaction

Binning not only supports visual comparisons but also captures non-linear relationships more effectively in some models.

4.4 Encoding Categorical Variables

Since machine learning models typically require numeric inputs, categorical features were encoded:

- **Binary features** such as Gender, HadComplaint, and HasUnresolvedComplaint were converted using simple binary encoding (0/1).
- **Multi-class categorical features** like ServiceUsage, IncomeLevel, and MaritalStatus were transformed using one-hot encoding.

This transformation ensured compatibility with downstream modeling pipelines without losing any category-specific information.

4.5 Time-Based Features

To capture the recency of customer activities, several time-based features were engineered:

- **Days Since Last Purchase**:
  Calculated as the difference between the most recent purchase date and the individual customer's last purchase date. This provides a measure of how long it has been since the customer last interacted with the company in terms of purchasing.
- **Days Since Last Login**:
  Calculated as the difference between the most recent login date and the individual customer's last login date.

- **Days Since Last Interaction**:
  Measured as the difference between the most recent interaction date and the customer's last interaction.

These time-based features offered a better understanding of customer engagement and helped assess whether there were any patterns associated with churn.

5. Segmentation and Clustering

This section focuses on customer segmentation using clustering techniques, aimed at grouping similar customers based on their behaviors, interactions, and spending patterns. Clustering allows for the identification of distinct customer segments, which is vital for targeted marketing strategies and understanding potential churn risks.

5.1 RFM Segmentation

To begin, we used the **RFM (Recency, Frequency, Monetary)** model to segment customers based on their recent interactions, transaction frequency, and spending habits. The RFM metrics were converted into scores, with each metric (Recency, Frequency, Monetary) receiving a score from 1 to 3. The following scores were assigned:

- **Recency** (R): Scored based on how recently the customer made a transaction, with more recent transactions receiving a higher score.

- **Frequency** (F): Scored based on how often the customer transacts, with more frequent transactions receiving a higher score.

- **Monetary** (M): Scored based on how much the customer has spent, with higher spending customers receiving a higher score.

The combined **RFM score** was then computed by concatenating the individual scores for Recency, Frequency, and Monetary values, resulting in a new feature called RFM_score_trans.

While this segmentation approach helped us categorize customers based on their transaction behavior, statistical tests (Chi-square statistic = 13.83, p-value = 0.8767) showed **no significant association** between the RFM segments and churn status, suggesting that RFM alone might not be a strong indicator of churn.

5.2 Customer Clustering

Following the RFM segmentation, we further refined customer segmentation through **unsupervised clustering**. This method allowed us to group customers with similar behaviors, engagement levels, and spending patterns.

- **Clustering Technique**: We applied **K-means clustering** to divide customers into distinct clusters. K-means works by grouping customers based on the similarity of their features, with each cluster representing a group of customers who exhibit similar behaviors.

- **Features Used for Clustering**: We utilized a combination of features, including **RFM scores**, customer interaction frequency, complaint status, spending behavior, and engagement levels.

- **Number of Clusters**: The optimal number of clusters was determined through the **Elbow Method**, which suggested that **3 clusters** was the best choice. This was based on the analysis of the within-cluster sum of squares (WCSS) and the interpretability of the clusters.

The result was the identification of **3 distinct customer segments**, which allowed for a deeper understanding of customer behavior and engagement levels.

5.3 Statistical Testing for Cluster Differences

After performing clustering, we conducted statistical tests to determine whether there were significant differences in churn risk and customer engagement across the identified clusters:

- We used **Kruskal-Wallis tests** to assess differences in churn across the 3 customer clusters. This non-parametric test was appropriate since it doesn't assume normality and is ideal for comparing multiple groups.

- The results revealed that there were **significant differences** between some clusters in terms of churn behavior. However, not all clusters showed a significant relationship with churn, suggesting that additional external factors may play a role in influencing churn.

6. User Personas and Strategic Recommendations

Following the clustering analysis, three distinct customer personas were identified based on behavioral patterns such as transaction activity, complaint history, service usage, and login frequency. These personas provide a practical framework for SmartBank to improve engagement, retention, and service personalization.

6.1 Cluster 0 – Lightly Engaged, Low-Issue Customers

Profile:

- Avg. Total Spent: $1,287.79
- Avg. Transactions: 5
- Avg. Interactions: 1
- Complaints: Low, mostly resolved
- Interaction Frequency: Very low
- Login Frequency: High

Insights:
This segment demonstrates consistent digital activity but minimal interaction with customer service. These customers rely on SmartBank for basic financial needs and rarely seek assistance or use advanced features.

Strategic Recommendations:

- Introduce cross-selling of value-adding products (e.g., investment plans, insurance)
- Use personalized nudges to raise awareness of unused app features
- Offer educational content to increase confidence in exploring additional services

6.2 Cluster 1 – Highly Engaged, High-Issue Customers

Profile:

- Avg. Total Spent: $1,338.73 (Highest)
- Avg. Transactions: 5.25
- Avg. Interactions: 2

- Complaints: Moderate, with a high rate of unresolved issues
- Interaction Span: Long
- Login Frequency: Moderate

Insights:
These customers engage across multiple channels and represent high value, but they also experience more service issues. They are long-term clients whose dissatisfaction may undermine loyalty.

Strategic Recommendations:

- Prioritize issue resolution via a dedicated VIP support team
- Implement proactive follow-up mechanisms through CRM
- Launch loyalty benefits (e.g., lower loan interest, fee waivers) to reinforce value

6.3 Cluster 2 – At-Risk Silent Users

Profile:

- Avg. Total Spent: $1,175.83 (Lowest)
- Avg. Transactions & Interactions: Minimal
- Complaints & Interaction Types: None
- Login & Interaction Frequency: Zero

Insights:
These users are disengaged and inactive, with little to no signs of recent interaction. They represent the highest churn risk despite having no formal complaints.

Strategic Recommendations:

- Send reactivation offers (e.g., cashback, temporary fee removal)
- Initiate exit surveys or feedback calls to understand disengagement causes
- Set up early warning systems to detect future silent churn behavior

6.4 Summary

While direct relationships between these personas and churn were not statistically significant, the behavior-driven segmentation presents clear strategic value.

By tailoring communication, services, and retention strategies to each persona's needs, SmartBank can proactively address customer satisfaction gaps, unlock cross-selling potential, and mitigate churn risk—even in the absence of direct churn predictors.

7. Conclusion and Next Steps

7.1 Summary of Insights

This analysis explored customer churn from multiple angles—behavioral patterns, transaction history, service interactions, and demographic factors. Key takeaways include:

- Some behavioral features like **Login Frequency**, **Interaction Frequency**, and **Complaint Resolution** show visible differences across churn groups. However, no single feature or segment consistently predicted churn.

- **RFM segmentation** and **time-based features** provided additional behavioral granularity, but statistical tests (e.g., chi-square) showed no significant relationship with churn.

- **Clustering** uncovered three distinct behavioral personas, each requiring tailored strategies. These personas offer a practical lens for customer engagement, retention, and growth—even without clear churn predictors.

Together, these findings suggest that **customer churn may be influenced by external factors** (e.g., market competition, personal financial changes) that are not captured in the current data.

7.2 Recommendations and Future Work

To build on the current work and improve churn management, the following next steps are recommended:

- **Integrate External Data**
  Enrich existing datasets with macroeconomic indicators (e.g., inflation rates, employment levels) or customer lifestyle signals (e.g., location-based trends) to better account for external drivers of churn.

- **Deploy Predictive Modeling**
  With the engineered features and enriched segmentation in place, the next step is to train supervised models (e.g., logistic regression, random forest) to identify subtle churn drivers and score churn risk at the individual level.

- **Implement Real-Time Monitoring**
  Develop dashboards to track interaction frequency, complaint trends, and login behavior. This can support early intervention for customers showing signs of disengagement.

- **Continue Customer Research**
  Conduct qualitative research (e.g., interviews, surveys) with churned and at-risk users to understand missing factors that data may not reveal.

- **Test Retention Interventions**
  Run A/B experiments on the strategic recommendations derived from the personas (e.g., exclusive offers for Cluster 1) to measure their impact on satisfaction and churn.

By evolving from descriptive insights to predictive and prescriptive actions, SmartBank can strengthen its customer retention strategy and unlock long-term value from its customer base.