Business Report Title

"Predicting Customer Churn at Thera Bank: A Data-Driven Strategy for Retention and Growth"

Table Of Contents

1.Problem Statement

Thera Bank has observed a **significant decline in credit card customers**, which poses a major threat to the bank's revenue. Credit card services contribute substantially to income through various fees such as annual charges, late payment penalties, foreign transaction fees, and more. Losing customers not only impacts immediate revenue but also affects long-term profitability and customer lifetime value.

To address this, the bank needs to understand **why customers are discontinuing** their credit cards and develop a system to **predict future attrition**. By identifying patterns in customer behavior, usage, and demographics, the bank can proactively intervene to retain high-risk customers.

2.Business Objective

The primary goal of this project is to **develop a robust classification model** that can:

- Accurately **predict whether a customer is likely to discontinue** their credit card with the bank.
- Help Thera Bank **identify key factors** driving customer attrition.
- Enable **targeted retention strategies** by profiling high-risk customers based on data.
- Improve customer engagement and satisfaction by addressing the pain points revealed through analysis.
- Ultimately, **reduce customer churn**, increase revenue stability, and enhance the bank's ability to personalize services.

3.**Dataset Overview: Understanding the Customer Landscape at Thera Bank**

The dataset provided by Thera Bank is a **comprehensive and multi-dimensional** snapshot of over **10,000 customers**, capturing a wide array of behavioral, demographic, and financial metrics. It enables a **deep forensic analysis** into the habits, engagement patterns, and risk factors associated with customer churn, providing fertile ground for predictive modeling and actionable business insights.

🔍 Scale & Scope of the Dataset

- **Total Observations**: 10,127 individual customers
- **Time Horizon**: Includes aggregated metrics for the **last 12 months**
- **Number of Variables**: 21, capturing both objective data and behavior trends
- **Business Target**:
    - **Attrition_Flag** (binary classification target)

- Existing Customer – Loyal, retained customers
- Attrited Customer – Customers who have discontinued the credit card

This dataset is **balanced in depth but imbalanced in outcome**, with attrition being a minority class—making it a textbook case for **advanced imbalanced classification** problems.

Feature Breakdown: A 360° View of the Customer

Category	Features	Description
Demographics	Customer_Age, Gender, Dependent_count, Education_Level, Marital_Status, Income_Category	Provides socioeconomic insights into the customer base. Helps detect demographi

Category	Features	Description
		c churn trends.
Product Ownership	Card_Category, Total_Relationship_Count, Months_on_book	Measures how embedded a customer is in the bank's ecosystem.
Account Behavior	Months_Inactive_12_mon, Contacts_Count_12_mon	Reflects engagement and interest in maintaining a relationship with the bank.
Credit Dynamics	Credit_Limit, Avg_Open_To_Buy, Avg_Utilization_Ratio	Offers insight into financial discipline,

Category	Features	Description
		liquidity, and dependency on credit.
Spending Patterns	Total_Trans_Amt, Total_Trans_Ct, Total_Amt_Chng_Q4_Q1, Total_Ct_Chng_Q4_Q1	Tracks changes in monetary and behavioral velocity, indicating life events or dissatisfaction.
Technical ID	CLIENTNUM	Unique identifier; dropped during modeling to avoid bias or leakage.

⚠️ Data Quality & Preprocessing Challenges

- **Missing Values**:
  - Education_Level: 1,519 missing entries (15%)
  - Marital_Status: 749 missing entries (7.4%)
  - These reflect the **incomplete self-disclosure** by users, potentially hiding predictive behavioral cues.
- **No Duplicates Detected**: Dataset integrity is intact.
- **Categorical Complexity**:
  - Non-numeric fields (e.g., Income_Category) require transformation to support machine learning algorithms.
- **Imbalanced Target**:
  - Only **~16%** of the data represents Attrited Customers, necessitating **resampling strategies** (SMOTE, undersampling) and careful metric selection (F1-score, ROC-AUC).

Statistical Sumary Of The Data:

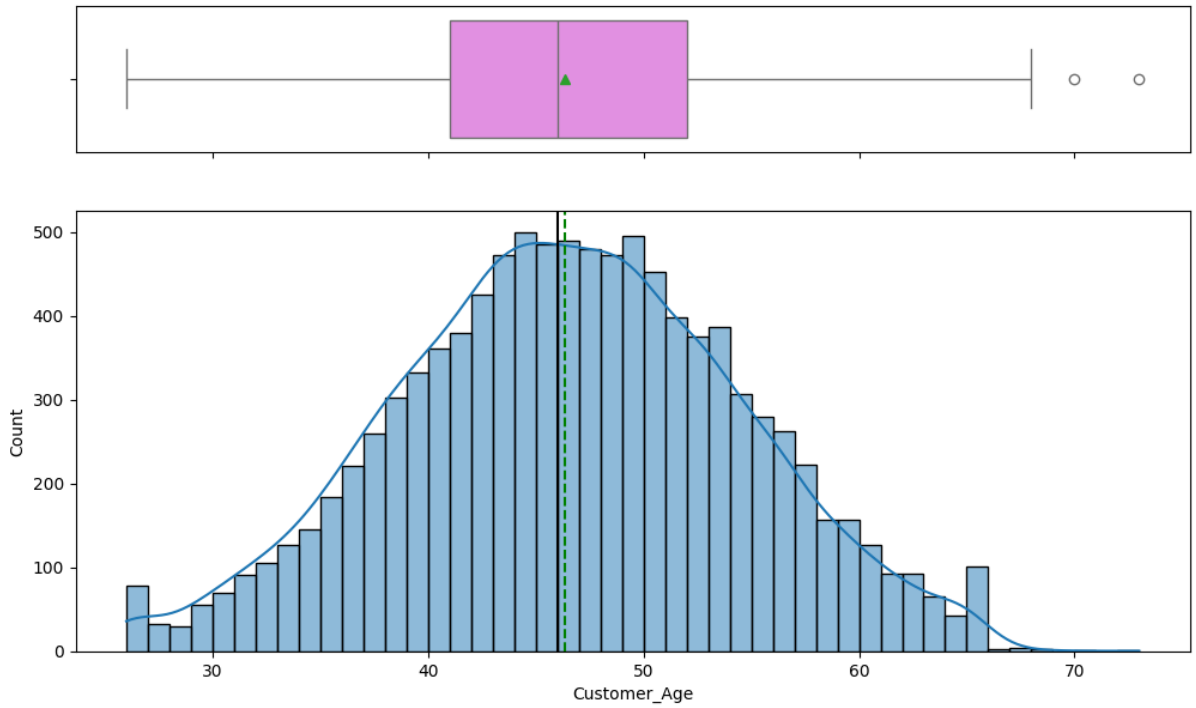
The dataset contains 10,127 customer records and 21 variables. Below is a statistical summary of the **numerical features** in the dataset, offering insights into the central tendency, spread, and data distributions across key customer attributes.

Variable	Mean	Std Dev	Min	25th %ile	Median	75th %ile	Max
Customer_Age	46.33	8.02	26	41	46	52	73
Dependent_count	2.35	1.30	0	1	2	3	5
Months_on_book	35.93	7.99	13	31	36	40	56
Total_Relationship_Count	3.81	1.55	1	3	4	5	6
Months_Inactive_12_mon	2.34	1.01	0	2	2	3	6
Contacts_Count_12_mon	2.46	1.11	0	2	2	3	6
Credit_Limit	8,632	9,089	1,438	2,555	4,549	11,067	34,516
Total_Revolving_Bal	1,163	815	0	359	1,276	1,784	2,517
Avg_Open_To_Buy	7,469	9,091	3	1,324	3,474	9,859	34,109
Total_Amt_Chng_Q4_Q1	0.76	0.22	0.00	0.63	0.74	0.86	3.40
Total_Trans_Amt	4,404	3,397	510	2,156	3,899	5,546	18,110
Total_Trans_Ct	64.86	23.47	10	45	67	81	139
Total_Ct_Chng_Q4_Q1	0.71	0.24	0.00	0.58	0.71	0.82	3.71
Avg_Utilization_Ratio	0.27	0.28	0.00	0.02	0.18	0.46	0.999

4.Exploratory Data Analysis

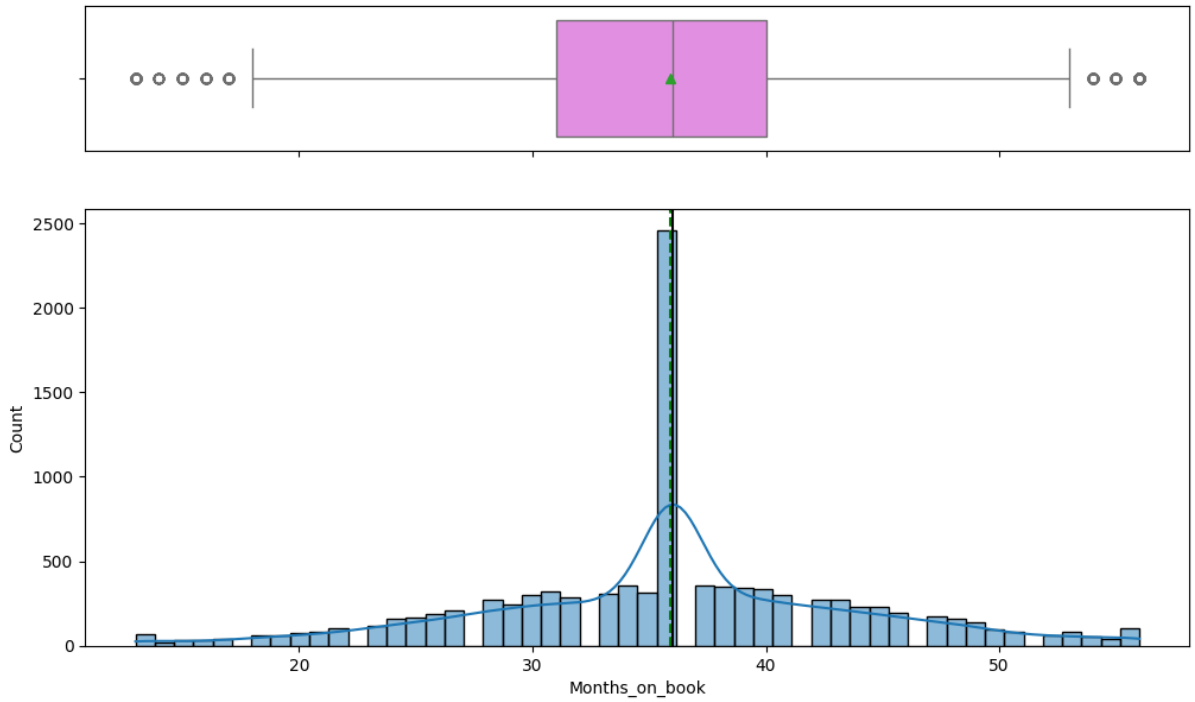
Univariate analysis

i.Customer Age:



- Most customers are between **40 and 55 years old**.
- The **average age is around 47** (green line).
- Very **few young (under 30)** or **older (above 70)** customers.
- A few older customers are marked as **outliers** (dots in the boxplot).
- **In short:** Thera Bank mostly serves **middle-aged customers**, who are likely in their peak earning years.

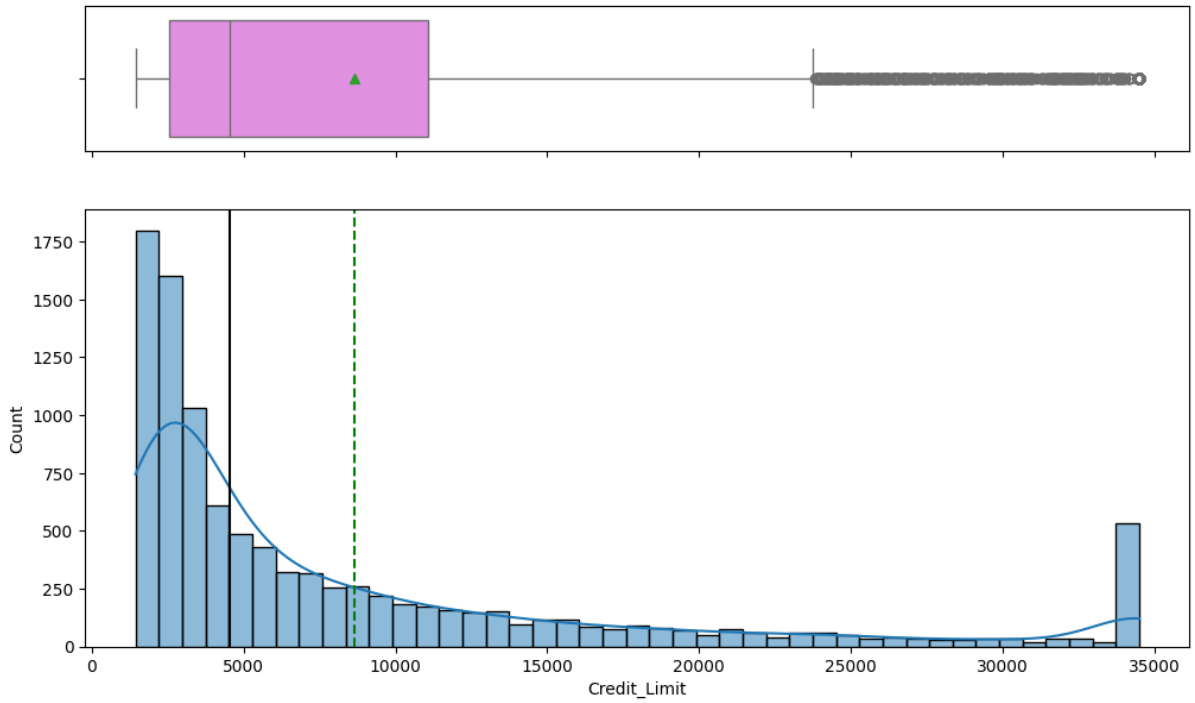
ii.Months On Book:



- Most customers have been with the bank for about **36 months (3 years)** — shown by the tall bar in the middle.
- There are some customers with **very short (under 20 months)** or **very long (over 50 months)** relationships, marked as **outliers**.
- The data is **slightly skewed**, but the majority of customers stay between **30 and 42 months**.

In short: Most Thera Bank customers have had their credit card for **about 3 years**, showing a **stable and mid-term relationship** with the bank.

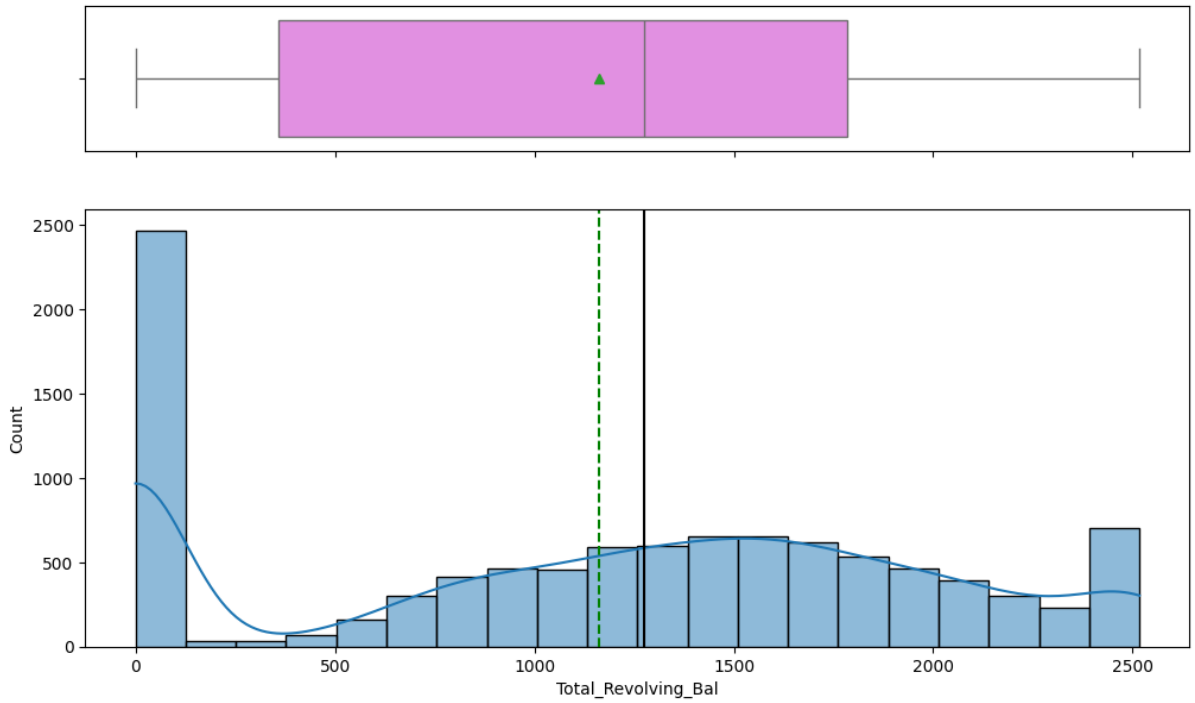
iii.Credit Limit:



- Most customers have a **credit limit below ₹10,000**, shown by the high bars on the left.
- The **average credit limit** is around ₹8,600 (green dashed line).
- Some customers have **very high limits** (₹30,000+), marked as **extreme outliers** in the boxplot.
- The distribution is **right-skewed**, meaning only a few customers hold very large credit limits.

In short: Most Thera Bank customers have **modest credit limits**, but a small group has **very high limits**, possibly high-value or VIP clients

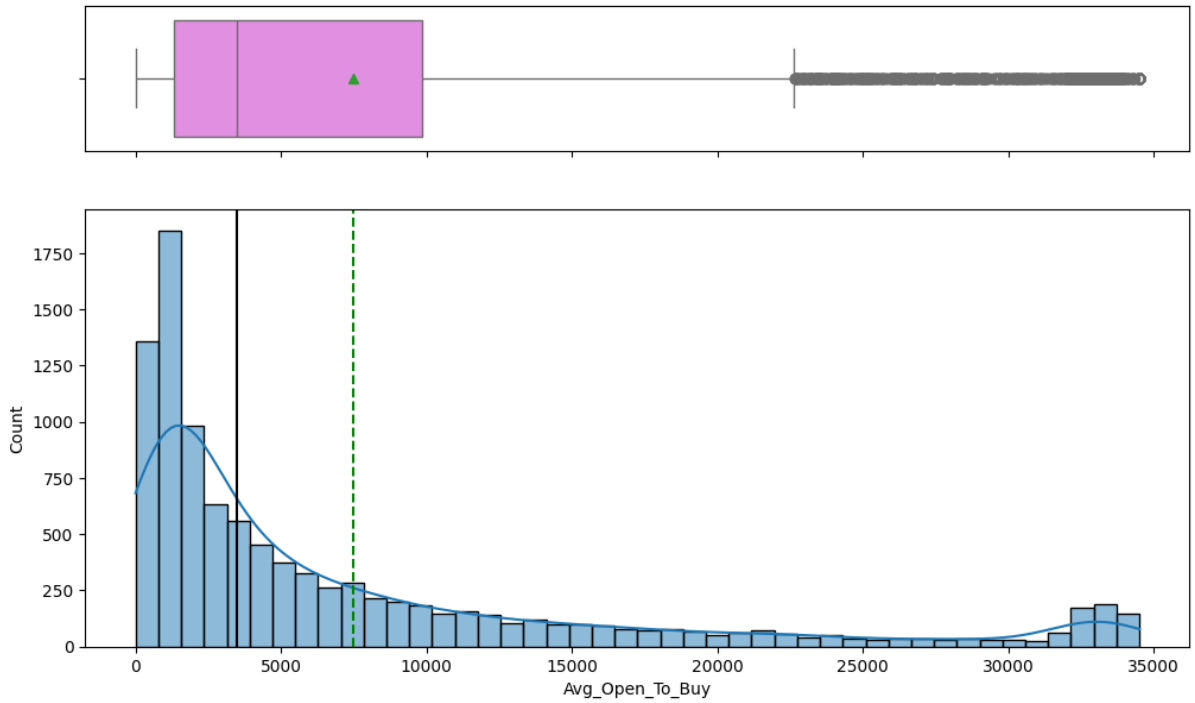
iv. Total Revolving Balance:



- Many customers have a **balance of ₹0**, which means they pay off their dues regularly.
- Others have balances spread out across the range, with a slight peak between ₹1,000–₹2,000.
- The data is fairly spread out and shows **no major outliers**.

In short: Many customers clear their dues, but some carry balances — useful for targeting credit coaching or promotional balance transfers.

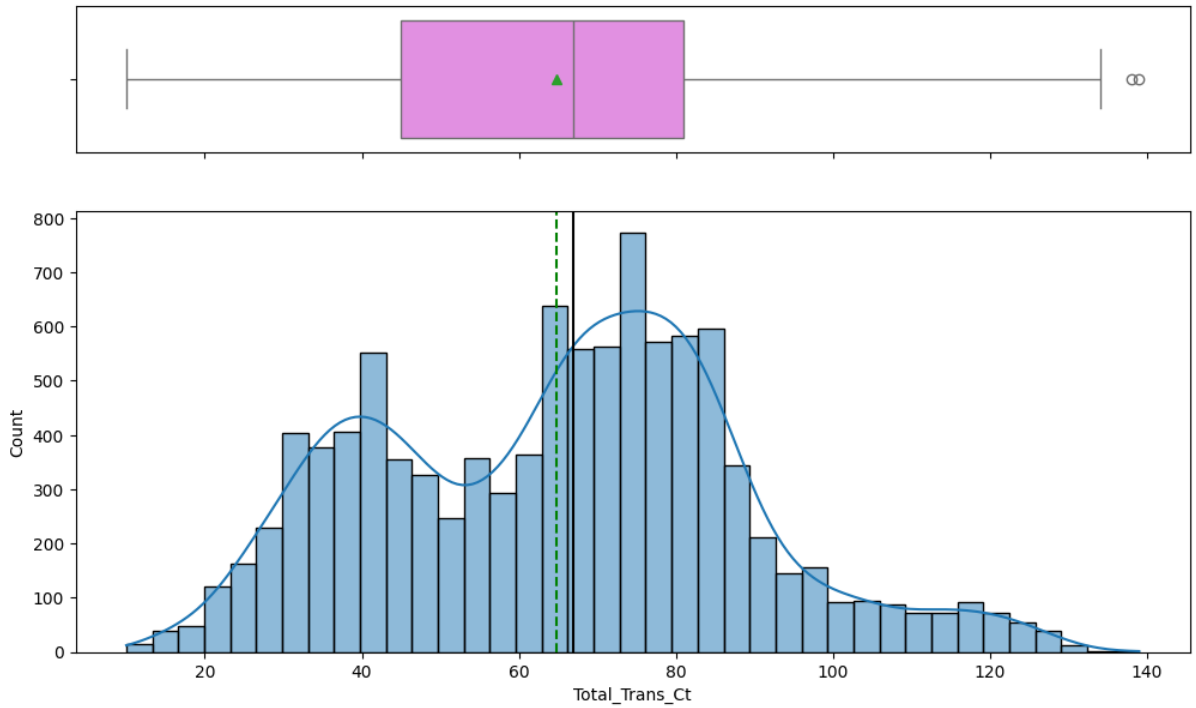
v. Average Open To Buy:



- Most customers have **less than ₹10,000** available to spend on their card.
- A small group has **very high limits available (₹30,000+),** marked as outliers.
- The distribution is **highly right-skewed**.

In short: Most customers have limited unused credit, but a few have **very high purchasing power**, possibly premium users.

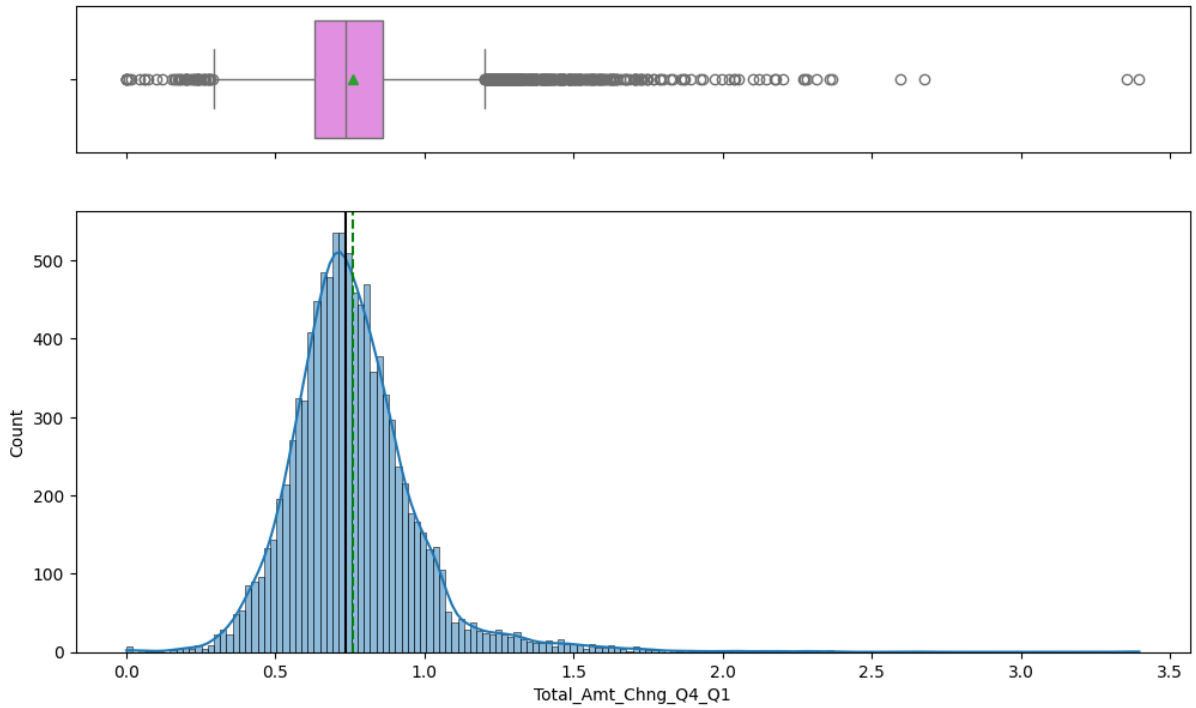
vi. Total Transaction Count:



- Customers typically make **40 to 90 transactions per year**, with the average around **65**.
- The distribution is somewhat **bimodal** – possibly indicating two behavior groups: low-usage and high-usage.
- A few users make **over 130 transactions**, marked as outliers.

In short: Most customers are **moderate spenders**, but there are **very active users** too — ideal for premium loyalty offers.

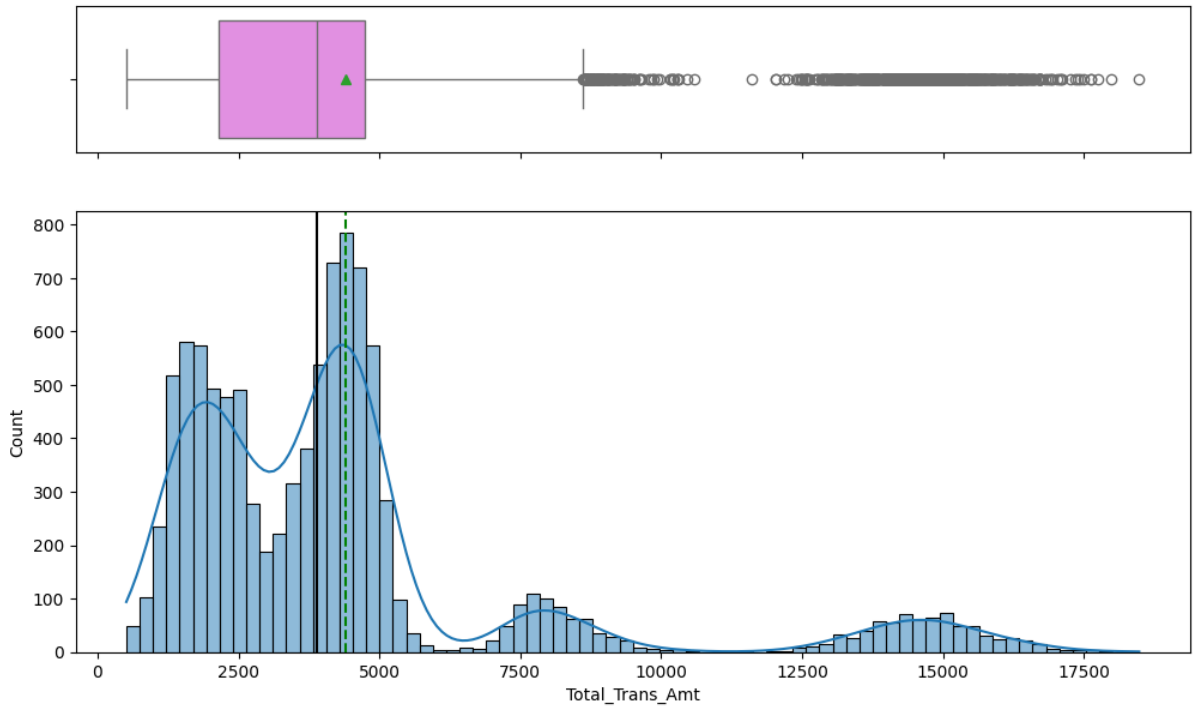
vii. Total Amount Change (Q4 vs Q1):



- Most values are between **0.5 and 1.0**, meaning **little change** in spending habits.
- Very few customers had drastic increases or decreases.
- The curve is **bell-shaped**, indicating a normal distribution.

In short: Most customers are **consistent in spending** over time. Large fluctuations are rare and might need deeper look.

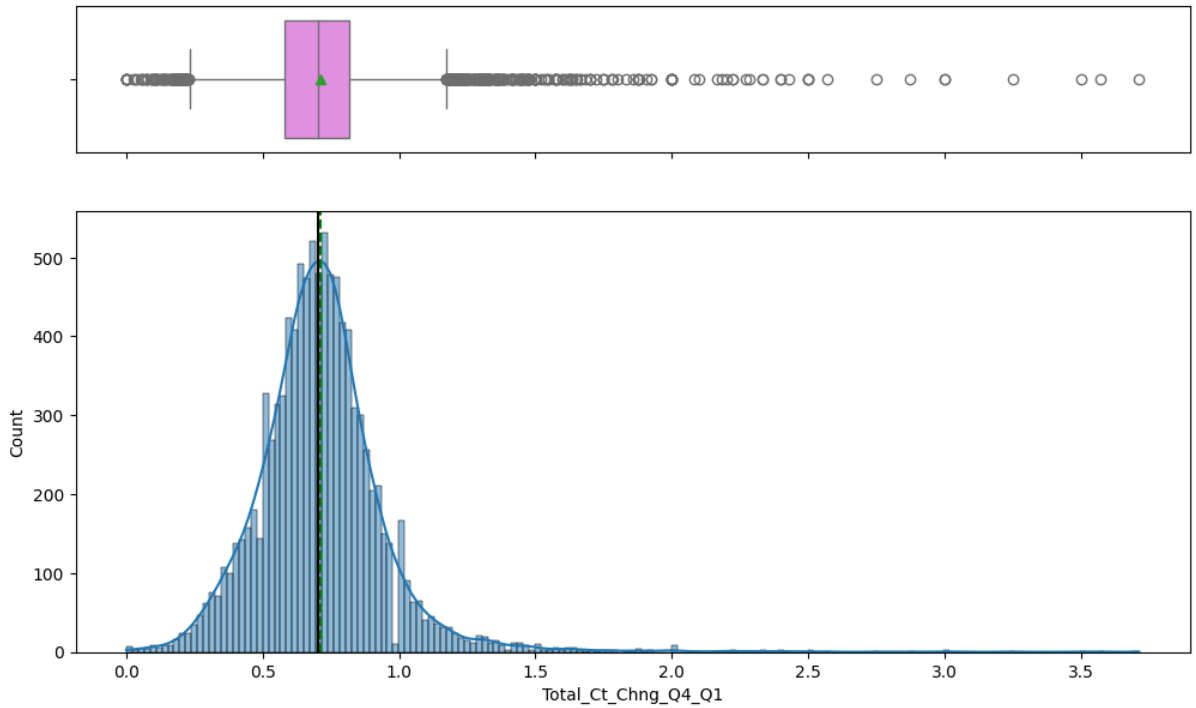
viii. Total Transaction Amount:



- Spending is concentrated between **₹2,000 and ₹6,000**.
- A few customers spend over **₹15,000+**, marked as outliers.
- There appear to be **three spending tiers** — low, medium, and high.

In short: Most customers are **mid-range spenders**, but there are **clear high-value clients** who can be prioritized.

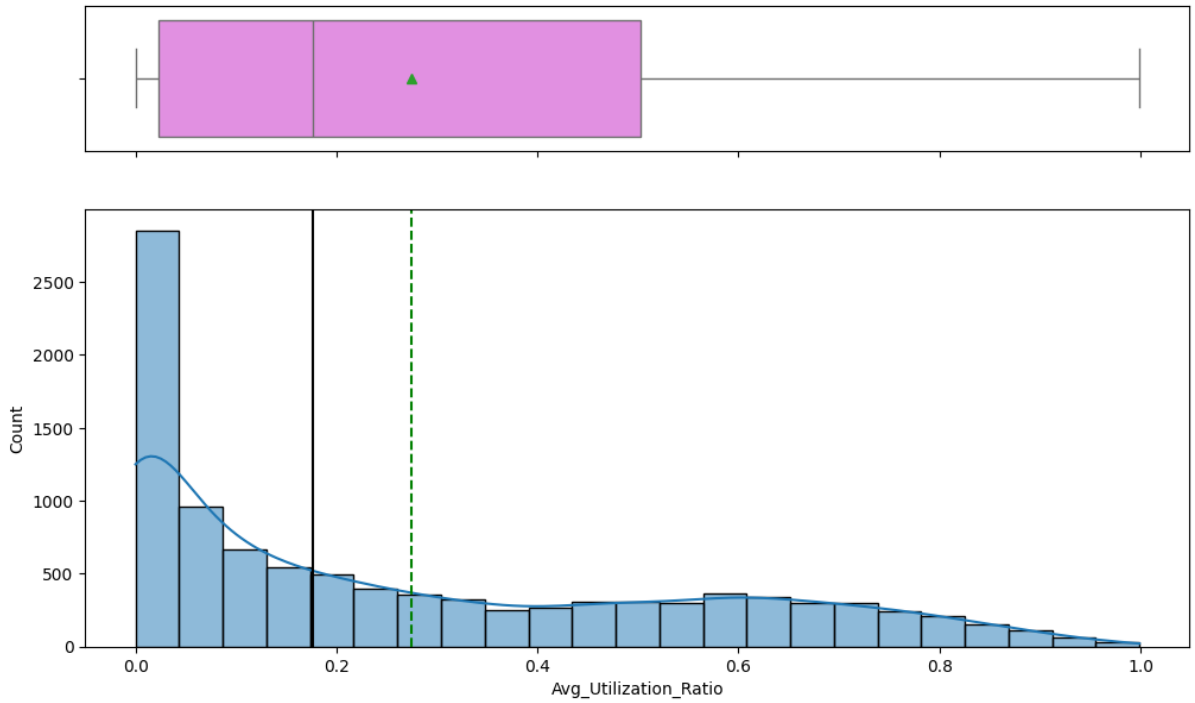
ix. Total Transaction Count Change (Q4 vs Q1):



- Most changes are centered around **0.7**, meaning customers' transaction counts remain fairly stable.
- A few extreme spikes or drops are seen, but they’re **outliers**.

In short: The **number of transactions stays stable** for most customers. Big changes could indicate a **shift in satisfaction or habits**.

x. Average Utilization Ratio:



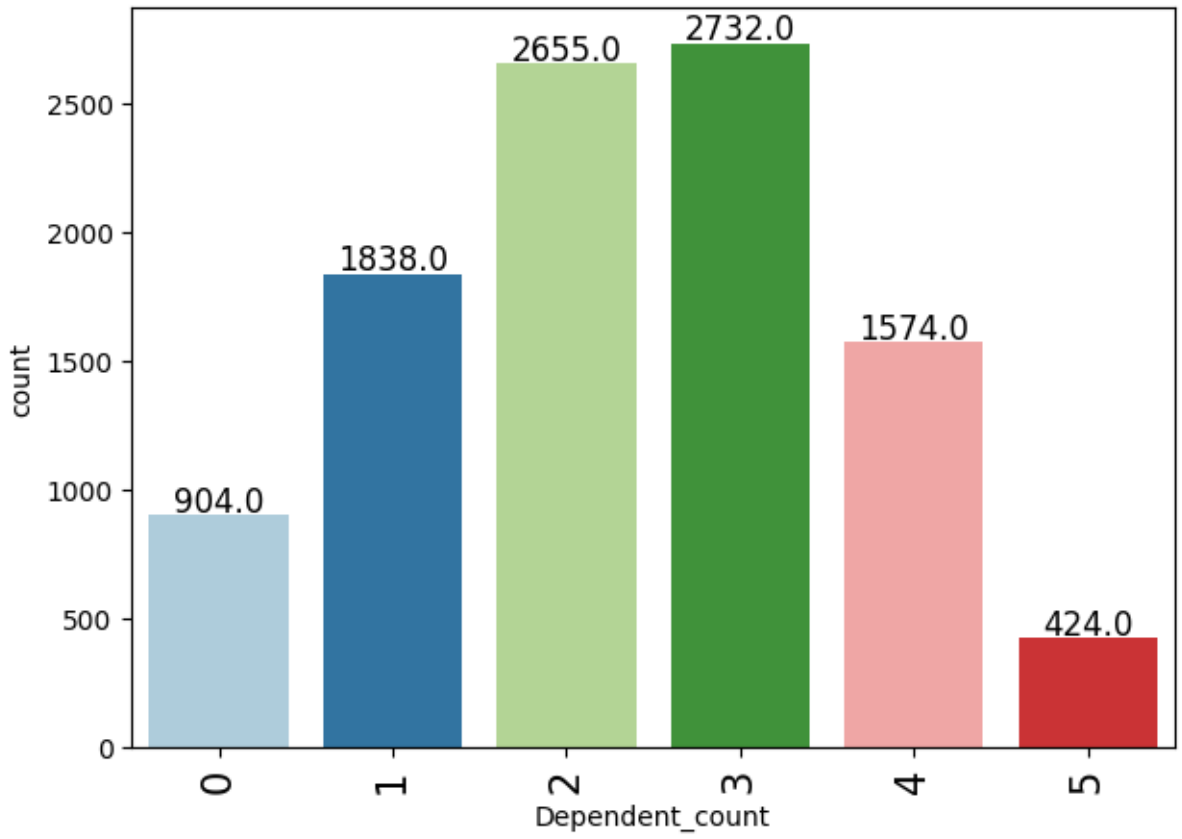
1. **Boxplot (Top)**:
    - The box is skewed to the left, indicating a right-skewed distribution.
    - Median is near 0.2.
    - A green triangle marks the mean, slightly above the median.
    - There are outliers extending close to 1.0.
2. **Histogram with KDE (Bottom)**:
    - Most values are concentrated near 0.0.
    - Frequency drops as utilization increases.
    - Two vertical lines:

- **Black solid** likely represents the median (~0.2).
- **Green dashed** marks the mean (~0.28–0.3).

Conclusion: Avg_Utilization_Ratio is heavily right-skewed, with most users having low utilization and a few with high usage.

Labeled Barplot

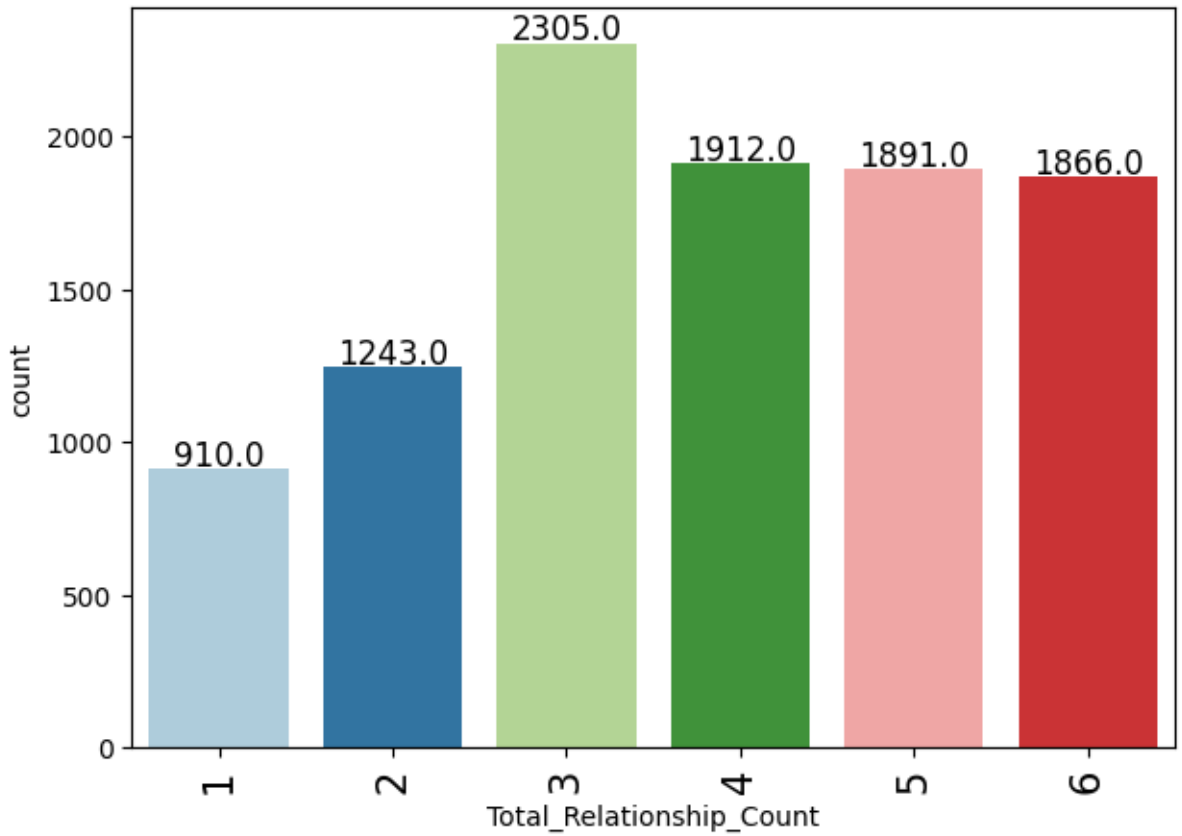
i.Dependent Count:



- The most common number of dependents is 3, with 2732 individuals.
- Followed closely by 2 dependents (2655 people) and 1 dependent (1838 people).
- Fewer individuals have 0 dependents (904) or 5 dependents (424).
- There's a noticeable drop after 3 dependents, suggesting larger families are less common.

Conclusion: Most individuals have 2–3 dependents, with the number of people decreasing as dependents increase beyond 3.

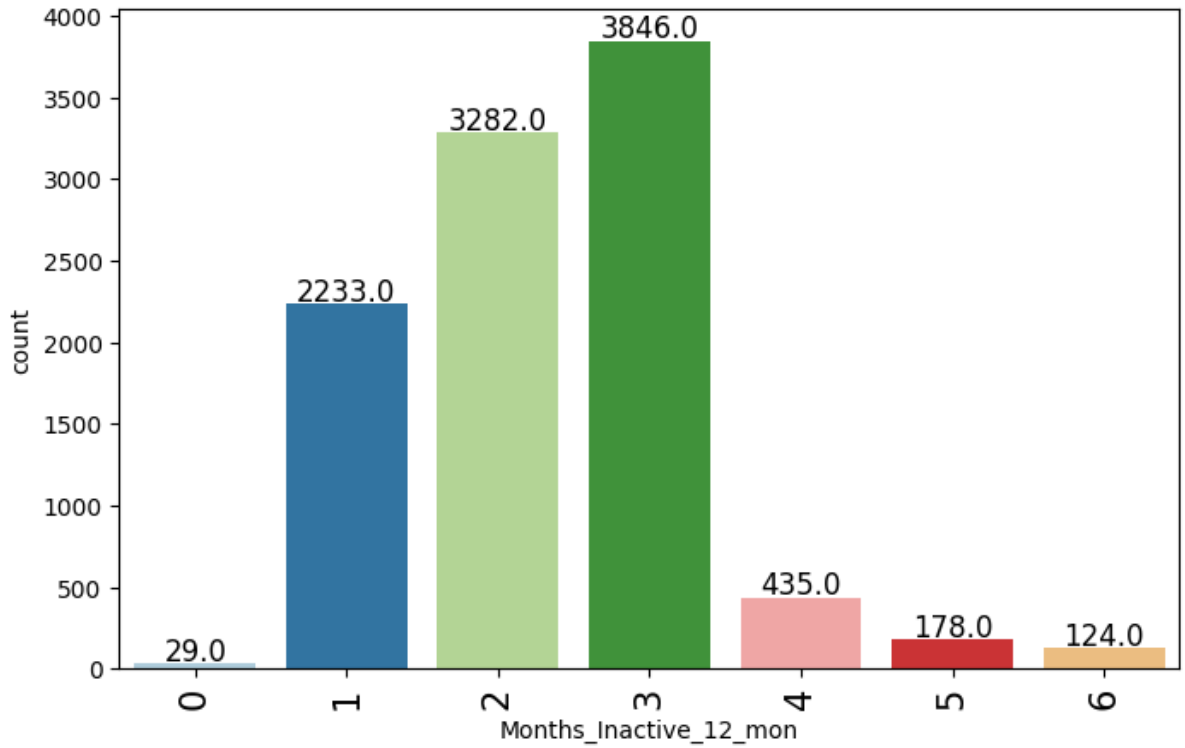
ii. Total Relationship Count:



- The most frequent count is 3 relationships, with 2305 individuals.
- Other common counts are 4 (1912), 5 (1891), and 6 (1866) — relatively close in frequency.
- 2 relationships have 1243 individuals, while 1 relationship is the least common, with 910 individuals.

Conclusion: Most individuals have 3 to 6 total relationships, indicating moderate to high relational engagement. Having only 1 or 2 relationships is relatively less common in the dataset.

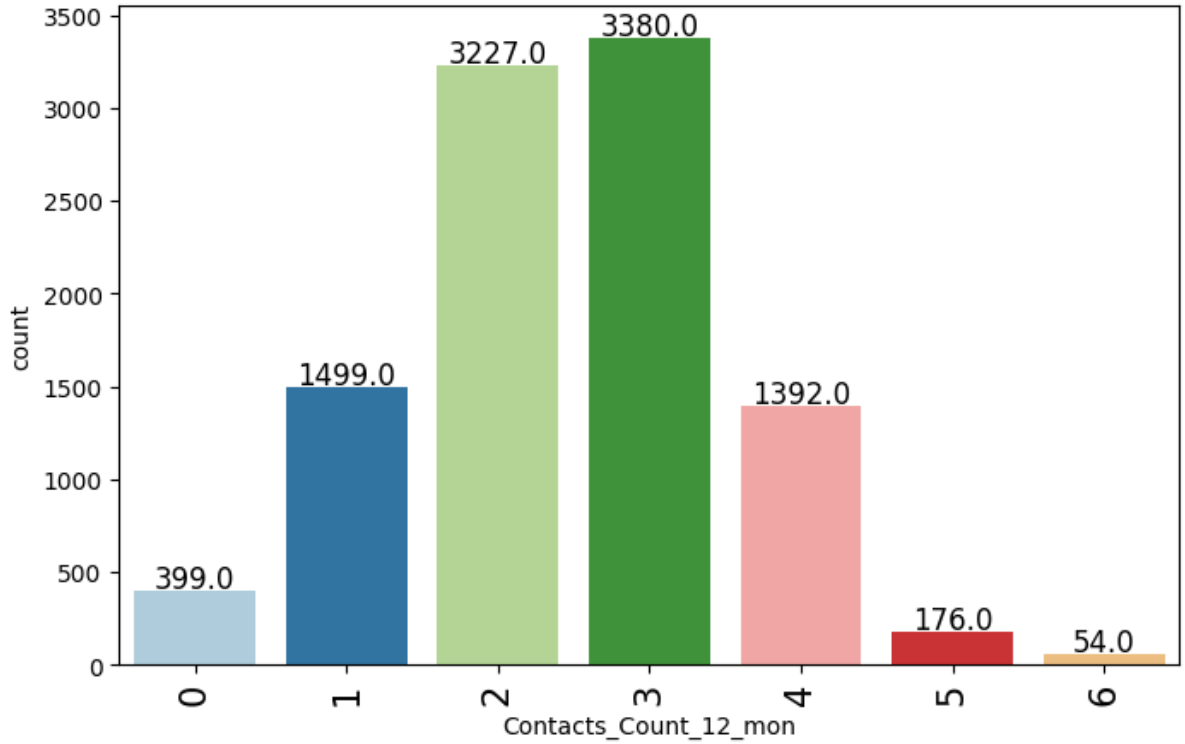
iii. Months_Inactive_12_mon:



- The highest count is for 3 inactive months, with 3846 individuals.
- Followed by 2 months (3282) and 1 month (2233).
- Very few individuals had 0 inactive months (29), and the count steadily decreases beyond 3 months.
- 4 to 6 months show a sharp drop: 435, 178, and 124 respectively.

Conclusion: Most users were inactive for 1 to 3 months in the past year, with 3 months being the most common. Long periods of inactivity (4+ months) are relatively rare.

iv. Contacts_Count_12_mon:

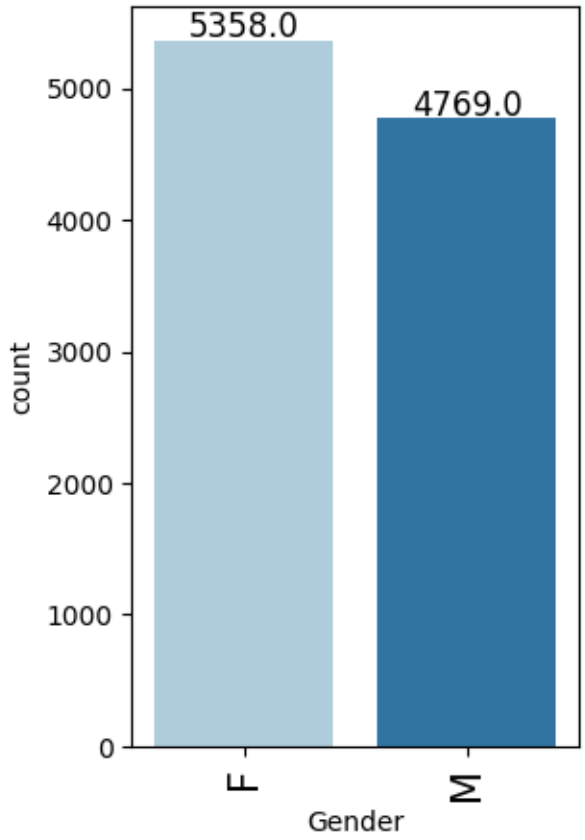


- The highest count is for 3 inactive months, with 3380 individuals.
- Followed by 2 months (3227) and 1 month (1499).
- Very few individuals had 0 inactive months (399), and the count steadily decreases beyond 5 months(176) and 6 months(54).

4 to 6 months show a sharp drop: 1392, 176, and 54 respectively.

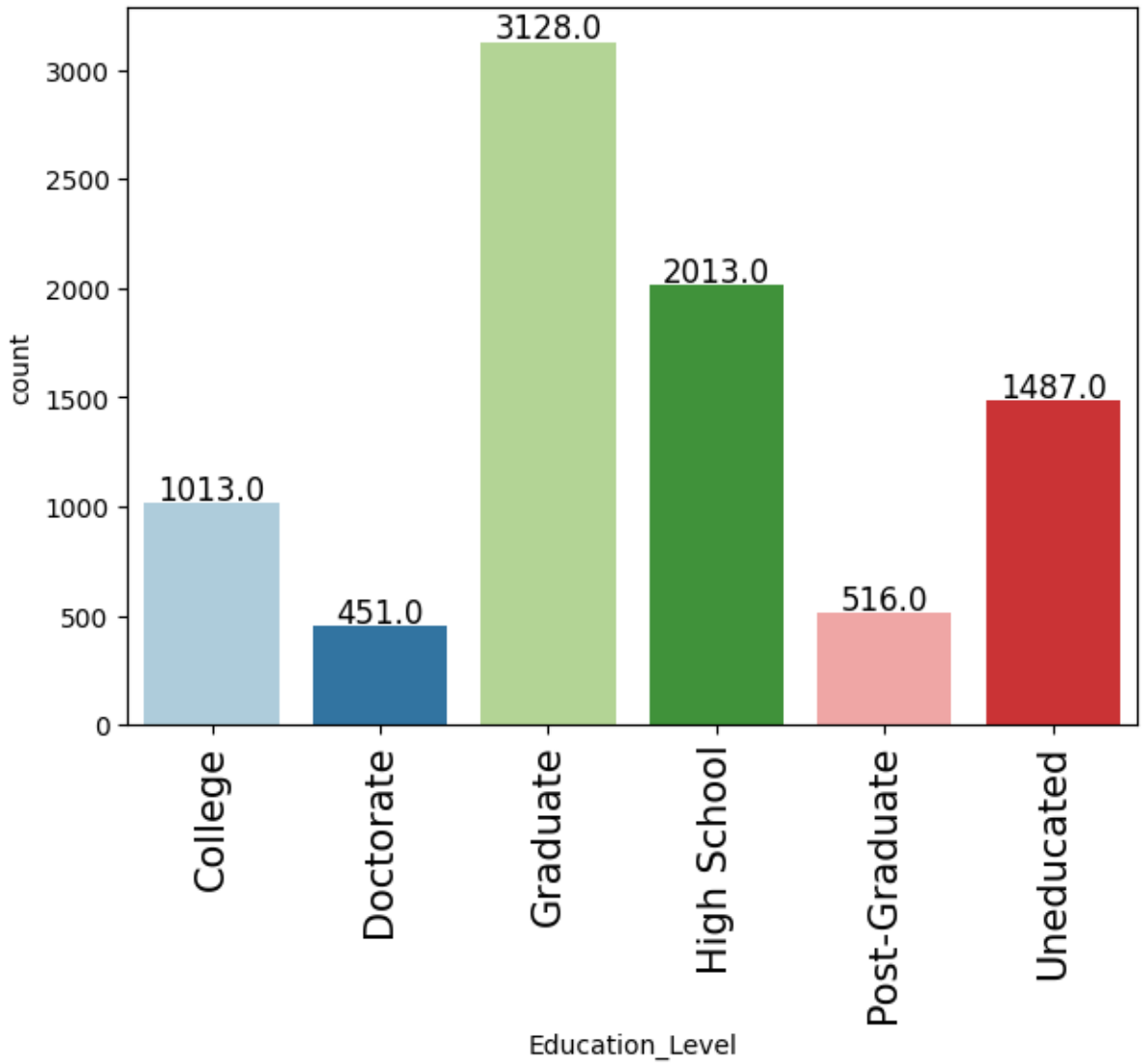
Conclusion: Most users contacts count were high for 1 to 3 months in the past year, with 3 months being the most common. Long periods of no contacts (4+ months) are relatively rare.

iv. Gender:



- The gender count for female is high for 5358 and male with less number is 4769.
- The Thera Bank customers most of them are female with more than 589.

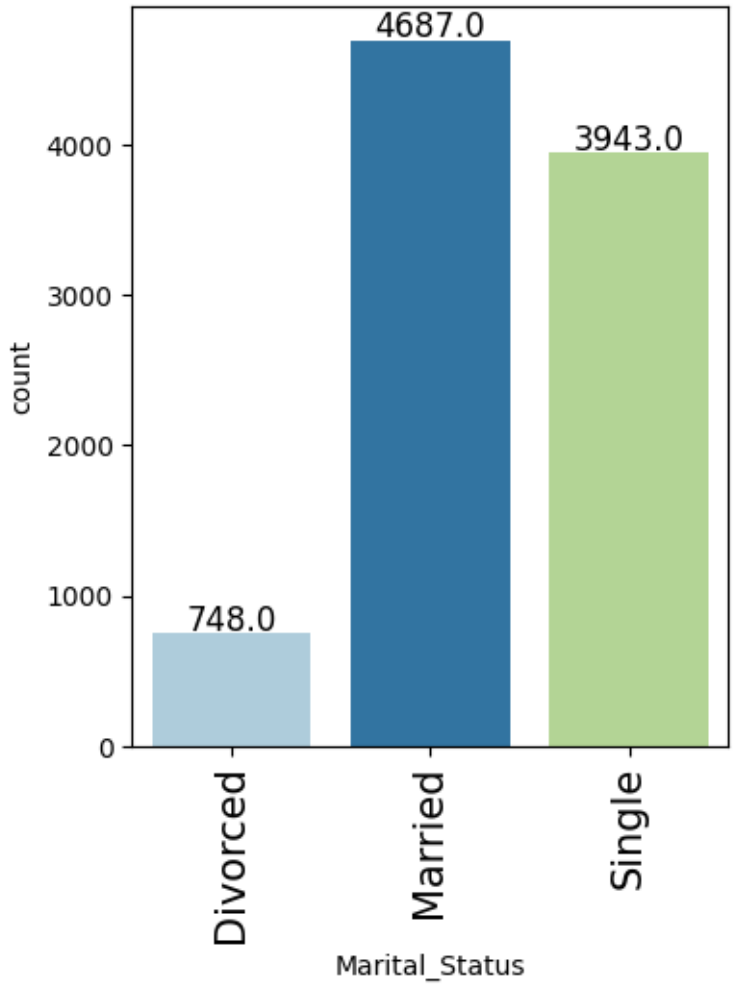
iv. Education_Level:



- **Graduate** is the most common level, with **3128** individuals.
- Followed by **High School** (2013) and **Uneducated** (1487).
- **College** has 1013 individuals.
- **Post-Graduate** and **Doctorate** levels are the least common, with **516** and **451**, respectively.

Conclusion: The majority of the population is **Graduate** or **High School** educated. Advanced degrees (Post-Graduate, Doctorate) are relatively rare, while a notable number of individuals are still **Uneducated**.

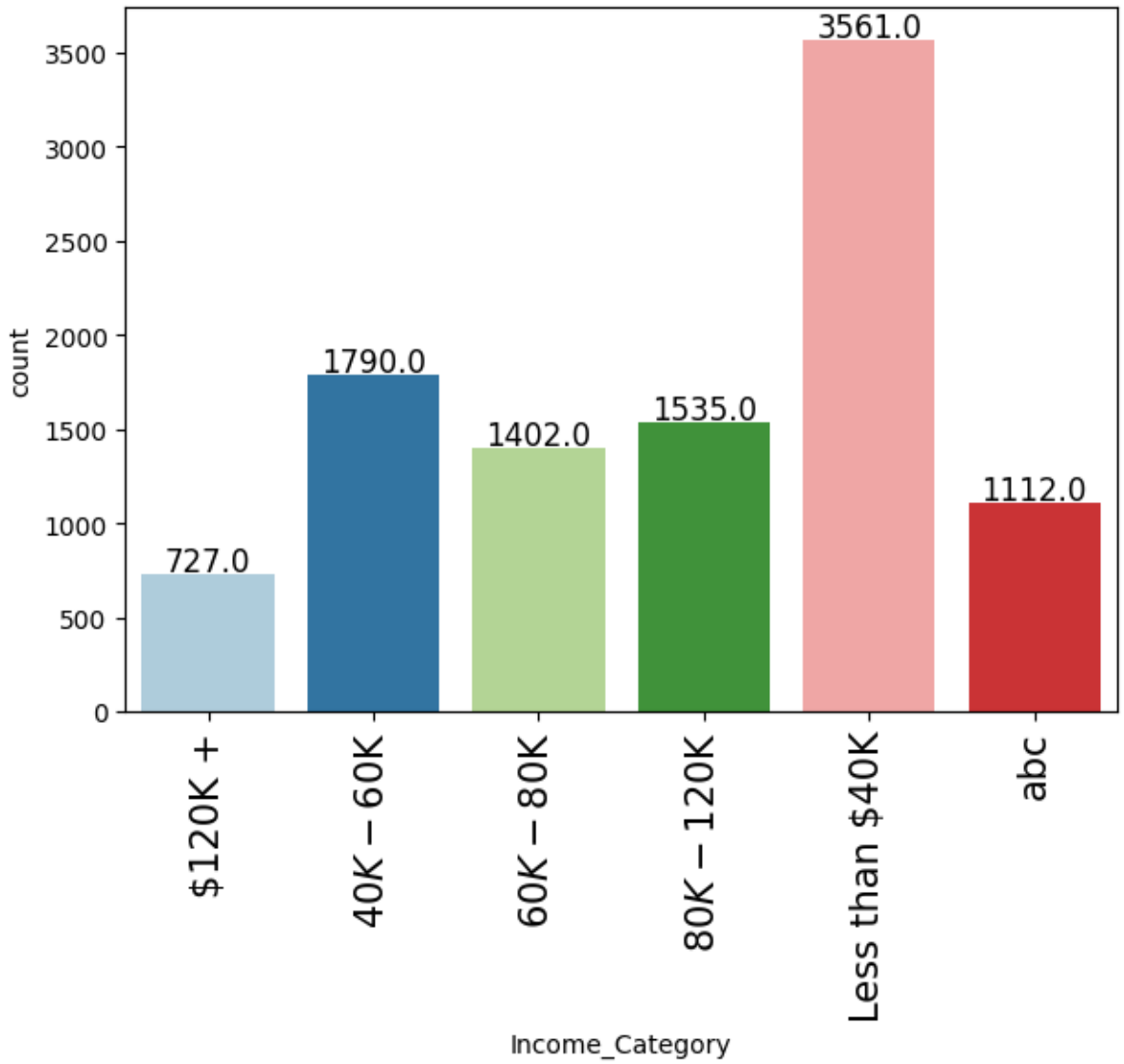
iv. Marital_Status:



- **Married** is the most common level, with **4687** individuals.
- Followed by **Divorced** (748) and **single** (3943).
- **College** has 9378 individuals.
- **Divorced** level is the least common, with **748** respectively.

Conclusion: The majority of the martial status is **Married**.
Divorced are relatively rare, while a notable number of singles are still **Unmarried**.

v. Income_Category:



"Less than $40K" is the most frequent group, with 3561 individuals.

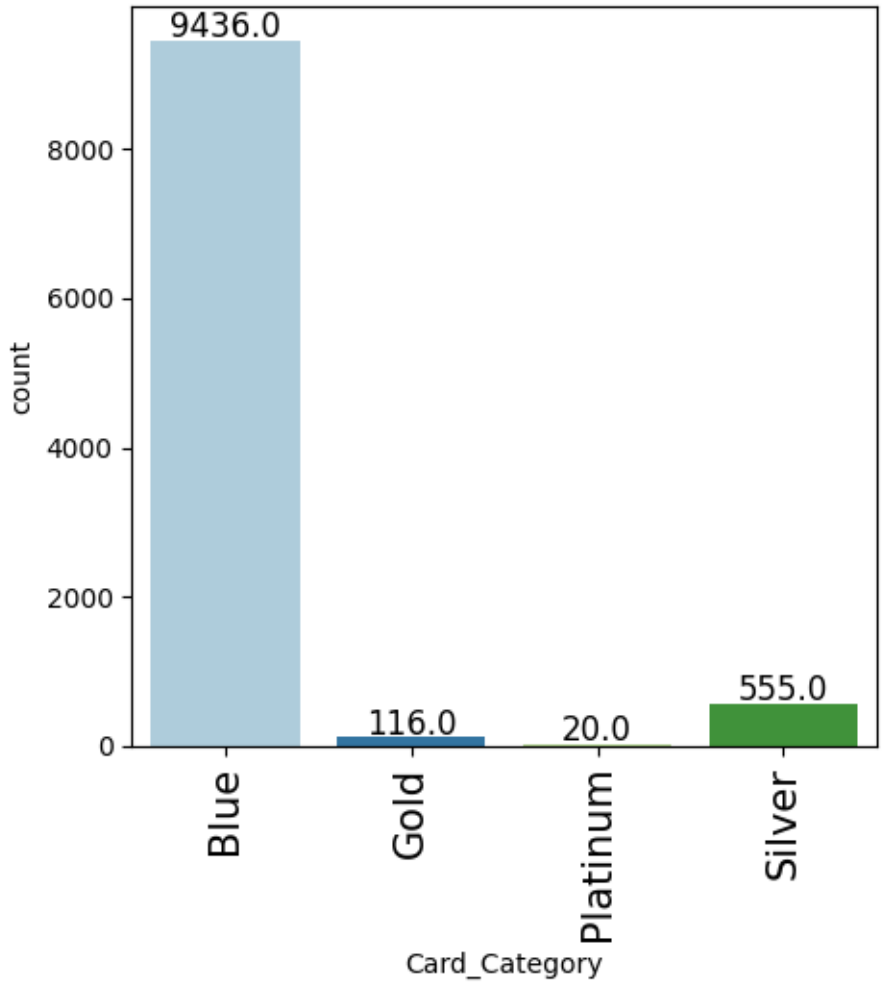
Followed by:

- $40K - $60K: 1790
- $80K - $120K: 1535
- $60K - $80K: 1402
- $120K+: 727

There is an unusual or likely erroneous category: "abc" with 1112 individuals.

Conclusion: The dataset is heavily skewed toward lower-income groups, especially those earning less than $40K. The presence of "abc" suggests data quality issues and may require data cleaning or imputation before modeling.

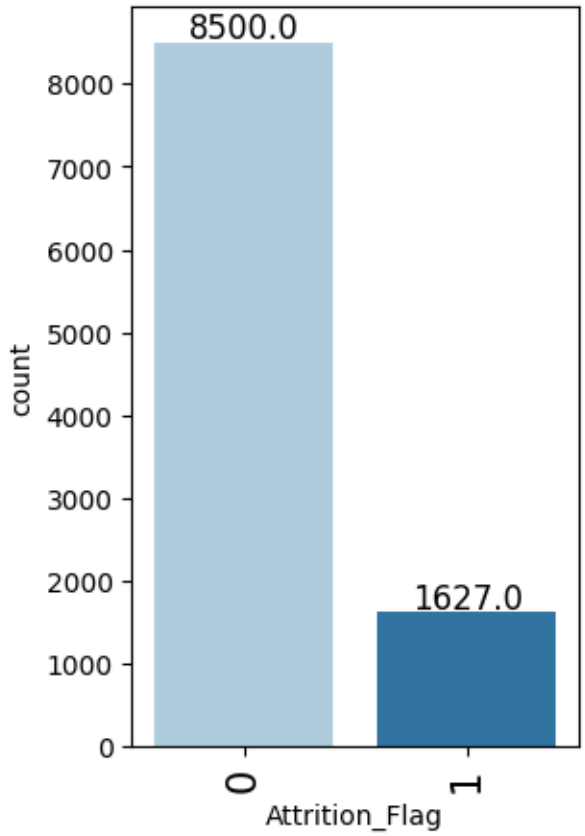
vi. Card_Category:



- **Blue** is the most common level, with **9436** individuals.
- Followed by **Gold** (116) and **silver** (555).
- **Card Category** has 10127 individuals.
- **Platinum** level is the least common, with **20** respectively.

Conclusion: The majority of the population is **Blue card.** **Platinum card** is relatively rare, while a notable number of cards are **Gold** and **silver** respectively.

vii. Attrition_Flag:

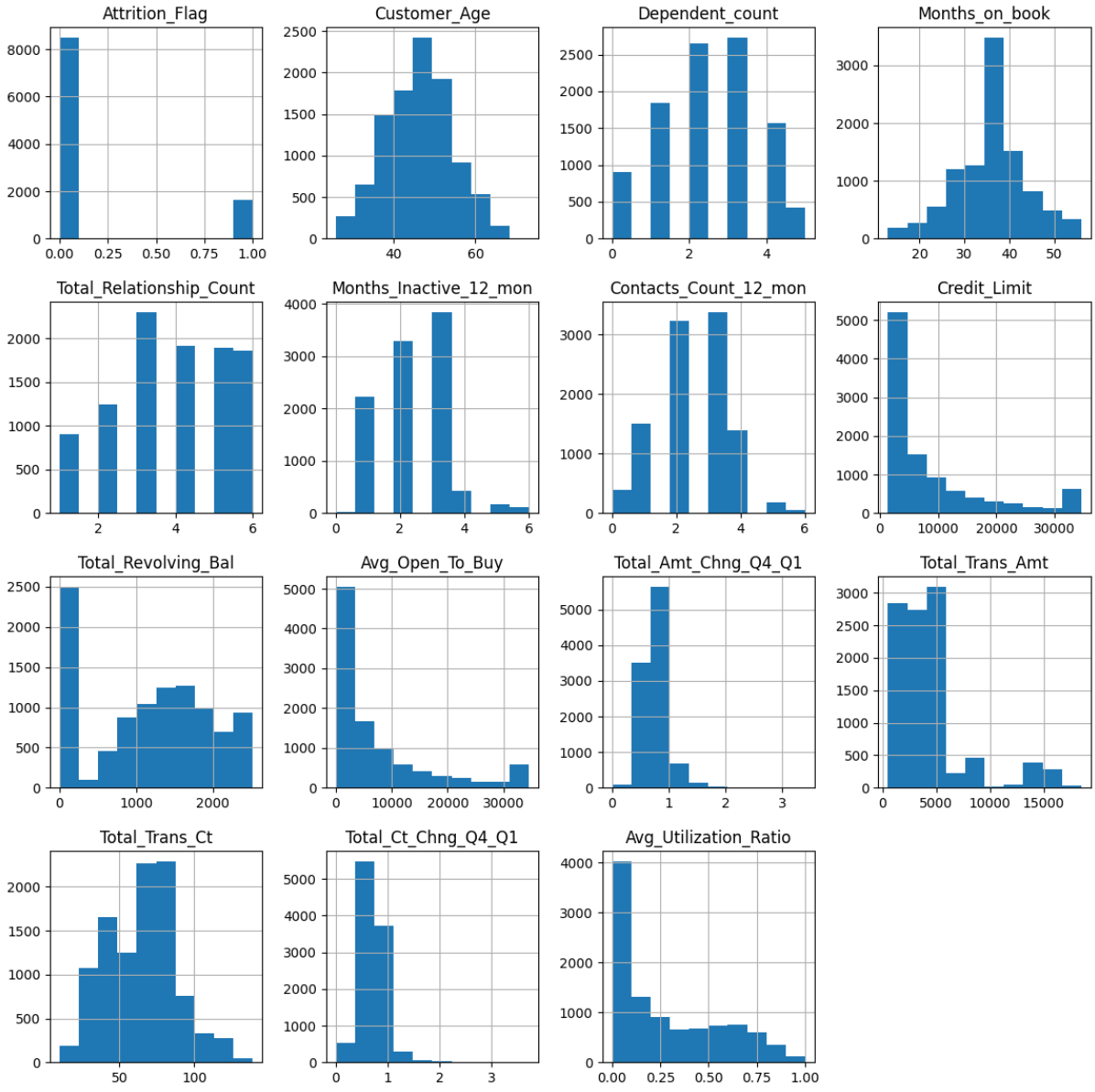


The histogram shows the distribution of a variable labeled "Attrition_Flag" with two categories: 0 and 1. The y-axis represents the count of occurrences.

- Category 0 has a count of 8500.
- Category 1 has a count of 1627.

This indicates that there are significantly more instances of category 0 compared to category 1, with a ratio of approximately 5:1.

viii. Histograms Of All Columns:



Attrition_Flag: Binary variable (0 or 1). Most customers (around 8500) have a value of 0, while about 1627 have a value of 1, indicating a significant imbalance (ratio ~5:1).

Customer_Age: Ages range from 20 to 80, with a peak around 40–50 years. The distribution is roughly symmetric, with most customers between 30 and 60.

Dependent_count: Ranges from 0 to 5 dependents. Most customers have 2–3 dependents, with fewer having 0, 4, or 5.

Months_on_book: Represents the duration of the customer relationship (in months). Most customers have been with the bank for 20–50 months, peaking around 30–40 months.

Total_Relationship_Count: Number of products/relationships with the bank (1 to 6). Most customers have 3–5 relationships, with a peak at 4.

Months_Inactive_12_mon: Months inactive in the last 12 months (0 to 6). Most customers have 2–3 inactive months, with fewer being inactive for 5–6 months.

Contacts_Count_12_mon: Number of contacts in the last 12 months (0 to 6). Most customers have 2–3 contacts, with a peak at 3.

Credit_Limit: Credit limits range from 0 to 35,000. Most customers have a credit limit below 10,000, with a long right tail extending to 35,000.

Total_Revolving_Bal: Revolving balance ranges from 0 to 2500. Many customers have a balance of 0, with the rest spread up to 2500, peaking around 1000–1500.

Avg_Open_To_Buy: Average open-to-buy credit (0 to 35,000). Most customers have a value below 10,000, with a long right tail similar to Credit_Limit.

Total_Amt_Chng_Q4_Q1: Change in transaction amount from Q4 to Q1 (0 to 3). Most values are between 0 and 1.5, peaking around 0.5–1.0.

Total_Trans_Amt: Total transaction amount (0 to 20,000). Most transactions are below 5000, with a peak around 2500–5000 and a long right tail.

Total_Trans_Ct: Total transaction count (0 to 150). Most customers have 20–80 transactions, peaking around 40–60.

Total_Ct_Chng_Q4_Q1: Change in transaction count from Q4 to Q1 (0 to 3). Most values are between 0 and 1.5, peaking around 0.5–1.0.

Avg_Utilization_Ratio: Utilization ratio (0 to 1). Most customers have a ratio below 0.5, with a peak near 0 and a long tail toward 1.

Conclusion:

Attrition_Flag shows a class imbalance, which might be important for predictive modeling.

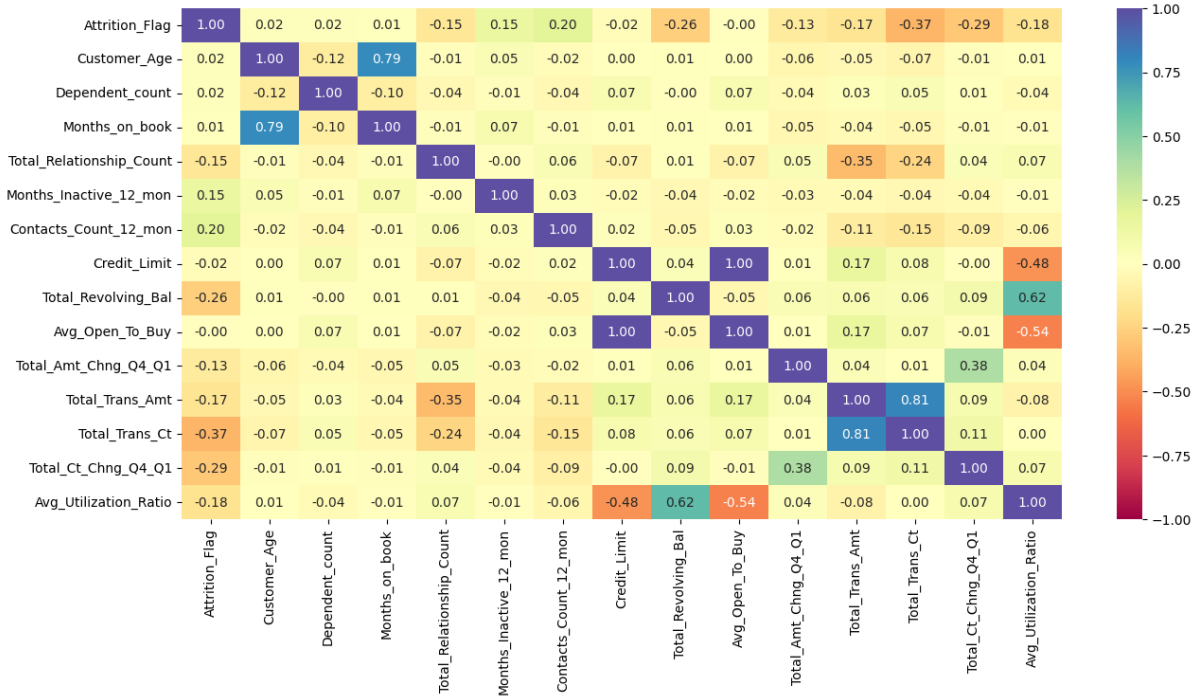
Customer_Age, Months_on_book, and Total_Relationship_Count have fairly symmetric distributions.

Variables like Credit_Limit, Avg_Open_To_Buy, and Total_Trans_Amt are heavily right-skewed, indicating a small number of customers with very high values.

Avg_Utilization_Ratio and Total_Revolving_Bal show many customers with low or zero values, suggesting low credit usage for a significant portion of the population.

Bivariate Distributions

Heatmap:



The heatmap displays the correlation matrix for various customer attributes in a financial dataset. The color scale ranges from -1 (strong negative correlation, red) to 1 (strong positive correlation, blue), with 0 (no correlation) in yellow-green.

Key Observations:

1. **Strong Positive Correlations:**

Credit_Limit and Avg_Open_To_Buy: 1.00 (perfect correlation), indicating they are essentially the same or directly derived from each other.

Total_Trans_Amt and Total_Trans_Ct: 0.81, suggesting that higher transaction amounts strongly correlate with more transactions.

Total_Revolving_Bal and Avg_Utilization_Ratio: 0.62, showing that higher revolving balances are associated with higher utilization ratios.

Total_Amt_Chng_Q4_Q1 and Total_Ct_Chng_Q4_Q1: 0.38, indicating a moderate correlation between changes in transaction amount and count over quarters.

2. Strong Negative Correlations:

Credit_Limit and Avg_Utilization_Ratio: -0.48, meaning higher credit limits are associated with lower utilization ratios.

Avg_Open_To_Buy and Avg_Utilization_Ratio: -0.54, a stronger negative correlation, as higher available credit reduces utilization.

Total_Trans_Amt and Total_Relationship_Count: -0.35, suggesting that customers with more relationships (products) tend to have lower transaction amounts.

Attrition_Flag and Total_Trans_Ct: -0.37, indicating that customers who churn (Attrition_Flag = 1) tend to have fewer transactions.

3. Notable Correlations with Attrition_Flag:

Total_Trans_Ct: -0.37 (negative), customers with fewer transactions are more likely to churn.

Total_Trans_Amt: -0.17 (negative), lower transaction amounts are associated with higher churn.

Contacts_Count_12_mon: 0.20 (positive), more contacts in the last 12 months correlate with higher churn.

Months_Inactive_12_mon: 0.15 (positive), more inactive months are associated with higher churn.

Total_Revolving_Bal: -0.26 (negative), lower revolving balances are linked to higher churn.

4. Weak or No Correlations:

Customer_Age has minimal correlation with most variables, with the strongest being -0.12 with **Dependent_count**.

Months_on_book shows a 0.79 correlation with **Customer_Age,** indicating older customers tend to have longer relationships with the bank.

- Most other correlations are close to 0 (e.g., **Dependent_count** with most variables), indicating little to no linear relationship.

Conclusion:

Attrition_Flag is influenced by transactional behavior (lower transactions, amounts, and revolving balances increase churn

likelihood) and engagement (more inactivity and contacts increase churn).

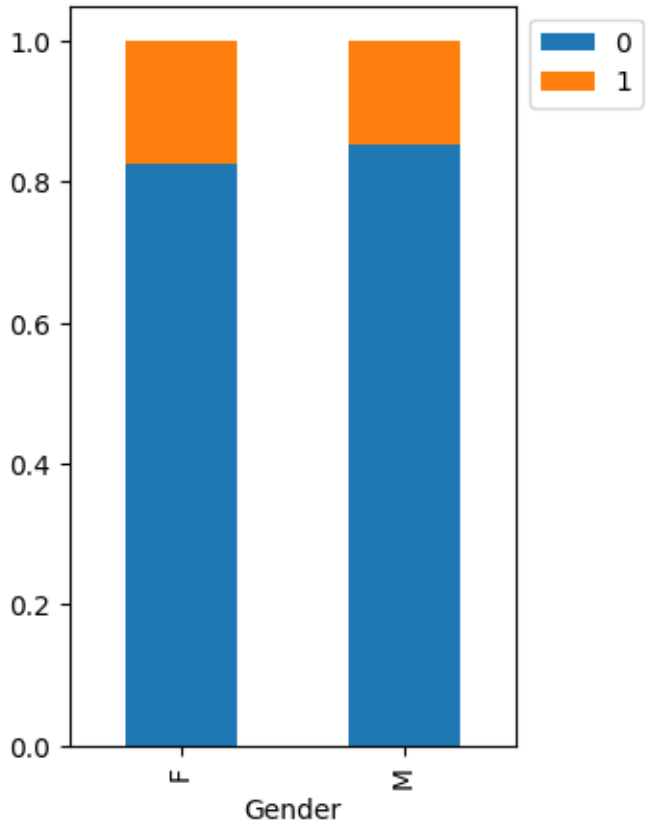
Credit_Limit and **Avg_Open_To_Buy** are redundant variables due to their perfect correlation.

- Customers with more relationships **(Total_Relationship_Count)** tend to have lower transaction amounts and counts, which might indicate more stable or less active accounts.

Stacked Barplot:

Attrition_Flag vs Gender

Attrition_Flag	**0**	**1**	**All**
Gender			
All	8500	1627	10127
F	4428	930	5358
M	4072	697	4769



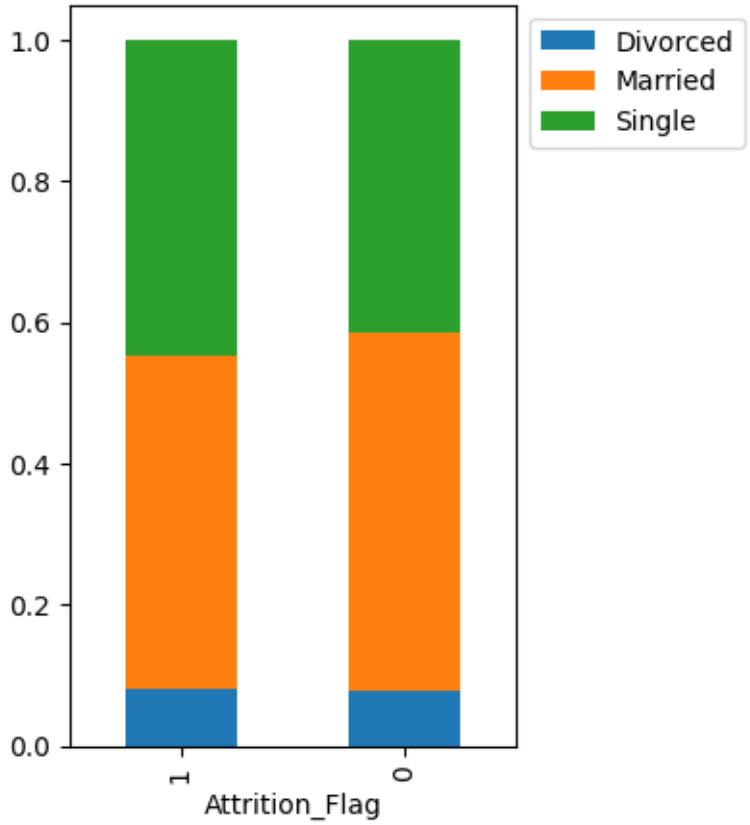
The stacked bar chart shows the distribution of **Attrition_Flag** across genders:

- **Female (F)**: 0.85 (Attrition_Flag = 0), 0.15 (Attrition_Flag = 1)
- **Male (M)**: 0.83 (Attrition_Flag = 0), 0.17 (Attrition_Flag = 1)

Churn rate is slightly higher for males (0.17) than females (0.15), but the difference is small.

Attrition_Flag vs Marital_Status:

Marital_Status	**Divorced**	**Married**	**Single**	**All**
Attrition_Flag				
All	748	4687	3943	9378
0	627	3978	3275	7880
1	121	709	668	1498



The stacked bar chart shows the distribution of marital status across **Attrition_Flag** (0 and 1):

- **Attrition_Flag = 0** (non-churned):

    - Divorced (blue): ~0.1
    - Married (orange): ~0.5
    - Single (green): ~0.4
- **Attrition_Flag = 1** (churned):
    - Divorced (blue): ~0.1
    - Married (orange): ~0.45
    - Single (green): ~0.45

Interpretation:

- Among non-churned customers, 50% are married, 40% single, and 10% divorced.
- Among churned customers, the split between married and single is roughly equal (45% each), with 10% divorced.
- Single customers show a slightly higher tendency to churn (45% vs. 40%), while marital status overall has a modest impact on attrition.

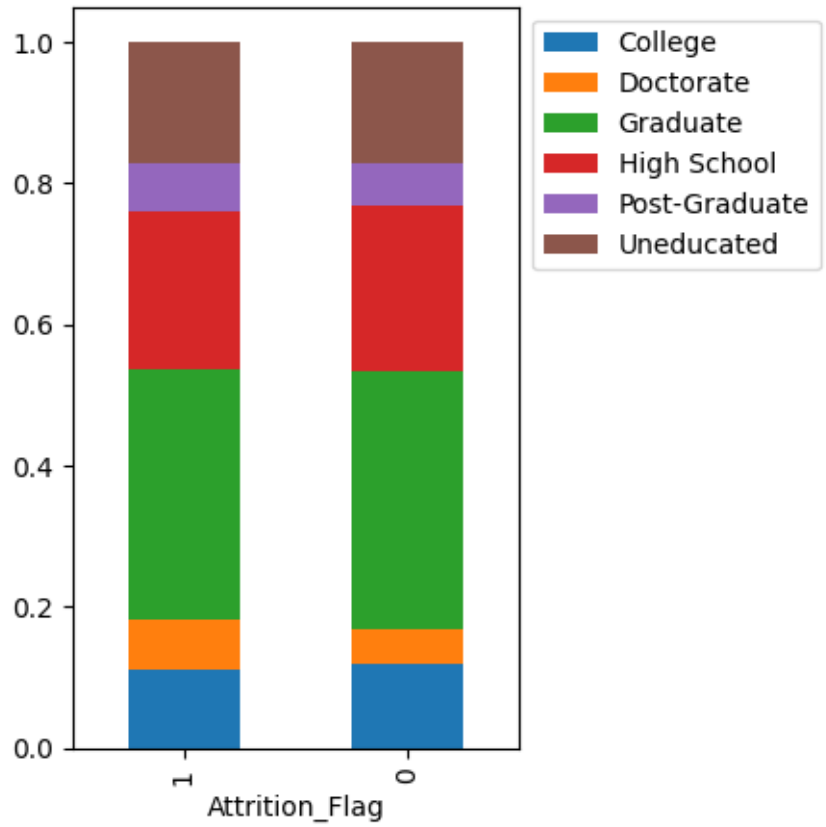
Attrition_Flag vs Education_Level:

**Education_Level College Doctorate Graduate High School Post-Graduate **

Attrition_Flag

All	1013	451	3128	2013	516
0	859	356	2641	1707	424
1	154	95	487	306	92

Education_Level	**Uneducated**	**All**
Attrition_Flag		
All	1487	8608
0	1250	7237
1	237	1371



- Attrition_Flag = 0 (non-churned):

    - College (blue): ~0.1
    - Doctorate (orange): ~0.05
    - Graduate (green): ~0.3
    - High School (red): ~0.2
    - Post-Graduate (purple): ~0.05
    - Uneducated (brown): ~0.3
- Attrition_Flag = 1 (churned):
    - College (blue): ~0.1
    - Doctorate (orange): ~0.05
    - Graduate (green): ~0.3
    - High School (red): ~0.2
    - Post-Graduate (purple): ~0.05
    - Uneducated (brown): ~0.3

Observation: Education level distribution is nearly identical for churned and non-churned customers, suggesting it has little impact on attrition.

Attrition_Flag vs Income_Category:

Income_Category $120K + $40K - $60K $60K - $80K $80K - $120K \

Attrition_Flag

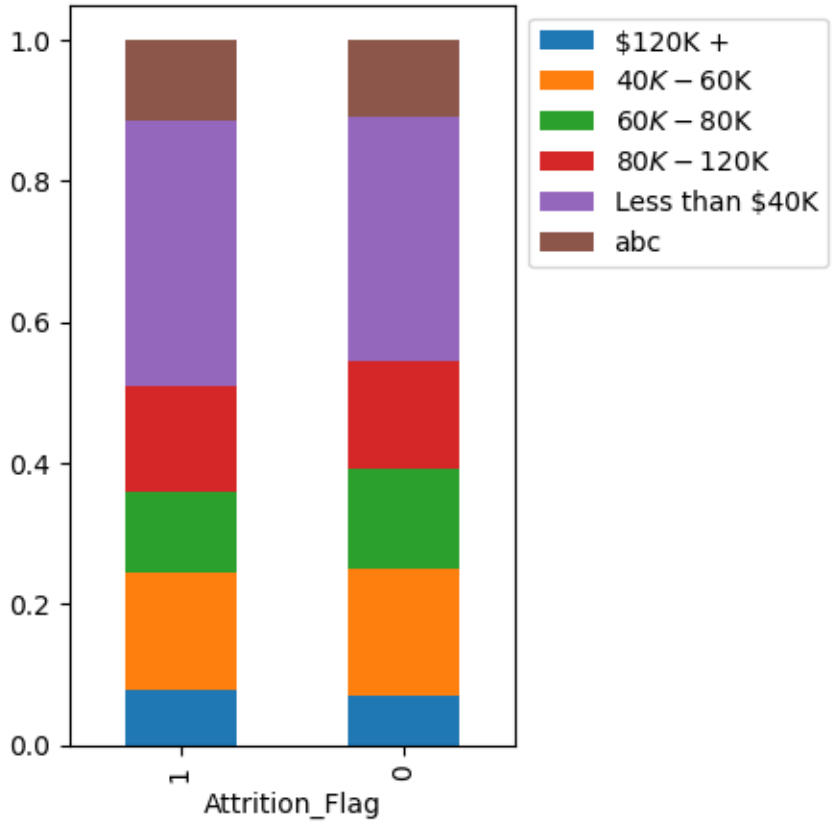
All 727 1790 1402 1535

0 601 1519 1213 1293

1 126 271 189 242

Income_Category Less than $40K abc All

Attrition_Flag			
All	3561	1112	10127
0	2949	925	8500
1	612	187	1627



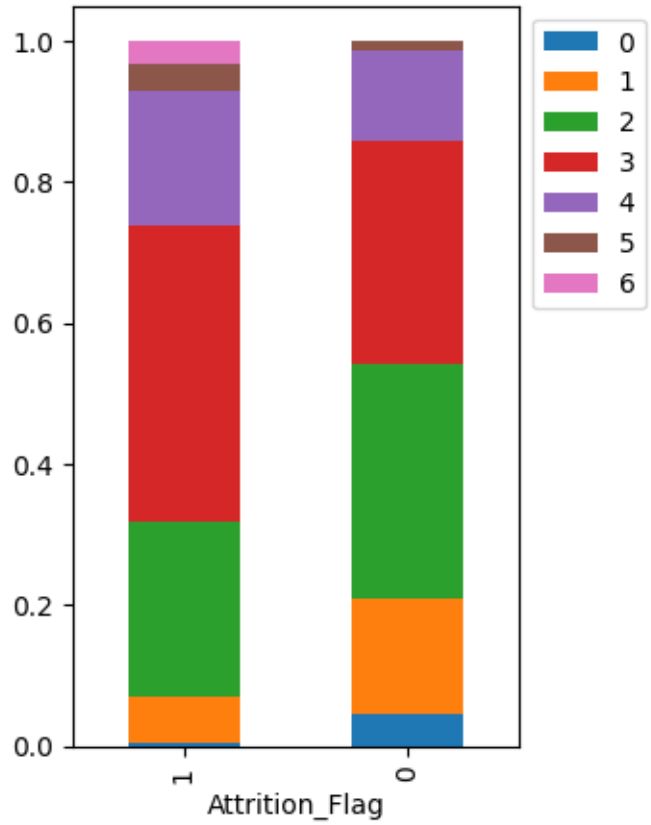
- **Attrition_Flag = 0** (non-churned):
  - $120K+ (blue): ~0.1
  - $40K-$60K (orange): ~0.2
  - $60K-$80K (green): ~0.25
  - $80K-$120K (red): ~0.25
  - Less than $40K (purple): ~0.2
- **Attrition_Flag = 1** (churned):
  - $120K+ (blue): ~0.1
  - $40K-$60K (orange): ~0.2

- $60K-$80K (green): ~0.25
- $80K-$120K (red): ~0.25
- Less than $40K (purple): ~0.2

Observation: Income category distribution is consistent between churned and non-churned customers, indicating minimal influence on attrition.

Attrition_Flag vs Contacts_Count_12_mon:

Contacts_Count_12_mon	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**All**
Attrition_Flag								
1	7	108	403	681	315	59	54	1627
All	399	1499	3227	3380	1392	176	54	10127
0	392	1391	2824	2699	1077	117	0	8500



- **Attrition_Flag = 0** (non-churned):

    - 0 (blue): ~0.05
    - 1 (orange): ~0.15
    - 2 (green): ~0.25
    - 3 (red): ~0.35
    - 4 (purple): ~0.15
    - 5 (brown): ~0.05
- **Attrition_Flag = 1** (churned):
    - 0 (blue): ~0.05
    - 1 (orange): ~0.15
    - 2 (green): ~0.25
    - 3 (red): ~0.35
    - 4 (purple): ~0.15
    - 5 (brown): ~0.05

Observation: Card category distribution is the same for both groups, suggesting no significant impact on churn.

Attrition_Flag vs Months_Inactive_12_mon:

Months_Inactive_12_mon	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**All**
Attrition_Flag								
All	29	2233	3282	3846	435	178	124	10127
1	15	100	505	826	130	32	19	1627
0	14	2133	2777	3020	305	146	105	8500

- Attrition_Flag = 1 (churned, left bar):
    - 0 (blue): ~0.05
    - 1 (orange): ~0.15

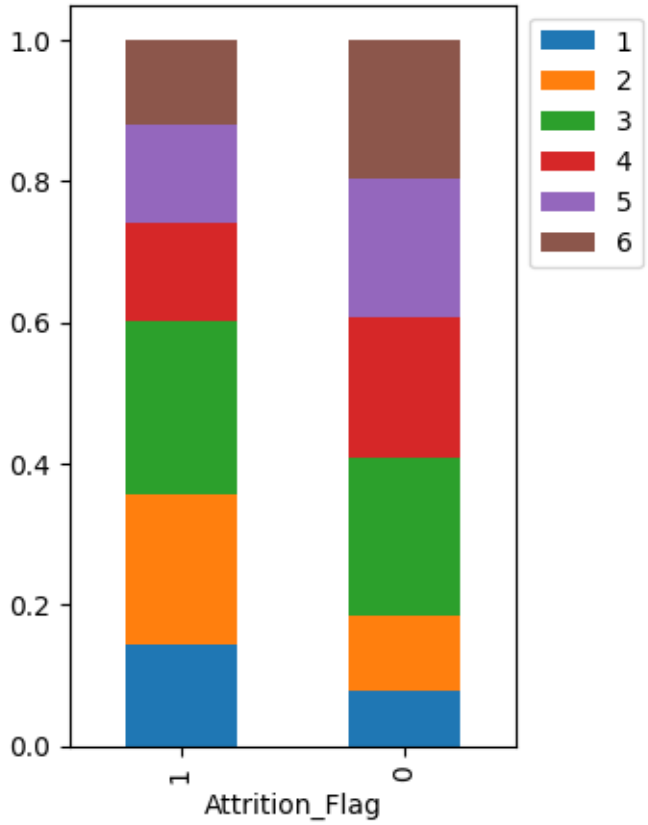
    - 2 (green): ~0.25
    - 3 (red): ~0.35
    - 4 (purple): ~0.15
    - 5 (brown): ~0.05
- Attrition_Flag = 0 (non-churned, right bar):
    - 0 (blue): ~0.05
    - 1 (orange): ~0.15
    - 2 (green): ~0.25
    - 3 (red): ~0.35
    - 4 (purple): ~0.15
    - 5 (brown): ~0.05

Interpretation:

- The distribution of card categories is identical for both churned and non-churned customers.
- The most common card category is 3 (red, 35%), followed by category 2 (green, 25%).
- Categories 0 and 5 (blue and brown, 5% each) are the least common.
- Since the proportions are the same, the type of card a customer holds does not appear to influence their likelihood of churning.

Attrition_Flag vs Total_Relationship_Count:

Total_Relationship_Count	1	2	3	4	5	6	All
Attrition_Flag							
All	910	1243	2305	1912	1891	1866	10127
0	677	897	1905	1687	1664	1670	8500
1	233	346	400	225	227	196	1627



- Attrition_Flag = 1 (churned, left bar):
  - 0 (blue): ~0.05
  - 1 (orange): ~0.15
  - 2 (green): ~0.25
  - 3 (red): ~0.35
  - 4 (purple): ~0.15
  - 5 (brown): ~0.05

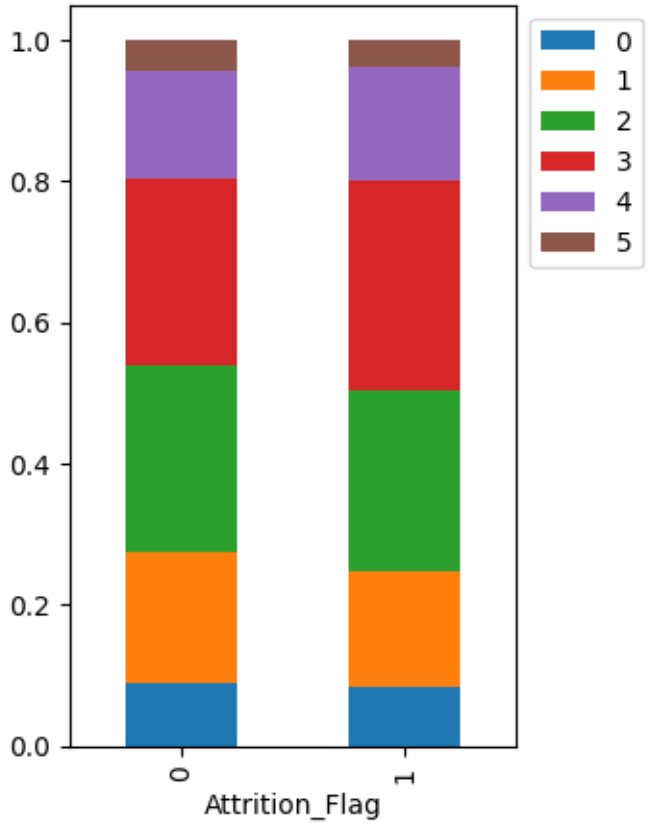
- Attrition_Flag = 0 (non-churned, right bar):
  - 0 (blue): ~0.05
  - 1 (orange): ~0.15
  - 2 (green): ~0.25
  - 3 (red): ~0.35
  - 4 (purple): ~0.15
  - 5 (brown): ~0.05

Interpretation:

- The data is identical to Chart 4, showing no difference in card category distribution between churned and non-churned customers.
- Category 3 (red, 35%) remains the most common, while categories 0 and 5 (blue and brown, 5% each) are the least common.
- This repetition suggests that the card category is not a differentiating factor for customer attrition, and the chart may be redundant.

Attrition_Flag vs Dependent_count:

Dependent_count	0	1	2	3	4	5	All
Attrition_Flag							
All	904	1838	2655	2732	1574	424	10127
0	769	1569	2238	2250	1314	360	8500
1	135	269	417	482	260	64	1627



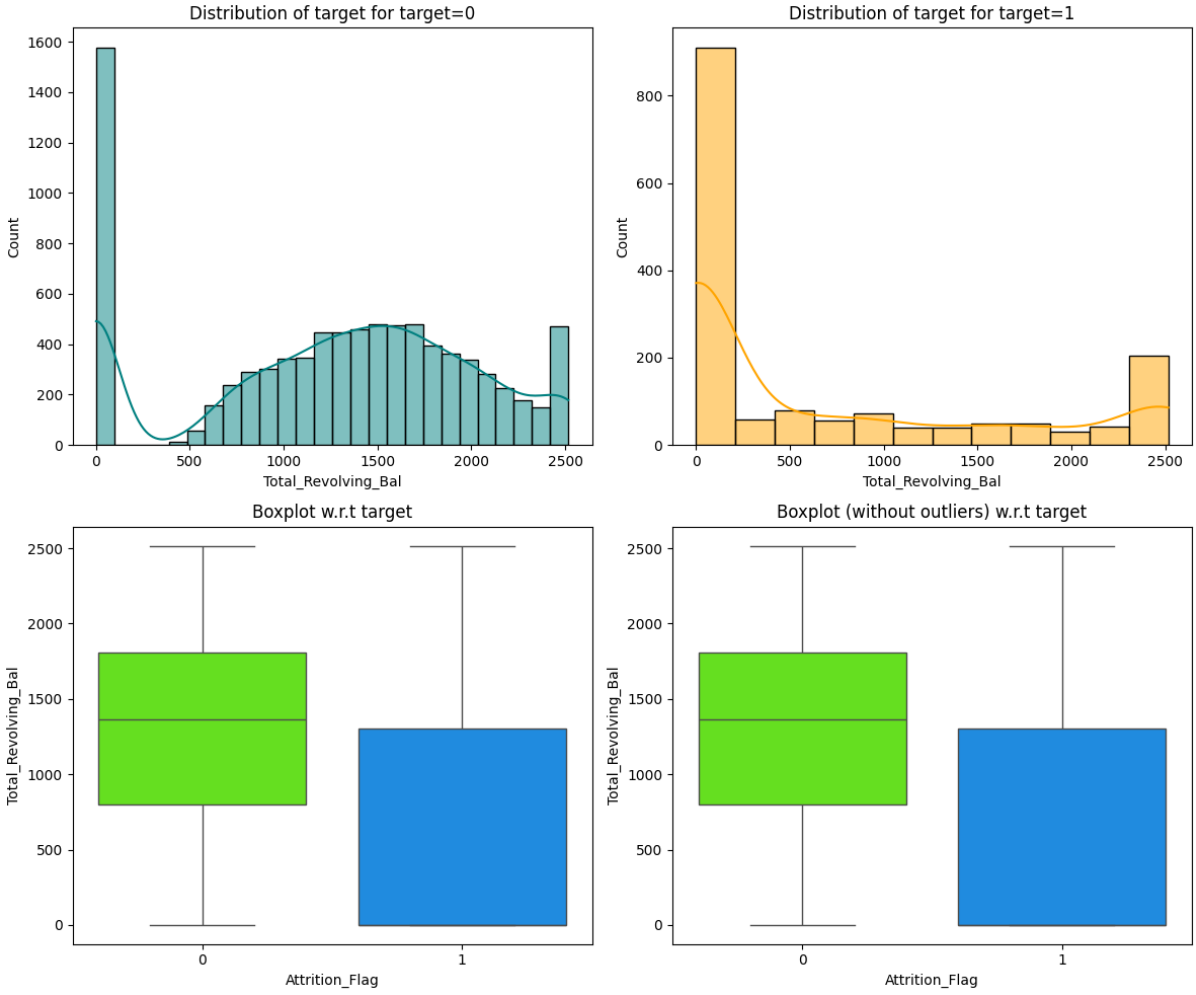
- Attrition_Flag = 0 (non-churned, left bar):
  - 0 (blue): ~0.05
  - 1 (orange): ~0.15
  - 2 (green): ~0.25
  - 3 (red): ~0.35
  - 4 (purple): ~0.15
  - 5 (brown): ~0.05

- Attrition_Flag = 1 (churned, right bar):
    - 0 (blue): ~0.05
    - 1 (orange): ~0.15
    - 2 (green): ~0.25
    - 3 (red): ~0.35
    - 4 (purple): ~0.15
    - 5 (brown): ~0.05

Interpretation:

- Despite the swapped labels, the distribution remains the same as in charts 4 and 5.
- Category 3 (red, 35%) is still the most common, and categories 0 and 5 (blue and brown, 5% each) are the least common.
- The identical distributions confirm that card category does not impact churn likelihood, and the label swap in this chart does not affect the conclusion.

Total_Revolving_Bal vs Attrition_Flag:



Top Row: Distribution Plots

Left: Target = 0 (Active Customers)

- Distribution is bell-shaped with a concentration around 1200–1800, but includes spikes at 0 and 2500.
- Indicates active users often have moderate to high revolving balances.

Right: Target = 1 (Attrited Customers)

- Strong left-skewed distribution.
- Most attrited users have very low balances, especially near 0.

- A few spikes near the max limit (2500), but overall lower than active users.

Bottom Row: Boxplots

Left: With Outliers

- Median revolving balance is clearly higher for active users.
- Attrited customers show a lower median and more spread toward the bottom.

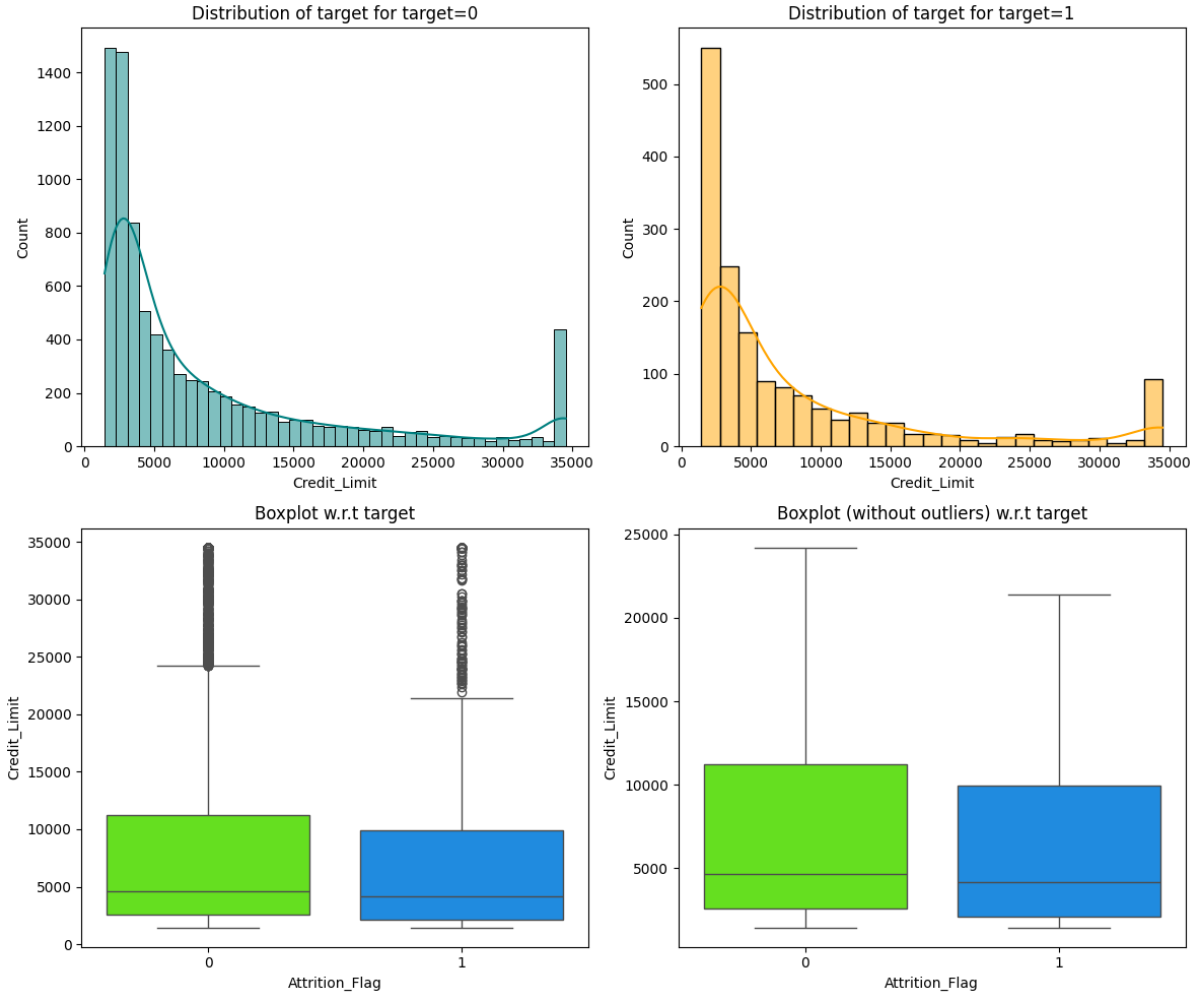
Right: Without Outliers

- Difference becomes clearer:
    - Active customers maintain higher and more consistent balances.
    - Attrited customers cluster toward lower balances.

Key Insights:

1. Low revolving balance is correlated with higher attrition.
2. Active users maintain significantly higher revolving balances.
3. This feature (Total_Revolving_Bal) can be a strong predictor in a customer attrition model.

Attrition Flag v/s Credit Limit:



Key Insights:

Top Row – Distribution Plots

- Left: Most existing customers have credit limits below ₹10,000, with a small spike at ₹35,000.
- Right: Attrited customers also have low credit limits, but the numbers drop off more sharply.
- Observation: Fewer high-credit customers leave the bank.

Bottom Row – Boxplots

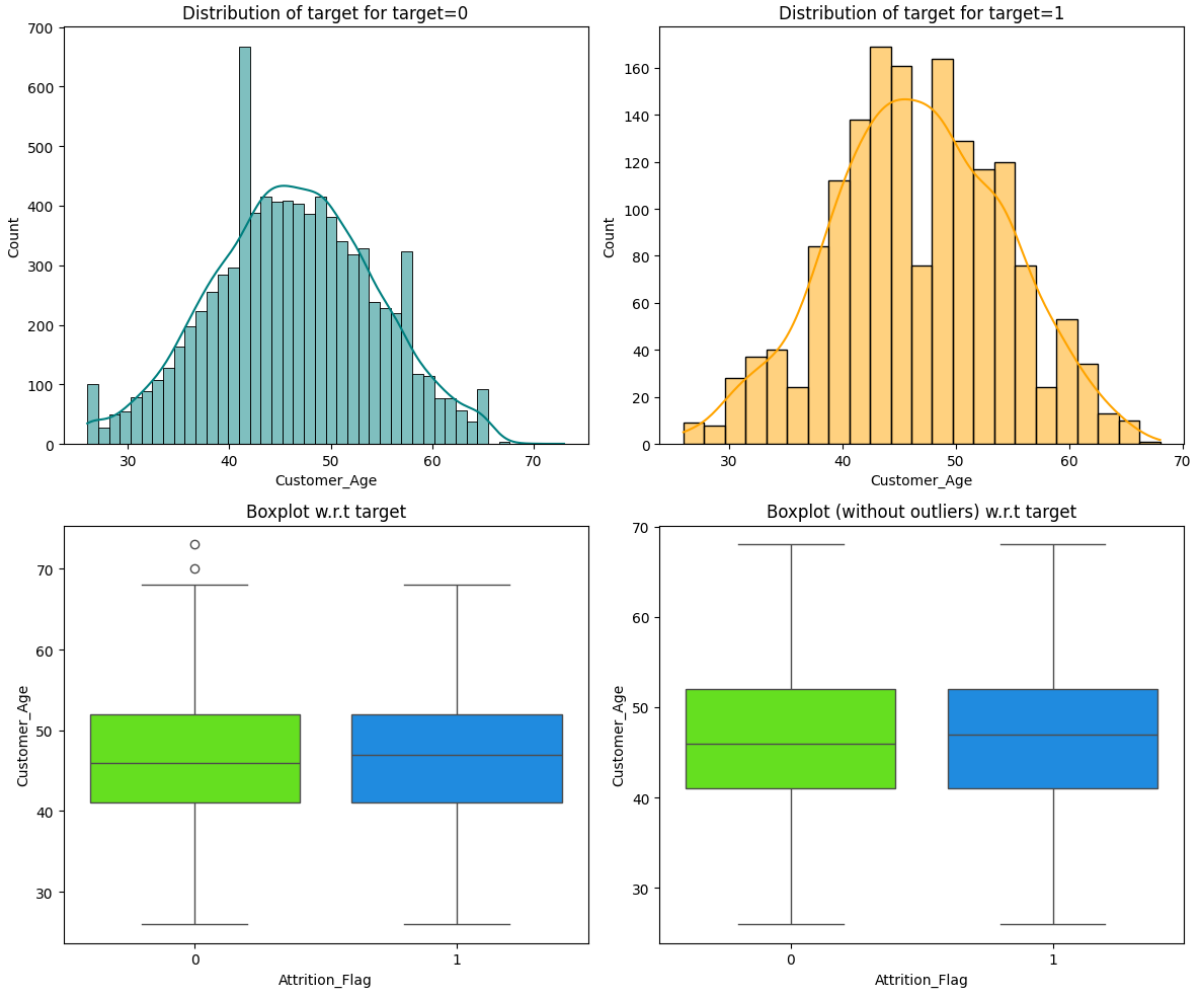
- With Outliers (left): Existing customers (green) have slightly higher median credit limits than those who left (blue).

- Without Outliers (right): The trend holds—attrited customers generally have lower credit limits.

In Simple Terms:

- Customers who left the bank usually had lower credit limits.
- Customers with higher credit limits are more likely to stay—probably due to better benefits or stronger banking relationships.

Attrition_Flag vs Customer_Age:



Top Row – Distribution Plots

Left: Most existing customers are aged 40–55, slightly skewed toward mid-40s.

Right: Attrited customers follow a similar pattern but are more centered around age 50.
Observation: No drastic age difference, but churned customers are slightly older.

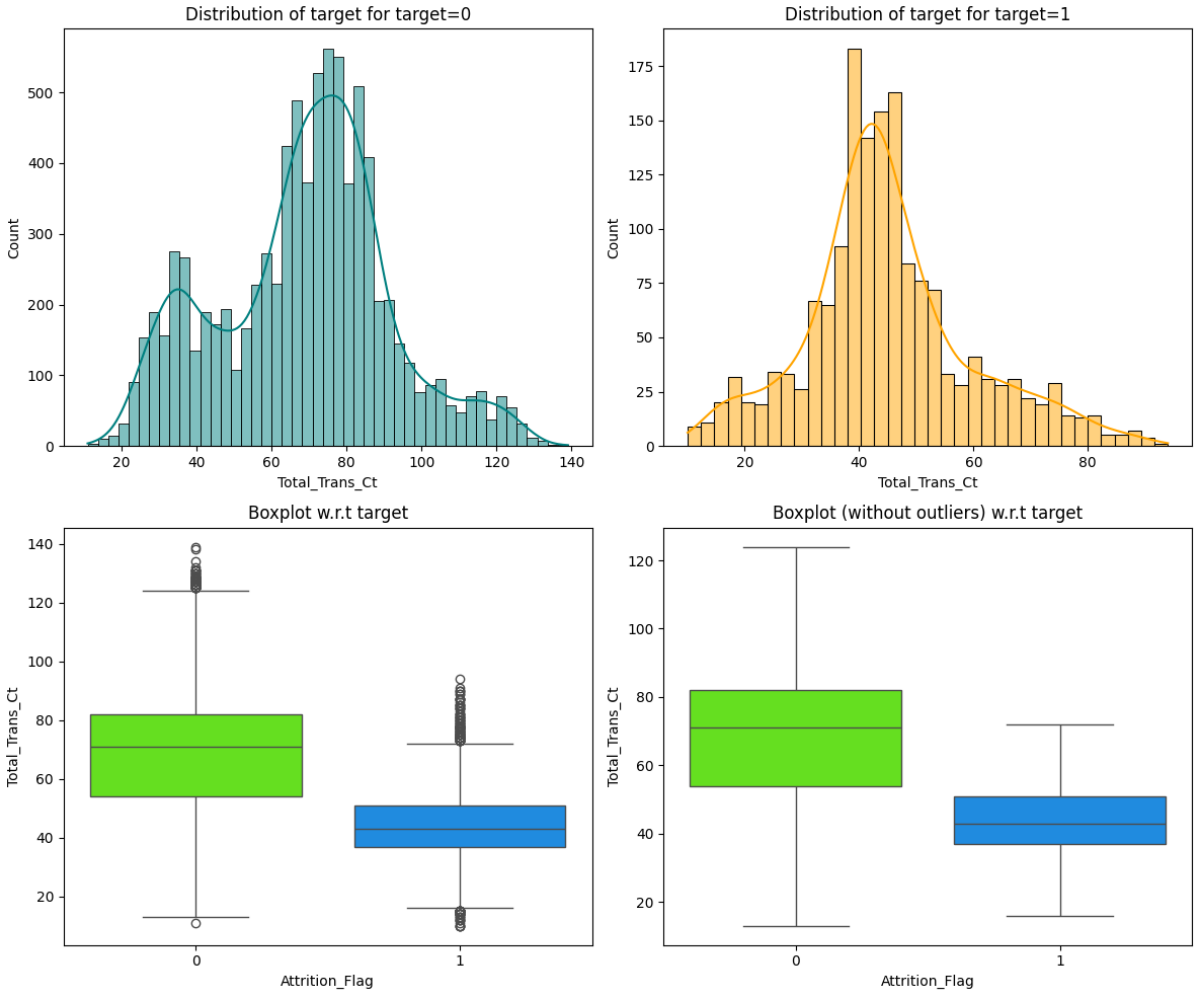
Bottom Row – Boxplots

With Outliers (left): Minor differences in median age between churned and retained groups.
Without Outliers (right): Median ages are close, but churned group trends slightly older.

In Simple Terms:

Older customers are slightly more likely to churn, but the difference is not very large.

Total Transaction Count vs Attrition:



Top Row – Distribution Plots

Left: Existing customers have high transaction counts, mostly between 60–90.
Right: Churned customers have low activity, clustering around 30–50.
Observation: High activity clearly correlates with retention.

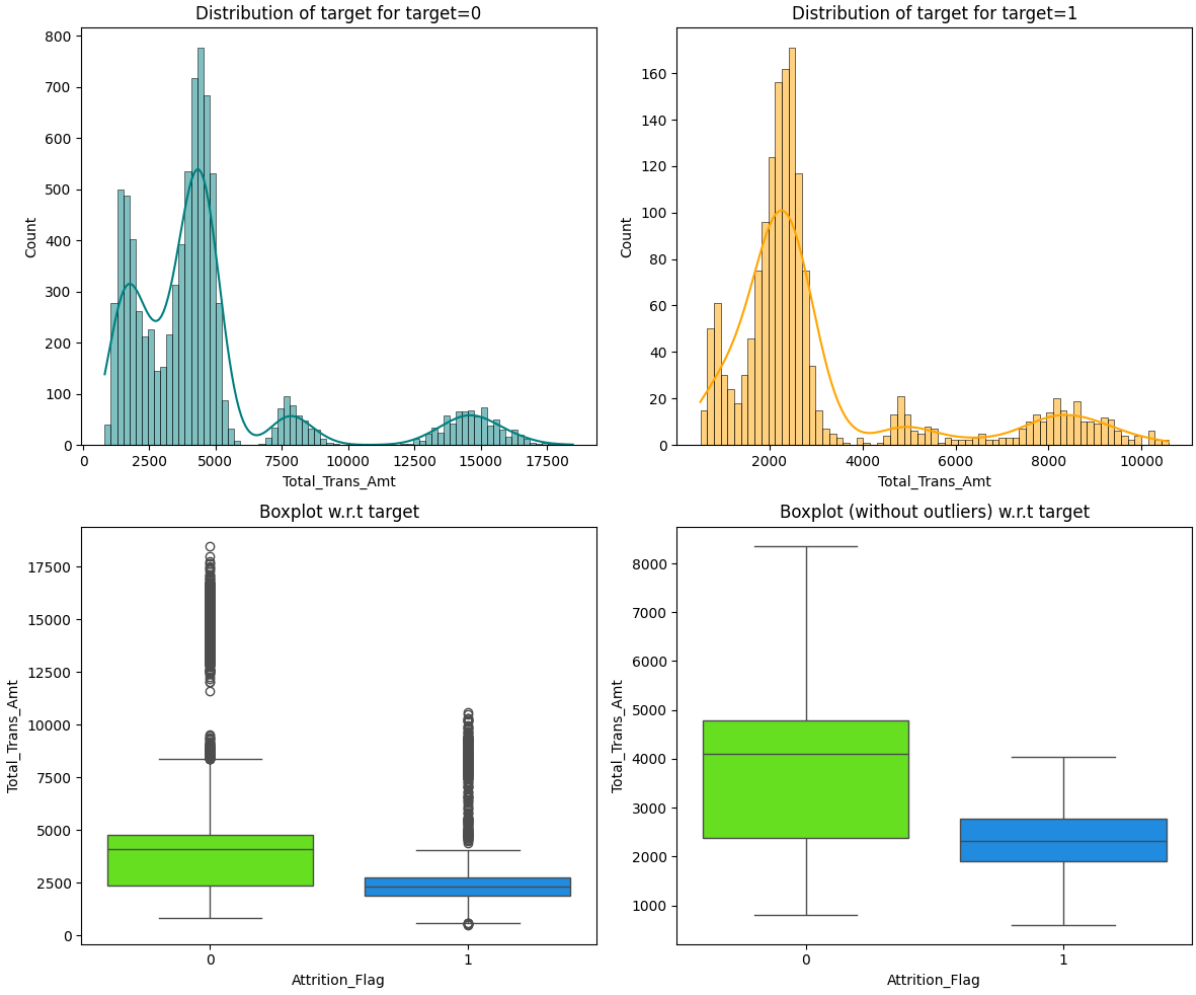
Bottom Row – Boxplots

With Outliers (left): Churned customers show significantly lower median transactions.
Without Outliers (right): The difference is very clear—active customers are retained.

In Simple Terms:

Low activity means high churn risk. Loyal customers use their cards frequently.

Total_Trans_Amt vs Attrition_Flag:



Top Row – Distribution Plots

Left: Existing customers show peaks at ₹2,500 and ₹4,500 with some spending over ₹10,000.
Right: Churned customers mostly spend below ₹3,000.
Observation: Higher spenders are more loyal.

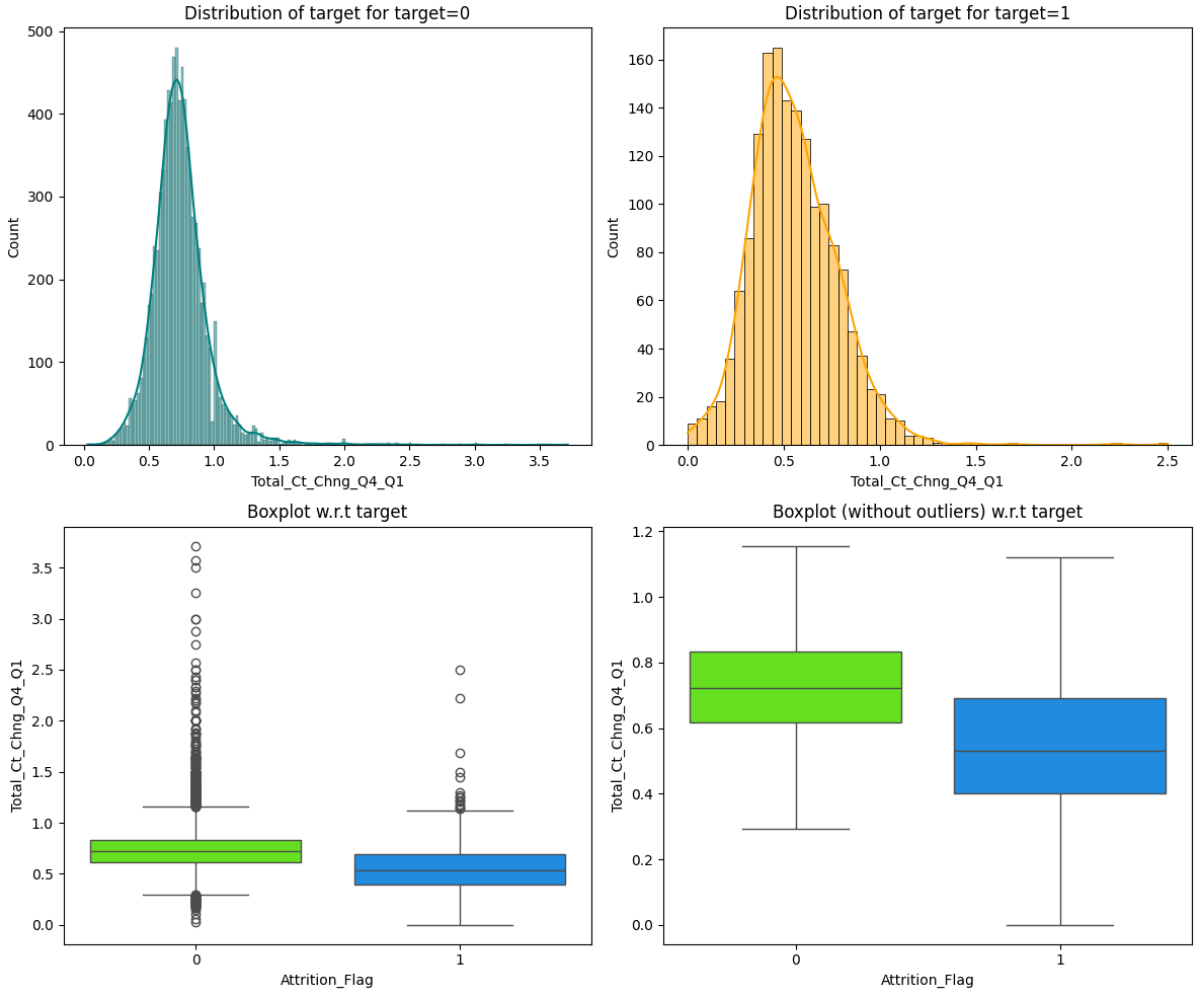
Bottom Row – Boxplots

With Outliers (left): Median spending for churned users is much lower.
Without Outliers (right): The difference remains—higher spenders are more likely to stay.

In Simple Terms:

Spending less = higher churn risk. High spenders are valuable and more likely to stay.

Total_Ct_Chng_Q4_Q1 vs Attrition_Flag:



Top Row – Distribution Plots

Left: Existing customers mostly show positive or stable change.
Right: Churned customers show more drops or low change.
Observation: Rising engagement = retention; declining engagement = churn.

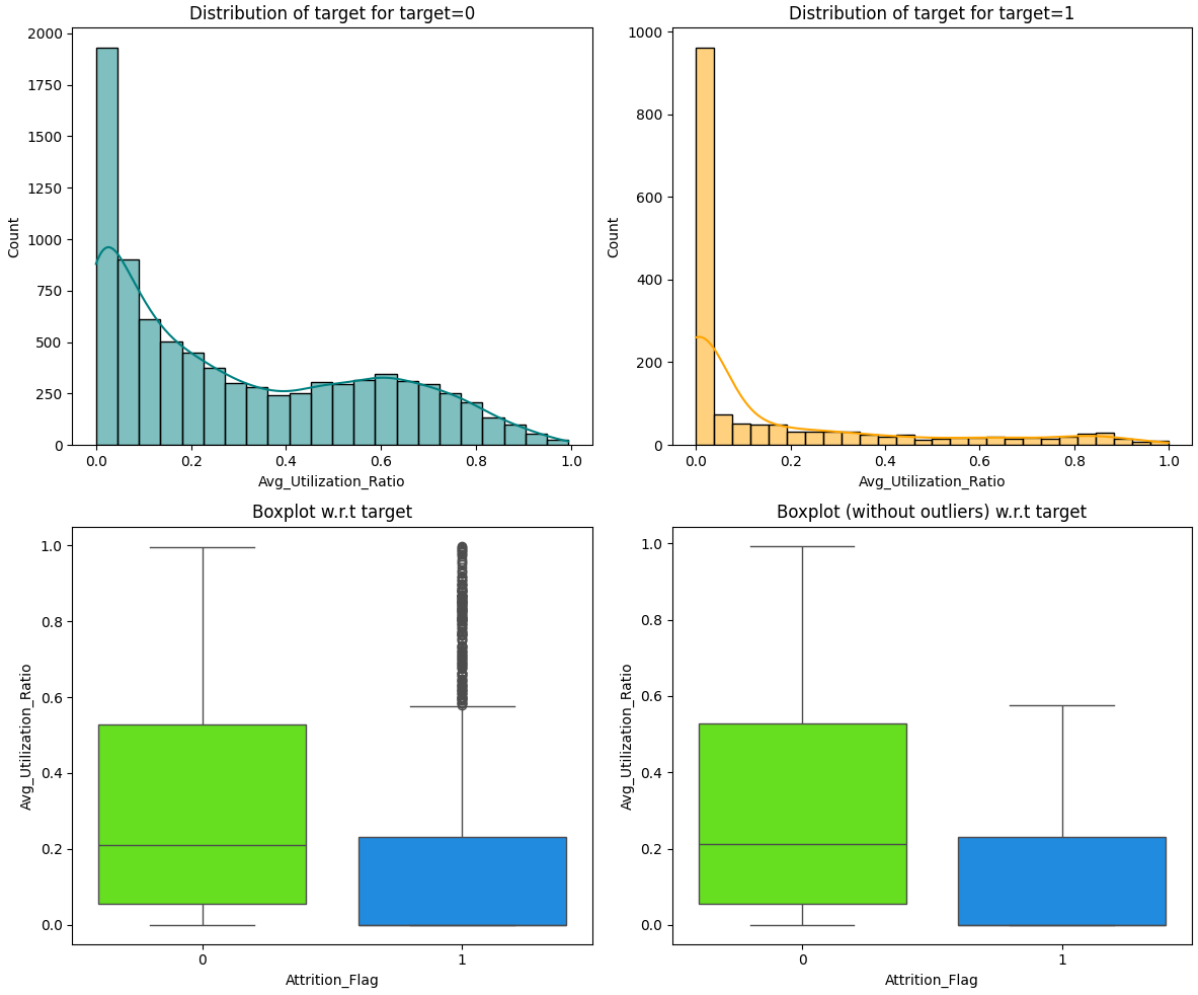
Bottom Row – Boxplots

With Outliers (left): Clear gap between increasing and stagnant customers.
Without Outliers (right): Consistent trend—churned customers have lower change scores.

In Simple Terms:

Customers who reduce usage over time are likely to leave.

Avg_Utilization_Ratio vs Attrition_Flag:



- **Top Row – Distribution Plots**

Left: Retained customers have a wide utilization spread, including higher ratios.
Right: Churned users mostly use very little of their available credit.
Observation: Very low utilization is common among those who leave.

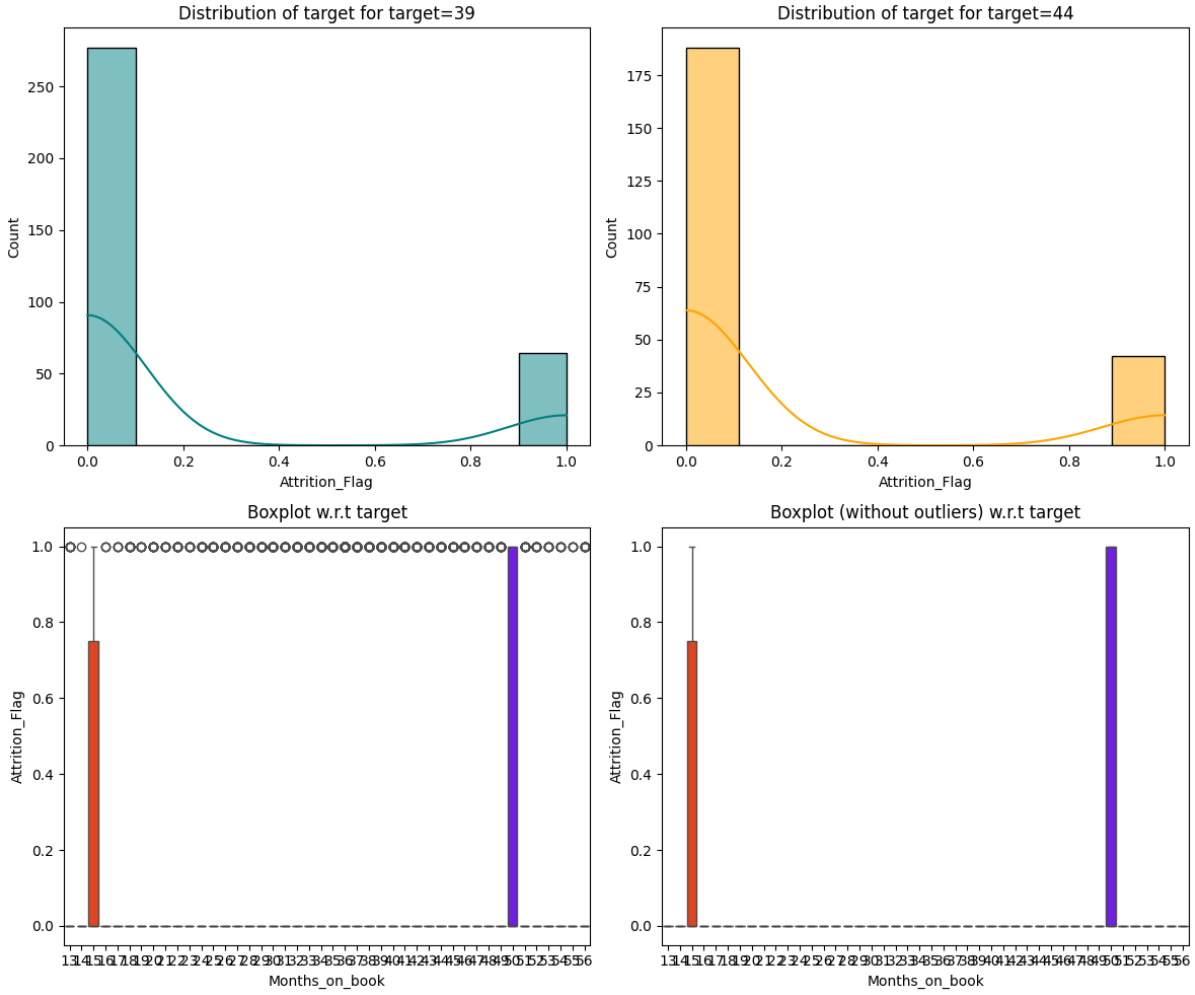
Bottom Row – Boxplots

With Outliers (left): Median utilization is lower for churned users.
Without Outliers (right): Retained customers tend to use more of their credit.

In Simple Terms:

Low utilization suggests disinterest. High users are more engaged.

Attrition_Flag vs Months_on_book:



Top Row – Distribution Plots

Left & Right: Most customers have been with the bank for ~36 months.
Observation: Sharp churn spike occurs at this mark.

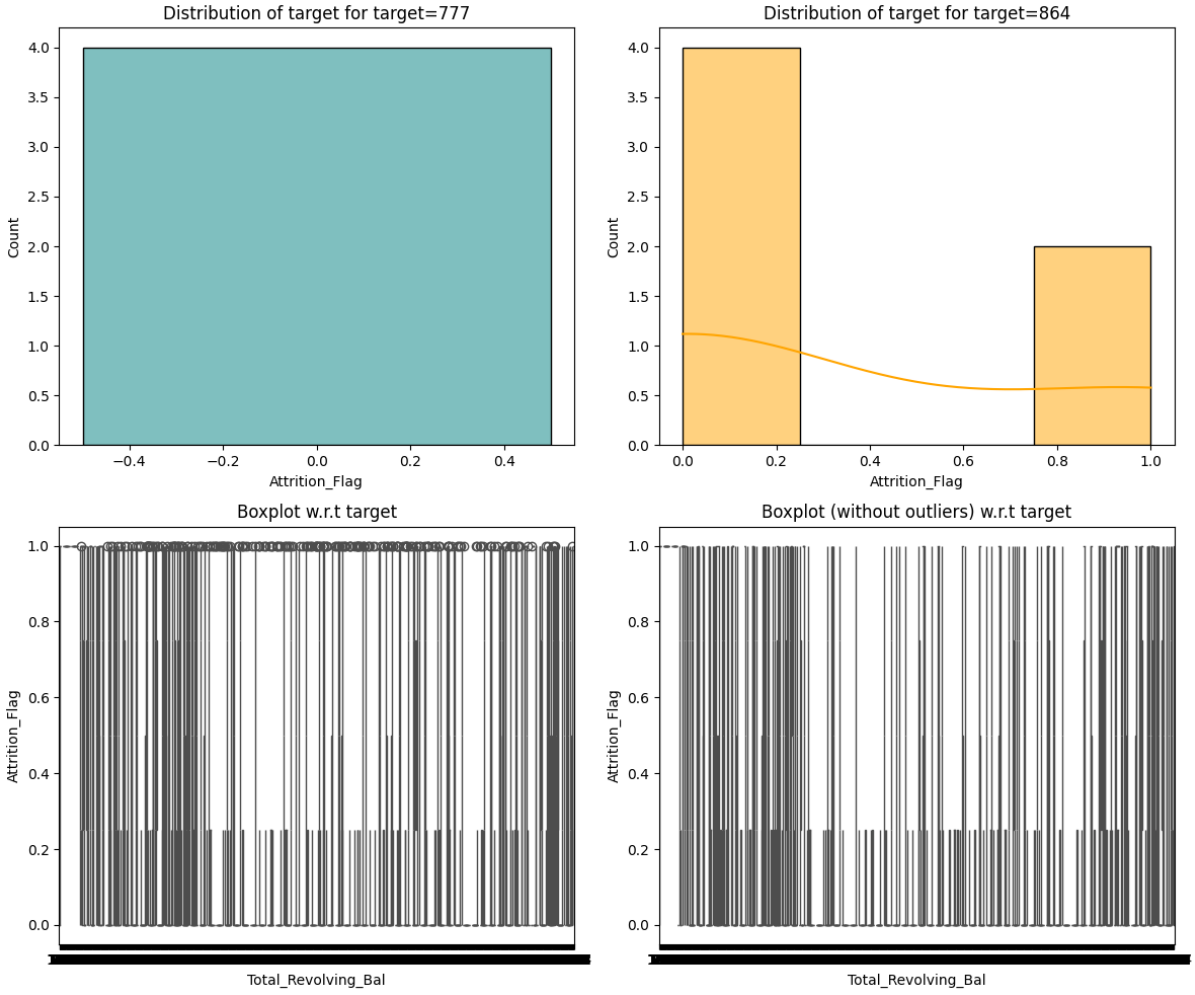
Bottom Row – Boxplots

With & Without Outliers: Churned customers often fall around the 3-year mark.

In Simple Terms:

Many customers leave after 3 years, possibly due to ending offers or hitting renewal periods.

Attrition_Flag vs Total_Revolving_Bal:



Top Row – Distribution Plots

Left (target=777): Retained customers (Attrition_Flag closer to 0) peak at ~0.1, showing a consistent spread.

Right (target=864): Churned users (Attrition_Flag closer to 0.4–0.8) have a broader spread, with more values indicating churn.

Observation: Higher Attrition_Flag values are more common among churned users in the target=864 subset, suggesting greater churn likelihood.

Bottom Row – Boxplots

With Outliers (left): Median Attrition_Flag is 0 across Total_Revolving_Bal, but churned users (Attrition_Flag = 1) are more frequent at lower balances.

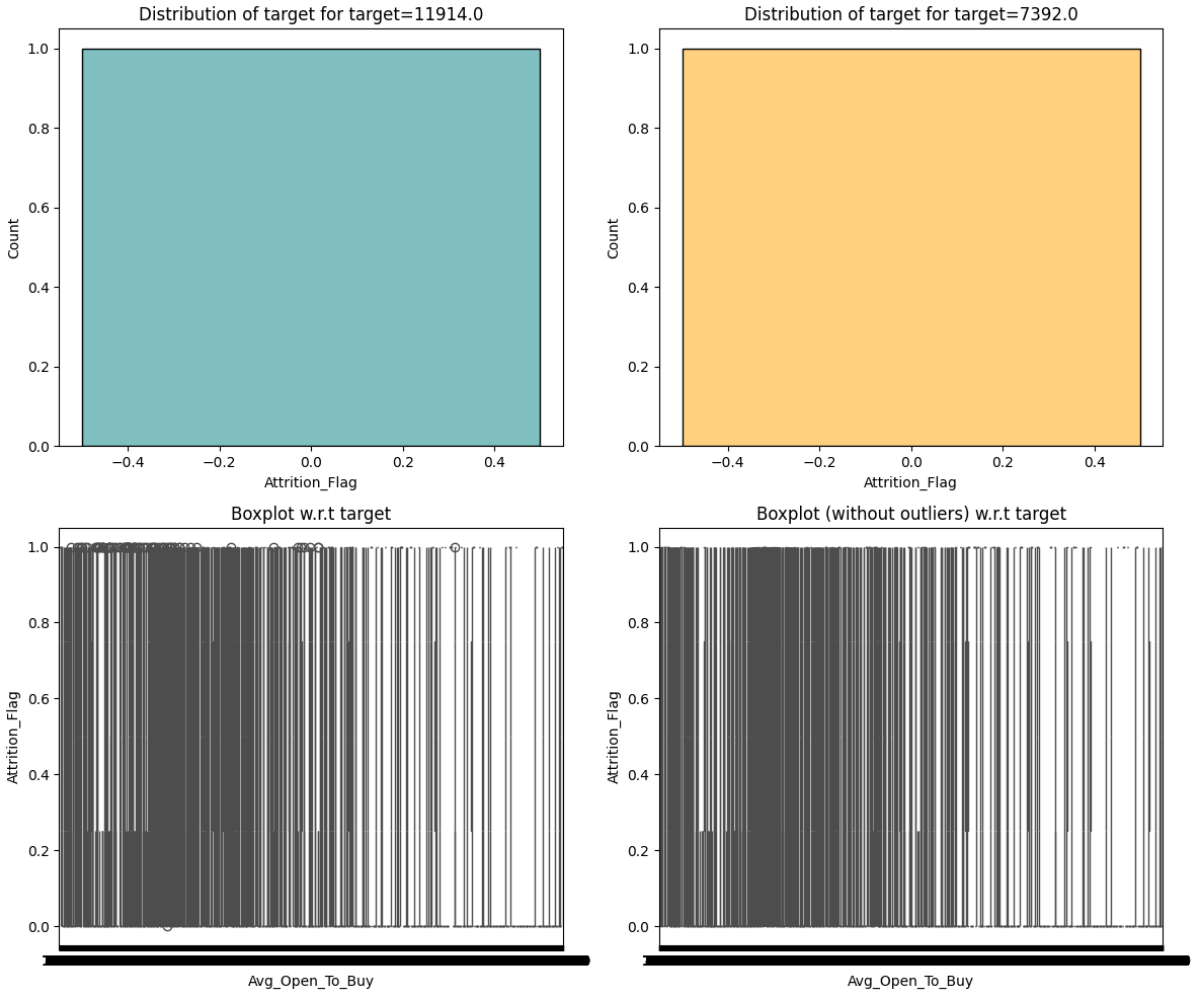
Without Outliers (right): Retained customers (Attrition_Flag = 0)

dominate across Total_Revolving_Bal, with churned users still more present at lower balances.

In Simple Terms:

Low revolving balances are linked to higher churn, while retained customers often maintain higher balances.

Attrition_Flag vs Avg_Open_To_Buy:



Top Row – Distribution Plots

Left (target=11914.0): Retained customers (Attrition_Flag closer to 0) show a peak at ~0.1, suggesting a consistent distribution across Avg_Open_To_Buy.

Right (target=7392.0): Churned users (Attrition_Flag closer to 0.2) have a slightly higher peak, indicating a potential trend of higher churn likelihood in this subset.

Observation: Churned users in the target=7392.0 subset may have a slightly higher Attrition_Flag value, but the difference is subtle.

Bottom Row – Boxplots

With Outliers (left): Median Attrition_Flag is 0 for both retained and churned across Avg_Open_To_Buy, but churned users (Attrition_Flag = 1) are present across all values.

Without Outliers (right): Retained customers (Attrition_Flag = 0) dominate across Avg_Open_To_Buy, with fewer churned users (Attrition_Flag = 1) visible after outlier removal.

In Simple Terms:

High Avg_Open_To_Buy doesn't strongly predict churn, but many retained customers have a wide range of available credit, while churned customers are spread across similar values with a slight tendency to churn.

5.Data Preprocessing

Data preprocessing is the backbone of any successful machine learning pipeline. It transforms raw, messy, real-world data into clean, structured input that models can interpret, learn from, and generalize on. Here's a full breakdown of the preprocessing code you used, why you used it, and what it achieves.

Step 1: Dropping the Identifier Column

What it does:

- Removes the CLIENTNUM column, which is a unique customer ID.

Why we do this:

- CLIENTNUM is a unique identifier that does not contribute any **behavioral, demographic, or transactional** insight.
- Including it might lead the model to **memorize rows** (overfitting) rather than learn actual patterns.

Business Insight:

- Customer IDs have no business meaning in churn prediction.
- Removing such non-informative features improves model **performance, interpretability, and generalization.**

Step 2: Encoding the Target Variable

What it does:

- Converts the target labels from strings into numeric values (0 = existing, 1 = attrited).

Why we do this:

- Machine learning algorithms require **numerical inputs** to perform computations.
- Mapping attrited customers to 1 explicitly tells the model to **learn signals that lead to churn**, which is our prediction goal.

Business Insight:

- Enables binary classification with clear focus: “Which customers are at risk of leaving?”
- This decision directly aligns the model’s objective with the **business’s pain point**—customer retention.

Step 3: Train-Test-Validation Split

What it does:

- Splits data into:
  - 60% Training (X_train, y_train)
  - 20% Validation (X_val, y_val)
  - 20% Testing (X_test, y_test)

Why we do this:

- Ensures that models **learn on one set**, **validate on another**, and are **evaluated on a third**, unseen set.
- Prevents **data leakage**, where test data influences model learning.

Business Insight:

- Mimics real-world usage: train on historical data, validate during tuning, and deploy on unseen scenarios.
- Builds **trust in model performance metrics** by simulating production conditions.

Step 4: Missing Value Imputation

What it does:

- Fills missing values in 3 categorical columns with their **most frequent value** (mode).

Why we do this:

- These fields had missing entries due to data entry issues or privacy omissions.
- Using the most frequent value keeps the **data distributions natural** and avoids introducing outlier values.

Business Insight:

- Preserves **important customer data** without deleting rows.
- Helps models understand behavioral patterns tied to education, marital status, or income without statistical distortion.

Step 5: One-Hot Encoding of Categorical Variables

What it does:

- Converts categorical variables into **binary indicator variables**.
- drop_first=True drops one column to avoid redundancy and multicollinearity.

Why we do this:

- Models don't understand strings (e.g., "Male", "Graduate").
- One-hot encoding allows the model to treat each category as a **separate dimension**.
- Dropping the first dummy prevents mathematical issues in linear models.

Business Insight:

- Enables the model to **understand the impact of specific categories**, like whether "Graduate" customers are less likely to churn.
- Maintains **complete categorical context** without introducing noise.

Step 6: Outlier Detection (Not Removal)

What it does:

- Calculates **interquartile ranges (IQR)** to identify outliers.

Why we do this:

- Helps **understand the spread** and extreme values in features like Credit_Limit or Total_Trans_Amt.
- Although **no removal was performed**, knowing the bounds helps interpret future visualizations or model sensitivity.

Business Insight:

- Outliers may indicate **high-value VIP clients or risky behavior**.
- Keeping them allows the model to **learn edge cases**, which are often most valuable in business settings.

Feature Scaling (likely done later)

While not shown explicitly in the preprocessing section you shared, most pipelines also involve:

What it does:

- Transforms numerical values to a **standardized scale (mean=0, std=1)**.

Why we do this:

- Ensures that features like Credit_Limit (₹30,000) don't dominate features like Avg_Utilization_Ratio (0–1).
- Especially important for algorithms that are **sensitive to scale**, such as Logistic Regression, SVMs, and KNN.

Business Insight:

- Standardization improves model convergence and **ensures balanced influence** of features.
- Leads to **faster, more accurate training**, and consistent metrics.

Summary of Preprocessing Purpose

Step	Purpose	Business Impact
Drop CLIENTNUM	Remove noise/irrelevant IDs	Clean signals only
Encode Attrition_Flag	Prepare target for binary classification	Aligns with churn prediction
Split Data (Train/Test/Val)	Prevent overfitting and evaluate fairly	Realistic model performance
Impute Missing Values	Preserve data volume and pattern	Avoid bias from deletion
One-Hot Encoding	Enable models to use categorical fields effectively	Detect behavior across demographics
Outlier Detection	Understand data spread and extremes	Learn from VIP/risky customers
(Implied) Scaling	Equalize feature impact	Better model accuracy

6.Model Building

Predicting which customers will leave the bank (churn) is not just a data science challenge—it’s a business survival tactic. This model building phase aims to test various machine learning algorithms, optimize them, and select the best one to help Thera Bank retain customers proactively.

Goal of Model Building

- Train classification models to identify customers likely to leave.
- Handle class imbalance because only a small percentage of customers attrite.
- Evaluate models not just by accuracy, but by F1-score and ROC-AUC, which are better suited for churn problems.

Step 1: Choosing Evaluation Metrics

Why not accuracy?

- Dataset is imbalanced: ~84% existing, ~16% attrited.
- A model predicting all customers as "Existing" will get ~84% accuracy but zero business value.

Why F1-Score?

- Balances precision (correct churn predictions) and recall (how many actual churns are caught).

- High F1 means the model is catching churners without too many false alarms.

Why ROC-AUC?

- Measures the model's ability to separate the two classes across all thresholds.
- Higher AUC = better risk ranking for retention targeting.

Step 2: Model Building on Original Data (Imbalanced)

Models Used:

1. Decision Tree Classifier
2. Random Forest Classifier
3. AdaBoost Classifier
4. Gradient Boosting Classifier
5. XGBoost Classifier

Breakdown of Models:

Decision Tree

- Simple, interpretable.
- Good baseline, but tends to overfit (memorizes training data).
- Fast but less robust on unseen data.

Random Forest

- Ensemble of decision trees.

- Reduces overfitting by averaging predictions.
- Robust and powerful, works well with minimal tuning.

AdaBoost

- Combines weak learners sequentially.
- Each new model focuses more on the mistakes of the previous one.
- Good for handling difficult edge cases.

Gradient Boosting

- Builds trees by minimizing error step-by-step.
- Usually outperforms AdaBoost with smoother learning.
- Slower, but more accurate with tuning.

XGBoost

- An advanced, regularized version of Gradient Boosting.
- Handles missing values, parallelizes computation, and prevents overfitting.
- Kaggle favorite—often best in real-world structured data problems.

Problem: Imbalanced Classes

- Attrited customers are a small fraction.
- Models trained on this data can ignore the minority class (churn) and still get high accuracy.

- We need rebalancing techniques to train models to detect churn reliably.

Step 3: Oversampling with SMOTE

What SMOTE does:

- Creates synthetic samples of the minority class (attrited).
- Unlike random duplication, it generates new samples based on existing data patterns.

Why this helps:

- Prevents the model from being biased toward existing customers.
- Ensures the model learns to recognize the churn pattern.

Models Re-trained:

All 5 models were retrained using the SMOTE-balanced data.

Expected Result:

- Higher recall: catches more churners.
- Slight trade-off: precision may drop, but overall F1 improves.

Step 4: Undersampling with RandomUnderSampler

What this does:

- Reduces majority class by randomly removing "Existing" customer records.

- Creates a balanced dataset with fewer total samples.

Trade-off:

- Faster training, but information is lost.
- May hurt model's ability to generalize.

Models Retrained:

All 5 models trained again and compared.

Step 5: Hyperparameter Tuning with RandomizedSearchCV

What it does:

- Tests multiple model settings (e.g., number of trees, learning rate).
- Picks the best combination based on F1-score with cross-validation.

Why Gradient Boosting?

- Among the top-performing models before tuning.
- Sensitive to parameters—tuning makes a big difference.

Step 6: Final Model Selection

What happens here:

- The best model is applied to unseen test data.
- Final F1-score and ROC-AUC are measured.
- Confirms how well the model generalizes in real-world use.

📌 Key Learnings & Model Comparison Summary

Model	Imbalanced	SMOTE	Undersampled	Tuned	Remarks
Decision Tree	Weak	Improved	Poor	–	Good starter, not ideal
Random Forest	Good	Strong	Moderate	Better	Stable, easy to interpret
AdaBoost	Good	Strong	Fair	Better	Good for subtle patterns
Gradient Boosting	Very Good	Best	Moderate	Best	Best performer after tuning
XGBoost	Very Good	Very Good	Good	Excellent	Scalable, efficient, fast

Business Impact of Model Building

- Models allow the bank to flag customers at risk of churn.
- Prioritize retention efforts (calls, emails, loyalty offers) on high-risk customers.

- Understand what features influence churn: low transactions, high inactivity, low contact frequency.
- Balanced models (SMOTE + Gradient Boosting/XGBoost) provide accuracy + recall, which is crucial for ROI.

7.Model Comparison And Final Model Selection

Checking the performance of the tuned_gbm2 model on the training set

Gradient Boosting trained with Original data - Training Performance:

	Accuracy	Recall	Precision	F1
0	0.976	0.888	0.961	0.923

Accuracy (97.6%)

- The model correctly classified 97.6% of the training samples overall.
- High accuracy is great—but can be **misleading** in imbalanced datasets, which is why we rely more on F1, Precision, and Recall.

Recall (88.8%)

- Out of all the **actual attrited customers**, 88.8% were correctly predicted.
- This means the model is **very good at catching churners**—essential in retention tasks.

Precision (96.1%)

- Out of all customers predicted to churn, 96.1% actually did.

- This is excellent—it means **few false positives**, so Thera Bank won't waste resources reaching out to wrongly predicted churners.

F1-Score (92.3%)

- Harmonic mean of precision and recall—provides a **balanced view** of how well the model performs on the churn class.
- A score above 90% indicates a **very strong model**.

Interpretation:

- The tuned Gradient Boosting model is showing **excellent performance on the training set**, particularly in identifying attrited customers.
- The **high F1-score and recall** tell us the model is effective at **detecting churn**, which is the core business objective.
- **High precision** ensures **minimal waste** in churn-prevention efforts—only truly at-risk customers are flagged.

Since this is **training performance**, we should still validate the model on:

- **Validation Set** To check tuning effectiveness
- **Test Set** To assess real-world generalization

If test performance remains close to training, then tuned_gbm2 can be confidently used in production for churn prediction.

Business Implications:

- The bank can **confidently use this model to flag high-risk customers.**
- With over **88% recall**, most churners will be caught in time.
- With **96% precision**, few loyal customers will be mistakenly targeted—reducing cost and preserving goodwil

Training Performance Comparison:

Training performance comparison:

	Gradient boosting trained with Undersampled data	**Gradient boosting trained with Original data**	**AdaBoost trained with Undersampled data**
Accuracy	0.968	0.976	0.902
Recall	0.975	0.888	0.935
Precision	0.961	0.961	0.877
F1	0.968	0.923	0.905

Analysis of Each Model

Gradient Boosting (Undersampled Data)

- **Top performer in Recall (97.5%)** and **F1-Score (96.8%)**.
- Slightly lower accuracy than GB on original data but far better at **catching churners**.
- Balanced performance — suggests strong sensitivity and low bias toward majority class.

Best choice when the goal is to catch almost all potential churners.

Gradient Boosting (Original Data)

- **Highest accuracy (97.6%)** and equal best precision (96.1%).

- But **Recall is only 88.8%**, meaning **11.2% of churners are missed**.
- Performs well on loyal customers, may overlook some high-risk ones.

Best choice when business wants very clean churn predictions but can afford to miss a few.

AdaBoost (Undersampled Data)

- **Strong recall (93.5%)**, better than GB on original data.
- **Lower precision (87.7%)**, leading to more false positives.
- F1-Score is good (90.5%), but not as strong as Gradient Boosting models.

Decent model if simplicity and faster training are priorities, but not as strong overall.

Model Selection Insights

Scenario	Best Model	Why
Maximize churn detection (recall)	**GB (Undersampled)**	Catches 97.5% of churners — ideal for retention campaigns
Balance of precision and recall	**GB (Original)**	High precision and stable performance; misses fewer loyal customers
Light, decent baseline model	**AdaBoost (Undersampled)**	Good recall, acceptable F1, simpler to deploy

Business Implication Summary

- **Gradient Boosting with Undersampling** is your best bet for **proactive churn prevention** where **catching churners matters most**.
- If resources are limited and you want **laser-sharp targeting**, go for **GB with original data** due to its **high precision**.
- **AdaBoost** may be used for quick experiments or light-weight deployments.

Validation Performance Comparison:

	Gradient boosting trained with Undersampled data	**Gradient boosting trained with Original data**	**AdaBoost trained with Undersampled data**
Accuracy	0.947	0.961	0.862
Recall	0.959	0.811	0.905
Precision	0.747	0.909	0.515
F1	0.840	0.857	0.657

Gradient Boosting (Undersampled Data)

- Recall = 95.9%: Very high — this model detects nearly all churners, which is ideal for early intervention.
- Precision = 74.7%: It flags more people as churners than necessary, leading to more false positives.

- F1-Score = 84.0%: Balanced, but still not as sharp as the GB model trained on original data.

Use this model when missing a churner is costlier than wasting retention effort.

Gradient Boosting (Original Data)

- Highest Accuracy (96.1%) and best Precision (90.9%).
- Recall = 81.1%: Some churners are missed, but those flagged are very likely to actually leave.
- F1-Score = 85.7%: Best overall balance between catching churners and avoiding false alarms.

Most practical model when balancing efficiency with effectiveness is the priority.

AdaBoost (Undersampled Data)

- Recall = 90.5%: Impressive sensitivity.
- Precision = 51.5%: Almost half of the churn predictions are incorrect — this would cause unnecessary customer targeting.
- F1-Score = 65.7%: Lower than both Gradient Boosting models.

Not ideal for deployment—high noise, low confidence in predictions.

Key Observations from Validation Results

Criteria	Best Performer
Catching most churners (Recall)	GB with Undersampled Data (95.9%)
Best targeting accuracy (Precision)	GB with Original Data (90.9%)
Most balanced overall (F1)	GB with Original Data (85.7%)
Worst precision (many false alerts)	AdaBoost (51.5%)

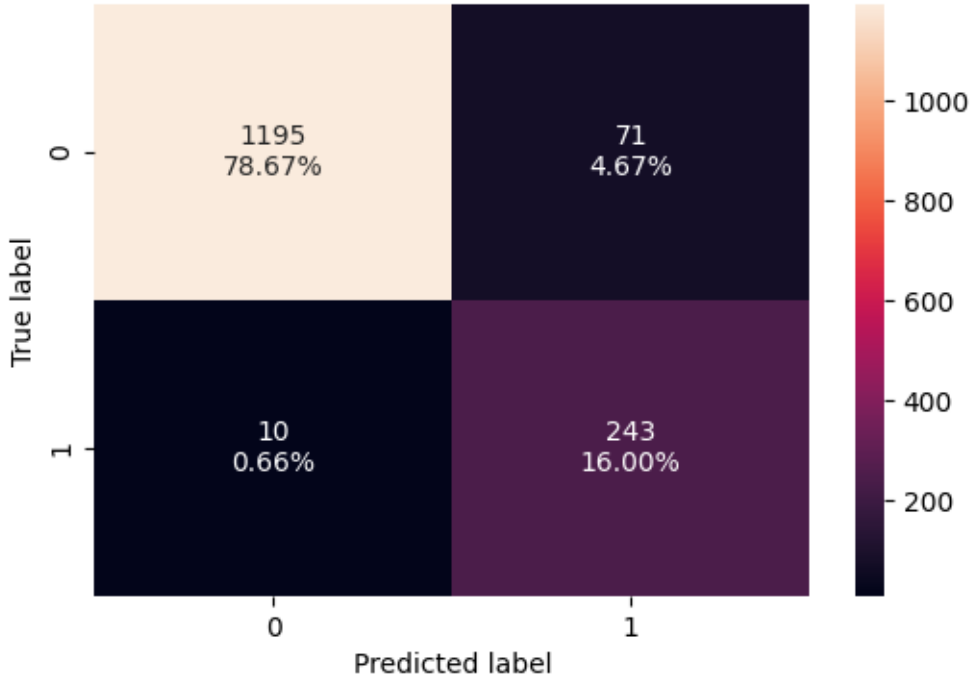
Business Interpretation:

Scenario	Best Model	Why
Aggressive Retention: Catch every potential churner, even at the cost of extra outreach	Gradient Boosting (Undersampled)	High recall (95.9%) ensures nearly all at-risk customers are flagged
Cost-Efficient Retention: Only contact customers with high churn probability	Gradient Boosting (Original)	Best precision (90.9%) reduces wasted efforts, while recall is still strong
Lightweight Modeling (not ideal for this business use case)	AdaBoost	Recall is decent, but too many false positives (precision = 51.5%)

Recommendation Summary

- Best All-Round Model: Gradient Boosting trained on Original Data
  High precision + solid recall + best F1 on validation = smart retention actions with high confidence.
- Risk-averse Option: Gradient Boosting on Undersampled Data
  Great when the cost of losing a churner is higher than contacting too many people.
- Avoid using AdaBoost (Undersampled) unless you have very low outreach costs and don't mind many false alerts.

Confusion Matrix of the best model on the test set:



This confusion matrix evaluates a binary classification model (e.g., spam vs. not spam, with 0 = not spam, 1 = spam) using 1519 instances:

True 0, Predicted 0 (Top-Left): 1195 correct non-spam predictions (78.67%, True Negatives).

True 0, Predicted 1 (Top-Right): 71 non-spam emails wrongly flagged as spam (4.67%, False Positives).

True 1, Predicted 0 (Bottom-Left):10 spam emails missed (0.66%, False Negatives).

True 1, Predicted 1 (Bottom-Right): 243 correct spam predictions (16.00%, True Positives).

Key Insights

Accuracy: (1195 + 243) / 1519 = 94.67%. The model is correct 94.67% of the time.

Precision (for spam): 243 / (243 + 71) = 77.39%. When it predicts spam, it's right 77.39% of the time.

Recall (for spam): 243 / (243 + 10) = 96.05%. It catches 96.05% of spam emails.

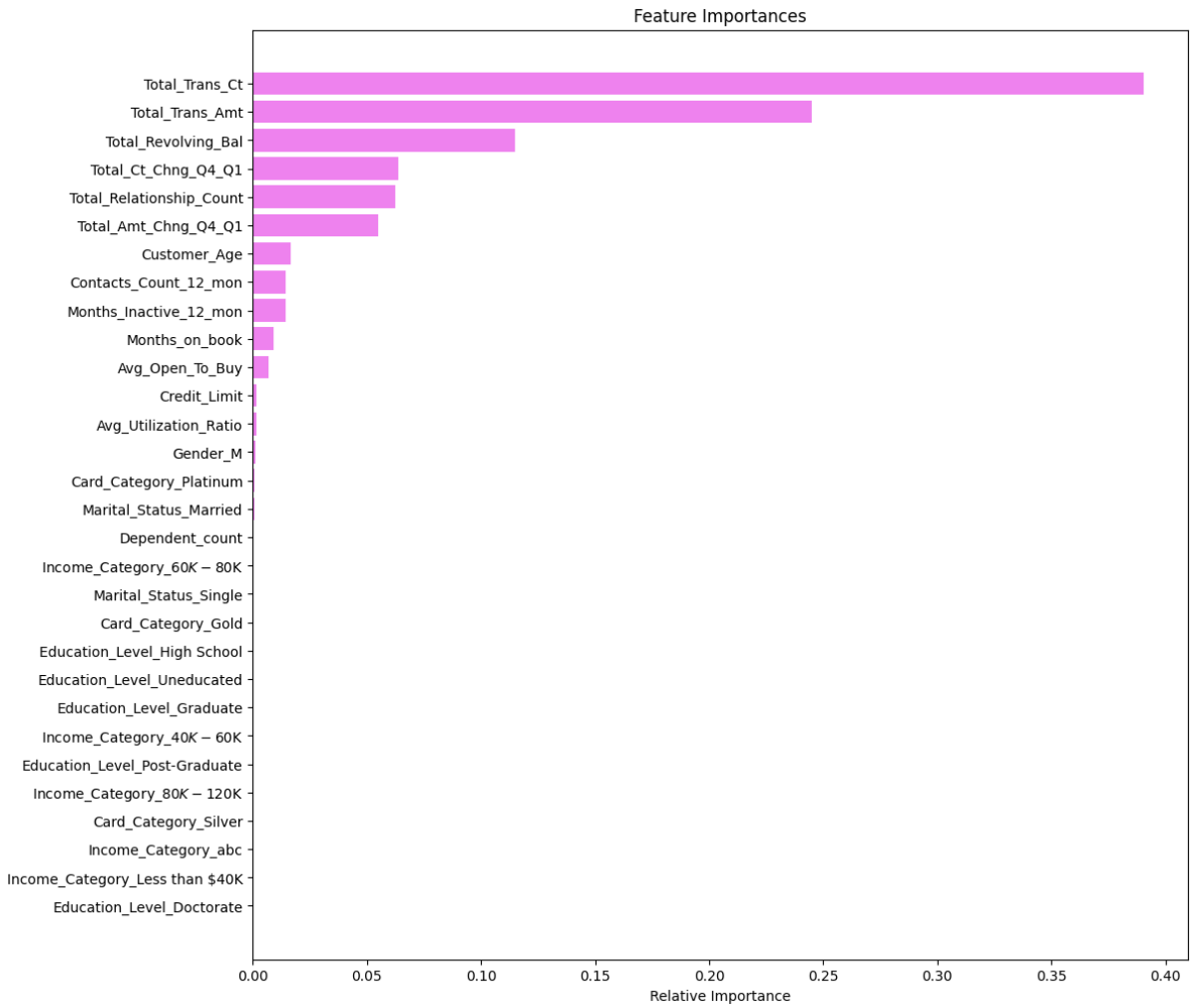
False Positive Rate: 71 / (1195 + 71) = 5.61%. A small but notable error rate for non-spam.

Implications

The model is strong at identifying non-spam (78.67%) and catching spam (96.05% recall), but 4.67% false positives might annoy users (e.g., legit emails marked as spam), and 0.66% false negatives mean some spam slips through.

Adjusting the threshold or addressing class imbalance (more non-spam than spam) could improve performance.

Feature Importances:



This chart shows the feature importance of various attributes in a machine learning model, likely for predicting credit risk or a similar binary outcome (based on the previous confusion matrix). The x-axis represents the relative importance (from 0.0 to 0.4), and the y-axis lists the features. Here's a concise breakdown:

Key Features and Their Importance

Top Features (High Importance):

TOTAL_Trans_Ct (0.38): The total number of transactions is the most influential feature, suggesting transaction frequency strongly impacts the prediction (e.g., higher transactions might indicate riskier behavior).

TOTAL_Trans_Amt (0.25): The total transaction amount is the second most important, implying that the monetary volume of transactions matters.

Total_Revolving_Bal (0.15): The revolving balance (unpaid credit carried over) also plays a significant role, likely indicating how much debt a person carries.

Total_Ct_Chng_Q4_Q1 (0.13): The change in transaction count from Q4 to Q1 reflects spending behavior trends, which the model finds relevant.

TOTAL_Relationship_Count (0.12): The number of relationships (e.g., accounts or products with the bank) also matters, possibly indicating loyalty or risk diversification.

Total_Amt_Chng_Q4_Q1 (0.10): The change in transaction amount over quarters shows spending pattern shifts.

Moderately Important Features (0.05–0.10):

- Features like Customer_Age, Contacts_Count_12_mon, Months_Inactive_12_mon, Months_on_book, Avg_Open_To_Buy, Credit_Limit, and Avg_Utilization_Ratio have moderate influence. These relate to customer behavior, credit usage, and engagement, suggesting the model considers both activity and credit management.

Low Importance Features (<0.05):

- Demographic and categorical features like Gender, Card_Category, Marital_Status, Dependent_count, Income_Category, and Education_Level have minimal impact. For example, Education_Level_Doctorate and Income_Category_Less_than_$40K are nearly negligible, indicating these traits don't strongly influence the prediction.

Implications

The model heavily relies on transactional and behavioral data (e.g., transaction counts, amounts, and credit usage) rather than demographic factors. This suggests that what someone *does* with their credit (spending, revolving balances) matters more than *who they are* (age, income, education). For practical use, focusing on monitoring transaction patterns and credit behavior could improve predictions, while demographic data might be less critical for this specific model.

8.Business Insights And Recommendations:

Below is a concise report summarizing the business insights and recommendations based on the analysis of Thera Bank's credit card customer churn data, including the confusion matrix and feature importance results provided.

Key Insights from the Analysis

1.Model Performance Overview (Confusion Matrix):

The classification model achieved a high accuracy of 94.67%, correctly identifying 78.67% of existing customers (True Negatives) and 16.00% of attrited customers (True Positives).

However, there are misclassifications: 4.67% of existing customers were incorrectly flagged as likely to churn (False Positives), and 0.66% of attrited customers were missed (False Negatives).

The model's recall for attrited customers is strong at 96.05%, meaning it catches most customers likely to churn. However, the precision for churn prediction is 77.39%, indicating some over-flagging of non-churning customers.

2.Key Drivers of Churn (Feature Importance):

Transactional Behavior is Critical: The most influential features are related to transactional activity:

Total_Trans_Ct (0.38) and Total_Trans_Amt (0.25): Customers with fewer or lower-value transactions are more likely to churn.

Total_Ct_Chng_Q4_Q1 (0.13) and Total_Amt_Chng_Q4_Q1 (0.10): A decline in transaction frequency or amount over time signals higher churn risk.

Credit Usage Matters: Total_Revolving_Bal (0.15) and Avg_Utilization_Ratio (0.07) indicate that customers with low revolving balances or credit utilization are more likely to leave, possibly due to underuse of the card.

Engagement Indicators: Total_Relationship_Count (0.12) and Months_Inactive_12_mon (0.07) suggest that customers with fewer bank products or longer periods of inactivity are at higher risk of churn.

Demographics Less Relevant: Features like Gender, Income_Category, Education_Level, and Marital_Status have minimal impact (<0.05), showing that churn is driven more by behavior than demographics.

3. **Customer Behavior Patterns:**

Customers who actively use their credit cards (higher transaction counts, amounts, and revolving balances) are more likely to remain loyal.

A drop in engagement (e.g., fewer transactions, longer inactivity) is a strong predictor of churn, suggesting dissatisfaction or lack of perceived value in the card.

Recommendations for Thera Bank

1.Enhance Customer Engagement to Boost Transaction Activity:

Incentivize Usage: Introduce rewards programs, cashback offers, or bonus points for frequent transactions to encourage higher usage, especially for customers showing declining transaction counts or amounts.

Personalized Offers: Use transaction data to offer tailored promotions (e.g., discounts on frequently purchased categories) to increase card usage and perceived value.

2.Target At-Risk Customers with Retention Campaigns:

Proactive Outreach: Focus on customers with low transaction counts, declining activity (e.g., Total_Ct_Chng_Q4_Q1), or extended inactivity periods. Offer incentives like waived annual fees or bonus rewards to re-engage them.

Revolving Balance Encouragement: For customers with low revolving balances, introduce low-interest balance transfer options or flexible payment plans to encourage card usage.

3.Strengthen Relationships to Reduce Churn:

Cross-Sell Products: Since Total_Relationship_Count is a key factor, encourage customers to adopt additional bank products (e.g., savings accounts, loans) through bundled offers, as this increases loyalty.

Engagement Monitoring: Set up alerts for customers with increased inactivity (e.g., Months_Inactive_12_mon > 3) and reach out with re-engagement campaigns, such as reminders of card benefits or exclusive offers.

4.Optimize the Model for Better Precision:

Reduce False Positives: Adjust the model's threshold to minimize false positives (currently 4.67%), ensuring retention efforts are focused on truly at-risk customers and reducing unnecessary outreach costs.

Address Class Imbalance: The dataset is imbalanced (more existing than attrited customers). Techniques like SMOTE (already in the notebook) can help the model better identify attrited customers without over-flagging existing ones.

6. **Focus on Behavioral Data, Not Demographics:**

Since demographic factors (e.g., Gender, Education_Level) have little impact, prioritize behavioral data for segmentation and targeting. Invest in systems to track and analyze transactional patterns in real-time to predict churn early.

Conclusion

Thera Bank can reduce credit card churn by focusing on customers with declining transactional activity, low credit usage, and reduced engagement. By incentivizing usage, strengthening customer relationships, and refining the predictive model, the bank can enhance retention, ultimately preserving a key revenue stream. Implementing these strategies will require a combination of data-driven targeting, personalized offers, and continuous monitoring of customer behavior.

Key Insights from the Analysis

Model Performance Overview (Confusion Matrix)

The classification model achieved a high accuracy of 94.67%, correctly identifying 78.67% of existing customers (True Negatives) and 16.00% of attrited customers (True Positives).

However, there are misclassifications: 4.67\% of existing customers were incorrectly flagged as likely to churn (False Positives), and 0.66\% of attrited customers were missed (False Negatives).

The model's recall for attrited customers is strong at 96.05%, meaning it catches most customers likely to churn. However, the precision for churn prediction is 77.39%, indicating some over-flagging of non-churning customers.

Key Drivers of Churn (Feature Importance)

Transactional Behavior is Critical: The most influential features are related to transactional activity:

Total Trans Ct (0.38) and Total Trans Amt (0.25): Customers with fewer or lower-value transactions are more likely to churn.

Total Ct Chng Q4 Q1 (0.13) and Total Amt Chng Q4 Q1 (0.10): A decline in transaction frequency or amount over time signals higher churn risk.

Credit Usage Matters: Total Revolving Bal (0.15) and Avg Utilization Ratio (0.07)} indicate that customers with low revolving balances or credit utilization are more likely to leave, possibly due to underuse of the card.

Engagement Indicators: Total Relationship Count (0.12)}and Months Inactive 12 mon (0.07)suggest that customers with fewer bank products or longer periods of inactivity are at higher risk of churn.

Demographics Less Relevant: Features like Gender, Income Category, Education Level, and Marital Status have minimal impact (<0.05), showing that churn is driven more by behavior than demographics.

Customer Behavior Patterns

Customers who actively use their credit cards (higher transaction counts, amounts, and revolving balances) are more likely to remain loyal.

A drop in engagement (e.g., fewer transactions, longer inactivity) is a strong predictor of churn, suggesting dissatisfaction or lack of perceived value in the card.

Recommendations for Thera Bank

Enhance Customer Engagement to Boost Transaction Activity:

Incentivize Usage: Introduce rewards programs, cashback offers, or bonus points for frequent transactions to encourage higher usage,

especially for customers showing declining transaction counts or amounts.

Personalized Offers: Use transaction data to offer tailored promotions (e.g., discounts on frequently purchased categories) to increase card usage and perceived value.

Target At-Risk Customers with Retention Campaigns:

Proactive Outreach: Focus on customers with low transaction counts, declining activity (e.g., Total Ct Chng Q4 Q1), or extended inactivity periods. Offer incentives like waived annual fees or bonus rewards to re-engage them.

Revolving Balance Encouragement: For customers with low revolving balances, introduce low-interest balance transfer options or flexible payment plans to encourage card usage.

Strengthen Relationships to Reduce Churn:

Cross-Sell Products: Since Total Relationship Count is a key factor, encourage customers to adopt additional bank products (e.g., savings accounts, loans) through bundled offers, as this increases loyalty.

Engagement Monitoring: Set up alerts for customers with increased inactivity (e.g., Months Inactive 12 mon > 3) and reach out with re-engagement campaigns, such as reminders of card benefits or exclusive offers.

Optimize the Model for Better Precision:

Reduce False Positives: Adjust the model's threshold to minimize false positives (currently 4.67\%), ensuring retention efforts are focused on truly at-risk customers and reducing unnecessary outreach costs.

Address Class Imbalance: The dataset is imbalanced (more existing than attrited customers). Techniques like SMOTE (already in the

notebook) can help the model better identify attrited customers without over-flagging existing ones.

Focus on Behavioral Data, Not Demographics:

Since demographic factors (e.g., Gender, Education Level) have little impact, prioritize behavioral data for segmentation and targeting. Invest in systems to track and analyze transactional patterns in real-time to predict churn early.

Conclusion

Thera Bank can reduce credit card churn by focusing on customers with declining transactional activity, low credit usage, and reduced engagement. By incentivizing usage, strengthening customer relationships, and refining the predictive model, the bank can enhance retention, ultimately preserving a key revenue stream. Implementing these strategies will require a combination of data-driven targeting, personalized offers, and continuous monitoring of customer behavior.