Credit Card Attrition Analysis:

Churn Clues: Unlocking Retention Secrets

Date: 05/12/2025 – 06/03/2025

By Rohit Raman:

Business Problem:

1. **Why are customers leaving (attrition/churn), and what factors contribute most to it?**
   (Exploratory Data Analysis, Logistic Regression, Random Forest, Chi-Square Test, T-Test)
2. **Which customer segments are most valuable, and how can we retain them?**
   (K-Means Clustering, Percentile Analysis, Bar Graphs and Tables, Principal Component Analysis)
3. **How do credit card usage patterns (e.g., credit limit, utilization, transactions) affect churn and revenue?**
   (Correlation Analysis, Bar Graphs and Tables, Random Forest Feature Importance, Boxplots)
4. **How can we optimize retention strategies based on demographic and behavioral data?**
   (Exploratory Data Analysis, Boxplots, Chi-Square Test, Bar Graphs)
5. **Which customer groups should we target for cross-selling or premium offerings to maximize revenue?**
   (ANOVA, Tukey HSD Test, Bar Graphs and Tables, K-Means Clustering)

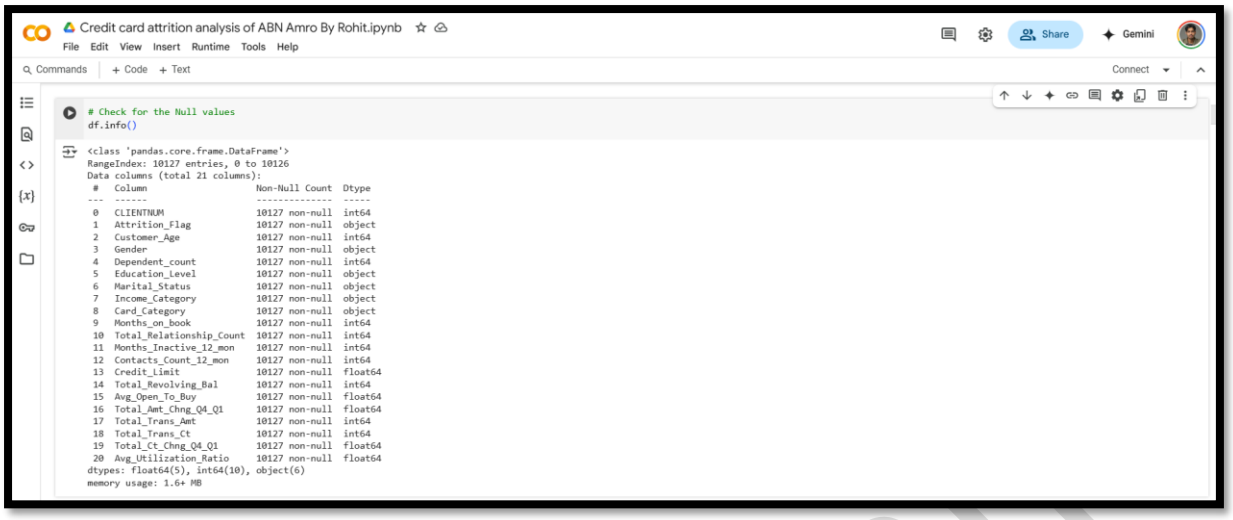
Basic overview of the datasets:

Fig1: Type of columns in the datasets

1. Types of columns in the datasets

2. **CLIENTNUM** – Unique identifier assigned to each customer.
3. **Attrition_Flag** – Indicates whether the customer has left the bank (Attrited Customer) or is still active (Existing Customer).
4. **Customer_Age** – Age of the customer.
5. **Gender** – Gender of the customer (Male or Female).
6. **Dependent_count** – Number of dependents (children or other family members financially dependent on the customer).
7. **Education_Level** – Highest education qualification (High School, Graduate, Doctorate, etc.).
8. **Marital_Status** – Marital status of the customer (Single, Married, Divorced).
9. **Income_Category** – Customer's annual income range ($40K - $60K, $80K+, etc.).
10. **Card_Category** – Type of credit card owned (Blue, Silver, Gold, Platinum).
11. **Months_on_book** – Total number of months the customer has been with the bank.
12. **Total_Relationship_Count** – Number of products the customer has with the bank (e.g., savings, checking, credit card).
13. **Months_Inactive_12_mon** – Number of months the customer was inactive in the last 12 months.
14. **Contacts_Count_12_mon** – Number of times the customer contacted the bank in the last 12 months.
15. **Credit_Limit** – Maximum credit amount assigned to the customer.
16. **Total_Revolving_Bal** – Total outstanding balance on the customer's credit card.
17. **Avg_Open_To_Buy** – Average amount available to spend (Credit Limit - Current Balance).
18. **Total_Amt_Chng_Q4_Q1** – Change in total transaction amount from Q4 to Q1.
19. **Total_Trans_Amt** – Total amount spent by the customer on transactions.
20. **Total_Trans_Ct** – Total number of transactions made by the customer.
21. **Total_Ct_Chng_Q4_Q1** – Change in the total number of transactions from Q4 to Q1.
22. **Avg_Utilization_Ratio** – Ratio of total balance to total credit limit (Total_Revolving_Bal / Credit_Limit).
23. **Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1** – Output of a **Naive Bayes classifier**, predicting attrition based on multiple factors.
24. **Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2** – Another prediction output from the **Naive Bayes classifier**, possibly representing probability scores for different classes

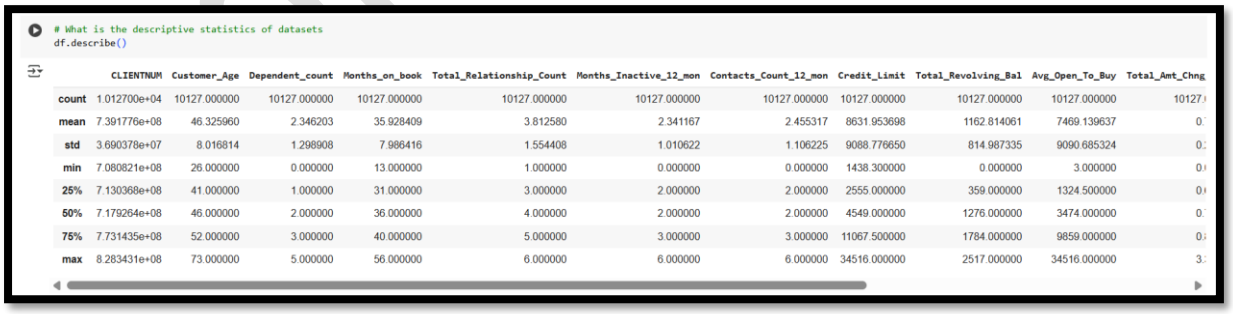
Check the Null Values:



2. Data type and null value check

This dataset contains information about **10,127 customers** and their credit card usage. Each row represents a customer, identified by **CLIENTNUM**. The dataset includes details like **age, gender, marital status, income category, and education level**. It also tracks **credit card usage**, such as **credit limit, balance, total transactions, and relationship duration with the bank**. Additionally, it records **customer behavior**, including **inactive months, contact count, and changes in spending habits**. The **Attrition_Flag** column indicates whether a customer is still active or has left the bank. This data can be used for **customer retention analysis, churn prediction, and financial risk assessment**.

Descriptive Statistics of the datasets:



	CLIENTNUM	Customer_Age	Dependent_count	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng
count	1.012700e+04	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.
mean	7.391776e+08	46.325960	2.346203	35.928409	3.812580	2.341167	2.455317	8631.953698	1162.814061	7469.139637	0.
std	3.690378e+07	8.016814	1.298908	7.986416	1.554408	1.010622	1.106225	9088.776650	814.987335	9090.685324	0.
min	7.080821e+08	26.000000	0.000000	13.000000	1.000000	0.000000	0.000000	1438.300000	0.000000	3.000000	0.
25%	7.130368e+08	41.000000	1.000000	31.000000	3.000000	2.000000	2.000000	2555.000000	359.000000	1324.500000	0.
50%	7.179264e+08	46.000000	2.000000	36.000000	4.000000	2.000000	2.000000	4549.000000	1276.000000	3474.000000	0.
75%	7.731435e+08	52.000000	3.000000	40.000000	5.000000	3.000000	3.000000	11067.500000	1784.000000	9859.000000	0.
max	8.283431e+08	73.000000	5.000000	56.000000	6.000000	6.000000	6.000000	34516.000000	2517.000000	34516.000000	3.

3. Descriptive Statistics of the datasets.

The descriptive statistics provide an overview of the dataset's numerical columns. The **Customer_Age** ranges from **26 to 73 years**, with an average of **46.3 years**. The **Dependent_count** varies between **0 and 5**, indicating different family responsibilities. Customers have been associated with the bank for **an average of 35.9 months**, with a minimum of **13 months** and a maximum of **56 months**. The **Total_Relationship_Count**

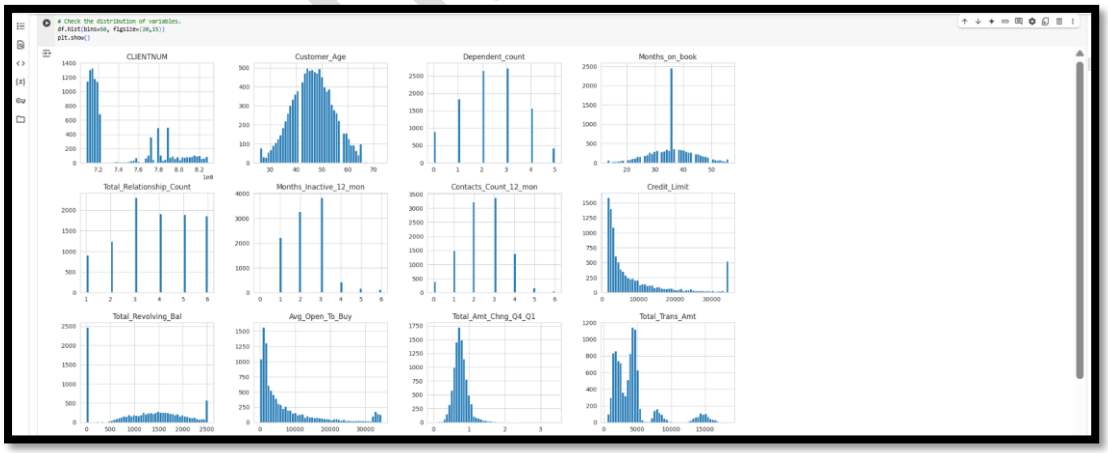
(number of accounts a customer has with the bank) ranges from **1 to 6**, with an average of **3.8**. The **Credit_Limit** has a wide range, from **$1,438 to $34,516**, with an average of **$8,632**. The **Total_Trans_Amt** (total transaction amount) varies significantly, with a median of **$3,899**, but some customers have transactions as high as **$18,484**. The **Avg_Utilization_Ratio**, which represents how much of the credit limit is used, has a median of **0.176**, meaning most customers use around **17.6% of their available credit**. These insights help understand customer demographics, financial behavior, and credit card usage patterns.

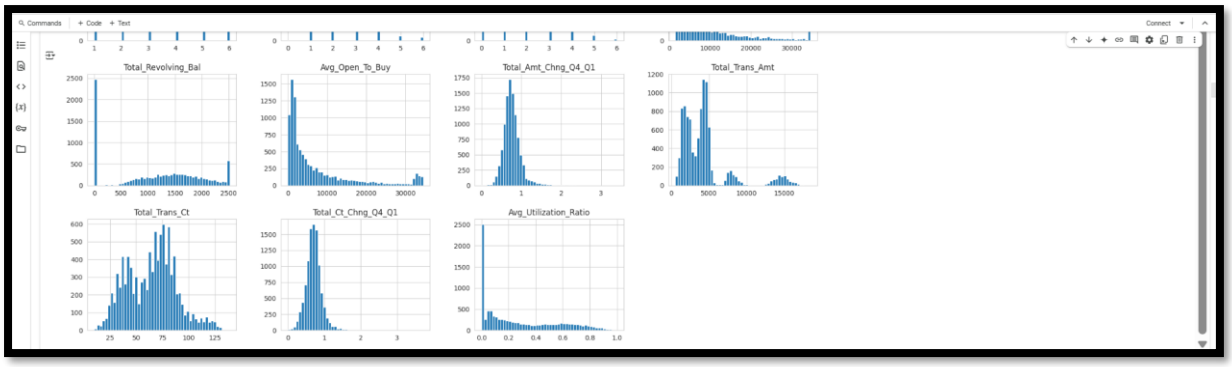
Missing values:

4. Checking the null value across columns.

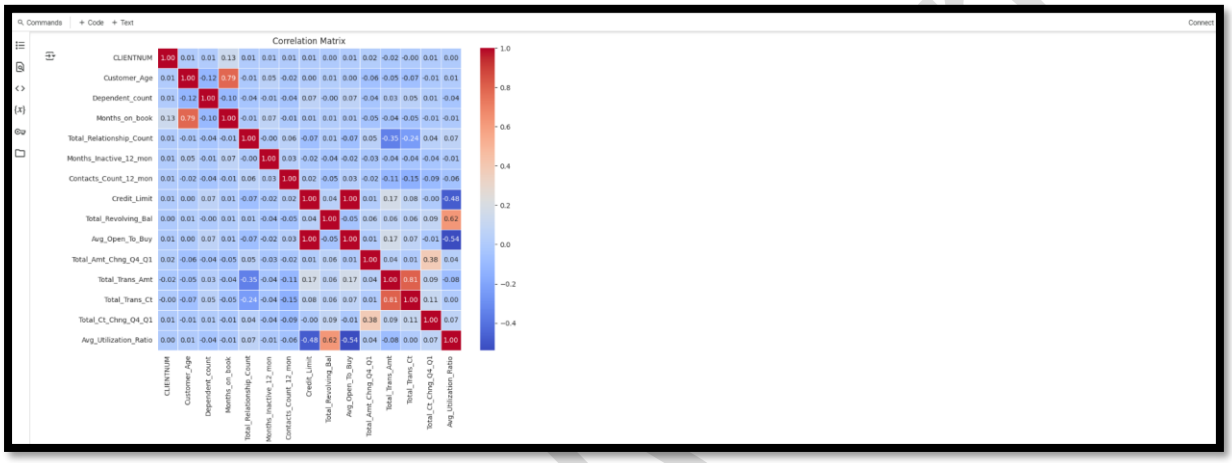
The dataset is clean we don't find any missing values.

Distribution of Variables:





5. Distribution of variables



6.Correaltion matrix

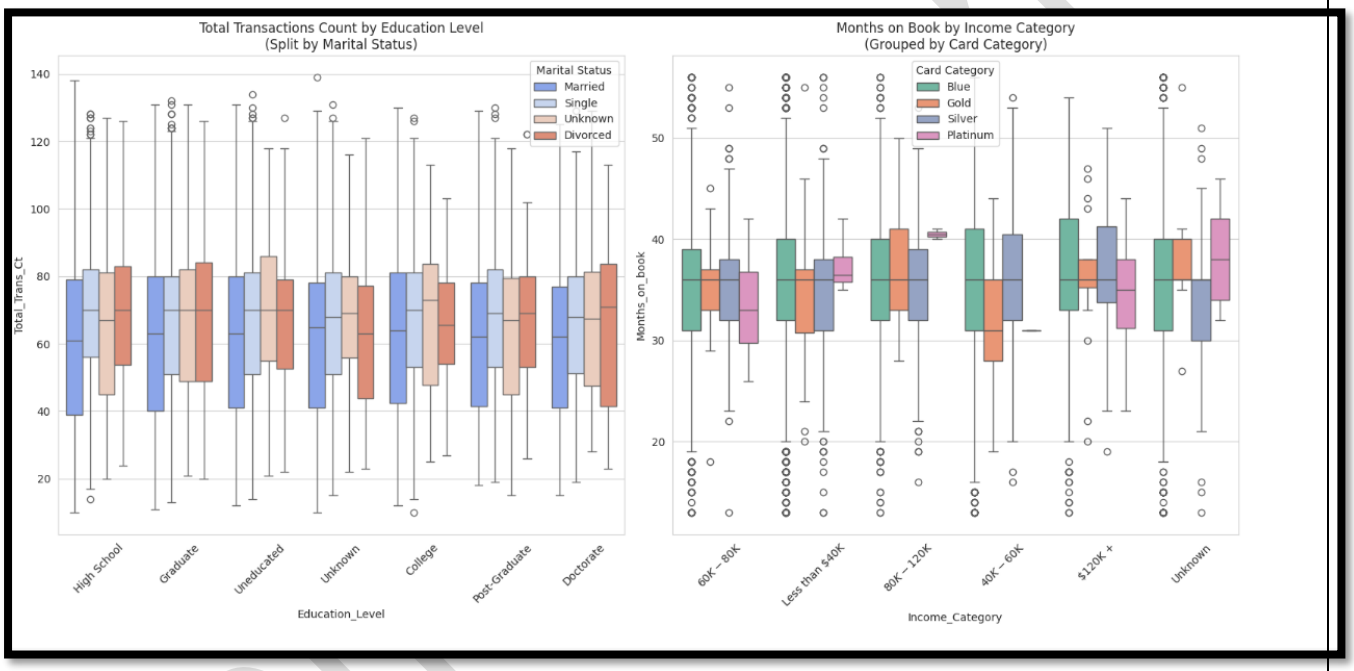
```
Top Positively Correlated Variables:
                Variable 1           Variable 2  Correlation
0        Avg_Open_To_Buy         Credit_Limit     0.995981
1         Total_Trans_Ct       Total_Trans_Amt     0.807192
2         Months_on_book         Customer_Age     0.788912
3  Avg_Utilization_Ratio   Total_Revolving_Bal     0.624022
4    Total_Ct_Chng_Q4_Q1  Total_Amt_Chng_Q4_Q1     0.384189
5        Total_Trans_Amt         Credit_Limit     0.171730
6        Total_Trans_Amt       Avg_Open_To_Buy     0.165923
7         Months_on_book             CLIENTNUM     0.134588
8    Total_Ct_Chng_Q4_Q1        Total_Trans_Ct     0.112324
9    Total_Ct_Chng_Q4_Q1   Total_Revolving_Bal     0.089861

Top Negatively Correlated Variables:
                Variable 1               Variable 2  Correlation
0  Avg_Utilization_Ratio          Avg_Open_To_Buy    -0.538808
1  Avg_Utilization_Ratio             Credit_Limit    -0.482965
2        Total_Trans_Amt  Total_Relationship_Count    -0.347229
3         Total_Trans_Ct  Total_Relationship_Count    -0.241891
4         Total_Trans_Ct     Contacts_Count_12_mon    -0.152213
5        Dependent_count             Customer_Age    -0.122254
6        Total_Trans_Amt     Contacts_Count_12_mon    -0.112774
7         Months_on_book          Dependent_count    -0.103062
8    Total_Ct_Chng_Q4_Q1     Contacts_Count_12_mon    -0.094997
9  Avg_Utilization_Ratio           Total_Trans_Amt    -0.083034
```

7. Correlation table.

The correlation matrix highlights relationships between various customer attributes. Positive correlations indicate that as one variable increases, the other tends to increase as well. For

example, **credit limit and available credit** are strongly correlated, meaning **customers with higher credit limits have more available credit**. Similarly, **total transaction count and total transaction amount show a strong relationship**, suggesting that frequent transactions lead to higher spending. **Negative correlations show** inverse relationships, such **as credit utilization ratio and available credit**, meaning that customers who use more of their credit have less remaining credit. Additionally, older customers tend to have fewer dependents, and customers with higher interactions with the bank tend to have slightly lower transaction amounts. These insights help in understanding customer behaviour, risk assessment, and financial decision-making.



8. Boxplot: Total trans count with education level (varied with martial status) and months on book with income categories (varied with card type)

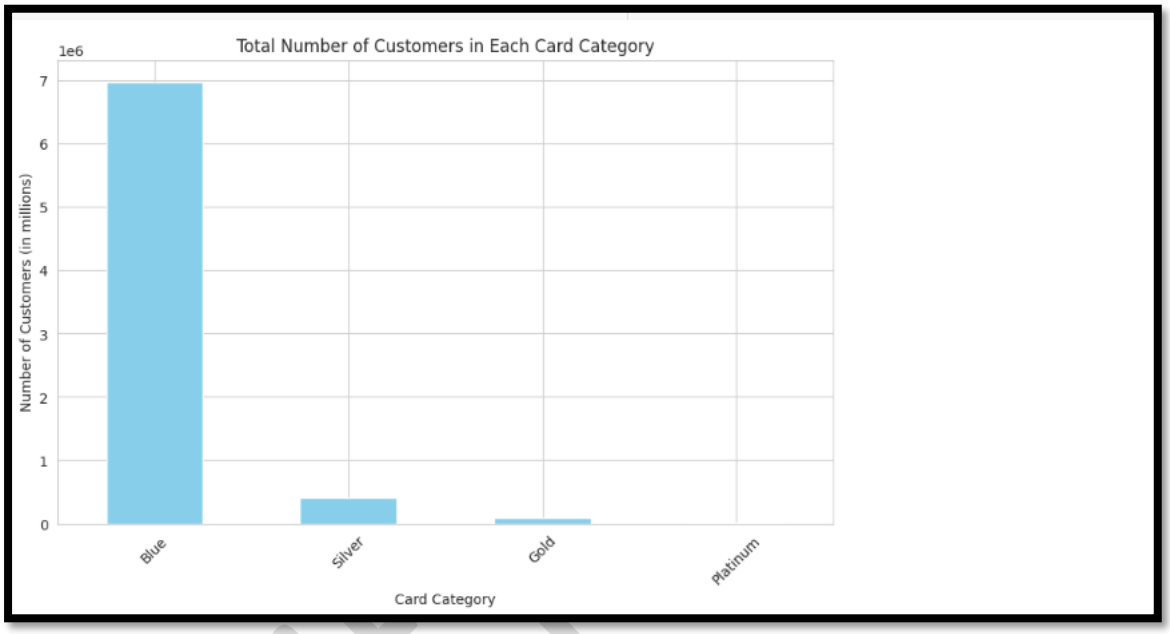
1. **The first boxplot** examines how the **total transaction count (Total_Trans_Ct)** varies by **education level**, with further categorization by **marital status**. This helps to identify if people with different education backgrounds and marital statuses show different spending behaviors.
2. **The second boxplot** explores how the **number of months a customer has been with the bank (Months_on_book)** is distributed across different **income categories**, with a breakdown by **card category**. This analysis can reveal if customers with certain income levels tend to stay longer and which card types, they prefer.

Exploratory Data Analysis:

1. Number of customers within each card category:

```
# Find the total number of customer in each card category and make a barplot:
df.groupby('Card_Category')['CLIENTNUM'].sum().sort_values(ascending=False)
```

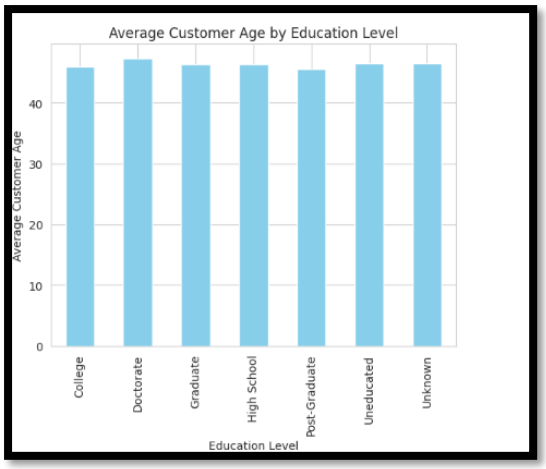
Card_Category	CLIENTNUM
Blue	6974338683888
Silver	410985651315
Gold	85623872853
Platinum	14703411285



9. Bar graph for number of customer in different card Category

This data represents the total transaction amounts for different **card categories**: **Blue, Silver, Gold, and Platinum**. The **Blue card** has the highest transaction volume, while **Platinum** has the lowest, indicating that most customers likely use Blue for their transactions.

Average Customer Age with education level.



10 . Bar graph: Avg age of customer by Education Level.

The average age of customers with a college degree is around 43, while those with a doctorate degree have an average age of 46. Customers with a graduate degree, high school education, or postgraduate degree have an average age of approximately 42.

Number of Married Customer who were Attrited , avg utilization ratio with month on book greater than 12 and ratio across gender

```
#6. Count the number of customers with a Marital_Status of 'Single' and Attrition_Flag of 'Attrited Customer'.

df[(df["Marital_Status"] == "Single") & (df['Attrition_Flag'] == 'Attrited Customer')].shape[0]
#Remember here count will count all the null values across the column, but shape would count all the rows based on the above conditions
```

```
668
```

```
#7. Calculate the average Avg_Utilization_Ratio for customers who have a Months_on_book greater than 12 months.
df[df['Months_on_book'] > 12]['Avg_Utilization_Ratio'].mean()
```

```
np.float64(0.2748935518909845)
```

```
#7. Calculate the average Avg_Utilization_Ratio for customers who have a Months_on_book greater than 12 months based on the gender.
df[df['Months_on_book']>12].groupby('Gender')['Avg_Utilization_Ratio'].mean()
```

Gender	Avg_Utilization_Ratio
F	0.341957
M	0.199548

dtype: float64

There are **668 customers** who are **single** and have **attrited**. The **average utilization ratio** for customers who have been with the bank for more than **12 months** is about **27%**. Specifically, the average utilization ratio is **34% for females** and **9% for males**.

The **average utilization ratio** (the percentage of available credit a customer is using) is **higher for females (0.34 or 34%)** compared to **males (0.20 or 20%)**. This suggests that, on average, female customers are using a larger portion of their available credit than male customers.

```
#9. "Find the total Total_Relationship_Count grouped by Attrition_Flag status based on the education level"

df.groupby(['Attrition_Flag', 'Education_Level'])['Total_Relationship_Count'].mean().sort_values(ascending= False)
```

Attrition_Flag	Education_Level	Total_Relationship_Count
Existing Customer	Post-Graduate	4.011792
	Uneducated	3.957600
	Unknown	3.920823
	High School	3.913298
	Graduate	3.911776
	College	3.838184
	Doctorate	3.837079
Attrited Customer	Doctorate	3.410526
	Graduate	3.355236
	Post-Graduate	3.347826
	College	3.253247
	Uneducated	3.240506
	High School	3.215686
	Unknown	3.191406

dtype: float64

11. Table to show the mean of total relationship count based on education level across attrition flag category.

Table: Total relationship counts vs attrition flag varied with education level

Total relationship counts across education level with attrition flag status.

The **total relationship count** represents the number of different relationships a customer has with the bank. Among existing customers, **post-graduate and uneducated individuals** have the highest relationship counts, with values of **4 and 3**, respectively. For **attrited customers**, those with a **doctorate or graduate degree** typically have **3 relationships** with the bank.

```
#9. Calculate the total Total_Trans_Amt for each Card_Category.
df.groupby('Card_Category')['Total_Trans_Amt'].sum()
```

Card_Category	Total_Trans_Amt
Blue	39870938
Gold	891531
Platinum	179995
Silver	3657718

dtype: int64

12. Total transaction across different card category.

This table represents the **total transaction amount** for customers based on their **card category**. Customers with the **Blue** card have the highest total transaction amount at **39.87 million**, followed by **silver** cardholders with **3.66 million**. **Gold** cardholders have transactions totalling **891,531**, while **Platinum** cardholders have the lowest total transaction amount at **179,995**. This suggests that **blue cardholders are the most active in terms of transactions**, while **Platinum cardholders have the least activity.**

```
#10.  Filter the customers who have been inactive for more than 6 months (Months_Inactive_12_mon > 6) and count them.

df[df['Months_Inactive_12_mon'] >5 ].shape[0]
```

```
124
```

```
inactive_customers = df[df['Months_Inactive_12_mon'] > 5]

# Count of inactive customers
inactive_count = inactive_customers.shape[0]

# Unique values and counts for card types, gender, and education
card_distribution = inactive_customers['Card_Category'].value_counts()
gender_distribution = inactive_customers['Gender'].value_counts()
education_distribution = inactive_customers['Education_Level'].value_counts()

print(f"Total inactive customers: {inactive_count}")
print("\nCard Type Distribution:\n", card_distribution)
print("\nGender Distribution:\n", gender_distribution)
print("\nEducation Background Distribution:\n", education_distribution)
```

```
Total inactive customers: 124

Card Type Distribution:
 Card_Category
Blue      119
Silver      3
Gold        2
Name: count, dtype: int64

Gender Distribution:
 Gender
F    70
M    54
Name: count, dtype: int64

Education Background Distribution:
 Education_Level
Graduate         41
Unknown          24
High School      23
Uneducated       14
Doctorate        10
College           9
Post-Graduate     3
Name: count, dtype: int64
```

Fig13: Code, table bar to filter out the customer who has been inactive more than 6 months

The analysis of **124 inactive customers** reveals key insights into their demographics and card preferences. The majority of these customers hold a **blue card (119 out of 124),** with only a small number using **Silver (3) and Gold (2) cards**, suggesting that premium cardholders may be more engaged. In terms of gender distribution, **70 inactive customers are female and 54 are male**, indicating a slightly higher attrition rate among women. Examining education levels, the highest number of inactive customers are **Graduates (41), followed by those with an Unknown education level (24) and High School graduates (23).** On the other hand, **Doctorate (10) and post-graduate (3) customers** have the lowest inactivity rates, which may suggest that **higher-educated individuals tend to be more engaged with the bank**

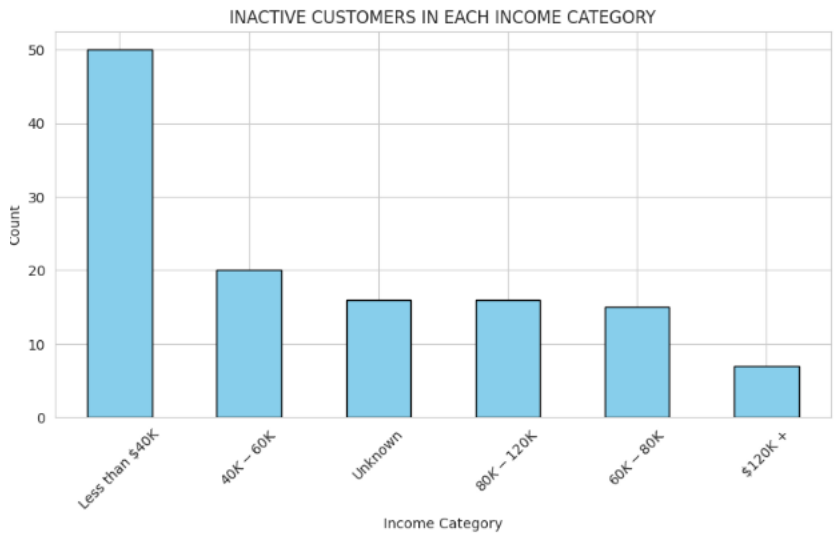


Fig 14. Bar Graph for inactive customer in each income category.

Customers who have been inactive for more than 6 months mostly have an income of **less than $40K**. The second-largest group falls within the **$40K to $60K** income range. On the other hand, customers earning **$120K or more** are the least inactive, with only **around 7 customers** staying inactive for more than 6 months.

```
# 11. Group the dataset by Income_Category and calculate the median Avg_Open_To_Buy for each group.
df.groupby('Income_Category')['Avg_Open_To_Buy'].median()
```

Income_Category	Avg_Open_To_Buy
$120K +	17117.0
$40K - $60K	2580.5
$60K - $80K	6418.5
$80K - $120K	11606.0
Less than $40K	1478.0
Unknown	5464.5

dtype: float64

15. Table for Income category against Avg open to buy.

The table shows the average "Open to Buy" amount, which is the available credit left after spending, across different income categories. Customers earning over $120K have the highest average available credit at $17,117, followed by those in the $80K–$120K range with $11,606. On the other hand, customers with incomes less than $40K have the lowest average open credit at $1,478. This indicates a clear trend where higher-income groups tend to have more available credit, while lower-income groups have less, likely due to lower credit limits or higher utilization.

```
# 12.Find the mean of  total Total_Trans_Amt for customers across the gender who have Total_Trans_Ct greater than 50.
df[df['Total_Trans_Ct']>50].groupby('Gender')['Total_Trans_Amt'].mean()
```

Gender	Total_Trans_Amt
F	5235.630631
M	5961.645472

dtype: float64

16. Avg of total transaction amount across the gender.

the table shows average total transaction amount by gender. On average, male customers spend about $5,961, while female customers spend around $5,236. This means male customers tend to make slightly higher total transactions than female customers.

```
#14.  Find the average Total_Revolving_Bal by Education_Level for customers who have a Card_Category of 'Blue'.

df[df['Card_Category']=='Blue'].groupby('Education_Level')['Total_Revolving_Bal'].mean()
```

Education_Level	Total_Revolving_Bal
College	1122.006383
Doctorate	1093.082938
Graduate	1160.268713
High School	1183.971398
Post-Graduate	1188.418067
Uneducated	1159.383896
Unknown	1149.156338

dtype: float64

17. Avg of Education level across different education level.

The average total revolving balance, which represents the unpaid balance carried by customers on their credit cards, varies slightly across different education levels. Customers with a post-graduate education have the highest average revolving balance at approximately $1,188, followed closely by those with a high school education at around $1,183. Those with a graduate or uneducated background also maintain similar balances, hovering around $1,159 to $1,160. Interestingly, customers with a doctorate have the lowest average revolving balance at about $1,093, suggesting they may manage their credit usage more conservatively compared to other groups. Overall, the differences are modest but show that education level may have a subtle impact on credit behaviour.

```
#15.  Calculate the percentage of customers who have Attrition_Flag as 'Attrited Customer' for each Income_Category.

attrited_percentage=df[df['Attrition_Flag']=="Attrited Customer"].groupby('Income_Category')['CLIENTNUM'].count()/len(df)*100
attrited_percentage
```

Income_Category	CLIENTNUM
$120K +	1.244199
$40K - $60K	2.676015
$60K - $80K	1.866298
$80K - $120K	2.389651
Less than $40K	6.043251
Unknown	1.846549

dtype: float64

18. Percent of attrited customer across different income category.

This data shows the average number of customers (represented by CLIENTNUM count per group) across different income categories. Customers earning **less than $40K** have the **highest representation**, averaging about **6 customers**, indicating that a larger portion of the customer base falls in this income range. Meanwhile, customers earning **$40K - $60K** and **$80K - $120K** also show moderate representation with averages of about **2.68** and **2.39** respectively. In contrast, higher income brackets like **$120K+** and **$60K - $80K**

have fewer customers on average, around **1.24** and **1.87**, suggesting that the bank serves more customers in the lower to middle-income segments than in higher ones.

```
#16.  Group customers by Marital_Status, and gender and calculate the sum of Contacts_Count_12_mon for each group.
df.groupby(['Marital_Status', "Gender"])['Contacts_Count_12_mon'].sum()
```

Marital_Status	Gender	Contacts_Count_12_mon
Divorced	F	970
	M	842
Married	F	5852
	M	5665
Single	F	5218
	M	4507
Unknown	F	892
	M	919

dtype: int64

19. Total number customer across different gender who contacted the bank for the customer service

The data shows the number of customer service contacts made in the past 12 months, categorized by marital status and gender. Married customers had the highest number of contacts, with married females reaching out 5,852 times and married males 5,665 times. Single customers followed, with single females making 5,218 contacts and single males 4,507. Divorced customers had significantly fewer contacts—970 for females and 842 for males. Interestingly, among customers whose marital status is unknown, males contacted slightly more (919) than females (892). This suggests that married and single customers tend to engage more with customer service, possibly indicating higher product usage or support needs.

Customer Attrition Prediction using Machine Learning:

1. Why are customers leaving (attrition/churn), and what factors contribute most to it?

Customer attrition, or churn, is a key challenge for businesses, particularly in the financial sector. Identifying customers at risk of leaving enables organizations to implement proactive retention strategies, such as personalized offers and improved customer engagement. This study employs machine learning techniques to predict customer attrition and determine key factors contributing to churn.

Data Preprocessing: To ensure the dataset was suitable for modelling, the following preprocessing steps were performed:

Handling Categorical Variables:

The target variable, **Attrition Flag**, was label-encoded: **Existing Customer → 1**,

Attrited Customer → 0,

Categorical features (Gender, Education_Level, Marital_Status, Income_Category, Card_Category) were also label-encoded for compatibility with machine learning algorithms. Feature Selection:

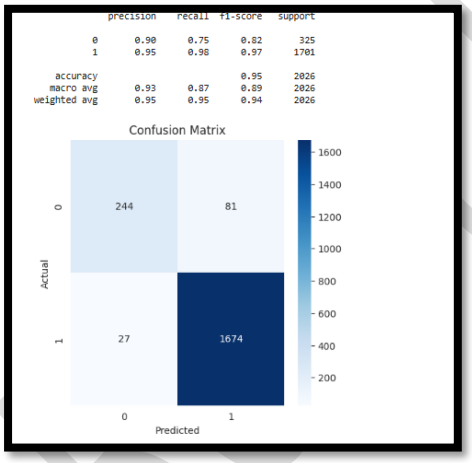
The column CLIENTNUM was dropped as it was only an identifier and did not provide predictive value. The remaining variables, including Credit Limit, Total Transaction Amount, Total Transaction Count, and Utilization Ratio, were retained based on their potential influence on attrition.

Train-Test Split:

The dataset was split into 80% training and 20% testing to evaluate model performance. Stratified sampling was applied to maintain class balance.

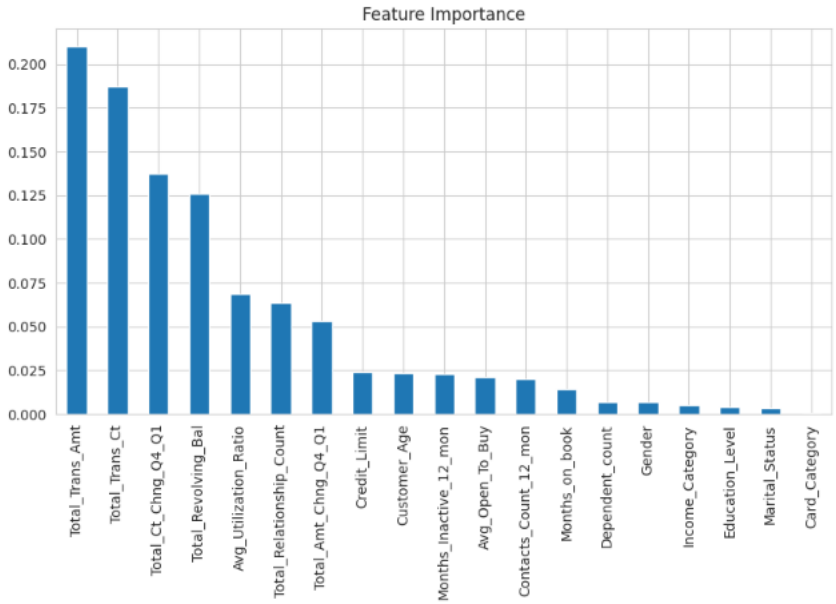
Feature Scaling:

Numerical variables were standardized using StandardScaler to normalize the data distribution and improve model performance.

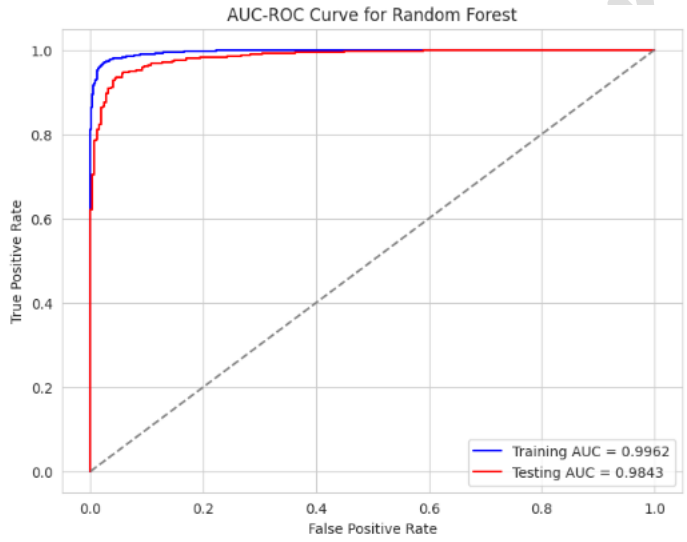


20. Confusion matrix for random forest.

The model performs very well overall with an **accuracy of 94.67%**. It is especially strong at identifying class **1** (likely the majority class), with high precision and recall, meaning it correctly predicts most of those cases. However, it is less effective at predicting class **0**, with a lower recall of **75%**, which means it misses more instances of that class. Overall, the model is reliable but slightly biased towards class 1.



21. Important variable that has influence on the attrition in the random forest model .




```
[ ]  print(f"Training AUC: {train_auc:.4f}, Testing AUC: {test_auc:.4f}")

Training AUC: 0.9962, Testing AUC: 0.9843
```

Ruc Curve for model performance and behaviour of training and testing data on model.

The Random Forest model shows that customers are more likely to leave (churn) if they don't use their credit card much—low transaction amounts (**Total_Trans_Amt**), fewer transactions **(Total_Trans_Ct**), and little or no balance **(Total_Revolving_Bal)** are big red flags. It also finds that how much customers use their credit limit **(Avg_Utilization_Ratio)** and changes in their spending over time **(Total_Amt_Chng_Q4_Q1)** matter too. With 94.67% accuracy, the model is great at predicting who stays (engaged customers) but misses some who leave, meaning these factors are key to keeping customers but don't catch every churn case. In short, to stop churn, get customers spending more and using their cards regularly!

Customer Segmentation Using K-Means Clustering

2. **Which customer segments are most valuable, and how can we retain them?**
3. **Which customer groups should we target for cross-selling or premium offerings to maximize revenue?**

Introduction Customer segmentation is a key data analysis technique used to categorize customers based on their behaviour, spending patterns, and financial activity. In this project, I applied K-Means clustering to group customers into meaningful segments based on their transaction history and credit activity.

Feature Selection

For clustering, I selected the following three features:

Total_Trans_Amt - The total amount of transactions made by a customer in a given period.

Total_Relationship_Count - The number of products/services the customer holds with the company.

Credit_Limit - The maximum amount of credit available to the customer. These features are crucial because they capture spending behavior, engagement level, and financial capacity, which help differentiate customers into distinct groups.

Data Preprocessing

Before applying clustering, I standardized the features using StandardScaler to ensure that all variables have a similar scale. This is important because K-Means relies on distance calculations, and features with larger ranges could dominate clustering decisions.

Finding the Optimal Number of Clusters (K)

To determine the optimal number of clusters, we used the Elbow Method, which analyzes the Within-Cluster Sum of Squares (WCSS). The elbow point in the plot represents the value of K where adding more clusters provides minimal improvement. Based on the analysis, we selected K = 3 as the optimal number of clusters.

Applying K-Means Clustering

With K = 3, we trained the K-Means model and assigned each customer to one of the three clusters. After clustering, I calculated the mean values of the selected features for each cluster to interpret their characteristics.

Assigning Meaningful Cluster Labels

By analyzing the cluster means, we assigned descriptive labels to each segment:

Low Spenders with High Credit (Cluster 0)

Low Total_Trans_Amt (spends conservatively) High Credit_Limit (trusted financially but does not spend much) Moderate Total_Relationship_Count (some level of engagement) These customers have high financial trust but do not utilize their credit as much, making them low spenders despite having a high credit limit.

Moderate Spenders with Medium Credit (Cluster 1)

Moderate Total_Trans_Amt (spends occasionally) Medium Credit_Limit (moderate financial trust) Moderate Total_Relationship_Count (some engagement with the company's offerings) These customers balance their spending and credit usage, falling into the mid-tier category.

High Spenders with Medium Credit (Cluster 2)

High Total_Trans_Amt (spends a lot) Medium Credit_Limit (moderate financial access) High Total_Relationship_Count (engaged with multiple products/services) These customers are active spenders who frequently engage with the company's offerings, despite having a medium credit limit.

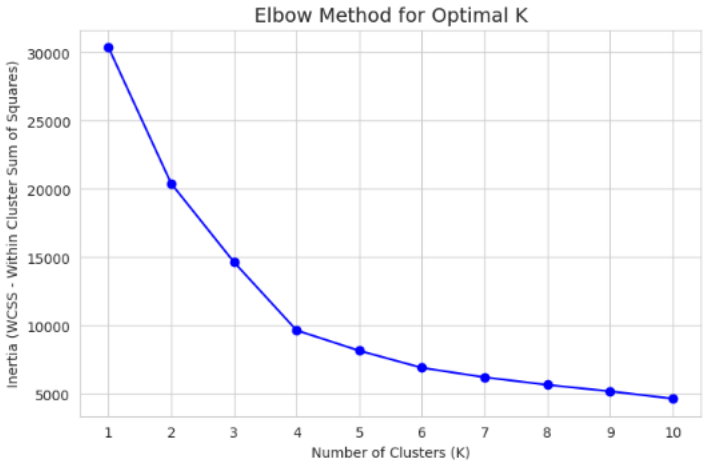
Visualizing Clusters Using PCA

Since I had three features, I used Principal Component Analysis (PCA) to reduce the data to two principal components (PCA1 and PCA2) for visualization. Each customer was plotted on a 2D scatter plot, where:

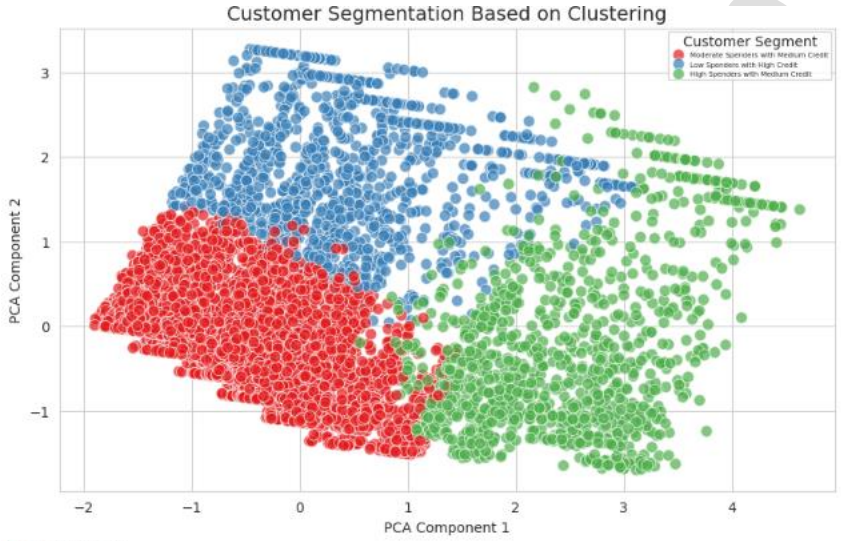
The color represents the assigned cluster. The spread of points shows how different clusters are separated. The legend explains the customer segment each cluster represents.

Why We Changed Labels from PC1 to Feature Names

Originally, the visualization used PCA1 and PCA2, which are mathematical transformations of the original features. However, PCA does not retain feature names, so these labels are abstract. To ensure interpretability, I re-examined the clustering results using the original feature names (Total_Trans_Amt, Total_Relationship_Count, Credit_Limit). This helped me derive real-world insights rather than relying on unnamed principal components.



22. Elbow graph to determine the optimal number of cluster



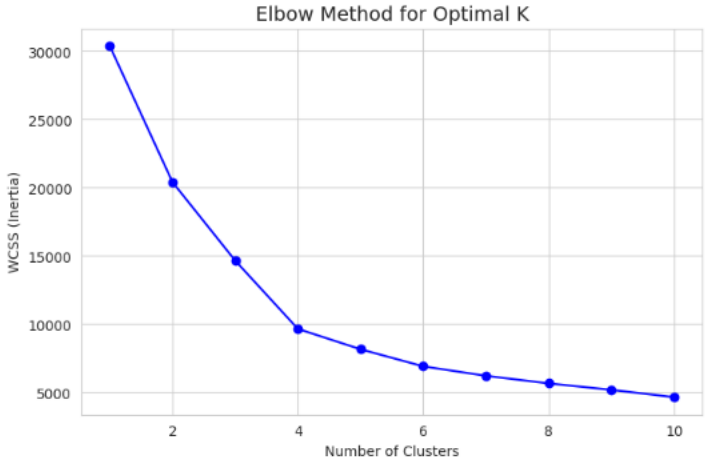

```
Cluster Summary:
 Cluster  Segment
0        Moderate Spenders with Medium Credit    7547
2        Low Spenders with High Credit           1396
1        High Spenders with Medium Credit        1184
Name: count, dtype: int64
Mean Values of Features per Cluster:
    Cluster  Total_Trans_Amt  Credit_Limit
2        2           3354.5       25657.5
0        0           3622.0        3252.0
1        1          13804.0        9300.5
```

23. Customer segmentation based on the PCA

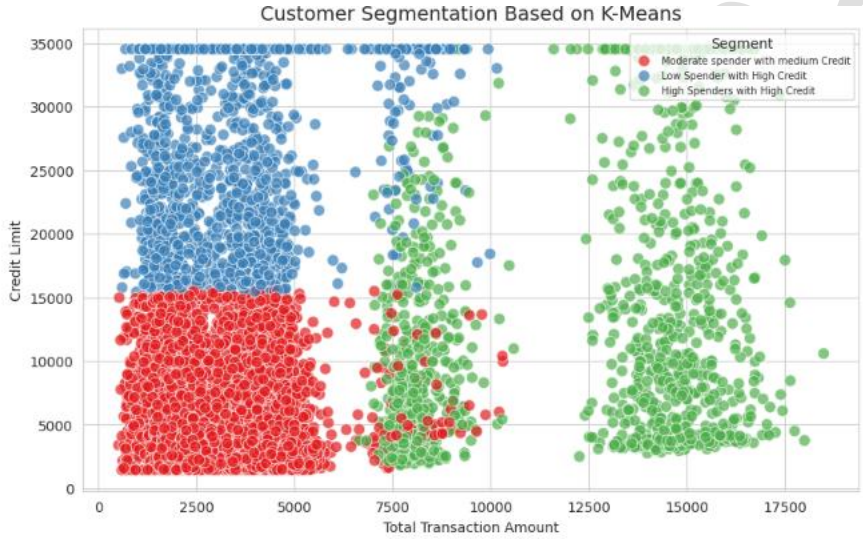
Most customers (around **7,500**) fall into **Cluster 0**, where people spend a **moderate** amount and have **medium credit limits**.

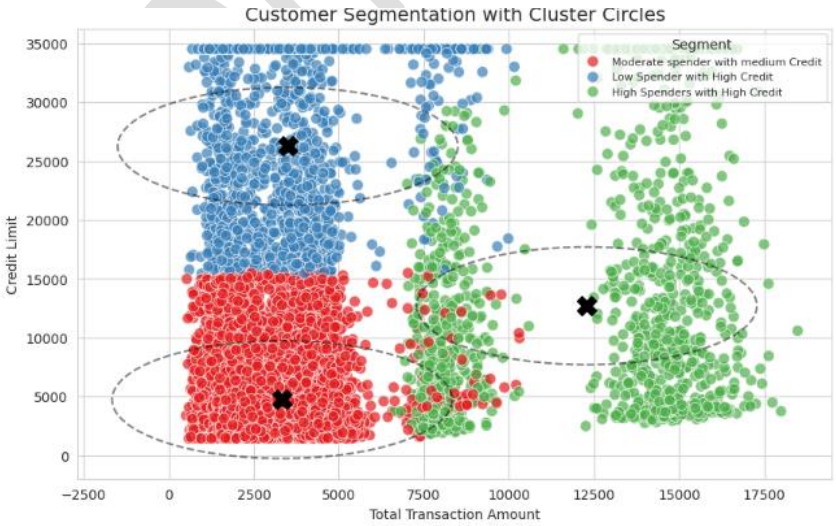
Cluster 2 includes about **1,400** people who are **low spenders** but have **high credit limits**.

Cluster 1 has about **1,200** customers who are **high spenders** with **medium credit limits**.




```
Cluster Summary:
 Cluster  Segment
0        Moderate spender with medium Credit    7547
2        Low Spender with High Credit           1396
1        High Spenders with High Credit         1184
Name: count, dtype: int64
```





25. K-means cluster for customer segmentation

Cluster 0: The largest group, with **7,547 customers**, are **moderate spenders** who have a **medium credit limit**. These are average users in terms of spending and available credit.

Cluster 2: Includes **1,396 customers** who are **low spenders** but have a **high credit limit**. They may not use their full credit potential.

Cluster 1: Contains **1,184 customers** who are **high spenders** with **high credit limits**, likely the most profitable or active users for the bank.

Summary of my cluster Analysis:

- **For "Which customer segments are most valuable, and how can I retain them?"**:
    - I discovered that **Cluster 1 (High Spenders with Medium Credit, 1,184 customers)** is my most valuable group because they spend a lot and stay engaged—they're my cash cows. **Cluster 0 (Moderate Spenders with Medium Credit, 7,547 customers)** is also valuable since it's the biggest bunch and keeps things steady. To hold onto them, I'd pamper Cluster 1 with rewards or fancy cards and push Cluster 0 to grab more products to up their worth.
- **For "Which customer groups should I target for cross-selling or premium offerings to maximize revenue?"**:
    - I'd aim premium offers, like top-tier cards, at **Cluster 1** since they're already big spenders. For **Cluster 0**, I'd cross-sell extra stuff like loans or accounts to get them spending more. And for **Cluster 2 (Low Spenders with High Credit, 1,396 customers)**, I'd tempt them with deals to tap into their huge credit limits, turning them into bigger earners for me.

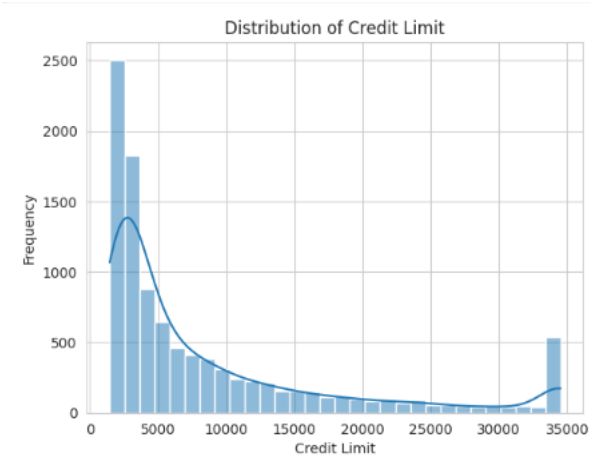
In short, I've figured out that my High Spenders (Cluster 1) are the stars I need to keep and upsell, my Moderate Spenders (Cluster 0) are a huge group I can grow with cross-selling, and my Low Spenders (Cluster 2) are a chance to boost revenue—all helping me hang onto my best customers and rake in more profits!

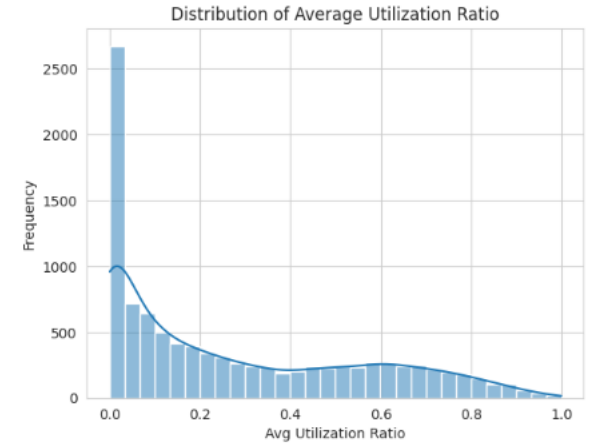
Credit utilization analysis:

4. **"How do credit card usage patterns (e.g., credit limit, utilization, transactions) affect churn and revenue?"**

Objective of analysis

The goal of this analysis is to examine the relationship between Credit Limit and Average Utilization Ratio to determine whether customers with higher credit limits use their credit more efficiently and its impacts on churn.

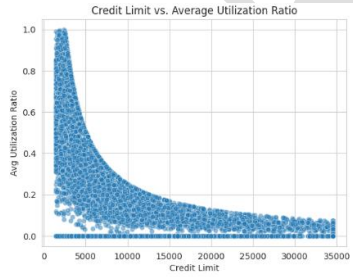


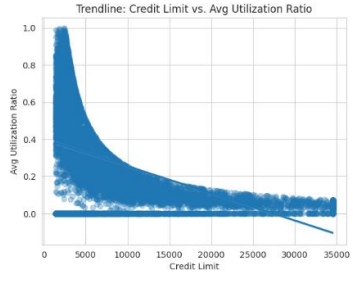


Correlation Matrix:

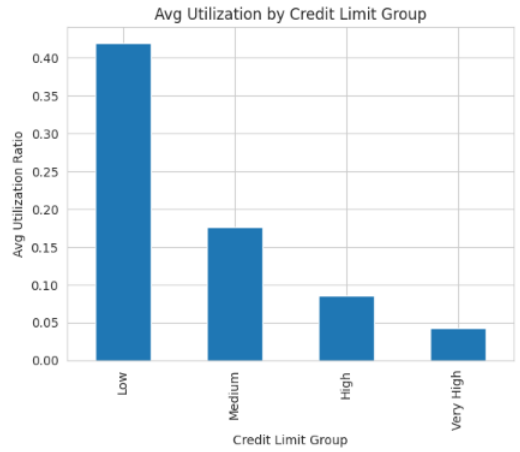
	Credit_Limit	Avg_Utilization_Ratio
Credit_Limit	1.000000	-0.482965
Avg_Utilization_Ratio	-0.482965	1.000000

26. Distribution of Credit limit and Avg Utilization ratios.





```
Credit_Limit_Group
Low          0.419684
Medium       0.175663
High         0.085227
Very High    0.042711
Name: Avg_Utilization_Ratio, dtype: float64
```



27. Credit limit vs utilization ratio relations.

Key Observations:

Negative Correlation (-0.48):

- A correlation of -0.4829 suggests a moderate negative relationship between Credit Limit and Average Utilization Ratio.
- As Credit Limit increases, the Utilization Ratio tends to decrease.
- This means customers with higher credit limits generally use a smaller percentage of their available credit.

Explanation:

- Customers with higher credit limits may have better financial stability and do not need to max out their credit.
- Those with lower credit limits might use a larger proportion of their available credit, leading to a higher utilization ratio.

Business Insight:

Low utilization + High Credit Limit = Responsible Borrowers

- These customers are low-risk, possibly prime customers for premium financial products.

High utilization + Low Credit Limit = Financially Constrained Customers

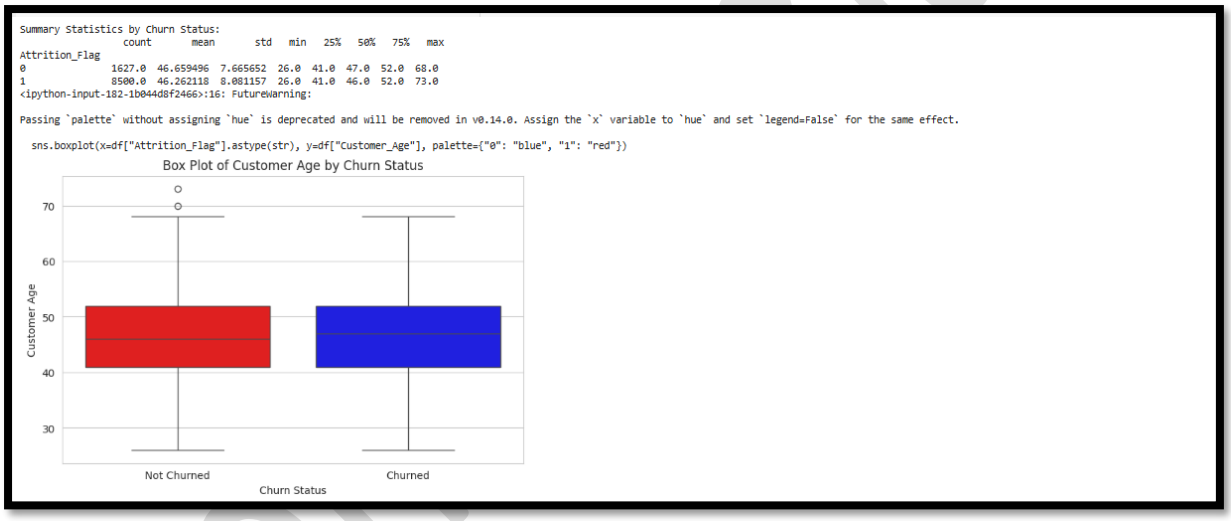
- They may need credit limit increases or financial assistance to avoid debt accumulation.

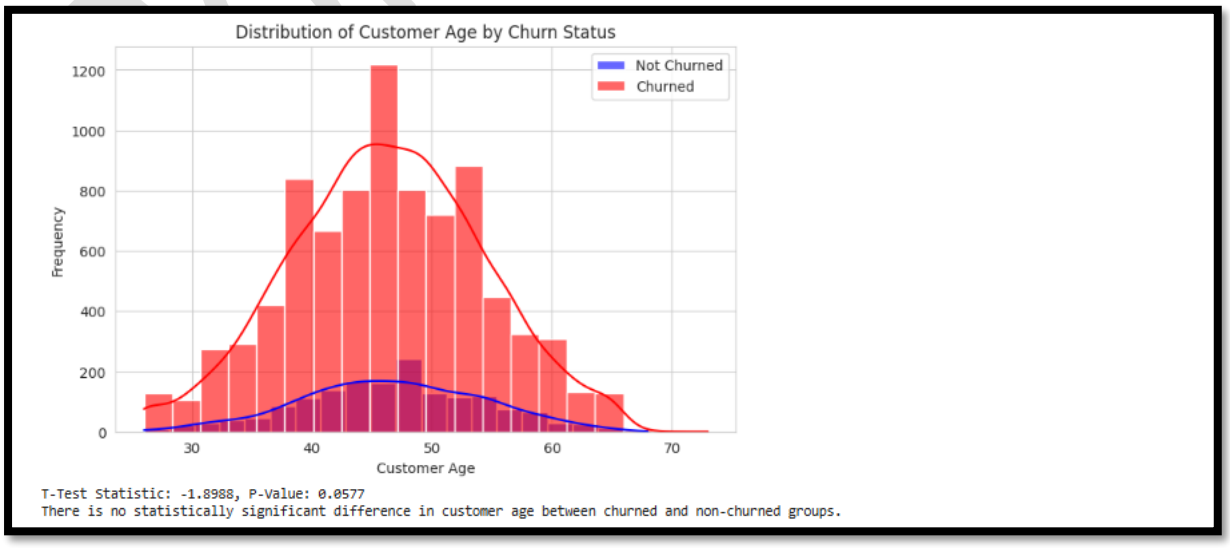
Conclusion:

I found a moderate negative correlation (-0.4829) between Credit_Limit and Avg_Utilization_Ratio, meaning that customers with higher credit limits tend to use a smaller percentage of their available credit, while those with lower limits use a larger share. This shows how credit limit and utilization patterns connect.

- **Low Utilization + High Credit Limit**: These customers use less of their credit (e.g., responsible borrowers), suggesting they're financially stable and less likely to churn, plus they could generate more revenue if encouraged to spend.
- **High Utilization + Low Credit Limit**: These customers max out their smaller limits, hinting at financial strain, which could increase churn risk and limit revenue due to debt concerns.

Churn Risk by Customer Age Analysis.





28. Distribution of customer Age with the churn status

T-Test Statistic: -1.8988

This means there is a slight difference in the average age of churned vs. non-churned customers, but it is not strong enough to be statistically significant.

P-Value: 0.0577

A p-value of 0.0577 is greater than 0.05, meaning we fail to reject the null hypothesis.

This suggests that there is no strong evidence that customer age significantly affects churn risk.

What This Means for Business Decisions:

Age alone is not a strong predictor of churn. This means other factors (e.g., transaction behaviour, credit utilization, engagement with services) might be better indicators of churn.

5. **"How can I optimize retention strategies based on demographic and behavioural data?"**

Analysis of Income Category vs. Churn (Attrition) Risk

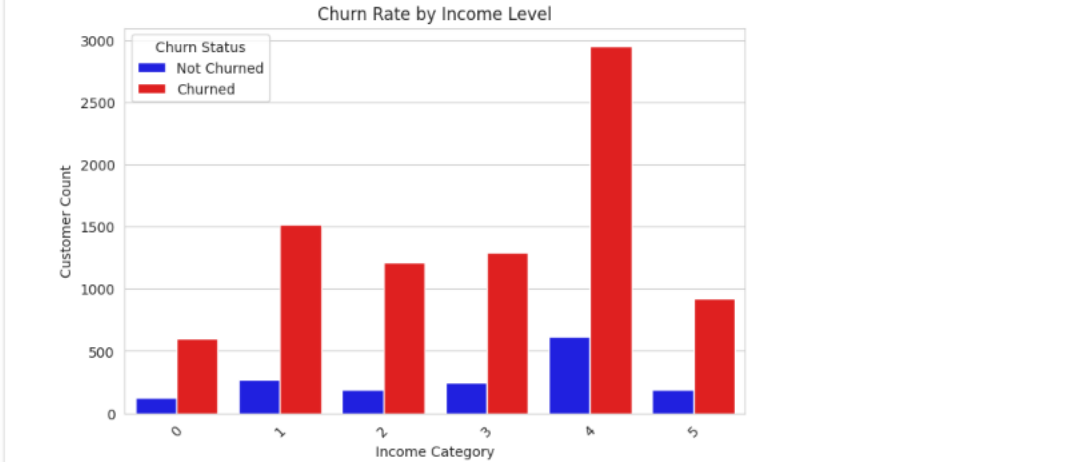
```
Chi-Square Contingency Table:

Attrition_Flag     0     1
Income_Category
0                126   601
1                271  1519
2                189  1213
3                242  1293
4                612  2949
5                187   925

Chi-Square Statistic: 12.8323, P-Value: 0.0250

Expected Frequencies Table:

Attrition_Flag            0           1
Income_Category
0                116.799546  610.200454
1                287.580725 1502.419275
2                225.244791 1176.755209
3                246.612521 1288.387479
4                572.108917 2988.891083
5                178.653501  933.346499
```



29. Chi-square testing on different income category.

Our analysis of customer churn reveals a critical business challenge: a consistently high churn rate across all income categories, with statistically significant variations linked to income levels. The dataset, encompassing 10,127 customers, shows an overall churn rate of approximately 84%, with specific income groups exhibiting rates ranging from 83% to 87%. The Chi-Square test (Chi-Square Statistic: 12.8323, P-Value: 0.0250) confirms a meaningful relationship between income category and churn, indicating that income influences customer retention, albeit subtly. The highest absolute churn occurs among high earners (>$120K), with 2,949 customers leaving, while the highest churn rate (87%) is observed in the $60K-$80K bracket.

The Business Problem

The business faces two interconnected issues:

1. **Widespread Churn**: An alarmingly high churn rate (83-87%) across all income segments suggests systemic dissatisfaction or competitive pressure, threatening long-term revenue and customer loyalty.
2. **Income-Specific Variations**: The statistically significant link between income and churn implies that different income groups may have distinct needs, expectations, or experiences driving their decision to leave. For instance:
    - **Low-Income Customers (<$40K)**: 601 of 727 churned (83%), possibly due to cost sensitivity or limited value perception.
    - **Middle-Income Customers ($60K-$80K)**: 1,213 of 1,402 churned (87%), the highest rate, potentially reflecting unmet expectations or better alternatives.
    - **High-Income Customers (>$120K)**: 2,949 of 3,561 churned (83%), the largest absolute loss, which could signal premium service gaps or lack of tailored offerings.

This dual challenge—high overall churn and income-specific differences—indicates that a one-size-fits-all retention strategy is insufficient, risking further customer loss and reduced profitability.

Business Implications

- **Revenue Impact**: Losing 84% of customers, especially high earners who likely contribute more per capita, erodes revenue and increases acquisition costs to replace them.
- **Market Positioning**: Persistent churn across income levels may weaken brand reputation and competitive standing, particularly if competitors better address segment-specific needs.

- **Opportunity Cost**: Failing to understand and address why churn varies by income (e.g., 87% in $60K-$80K vs. 83% in >$120K) prevents the business from targeting retention efforts effectively.

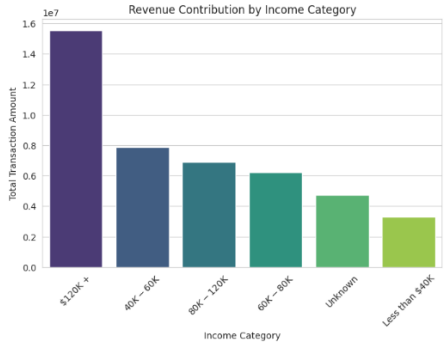
Revenue Contribution by Income Category:

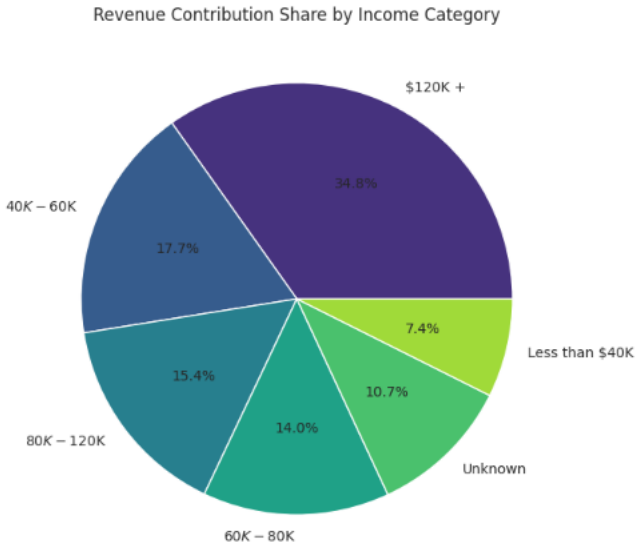
```
Revenue Contribution by Income Category:

  Income_Category  Total_Trans_Amt  Contribution (%)
0        $120K +         15538139         34.838735
1     $40K - $60K          7886058         17.681672
3     $80K - $120K         6881344         15.428959
2     $60K - $80K          6239899         13.990748
5         Unknown          4761819         10.676681
4  Less than $40K          3292923          7.383205
<ipython-input-185-7bef081af6f8>:33: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False

  sns.barplot(data=income_revenue, x="Income_Category", y="Total_Trans_Amt", palette="viridis")
```





30. Revenue by Income category.

Summary:

Customers earning more than $120K contribute the most to revenue, accounting for about 35%. They are followed by those in the $40K–$60K income range, who contribute around 18%, and customers earning between $80K–$120K, who make up roughly 15% of the total revenue. The customers with 40k to 60k contributing minimal with only 18 precent.

High value customer identifications

Step 1: Find the Top Spenders

I looked at the Total_Trans_Amt (total money spent by each customer). I found the top 10% highest spenders using the 90th percentile.

Step 2: Label Customers

I created a new category:

"High-Value" → If their spending is in the top 10%. "Regular" → If they spend less than the top 10%.

Step 3: Compare Spending Patterns

I calculated average spending for both groups. I checked if high-value customers behave differently from regular ones.

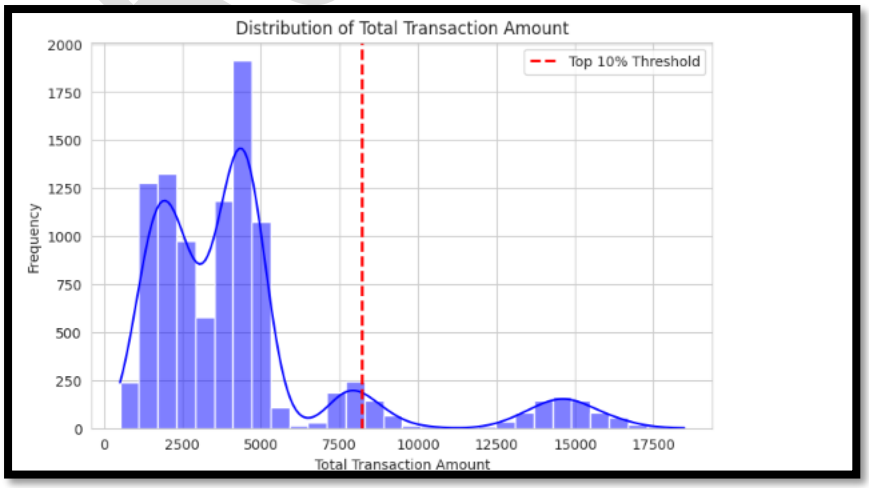
Step 4: Create Graphs to Visualize the Data

Histogram → Shows how customer spending is distributed, with a red line marking the top 10% threshold. Bar Chart → Compares the number of high-value vs. regular customers. Box Plot → Helps see the spending difference between the two groups.

Step 5: Analyse & Take Action

I saw that high-value customers contribute the most to revenue. Businesses should focus on keeping these customers happy by offering:

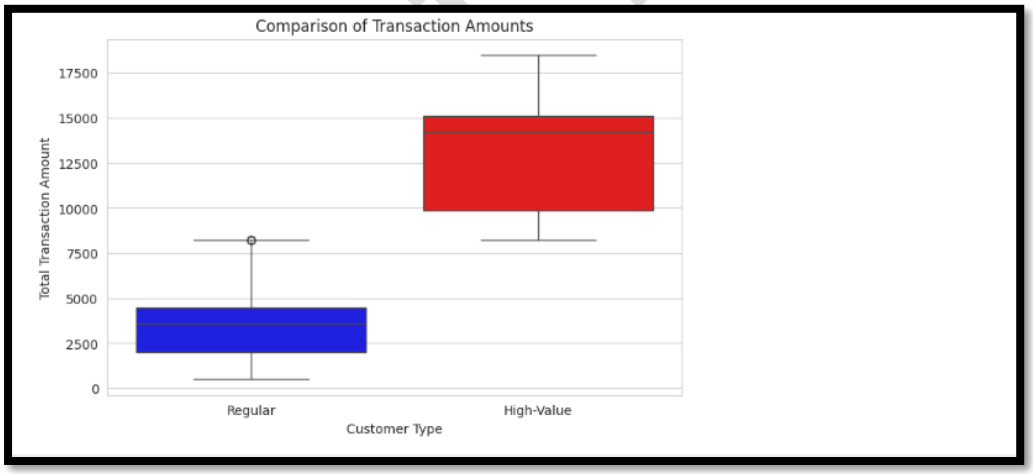
Exclusive rewards & discounts Loyalty programs Personalized offers

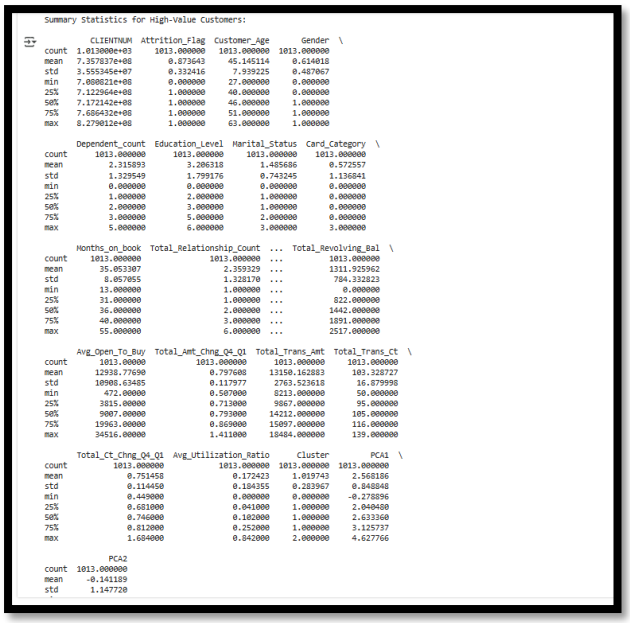


the high-value customers are those after the red line in the histogram. Here's why:

The red line marks the 90th percentile of total spending. Customers above this red line fall into the top 10% of spenders, meaning they spend more than 90% of all other customers. These are the "High-Value" customers because they contribute significantly more to the business's revenue. On the other hand, those before the red line are considered "Regular" customers.

In simple terms:
After the red line = high spenders = high-value customers
Before the red line = lower spenders = regular customers





```
Summary Statistics for High-Value Customers:

           CLIENTNUM  Attrition_Flag  Customer_Age       Gender  \
count  1.013000e+03     1013.000000   1013.000000  1013.000000
mean   7.357837e+08        0.873643     45.145114     0.614018
std    3.555345e+07        0.332416      7.939225     0.487067
min    7.080821e+08        0.000000     27.000000     0.000000
25%    7.122964e+08        1.000000     40.000000     0.000000
50%    7.172142e+08        1.000000     46.000000     1.000000
75%    7.686432e+08        1.000000     51.000000     1.000000
max    8.279012e+08        1.000000     63.000000     1.000000

       Dependent_count  Education_Level  Marital_Status  Card_Category  \
count      1013.000000      1013.000000     1013.000000    1013.000000
mean          2.315893         3.206318        1.485686       0.572557
std           1.329549         1.799176        0.743245       1.136841
min           0.000000         0.000000        0.000000       0.000000
25%           1.000000         2.000000        1.000000       0.000000
50%           2.000000         3.000000        1.000000       0.000000
75%           3.000000         5.000000        2.000000       0.000000
max           5.000000         6.000000        3.000000       3.000000

       Months_on_book  Total_Relationship_Count  ...  Total_Revolving_Bal  \
count     1013.000000               1013.000000  ...          1013.000000
mean        35.053307                  2.359329  ...          1311.925962
std          8.057055                  1.328170  ...           784.332823
min         13.000000                  1.000000  ...             0.000000
25%         31.000000                  1.000000  ...           822.000000
50%         36.000000                  2.000000  ...          1442.000000
75%         40.000000                  3.000000  ...          1891.000000
max         55.000000                  6.000000  ...          2517.000000

       Avg_Open_To_Buy  Total_Amt_Chng_Q4_Q1  Total_Trans_Amt  Total_Trans_Ct  \
count      1013.000000           1013.000000      1013.000000     1013.000000
mean      12938.776900              0.797608     13150.162883      103.328727
std       18908.634852              0.117977      2763.523618       16.879998
min          472.000000              0.507000      8213.000000       50.000000
25%         3815.000000              0.713000      9867.000000       95.000000
50%         9007.000000              0.793000     14212.000000      105.000000
75%        19963.000000              0.869000     15097.000000      116.000000
max        34516.000000              1.411000     18484.000000      139.000000

       Total_Ct_Chng_Q4_Q1  Avg_Utilization_Ratio      Cluster         PCA1  \
count          1013.000000            1013.000000  1013.000000  1013.000000
mean              0.751458               0.172423     1.019743     2.568186
std               0.114450               0.184355     0.283967     0.848848
min               0.449000               0.000000     0.000000    -0.278896
25%               0.681000               0.041000     1.000000     2.040480
50%               0.746000               0.102000     1.000000     2.633360
75%               0.812000               0.252000     1.000000     3.125737
max               1.684000               0.842000     2.000000     4.627766

              PCA2
count  1013.000000
mean     -0.141189
std       1.147720
```

31. Summary statistics and comparison analysis for high and low value customer.

Business Problem:

1. **"Which customer groups should I target for cross-selling or premium offerings to maximize revenue?"**

Customer Spending Differences Using ANOVA and Tukey HSD Test

Objective:

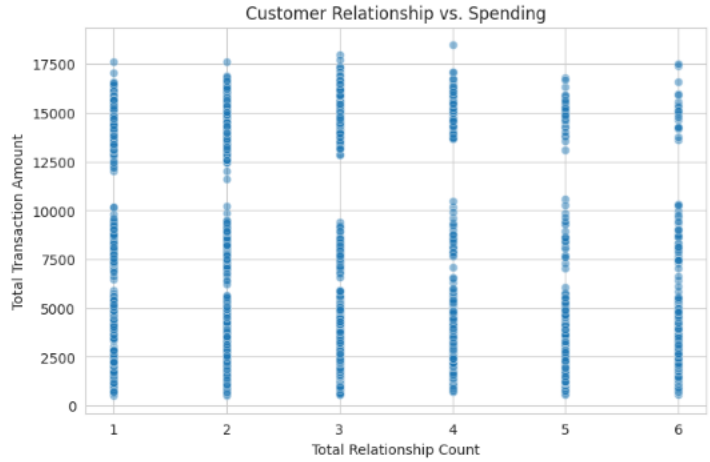
The goal of this analysis is to examine whether customer spending (Total_Trans_Amt) differs based on customer relationship count (Total_Relationship_Count) and determine which customer groups the bank should focus on.

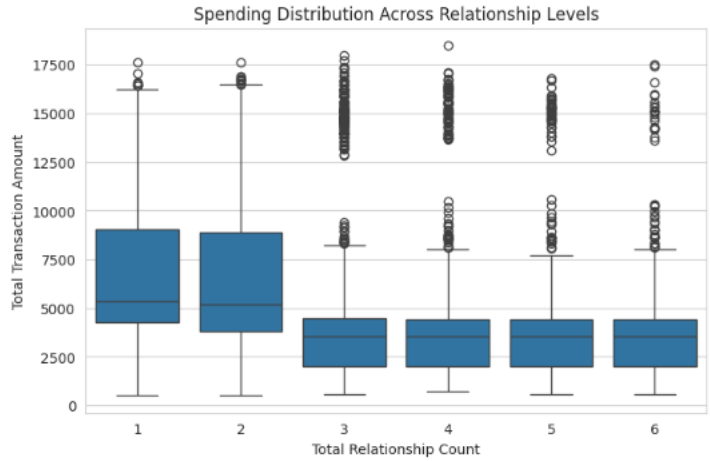
Step 1: ANOVA Test What is ANOVA?

ANOVA (Analysis of Variance) is used to check if there is a significant difference in spending between different relationship groups. It does not tell us which groups are different, just that at least one group differs.

```
       Total_Relationship_Count  Total_Trans_Amt
count                 10127.000000     10127.000000
mean                      3.812580      4404.086304
std                       1.554408      3397.129254
min                       1.000000       510.000000
25%                       3.000000      2155.500000
50%                       4.000000      3899.000000
75%                       5.000000      4741.000000
max                       6.000000     18484.000000

Correlation between Total_Relationship_Count and Total_Trans_Amt: -0.3472
```






```
ANOVA Test Statistic: 430.8849, P-Value: 0.0000
There is a significant difference in spending based on relationship count.
```

32. Anova test to identify the realations between total transaction amount and number of relation customer have with banks.

Results Interpretation:

ANOVA Test Statistic: 430.8849 P-Value: 0.0000

Since the p-value is very small (less than 0.05), we conclude that there is a significant difference in spending between at least two customer groups. However, ANOVA does not tell us which groups differ, so we use Tukey's HSD test to investigate further.

Step 2: Tukey HSD Test

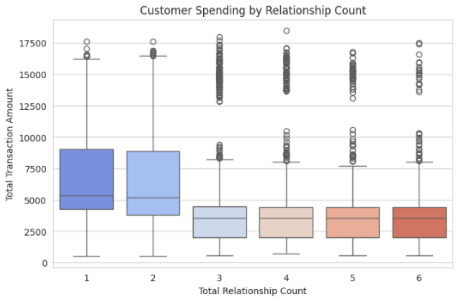
What is Tukey's HSD Test?

Tukey's Honestly Significant Difference (HSD) test compares every possible pair of groups to see where significant differences exist.

```
Tukey HSD Test Results:
     Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2  meandiff  p-adj    lower      upper    reject
-----------------------------------------------------
     1      2 -233.7696 0.5074  -617.4431    149.904  False
     1      3 -3316.6382    0.0 -3660.933 -2972.3434   True
     1      4 -3497.9469    0.0 -3852.1154 -3143.7783  True
     1      5 -3737.585     0.0 -4092.3871 -3382.7828  True
     1      6 -3836.4631    0.0 -4192.0366 -3480.8897  True
     2      3 -3082.8686    0.0 -3392.3375 -2773.3998  True
     2      4 -3264.1773    0.0 -3584.5949 -2943.7598  True
     2      5 -3503.8154    0.0 -3824.9331 -3182.6977  True
     2      6 -3602.6936    0.0 -3924.6634 -3280.7238  True
     3      4  -181.3087 0.4023  -453.3398    90.7224  False
     3      5  -420.9468 0.0002  -693.8022  -148.0913  True
     3      6  -519.8249    0.0  -793.6826  -245.9673  True
     4      5  -239.6381  0.158  -524.8513    45.5751  False
     4      6  -338.5163 0.0098  -624.6884   -52.3441  True
     5      6   -98.8782 0.9238  -385.8341   188.0777  False
-----------------------------------------------------
<ipython-input-188-73094e1bf5ea>:22: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(x=df["Total_Relationship_Count"], y=df["Total_Trans_Amt"], palette="coolwarm")
```



33. Tukey HSD test on relationship counts with total transaction amount.

Overview

The bank needed to determine if customer spending (Total_Trans_Amt) differs by relationship count (Total_Relationship_Count) and identify actionable group differences. Uncertainty about spending patterns risked inefficient resource use and missed revenue opportunities. ANOVA (Statistic: 430.8849, P-Value: 0.0000) confirmed spending varies significantly across groups, but didn't specify *which* groups.

Business Problem

- **Unclear Differences**: Without knowing which relationship groups (1-6) differ in spending, the bank couldn't target strategies effectively.
- **Revenue Risk**: Failing to address low spenders or leverage high spenders could limit profits.

Hypotheses

- **Null Hypothesis (H_o)**: "Average spending is the same across all relationship groups."
- **Alternative Hypothesis (H_1)**: "Average spending differs across at least some relationship groups."
- **ANOVA Result**: P = 0.0000 rejects H_o, supporting H_1—spending varies.

Tukey HSD: Solving the "Which Groups?" Problem

Tukey HSD tested pairwise differences:

- **Pairwise H_o**: "Spending is the same between these two groups (e.g., Group 1 = Group 3)."
- **Pairwise H_1**: "Spending differs between these two groups."

Results Interpretation:

Key Findings

- **Rejected H_o (Significant Differences, reject = True)**:
    - 1-3: Group 3 spends $3316.64 less (p-adj = 0.0).
    - 1-4: Group 4 spends $3497.95 less (p-adj = 0.0).
    - 1-5: Group 5 spends $3737.59 less (p-adj = 0.0).
    - 2-3: Group 3 spends $3082.87 less (p-adj = 0.0).
    - 4-6: Group 6 spends $338.52 less (p-adj = 0.0098).
- **Failed to Reject H_o (No Difference, reject = False)**:
    - 1-2, 3-4, 5-6 (p-adj > 0.05).

What Tukey HSD Solved

- **Identified Key Gaps**: Group 1 (1 relationship) spends the most, significantly more than Groups 3-5. Group 6 outspends Group 4.
- **Resolved Ambiguity**: Pinpointed where H_o fails, showing actionable differences (e.g., 1-5: $3737 gap) vs. similarities (e.g., 5-6: $98, not significant).

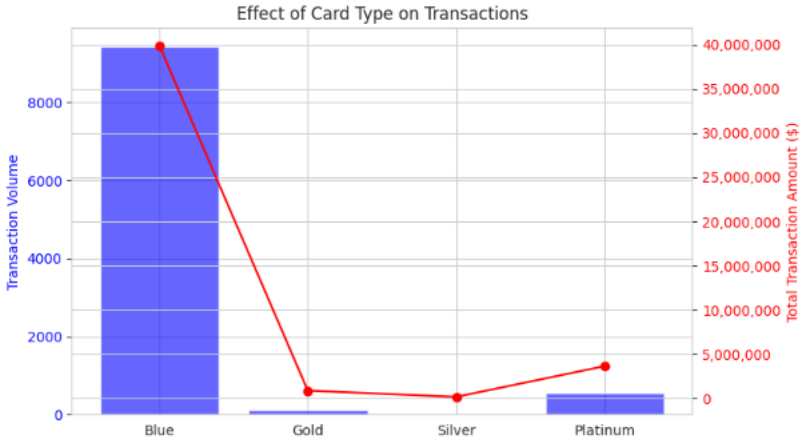
Recommendations

1. **Group 1**: Retain high spenders with single-product perks.
2. **Groups 3-5**: Boost mid-tier spending via cross-selling.
3. **Group 6**: Reward high-relationship spenders to grow this segment.

Conclusion

The problem was unclear spending differences across relationship groups. Tukey HSD solved it by rejecting H_o for 1-3, 1-4, 1-5, 2-3, and 4-6, proving significant gaps (e.g., Group 1 outspends Group 5 by $3737). The bank can now focus on retaining Group 1, lifting Groups 3-5, and rewarding Group 6 to optimize revenue.

Effects of Card type on transactions:



34. Total transaction with different card type

1. Blue cards drive the highest transaction volume (9,436 transactions) and contribute the most to total revenue.
2. Silver and Gold cards have the highest average transaction values (8,999 & $7,685), but they are rarely used.
3. Platinum cards balance between volume and value, making them valuable for premium customers.

To maximize revenue, increase high-value transactions for Gold & Platinum users while maintaining blue card usage.

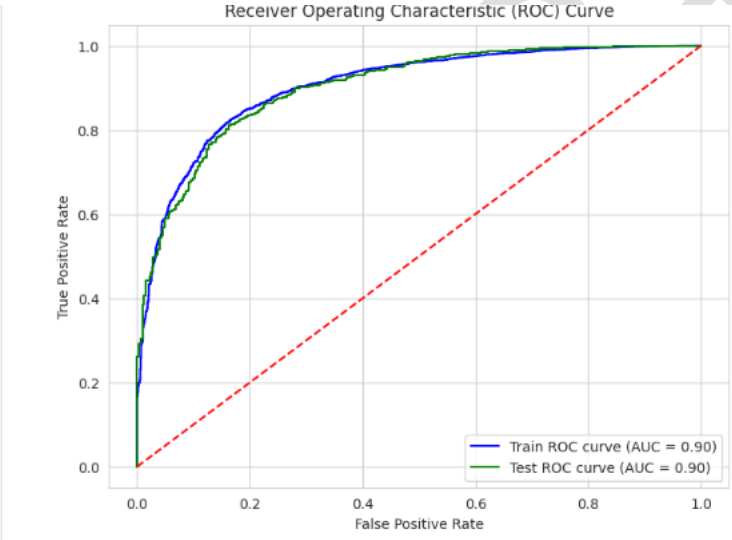
2. Why are customer leaving, what factor contribution to attrition

Logistic Regression:

The Logit Regression model analysed why customers leave (Attrition_Flag) using data from 10,127 customers. It finished successfully after 8 iterations with a function value of 0.260572. The model explains 41% of customer attrition (Pseudo R-squared: 0.4088) and is reliable (converged: True, LLR p-value: 0.000).

```
Optimization terminated successfully.
         Current function value: 0.260572
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:         Attrition_Flag   No. Observations:                10127
Model:                          Logit   Df Residuals:                    10115
Method:                           MLE   Df Model:                           11
Date:                Sat, 22 Mar 2025   Pseudo R-squ.:                 0.4088
Time:                        15:48:52   Log-Likelihood:                -2638.8
converged:                       True   LL-Null:                       -4463.6
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         -3.2267      0.308    -10.492      0.000      -3.829      -2.624
Customer_Age                   0.0052      0.007      0.735      0.462      -0.009       0.019
Dependent_count               -0.1401      0.028     -5.020      0.000      -0.195      -0.085
Months_on_book                 0.0038      0.007      0.537      0.592      -0.010       0.018
Total_Relationship_Count       0.4696      0.026     18.136      0.000       0.419       0.520
Months_Inactive_12_mon        -0.4877      0.036    -13.628      0.000      -0.558      -0.418
Contacts_Count_12_mon         -0.5153      0.034    -15.142      0.000      -0.582      -0.449
Credit_Limit               1.362e-05   4.92e-06      2.771      0.006    3.99e-06    2.33e-05
Total_Revolving_Bal            0.0010   6.77e-05     15.120      0.000       0.001       0.001
Total_Trans_Amt               -0.0004   2.14e-05    -19.514      0.000      -0.000      -0.000
Total_Trans_Ct                 0.1106      0.003     32.999      0.000       0.104       0.117
Avg_Utilization_Ratio         -0.0301      0.230     -0.131      0.896      -0.481       0.421
==============================================================================================

Odds Ratios:
 const                      0.039687
Customer_Age                1.005215
Dependent_count             0.869250
Months_on_book              1.003845
Total_Relationship_Count    1.599364
Months_Inactive_12_mon      0.614050
Contacts_Count_12_mon       0.597303
Credit_Limit                1.000014
Total_Revolving_Bal         1.001025
Total_Trans_Amt             0.999582
Total_Trans_Ct              1.116902
Avg_Utilization_Ratio       0.970309
dtype: float64
```




```
Summary Table:
       Metric  Train Set  Test Set
0    Accuracy   0.886187  0.890918
1   Precision   0.905590  0.908840
2      Recall   0.964995  0.967078
3    F1 Score   0.934349  0.937055
4     ROC AUC   0.903709  0.901367

Confusion Matrix (Train):
 [[ 618  684]
 [ 238 6561]]

Confusion Matrix (Test):
 [[ 160  165]
 [  56 1645]]
```

Variables and Their Impact on Attrition_Flag:

- **Total_Relationship_Count (coef: 0.4696)**: More products/services strongly increase staying (positive effect).

- **Total_Trans_Ct (coef: 0.1106)**: Higher transaction counts boost staying (positive effect).
- **Months_Inactive_12_mon (coef: -0.4877)**: More inactive months strongly increase leaving (negative effect).
- **Contacts_Count_12_mon (coef: -0.5153)**: More contacts strongly raise leaving (negative effect).
- **Total_Trans_Amt (coef: -0.0004)**: Higher transaction amounts slightly increase leaving (negative effect).
- **Dependent_count (coef: -0.1401)**: More dependents slightly reduce leaving (negative effect).
- **Credit_Limit (coef: 0.0000136)**: Higher limits slightly increase staying (small positive effect).
- **Total_Revolving_Bal (coef: 0.0010)**: Higher revolving balances slightly boost staying (positive effect).
- **Customer_Age (coef: 0.0052)**: Almost no effect on leaving or staying.
- **Months_on_book (coef: 0.0038)**: Time as a customer has little impact.
- **Avg_Utilization_Ratio (coef: -0.0301)**: Credit usage barely affects leaving or staying.

Key Insights:
Strongest factors for staying are more relationships and transactions. Inactivity and frequent contacts push customers to leave the most. Age, time as a customer, and credit usage don't matter much. Other factors like dependents, credit limit, and revolving balance have smaller roles.

Summary:

The Attrition Analysis investigates customer churn using a dataset of 10,127 customers, focusing on why customers leave, how to retain valuable segments, and how credit card usage impacts churn and revenue. The analysis employs a variety of statistical and machine learning techniques, including Exploratory Data Analysis (EDA), Logistic Regression, Random Forest, K-Means Clustering, ANOVA, and Chi-Square tests, to address five key business questions. Below is a concise summary of the findings and insights.

Dataset Overview

- Size and Scope: 10,127 customers with no missing values, tracked via unique CLIENTNUM.

- Key Variables: Includes Attrition_Flag (churn status), demographic data (age, gender, income, education), behavioral data (inactivity, contacts), and credit usage metrics (credit limit, transactions, utilization).
- Descriptive Insights:
    - Average customer age: 46.3 years (range: 26–73).
    - Credit limits range from $1,438 to $34,516 (mean: $8,632).
    - Median transaction amount: $3,899; median utilization ratio: 17.6%.
    - Customers average 35.9 months with the bank and hold 3.8 products.

Key Business Questions and Findings

1. **Why Are Customers Leaving, and What Factors Contribute Most to Churn?**
    - Methods: EDA, Logistic Regression (41% variance explained), Random Forest (94.67% accuracy), Chi-Square, T-tests.
    - Key Drivers:
        - Inactivity (Months_Inactive_12_mon, coef: -0.4877) and frequent contacts (Contacts_Count_12_mon, coef: -0.5153) strongly increase churn, reflecting disengagement or dissatisfaction.
        - Low engagement (Total_Relationship_Count, coef: 0.4696; Total_Trans_Ct, coef: 0.1106) heightens churn risk.
    - Minor Factors: Age (p = 0.0577), tenure, and utilization ratio have little impact.
    - Income Effect: Churn rates range from 83% (<$40K) to 87% ($60K–$80K), with statistical significance (Chi-Square p = 0.0250).
    - Insight: High churn (84% overall) suggests systemic issues; inactivity and poor service experiences are critical triggers.
2. **Which Customer Segments Are Most Valuable, and How Can We Retain Them?**
    - Methods: K-Means Clustering (K=3), percentile analysis for high-value customers, revenue by income.
    - Segments:
        - Cluster 0 (Moderate Spenders): 7,547 customers, medium spenders with medium credit limits.

        - Cluster 1 (High Spenders): 1,184 customers, high transaction amounts, medium credit, revenue drivers.
        - Cluster 2 (Low Spenders, High Credit): 1,396 customers, conservative spenders with high limits.
    - High-Value Customers: Top 10% spenders (above 90th percentile of Total_Trans_Amt) contribute disproportionately to revenue.
    - Revenue Leaders: >$120K income group (35%), followed by $40K–$60K (18%) and $80K–$120K (15%).
    - **Insight: High spenders and high-income customers are most valuable; low spenders with high credit are untapped potential.**

3. **How Do Credit Card Usage Patterns Affect Churn and Revenue?**
    - Methods: Correlation analysis (-0.4829 between credit limit and utilization), Random Forest feature importance, card type analysis.
    - Findings:
        - Higher credit limits correlate with lower utilization (responsible borrowing); low limits link to higher utilization (financial strain).
        - Blue cards lead in volume (9,436 transactions), while Silver/Gold have higher average values ($8,999/$7,685) but low usage.
        - Low transaction activity increases churn; higher revolving balances slightly reduce it.
    - **Insight: Usage patterns signal churn risk and revenue potential; premium cards underperform in volume.**

4. **How Can We Optimize Retention Strategies Based on Demographic and Behavioral Data?**
    - Methods: EDA, boxplots, Chi-Square (income vs. churn), bar graphs.
    - Findings:
        - Demographics: Income impacts churn (highest at $60K–$80K, 87%), age does not ($p > 0.05$); females show higher utilization (34% vs. 9% for males).
        - Behavior: Inactive customers (>6 months) are mostly low-income blue cardholders; frequent contacts predict churn.
    - Insight: Tailored strategies are needed—low-income customers require affordability, inactive ones need re-engagement, and females may benefit -

Solution: Offer affordable bundles for low-income groups, re-engage inactive customers with campaigns, and provide flexible plans for high-utilization females.

5. **Which Customer Groups Should We Target for Cross-Selling or Premium Offerings?**
    - Methods: ANOVA (p = 0.000), Tukey HSD, clustering, card type analysis.
    - Findings:
        - Customers with more relationships (Groups 5–6) spend significantly more than those with fewer (Groups 1–2).
        - Platinum/Gold cards yield high-value transactions; Blue dominates volume.
        - High (Cluster 1) and moderate spenders (Cluster 0) show cross-sell potential.
    - Insight: Target low-relationship customers (Groups 1–2) for cross-selling and high spenders (Cluster 1, Groups 5–6) for premium upgrades.

Key Insights and Recommendations

- Churn Drivers: Inactivity, frequent contacts, and low engagement are the top reasons for churn. Address via re-engagement campaigns, improved service, and product bundling.
- Valuable Segments: High spenders (Cluster 1) and >$120K income customers drive revenue. Retain with loyalty programs and premium offerings.
- Credit Usage: Low utilization with high limits signals responsible borrowers (prime for premium products); high utilization with low limits indicates risk. Adjust credit limits accordingly.
- Retention Optimization: Tailor strategies by income (affordability for low-income), inactivity (re-engage), and gender (flexible plans for females).
- Cross-Selling/Premium Targets: Focus on low-relationship and high-spending groups to boost engagement and revenue.

Business Implications

- Revenue Protection: High churn (84%) and loss of high earners threaten profitability. Prioritize retention of top spenders and high-income groups.
- Strategic Focus: Use segment-specific insights to reduce churn, increase spending, and optimize card usage, ensuring competitive positioning and long-term growth.

This analysis provides actionable strategies to mitigate churn, retain valuable customers, and maximize revenue for banks credit card business