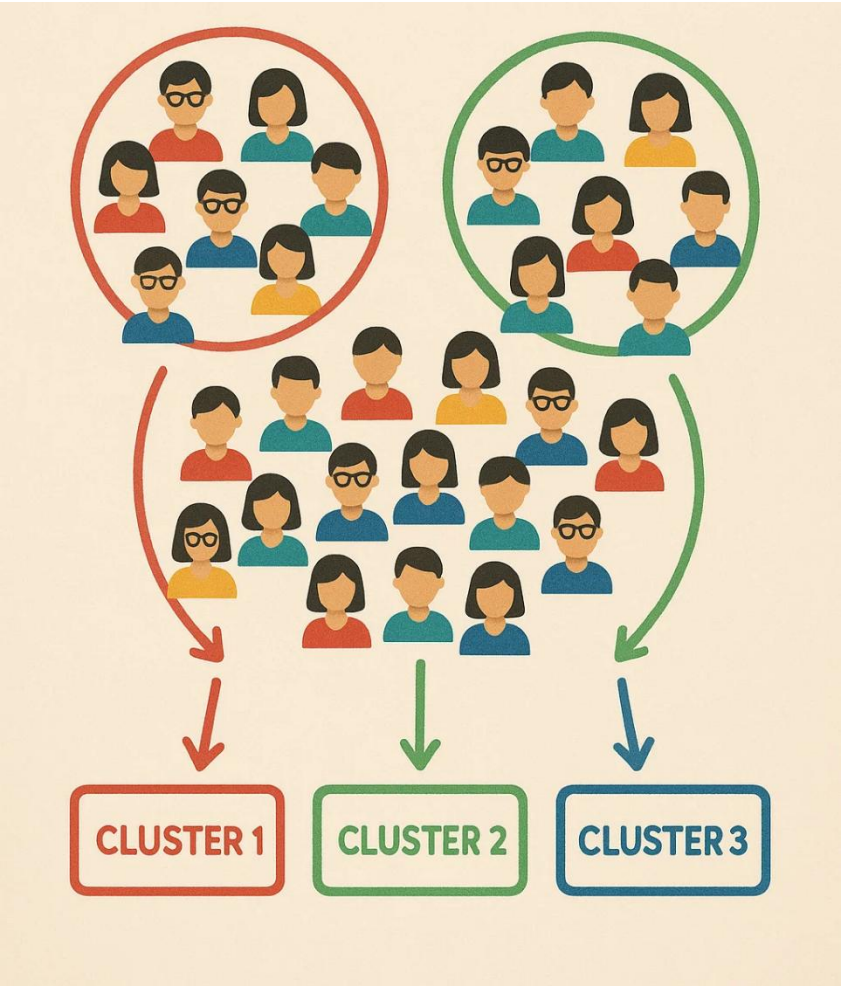
Customer Segmentation for Credit Card Customers



Machine Learning 2 – Final Project

Hend Khaled Abdelhamid Mohamed M. Aly

Table of Contents

1.Executive Summary

This project focused on segmenting credit card customers to better understand diverse spending behaviors and support data-driven business strategies. Using clustering approaches, the objective was to group customers into meaningful segments based on features such as balance, purchase patterns, payment behavior, and cash advance usage.

After thorough data exploration and preprocessing, four distinct customer clusters were identified:

1. **Cash Advance Dependent Users**
   Low purchasing activity but heavy reliance on cash advances. Weak repayment behavior indicates higher financial risk and a potential need for credit monitoring or financial support.
2. **Moderate Spenders with Balanced Use**
   Display moderate, well-balanced spending across one-off and installment purchases, with consistent payment habits. A stable segment with strong potential for product upselling.
3. **High-Spending Full Payers**
   The most profitable segment, characterized by high one-off spending, minimal cash advance usage, and excellent repayment discipline. Ideal candidates for premium products and loyalty incentives.
4. **Installment-Focused Budget Users**
   Prefer installment purchases and avoid cash advances. Reliable payers with lower credit limits, making them suitable for installment-based promotions and long-term retention strategies.

These insights are valuable for tailoring marketing campaigns, personalizing offers, and improving credit risk management. Key strategic recommendations include:

- Incentivize full payers with premium rewards or loyalty programs
- Offer installment-based promotions for budget-conscious users

- Monitor and support cash advance-reliant users to mitigate risk
- Launch cross-sell campaigns targeting balanced users with growth potential

Overall, this segmentation enables the business to shift from a one-size-fits-all approach to a more targeted, data-driven strategy for customer engagement and financial product development.

2.Introduction

2.1. Background

Customer segmentation enables financial institutions to analyze customer behavior and design targeted strategies. In credit card usage, clustering customers based on transaction patterns, payment habits, and credit usage provides actionable insights for personalization and risk management.

2.2. Objective

The goal of this project is to apply unsupervised machine learning techniques to segment credit card customers into distinct groups. This involves:

- Preprocessing and analyzing customer behavioral data
- Identifying the optimal number of clusters
- Interpreting and profiling each segment
- Deriving business insights from clustering results

2.3. Dataset Overview

- Source: customer-segmentation-credit-cards
- Records & Features: 8,950 entries, 18 numerical features

3.Data Exploration and Preprocessing

3.1. Initial Observations

3.1.1. Features

- **Account Status and Limits:**
  - BALANCE: Remaining account balance
  - CREDIT_LIMIT: Maximum credit assigned
  - BALANCE_FREQUENCY: Frequency of balance updates (0 to 1)
- **Purchase Behavior:**
  - PURCHASES: Total purchase amount
  - ONEOFF_PURCHASES: Amount from one-time transactions
  - INSTALLMENTS_PURCHASES: Amount from installment-based purchases
  - PURCHASES_FREQUENCY: Frequency of any purchase activity
  - ONEOFF_PURCHASES_FREQUENCY: Frequency of one-off purchases
  - PURCHASES_INSTALLMENTS_FREQUENCY: Frequency of installment-based purchases
  - PURCHASES_TRX: Number of purchase transactions
- **Cash Advance Behavior:**
  - CASH_ADVANCE: Amount taken as cash in advance
  - CASH_ADVANCE_FREQUENCY: Frequency of cash advances
  - CASH_ADVANCE_TRX: Number of cash advance transactions
- **Repayment & Tenure:**
  - PAYMENTS: Total amount paid
  - MINIMUM_PAYMENTS: Minimum required payment made
  - PRC_FULL_PAYMENT: Percentage of times full balance was paid
  - TENURE: Length of relationship with the card issuer (months)

3.1.2. Observations

- 1 null at CREDIT_LIMIT column and 313 nulls MINIMUM_PAYMENTS column.
- Skewed distributions in the following columns: BALANCE, PURCHASES, ONEOFF_PURCHASES, INSTALLMENTS_PURCHASES, CASH_ADVANCE, CASH_ADVANCE_TRX, PURCHASES_TRX, CREDIT_LIMIT, PAYMENTS, and MINIMUM_PAYMENTS, this indicates high variance in spending patterns.

- Some features contain zero values where they logically could be absent (e.g. ONOFF_PURCHASES = 0 for installment-only customers), but not necessarily anomalies.

3.2. Data Cleaning

- **Handling Missing Values:** filled them with median due to skewness in the data
- **Handling of Skewness and Zeros:** Applied log transformation to skewed features to stabilize variance.
- **Outlier Treatment (Capping):** applied percentile-based (1st and 99th percentiles) to suppress the effect of extreme values without dropping data.
- **Scaling:** all numerical features were standardized using StandardScaler to ensure uniformity in range and influence during clustering.

3.3. Feature Engineering

Derived new features to capture key customer behaviors:

- **`PAYMENT_RATIO = PAYMENTS / MINIMUM_PAYMENTS`**

Indicates how much a customer pays relative to the minimum required, reflecting payment discipline.

- **`ONEOFF_PURCHASE_RATIO = ONEOFF_PURCHASES / PURCHASES`**

Shows the proportion of total purchases made as one-time payments, indicating spending style.

- **`INSTALLMENT_PURCHASE_RATIO = INSTALLMENTS_PURCHASES / PURCHASES`**

Reflects the share of purchases made in installments, suggesting reliance on deferred payment.

- **CASH_ADVANCE_RATIO = CASH_ADVANCE / BALANCE**

Reveals the extent to which the credit limit is used for cash withdrawals, often signaling financial stress.

- **UTILIZATION_RATIO = BALANCE / CREDIT_LIMIT**

Measures how much of the available credit a customer is using, highlighting credit dependency.

- **AVG_PURCHASE_TRX = PURCHASES / PURCHASES_TRX**

Represents the average value of each purchase transaction, differentiating small vs. large spenders.

- **AVG_CASHADVANCE_TRX = CASH_ADVANCE / CASH_ADVANCE_TRX**

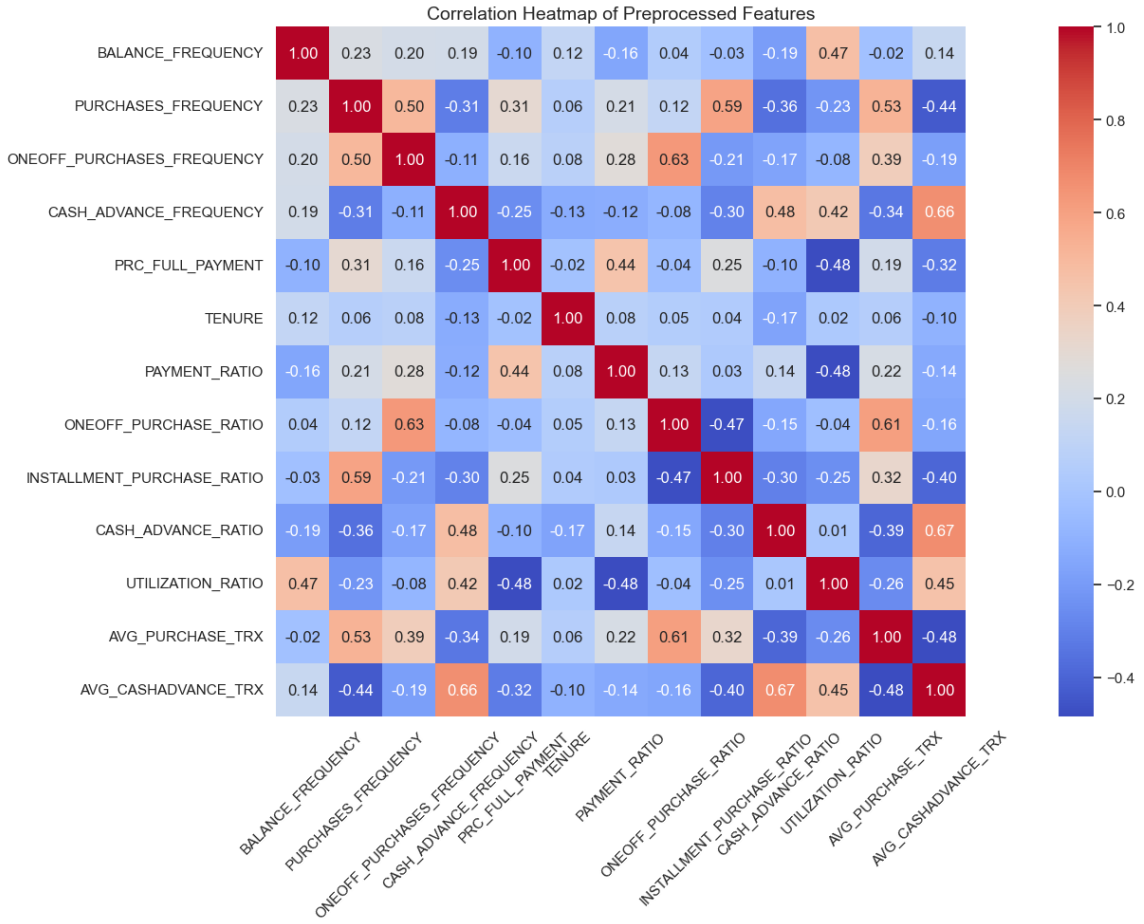
Indicates the average cash amount withdrawn per transaction, capturing typical cash advance behavior.

3.4. Insights and Visualization

After finishing data cleaning and preprocessing, visualization is mandatory to make sure the data is ready for determining the optimal number of clusters, this is done using the following visuals:

- **Correlation Heatmap:**

To identify relationships between features and assess multicollinearity before clustering.

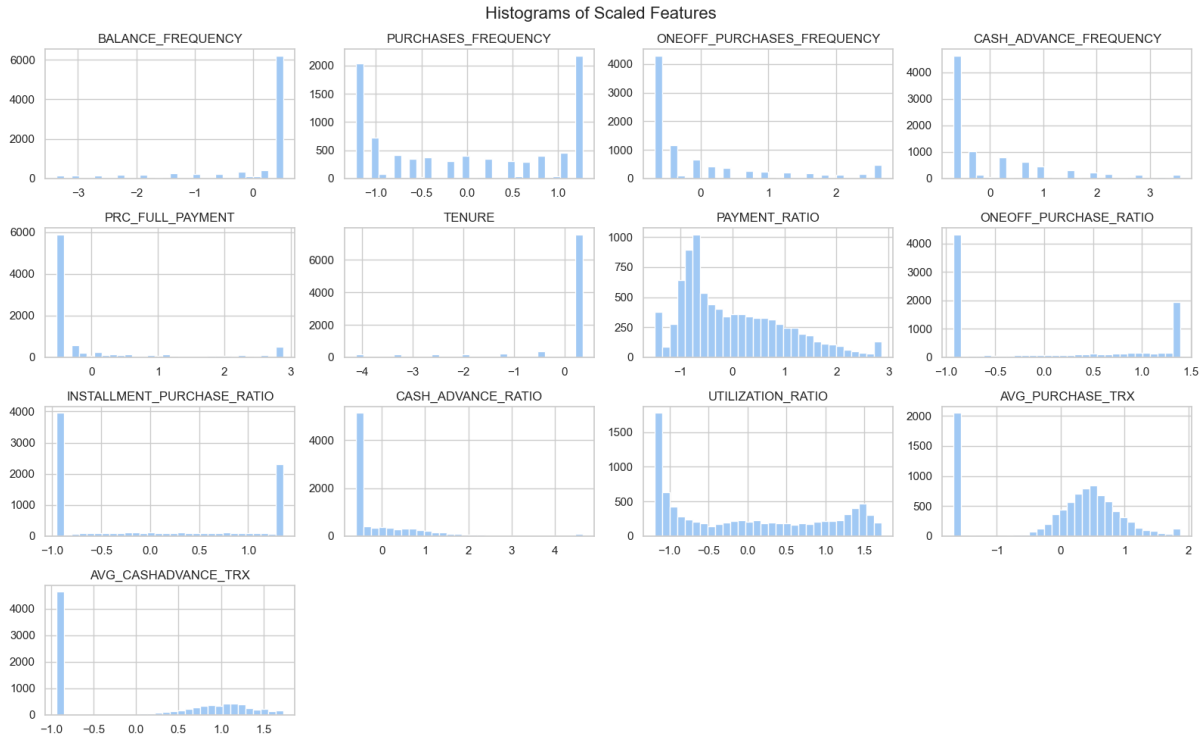


Observations:

- Moderate to strong correlations exist between features like PURCHASES_FREQUENCY, ONEOFF_PURCHASES_FREQUENCY, and AVG_PURCHASE_TRX, suggesting consistent spending behavior among certain customer groups.
- UTILIZATION_RATIO and CASH_ADVANCE_RATIO are negatively correlated with repayment features (PRC_FULL_PAYMENT, PAYMENT_RATIO), indicating potential riskier financial behavior.
- No extreme multicollinearity is observed, confirming that all selected features contribute uniquely to the clustering process.

- **Histograms:**

To examine the distribution of customer behavior features after data transformation and scaling.



Observations:

- Most features, such as PAYMENT_RATIO, UTILIZATION_RATIO, and AVG_PURCHASE_TRX, exhibit a more normalized distribution after log transformation and standardization.
- Despite standardization, some features still show high variance, indicating the presence of diverse spending and repayment behaviors across customers.
- Distributions suggest the dataset likely includes distinct behavioral segments, such as low-activity users vs. high-usage customers.

4. Determining Optimal Clusters

To determine the best number of clusters, a multi-step approach was applied:

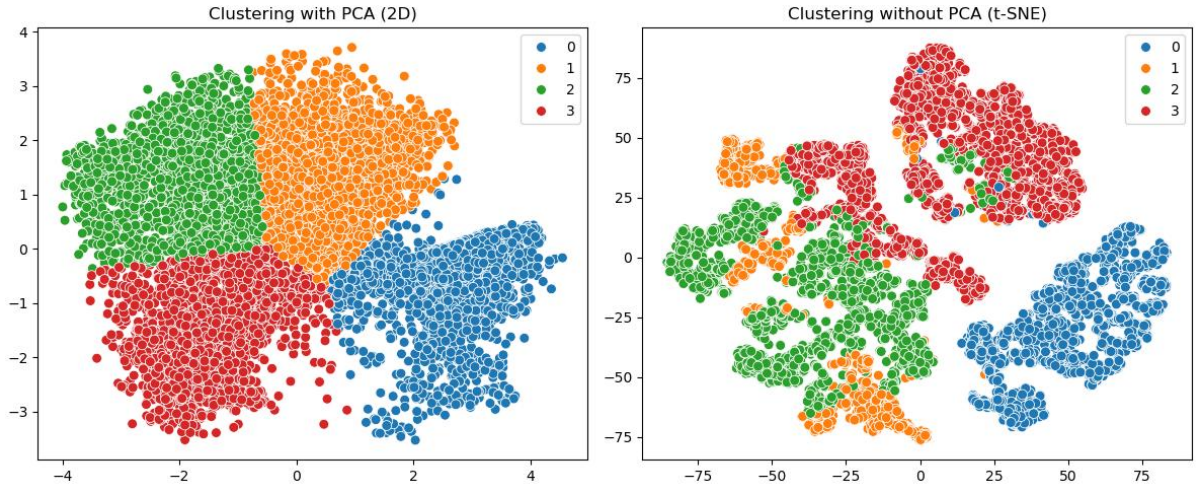
- **Initial Trial and Error:**

A few values of *k* were tested on the original scaled data using different clustering visualizations (with and without dimensionality reduction) to understand how well-separated the clusters appeared.

- **Dimensionality Reduction (PCA and t-SNE):**

Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were used to reduce dimensionality and visualize clustering quality.

- Clustering with PCA provided clearer and more compact groupings compared to t-SNE.
- As a result, PCA-transformed data (X_pca) was chosen as the basis for the final clustering.

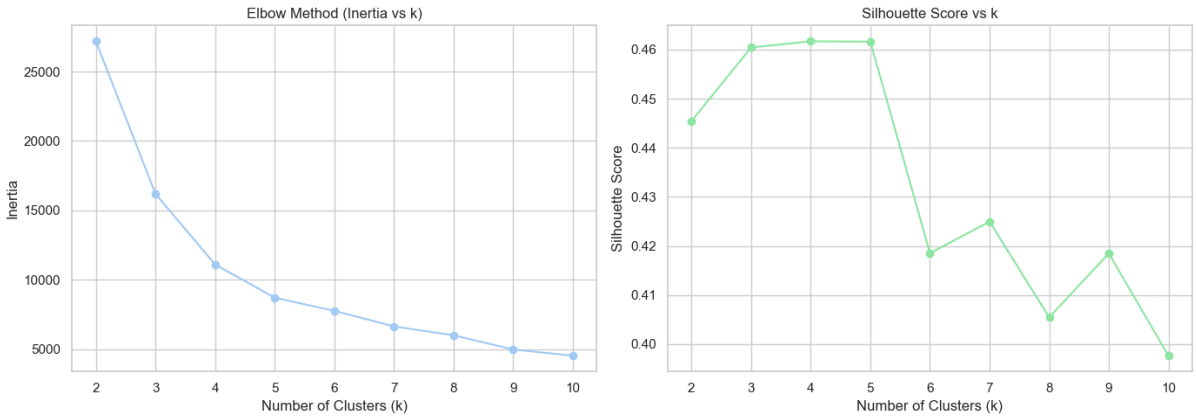


- **Elbow Method:**

The Within-Cluster Sum of Squares (WCSS) was plotted for different values of k. The "elbow" point, where WCSS reduction slows indicated a suitable number of clusters.

- **Silhouette Score:**

The silhouette score was computed for a range of cluster values to assess cluster cohesion and separation. The optimal number corresponded to the highest or a plateauing silhouette score.



- **Observations:**

Based on PCA, the Elbow Method, and Silhouette Scores, **k = 4** clusters were selected as optimal.

This approach ensures that the clusters:

- Are visually well-separated in PCA space,
- Have good internal cohesion and external separation,
- Are interpretable from a business perspective.

This combination of dimensionality reduction and clustering evaluation provided a robust foundation for effective customer segmentation.

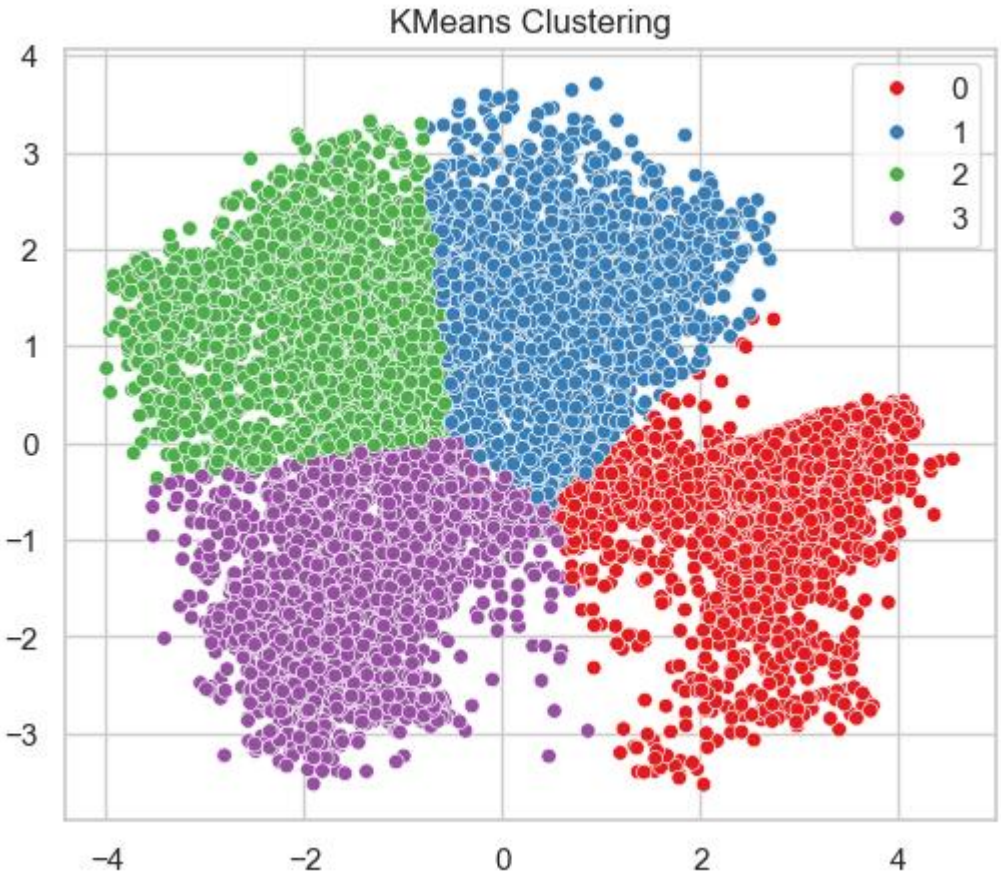
5. Customer Segmentation (Clustering)

5.1. Clustering Techniques:

To identify the most suitable clustering approach for customer segmentation, we experimented with three popular clustering algorithms, all configured to generate the same number of clusters.

5.1.1. Kmeans

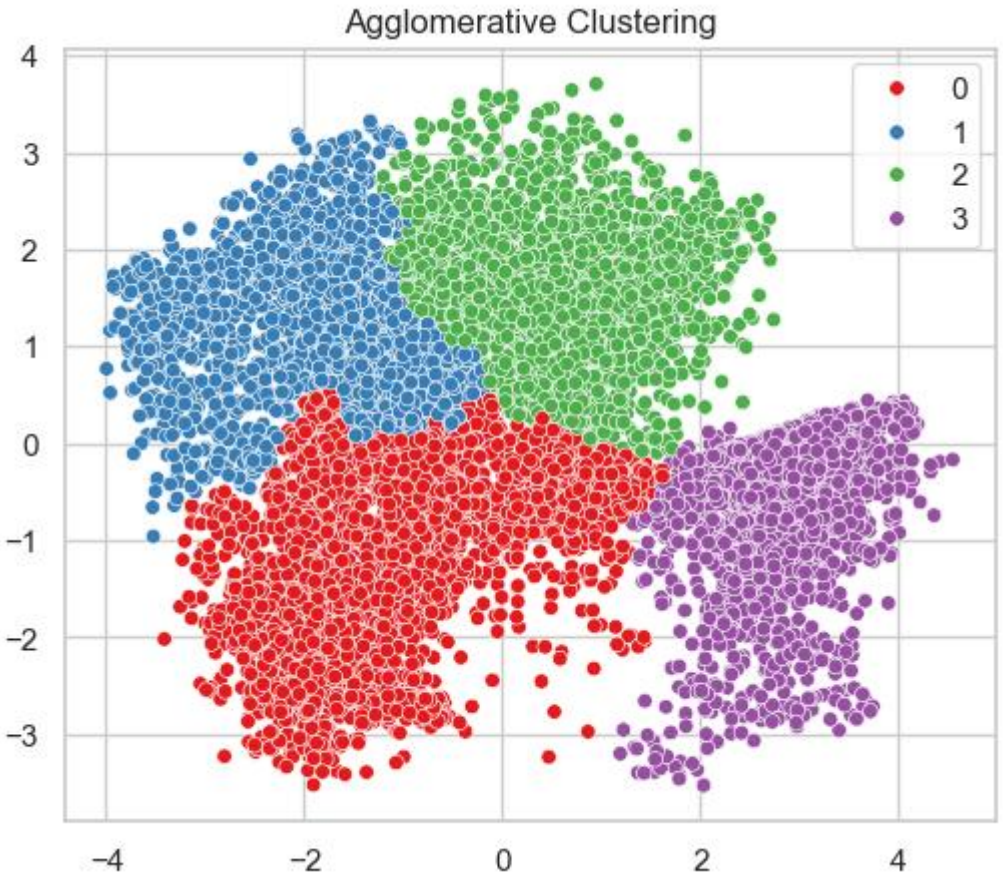
Fast and widely used clustering method that assigns each data point to the nearest centroid. It works well with well-separated, spherical clusters. In this project, it produced clearly defined customer groups with good compactness, making it a strong option despite being sensitive to outliers.



Silhouette Score: 0.4617

5.1.2. Agglomerative Clustering

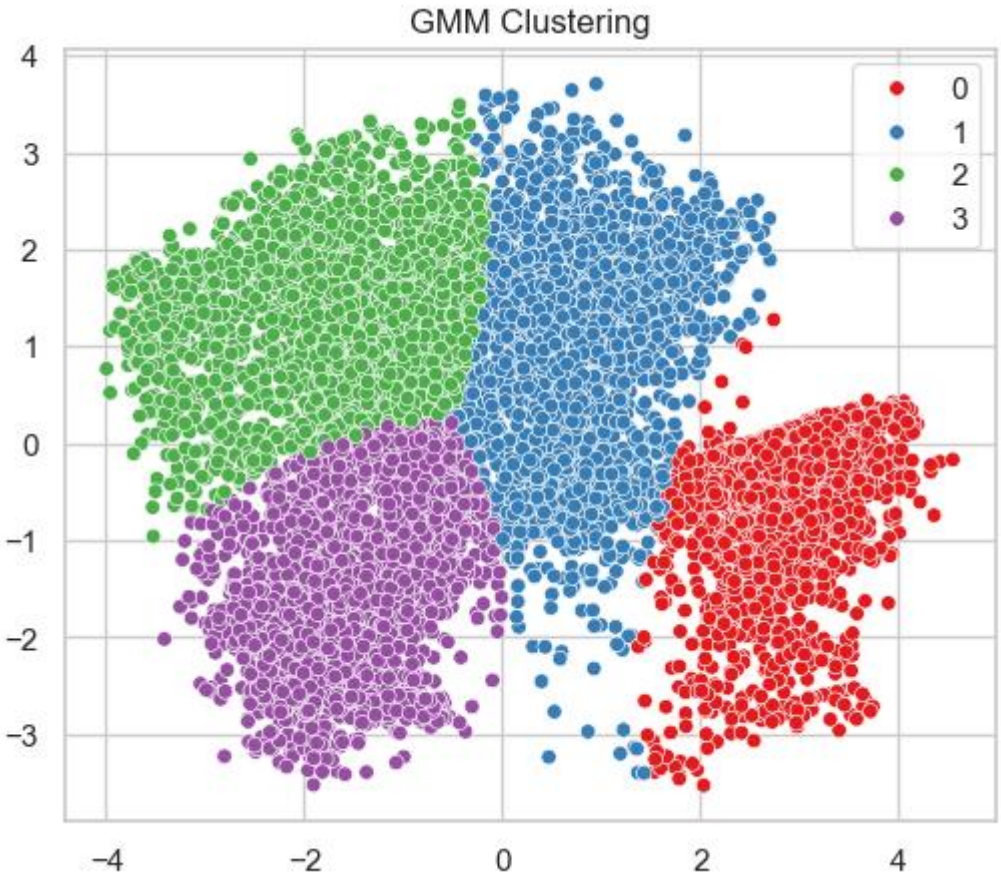
Agglomerative Clustering builds clusters by progressively merging the closest pairs. It's useful for identifying hierarchical relationships but is less scalable for larger datasets. In this analysis, it showed reasonable results but lacked the cluster clarity observed with KMeans.



Silhouette Score: 0.4349

5.1.3. Gaussian Mixture Model (GMM)

GMM models data as a mixture of Gaussian distributions, allowing soft cluster memberships. It handles overlapping and non-spherical clusters well. In this case, it offered more nuanced segmentation but was more computationally intensive and less interpretable than KMeans.



Silhouette Score: 0.4453

5.1.4. Observation

After evaluating all three clustering techniques using the Silhouette Score, **KMeans** was selected as the final model. It achieved the highest silhouette score among the methods tested and offered a good balance of simplicity, scalability, and well-defined customer clusters, making it the most suitable choice for our segmentation task.

5.2. Cluster Summary Table

The following table is cluster summary table is computed by applying the clusters found to the whole dataset, then grouping the rows by cluster number to get the mean of each feature

And this helped in getting the cluster profiles that is discussed in the next section

	Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3
0	BALANCE	2229.381678	2656.881903	1201.494775	400.808199
1	BALANCE_FREQUENCY	0.887589	0.966731	0.922907	0.765323
2	PURCHASES	52.402291	1044.011586	2541.059208	674.437577
3	ONEOFF_PURCHASES	2.839068	698.090446	1873.344090	77.942904
4	INSTALLMENTS_PURCHASES	49.724486	346.130664	667.739829	597.201708
5	CASH_ADVANCE	2123.156958	1775.758192	36.946229	39.810429
6	PURCHASES_FREQUENCY	0.077794	0.493152	0.740505	0.677173
7	ONEOFF_PURCHASES_FREQUENCY	0.002920	0.283989	0.572971	0.041621
8	PURCHASES_INSTALLMENTS_FREQUENCY	0.068853	0.300858	0.437384	0.629151
9	CASH_ADVANCE_FREQUENCY	0.280522	0.253628	0.009543	0.007224
10	CASH_ADVANCE_TRX	6.643395	6.355117	0.154312	0.128462
11	PURCHASES_TRX	1.392140	16.179838	31.060514	13.381154
12	CREDIT_LIMIT	4115.746934	4697.773326	5964.336405	3567.414549
13	PAYMENTS	1712.657480	2061.991681	2422.342923	977.336211
14	MINIMUM_PAYMENTS	1021.107204	1303.029547	510.437800	590.410823
15	PRC_FULL_PAYMENT	0.042631	0.025977	0.227123	0.296965
16	TENURE	11.266722	11.553698	11.777105	11.524231
17	PAYMENT_RATIO	9.173954	2.961414	8.655738	13.892212
18	ONEOFF_PURCHASE_RATIO	0.013782	0.762208	0.778885	0.121682
19	INSTALLMENT_PURCHASE_RATIO	3635.166012	0.238684	0.221122	0.880080
20	CASH_ADVANCE_RATIO	67896.064770	5235.714258	0.085284	0.296296
21	UTILIZATION_RATIO	0.577066	0.621464	0.230973	0.159759
22	AVG_PURCHASE_TRX	6.928913	109.436198	141.978467	30293.459974
23	AVG_CASHADVANCE_TRX	489.924976	306.467250	19.855969	20.485820

5.3. Cluster Profiles

From the cluster summary table, we can describe the cluster profiles as follows:

- **Cluster 0: Cash Advance Dependent Users**

Description:

These customers rarely use their cards for purchases but take a lot of cash advances and then only repay small portions of their balance. This pattern suggests financial stress or risky credit behavior.

Key Traits:

- Low Purchase Volume: ~52 transactions total
- High Cash Advance Usage: ~6.6 cash-advance transactions, average advance amount ≈ $2,123
- Poor Repayment: Only ~4 % full-payment rate
- Utilization Ratio: Moderate (≈ 0.58)

- **Cluster 1: Moderate Spenders with Balanced Use**

Description:

They mix one-off and installment purchases at a steady pace. Payments aren't always in full, but their activity is consistent and balanced.

Key Traits:

- Medium Purchase Volume: ~1,044 total transactions
- Split Spend: ~698 one-off vs. ~346 installment purchases
- Cash Advances: ~6.4 transactions, avg ≈ $1,776
- Full-Payment Rate: ≈ 2.6 %

- **Cluster 2: High-Spending Full Payers**

Description:

This is your top “ideal” segment—big spenders who almost never use cash advances and almost always pay in full.

Key Traits:

- High Purchase Volume: ~2,541 transactions (≈ 1,873 one-off)
- Minimal Cash Advances: ~0.15 transaction frequency, avg ≈ $37
- Strong Repayment: ~22.7 % full-payment rate
- High Credit Limits: ≈ $5,964

- **Cluster 3: Installment-Focused Budget Users**

Description:

These users prefer spreading payments over installments and barely touch cash advances. They spend less overall but are reliable payers.

Key Traits:

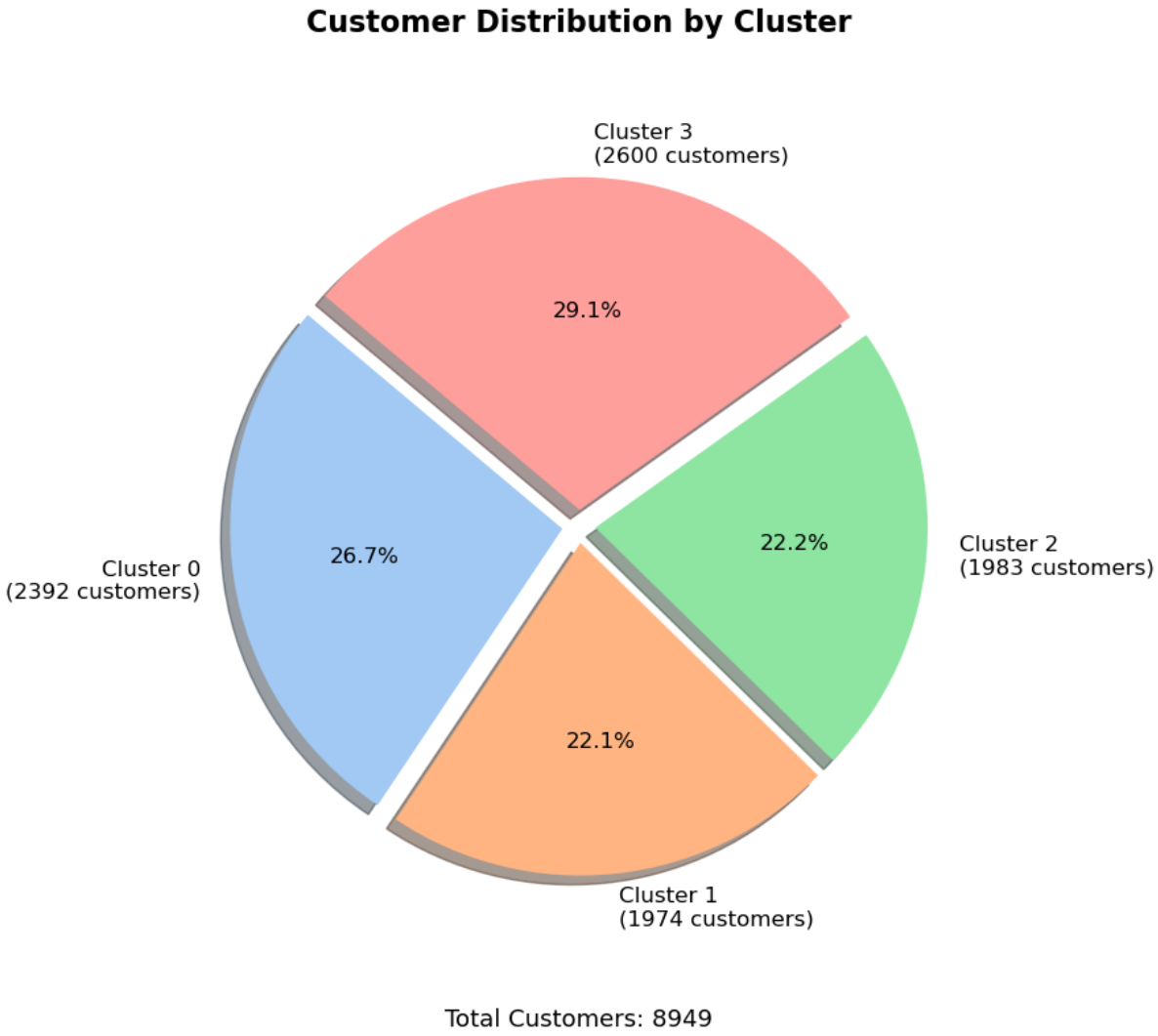
- Moderate Purchase Volume: ~674 transactions (≈ 597 installments)
- Very Low One-Off Spend: ~78 transactions
- Low Cash Advances: ~0.13 frequency, avg ≈ $40
- Highest Full-Payment Rate: ≈ 29.7 %

- **Summary:**

The clustering results reveal four distinct customer segments, each with unique financial behaviors and credit usage patterns. From high-value full payers to cash-advance-reliant users, these insights help highlight opportunities for tailored marketing, risk management, and customer engagement strategies. By understanding what defines each group, whether it's their spending habits, repayment behavior, or preferred transaction types, businesses can design more personalized financial products, improve retention, and reduce credit risk.

6. Visualization and Analysis

Now we can visualize the clusters to understand the differences and significance.



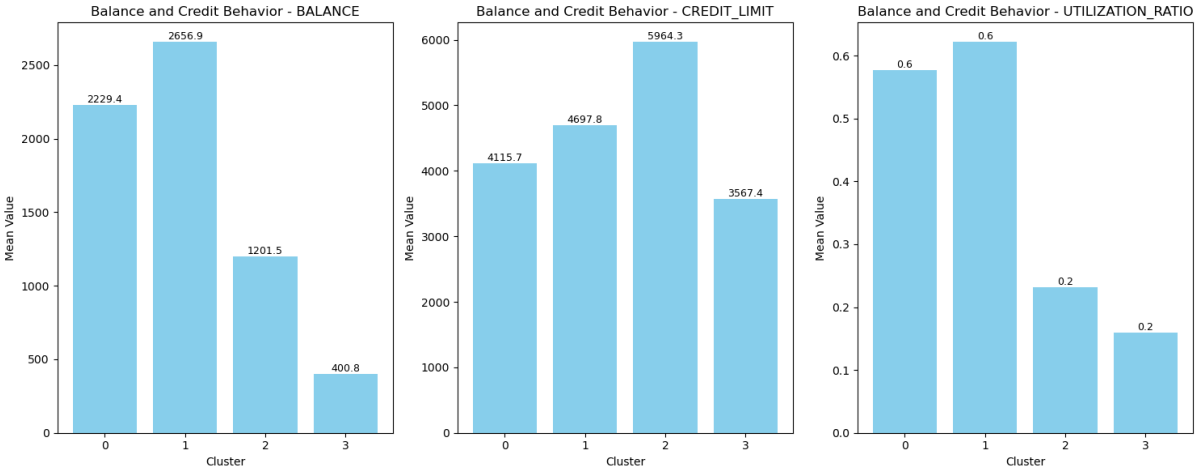
The pie chart shows the distribution of customers across four clusters. Cluster 0 (26.7%) consists of cash advance-dependent users. Cluster 1 (22.1%) represents moderate spenders with balanced credit use. Cluster 2 (22.2%) includes high-spending full payers, while Cluster 3 (29.1%) comprises installment-focused budget users. These insights highlight diverse customer behaviors, allowing for targeted marketing and risk management strategies.

7. Business Insights and Recommendations

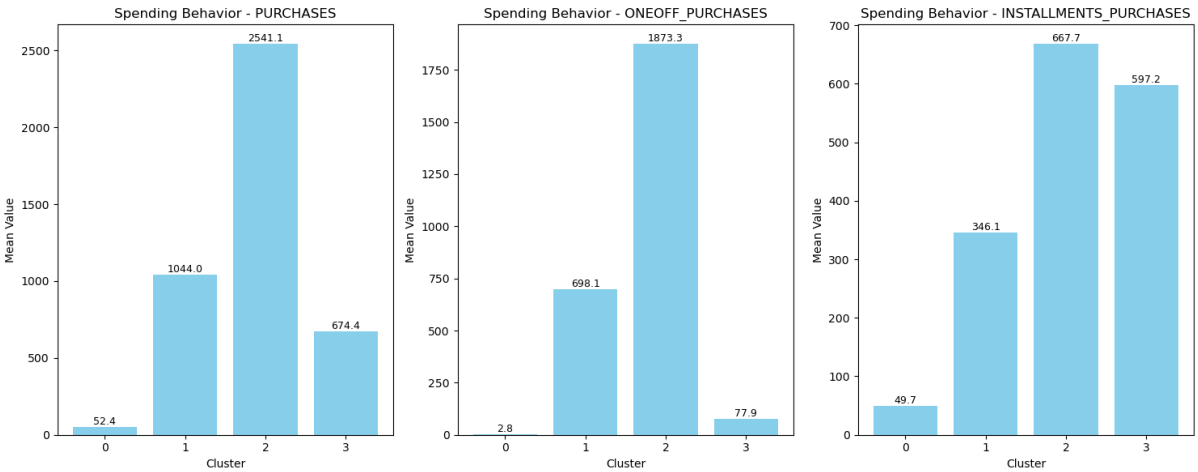
7.1. Comparisons of Key Features Across Clusters

Divided the columns to some categories as follows

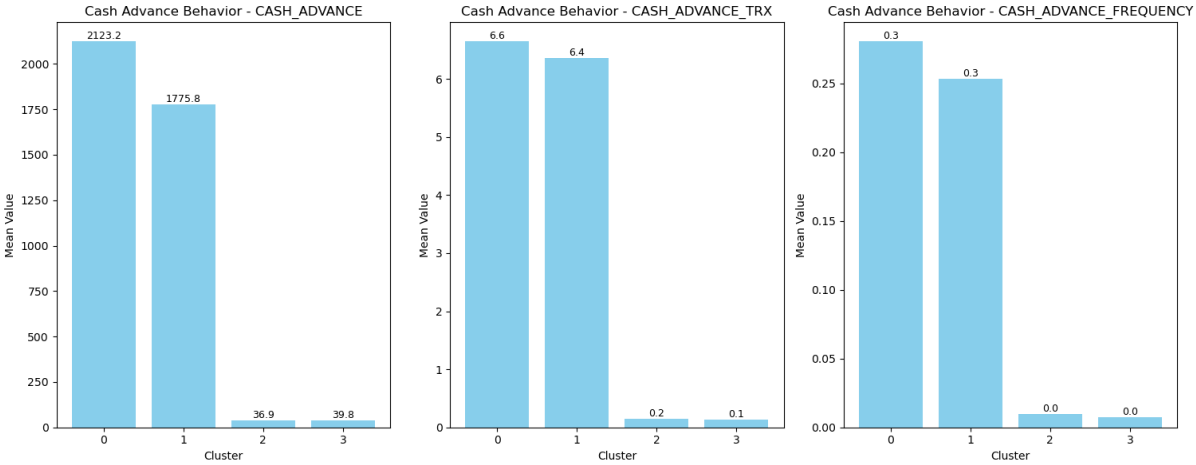
7.1.1. Balance and Credit Behavior:



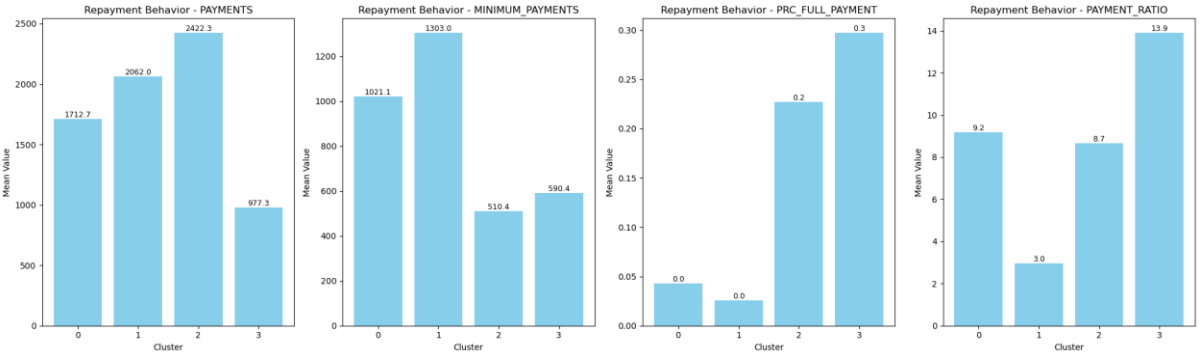
7.1.2. Spending Behavior:



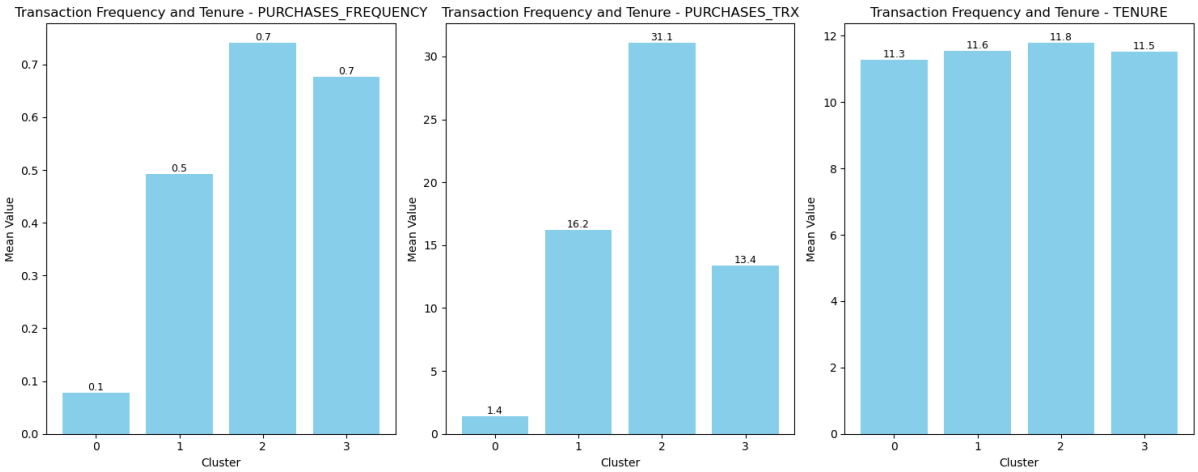
7.1.3. Cash Advance Behavior:



7.1.4. Repayment Behavior:



7.1.5. Transaction Frequency and Tenure:



7.2. Recommendations

- **Offer personalized credit card products:**

Customers show different spending and payment habits, some prefer one-off purchases, others lean toward installments, and some rely heavily on cash advances. Designing credit card plans that match these distinct behaviors can improve customer satisfaction and engagement.

- **Promote financial wellness:**

High reliance on cash advances suggests that some customers may be facing financial stress. Introducing educational campaigns or in-app budgeting tools can encourage healthier financial habits and reduce potential risk.

- **Optimize credit limits and risk management:**

By understanding typical utilization and payment behavior in each group, the bank can make better decisions on adjusting credit limits, setting payment reminders, or identifying early signs of risk.

- **Enhance rewards and loyalty programs:**

For customers who consistently pay on time and use their credit cards actively, targeted rewards and loyalty perks can help maintain strong relationships and reduce churn.

- **Use data to drive communication:**

Segmenting customers based on behavior allows for more meaningful and relevant communication, whether promoting offers, providing advice, or notifying users of beneficial account features.

8. Conclusion

This project aimed to uncover patterns in customer credit card usage using clustering techniques. By analyzing key features like spending habits, payment behavior, and credit utilization, we identified distinct groups of users with unique financial behaviors. These insights not only help in understanding the current customer base but also offer a solid foundation for improving services, reducing risk, and building stronger relationships with customers. Moving forward, leveraging this data in day-to-day business strategies can help deliver more value to both the company and its clients.