Customer Segmentation for Credit Card Customers

Machine Learning 2 — Project Report

Dataset: CC General — Credit Card Customer Behaviour
Algorithms: K-Means, Gaussian Mixture Model, PCA, t-SNE

Omar Gamal ElKady

February 2026

0 Contents

1 Executive Summary

This report documents the end-to-end implementation of an unsupervised machine learning pipeline that segments 8,950 credit card customers into seven behaviourally distinct groups. The analysis follows five phases prescribed by the project specification: data exploration and preprocessing, determination of the optimal cluster count, customer segmentation, visualisation and analysis, and actionable business recommendations.

Key outcomes are summarised below.

Phase	Key Outcome
Data Exploration	8,950 customers, 17 numeric features; 2 features with missingness; strong right-skew; high collinearity in purchase-related features.
Preprocessing	Median imputation; log1p transformation on 10 monetary/count features; PCA retained 95%+ variance in 6 components.
Optimal k	Elbow method suggested $k = 4$; silhouette score peaked at $k = 7$; $k = 7$ selected.
Segmentation	K-Means outperformed GMM (silhouette 0.4477 vs. 0.3259); seven well-separated clusters identified.
Visualisation	t-SNE, radar charts, and heatmaps confirm distinct cluster fingerprints.

2 Phase 1 — Data Exploration & Preprocessing

2.1 Dataset Overview

The dataset contains **8,950 credit card customers** and **18 columns** (one customer identifier + 17 numeric behavioural features). The features capture six behavioural dimensions:

- **Balance behaviour:** `BALANCE`, `BALANCE_FREQUENCY`
- **Purchase behaviour:** `PURCHASES`, `ONEOFF_PURCHASES`, `INSTALLMENTS_PURCHASES`, `PURCHASES_FREQUENCY`, `ONEOFF_PURCHASES_FREQUENCY`, `PURCHASES_INSTALLMENTS_FREQUENCY`, `PURCHASES_TRX`
- **Cash advance behaviour:** `CASH_ADVANCE`, `CASH_ADVANCE_FREQUENCY`, `CASH_ADVANCE_TRX`
- **Credit & payments:** `CREDIT_LIMIT`, `PAYMENTS`, `MINIMUM_PAYMENTS`, `PRC_FULL_PAYMENT`
- **Tenure:** `TENURE`

2.2 Missing Values

Only two features contain missing values:

Feature	Missing Count	Missing %
`CREDIT_LIMIT`	1	0.01%
`MINIMUM_PAYMENTS`	313	3.50%

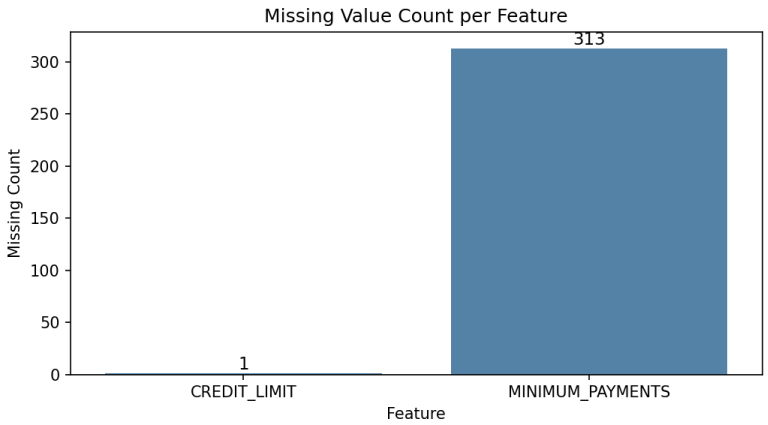

Figure 1: Count of missing values per feature.

Business Insight: Both features are right-skewed monetary amounts. Median imputation was applied rather than mean imputation to avoid inflating values and introducing bias into the cluster centroids. The `CUST_ID` column was dropped before any modelling as it carries no behavioural information.

2.3 Feature Distributions & Skewness

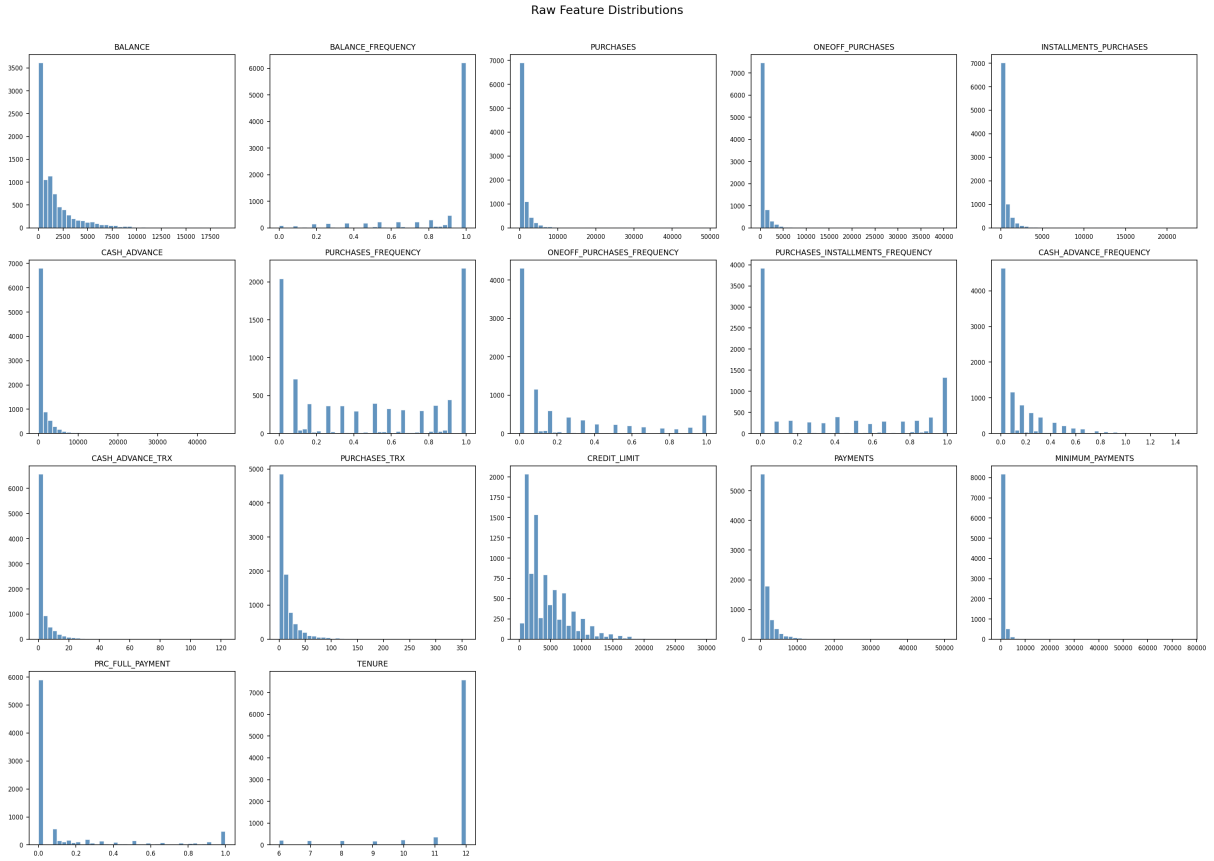

Figure 2: Raw distributions of all 17 numeric features before transformation.

Skewness analysis reveals that **15 out of 17 features** have $|\text{skew}| > 1$. The most extreme

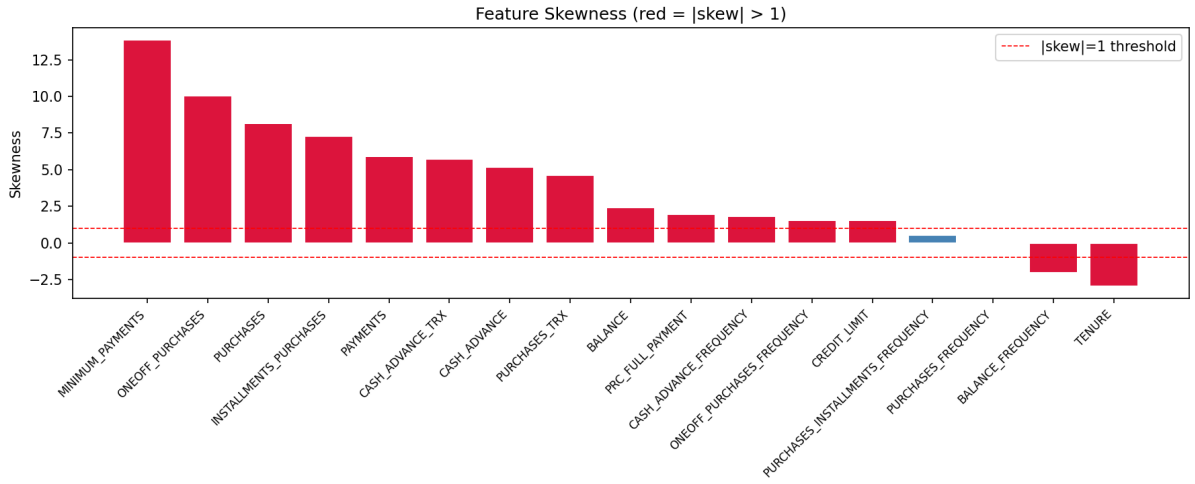

Figure 3: Skewness of each feature (red bars: $|\text{skew}| > 1$).

cases are:

Feature	**Skewness**
MINIMUM_PAYMENTS	13.85
ONEOFF_PURCHASES	10.05
PURCHASES	8.14
INSTALLMENTS_PURCHASES	7.30
PAYMENTS	5.91
CASH_ADVANCE_TRX	5.72
CASH_ADVANCE	5.17

Business Insight: Most customers carry moderate balances and purchase amounts, while a minority of high-value customers drives extreme values. This long right tail is characteristic of real-world financial data and motivates the use of a log transformation to compress outlier influence without discarding those customers entirely.

2.4 Correlation Analysis

Key correlation findings:

- PURCHASES and ONEOFF_PURCHASES: $r \approx 0.92$ — one-off purchases constitute the majority of total purchases for most customers.
- PURCHASES_FREQUENCY and PURCHASES_INSTALLMENTS_FREQUENCY: $r \approx 0.86$ — frequent purchasers tend to buy in instalments.
- CASH_ADVANCE and CASH_ADVANCE_TRX: $r \approx 0.91$ — higher cash-advance amounts correlate with higher transaction counts.

Business Insight: High collinearity between purchase-related features adds redundant information to the feature space and can distort distance-based clustering. PCA (applied in Section 2.7) collapses these correlated dimensions into fewer orthogonal components, improving cluster quality.

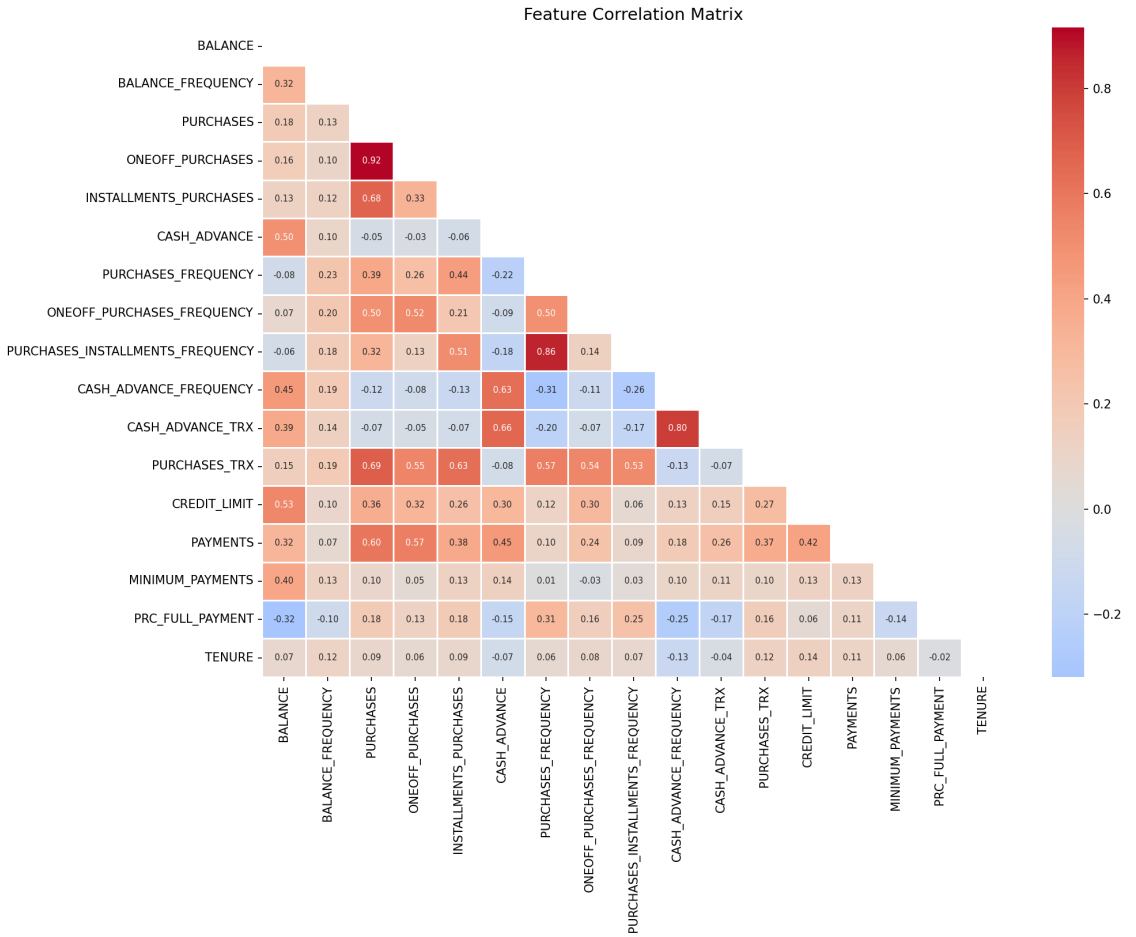

Figure 4: Lower-triangular correlation heatmap of all numeric features.

2.5 Outlier Detection

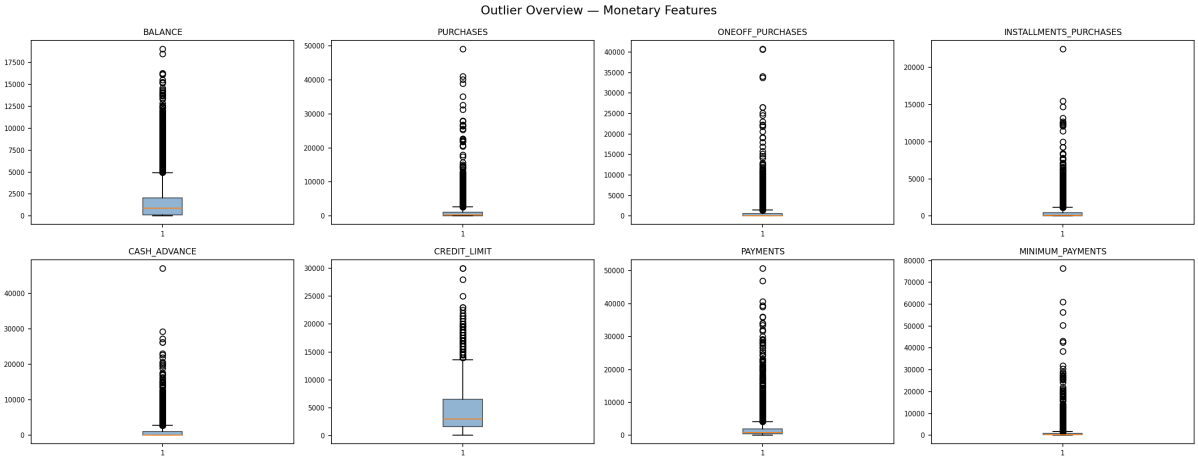

Figure 5: Box plots of monetary features showing extreme outliers.

Business Insight: Extreme outliers — customers with `CASH_ADVANCE` exceeding $40,000 or `CREDIT_LIMIT` up to $30,000 — represent real VIP and high-risk customer segments, not data errors. Removing them would eliminate entire behavioural groups.

Log transformation compresses their scale while retaining them in the analysis.

2.6 Preprocessing — Log$_1$p Transformation

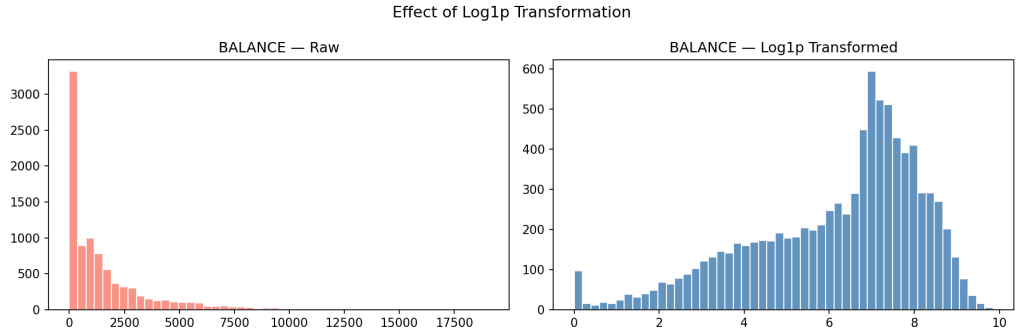

Figure 6: Effect of log$_1$p transformation on the `BALANCE` feature.

Log$_1$p ($\log(1+x)$) transformation was applied to the 10 most skewed monetary and count features:

`BALANCE`, `PURCHASES`, `ONEOFF_PURCHASES`, `INSTALLMENTS_PURCHASES`, `CASH_ADVANCE`, `CREDIT_LIMIT`, `PAYMENTS`, `MINIMUM_PAYMENTS`, `CASH_ADVANCE_TRX`, `PURCHASES_TRX`

The +1 offset ensures safe handling of zero values ($\log(0)$ is undefined). No StandardScaler was applied because PCA is sensitive to variance, and preserving the relative scale differences between features carries meaningful behavioural information.

2.7 Dimensionality Reduction — PCA

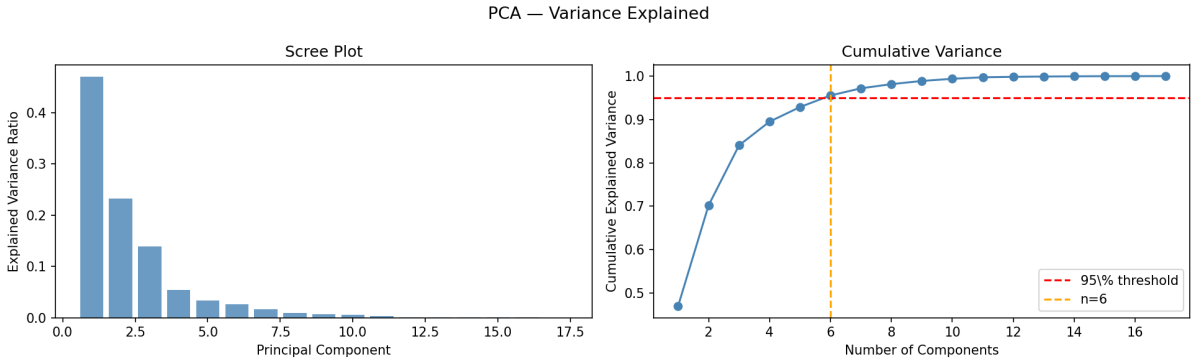

Figure 7: Scree plot (left) and cumulative explained variance (right) for PCA.

Principal Component Analysis was applied to the log-transformed feature matrix ($8{,}950 \times 17$). **6 principal components** were sufficient to capture **95% of the total variance**, reducing the feature dimensionality from 17 to 6 while eliminating correlated noise. This compressed representation was used as input to all clustering algorithms.

3 Phase 2 — Determining the Optimal Number of Clusters

To determine the optimal number of clusters, two complementary metrics were evaluated over $k \in \{2, 3, \ldots, 10\}$:

- **Elbow Method (Inertia):** Measures within-cluster sum of squared distances. The "elbow" — the point of diminishing returns — indicates k.
- **Silhouette Score:** Measures how similar each point is to its own cluster compared to others (-1 to $+1$; higher is better).

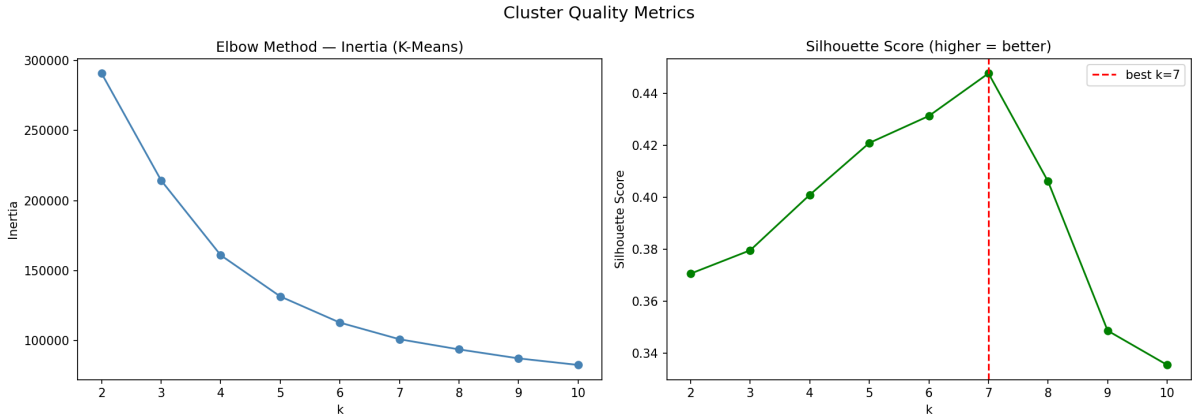

Figure 8: Elbow curve (inertia) and silhouette scores across $k = 2$ to $k = 10$.

k	Inertia	Silhouette Score
2	290,787	0.3706
3	214,367	0.3796
4	161,004	0.4010
5	131,575	0.4209
6	112,908	0.4313
7	**101,037**	**0.4477**
8	93,773	0.4061
9	87,372	0.3487
10	82,687	0.3355

Decision justification:

- The second derivative of the inertia curve peaks at $k = 4$, indicating the traditional elbow point.
- The silhouette score peaks at $k = \mathbf{7}$ (0.4477) and drops sharply at $k = 8$.
- $k = 7$ was chosen because it maximises the silhouette score among the candidate values ($\{4, 7\}$), indicating the most internally coherent and externally well-separated clusters.

Business Insight: While $k = 4$ offers a simpler model, it merges customer groups with meaningfully different behaviours (e.g., cash-advance revolvers vs. installment shoppers). $k = 7$ provides sufficient granularity to design distinct marketing strategies for each segment without over-fragmenting the customer base.

4 Phase 3 — Customer Segmentation

4.1 Algorithm Selection: K-Means vs. GMM

Both K-Means and Gaussian Mixture Model (GMM) were trained with $k = 7$ on the 6-dimensional PCA-transformed data.

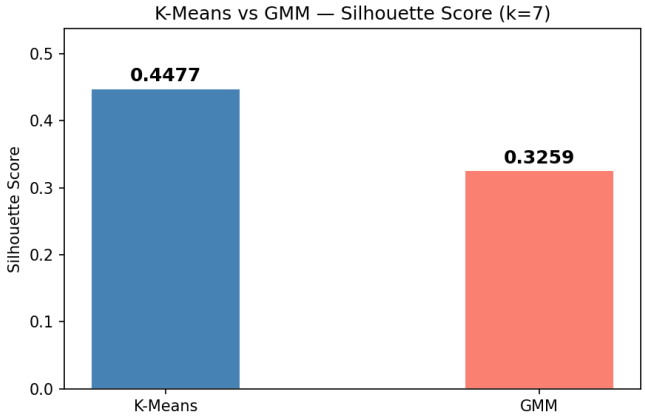

Figure 9: Silhouette score comparison: K-Means vs. GMM at $k = 7$.

Model	Silhouette Score	Selected?
K-Means ($k = 7$)	**0.4477**	✓
GMM (full covariance, $k = 7$)	0.3259	

K-Means was selected as the final model. This is consistent with the roughly spherical cluster shapes visible in the t-SNE projection (Figure 10), for which K-Means is well suited. GMM's flexibility in modelling elliptical shapes did not translate into better separability here because the PCA-compressed space already produces compact, isotropic clusters.

4.2 Cluster Profiles

The final K-Means model with $k = 7$ assigns every customer to one of seven segments. Below is a detailed narrative profile of each segment.

Cluster 0 — Big Spenders / VIP (18.2%, 1,633 customers) The highest-purchasing segment with a mean of \$2,683 in purchases per period, the highest credit limits, and a notable 26% full-payment rate. These customers shop frequently (purchase frequency 0.84) and split purchases across both one-off and instalment channels. They carry moderate balances and make substantial payments.

Cluster 1 — Cash-Only Dependents (23.1%, 2,068 customers) The largest and most at-risk segment. Mean purchases are effectively \$0 — these customers never use the card for shopping. Instead they rely entirely on cash advances (\$1,994 mean), carry high balances (\$2,151), and have a near-zero full-payment rate (4%). This is the bank's highest-risk concentration.

Cluster 2 — Installment Shoppers (21.2%, 1,895 customers) The most financially prudent segment. Low balances (\$365), zero cash-advance usage, and the highest

full-payment rate (30%). Purchases are moderate ($499) and predominantly in instalments. This group presents the lowest credit risk.

Cluster 3 — High-Risk Heavy Users (10.5%, 938 customers) The most financially active and highest-risk group. Highest balances ($2,932), highest cash advances ($2,200), second-highest purchases ($2,068), and highest credit limits ($5,964). Near-zero full-payment rate (7%). This segment generates maximum revenue but also maximum default risk.

Cluster 4 — One-off & Cash Advance Revolvers (8.9%, 798 customers) Characterised by one-off purchase behaviour (frequency 0.28) combined with heavy cash advance usage ($2,023). High balances ($2,354) and low full-payment rates (5%) indicate revolving debt. A hybrid risk profile.

Cluster 5 — One-off Shoppers (12.7%, 1,138 customers) Moderate activity with a focus on one-off purchases ($836 mean, all one-off). No cash advance usage. Moderate balances ($747) and a 14% full-payment rate. A transitional segment with potential to migrate toward VIP behaviour.

Cluster 6 — Installment & Cash Advance Revolvers (5.4%, 480 customers) The smallest and most complex segment. Combines instalment purchase habits ($523 mean, predominantly instalment) with significant cash advance usage ($1,994). High balances ($2,580) and near-zero full-payment rate (4%) — a financially stretched group.

5 Phase 4 — Visualisation & Analysis

5.1 t-SNE 2D Cluster Projection

t-SNE (t-Distributed Stochastic Neighbour Embedding) projects the 6-dimensional PCA space into 2D while preserving local neighbourhood structure. The seven clusters appear as clearly separated colour blobs with minimal overlap, which is consistent with a silhouette score of 0.4477.

> **Business Insight:** The spatial separation confirms that the 7 groups are genuinely distinct in their financial behaviour — not arbitrary partitions. Any overlap between adjacent clusters (e.g. C3 and C0) reflects customers who share partial behaviours such as high purchases in both groups, but differ in cash-advance usage.

5.2 Cluster Size Distribution

> **Business Insight:** Cash-Only Dependents (C1) is the largest segment at 23.1% — representing the bank's greatest concentration of default risk in a single group. Installment & Cash Advance Revolvers (C6) at only 5.4% is the smallest but most financially complex segment, requiring specialised monitoring despite its small size.

5.3 Feature Distributions by Cluster

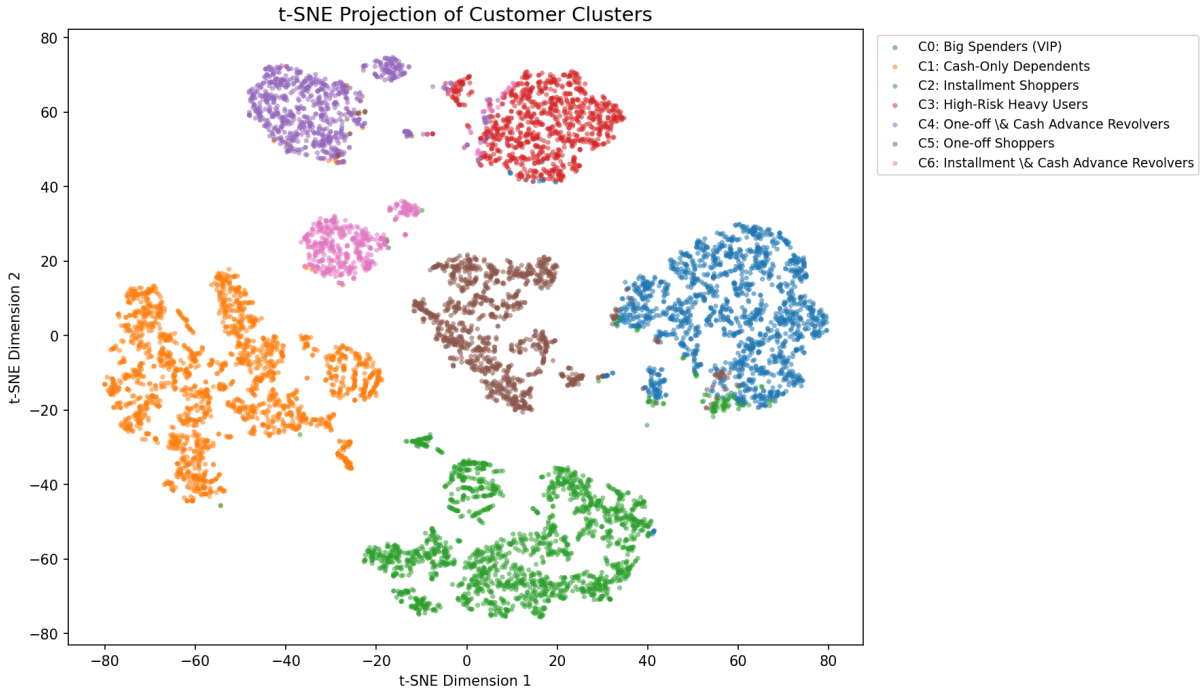

Figure 10: t-SNE 2D projection of the 8,950 customers coloured by cluster assignment.

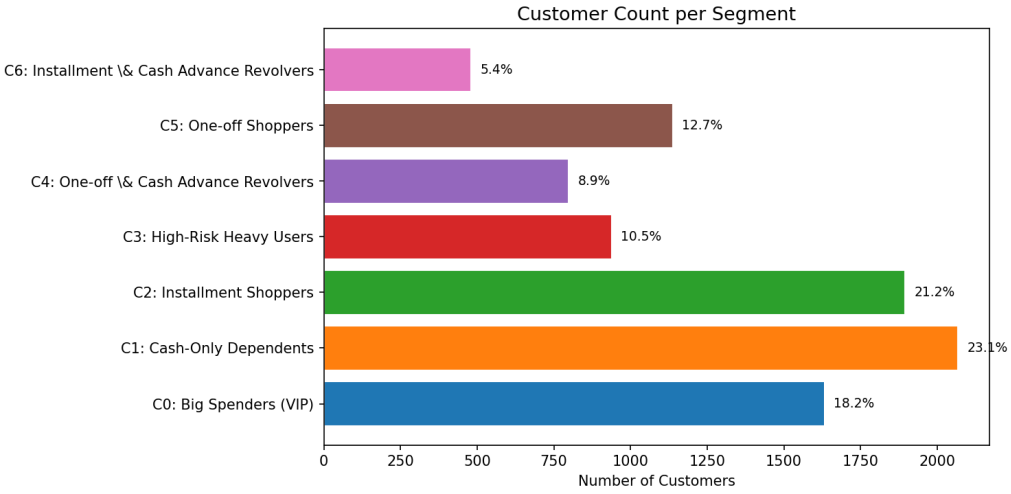

Figure 11: Number of customers per segment with percentage labels.

Business Insight: `PURCHASES` and `CASH_ADVANCE` provide the clearest cluster separation. C1 spikes at exactly zero for purchases, while C0 spreads far to the right. `TENURE` shows near-complete overlap across all clusters, confirming that account age carries no meaningful discriminatory power for segmentation.

5.4 Cluster Heatmap — Mean Feature Values

Business Insight: C3 (High-Risk Heavy Users) is the darkest row across `BALANCE`, `CASH_ADVANCE`, and `PAYMENTS` simultaneously — the most financially active and highest-risk segment. C2 (Installment Shoppers) is the lightest row overall — lowest activity

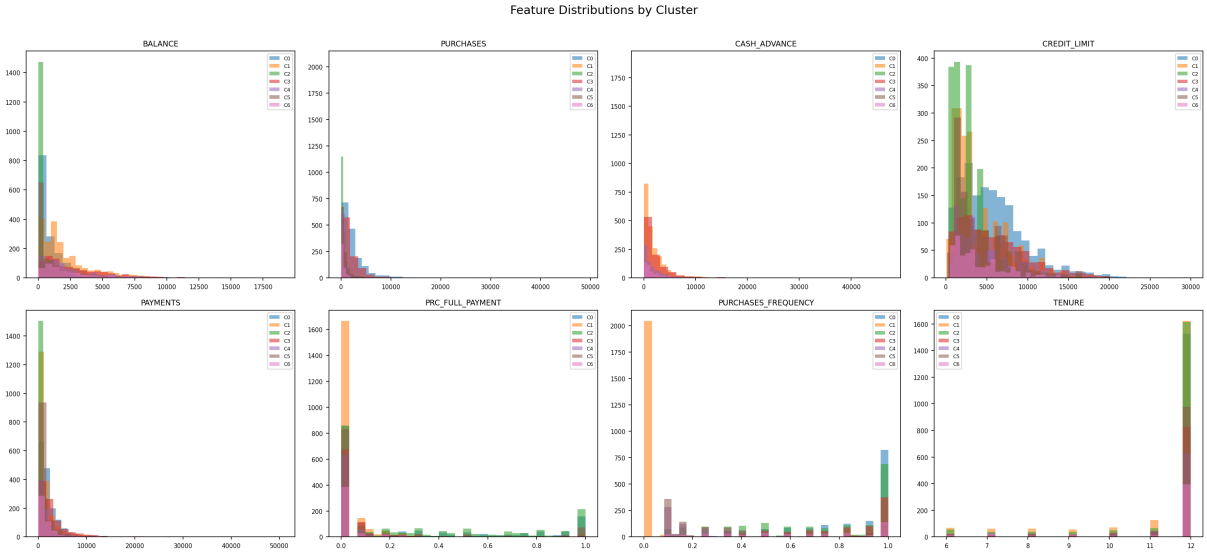

Figure 12: Overlaid histograms of 8 key features coloured by cluster.

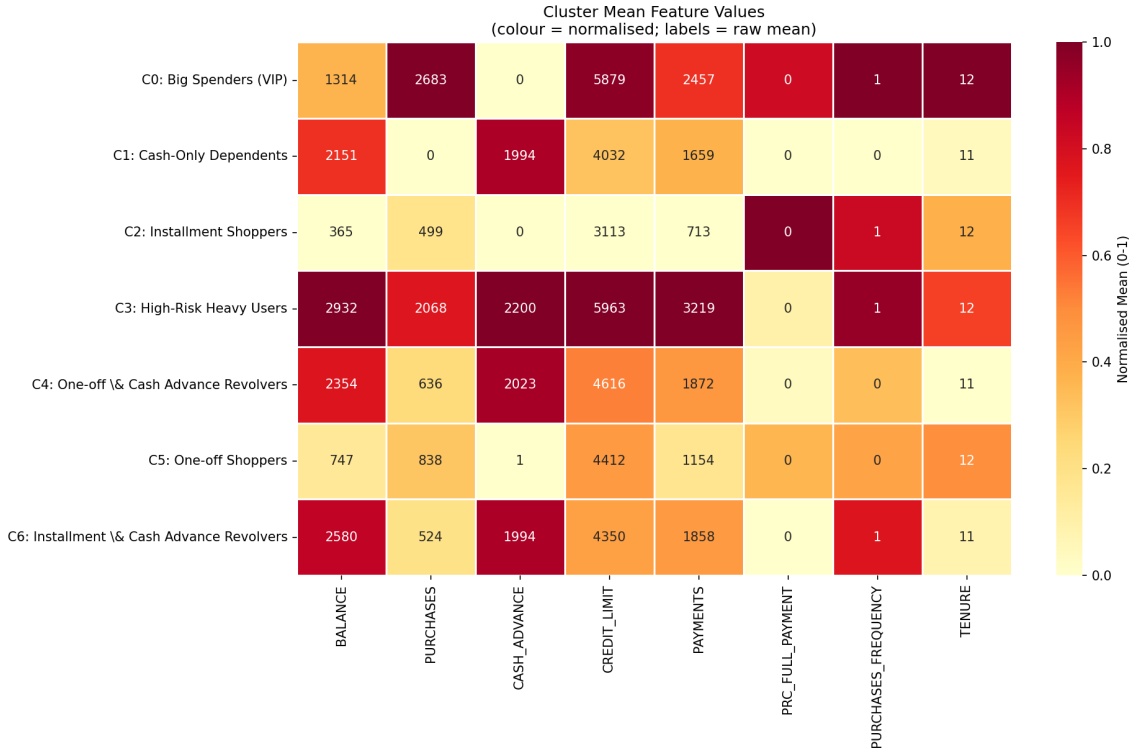

Figure 13: Heatmap of normalised mean feature values per cluster. Raw means shown as annotations.

with zero cash-advance usage. The `PRC_FULL_PAYMENT` column is meaningfully above zero only for C2, confirming it is the only segment with a tendency to pay balances in full.

5.5 Radar Charts — Cluster Fingerprints

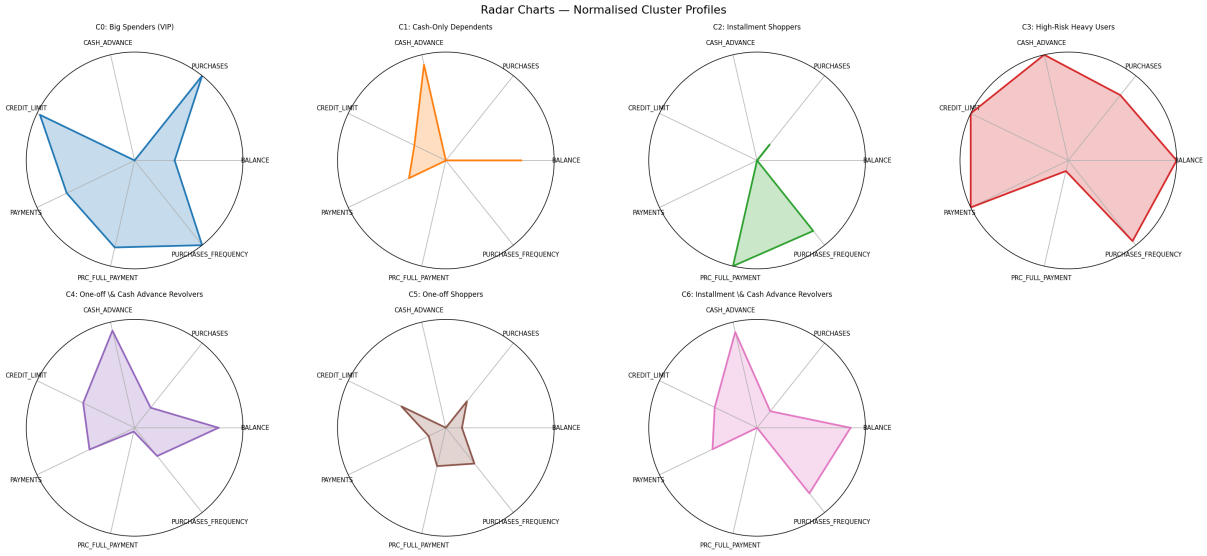

Figure 14: Normalised radar charts showing the behavioural profile of each cluster.

Business Insight: C3 has the largest radar polygon, spiking simultaneously across `PURCHASES`, `CASH_ADVANCE`, and `BALANCE` — the only cluster combining all three risk dimensions. C1 has a highly asymmetric shape — large only on `CASH_ADVANCE` and `BALANCE`, completely flat on `PURCHASES`. C2 has the smallest polygon overall, reflecting the lowest financial activity across all features.

5.6 Box Plots — Key Financial Features by Segment

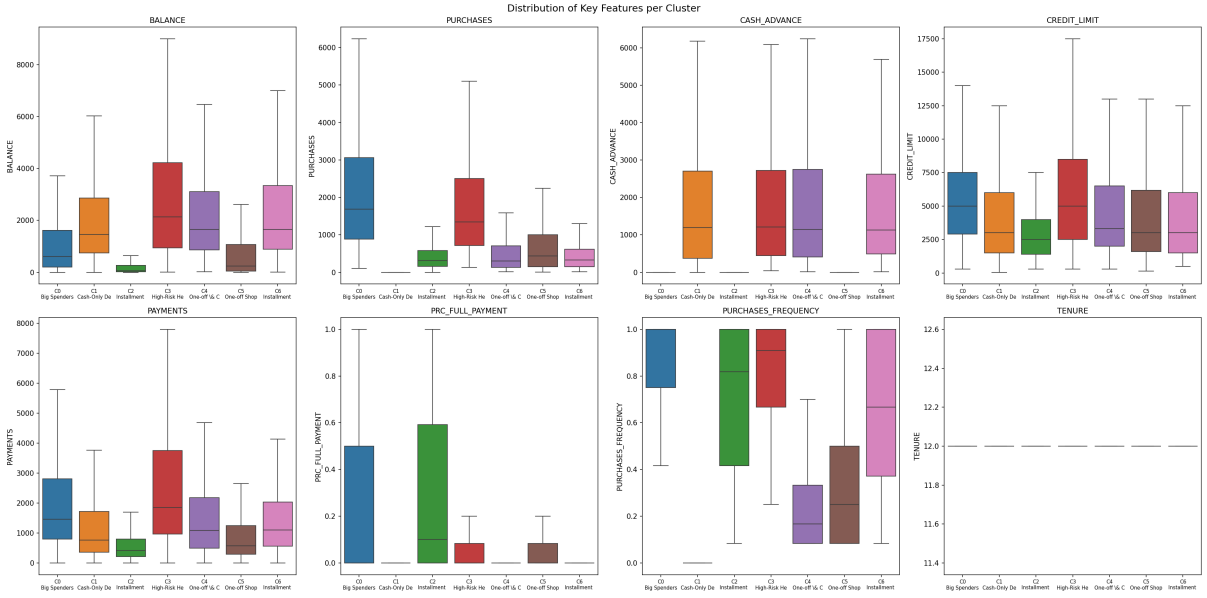

Figure 15: Box plots (outliers hidden) of 8 key features disaggregated by cluster.

Business Insight:

- **BALANCE** — C3 and C6 have the highest and widest boxes, confirming heavy debt carriers. C2 sits near zero with a narrow box.

- **PURCHASES** — C0 has the highest median; C1 has a flat box at zero — every customer in this segment truly never purchases.
- **CASH_ADVANCE** — C1, C3, C4, C6 all have elevated boxes; C0 and C2 sit at zero.
- **CREDIT_LIMIT** — C0 and C3 have the highest limits, reflecting the bank rewarding high spenders with more credit.
- **PRC_FULL_PAYMENT** — Almost all clusters are near zero; C2 is the only segment with a visible upward spread.
- **PURCHASES_FREQUENCY** — C0 is narrow and near 1.0 (buys every month); C1 is flat at 0.0 (never buys). The sharpest contrast of all features.
- **TENURE** — All clusters have nearly identical boxes — account age provides no discriminatory value.

6 Phase 5 — Business Insights & Strategic Recommendations

6.1 Segment-Level Recommendations

Segment	Business Priority	Recommended Actions
C0 — Big Spenders (VIP) (18.2%)	Revenue maximisation	Cross-sell premium travel and lifestyle cards. Offer VIP concierge services, airport lounge access, and elevated reward multipliers. Maintain credit-limit headroom to enable continued high-spend behaviour. Monitor for competitor card switching.
C1 — Cash-Only Dependents (23.1%)	Risk mitigation	Flag for proactive credit counselling. Offer personal loan products at lower interest rates than cash-advance fees as a debt-restructuring incentive. Set conservative credit-limit increase thresholds. Trigger alerts when cash-advance frequency rises.
C2 — Instalment Shoppers (21.2%)	Cross-sell & activation	These low-risk customers are under-utilised. Offer instalment-specific promotional rates (e.g., 0% instalment for 12 months). Upsell to premium card tiers. Target with partner merchant offers to increase purchase frequency.
C3 — High-Risk Heavy Users (10.5%)	Dual-track: revenue + risk	Highest revenue potential but also highest default risk. Reward high spend with premium benefits while simultaneously stress-testing credit exposure. Implement early-warning delinquency models calibrated specifically to this segment. Limit unsolicited credit-limit increases.

Continued on next page...

Segment	Business Priority	Recommended Actions
C4 — One-off & Cash Advance Revolvers (8.9%)	Debt management	Offer structured repayment plans to reduce revolving cash-advance balances. Market balance-transfer promotions. Educate on the cost differential between purchase interest and cash-advance interest rates.
C5 — One-off Shoppers (12.7%)	Engagement & migration	Activate with bonus rewards on one-off spend categories (e.g., dining, travel). Introduce instalment features to convert single large purchases into recurring instalment transactions, deepening engagement. Migrate aspirational customers toward VIP tier.
C6 — Installment & Cash Advance Revolvers (5.4%)	Specialist risk monitoring	Despite small size, this segment's combined instalment and cash-advance usage creates complex credit exposure. Assign dedicated relationship managers. Offer debt consolidation products. Apply stricter credit review cycles.

6.2 Portfolio-Level Insights

1. **Risk concentration:** 44.9% of customers (C1 + C3 + C6) carry high balances combined with heavy cash-advance reliance and near-zero full-payment rates. This is the primary default-risk concentration in the portfolio.
2. **Revenue concentration:** 28.7% of customers (C0 + C3) generate the highest payment and purchase volumes. Retaining these customers and expanding their credit products is the primary revenue-protection priority.
3. **Under-leveraged segment:** C2 Installment Shoppers (21.2%) are the safest customers yet have the lowest credit limits and lowest utilisation. They represent the greatest opportunity for risk-free credit expansion.
4. **Tenure is irrelevant:** Account age does not differentiate customer behaviour. Loyalty programmes should be behaviour-based (spend patterns, payment discipline) rather than tenure-based.
5. **Two-dimensional risk model:** The clearest risk stratification comes from combining `CASH_ADVANCE` (liquidity stress indicator) with `PRC_FULL_PAYMENT` (repayment discipline indicator). Customers high on the former and low on the latter form the highest-risk cohort (C1, C3, C6).

7 Summary of Findings

Metric	Value / Outcome
Dataset size	8,950 customers, 17 behavioural features
Missing values	2 features; median-imputed
Preprocessing	Log_1p on 10 features; PCA to 6 components (95% variance)
Optimal k	7 (silhouette 0.4477, elbow at 4)
Best model	K-Means (silhouette 0.4477 vs GMM 0.3259)
Cluster count	7 distinct segments
Largest segment	C1 Cash-Only Dependents (23.1%)
Highest-risk segment	C3 High-Risk Heavy Users + C1 Cash-Only Dependents
Highest-revenue segment	C0 Big Spenders (VIP) + C3 High-Risk Heavy Users
Safest segment	C2 Installment Shoppers (30% full-payment rate)
Key discriminators	PURCHASES, CASH_ADVANCE, PRC_FULL_PAYMENT, PURCHASES_FREQUENCY
Non-discriminating feature	TENURE (identical distribution across all clusters)