CREDIT SCORE CLASSIFICATION

By Group 2

Tishani Wijekoon(S16379), Chami Sewwandi(S16028), W.K.Hiruni Hasara(S16210), S.Luxan(s16329)

ABSTRACT

This project investigates the patterns and relationships among customer attributes that influence credit score classification using a comprehensive dataset of 100,000 observations and 28 variables. Through extensive exploratory data analysis (EDA), we examined how demographic, financial, and behavioral variables are associated with credit scores categorized as Good, Standard, and Poor. Our analysis revealed significant class imbalance in the target variable, with Standard credit scores comprising 53.2% of observations, Poor 29.0%, and Good 17.8%. Key findings demonstrate strong associations between credit scores and variables such as number of credit inquiries, outstanding debt, payment delays, and credit utilization patterns. Statistical validation through ANOVA and Chi-square tests confirmed the significance of these relationships. The analysis revealed that poor credit scores are characterized by higher debt loads, more credit accounts, increased payment delays, and riskier financial behaviors, while good credit scores are associated with more conservative credit management practices. These insights provide a foundation for future predictive modeling efforts and offer valuable guidance for financial institutions in risk assessment and customer evaluation.

TABLE OF CONTENTS

Abstract 1
List of Figures 2
List of Tables 2
Introduction 3
The question we are going to answer 3
Data set 4

Data Pre-processing and Feature Selection 5
Main results of the descriptive analysis 7
Target Variable Distribution and Characteristics 7
Univariate Analysis 7
Bivariate Analysis 8
Temporal Pattern Analysis 9
Famd And Clustering 10
Suggestions for a quality advanced analysis 11
Appendix 12
Bibliography 12

LIST OF FIGURES

Figure 1-Outlier Detection 5
Figure 2-Bar chart of the recategorized occupation 5
Figure 3-Bar chart of the recategorized Payment_Behaviour 6
Figure 4-Correlation heatmap 6
Figure 5-Cramér's V correlation matrix 6
Figure 6-Distribution of the Credit Score 7
Figure 7-Boxplot of Outstanding_Debt 7
Figure 8-Boxplot of Amount_Invested_Monthly 7
Figure 9- No of credit cards Vs Credit score 8
Figure 10-Delays from due date Vs Credit score 8
Figure 11- No of loan Vs Credit score 8
Figure 12-Month Vs Credit score 9
Figure 13-Payment of Min_Amount VS credit score 10
Figure 14-The scree plot of explained inertia 10
Figure 15-Silhouette Score for Different k 10
Figure 16- 3D clusters on 1st three PCs 11

LIST OF TABLES

Table 1- Data set description 4
Table 2-Recategorized Occupation 5
Table 3- Recategorized Payment_Behaviour 6

INTRODUCTION

Credit scoring is a fundamental component of modern financial risk assessment, serving as a critical tool for lenders to evaluate the creditworthiness of potential borrowers. The ability to accurately classify individuals into appropriate credit risk categories directly impacts lending decisions, interest rates, and overall portfolio management strategies. Understanding the underlying factors that contribute to credit score classifications is essential for both financial institutions and consumers seeking to improve their financial standing. *(Arram, 2023)*

Traditional credit scoring models rely heavily on historical payment patterns and basic demographic information. However, with the increasing availability of comprehensive financial data, there is an opportunity to explore more nuanced relationships between various customer attributes and credit outcomes. This analysis aims to provide insights into these complex relationships through systematic exploratory data analysis.

The significance of this research extends beyond academic interest, as improved understanding of credit score determinants can lead to more fair and accurate risk assessment tools, better financial products, and enhanced consumer education initiatives. By identifying the most influential factors in credit score classification, financial institutions can develop more targeted strategies for risk management and customer acquisition.

THE QUESTION WE ARE GOING TO ANSWER

"What demographic, financial, and behavioral variables are most influential in differentiating customers with Good, Standard, and Poor credit scores?"

This research question addresses several key aspects:

- How do debt burdens and payment delays vary across score groups?
- Are certain occupations or age bands over-represented in Poor scores?
- Does the distribution of credit scores shift between January and August?

The question is designed to guide a comprehensive exploratory analysis that will identify the most significant predictors of credit score categories, ultimately providing insights for both theoretical understanding and practical application in credit risk assessment.

DATA SET

The dataset utilized in this analysis originates from Kaggle's "Credit Score Classification" dataset. This comprehensive dataset provides a unique longitudinal perspective on credit behavior, containing exactly 100,000 monthly observations representing 12,500 unique customers tracked across eight consecutive months from January to August.

Table 1- Data set description

Demographic Variables	
Variable	*Description*
Age	Customer age
Occupation	Professional category (16 distinct occupations)
Name	Customer name
SSN	Represents the social security number of a person
Customer_id	Represents a unique identification of a person

Financial Variables	
Variable	*Description*
Annual_Income	Yearly income
Monthly_Inhand_Salary:	Monthly disposable income
Outstanding_Debt	Current debt obligations
Credit_Utilization_Ratio:	Percentage of credit limit used

Credit Account Information	
Variable	*Description*
Total_EMI per month	The monthly EMI payments (in USD)
Num_Bank_Accounts	Number of bank accounts
Num of credit inquiries	Represents the number of credit card inquiries
Credit history age	The age of credit history of the person
Num_Credit_Card	Number of credit cards
Num_of_Loan	Number of active loans
Interest_Rate	Credit card interest rate
Types of loan	The types of loan taken by a person
Payment behavior	The payment behavior of the customer (in USD)

Behavioral Variables	
Delay_from_due_date:	Average payment delay in days
Payment_of_Min_Amount:	Whether minimum payments are made
Payment_Behaviour:	Spending and payment patterns
Credit_Mix:	Quality of credit portfolio
Changed_Credit_limit	The percentage change in credit card limit
Num of delayed payments	The average number of payments delayed by a person
Monthly balance	The monthly balance amount of the customer

Target Variable	
Variable	*Description*
Credit_Score	**Categorical outcome (Good, Standard, Poor)**

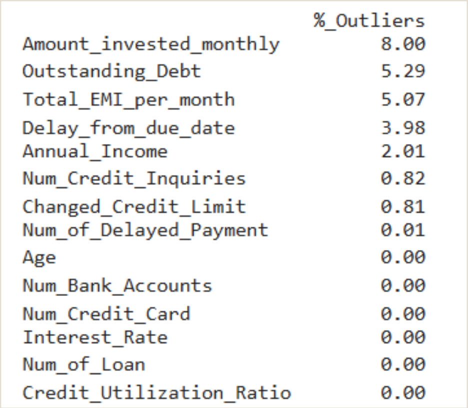
DATA PRE-PROCESSING AND FEATURE SELECTION

Missing Value Imputation

- Customer-consistent variables: Forward-fill within *Customer_ID* groups
- Monthly-varying variables: Group-wise mean imputation by *Customer_ID*
- Remaining missing values: Overall mean/mode imputation

Data Validation

- Corrected negative age values through customer-specific imputation
- Outlier Detection:
  - Used the IQR method to identify outliers
  - Outliers were retained in the data, as reasons for the outliers were not known
  - Plan to apply transformations in advanced analysis to mitigate the impact



```
                            %_Outliers
Amount_invested_monthly           8.00
Outstanding_Debt                  5.29
Total_EMI_per_month               5.07
Delay_from_due_date               3.98
Annual_Income                     2.01
Num_Credit_Inquiries              0.82
Changed_Credit_Limit              0.81
Num_of_Delayed_Payment            0.01
Age                               0.00
Num_Bank_Accounts                 0.00
Num_Credit_Card                   0.00
Interest_Rate                     0.00
Num_of_Loan                       0.00
Credit_Utilization_Ratio          0.00
```

Figure 1-Outlier Detection

Feature Selection

- Removed identifier variables (*ID, Customer_ID, Name, SSN*)
- Eliminated *Credit_History_Age* due to inconsistent formatting
- Dropped *Type_of_Loan*
- Occupation consolidation: Grouped 15 occupations into 3 socioeconomic categories *(Figure 2)*

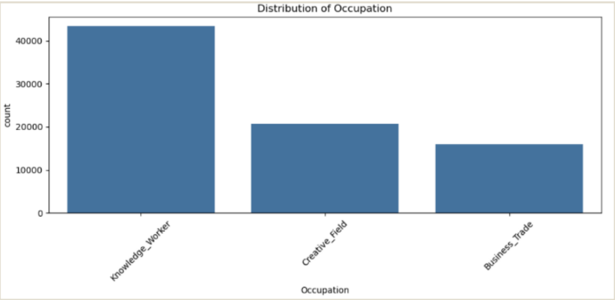


Figure 2-Bar chart of the recategorized occupation

Table 2-Recategorized Occupation

New Category	Original Occupations
Knowledge_Worker	Lawyer, Engineer, Scientist, Architect, Developer, Doctor, Teacher, Accountant
Creative_Field	Journalist, Musician, Writer, Media_Manager
Business_Trade	Manager, Entrepreneur, Mechanic

- *Payment_Behaviour* simplification: Reduced 6 categories to 3 based on payment value *(Figure3)*

Table 3- Recategorized Payment_Behaviour

New Category	Original Categories
Small_Payments	Low_spent_Small_value_payments, High_spent_Small_value_payments
Medium_Payments	Low_spent_Medium_value_payments, High_spent_Medium_value_payments
Large_Payments	Low_spent_Large_value_payments, High_spent_Large_value_payments

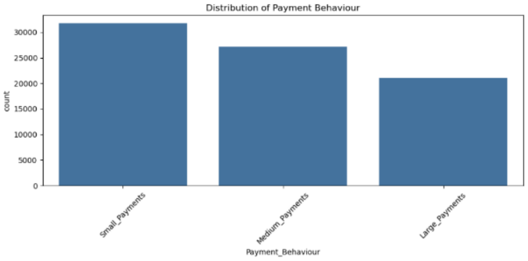


Figure 3-Bar chart of the recategorized Payment_Behaviour

- Split the dataset into training and testing. (Train set-80 000 , Test set-20 000)

Statistical Validation

- Performed correlation analysis:
  - — Correlation heatmap revealed high multicollinearity between *Annual_Income* and *Monthly_Inhand_Salary*, so *Monthly_Inhand_Salary* was dropped *(Figure 4)*
  - — Cramér's V test showed a strong association (0.81) between *Credit_Mix* and *Payment_of_Min_Amount*, so we retained only the latter *(Figure 5)*

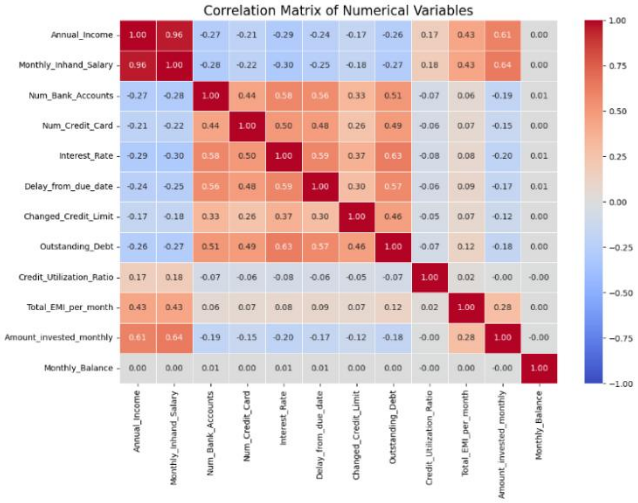


Figure 4-Correlation heatmap

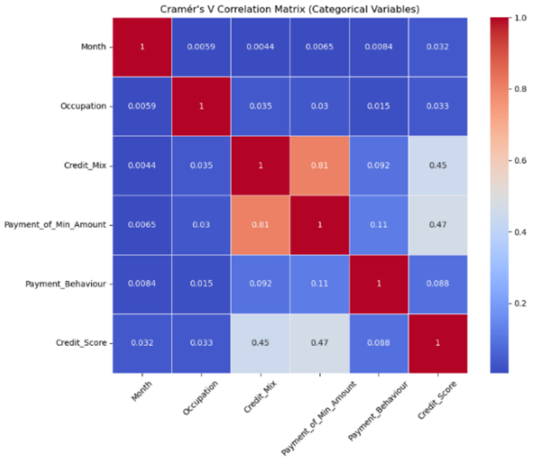


Figure 5-Cramér's V correlation matrix

- Conducted preliminary statistical tests for variable significance:
    - Chi-square tests showed all selected categorical variables were significantly associated with *Credit_Score*
    - ANOVA revealed *Monthly_Balance* was not significant and was dropped (p-value = 0.6033)

MAIN RESULTS OF THE DESCRIPTIVE ANALYSIS

TARGET VARIABLE DISTRIBUTION AND CHARACTERISTICS

The target variable, Credit Score, was found to be imbalanced with the Standard category comprising 53.2% of the observations, followed by Poor (29.0%) and Good (17.8%) *(Figure 6)*. This suggests that most customers fall into the average credit score range, with fewer customers achieving a good credit rating. This imbalance may require attention in future predictive modeling through methods such as resampling or class weight adjustments.

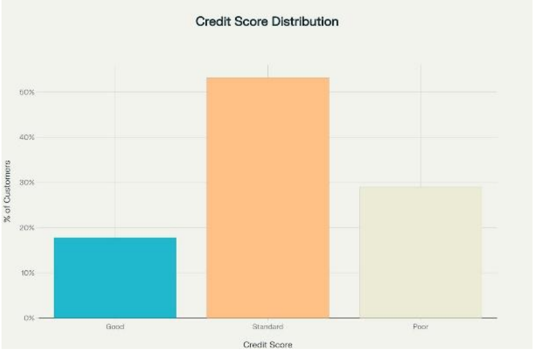


Figure 6-Distribution of the Credit Score

UNIVARIATE ANALYSIS

In the univariate analysis, we explored the distributional characteristics of key numerical variables. Variables such as *Age, Outstanding_Debt,* and *Number_of_Credit_Cards* exhibited right-skewed distributions, with a noticeable presence of extreme values. For instance, *Amount_Invested_Monthly* and *Outstanding_Debt (Figures 7,8)* contained substantial outliers. These were **not removed**, as they may reflect genuine financial behavior and will be addressed through transformation techniques during advanced modeling.

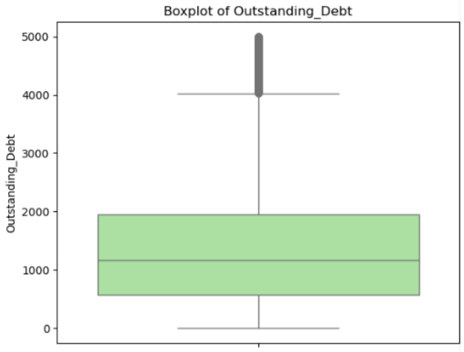


Figure 7-Boxplot of Outstanding_Debt

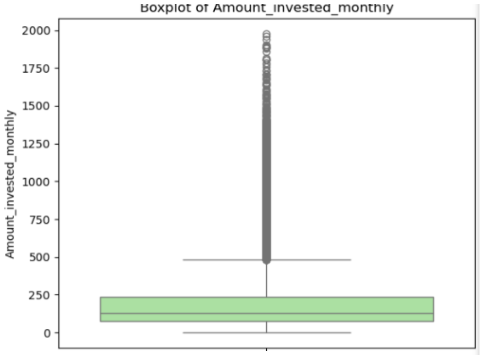


Figure 8-Boxplot of Amount_Invested_Monthly

BIVARIATE ANALYSIS

Bivariate analysis was conducted to explore how both numerical and categorical features relate to the *Credit_Score* target variable. Several key associations were observed.

For numerical variables, we found clear patterns across the credit score categories:

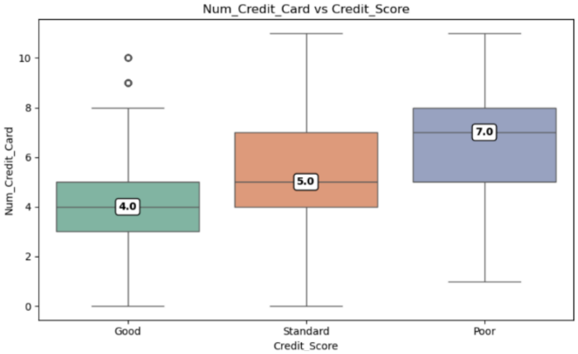


Figure 9- No of credit cards Vs Credit score

- Number of Credit Cards: The median number increased with worsening credit scores—4 for Good, 5 for Standard, and 7 for Poor credit scores *(Figure 9)*.
- Number of Loans: Customers with Good credit scores had a median of 2 loans, while those with Standard and Poor scores had medians of 3 and 5 loans respectively *(Figure 10)*.
- Delays from Due Date: Payment delay increased consistently with deteriorating credit scores. The median delay was 10 days for Good, 18 days for Standard, and 27 days for Poor credit scores *(Figure 11)*.

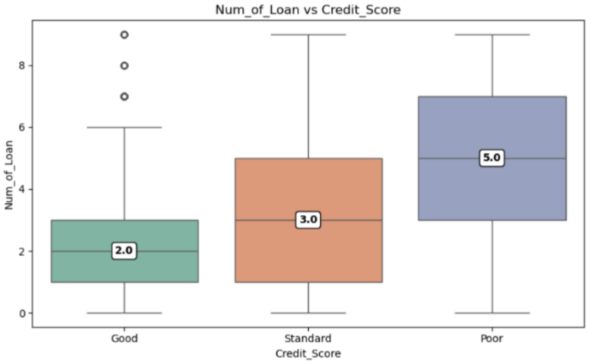


Figure 11- No of loan Vs Credit score

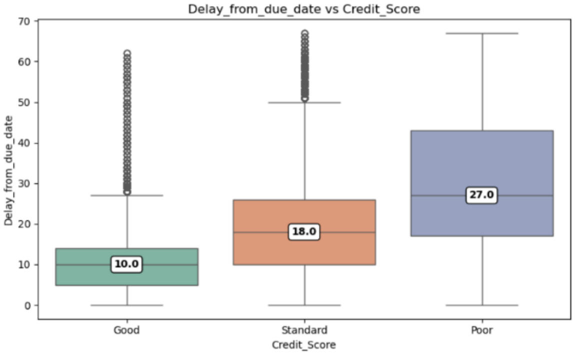


Figure 10-Delays from due date Vs Credit score

These findings indicate a **correlation** between higher debt load and payment delays with lower credit scores. However, it is important to note that **causal relationships cannot be inferred** from these associations alone, as this is purely observational analysis.

For **categorical variables**, *Payment_of_Min_Amount* emerged as one of the most influential predictors. Among those who responded "No," 38% had Good credit scores, whereas only 3.7% of customers who answered "Yes" belonged to the Good category. In contrast, 39.6% of the

"Yes" group were classified as Poor, suggesting that consistently paying only the minimum amount is strongly associated with poor creditworthiness.

Overall, the bivariate analysis highlights that **poorer credit scores tend to be associated with greater financial burden**, more frequent use of credit, and less responsible repayment behavior. These insights provide a critical foundation for identifying the most predictive variables for future classification models.

TEMPORAL PATTERN ANALYSIS

To examine whether credit score distributions vary over time, we analyzed the temporal behavior of the *Credit_Score* variable across the eight months from January to August. Overall, the distribution remained relatively stable throughout the period, with the "Standard" category consistently representing the majority (>50%) of observations each month *(Figure 12)*..

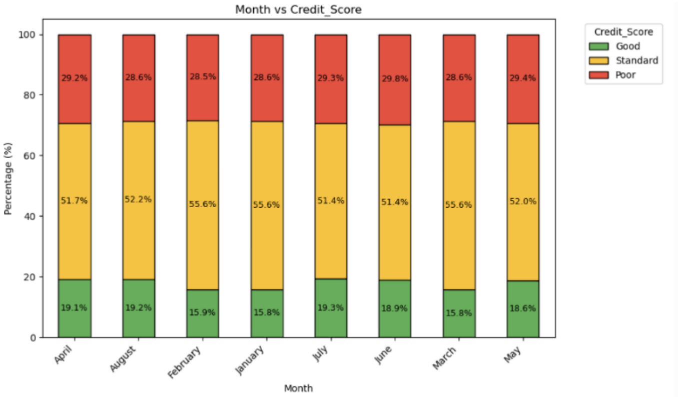


Figure 12-Month Vs Credit score

Notably, July recorded the highest proportion of customers with "Good" credit scores (19.3%), whereas the average across other months ranged between 16.5% and 18.5%. Meanwhile, the proportion of Poor credit scores remained steady at approximately 29% across all months. These minor month-to-month fluctuations suggest potential seasonal effects or reporting lags in credit behavior.

To assess the statistical significance of these differences, we conducted a Chi-square test of independence between the *Month* and *Credit_Score* variables. The test yielded a **p-value < 0.05**, indicating that the variation in credit score distribution across months is statistically significant, despite appearing visually modest. This supports the hypothesis that **time-dependent patterns** in financial behavior may influence creditworthiness.

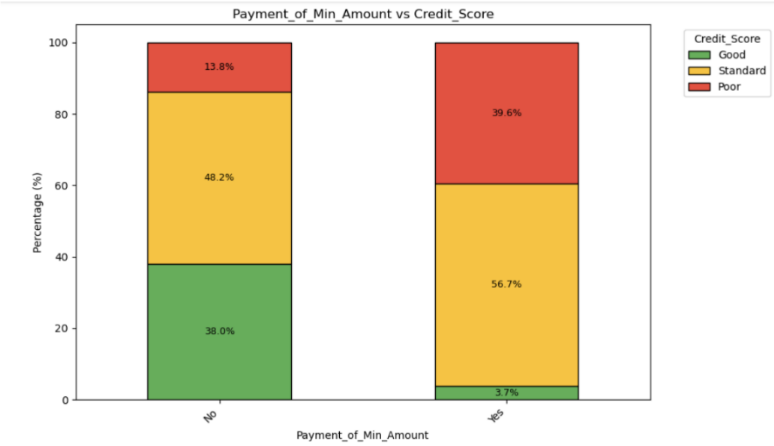


Figure 13-Payment of Min_Amount VS credit score

While these patterns are not drastic, the large sample size (n = 100,000) means even small shifts can be meaningful. In future work, time-series analysis or longitudinal modeling (e.g., mixed-effects models or survival analysis) could further explore how customer credit scores evolve over time in response to behavioral or economic triggers.

FAMD AND CLUSTERING

Factor Analysis of Mixed Data (FAMD):

To uncover latent patterns in mixed-type data, we employed Factor Analysis of Mixed Data (FAMD) using PCA as a proxy after one-hot encoding categorical variables and standardizing numerical features. This dimensionality reduction was followed by clustering to investigate potential subgroups within the customer population.

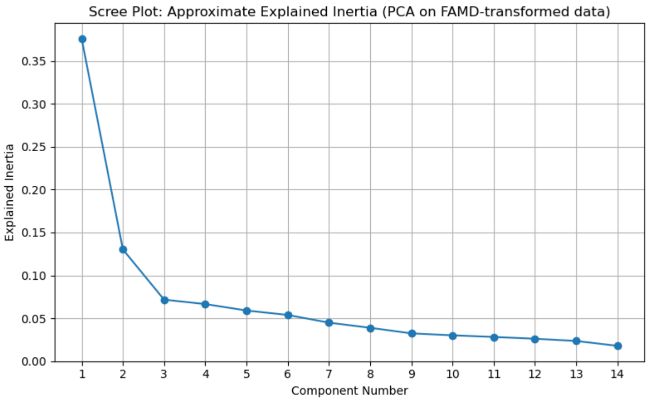


Figure 14-The scree plot of explained inertia

The scree plot of explained inertia *(Figure 14)* indicates that the PC1 alone accounts for approximately 36% of the variance, while the **first three components together explain around 58%** of the total variability. After the third component, the marginal gain from additional components diminishes, justifying our decision to visualize clusters in 3D using PC1, PC2, and PC3.

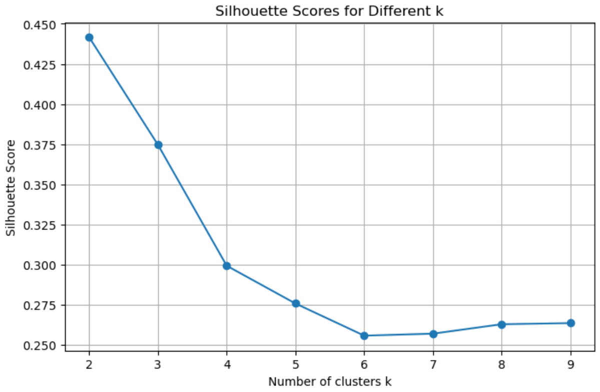


Figure 15-Silhouette Score for Different k

KMeans Clustering:

We applied KMeans clustering on the PCA-transformed data, testing cluster sizes from k = 2 to 9. The silhouette score peaked at **k = 2** with a value of **0.4416** *(Figure 15),* suggesting that a two-cluster solution offers the best balance of cohesion and separation. However, this silhouette score is moderate, indicating only limited structure in the data.

The 3D visualization *(Figure 16)* of the clusters in PCA space supports this result: while two clusters are visually discernible, they overlap significantly, and the separation boundary is not sharply defined. This suggests that **customer credit behavior does not strongly cluster** into distinct types based on the variables used in this analysis.

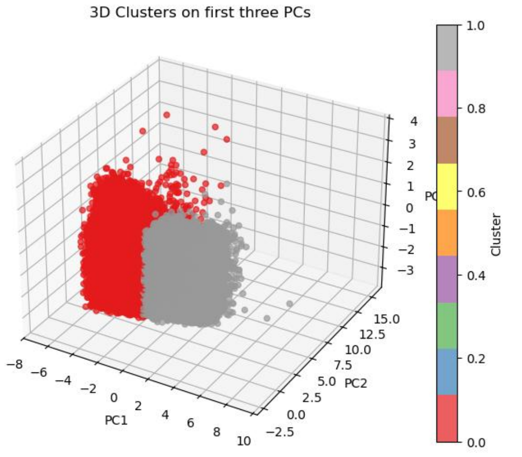


Figure 16- 3D clusters on 1st three PCs

Interpretation and Recommendation:

The FAMD-based clustering approach uncovered some latent structure, but the **weak cluster cohesion** implies that segmenting customers into distinct behavioral groups may not be effective with unsupervised methods alone. These results reinforce the decision to rely on supervised classification models in future analysis, where labeled credit score outcomes can guide learning.

Further improvements could involve:

- Applying **density-based clustering** (e.g., DBSCAN) to capture non-linear structures.
- Exploring **dynamic segmentation** using longitudinal behaviors over time.
- Re-assessing clustering after additional feature engineering or feature selection.

SUGGESTIONS FOR A QUALITY ADVANCED ANALYSIS

To extend this analysis into predictive modeling and improved insights:

- Implement classification models such as Logistic Regression, Decision Trees, Random Forests, SVM, or XGBoost to predict Credit Score
- Apply scaling and transformation techniques (e.g., log, Box-Cox) to handle outliers
- Perform cross-validation to evaluate model performance and generalization
- Use SHAP or LIME for model explainability and to understand individual-level predictions
- Investigate class imbalance handling techniques such as SMOTE

APPENDIX

EDA code

BIBLIOGRAPHY

Arram, A. a. (2023). Credit card score prediction using machine learning models: A new dataset. *arXiv preprint arXiv:2310.02956*.

Zen, M. F. (2025, February 14). Building a credit score model: understanding the business context and dataset. Medium. https://medium.com/@zaynmuhammad20/building-a-credit-score-model-understanding-the-business-context-and-dataset-426245d83e63

Zen, M. F. (2025, February 18). Building a credit score model: handling missing values and outliers. Medium. https://medium.com/@zaynmuhammad20/building-a-credit-score-model-handling-missing-values-and-outliers-ceb501b3b7b7

Zen, M. F. (2025, February 20). Building a credit score model: feature engineering and encoding. Medium. https://medium.com/@zaynmuhammad20/building-a-credit-score-model-feature-engineering-and-encoding-999373e0b9bb

Zen, M. F. (2025, February 20). Building a credit score model: Feature Selection - Muhammad Faizin Zen - Medium. Medium. https://medium.com/@zaynmuhammad20/building-a-credit-score-model-feature-selection-516e5c54ed53