Institute of Technology of Cambodia

Department of Applied Mathematics and Statistics

EDA AND UNSUPERVISED LEARNING REPORT

"Credit Risk EDA"

GROUP: I5-AMS-B (7)

Name	ID	Contributions
PHETH Soriyuon	e20210674	Data Cleaning and Processing
PHOEURN Kimhor	e20210823	EDA on Quantitative & Quantitative, Qualitative & Qualitative Columns.
Yin Sambat	e20210138	EDA on Quantitative & Qualitative Columns, Clustering.

Lecturer: **Dr. HAS Sothea**

Academic Year 2025-2026

Table of Content

I. INTRODUCTION

1.1. Objectives of the Analysis

The objective of this project is to analyze credit risk data to identify meaningful borrower segments and patterns associated with loan default. Specifically, the study aims to:

1. Ensure data quality by addressing missing values, duplicates, and outliers.
2. Explore borrower and loan characteristics through univariate and bivariate analyses.
3. Apply unsupervised learning methods, such as K-Means and Hierarchical Clustering, to group borrowers with similar profiles.
4. Use dimensionality reduction techniques to simplify visualization and interpretation of clusters.

1.2. Dataset Overview and Descriptive Statistics

The data used in this analysis is sourced from the Credit Risk Dataset available on OpenML (ID: 43454). This dataset is a standard benchmark in financial analytics, designed to help institutions evaluate the likelihood of loan defaults based on a borrower's financial and demographic profile. It contains historical information on loan applications, encompassing borrower demographics, financial standing, and loan-specific attributes.

	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status
0	22	59000	RENT	123.0	PERSONAL	D	35000	16.02	1
1	21	9600	OWN	5.0	EDUCATION	B	1000	11.14	0
2	25	9600	MORTGAGE	1.0	MEDICAL	C	5500	12.87	1
3	23	65500	RENT	4.0	MEDICAL	C	35000	15.23	1
4	24	54400	RENT	8.0	MEDICAL	C	35000	14.27	1

Figure 1: Five Rows of Data

The dataset contains 32,581 observations and 12 features that provide a comprehensive view of each applicant's creditworthiness:

Table 1: Features Overview

Feature Name	Description
person_age	Age of the borrower in years.
person_income	Annual income of the borrower.
person_home_ownership	Current housing situation.
person_emp_length	Total years of employment.
loan_intent	The specific purpose for the loan.
loan_grade	Risk-based grade assigned to the loan (A to G).
loan_amnt	Total amount of credit requested.
loan_int_rate	The annual interest rate for the loan.
loan_status	Target Variable: Indicates if a loan defaulted (1) or not (0).
loan_percent_income	The percentage of the borrower's income represented by the loan.

cb_person_default_on_file	Indicates if the borrower has a history of prior defaults.
cb_person_cred_hist_length	Number of years the borrower has had active credit history.

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
person_age	32581.0	27.734600	6.348078	20.00	23.00	26.00	30.00	144.00
person_income	32581.0	66074.848470	61983.119168	4000.00	38500.00	55000.00	79200.00	6000000.00
person_emp_length	31686.0	4.789686	4.142630	0.00	2.00	4.00	7.00	123.00
loan_amnt	32581.0	9589.371106	6322.086646	500.00	5000.00	8000.00	12200.00	35000.00
loan_int_rate	29465.0	11.011695	3.240459	5.42	7.90	10.99	13.47	23.22
loan_percent_income	32581.0	0.170203	0.106782	0.00	0.09	0.15	0.23	0.83
cb_person_cred_hist_length	32581.0	5.804211	4.055001	2.00	3.00	4.00	8.00	30.00

Figure 2: Descriptive Statistics of Data

From Figure 2, highlights data quality issues, with 895 missing values in *person_emp_length* and 3,116 missing values in *loan_int_rate*. Additionally, several variables show outliers. For instance, *person_age* has a mean of 27.73 and a median of 26, with a maximum value of 144, which is clearly implausible. Similarly, *person_income* reaches a maximum of $6,000,000, far above the 75th percentile of $79,200. These observations suggest the need for data cleaning.

II. DATA PREPROCESSING

2.1. Handling Missing Values

```
                   Missing Count  Percentage
person_emp_length            895    2.747000
loan_int_rate               3116    9.563856
```

Figure 3: Output from Checking Missing Value Code

Based on the output shown in Figure 3, there are 895 missing values in *person_emp_length* and 3,316 missing values in *loan_int_rate*. To assess the potential impact of removing these records, we examined how dropping the missing values affects the distributions of both numerical and categorical variables.

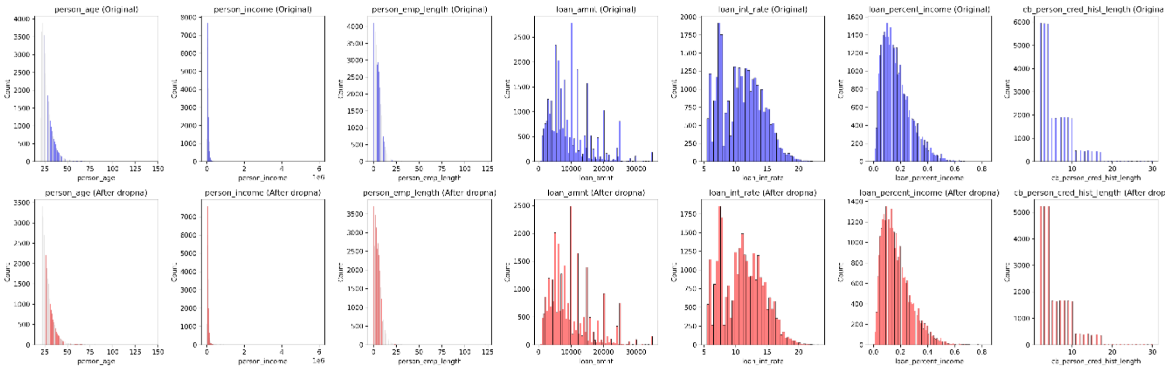


Figure 4: Comparison of Numerical Column Distribution before and after Dropping Missing Values

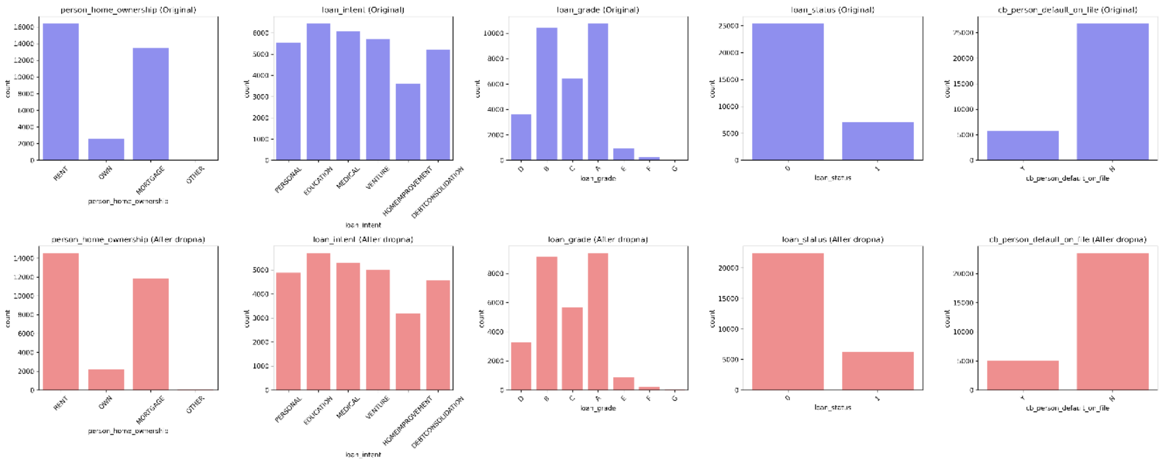


Figure 5: Comparison of Categorical Columns before and after Dropping Missing Values

Based on Figure 4 and Figure 5, dropping the missing values shows no significant impact on the distributions of the data. Therefore, the missing values were removed from the dataset.

2.2. Duplicate Removal

Ensuring the uniqueness of observations is critical to prevent bias in frequency counts and correlation coefficients. The dataset was scanned for identical entries across all 12 features. A total of 165 duplicate records were identified. These redundant rows were removed, resulting in a cleaner dataset where each row represents a unique loan application. This step ensures that the subsequent exploratory analysis is not skewed by over-represented borrower profiles.

2.3. Outlier Detection and Treatment

Using the boxplot in Figure 6, we investigate the extreme values in person_age, person_income, and person_emp_length.

Figure 6: Boxplot of Numerical Columns

Figure 7 highlights six extreme outliers in the dataset: four rows with ages over 100 years but short employment lengths (≤4 years), and two rows with implausibly long employment tenures (123 years) despite ages of 22 or younger. These six records, representing less than 0.1% of the dataset, were removed to maintain data integrity.

```python
df.query('person_age > 100')
```

✓ 0.0s Python

	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_statu
74	144	250000	RENT	4.0	VENTURE	C	4800	13.57	
163	144	200000	MORTGAGE	4.0	EDUCATION	B	6000	11.86	
508	123	80004	RENT	2.0	EDUCATION	B	20400	10.25	
28388	144	6000000	MORTGAGE	12.0	PERSONAL	C	5000	12.73	

```python
print("age only 21-22 but employment length is 123 years")
df.query('person_emp_length > 45')
```

✓ 0.0s Python

```
age only 21-22 but employment length is 123 years
```

	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status
0	22	59000	RENT	123.0	PERSONAL	D	35000	16.02	1
186	21	192000	MORTGAGE	123.0	VENTURE	A	20000	6.54	0

Figure 7: Rows with Extreme Values in person_age and person_emp_length

III. EXPLORATORY DATA ANALYSIS

3.1. Quantitative vs Quantitative

A pair plot shown in Figure 8 was used to visualize relationships among key numerical variables, including age, income, loan amount, loan interest rate, and credit history length. The plot highlights both variable distributions and inter-variable correlations.

Key Insights:

- **Age and Credit History:** A strong positive relationship is observed, as older individuals tend to have longer credit histories.
- **Loan Percent of Income:** Borrowers with lower incomes exhibit higher loan-to-income ratios, indicating greater lending risk.
- **Skewed Distributions:** Most variables are right-skewed, with concentrations at lower values for age, income, and employment length.

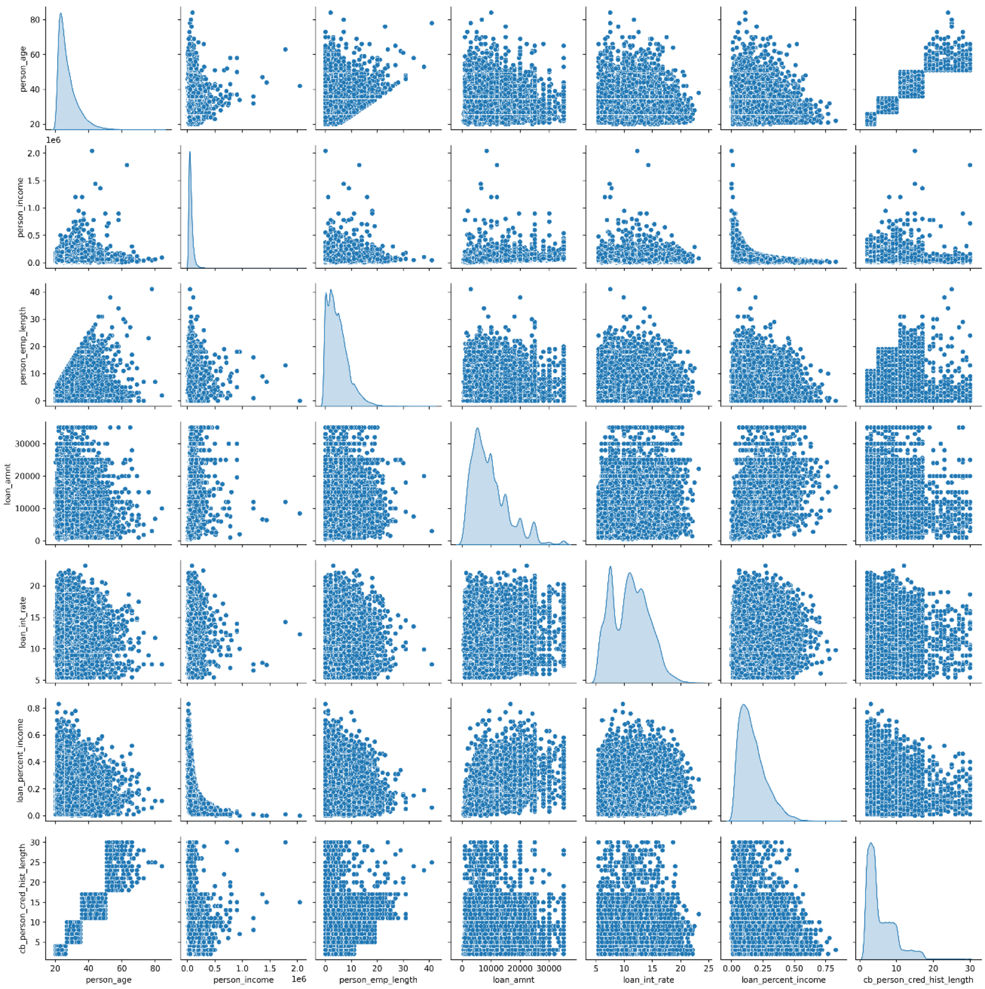


Figure 8: Pairplot of Quantitative Columns

3.2. Quantitative vs Qualitative

Before applying the Eta-squared test statistic to examine the relationship between numerical and categorical variables, it is necessary to first check for the presence of outliers, as the test is sensitive to extreme values. The results below show the outliers detected in the numerical columns of the dataset.

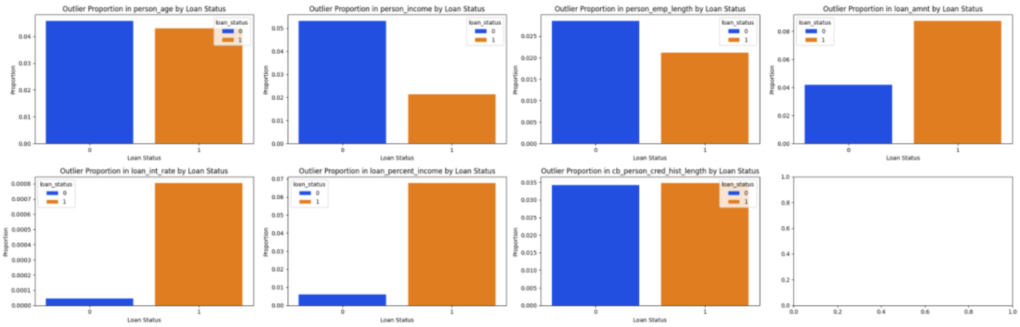


Figure 9: Outliers Ratio from Numerical Classes

From Figure 9, we observe that only Age and Credit History Length do not show dominance or imbalance between the two classes of Loan Status. In contrast, the remaining numerical variables exhibit noticeable imbalance between the two loan status classes. Therefore, we decided to address the outliers by capping the imbalanced classes at their lower and upper bounds, while removing outliers for the *Age* and *Credit History Length* variables. After this preprocessing step, we applied one-way ANOVA, and the results are shown below. It should be noted that we also experimented with removing outliers directly from all numerical columns; however, the results were consistent with those obtained using the capping approach.

Numeric Variable	Categorical Variable	Eta Squared	p-value
loan_int_rate	loan_grade	0.902809	0.00E+00
loan_int_rate	cb_person_default_on_file	0.250575	0.00E+00
loan_percent_income	loan_status	0.141794	0.00E+00
loan_int_rate	loan_status	0.114866	0.00E+00
person_income	person_home_ownership	0.108553	0.00E+00

The analysis indicates that Loan Interest Rate is strongly influenced by Loan Grade, as clearly illustrated in the conditional box plot in Figure 10.

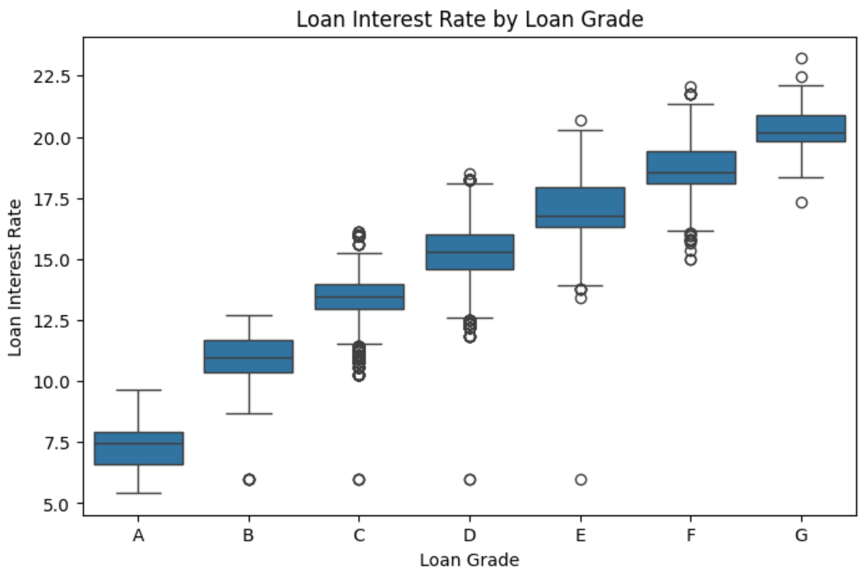
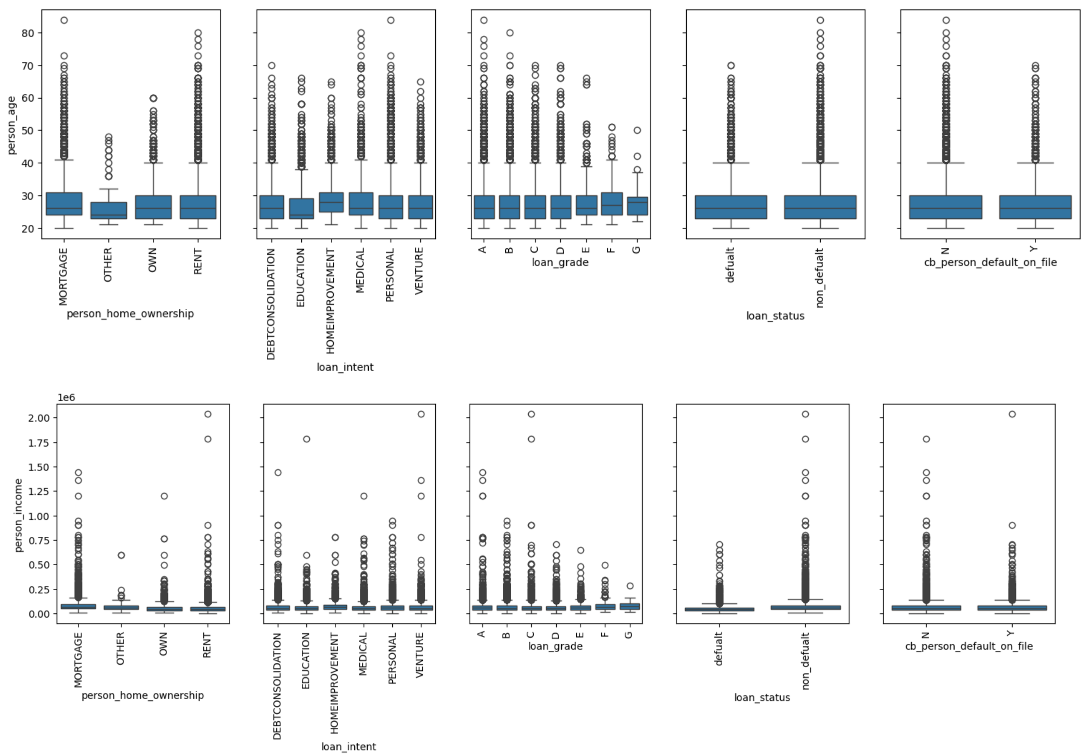


Figure 10: Relation of Rate and Grade

The conditional boxplots for the remaining variables indicate that the Eta-squared results are consistent with the visual analysis. Most categorical–numerical relationships appear weak, as the distributions across categories largely overlap. Notably, Loan Interest Rate shows a comparatively stronger association with Loan Grade, where clear differences in medians and spread can be observed across categories.



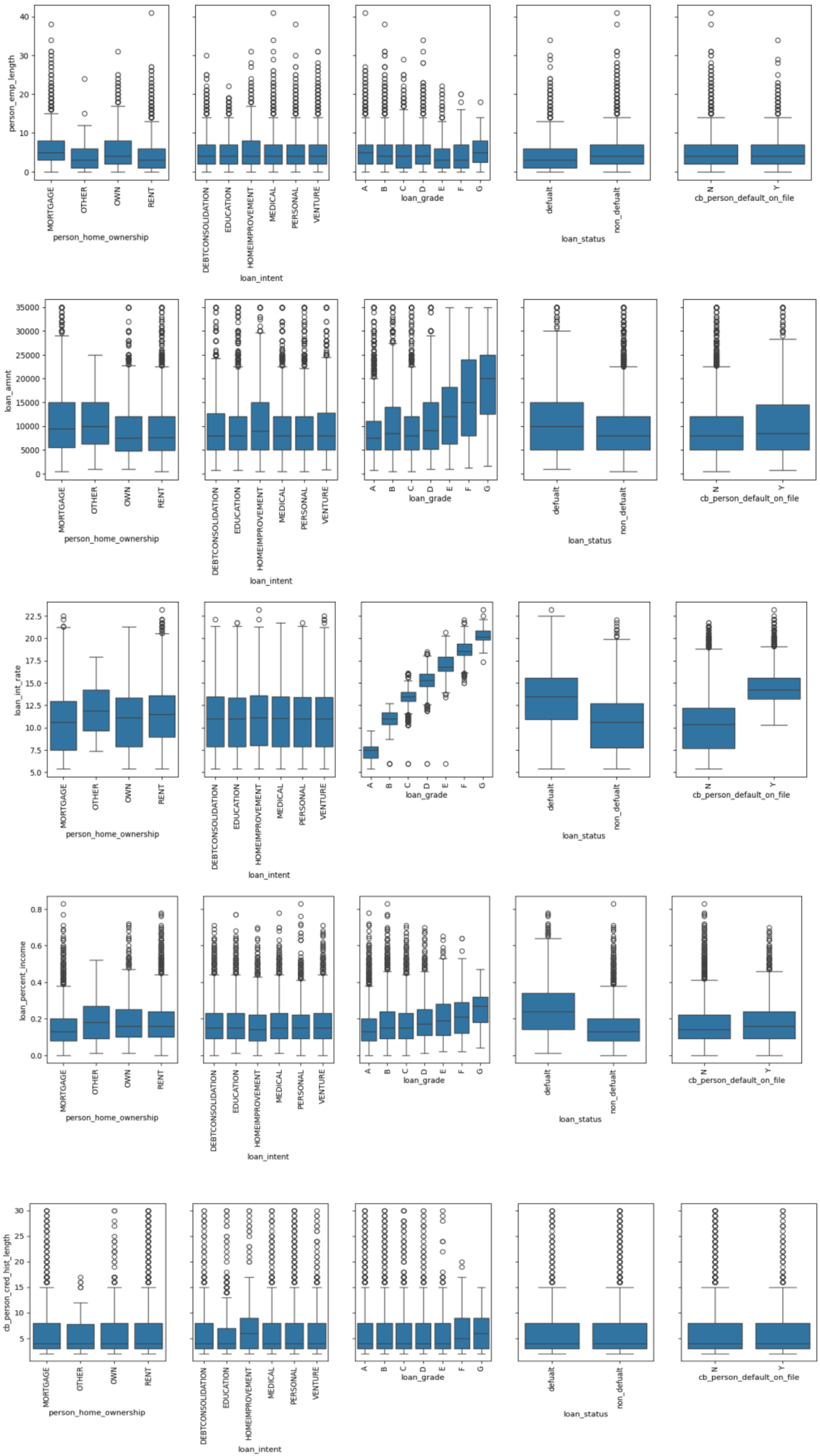


Figure 11: Conditional Boxplot across Category vs Numeric

Note that although outliers are visible in some boxplots in Figure 11, these represent within-group outliers rather than global outliers. Global outliers were previously identified and treated using the applied outlier handling techniques. Therefore, the remaining extreme values reflect natural variability within each categorical group and do not undermine the overall analysis.

Finally, the correlation plot below presents the relationships among the remaining numerical variables after preprocessing.

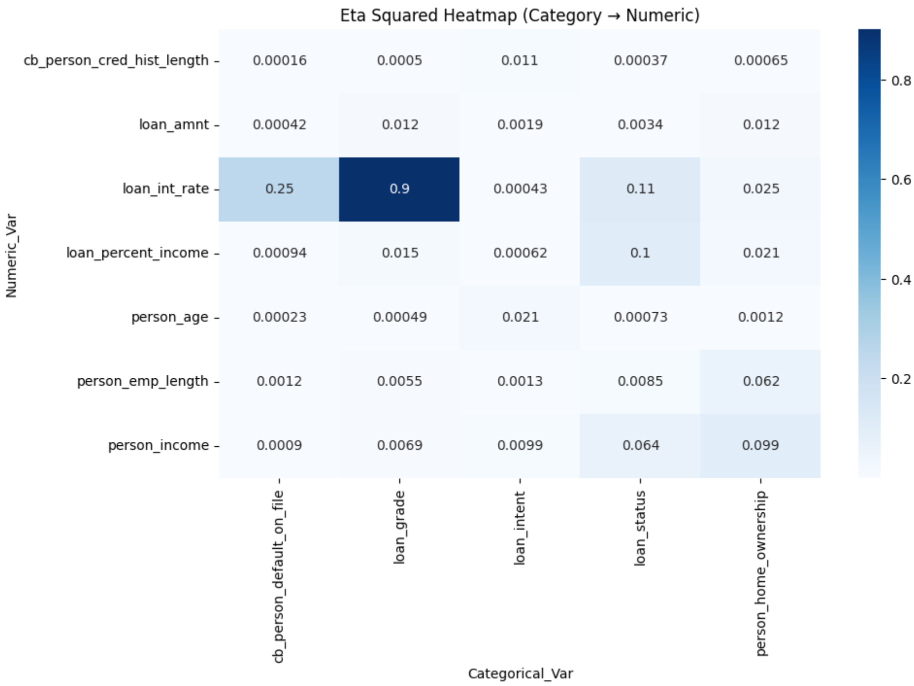


Figure 12: Heatmap of Relation Category vs Numerical

3.3. Qualitative vs Qualitative

As shown in Figure 13, grouped bar charts illustrate the relationships between categorical variables and loan status (default vs. non-default).

Key Findings:

- **Previous Default History:** Borrowers with a prior default record (*'Y'*) show a noticeably higher tendency to default again compared to those without previous defaults (*'N'*).
- **Loan Grade**: As the grade moves from **A → G**, the proportion of defaults increases significantly. In categories **D, E, F, and G**, the red bars (defaults) actually become comparable to or even surpass the blue bars in height, indicating much higher risk.
- **Loan Intent:** Loans for debt consolidation and medical purposes have higher default proportions, whereas education and venture loans exhibit lower risk.
- **Home Ownership: Renters** demonstrate the highest default rates, while **mortgage holders** and **owners** show greater repayment stability.

Overall, the figure suggests that home ownership and previous default history are strong qualitative indicators of credit risk, with loan intent acting as a moderate differentiating factor among borrowers.

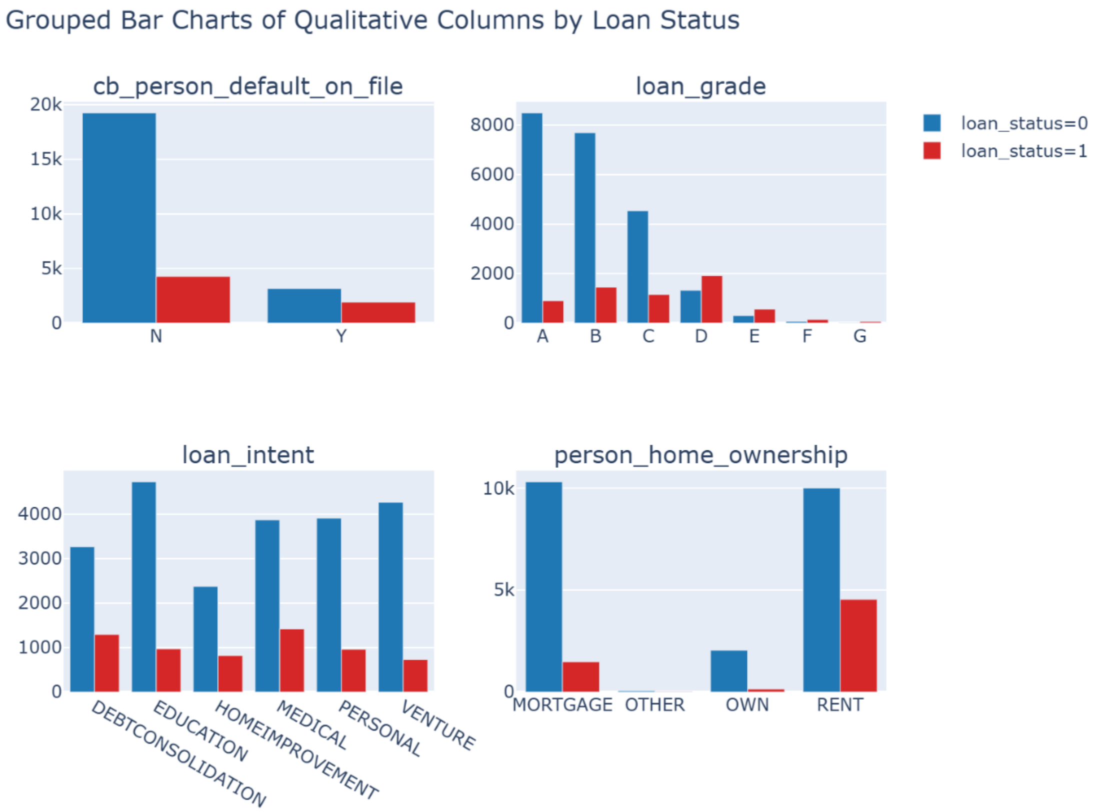


Figure 13: Grouped Bar of Qualitative Columns by Loan Status

Next, a Chi-square test of independence was performed between the categorical features to assess whether they are statistically associated. As shown in Figure 14, which displays the test output from the code, all categorical variable pairs have significant associations but only cb_person_default_on_file and loan_grade are strongly associated.

Var1	Var2	Chi2	p-value	CramersV	Significant (α=0.05)
cb_person_default_on_file	loan_grade	11437.053982	0.0000e+00	0.6319	True
loan_intent	person_home_ownership	652.866894	2.0323e-129	0.0862	True
loan_grade	person_home_ownership	623.069539	1.1375e-120	0.0839	True
cb_person_default_on_file	person_home_ownership	115.848172	6.0461e-25	0.0628	True
loan_grade	loan_intent	65.666394	1.8013e-04	0.0158	True
cb_person_default_on_file	loan_intent	7.403335	0.1923	0.0092	False

Figure 14: Output of the Chi-squared and Cramer's V tests between categorical columns

As shown in Figure 15, the symmetric biplot from Correspondence Analysis visualizes associations between history default and loan grade.

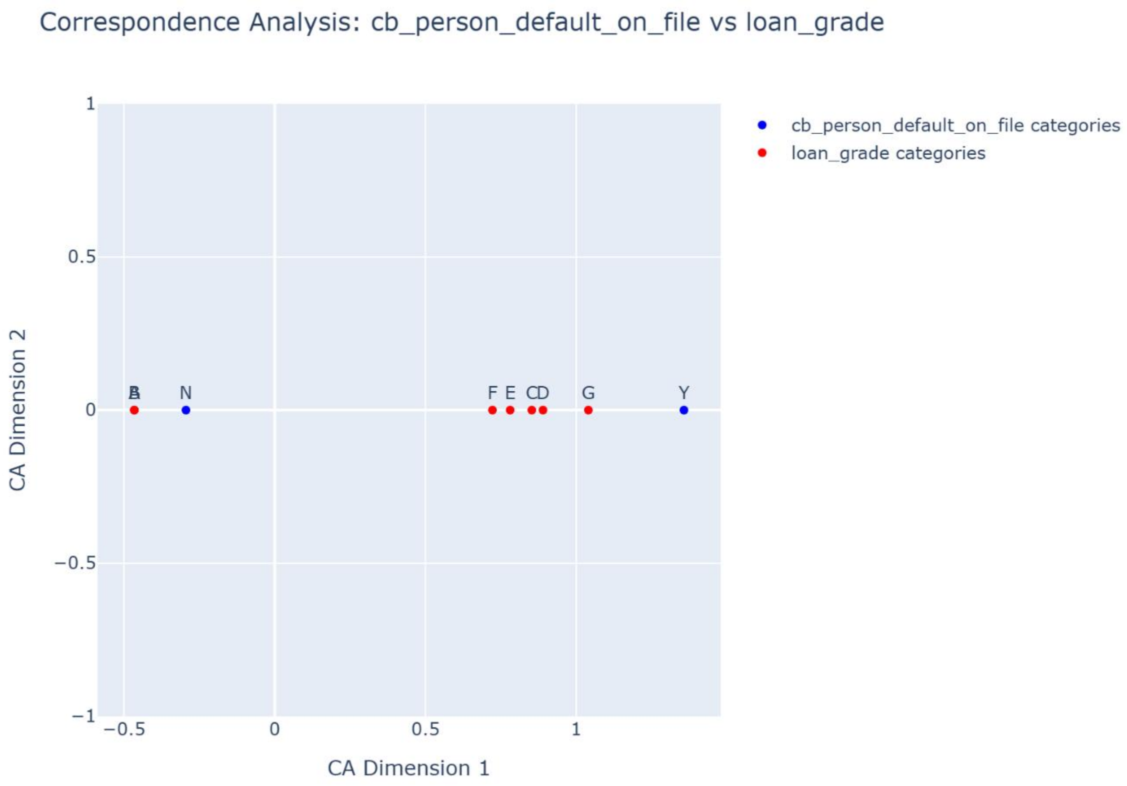


Figure 15: CA Symmetric Biplot

IV. UNSUPERVISED LEARNING

4.1. Data Preparation for Clustering

Outlier Management

Two outlier-handling strategies were implemented:

- **Capping:** The Interquartile Range (IQR) method was used to cap extreme values for variables such as person_income and loan_amnt at their respective upper and lower bounds. This approach is consistent with the preprocessing applied during the Eta-Square analysis.
- **Trimming:** Extreme outliers were removed from person_age and cb_person_cred_hist_length to prevent unusually large values from disproportionately influencing cluster centroids.

Feature Engineering

- **Encoding:** Categorical variables were transformed using One-Hot Encoding, converting them into dummy variables suitable for distance-based clustering algorithms.
- **Scaling:** Numerical features were standardized using StandardScaler to ensure that all variables contribute equally to distance calculations, which is particularly important for K-Means clustering.

After preprocessing, the scaled numerical features and encoded categorical features were **combined into a single feature matrix** for clustering.

4.2. Choice of Clustering Algorithm

4.2.1. K-Means Clustering

K-Means clustering is a widely used unsupervised learning algorithm that partitions data into k distinct clusters by minimizing the within-cluster sum of squared distances between data points and their assigned cluster centroids.

The algorithm iteratively assigns data points to the nearest centroid and updates centroid positions until convergence. K-Means is computationally efficient and commonly used for applications such as customer segmentation and pattern discovery.

4.2.2. Hierarchical Clustering

Hierarchical clustering is an unsupervised learning technique that organizes data points into a nested hierarchy of clusters, typically visualized using a dendrogram.

Unlike K-Means, it does not require the number of clusters to be specified in advance. This study employs agglomerative (bottom-up) hierarchical clustering, where individual data points are progressively merged based on similarity until all points form a single cluster.

Hierarchical clustering is useful for revealing natural relationships and cluster structures within the data.

4.2.3. Spectral Clustering

Spectral clustering is a graph-based clustering method capable of identifying clusters with complex or non-convex shapes. It works by constructing a similarity graph from the data, computing the graph Laplacian, and projecting the data into a lower-dimensional space using its eigenvectors.

Standard clustering algorithms, such as K-Means, are then applied in this transformed space. Spectral clustering is particularly effective when cluster boundaries are not well separated in the original feature space.

4.3. Determining Optimal Number of Clusters

4.3.1. Elbow Method

The Elbow Method is a visual technique commonly applied to K-Means clustering. It plots the Within-Cluster Sum of Squares (WCSS) against different values of k.

The optimal number of clusters is identified at the "elbow" point, where the reduction in WCSS begins to slow significantly, indicating diminishing returns from adding more clusters.

4.3.2. Silhouette Score

The Silhouette Score measures the quality of clustering by comparing a data point's similarity to its own cluster (cohesion) with its similarity to other clusters (separation).

The score ranges from **-1 to +1**, where higher values indicate better-defined clusters. Values close to zero suggest overlapping clusters.

Although useful for selecting the optimal number of clusters, silhouette scores tend to work best for compact and well-separated clusters.

4.4. Cluster Profiling and Interpretation

Based on the K-Means results, the Elbow Method suggests that increasing the number of clusters continually reduces WCSS, without a clear elbow point. However, the Silhouette Score reaches its maximum at K = 2, indicating that a two-cluster solution provides the best separation among the tested values. Despite this, the highest silhouette score achieved is approximately 0.17, which is relatively low. This suggests that the clusters are weakly separated, and the underlying structure in the data may not be strongly clusterable.

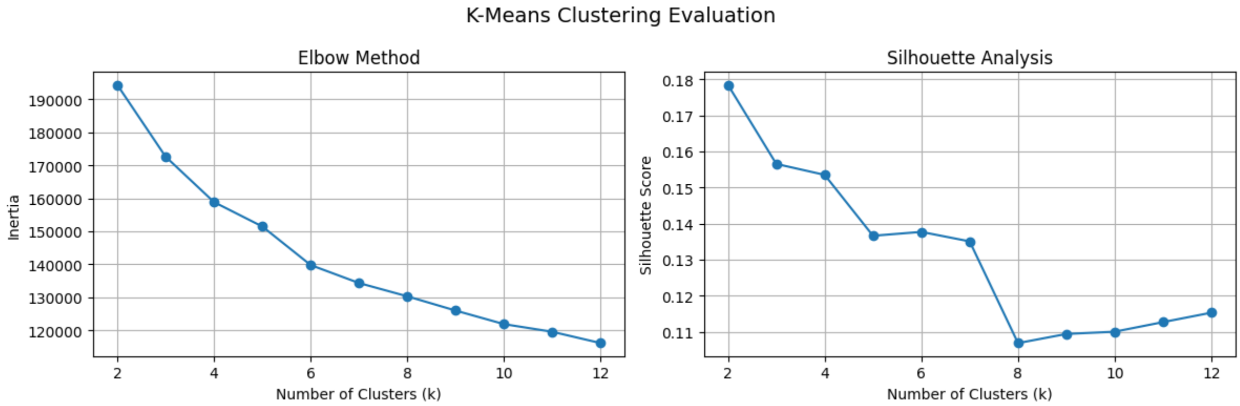


Figure 16: Elbow and Silhouette of K-Means

Applying Spectral Clustering yields similar results, with K = 2 again producing the highest silhouette score, though the value is even lower (approximately 0.12), reinforcing the observation of limited natural separation in the dataset.

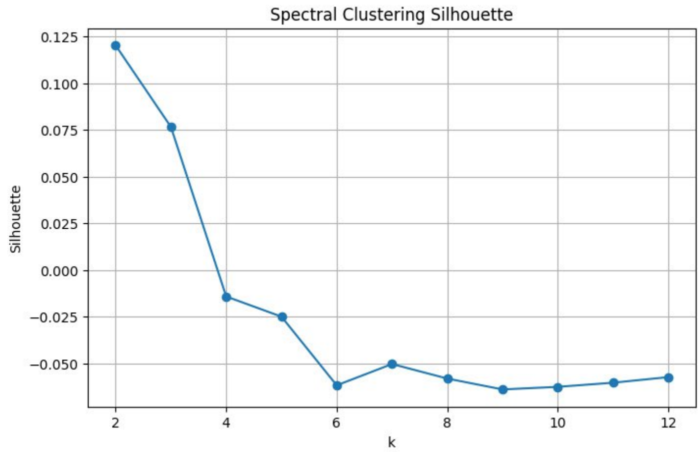


Figure 17: Silhouette score of Spectral Cluster

In contrast, Hierarchical Clustering produces a different outcome. By examining the dendrogram, a noticeable increase in linkage distance occurs around **150**, suggesting a cut that results in **three clusters**. This indicates that hierarchical clustering may capture structural relationships in the data that are not evident using centroid-based methods.

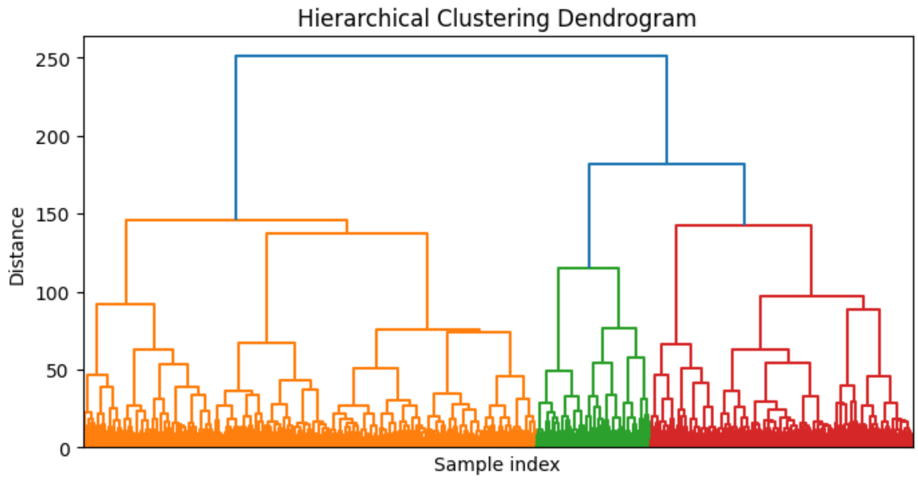


Figure 18: Dendrogram of Hierarchical Cluster

It is also worth noting that experiments were conducted using only a subset of selected features for clustering; however, these attempts did not lead to any noticeable improvement in clustering performance and, in some cases, resulted in poorer outcomes.

V. CONCLUSION

5.1. Summary of Key Findings

The exploratory analysis revealed important insights regarding data quality, borrower characteristics, and risk indicators. Missing values were identified in employment length (895 records) and loan interest rates (3,316 records) and were removed without significantly affecting overall data distributions. Extreme outliers, such as implausible ages up to 144 years and employment lengths of 123 years, were capped or removed to maintain data integrity.

Demographically, borrower age was positively correlated with the length of credit history, and lower-income borrowers typically exhibited higher loan-to-income ratios, indicating increased lending risk. Loan grades strongly influenced interest rates, with clear differences in medians across grades (Eta-squared = 0.90).

Key risk indicators included loan grade, prior default history, home ownership, and loan purpose. Borrowers with lower loan grades (D–G) and those with prior defaults were more likely to default again. Renters had the highest default rates, while mortgage holders and homeowners demonstrated greater repayment stability. Loans for debt consolidation or medical purposes carried higher risk compared to loans for education or business ventures.

Unsupervised learning analyses suggested that K-Means and Spectral clustering produced a two-cluster solution, but low silhouette scores (~0.17) indicated weak natural separation among borrowers. Hierarchical clustering, however, identified three clusters, capturing potential structural relationships that centroid-based methods may have missed.

5.2. Limitations of the Analysis

- The study relied solely on the OpenML Credit Risk Dataset, which may not fully represent real-world lending scenarios across different regions or institutions.
- Unsupervised methods struggled to identify distinct segments.

REFERENCES

Data Source:
https://www.openml.org/search?type=data&id=43454&sort=runs&status=active

Source Code: https://github.com/phoeurnkimhor/credit-risk-eda