Exploratory Data Analysis (EDA) Summary

1. Introduction

This report provides an initial exploratory analysis of Geldium's credit delinquency dataset, focusing on data quality, missing value handling, and the identification of early risk indicators relevant for predictive modeling.

2. Dataset Overview

- Number of records: 500
- Key variables: Customer_ID, Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Delinquent_Account, Loan_Balance, Debt_to_Income_Ratio, Employment_Status, Account_Tenure, Credit_Card_Type, Location, Month_1, Month_2, Month_3, Month_4, Month_5, Month_6
- Data types:
    - Numerical: Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Delinquent_Account, Loan_Balance, Debt_to_Income_Ratio, Account_Tenure
    - Categorical: Customer_ID, Employment_Status, Credit_Card_Type, Location, Month_1, Month_2, Month_3, Month_4, Month_5, Month_6

3. Missing Data Analysis

Key missing data findings:

- Income: 7.8% missing
- Loan_Balance: 5.8% missing
- Credit_Score: 0.4% missing

Missing data treatment:

- Income: Median impute
- Loan_Balance: Median impute
- Credit_Score: KNN imputation for small missing fraction

Missing Data Issue	Handling Method	Rationale
Income missing (~7.8%)	Median impute	Maintains central tendency and is robust to rare values
Loan_Balance missing (~5.8%)	Median impute	Simple approach preserves representative values
Credit_Score missing (~0.4%)	KNN impute	Small fraction; uses nearest neighbors and existing relationships

4. Key Findings and Risk Indicators

Notable missing or inconsistent data:

- Income and Loan_Balance have highest missingness; Employment_Status needs standardization
- Credit_Utilization mean is 0.49; review for scale (0–1 vs 0–100)

Key anomalies:

- Occasional outliers in Account_Tenure values; inconsistent categorical labeling

Early indicators of delinquency risk:

- High CreditUtilization: Signals overextension; delinquency odds rise beyond ~50% utilization
- Recent Missed/Late Payments: Direct evidence of repayment stress; strong predictor
- High Debt-to-Income Ratio: Elevated repayment burden relative to income
- Multiple Recent Inquiries: May indicate liquidity strain or credit shopping
- Low/Unstable Income or Unemployment: Reduces repayment reliability

5. GenAI & Automated Analytics Usage

GenAI tools were leveraged to systematically summarize missing values, suggest appropriate imputation strategies, and identify features with the highest potential risk for prioritization. All outputs were cross-validated against industry-standard practices and domain knowledge to ensure reliability and relevance. This approach accelerated insight generation while maintaining analytical rigor.

6. Conclusion & Next Steps

The initial EDA highlights that targeted data cleaning is needed, particularly for **Income** and **Loan_Balance**, along with standardization of categorical fields to ensure consistency before modeling. For feature engineering, focus should be on **Credit_Utilization**, **Missed_Payments**, and **Debt_to_Income_Ratio**, as they are likely strong predictors of risk.

Additionally, consider **segmenting customers based on account tenure**. Different tenure groups (e.g., new, medium-term, and long-term customers) often exhibit distinct patterns in payment behavior and credit usage. Tailoring models or analyses for these segments can improve predictive accuracy and enable more precise risk management strategies.