# Geldium — Exploratory Data Analysis (EDA) Summary

Prepared for: Tata iQ analytics team & Geldium decision-makers

## 1. Introduction

Purpose: Initial EDA of Geldium's delinquency dataset to (1) evaluate data quality and completeness, (2) identify early risk indicators for delinquency modelling, and (3) recommend preprocessing and imputation steps prior to model training.

## 2. Dataset Overview

Number of records: 500

Number of columns: 19

Key variables and inferred data types:

* - Customer\_ID: object
* - Age: int64
* - Income: float64
* - Credit\_Score: float64
* - Credit\_Utilization: float64
* - Missed\_Payments: int64
* - Delinquent\_Account: int64
* - Loan\_Balance: float64
* - Debt\_to\_Income\_Ratio: float64
* - Employment\_Status: object
* - Account\_Tenure: int64
* - Credit\_Card\_Type: object
* - Location: object
* - Month\_1: object
* - Month\_2: object
* - Month\_3: object
* - Month\_4: object
* - Month\_5: object
* - Month\_6: object

Number of duplicate rows: 0

## 3. Missing Data Analysis

Columns with missing values (column: missing\_count — %missing):

* - Income: 39 — 7.8%
* - Loan\_Balance: 29 — 5.8%
* - Credit\_Score: 2 — 0.4%

Proposed handling (examples):

* - Income: 39 missing (7.8%) — Suggested: Impute with median (robust to outliers) or use predictive imputation.
* - Loan\_Balance: 29 missing (5.8%) — Suggested: Impute with median (robust to outliers) or use predictive imputation.
* - Credit\_Score: 2 missing (0.4%) — Suggested: Impute with median (robust to outliers) or use predictive imputation.

## 4. Key Findings and Risk Indicators

Summary statistics for numeric variables (selected):

* - Age: mean=46.27, std=16.19, min=18.0, 25%=33.0, 50%=46.5, 75%=59.25, max=74.0
* - Income: mean=108379.89, std=53662.72, min=15404.0, 25%=62295.0, 50%=107658.0, 75%=155734.0, max=199943.0
* - Credit\_Score: mean=577.72, std=168.88, min=301.0, 25%=418.25, 50%=586.0, 75%=727.25, max=847.0
* - Credit\_Utilization: mean=0.49, std=0.20, min=0.05, 25%=0.356486053, 50%=0.48563557700000004, 75%=0.63444035075, max=1.025842526
* - Missed\_Payments: mean=2.97, std=1.95, min=0.0, 25%=1.0, 50%=3.0, 75%=5.0, max=6.0
* - Delinquent\_Account: mean=0.16, std=0.37, min=0.0, 25%=0.0, 50%=0.0, 75%=0.0, max=1.0
* - Loan\_Balance: mean=48654.43, std=29395.54, min=612.0, 25%=23716.5, 50%=45776.0, 75%=75546.5, max=99620.0
* - Debt\_to\_Income\_Ratio: mean=0.30, std=0.09, min=0.1, 25%=0.23363875274999998, 50%=0.301633935, 75%=0.36273739024999996, max=0.552956451

Top predictors (quick RandomForest feature importances):

* - Credit\_Score: importance=0.090
* - Debt\_to\_Income\_Ratio: importance=0.090
* - Customer\_ID: importance=0.086
* - Loan\_Balance: importance=0.085
* - Credit\_Utilization: importance=0.085
* - Income: importance=0.081
* - Age: importance=0.076
* - Account\_Tenure: importance=0.075
* - Employment\_Status: importance=0.049
* - Credit\_Card\_Type: importance=0.044

Notable anomalies detected:

- No major anomalies detected based on simple threshold checks.

## 5. Risk Indicators (Early)

* - Missed\_Payments: Past missed payments are a strong predictor of future delinquency.

## 6. Missing Data Treatment Table (examples)

|  |  |  |
| --- | --- | --- |
| Issue (column) | Proposed Handling | Justification |
| Income (39 missing, 7.8%) | Impute with median (robust to outliers) or use predictive imputation. | Chosen to preserve signal and avoid bias; robust to outliers when using median. |
| Loan\_Balance (29 missing, 5.8%) | Impute with median (robust to outliers) or use predictive imputation. | Chosen to preserve signal and avoid bias; robust to outliers when using median. |
| Credit\_Score (2 missing, 0.4%) | Impute with median (robust to outliers) or use predictive imputation. | Chosen to preserve signal and avoid bias; robust to outliers when using median. |

## 7. AI & GenAI Usage

GenAI-assisted steps used:

- Summarized dataset patterns and missing-value issues programmatically.

- Suggested imputation strategies and synthetic-data options for high-missing columns.

Example prompts used: 'Summarize key patterns, outliers, and missing values in this dataset. Highlight fields that might present problems for modeling delinquency.'

## 8. Conclusion & Next Steps

The dataset shows several missing-value issues and a small number of duplicates. Numeric fields generally have plausible ranges, but there are instances of unusually high credit utilization and potential negative incomes that require investigation. Early risk indicators include credit\_utilization, missed payment history, and credit\_score — these variables should be prioritized for feature engineering and model training.

Recommended next steps:

- Perform targeted imputation (median for numeric, mode or predictive for categorical) and create missing-indicator flags for important fields.

- Engineer features: rolling payment patterns (from Month\_1..Month\_6), recent changes in utilization, and employment stability.

- Run a proper model pipeline with cross-validation, calibrate thresholds, and analyze fairness checks across demographic groups.