Exploratory Data Analysis (EDA) Summary Report

1. Introduction

This report summarizes an initial exploratory data analysis (EDA) of Geldium's customer dataset to assess data quality and identify early risk indicators for credit card delinquency. The findings will guide missing-value treatment and help prioritize variables for delinquency risk modeling and intervention strategies.

2. Dataset Overview

This section summarizes the dataset, including the number of records, key variables, and data types. It also highlights any anomalies, duplicates, or inconsistencies observed during the initial review.

Key dataset attributes:

- Number of records: 500 customers

- Number of columns: 19

- Income missing = 39

- Credit_Score missing = 2

- Target variable: Delinquent_Account (0 = No, 1 = Yes)

- Target distribution: 420 non-delinquent (84%), 80 delinquent (16%)

- Data types: 9 numerical, 10 categorical (incl. Month_1–Month_6 payment history)

- Key variables:

Demographics/financial: Age, Income, Debt_to_Income_Ratio

Credit profile: Credit_Score, Credit_Utilization

Behavior/payment: Missed_Payments, Month_1–Month_6

Account context: Account_Tenure, Loan_Balance, Credit_Card_Type, Employment_Status, Location

3. Missing Data Analysis

Identifying and addressing missing data is critical to ensuring model accuracy. This section outlines missing values in the dataset, the approach taken to handle them, and justifications for the chosen method.

Key missing data findings:

- Variables with missing values:1. Income: 39 missing (~7.8%)

2. Credit_Score: 2 missing (~0.4%)

- Missing data treatment:

1. Income: impute using **median Income** (robust to outliers) OR median by **Employment_Status** if income differs strongly by employment group.

2. Credit_Score: impute using **median Credit_Score** (only 2 values missing, low risk).

Missingness is limited and concentrated in two fields; median imputation preserves dataset size and reduces bias from dropping records.

4. Key Findings and Risk Indicators

This section identifies trends and patterns that may indicate risk factors for delinquency. Feature relationships and statistical correlations are explored to uncover insights relevant to predictive modeling.

Key findings:

Correlations / relationships to check (risk signals)

1. **Payment history (Month_1–Month_6):** Customers with repeated **Late/Missed** statuses across recent months are more likely to be delinquent.
2. **Missed_Payments:** A higher count of missed payments in the last 12 months is a strong early warning indicator.
3. **Credit_Utilization:** Higher utilization (using a large % of available credit) typically indicates financial stress and may correlate with delinquency.

4. **Debt_to_Income_Ratio:** Higher DTI suggests reduced ability to repay and is a key risk indicator.
5. **Credit_Score:** Lower credit scores generally indicate higher delinquency risk.

Top 3 variables likely to predict delinquency (with brief reasoning)

1. **Month_1–Month_6 payment history** — captures recent repayment behavior directly (most predictive for near-term delinquency).
2. **Missed_Payments** — summarises historical missed events over 12 months (strong signal of repeat behavior).
3. **Credit_Utilization / Debt_to_Income_Ratio** — reflects financial stress and repayment capacity (choose one as #3 if you want only one).

Unexpected anomalies to flag for investigation (data quality)

- Validate **Credit_Utilization** is within expected bounds (e.g., 0–1 or 0–100%) and has no negative or extreme outliers.
- Validate **Debt_to_Income_Ratio** is within expected bounds (0–1 or 0–100%).
- Validate **Credit_Score** values fall within a realistic range (typically 300–850).
- Check for any customers with **high Income + high Missed_Payments** (could be genuine but worth validating).

5. AI & GenAI Usage

Generative AI tools were used to summarize the dataset, impute missing data, and detect patterns. This section documents AI-generated insights and the prompts used to obtain results.

How GenAI was used (safely):

- Used GenAI to summarize the dataset structure, list key variables relevant to delinquency, and suggest best-practice missing value handling (no sensitive customer data was shared—only column names and aggregated counts).

- Used GenAI to help draft clear, executive-friendly bullet points for risk indicators and data quality observations.

Prompts used (examples):

1. “Given these columns (Age, Income, Credit_Score, Credit_Utilization, missed_Payments, Debt_to_Income_Ratio, Month_1–Month_6, Delinquent_Account), identify the top predictors of delinquency and explain why.”

2. “Income has 39 missing values and Credit_Score has 2 missing values out of 500 records. Recommend an imputation strategy and explain the trade-offs.”

3. “Draft a concise EDA summary for a Collections team: key data quality issues, key risk indicators, and next steps.”

6. Conclusion & Next Steps

Geldium’s dataset is largely complete and suitable for delinquency risk analysis, with missing values limited to Income (39 records) and Credit_Score (2 records). The target outcome shows moderate imbalance (16% delinquent), which should be considered during model evaluation. The most relevant risk indicators to prioritize are recent payment history (Month_1–Month_6), Missed_Payments, and financial stress measures such as Credit_Utilization and Debt_to_Income_Ratio. Next steps are to impute missing Income and Credit_Score values using median-based methods, validate numeric ranges for key fields, and then proceed to model development with appropriate metrics (e.g., precision/recall) to support targeted Collections interventions.