# Exploratory Data Analysis (EDA) Summary Report

## 1. Introduction

This report presents the findings from an Exploratory Data Analysis (EDA) conducted on Geldium Finance's customer dataset. The primary objective is to identify patterns and risk indicators related to credit card delinquency, address data quality issues, and guide the development of AI-powered predictive models.

## 2. Dataset Overview

The dataset contains 501 records with key attributes relevant to assessing delinquency risk.

**Key dataset attributes:**

- Number of records: 501
- Key variables: Age, Monthly Income, Credit Score, Credit Utilization, Loan Balance, and Delinquency Status
- Data types: Age (Numerical), Monthly Income (Numerical), Credit Score (Numerical), Credit Utilization (Numerical), Loan Balance (Numerical), Delinquency Status (Categorical)

**Initial observations:**

- Monthly Income has 39 missing entries
- Loan Balance has 29 missing entries
- Credit Score has 2 missing entries
- Data is largely clean with no duplicate records or invalid types

## 3. Missing Data Analysis

**Key missing data findings:**

- Variables with missing values: Monthly Income, Credit Score, and Loan Balance
- Missing data treatment:

- Monthly Income and Loan Balance were filled using median imputation to preserve the central tendency without skewing the data.
- The 2 rows with missing Credit Score were removed to avoid introducing potential bias with imputation due to the very low number of missing values.

## 4. Key Findings and Risk Indicators

**Key findings:**

- High credit utilization is strongly correlated with higher delinquency likelihood.
- Lower credit scores are a consistent risk factor.
- Customers with high loan balances and low income are more likely to miss payments.
- A few high-income customers still show delinquency, indicating other behavioral factors.

**Unexpected anomalies:**

- Some customers with stable income and credit score still appear as delinquent, warranting further behavioral analysis.

## 5. AI & GenAI Usage

Generative AI tools like ChatGPT were used to analyze patterns, handle missing data, and surface key risk indicators. The following prompts were utilized:

- 'Summarize key patterns in the dataset and identify anomalies.'
- 'Suggest an imputation strategy for missing income values based on industry best practices.'
- 'List top features that may indicate customer delinquency based on historical data patterns.'

## 6. Conclusion & Next Steps

The EDA process revealed clear relationships between high credit utilization, low credit scores, and increased risk of delinquency. By addressing missing data and identifying key risk indicators, the foundation has been laid for building accurate AI models. Next steps include developing and validating predictive models and designing intervention strategies for at-risk customers.

**Next Steps:**

- Use the cleaned dataset to train AI models that predict delinquency.
- Validate key predictors like credit utilization across sub-groups.
- Investigate outliers (e.g., high income but delinquent cases) for deeper insights.
- Collaborate with Geldium's Collections team to create data-backed intervention strategies.