**Exploratory Data Analysis (EDA) Summary**   
 **Report**

# 1. Introduction

The purpose of this report is to conduct an Exploratory Data Analysis (EDA) on Geldium’s dataset for Tata iQ’s analytics team. The goal is to assess data quality, identify risk indicators, and prepare for building a delinquency prediction model.

# 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

- Key variables: - Customer\_ID: Unique identifier for each customer  
- Age: Customer's age  
- Income: Reported annual income  
- Credit\_Score: Numeric representation of creditworthiness  
- Credit\_Utilization: Ratio of used credit to total available credit  
- Missed\_Payments: Count of missed payments  
- Delinquent\_Account: Indicator of whether the account is delinquent  
- Loan\_Balance: Outstanding loan balance  
- Debt\_to\_Income\_Ratio: Financial risk measure  
- Employment\_Status, Credit\_Card\_Type, Location: Categorical features  
- Month\_1 to Month\_6: Payment history over six months

- Data types: Numerical (Age, Income, Credit\_Score, Credit\_Utilization, Missed\_Payments, Loan\_Balance, Debt\_to\_Income\_Ratio), Categorical (Employment\_Status, Credit\_Card\_Type, Location, Month\_1 to Month\_6)

# 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: - Income: 39 missing values  
- Credit\_Score: 2 missing values  
- Loan\_Balance: 29 missing values

- Missing data treatment: - Income: Imputed using median due to skewed distribution  
- Credit\_Score: Imputed using mean due to low missing count  
- Loan\_Balance: Imputed using median to preserve distribution and mitigate outliers

# 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 observed between key variables: - High Missed\_Payments correlates with Delinquent\_Account  
- Lower Credit\_Score often appears in delinquent cases  
- High Credit\_Utilization and high Debt\_to\_Income\_Ratio are strong risk indicators

- Unexpected anomalies: - Customers with 0 Account\_Tenure and multiple missed payments  
- Inconsistent monthly payment histories that conflict with Delinquent\_Account = 0

# 5. AI & GenAI Usage

Generative AI tools (ChatGPT) were used to summarize the dataset, detect anomalies, recommend imputation strategies, and highlight risk factors. AI provided rapid insights and supported hypothesis generation.

Example AI prompts used:

- 'Summarize key patterns in the dataset and identify anomalies.'

- 'Suggest an imputation strategy for missing income values based on industry best practices.'

# 6. Conclusion & Next Steps

The dataset is largely complete and shows strong potential for predicting delinquency based on key variables like Credit\_Score, Missed\_Payments, and Credit\_Utilization. Minor missing data issues were resolved using imputation. Next steps include feature encoding, deeper statistical analysis, and model development for predicting delinquency.