Exploratory Data Analysis (EDA) Summary Report

1. Introduction

This report presents an exploratory data analysis of Geldium's customer dataset to support Tata iQ's efforts in refining its delinquency risk model. The goal is to identify data quality issues, surface early risk indicators, and provide recommendations for data preparation before predictive modeling.

---

2. Dataset Overview

This section summarises the dataset, including record counts, important columns, and structure.

- **Number of records:** 10 (out of 50 full dataset records)
- **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`, and monthly payment history (`Month_1` to `Month_6`)
- **Data types:**
    - Numerical: `Age`, `Income`, `Credit_Score`, `Credit_Utilization`, `Missed_Payments`, `Loan_Balance`, `Debt_to_Income_Ratio`, `Account_Tenure`
    - Categorical: `Customer_ID`, `Employment_Status`, `Credit_Card_Type`, `Location`, `Month_1`–`Month_6`
    - Binary: `Delinquent_Account` (0 = No, 1 = Yes)

---

3. Missing Data Analysis

Identifying and treating missing values is critical for maintaining model accuracy.

- **Variables with missing values:**
    - `Loan_Balance and Income`: missing value observed in the sample
- **Missing data treatment:**

Variable	Method	Justification
`Loan_Balance`	Median Imputation	Maintains distribution, less affected by outliers
`Income`	Median imputation	Maintains Distribution, normal affected by Outliers

---

4. Key Findings and Risk Indicators

Feature Insights:

- **High Missed Payments (≥4)** correlate with delinquency in many cases.
- **High Credit Utilization (> 80%)** is observed in riskier accounts.
- **Low Credit Score (< 500)** often appears with `Delinquent_Account = 1`.
- **Debt_to_Income_Ratio > 0.4** is a common factor in struggling borrowers.
- Customers with `Employment_Status = Unemployed` or `Self-employed` show higher risk.

Unexpected Anomalies:

- `Account_Tenure` = 0 observed, which may indicate newly onboarded or incorrect records.
- Repetitive "Missed" or "Late" values in payment history are red flags.

- Some records have missing loan balances while still showing debt ratios.

---

5. AI & GenAI Usage

The following GenAI prompts were used to analyze the dataset and detect risks:

- “Summarize key patterns and missing values in this dataset.”
- “Identify the top 3 variables likely to predict delinquency.”
- “Suggest imputation strategies for incomplete financial features.”
- “Detect patterns of risk in delinquent accounts.”

---

6. Conclusion & Next Steps

This EDA identified critical data gaps (e.g., missing loan amounts, income), early indicators of risk (e.g., high missed payments, credit utilisation), and potential predictors of delinquency. The next steps include:

- Applying imputation strategies to fill missing values.
- Conducting feature engineering and scaling.
- Feeding the cleaned data into predictive modeling pipelines.