Exploratory Data Analysis (EDA)

Summary Report

1. Introduction

This report presents the findings from an Exploratory Data Analysis (EDA) conducted on the **Geldium Delinquency Prediction Dataset**. The primary goal of this analysis is to evaluate the dataset's quality, completeness, and potential risk indicators in preparation for building a delinquency risk model. This effort supports Tata iQ's analytics team and Geldium's leadership in refining their intervention strategies to better mitigate financial risk.

2. Dataset Overview

This section summarizes the structure, characteristics, and early anomalies within the dataset.

Key dataset attributes:

- **Number of records:** 500
- **Key variables:**
    - Customer_ID: Unique identifier
    - Age: Age of the customer
    - Income: Annual income in currency
    - Credit_Score: A numerical measure of credit health
    - Credit_Utilization: Ratio of used credit to total available
    - Missed_Payments: Count of missed payments
    - Delinquent_Account: Target variable (1 = delinquent, 0 = non-delinquent)
    - Loan_Balance: Total outstanding loan amount
    - Debt_to_Income_Ratio: Financial stability metric
    - Employment_Status: Employment category (e.g., EMP, Self-employed)
    - Account_Tenure: Duration (in months) of account history
    - Credit_Card_Type: Type of card (e.g., Student, Standard, Platinum)
    - Location: City or region of the customer
    - Month_1 to Month_6: Categorical variables representing repayment status each month (On-time, Late, Missed)

Data types:
A mix of **numerical (int, float)** and **categorical (object)** features are present.

Anomalies & Inconsistencies:

- Some customers show **zero account tenure** — this could be a legitimate new user or a data entry issue.
- Six month-wise repayment status fields are in text format and will require preprocessing.
- No duplicate records were found.

---

3. Missing Data Analysis

Identifying and handling missing values is crucial to avoid bias or inaccuracies in modeling.

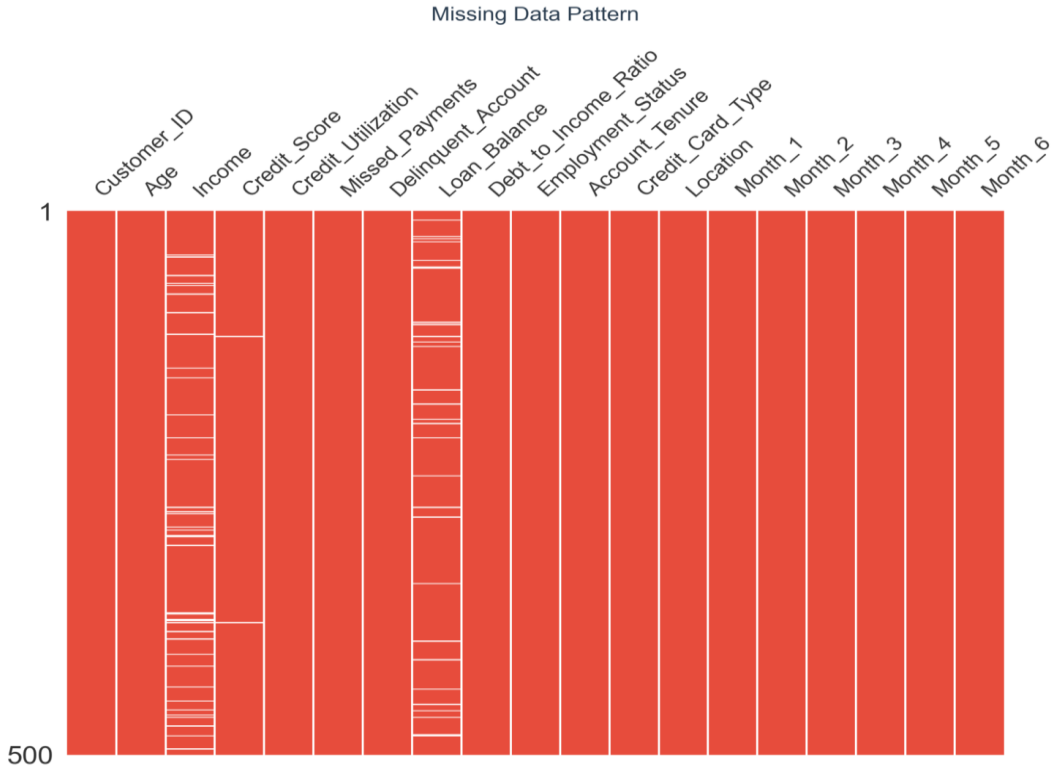
Key missing data findings:

- **Income: 39 missing values (~7.8%)**
- **Credit_Score: 2 missing values (~0.4%)**
- **Loan_Balance: 29 missing values (~5.8%)**

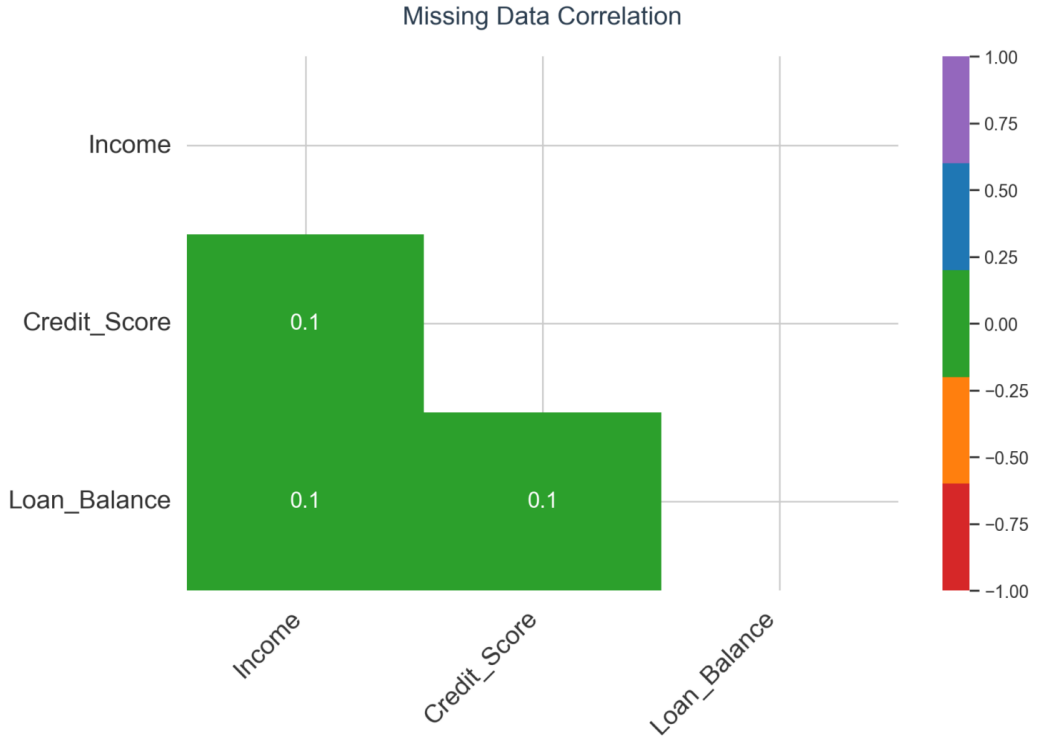
Missing data treatment:

Variable	Handling Method	Justification
Credit_Score	Mean Imputation	Minimal missingness; distribution allows for average-based fill
Income	Predictive Imputation	Significant feature; correlate with Age, Employment_Status
Loan_Balance	Median Imputation	Median is robust against skew, suitable for financial amounts

- **Missing value Column wise Graph**



- **Missing Value Correlation**



4. Key Findings and Risk Indicators

This section highlights observed patterns and variables most predictive of delinquency.

Key correlations & patterns:

- **Missed_Payments** shows a direct and strong relationship with Delinquent_Account.
- High **Credit_Utilization** and **Debt_to_Income_Ratio** frequently appear in delinquent records.
- Time-series debt behavior (Month_1–Month_6) shows that **sequential lateness or missed payments** is a strong indicator of risk.
- Some cases exhibit high Income yet still appear delinquent — these need deeper behavioral or fraud-related investigation.

High-risk indicators identified:

1. Missed_Payments — Primary contributor to delinquency.
2. Credit_Utilization — Indicates spending habits relative to limits.
3. Debt_to_Income_Ratio — Reflects capacity to handle further credit.
4. Month_1 to Month_6 — Progressive risk signal in repayment behavior.

Unexpected anomalies:

- Several high-income customers flagged as delinquent.
- Zero Account_Tenure may represent new accounts or data issues.

5. AI & GenAI Usage

Generative AI tools such as **ChatGPT** were used to:

- Summarize the dataset structure and quality
- Identify potential risk indicators
- Recommend missing data handling strategies

Prompts used include:

- *"Summarize key patterns, outliers, and missing values in this dataset."*
- *"Identify the top 3 variables most likely to predict delinquency."*
- *"Suggest an imputation strategy for missing income values based on industry best practices."*

These tools accelerated insight generation and supported evidence-based recommendations.

---

6. Conclusion & Next Steps

This EDA has provided a strong foundation for building a delinquency risk model. While the dataset is largely complete and clean, several variables require thoughtful handling due to missingness or categorical formatting. Significant predictors such as **Missed_Payments, Credit_Utilization, and historical repayment** behavior should be prioritized in model development.

Next steps:

- Finalize imputation and data cleaning
- Engineer features from Month_1 to Month_6
- Encode categorical fields appropriately
- Proceed with baseline model creation and cross-validation