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

**1️⃣ Introduction**

This report presents the Exploratory Data Analysis (EDA) conducted on Geldium’s customer dataset as part of Tata iQ’s initiative to refine delinquency risk models. The goal of the analysis was to assess the dataset’s quality, identify data gaps and anomalies, and uncover early indicators of delinquency risk that could shape future predictive modeling and intervention strategies.

**2️⃣ Dataset Overview**

The dataset consists of 755 records representing customers, with variables covering demographic, financial, and behavioral attributes.

**Key variables and data types:**

* Customer\_ID (Categorical): Unique customer identifier
* Age (Numerical): Age in years
* Income (Numerical): Annual income in USD
* Credit\_Score (Numerical): Credit score (300-850)
* Credit\_Utilization (Numerical %): % of available credit used
* Missed\_Payments (Numerical): Count of missed payments in 12 months
* Delinquent\_Account (Binary): 0 = No, 1 = Yes
* Loan\_Balance (Numerical): Outstanding loan balance in USD
* Debt\_to\_Income\_Ratio (Numerical %): Debt as % of income
* Employment\_Status (Categorical): Employment type
* Account\_Tenure (Numerical): Account age in years
* Credit\_Card\_Type (Categorical): Card type
* Location (Categorical): Region/city
* Month\_1 to Month\_6 (Categorical): Recent payment history

✅ The dataset was clean overall, with no extreme or illogical values (e.g., no Credit Utilization >100%, no negative or unrealistic Income, Loan Balance, or Debt-to-Income Ratio values).

3️⃣ Missing Data Analysis

**Variables with missing values:**

* Income: 39 missing rows (5.17%)
* Loan\_Balance: 29 missing rows (3.84%)
* Credit\_Utilization: 2 missing rows (0.26%)

**Missing data treatment decisions:**

| **Variable** | **Handling Method** | **Justification** |
| --- | --- | --- |
| Income | Impute median | Median reduces skew impact; Income is critical for risk modeling |
| Loan\_Balance | Impute median | Important feature, low % missing, median preserves distribution |
| Credit\_Utilization | Impute with mean or predictive model | Very few missing, minimal bias risk |

**4️⃣ Key Findings and Risk Indicators**

* 319 customers (42.3%) have 2 or more missed payments → high risk of delinquency.
* No illogical or extreme values were detected in the dataset.
* Further analysis will explore interactions between high credit utilization and debt-to-income ratios.

**5️⃣ AI & GenAI Usage**

Generative AI was used to:

* Summarize dataset structure, missing data, and key patterns
* Suggest appropriate imputation methods for missing values
* Guide the EDA workflow and prompt design

**Prompts used included:**

* "Summarize key patterns, outliers, and missing values in this dataset. Highlight any fields that might present problems for modeling delinquency."
* "Suggest an imputation strategy for missing values in this dataset based on industry best practices."
* "Help draft a clean EDA summary report suitable for submission."

**6️⃣ Conclusion & Next Steps**

The EDA confirms that Geldium’s dataset is generally clean and suitable for predictive modeling, with minor missing data issues addressed through standard imputation techniques. Key risk indicators such as high missed payments will inform model training and intervention strategy design. The next steps include building and validating delinquency prediction models and incorporating these insights into risk mitigation plans.