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

# 1. Introduction

# The purpose of this report is to perform an exploratory data analysis (EDA) on Geldium’s customer financial dataset to assist Tata iQ’s analytics team in refining their delinquency risk model. The goal is to understand the structure and quality of the data, detect any missing or inconsistent values, and identify early risk indicators that could impact credit delinquency predictions.

# 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:

* Delinquent\_Account: Target variable (0 = non-delinquent, 1 = delinquent)
* Income, Credit\_Score, Credit\_Utilization: Financial metrics
* Missed\_Payments: Historical behavior
* Employment\_Status, Location: Demographic factors
* Month\_1 to Month\_6: Possibly repayment status over 6 months

- Data types:

* Categorical: Employment\_Status, Credit\_Card\_Type, Location, Month\_1 to Month\_6
* Numerical: Age, Income, Credit\_Score, Credit\_Utilization, etc.

Anomalies:

* Some rows have missing or null values
* Minor inconsistency in Loan\_Balance and Credit\_Score

# 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:

| **Variable** | **Handling Method** | **Justification** |
| --- | --- | --- |
| Income | Imputation (mean or median) | Reasonable distribution; common in modeling |
| Credit\_Score | Imputation (mean) | Only 2 missing; low impact |
| Loan\_Balance | Imputation or model-based fill | 29 missing; important variable |

# 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 Credit\_Utilization is associated with increased delinquency.
* Customers with more Missed\_Payments have a higher chance of defaulting.
* Lower Credit\_Score and higher Debt\_to\_Income\_Ratio are linked to delinquency.
* Employment\_Status may also influence ability to repay loans.
* Any surprising trend? (e.g., people with good credit scores but still delinquent)

# 5. AI & GenAI Usage

Generative AI tools were used to summarize the dataset, impute missing data, and detect patterns. This section documents AI-generated insights and the prompts used to obtain results.

Example AI prompts used:

* "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 income values based on industry best practices."
* "Identify the top 3 variables most likely to predict delinquency based on this dataset. Provide brief reasoning."
* "Propose how to handle missing values in loan balance while maintaining fairness in prediction."
* "Explain how missed payments across months relate to future delinquency risk."

# 6. Conclusion & Next Steps

This EDA revealed several key risk indicators for loan delinquency, including missed payments, high credit utilization, and high debt-to-income ratios. Missing data was managed using imputation methods to maintain dataset integrity. Next steps include building a predictive model using the cleaned dataset, validating key variables through feature selection, and exploring time series behavior through the monthly columns (Month\_1 to Month\_6) for trend-based modeling.