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

# 1. Introduction

# This report presents the findings from the Exploratory Data Analysis (EDA) of Geldium’s customer dataset. The primary objective is to assess data quality, identify patterns that may signal credit risk or delinquency, and prepare the dataset for predictive modeling. The analysis aims to support Tata iQ and Geldium in enhancing their delinquency risk model and refining intervention strategies.

# 2. Dataset Overview

The dataset consists of 500 customer records and 19 columns, covering demographics, credit usage, repayment behavior, and monthly payment performance.

**Key dataset attributes:**

* **Number of records**: 500
* **Key variables**:
  + Age, Income, Credit\_Score, Credit\_Utilization, Missed\_Payments, Loan\_Balance, Debt\_to\_Income\_Ratio
  + Employment\_Status, Account\_Tenure, Credit\_Card\_Type, Location
  + Monthly payment behavior: 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**: Employment\_Status, Credit\_Card\_Type, Location
  + **Target variable**: Delinquent\_Account (Binary)

**Initial observations:**

* No duplicate rows or customer IDs.
* 3 numerical variables had missing values: Income, Loan\_Balance, and Credit\_Score.
* Credit\_Utilization had values exceeding 1.0, which is unusual and may need further investigation.

# 3. Missing Data Analysis

**Key missing data findings:**

* **Income**: 7.8% missing
* **Loan\_Balance**: 5.8% missing
* **Credit\_Score**: 0.4% missing

**Missing data treatment:**

| **Variable** | **Missing %** | **Imputation Strategy** | **Justification** |
| --- | --- | --- | --- |
| Income | 7.8% | Median Imputation | Income is skewed; median is robust to outliers. |
| Loan\_Balance | 5.8% | Mean Imputation | Distribution is relatively normal; mean maintains consistency. |
| Credit\_Score | 0.4% | Median Imputation | Credit scores are clustered; median provides stability. |

All missing values were successfully imputed. Post-imputation validation confirms a clean dataset.

# 4. Key Findings and Risk Indicators

**Correlation & Risk Patterns:**

* Missed\_Payments has a visible association with Delinquent\_Account.
* Credit\_Utilization is higher among delinquent accounts.
* Credit\_Score is slightly lower among delinquents, though correlation is weak.
* Employment\_Status shows that unemployed customers have a higher proportion of delinquency.

**Noteworthy observations:**

* Weak correlations overall, suggesting need for non-linear models or interaction terms.
* Delinquency is imbalanced (only 16% of accounts are delinquent), which may affect model performance.
* Some variables (like Account\_Tenure) appear inversely related to delinquency, but weakly.

**Key risk indicators:**

1. **Missed\_Payments** – Directly reflects repayment issues.
2. **Credit\_Utilization** – Higher usage may signal financial stress.
3. **Credit\_Score** – Though correlation is low, low scores generally indicate higher risk.

# 5. AI & GenAI Usage

GenAI tools (ChatGPT) were used to summarize patterns, suggest imputation strategies, and interpret plots.

**AI Prompts Used:**

* “Summarize key patterns, missing values, and potential anomalies in the dataset.”
* “Identify the top 3 variables most likely to predict delinquency.”
* “Suggest an imputation strategy for missing values in this dataset based on industry best practices.”
* “What risk indicators can be derived from the correlation heatmap and delinquency distribution plots?”

**Summary of GenAI Insights:**

* Validated skewness in income and selected imputation accordingly.
* Detected unusual values in Credit\_Utilization.
* Highlighted imbalance in Delinquent\_Account variable.
* Prioritized variables for risk prediction and modeling.

# 6. Conclusion & Next Steps

The EDA successfully uncovered key risk factors and cleaned the dataset for modeling. The top risk indicators identified were Missed\_Payments, Credit\_Utilization, and Credit\_Score. Missing values were handled using industry-recommended imputation strategies. The dataset is now ready for feature engineering and model development.

**Next steps**:

* Perform feature transformation and encoding.
* Address class imbalance (e.g., SMOTE, weighting).
* Begin supervised model training (e.g., logistic regression, random forest).