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

# 1. Introduction

# This report details an Exploratory Data Analysis (EDA) of Geldium's customer dataset. The primary purpose is to assess data quality, identify key patterns, and uncover potential risk indicators for delinquency. The findings will be used to inform and improve Geldium's predictive modeling and intervention strategies.

# 2. Dataset Overview

This dataset contains customer information related to loan and credit card accounts. An initial review revealed several key attributes and some data quality issues that need to be addressed before any predictive modeling can begin.

Key dataset attributes:

Number of records: 500

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 Month\_1 to Month\_6 payment history.

Data types: The dataset includes numerical, categorical, and text-based data. For example, Age, Income, and Credit\_Score are numerical, while Employment\_Status and Credit\_Card\_Type are categorical. The monthly payment history is also categorical (Late, Missed, On-time).

Initial Observations:

Missing or Inconsistent Data: Several key columns, including Income, Credit\_Score, Loan\_Balance, and Debt\_to\_Income\_Ratio, have missing values. Additionally, there are inconsistencies in the Employment\_Status column (e.g., both "employed" and "Employed" are used).

Anomalies: Some Credit\_Utilization values are unusually low (e.g., 0.05) or even above 1.0, which may indicate data entry errors or require specific handling. There are also instances of 0-month Account\_Tenure, which may warrant further investigation.

# 3. Missing Data Analysis

To ensure the integrity of any future predictive model, it is crucial to handle missing data appropriately. The following table outlines the missing data issues and the chosen methods for treatment.

| Variable with Missing Values | Chosen Handling Method | Justification |
| --- | --- | --- |
| Income | Imputation using the median | The median is a robust measure that is less sensitive to outliers, providing a more reliable estimate for filling in missing income values. |
| Loan\_Balance | Imputation using the mean | The mean is a suitable method for filling in missing numerical data when the distribution is not heavily skewed. |
| Credit\_Score | Imputation using the median | Like income, the median is preferred to mitigate the impact of any extreme outliers that may exist in the credit score data. |

# 4. Key Findings and Risk Indicators

My analysis revealed several strong correlations and risk indicators for delinquency:

Key findings:

Missed Payments and Delinquency: A very strong positive correlation exists between the number of Missed\_Payments and a Delinquent\_Account status. Customers with 4 or more missed payments are significantly more likely to have a delinquent account.

Credit Utilization and Delinquency: High Credit\_Utilization rates are a clear indicator of financial stress and are strongly associated with delinquency. Customers with a credit utilization rate above 0.7 are at a higher risk.

Credit Score and Delinquency: A low Credit\_Score is a significant risk factor. Customers with scores below 500 show a much higher propensity for delinquency.

Unexpected anomalies: The existence of Credit\_Utilization values greater than 1.0. This suggests potential data entry errors and would need to be addressed with the data collection team. A value above 1.0 indicates a customer has spent more than their credit limit, which may be a legitimate indicator of high risk but should be verified.

Employment\_Status Inconsistencies: The Employment\_Status column contains multiple labels for the same status (e.g., "Employed" and "employed," "EMP" and "emp"). These would need to be cleaned and standardized for accurate modeling.

# 5. AI & GenAI Usage

Generative AI tools were invaluable in streamlining this EDA process. They were used to summarize initial findings, suggest imputation strategies, and identify key patterns.

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 values in this dataset based on industry best practices.

'Propose best-practice methods to handle missing credit utilization data for predictive modeling.'

# 6. Conclusion & Next Steps

This EDA has successfully identified key data quality issues and several strong risk indicators. The data is largely complete, but critical variables like Income, Loan\_Balance, and Credit\_Score require imputation to maximize model accuracy.

**Recommended next steps:**

1. **Data Cleaning:** Clean the Employment\_Status column to standardize all entries.
2. **Imputation:** Implement the proposed imputation strategy for missing values in Income, Loan\_Balance, and Credit\_Score.
3. **Feature Engineering:** Consider creating new features from the existing data, such as a Payment\_Consistency\_Score based on the monthly payment history.
4. **Modeling:** With the cleaned and prepared data, proceed with building and training a predictive delinquency model.
5. **Data Verification:** Engage with the data collection team to investigate the anomalies, such as Credit\_Utilization values > 1.0, to understand if they are genuine risk signals or data errors.