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

# 1. Introduction

The purpose of this report is to present the findings from an Exploratory Data Analysis (EDA) conducted on Geldium’s customer credit dataset. The goal is to assess data completeness, detect inconsistencies, and identify early warning indicators for credit card delinquency. These insights will support the development of a predictive model and help refine customer risk assessment and intervention strategies.

# 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:**

Customer\_ID – Unique identifier for each customer

Age – Customer’s age in years

Income – Annual income (USD), may contain missing values

Credit\_Score – Credit rating (300 to 850)

Credit\_Utilization – Percentage of credit currently in use

Missed\_Payments – Missed payments in the past 12 months

Delinquent\_Account – Binary indicator (0 = No, 1 = Yes)

Loan\_Balance – Total current loan balance

Debt\_to\_Income\_Ratio – Ratio of debt to income (0-100%)

Employment\_Status, Credit\_Card\_Type, Location – Categorical fields

Month\_1 **to** Month\_6 – Monthly payment history (0 = on-time, 1 = late, 2 = missed)

****Data types:****

****Numerical:**** Age, Income, Credit Score, Credit Utilization, DTI, etc.

****Categorical:**** Employment Status, Credit Card Type, Location, Monthly History

****Binary:**** Delinquent\_Account

****Anomalies & Inconsistencies:****

· No duplicate Customer\_ID’s found.

· No values below age 18.

· No credit utilization values above 100%.

· Minor inconsistencies found in text formats of categorical fields.

# 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:****

· Income: 39 missing entries

· Credit\_Score: 2 missing entries

· Loan\_Balance: 29 missing entries

****Missing data treatment:****

****Income:**** Regression-based imputation using related financial fields (DTI, Credit Score, Age)

****Credit\_Score:**** Median imputation, due to low number of missing values

****Loan\_Balance:**** Predictive imputation using Debt\_to\_Income\_Ratio and Account\_Tenure

No columns were dropped as the percentage of missing data was manageable and data could be reliably imputed.

# 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.

****Correlations observed between key variables:****

· Income, Credit\_Utilization, and Debt\_to\_Income\_Ratio showed weak but positive correlation with Delinquent\_Account.

· Missed\_Payments and Account\_Tenure had slight negative correlations with delinquency.

· Strongest (though still weak) correlation was from Income (+0.045) and Credit\_Utilization (+0.034).

****Unexpected anomalies:****

· A few customers had very high credit scores (>750) but still had late payments.

· Customers with 0 income reported, yet no signs of delinquency—possibly unrecorded earnings.

****Early indicators of delinquency risk:****

· High Credit\_Utilization: indicates reliance on credit, possible financial strain.

· High Debt\_to\_Income\_Ratio: reflects difficulty in managing debts.

· Lower Credit\_Score: typical indicator of past financial mismanagement.

· Short Account\_Tenure: newer customers tend to have higher uncertainty and less repayment history

# 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 in the dataset and identify anomalies.

· Suggest an imputation strategy for missing income values based on industry best practices.

· Which features are most correlated with credit card delinquency?

· Generate synthetic loan balance values that align with observed data distribution.

# 6. Conclusion & Next Steps

The dataset is in generally good condition with a few manageable missing values and minor anomalies. Through statistical imputation and pattern analysis, several early risk indicators have been identified that will support future delinquency prediction models. The next step is to finalize feature selection, prepare the cleaned dataset for modeling, and begin training AI-based classification models using the most predictive variables: Credit\_Utilization,Debt\_to\_Income\_Ratio,Income,and Missed\_Payments.