Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

The purpose of this report is to conduct an initial exploratory data analysis (EDA) of Geldium's customer dataset to assess data quality, detect anomalies, and uncover early risk indicators related to delinquency. The goal is to prepare the dataset for predictive modeling by identifying missing values, outliers, and key patterns that influence customer delinquency risk. This analysis will guide data cleaning strategies and highlight variables most relevant for accurate risk prediction.

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:

Key Columns	**Description**
Age	Customer's age
Income	Annual income in USD
Credit_Score	Customer's credit score
Credit_Utilization	% of credit used
Missed_Payments	Missed payments in past year
Delinquent_Account	Indicator of whether the customer has a delinquent account
Loan_Balance	Total outstanding loan
Debt_to_Income_Ratio	Debt/income percentage

Employment_Status	Job status
Credit_Card_Type	Type of card
Month_1 to Month_6	Monthly payment behavior

- 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, Month_1 to Month_6
- **Binary:** Delinquent_Account
- **ID:** Customer_ID (Categorical/Identifier)

- Anomalies and inconsistencies:

- Credit_Utilization exceeds 1.0 in one case (103%) — should be capped at 1.0.
- Income, Credit_Score, and Loan_Balance contain missing values.
- No duplicate Customer_IDs were found, but Employment_Status has inconsistent formatting (e.g., “EMP” instead of “Employed”).

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:

- Income: 39 missing entries (7.8%)
- Loan_Balance: 29 missing entries (5.8%)
- Credit_Score: 2 missing entries (0.4%)

- Missing data treatment:

Column	Handling Method	Justification
Income	Imputation (Median)	Preserves distribution; critical feature for affordability modeling

Loan_Balance	Imputation (Mean)	Important for debt assessment; low enough missing rate for statistical fill
Credit_Score	Imputation (KNN or Mean)	Only 2 missing, minimal effect; can use mean or nearest neighbors

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:

- Customers with **higher credit utilization** (above 0.6) tend to have a **greater chance of delinquency**.
- **Missed_Payments** is highly associated with delinquency — customers with 4+ missed payments are disproportionately delinquent.
- **Debt_to_Income_Ratio** above 0.35 appears to align with higher delinquency rates, suggesting over-leveraged customers are at greater risk.
- Preliminary patterns suggest **unemployment or self-employment** may also be associated with higher default risk.

- Unexpected anomalies:

- “Credit_Utilization” has a value exceeding 1.0 (103%), which violates the expected 0–100% range.
- “Income” ranges widely from ~$15,000 to ~$200,000 — further segmentation may be needed to detect income-based risk tiers.
- Inconsistent formatting in “Employment_Status” (e.g., “EMP”) may introduce noise in categorical encoding.
- “Missed_Payments” = 0 but “Delinquent_Account” = 1 for a few entries — may indicate lags or inconsistencies in reporting history.

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 in the dataset and identify anomalies.'

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

6. Conclusion & Next Steps

Key Findings:

- The dataset contains 500 customer records with a mix of financial, behavioral, and demographic features.
- Missing values are present in key columns such as **Income**, **Loan_Balance**, and **Credit_Score**, which were addressed through imputation techniques tailored to each variable.
- High **credit utilization**, frequent **missed payments**, and elevated **debt-to-income ratios** are strong predictors of delinquency.
- Minor anomalies and inconsistencies were identified, including out-of-range values and categorical label inconsistencies, which require attention before modeling.

Recommended Next Steps:

- Complete imputation for missing values using the proposed strategies.
- Normalize or cap **Credit_Utilization** values exceeding 100%.
- Clean categorical fields like **Employment_Status** to ensure consistency.
- Encode categorical variables and engineer relevant features (e.g., total missed payments from monthly history).
- Proceed to model development using identified predictors and validate against delinquency labels.