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

# 1. Introduction

This report summarizes an exploratory analysis of Geldium’s customer dataset. The goal is to assess data quality, surface patterns and risk indicators associated with credit default, and propose preparatory steps to make the data modeling ready for risk prediction.

# 2. Dataset Overview

The dataset contains 500 customer records with a mix of categorical and numerical attributes relevant to credit behaviour and risk assessment.

* **Total records:** 500
* **Key attributes:** Age, Income, Credit Score, Credit Utilization, Missed Payments, Debt to Income Ratio
* **Data types:**
  + Categorical: Employment Status, Credit Card Type
  + Numerical: Income, Loan Balance, Credit Utilization, Debt to Income Ratio

# 3. Missing Data Analysis

Missingness exists in important numeric fields and requires careful treatment to avoid biased model results.

* **Missing counts observed:**
  + Income: 50 missing values
  + Loan Balance: 30 missing values
* **Implications:** Missing values in income and loan balance can skew distribution-based features and affect models that rely on income-derived ratios.
* **Proposed handling approach:**
  + For numeric fields with moderate missingness, impute using robust central tendency (median) to limit influence from extreme values.
  + For Loan Balance, consider controlled synthetic data generation only where domain rules can be enforced and after validating that synthetic examples preserve realistic relationships with other fields.
  + Evaluate more advanced imputation (e.g., multiple imputation or model-based imputation) if missingness is non-random or correlated with target variables.

# 4. Key Findings and Risk Indicators

The exploratory analysis highlights behavioral and utilization patterns strongly correlated with delinquency.

* **Credit utilization:** Customers with utilization > 50% show a substantially elevated risk of delinquency.
* **Payment history:** Having 3 or more missed instalments within the previous six months corresponds to a noticeably higher default probability.
* **Notable anomalies:** A small subset of customers combines high reported income with low credit scores these cases require investigation for potential reporting errors or unobserved risk drivers.
* **Takeaway:** High credit usage and repeated missed payments are the most consistent indicators of elevated default risk in this dataset.

# 5. AI & GenAI Usage

Generative and analytical AI tools were used to accelerate data profiling and pattern detection while cross-checking findings against standard credit-risk metrics.

* **What AI helped with:** automated summarization of missingness, initial clustering of risk groups, and quick flagging of outliers.
* **Examples of prompts/queries used:**
  + “Summarize column-level missingness and highlight fields with potential impact on modeling.”
  + “Identify behavioral features (e.g., utilization, missed payments) that most strongly separate delinquent vs non-delinquent customers.”
* **Validation:** AI-derived observations were validated against conventional risk indicators to ensure alignment with financial domain expectations.

# 6. Conclusion & Next Steps

The EDA uncovers clear signals useful for credit risk modeling and highlights data quality issues that must be addressed prior to building predictive models.

**Conclusions:**

* Missing income and loan-balance values could materially affect downstream model performance if not addressed.
* High credit utilization and repeated missed payments are reliable predictors of default risk.
* Some records exhibit inconsistent profiles (e.g., high income but low credit score) and should be reviewed.

**Actionable recommendations:**

1. **Imputation strategy:** Apply median imputation for simple baselines and evaluate multiple imputation or predictive model imputation for improved accuracy avoid blanket synthetic generation unless validated.
2. **Subgroup validation:** Test whether the key risk factors (utilization, missed payments) hold consistently across subpopulations (by employment status, card type, income bracket).
3. **Outlier investigation:** Manually review or flag suspicious combinations (e.g., high income + low credit score) to confirm data correctness or discover hidden risk attributes.
4. **Feature engineering:** Create derived variables such as rolling missed payment counts, utilization bands (>50%), and debt to income buckets to improve model discriminative power.
5. **Model readiness checks:** After cleaning and imputation, run correlation analysis, check class balance, and perform baseline modeling (logistic regression / tree-based) to validate predictive signal strength.
6. **Governance for synthetic data:** If using AI to synthesize loan-balance values, establish constraints and sanity checks that preserve statistical relationships with income and utilization.

Implementing these steps will strengthen data reliability and improve the robustness of subsequent credit-risk models for Geldium.