Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

This report documents the exploratory data analysis (EDA) performed on the **'Delinquency_prediction_dataset'**. The goal is to uncover structure, quality issues, and early risk signals to guide feature engineering and model development for predicting customer delinquency.

2. Dataset Overview

- **Notable missing or inconsistent data:** Income missing 7.8 %, Loan_Balance missing 5.8 %, and inconsistent Employment_Status labels ("EMP" vs "Employed") resolved.
- **Key anomalies:** 4 records show Credit_Utilization > 1.0, and a small set of extreme-high incomes (> $200 k).
- **Early indicators of delinquency risk:** utilization ratios > 70 %, DTI > 40 %, credit score < 500, and ≥ 2 recent late/missed payments.

Initial data quality checks reveal generally well-structured information with modest missingness concentrated in three numeric fields. Most anomalies are limited and correctable, though the small set of utilization ratios exceeding 100 % warrant manual review. After standardizing categorical labels and imputing numeric gaps, the dataset now appears consistent and ready for modeling. No duplicate records or irreparable inconsistencies were discovered.

3. Missing Data Analysis

Issue	Handling Method	Justification
Income (7.8 % missing)	Log-BayesianRidge imputation on log-income + missing flag	Preserves skewed distribution and keeps predictive signal of missingness
Loan_Balance (5.8 % missing)	Median per Employment_Status + flag	Simple, robust and leverages job category to reflect repayment capacity
Credit_Score (0.4 % missing)	Global median imputation + flag	Tiny gap; median avoids distortion with negligible information loss

4. Key Findings and Risk Indicators

- Missing income flag = 1 – data withholding itself correlates with higher risk (17 % vs 15 %).
- Unemployment status – irregular income streams drive the highest observed delinquency share (19 %).
- Short account tenure (< 12 months) – limited history increases uncertainty; delinquency median tenure is 8 months.
- ≥ 2 late or missed payments in last six months – recent behaviour momentum predicts near-term default (22 % rate).
- Low credit score (< 500) – signals past repayment issues; bottom quartile exhibits 1.4 × portfolio risk.
- Debt-to-income ratio above 40 % – heavy fixed obligations shrink payment buffer, lifting delinquency odds by 30 %.
- High credit-utilization ratio (> 70 %) – borrowers near their limit default more often; rate rises ≈ 5 pp in this band.

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

EDA confirms that recent repayment behaviour, credit utilization, debt burden, and credit quality are strong delinquency drivers. The dataset is now free of missing numeric values, and risk-signal flags have been engineered. Next steps include: (1) scaling & encoding features, (2) handling class imbalance, (3) training baseline logistic and tree-based models, and (4) validating model stability with cross-validation and out-of-time testing.