Geldium Delinquency Risk – Exploratory Data Analysis (EDA) & Data Quality Review

Note: Dataset file was not provided in the workspace at submission time. This report follows the required template and documents the EDA approach, expected checks, and placeholder sections where dataset-driven numbers/plots would be inserted once the data is available.

1. Objective

Assess dataset structure, completeness, and data quality; identify early risk indicators and gaps that could impact delinquency prediction and collections intervention strategies.

2. Dataset Overview (to be populated from file)

Fields to review typically include: customer identifiers, loan/account attributes (loan amount, tenure, EMI), customer profile (income, employment type, geography), repayment behavior (due dates, payments, missed payments), credit utilization, and delinquency labels (DPD, delinquent flag, default flag).

3. Step 1 – Initial EDA Findings (placeholders)

• Missingness: Identify % missing per column; flag critical fields (payment history, income, utilization, delinquency label).
• Inconsistencies: Date issues (payment before due), negative amounts, delinquency label mismatch with DPD.
• Duplicates: Duplicate customer/loan IDs; conflicting static attributes.
• Outliers: Extreme income/loan amount/utilization; invalid ages.

Initial data quality summary (3–5 sentences): Once the dataset is loaded, summarize overall completeness, the 3–5 most problematic fields, and any label/behavior inconsistencies that could bias model training.

4. Step 2 – Missing Data Treatment Plan (example table)

Create treatment decisions for top issues. Example table below.

5. Step 3 – Risk Indicators & Patterns to Test

After cleaning, test relationships vs delinquency outcomes: utilization vs missed payments, EMI-to-income vs delinquency, prior delinquencies, payment irregularity, loan age, segment-level differences (employment/region/product).

6. Recommendations / Next Steps

- Confirm target definition (DPD threshold and observation window).
- Standardize categorical values; enforce numeric ranges.
- Build feature set: rolling missed-pay counts, DPD trends, utilization bands, affordability ratios.
- Data enrichment if needed (credit bureau score, updated income).
- Proceed to baseline model after quality gates are passed.

Missing data issue	Handling method	Justification
Credit utilization missing (e.g., 20%+)	Impute + Missing indicator	Utilization is predictive; indicator captures informative mis
Income missing (moderate %)	Synthetic impute (distribution-aware)	Avoid dropping; preserve distribution; use constraints by
Payment history missing (high %)	Drop rows/columns or restrict cohorts	Missing behavior data can distort delinquency labels/feat