Geldium — EDA Report (Steps 1 to 3)

Step 1: Data Structure and Quality Review

Dataset contains ~500 records and 19 columns with numeric and categorical fields. The primary target variable is **Delinquent_Account** (binary 0/1). Core columns include Credit_Score, Credit_Utilization, Debt_to_Income_Ratio, Age, Income, Employment_Status, Account_Tenure, Credit_Card_Type, Location, Loan_Balance, Missed_Payments and a six-month payment history (Month_1 to Month_6).

Key Patterns

- Monthly payment history (Month_1 to Month_6) provides a short-term behavioral signal valuable for recent risk assessment.
- Missed_Payments aggregates past 12-month failures — a strong historical predictor of future delinquency.
- Credit_Utilization and Debt_to_Income_Ratio capture current financial stress and leverage.

Outliers and Missing Values

- Extreme low credit scores (in the 300s) observed — valid but represent high-risk outliers.
- Account_Tenure contains a value of 0 which conflicts with presence of six months of payment history (logical anomaly).
- No explicit NaNs visible in sampled data, but inconsistent categorical formats and implicit missing values (e.g., 'N/A', empty strings) may exist.

Columns with Data Quality Issues

Column Name	Issue Type	Description
Employment_Status	Inconsistency	Varied spellings/abbreviations (e.g., 'EMP', 'employed', 'Self-employ
Account_Tenure	Inconsistency/Anomaly	Contains 0 which is inconsistent with six months of payment history
Month_1 to Month_6	Categorical Definition	Ensure consistent definitions for 0=On-time, 1=Late, 2=Missed and

Top 3 Variables Most Likely to Predict Delinquency

- **Missed_Payments**: Direct historical count of missed payments — strongest immediate predictor.
- **Credit_Score**: Standardized measure of creditworthiness; lower scores linked to higher delinquency risk.
- **Credit_Utilization**: High utilization indicates financial strain and correlates with missed payments.

Summary (3–5 sentences)

Initial review shows fair data quality with notable issues that must be addressed before modeling. Key tasks include standardizing Employment_Status, correcting logical anomalies such as Account_Tenure=0, and scanning for implicit missing values. Early indicators—Missed_Payments, Credit_Utilization, and Credit_Score—are likely to be the most influential features for predicting delinquency.

Step 2: Address Missing Data and Data Quality Issues

This section lists key missing/inconsistent data issues and recommended handling methods to prepare the dataset for predictive modeling.

Column Name	Issue Description	Handling Method	Justification
Employment_Status	Contains inconsistent labels (e.g., 'EMP', 'employed', 'Self-employed')	Standardize categories (map variants to 'Employed', 'Unemployed', 'Se	Ensures consistent categorical e
Income	Missing values and possible outliers	Median imputation by Employment_Status group	Median is robust to skew; group-ba
Account_Tenure	Contains 0 which conflicts with available payment history	Impute using median tenure of customers with similar profiles	Preserves records while avoiding g

Summary of Handling Decisions

Chosen methods prioritize data consistency and minimize information loss. Standardizing Employment_Status improves encoding, median imputation for Income limits distortion from outliers, and logical correction/imputation for Account_Tenure avoids dropping valid payment histories.

Step 3: Detect Patterns and Risk Factors

After cleaning, analyze relationships between features and delinquency. Below are identified high-risk indicators, unexpected findings, and recommendations for further analysis and feature engineering.

High-Risk Indicators (one-line each)

- **High Credit_Utilization (>=80%)**: Indicates customers are near credit limits and may struggle to make payments.
- **Recent Missed or Late Payments (Month_5/Month_6)**: Recent delinquencies strongly signal impending defaults.
- **High Missed_Payments count**: Past missed payments are highly predictive of future delinquency.
- **Low Credit_Score (<580)**: Lower scores correspond to higher default probabilities.
- **High Debt_to_Income_Ratio**: Large debt burden relative to income reduces repayment capacity.
- **Rapid Increase in Loan_Balance**: May indicate new borrowing due to stress, elevating risk.

Unexpected Findings & Follow-up Investigations

- Account_Tenure=0 despite six months of payment history — investigate data entry rules and source systems.
- 0 balance but flagged as delinquent in some rows — check labeling rules and possible data merges.
- Inconsistent coding in Month_1..Month_6 may hide true payment behavior; validate against raw transaction logs if available.

Feature Engineering & Modeling Recommendations

- Create rolling-window features from Month_1..Month_6 (e.g., number of misses in last 3 months).
- Derive utilization change features (month-over-month) to capture sudden stress.
- Encode categorical employment and credit card tiers with target-guided encoding or frequency encoding.
- Flag imputed values with binary indicators so model can learn missingness patterns.

Final Recommendations

1. Clean and standardize categorical fields, correct logical anomalies, and impute missing values as specified. 2. Engineer temporal features from the payment history and include missingness flags. 3. Retrain models with robust cross-validation and monitor model performance specifically for high-risk segments (e.g., low credit score, high utilization).

This combined EDA (Steps 1–3) provides a roadmap to prepare Geldium's dataset for delinquency modeling. Implementing the data cleaning steps and feature recommendations above will improve model stability and predictive power.