EDA Report for Geldium – Credit Card Delinquency Dataset

Step 1: Review dataset & identify key insights

Notable missing or inconsistent data

- Payment history field contains several missing entries → critical for predicting delinquency.
- Income and employment status have inconsistent formats (e.g., missing, outliers, or text where numbers should be).
- Credit utilization ratio has extreme outliers (values above 100% → indicates potential data entry errors).

Key anomalies

- Some customers show very high credit limits compared to income (possible anomaly or fraud).
- Duplicate entries for a few customer IDs.
- Unrealistic values in Age (e.g., below 18 or above 90).

Early indicators of delinquency risk

- High credit utilization ratio (>70%).
- Consistently late payments (past due amounts).
- Low or irregular income relative to credit exposure.
- Short credit history (new customers default more often).

Summary (3–5 sentences):
The dataset provides useful predictors for delinquency but contains gaps in critical variables such as payment history and income. Several anomalies, including unrealistic age values and extreme credit utilization, may skew analysis if left untreated. Early review suggests that utilization, income stability, and past due records will be strong predictors of delinquency. Cleaning and imputation are required before modelling.

Step 2: Address missing data and data quality issues

Missing Data Issue	Handling Method	Justification
Missing Payment History	Regression-based imputation using utilization, income, and past balances	Payment history is too important to drop, and correlated variables can predict it.
Missing Income Values	Impute using median by employment type & credit utilization	Preserves fairness and avoids skew from extreme values.
Missing Credit Utilization	Replace with synthetic values (normal distribution centered on customer's segment)	Maintains realistic utilization patterns needed for risk modelling.

Step 3: Detect patterns & risk factors

High-risk indicators

- **Credit Utilization >70%** → Strongly correlated with missed payments due to financial stress.
- **Low Income-to-Credit Ratio** → Customers with low income but high exposure show higher delinquency risk.
- **Frequent 30+ day late payments** → Strong historical predictor of future delinquency.
- **Short Credit History (<1 year)** → New borrowers have less stability and higher default rates.
- **High Balance Growth Month-on-Month** → Suggests over-leveraging, increasing delinquency risk.

Unexpected/Interesting findings

- Some high-income customers still delinquent, possibly due to behavioral factors (spending habits, lack of discipline).
- Younger borrowers show slightly higher delinquency, not purely due to income but possibly due to lower financial literacy.

Step 4: EDA Report Summary

Key Patterns & Anomalies
The dataset reveals missing payment history and income inconsistencies, extreme utilization rates, and duplicate customer records. Outliers like unrealistic age values and anomalously high credit limits require cleaning.

Summary of Missing Values & Treatment
Critical variables like payment history and income will be imputed using regression/median strategies, while synthetic distributions will be applied for utilization. Dropping is only recommended for extreme anomalies or duplicates.

Risk Indicators
Major delinquency risk drivers include high credit utilization, poor income-to-credit ratio, frequent late payments, short credit histories, and rapid balance increases. These align with industry observations and should be central in predictive modeling.