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

# 1. Introduction

The purpose of this report is to assess the quality and structure of Geldium’s customer dataset before building a predictive model for delinquency. This Exploratory Data Analysis (EDA) aims to identify missing or inconsistent data, understand key variables, detect early risk indicators, and highlight patterns that could influence default prediction. These insights will guide the modeling and risk profiling efforts of Tata iQ's analytics team.

# 2. Dataset Overview

This section summarizes the dataset, including the number of records, key variables, and data types. It also highlights any anomalies, duplicates, or inconsistencies observed during the initial review.

Key dataset attributes:

- Number of records: 500 (CUST0001 to CUST0500)

- Key variables:

* **Customer\_ID**: Unique identifier for each customer.
* **Age**: Customer's age.
* **Income**: Annual income (some missing values).
* **Credit\_Score**: Credit score of the customer.
* **Credit\_Utilization**: Ratio of credit used to credit available.
* **Missed\_Payments**: Number of missed payments in the past.
* **Delinquent\_Status**: Binary indicator (0 = not delinquent, 1 = delinquent).
* **Loan\_Balance**: Outstanding loan balance (some missing values).
* **Debt\_to\_Income\_Ratio**: Debt-to-income ratio.
* **Employment\_Status**: Employment category (e.g., Employed, Self-employed, Unemployed).
* **Account\_Tenure**: Duration of the account in months.
* **Credit\_Card\_Type**: Type of credit card (e.g., Standard, Gold, Platinum).
* **Location**: Customer's city (e.g., Los Angeles, Phoenix).
* **Month\_1 to Month\_6**: Payment status for 6 months (On-time, Late, Missed).

- Data types:

* **Customer\_ID**: String (Categorical).
* **Age**: Integer.
* **Income**: Float (with missing values).
* **Credit\_Score**: Integer.
* **Credit\_Utilization**: Float.
* **Missed\_Payments**: Integer.
* **Delinquent\_Status**: Binary (Integer: 0/1).
* **Loan\_Balance**: Float (with missing values).
* **Debt\_to\_Income\_Ratio**: Float.
* **Employment\_Status**: String (Categorical).
* **Account\_Tenure**: Integer.
* **Credit\_Card\_Type**: String (Categorical).
* **Location**: String (Categorical).
* **Month\_1 to Month\_6**: String (Categorical: "On-time", "Late", "Missed").

# 3. Missing Data Analysis

Identifying and addressing missing data is critical to ensuring model accuracy. This section outlines missing values in the dataset, the approach taken to handle them, and justifications for the chosen method.

Key missing data findings:

- Variables with missing values:

1. **Income**
2. **Loan\_Balance**
3. **Employment\_Status**
4. **Debt\_to\_Income\_Ratio**
5. **Credit\_Card\_Type**

- Missing data treatment: [Deletion, Imputation, Synthetic Data, etc.]

**1. Income (Numerical, ~20+ missing)**

* **Best Treatment**: **Multiple Imputation (MICE)**
  + **Why?** Income is likely correlated with other variables (e.g., Credit\_Score, Employment\_Status). MICE preserves relationships and uncertainty better than mean/median imputation.
  + **Alternatives**:
    - *Regression Imputation*: If missingness is MAR (Missing at Random).
    - *Median Imputation*: If data is skewed and outliers exist.

**2. Loan\_Balance (Numerical, ~15+ missing)**

* **Best Treatment**: **Model-Based Imputation (Random Forest/KNN)**
  + **Why?** Loan balance may depend on Income, Credit\_Utilization, and Debt\_to\_Income\_Ratio. Machine learning models capture nonlinear relationships.
  + **Alternatives**:
    - *Mean/Median Imputation*: If missingness is minimal and random.
    - *Deletion*: Only if missingness is <5% of rows.

**3. Employment\_Status (Categorical, few missing)**

**Best Treatment**: **Mode Imputation (Most Frequent Category)**

* + - **Why?** Simple and preserves categorical distribution. If "Employed" is the mode, use it.
    - **Alternatives**:
    - *Add "Unknown" Category*: If missingness is meaningful (e.g., unemployed customers hiding status).

**4. Debt\_to\_Income\_Ratio (Numerical, few missing)**

* **Best Treatment**: **Mean/Median Imputation**
  + **Why?** Likely MCAR (Missing Completely at Random). Mean if normally distributed, median if skewed.

**5. Credit\_Card\_Type (Categorical, few missing)**

* **Best Treatment**: **Mode Imputation**
  + **Why?** Assume missing values are "Standard" (most common type).

# 4. Key Findings and Risk Indicators

This section identifies trends and patterns that may indicate risk factors for delinquency. Feature relationships and statistical correlations are explored to uncover insights relevant to predictive modeling.

Key findings:

- Correlations observed between key variables:

**Strong Positive Correlations**

1. **Credit\_Score 🡪 Income**
   * Higher incomes often correlate with higher credit scores.
2. **Credit\_Utilization 🡪 Missed\_Payments**
   * Customers using more credit (%) tend to miss more payments.
3. **Debt\_to\_Income\_Ratio 🡪 Delinquent\_Status**
   * Higher debt-to-income ratios likely link to higher delinquency risk.

**Moderate/Negative Correlations**

1. **Age 🡪 Delinquent\_Status**
   * Older customers may show slightly lower delinquency rates.
2. **Income 🡪 Missed\_Payments**
   * Lower incomes might correlate with more missed payments.
3. **Loan\_Balance 🡪 Credit\_Utilization**
   * Higher loan balances could drive up credit utilization.

**Weak or Context-Dependent Correlations**

1. **Employment\_Status 🡪 Delinquency**
   * "Unemployed" or "Self-employed" may weakly link to higher risk.
2. **Credit\_Card\_Type (Platinum) 🡪 Income**
   * Premium cards might weakly associate with higher incomes.

**Key Insights for Modeling**

* **Top Predictors of Delinquency**:  
  Credit\_Utilization, Missed\_Payments, Debt\_to\_Income\_Ratio.
* **Potential Confounders**:  
  Income and Age may mediate other relationships (e.g., older customers with higher incomes default less).
* **Feature Engineering Opportunities**:  
  Combine Credit\_Utilization + Missed\_Payments into a "Risk\_Score".

- Unexpected anomalies:

**1. Income Anomalies**

* **Anomaly**:
  + Some customers with **high incomes (>$150K)** have **low credit scores (<400)** or **delinquent status (1)**.
  + A few entries show **$0 or implausibly low incomes** (e.g., $694) alongside high credit utilization.
* **Potential Causes**:
  + Data entry errors (e.g., missing decimal points, unit mismatches).
  + "High-income" delinquents might be freelancers with volatile cash flow or tax evasion red flags.
* **Action**:
  + Audit rows where Income < 10,000 or Income > 200,000 with Delinquent\_Status = 1.
  + Cross-check with Employment\_Status (e.g., "Self-employed" vs. "Unemployed").

**2. Credit Utilization Outliers**

* **Anomaly**:
  + Customers with **0% credit utilization** but **multiple missed payments**.
  + A few show **>100% utilization** (e.g., 102.5%), suggesting data errors or credit limit breaches.
* **Potential Causes**:
  + Missing credit limit data or system errors in utilization calculations.
  + Closed accounts misclassified as active.
* **Action**:
  + Flag accounts with Credit\_Utilization > 100% or = 0% + Missed\_Payments > 0.
  + Verify if Loan\_Balance is null for these cases.

**3. Employment Status Inconsistencies**

* **Anomaly**:
  + "Retired" customers with **ages < 40** or "Employed" with **$0 income**.
  + Blank/missing Employment\_Status for high-risk delinquents.
* **Potential Causes**:
  + Misclassified self-employed/gig workers.
  + Fraudulent applications (e.g., fake employment data).
* **Action**:
  + Cross-tabulate Employment\_Status with Age and Income.
  + Investigate if missing employment data correlates with delinquency.

**4. Payment Behavior Contradictions**

* **Anomaly**:
  + Customers marked **"On-time" for all 6 months** but labeled **delinquent (**Delinquent\_Status = 1**)**.
  + Missed\_Payments > 0 but no late payments in monthly logs.
* **Potential Causes**:
  + Definition mismatches (e.g., delinquency based on external data).
  + Time lag between missed payments and reporting.
* **Action**:
  + Reconcile Missed\_Payments with monthly logs (Month\_1 to Month\_6).
  + Check if delinquencies are triggered by non-payment variables (e.g., collections).

**5. Geographic Oddities**

* **Anomaly**:
  + Certain locations (e.g., "Phoenix") show **higher delinquency rates** despite average incomes/credit scores.
  + "New York" customers with low utilization but high missed payments.
* **Potential Causes**:
  + Regional economic factors (e.g., job market crises).
  + Fraud rings or systemic reporting issues in specific branches.
* **Action**:
  + Calculate delinquency rates by Location and compare to regional economic data.

**6. Loan Balance Discrepancies**

* **Anomaly**:
  + Null Loan\_Balance for customers with high Debt\_to\_Income\_Ratio.
  + Loan\_Balance = 0 but active Credit\_Utilization.
* **Potential Causes**:
  + Unrecorded loans (e.g., payday loans) affecting debt ratios.
  + Data pipeline failures.
* **Action**:
  + Investigate if missing Loan\_Balance correlates with alternative data sources (e.g., credit bureau links).

# 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:

- *"How many records (rows) are in the dataset?”*

- *"What are the key variables (columns) in the dataset?"*

*- "Provide the data types for all columns in the dataset."*

*- “Identify the variables with missing values in the dataset.”*

*- “Choose the best treatment for the identified missing values from the dataset.”*

*- “Summarize the correlations observed between the key variables.”*

*- “what unexpected anomalies have emerged in the data, and which specific points warrant further investigation to uncover potential underlying causes?”*

# 6. Conclusion & Next Steps

This **Exploratory Data Analysis (EDA)** uncovered critical insights, anomalies, and data quality issues in the delinquency prediction dataset. Below is a consolidated summary of findings and actionable recommendations:

1. Data Quality & Missing Values

* Missing Data: Found in Income (~20+), Loan\_Balance (~15+), and Employment\_Status.
* Anomalies:
  + Implausible values (e.g., $0 income, >100% credit utilization).
  + Logical mismatches (e.g., "On-time" payments but marked delinquent).
* Correlations:
  + Strong: Credit\_Utilization 🡪 Missed\_Payments 🡪 Delinquent\_Status.
  + Moderate: Income 🡪 Credit\_Score, Age 🡪 Lower delinquency.

2. Unexpected Patterns

* High-Income Delinquents: Some high earners had poor credit behavior.
* Employment Inconsistencies: "Retired" at age <40, missing employment data.
* Geographic Bias: Certain cities (e.g., Phoenix) showed higher delinquency rates.

3.Feature Importance for Modeling

Top Predictors: Credit\_Utilization, Missed\_Payments, Debt\_to\_Income\_Ratio.

* Weak Signals: Credit\_Card\_Type, Location (needs feature engineering).

**Recommended Next Steps**

1. Data Cleaning & Imputation

* + Impute Missing Values:
    - Numerical (Income, Loan\_Balance): Use MICE or KNN imputation.
    - Categorical (Employment\_Status): Mode impute or flag as "Unknown".
  + Fix Anomalies:
    - Cap Credit\_Utilization at 100% or validate with business rules.
    - Remove/flag $0 income entries if unjustified.

2. Feature Engineering

* + Derive New Features:
    - Risk Score: Combine Missed\_Payments + Credit\_Utilization.
    - Income-to-Debt Buckets: Categorize Debt\_to\_Income\_Ratio (e.g., Low/Medium/High Risk).
  + Enhance Location Data: Merge external economic data (e.g., unemployment rates by city).

3. Model Preparation

* + Train-Test Split: Ensure temporal consistency (if time-based).
  + Class Imbalance Check: If Delinquent\_Status is skewed (e.g., 90% Non-Delinquent), use SMOTE or weighted loss.
  + Feature Selection: Drop weak predictors (Account\_Tenure?) after correlation analysis.

4. Validation & Monitoring

* + Track Imputation Impact: Compare model performance (AUC-ROC) before/after fixes.
  + Monitor Drift: Post-deployment, check for shifts in Income or Credit\_Utilization distributions.