# Exploratory Data Analysis (EDA) Summary Report

## 1. Introduction

The objective of this Exploratory Data Analysis (EDA) report is to analyze the financial and behavioral attributes of customers within the Delinquency Prediction Dataset provided by Tata iQ. The primary goal is to uncover data quality issues, patterns, and relationships that can help identify risk indicators for customer delinquency. This analysis prepares the dataset for predictive modeling and ensures data readiness.

## 2. Dataset Overview

The dataset contains detailed financial and behavioral attributes of customers used to predict delinquency risk. It includes both numerical and categorical variables that describe demographic information, credit behavior, and payment history.

Key dataset attributes:

* - Number of records: 10,000 (approx.)
* - Key variables: Customer\_ID, Age, Income, Credit\_Score, Credit\_Utilization, Missed\_Payments, Delinquent\_Account, Loan\_Balance, Debt\_to\_Income\_Ratio, Employment\_Status, Account\_Tenure, Credit\_Card\_Type, Location, Month\_1–Month\_6.
* - Data types: Numerical (Age, Income, Credit\_Score, Loan\_Balance, Debt\_to\_Income\_Ratio), Categorical (Employment\_Status, Credit\_Card\_Type, Location, Month\_1–Month\_6), Binary (Delinquent\_Account).

Initial review revealed a few anomalies such as outliers in Credit\_Utilization and Loan\_Balance, and missing values in Income and Credit\_Score columns.

## 3. Missing Data Analysis

Missing values were primarily observed in 'Income' and 'Credit\_Score'. A small number of records also had missing entries in 'Employment\_Status'. The percentage of missing data was below 10% across these variables.

Treatment approach:

* - Numerical variables (Income, Credit\_Score): Imputed using median values to maintain data distribution.
* - Categorical variables (Employment\_Status): Filled using the mode (most frequent category).

This approach ensures minimal distortion while preserving the integrity of statistical relationships.

## 4. Key Findings and Risk Indicators

EDA revealed several relationships and patterns that are likely to be important predictors of delinquency:

* - Customers with higher Credit\_Utilization (>80%) and lower Credit\_Score (<600) show increased probability of delinquency.
* - A negative correlation was observed between Income and Delinquent\_Account, suggesting that customers with lower income are more prone to default.
* - Missed\_Payments and Debt\_to\_Income\_Ratio have strong positive correlations with delinquency risk.
* - Employment\_Status also influences delinquency — unemployed or self-employed customers showed higher risk.

Outliers were identified in Loan\_Balance and Credit\_Utilization columns, which were capped using the IQR method.

## 5. AI & GenAI Usage

Generative AI tools (ChatGPT and Python-based data analysis assistants) were used to automate and enhance insights. AI assisted in summarizing dataset characteristics, identifying missing data patterns, and suggesting appropriate imputation techniques. The tools were also leveraged to interpret correlation results and generate narrative summaries for the report.

Example AI prompts used:

* - 'Summarize key patterns in the delinquency dataset and identify major correlations.'
* - 'Suggest imputation methods for missing financial values like Income and Credit\_Score based on best practices.'
* - 'Interpret the relationship between Credit\_Utilization and Delinquent\_Account using correlation and visualization results.'

## 6. Conclusion & Next Steps

The EDA confirmed that the dataset is well-structured for predictive modeling after cleaning and transformation. Key indicators of delinquency include high credit utilization, low credit score, and frequent missed payments. Future steps include feature engineering, scaling numerical variables, encoding categorical variables, and building classification models (e.g., Logistic Regression, Random Forest) to predict delinquency likelihood.