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

# 1. Introduction

[By reviewing the customer data set to identify key indicators and sorting the missing information and detecting the risk factors that affect delinquency.]

# 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, No of columns – 19

- 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 to Month 6]

- Data types: [Categorical variables: Customer ID, Employment Status, Credit Card Type, Location, Month 1 to Month 6]

[Numerical: Age, Income, Credit Score, Credit Utilization, Missed Payments, Delinquent Account, Loan Balance, Debt to Income Ratio, Account Tenure]

# 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: [Income (7.8%missing), Credit Score (0.4%missing), Loan Balance (5.8%missing)]

- Missing data treatment: [Income- *Advanced Imputation* (e.g., KNN, regression), due to a moderate missing rate (7.8%).

Credit Score- *Simple Imputation* (mean/median/mode), as the missing rate is very low (0.4%).

Loan Balance- *Advanced Imputation* (e.g., KNN, regression), due to a moderate missing rate (5.8%).]

# 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: [The correlations between numerical features and the target variable Delinquent Account are **weak**, indicating that no single numerical variable strongly predicts delinquency on its own.

* **Income** has a very slight positive correlation (corr = **0.045**), suggesting that higher income may be **slightly associated** with increased delinquency risk, though the relationship is very weak and may not be meaningful.
* **Account Tenure** shows a slight negative correlation (corr = - **0.040**), indicating that customers with longer account histories may be **slightly less likely** to be delinquent.
* **Credit Score**, somewhat surprisingly, has a small positive correlation with delinquency (corr = **0.035**). This might suggest **data issues** or underlying factors like recent borrowing behavior that aren’t captured directly.
* **Debt-to-Income Ratio** and **Credit Utilization** both show weak positive correlations (corr = **0.034)**, hinting that higher debt burden and credit usage may be **marginally linked** to delinquency, but again, the effect is minimal.

Categorical risk indicators -

* **Employment Status**:
  + Highest delinquency among **Unemployed (19.4%)** and inconsistent labels like *"employed"* vs *"Employed"* indicate need for **data cleaning**.
* **Credit Card Type**:
  + **Business (21.3%)** and **Student (17.9%)** cardholders have higher delinquency rates.
* **Location**:
  + **Los Angeles (19.6%)** and **Houston (16.8%)** show higher delinquency compared to other cities.

- Unexpected anomalies: [No numerical anomalies are detected. All values fall within ±3 standard deviations of the mean.]

# 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:

- ' Identify indicators like no of records, key variables, data types like categorical and numerical.'

- ‘Find variables with missing values and suggest suitable missing data treatment for these variables like deletion, Imputation, Synthetic Data etc.’

- ‘Summarize correlation observed between key variables and highlight unexpected anomalies requiring further investigation.’

# 6. Conclusion & Next Steps

The analysis of the delinquency prediction dataset revealed several key insights. Missing data was present in three numerical variables: Income (7.8%), Credit Score (0.4%), and Loan Balance (5.8%). It is recommended to use advanced imputation methods (like KNN or regression) for Income and Loan Balance, and simple imputation (mean or median) for Credit Score. Correlation analysis showed that no single numerical variable strongly predicts delinquency—Income, Account Tenure, Credit Score, Debt to Income Ratio, and Credit Utilization all had very weak correlations (absolute values below 0.05) with the delinquency outcome. Categorical analysis revealed higher delinquency rates among the Unemployed (19.4%), Business credit card holders (21.3%), and residents of Los Angeles (19.6%). Inconsistent labeling in the Employment Status field (e.g., “Employed”, “employed”, “EMP”) indicates a need for data cleaning. No significant numerical anomalies or outliers were detected.

Based on these findings, the recommended next steps are clean and standardize categorical labels, impute missing values using the appropriate techniques, engineer new features (e.g., credit usage tiers or behavioral scores), explore predictive modeling to uncover non-obvious patterns, and conduct deeper analysis into high-risk geographic or demographic segments. These actions will support a more accurate and actionable delinquency risk assessment framework.