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

**1. Introduction**

Purpose and Goals of the Report:

The purpose of this report is to analyze customer financial data to identify risk factors associated with loan delinquency. The goal is to prepare the dataset for predictive modeling by addressing missing values, detecting anomalies, and uncovering feature patterns that influence delinquency behavior.

**2. Dataset Overview**

Categorical Variables like Employment\_Status, Credit\_Card\_Type, Location, and monthly payment statuses (Month\_1 to Month\_6) have multiple unique values.

These will need proper encoding (e.g., one-hot or ordinal encoding).

Target Imbalance: Only 16% are delinquent. Modeling might require balancing techniques like SMOTE or weighted loss.

Monthly Status Columns (Month\_1–Month\_6): These are categorical (On-time, Late, Missed) and offer a temporal behavioral trend. Must be encoded carefully (ordinal or frequency-based).

**3. Missing Data Analysis**

Income: 39 missing values

Credit\_Score: 2 missing values

Loan\_Balance: 29 missing values

Imputation Strategy for Missing Values

Income (39 missing values)

Type: Continuous (numeric), positively skewed

Strategy: Fill missing values using the median income for each Employment\_Status group.

Reason: Median is not affected by outliers and gives better accuracy for skewed financial data.

Additional Step: Add a new column called is\_income\_missing with value 1 for missing rows and 0 otherwise.

Credit\_Score (2 missing values)

Type: Continuous (numeric)

Strategy: Fill missing values with the median credit score for people in the same Age or Income range.

Reason: Credit score is often related to age and income, so this keeps it consistent.

Additional Step: Add a column is\_credit\_score\_missing.

Loan\_Balance (29 missing values)

Type: Continuous (numeric)

Strategy: If both Income and Debt\_to\_Income\_Ratio are present, calculate missing Loan\_Balance as:

Loan\_Balance = Income × Debt\_to\_Income\_Ratio

Otherwise, use the median loan balance for imputation.

Additional Step: Add a column is\_loan\_balance\_missing.

Other Notes:

Do not use the mean for imputation, as it is sensitive to outliers.

For advanced models, techniques like KNN Imputation can be used, but median-based methods are simple and effective.

Always keep track of which values were filled in by creating "missing flags" for each imputed field.

**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:

Although no strong linear correlations were observed with Delinquent\_Account, weak positive associations were found with:

Income (correlation ≈ 0.045)

Credit\_Score (≈ 0.035)

Debt\_to\_Income\_Ratio (≈ 0.034)

Credit\_Utilization (≈ 0.034)

Slight negative correlations were observed with:

Missed\_Payments (≈ -0.026)

Account\_Tenure (≈ -0.040)

These weak correlations suggest that delinquency is influenced by a complex interaction of features, not dominated by any single variable.

Unexpected anomalies:

Several customers are marked as delinquent despite having zero missed payments, including:

CUST0018, CUST0033, CUST0043, and others.

A few customers have Credit Utilization exceeding 1.0 (e.g., CUST0090, CUST0266, CUST0293, and CUST0427), which is unusual and may indicate data entry errors or extreme financial behavior.

These cases warrant further investigation or possible exclusion/imputation during data cleaning.

**5. AI & GenAI Usage**

Use of Generative AI in Dataset Analysis

Generative AI tools were utilized to accelerate the process of data understanding, imputation, and risk pattern identification. Below are the key AI prompts issued during the analysis, along with the corresponding AI-generated summaries:

Prompt 1:

"Summarize key patterns, outliers, and missing values in this dataset. Highlight any fields that might present problems for modeling delinquency."

Result Summary:

The dataset has missing values in Income (39), Credit\_Score (2), and Loan\_Balance (29). No major outliers were detected using the IQR method, but the Delinquent\_Account field is imbalanced (only 16% delinquent). Categorical variables like Employment\_Status and Credit\_Card\_Type will need encoding. Monthly payment status columns could offer behavior-based insights but require feature engineering. Fields with missing or imbalanced data may pose challenges for modeling.

Prompt 2:

"Suggest an imputation strategy for missing values in this dataset based on industry best practices."

Result Summary:

Income: Impute using median income by Employment\_Status, with a binary flag for missingness.

Credit\_Score: Fill with median credit score by Age or Income range.

Loan\_Balance: Preferably compute using Income × Debt\_to\_Income\_Ratio; otherwise, use median.

Mean imputation was avoided due to outlier sensitivity. Binary indicators for missing fields were recommended for modeling transparency.

Prompt 3:

\*"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: [Summarize findings]

Unexpected anomalies: [Highlight data points requiring further investigation]"\*

Result Summary:

No strong linear correlations with delinquency were found. Weak positive correlations exist with Income, Credit\_Score, and Credit\_Utilization. A few customers showed delinquency despite zero missed payments. Others had Credit\_Utilization > 1.0, signaling potential anomalies or data entry errors that require further review.

**6. Conclusion & Next Steps**

Summary of Key Findings and Recommended Next Steps

Key Findings:

Missing Data: Critical fields like Income, Credit\_Score, and Loan\_Balance contain missing values that must be addressed before modeling. Median and formula-based imputations were recommended.

Data Imbalance: The target variable Delinquent\_Account is highly imbalanced (only 16% delinquent), which may bias models if left unaddressed.

Weak Correlations: No strong linear predictors of delinquency were found. However, mild correlations with Income, Credit\_Utilization, and Credit\_Score suggest they may still contribute to prediction in combination with other features.

Anomalies Detected: Cases were found where customers had no missed payments but were marked as delinquent, and others with unrealistic credit utilization above 1.0. These outliers require validation or exclusion.

Categorical Variables: Fields like Employment\_Status, Credit\_Card\_Type, and monthly payment statuses require proper encoding. The monthly data presents an opportunity for feature engineering to capture behavioral trends.

Recommended Next Steps:

Impute Missing Values

Apply median-based imputation and formula-driven approaches (e.g., Income × Debt Ratio for Loan Balance).

Include binary flags for imputed fields.

Handle Class Imbalance

Apply techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or assign class weights during model training.

Feature Engineering

Aggregate monthly payment behavior (e.g., total missed or late payments across 6 months).

Encode categorical fields using appropriate methods (e.g., one-hot or ordinal encoding).

Outlier Review

Investigate and clean data points with impossible or unusual values like credit utilization > 1.0.

Validate delinquency status where no missed payments are recorded.

Model Development

Begin baseline modeling with logistic regression or decision trees.

Use cross-validation and AUC/ROC metrics for evaluation.

Interpretability and Reporting

Leverage SHAP or feature importance techniques for model explanation.

Track performance across different customer segments.