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

# 1. Introduction

[Insert purpose and goal(s) of the report.]

# 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: [Insert count]

- Key variables: [List key columns and descriptions]

- Data types: [Categorical, Numerical, etc.]

# 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** (39 missing) – Continuous, right-skew likely.

**Best practice**: Use **predictive mean matching (PMM)** or **multiple imputation** with correlated predictors (e.g., Age, Employment\_Status, Debt\_to\_Income\_Ratio). This preserves variability and reduces bias compared to mean/median imputation.

Simpler fallback: Grouped median imputation (e.g., median per Employment\_Status & Location).

**2. Loan\_Balance** (29 missing) – Continuous, possibly correlated with Income & Credit\_Score.

**Best practice**: Regression-based imputation (e.g., random forest regression) using correlated variables like Credit\_Score, Debt\_to\_Income\_Ratio.

Simpler fallback: Grouped median by Credit\_Card\_Type & Account\_Tenure.

**3. Credit\_Score** (2 missing) – Continuous, low missingness.

**Best practice**: Median imputation or regression using Age, Loan\_Balance, Credit\_Utilization.

Since missing count is small, imputation impact will be negligible.

- Missing data treatment:

* Income: Regression or grouped median imputation; synthetic normal-distribution for simulation studies.
* Loan\_Balance: Regression or grouped median by Credit\_Card\_Type and Account\_Tenure.
* Credit\_Score: Median imputation due to very low missingness.

# 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:

* Higher **Credit Utilization** is moderately correlated with increased **Missed Payments** and **Delinquent Account** status.
* Lower **Credit Score** is strongly correlated with higher **Debt-to-Income Ratio** and more frequent **Missed Payments**.
* **Income** shows a mild negative correlation with delinquency—lower-income customers have higher delinquency rates.
* Longer **Account Tenure** is associated with lower delinquency risk.

- Unexpected anomalies:

* A subset of customers with **high Credit Scores** still exhibit high **Missed Payments**, suggesting potential recent financial distress or reporting issues.
* Several records have **unusually high Loan\_Balance** despite low Debt-to-Income ratios, indicating possible data entry errors or unique loan structures.
* Extreme **Income** outliers that do not align with credit usage patterns require further review.

# 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:

# *"Summarize key patterns in the dataset and identify anomalies."*

# *"Suggest an imputation strategy for missing income values based on industry best practices."*

# *"Propose best-practice methods to handle missing credit utilization data for predictive modelling."*

# *"Generate realistic synthetic income values for missing entries using normal distribution assumptions."*

# *"Identify correlations between key variables and highlight unusual anomalies that may require further investigation."*

# 6. Conclusion & Next Steps

The analysis identified clear relationships between financial behavior and delinquency risk. Higher credit utilization, lower credit scores, and elevated debt-to-income ratios emerged as key predictors of delinquency. Anomalies such as customers with high credit scores but frequent missed payments, and unusual loan balance patterns, require further investigation.

**Recommended next steps:**

* Conduct a deeper statistical analysis and correlation validation using the full dataset.
* Implement advanced imputation techniques (e.g., regression, MICE) for missing income and loan balance data.
* Review and validate anomalous records to confirm data accuracy.
* Incorporate flagged anomalies and missingness indicators into the predictive modelling process.
* Develop and test a delinquency prediction model using the identified risk factors.