**Exploratory Data Analysis (EDA)**

**Summary Report**

# Introduction

The objective of this analysis is to perform an **Exploratory Data Analysis (EDA)** on the *Delinquency Prediction Dataset* to identify key financial and behavioral indicators that influence loan delinquency.

The goal is to prepare the dataset for predictive modeling by cleaning, imputing missing data, and uncovering significant correlations that help in understanding customer credit risk.

# Dataset Overview

The dataset contains customer-level financial and demographic details used to predict whether a customer becomes **delinquent** (misses loan payments).

**Key attributes:**

* **Number of records:** 500
* **Number of variables:** 19
* **Data types:**

**Numerical:** 11 (e.g., Age, Income, Credit\_Score, Debt\_to\_Income\_Ratio)

**Categorical:** 8 (e.g., Employment\_Status, Credit\_Card\_Type, Location, Month\_1–Month\_6)

**Observations:**

* No duplicate records detected.
* Outliers were observed in Income, Loan\_Balance, and Credit\_Utilization (values > 1).
* Dataset now has **no missing values** after imputation.

# Missing Data Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Missing Values** | **Treatment** | **Justification** |
| Credit\_Score | 50 | Group-wise median by  Employment\_Status&Credit\_Card\_Type | Median is robust to outliers and respects segment variations. |
| Income | 39 | Group median by Employment\_Status | Avoids distortion from skewed salary distribution. |
| Loan\_Balance | 29 | Group median by Account\_Tenure | Captures logical link between loan balance and account age. |

# Key Findings and Risk Indicators

**Correlations Observed:**

* **Missed\_Payments (r ≈ +0.62)** — Strong positive correlation with delinquency, indicating that past payment behavior is a major predictor.
* **Credit\_Utilization (r ≈ +0.47)** — Higher utilization ratios signal greater credit dependency and risk.
* **Debt\_to\_Income\_Ratio (r ≈ +0.36)** — Higher DTI values associate with limited repayment capacity.
* **Credit\_Score (r ≈ −0.55)** — Lower scores correlate with increased delinquency likelihood.
* **Loan\_Balance (r ≈ +0.29)** — Larger outstanding balances moderately increase delinquency risk.

**Categorical Insights:**

* **Employment\_Status:** Part-time and self-employed individuals show higher delinquency rates.
* **Credit\_Card\_Type:** Customers with “Silver” cards are more prone to delinquency, possibly due to lower income brackets.
* **Location:** Urban customers show slightly higher delinquency, potentially linked to higher living expenses.

**Unexpected Anomalies:**

* A few records with Credit\_Utilization > 1.0 were detected — representing overextended credit, which itself may be a strong risk signal.
* Some very high Loan\_Balance and Income values may be outliers but retained for model robustness.

# AI & GenAI Usage

**Generative AI Tools (ChatGPT & Pandas AI)** were used to:

* Summarize statistical trends and missing data patterns.
* Recommend imputation strategies aligned with financial modeling best practices.
* Interpret correlation results to identify key delinquency risk indicators.

**Example prompts used:**

* Summarize key patterns, outliers, and missing values in this dataset based on `.info()`,`.describe()` and `.isnull()`,`.sum()` outputs.
* Suggest an imputation strategy for the Credit\_Score column based on financial industry best practices.
* Identify top 3 variables most likely to predict delinquency and explain their importance.

# Conclusion & Next Steps

The dataset is now fully cleaned, structured, and ready for predictive modeling.

Key risk indicators identified include:

1. Missed\_Payments
2. Credit\_Utilization
3. Debt\_to\_Income\_Ratio
4. Credit\_Score

These variables show the strongest relationship with loan delinquency and will be essential in feature selection.