**Exploratory Data Analysis (EDA) Summary**

**1. Introduction**

This report intends to analyze the credit and payment behavior dataset to identify key risk factors associated with delinquency. The goal is to uncover patterns and insights that will be used in predictive modeling efforts for credit risk assessment.

All additional files referenced in this document can be found in the same location as this document. The csv file is in the folder called ‘data’

**2. Dataset Overview**

This dataset contains demographic, financial and behavioral attributes of customers. It is used to predict risk of delinquency based on past financial activities, credit history, and employment status. Below is a detailed explanation of each column.

Dataset File Name: Loan\_Delinquency\_0610.csv

**Key dataset attributes**:

* Number of records: 500
* Key variables:
  + loanID: unique customer identifier (alpha-numeric)
  + age: customer age in years (numerical)
  + income: annual income in USD (numerical)
  + loan\_balance: outstanding loan amount in USD (numerical)
  + credit\_score: creditworthiness score, typically ranging from 300 to 850 (numerical)
  + months\_employed: number of months that a customer has had employment (numerical)
  + credit\_utilization: ratio of used credit to credit limit (numerical)
  + interest\_rate: the interest rate on the loan (numerical)
  + loan\_term: the length of time allowed for repayment of the loan (months) (numerical)
  + debt\_to\_income\_ratio: ratio of debt to income (numerical)
  + highest\_education\_level: highest level of education (e.g., ‘High School’, ‘Bachelor's’, ' Master's', ‘PhD’) (categorical)
  + employment\_status: current employment status (e.g., ‘Full-time’, Part-time’, 'Self-Employed', ‘Unemployed’) (categorical)
  + marital\_status: is customer married? (e.g., ‘Married’, ‘Single’, 'Divorced') (categorical)
  + has\_mortgage: does customer have a mortgage? (Binary: 0=No, 1=Yes)
  + has\_dependents: does customer have dependents? (Binary: 0=No, 1=Yes)
  + loan\_purpose: how a the loan will be used? (e.g. 'Auto',‘Business’,‘Education’, 'Home',‘Other’) (categorical)
  + has\_cosigner: indicator of the loan having a co-signer (Binary: 0=No, 1=Yes)
  + delinquent\_account: indicator of whether the customer has a delinquent account (Binary: 0=No, 1=Yes)
  + account\_tenure: number of years customer has had an active account (numerical)
  + location: geographic location (categorical)
  + month\_1 to month\_6: payment status over six months (e.g. ‘On-time’=0,‘Late’=1,‘Missed’=2) (numerical)
* Data types: Mix of numerical and categorical features.

Initial review confirmed no duplicate records.

**3. Missing Data Analysis**

**Prompt: “**Summarize key patterns, outliers, missing values and cardinality in this dataset. Highlight any fields that might present problems for modeling delinquency.”

Python code suggested for missing data analysis.txt – Python code for Missing Data

Here is a structured summary of the dataset **Loan\_Delinquency\_0610.csv**:

**🔹 Basic Info**

* **Total rows:** 500
* **Total columns:** 26

**🔹 Missing Values**

Several columns have missing values:

* loanID: 8 missing
* age, income, loan\_balance, credit\_utilization, interest\_rate: 3 each
* credit\_score, debt\_to\_income\_ratio: 7 each
* loan\_term: 1
* employment\_status: 4
* loan\_purpose: 3
* Monthly status (month\_1 to month\_6): up to 6 missing per column

**🔹 Column Types**

* **Numerical:** age, income, loan\_balance, credit\_score, months\_employed, credit\_utilization, interest\_rate, loan\_term, debt\_to\_income\_ratio, delinquent\_account, account\_tenure
* **Categorical/Object:** loanID, education/employment/marital status, mortgage/cosigner/dependents flags, loan purpose, location, monthly statuses

**🔹 Key Statistics (Numerical Variables)**

* age: Mean ~44.2 (range 18–69)
* income: Mean ~$83.5K (range ~$15K–$150K)
* loan\_balance: Mean ~$131.5K
* credit\_score: Mean ~571 (range 301–849)
* months\_employed: Mean ~58.7 months
* credit\_utilization: Mean ~0.49
* interest\_rate: Mean ~13.26%
* delinquent\_account: 10.2% have delinquency

**Prompt**: “Suggest an imputation strategy for missing values”

For the dataset **Loan\_Delinquency\_0610.csv**, here are **imputation strategies** tailored to the nature and distribution of each column with missing values:

**🔹 Numerical Variables**

| **Column** | **Missing Count** | **Suggested Strategy** |
| --- | --- | --- |
| age | 3 | **Median** (robust to outliers) |
| income | 3 | **Median** or **regression imputation** using age, employment, etc. |
| loan\_balance | 3 | **Median** or **model-based** |
| credit\_score | 7 | **KNN imputation** using age, income, loan\_balance, etc. |
| credit\_utilization | 3 | **Mean** or **median** |
| interest\_rate | 3 | **Median** (typical for skewed distributions) |
| loan\_term | 1 | **Mode** (likely categorical-like: 12/24/36/60 months) |
| debt\_to\_income\_ratio | 7 | **Regression imputation** (based on income and loan\_balance) |

**🔹 Categorical Variables**

| **Column** | **Missing Count** | **Suggested Strategy** |
| --- | --- | --- |
| loanID | 8 | **Drop rows** (assuming unique identifiers) |
| employment\_status | 4 | **Mode** or **predictive modeling** (e.g., decision tree) |
| loan\_purpose | 3 | **Mode** |

**🔹 Time-based Status Variables**

| **Column** | **Missing Count** | **Suggested Strategy** |
| --- | --- | --- |
| month\_1–month\_6 | up to 6 each | **Forward/Backward fill**, or **Mode per customer** |

**🔸 Recommended Approach in Python**

Python imputation suggested code.txt – Python code for Imputation

**4. Key Findings and Risk Indicators**

**a. Prompt 4a:** Analyze the correlation between customer income and delinquency risk, summarizing key findings in simple terms.

Python code suggested for income vs delinquency.txt

**A chart with a few squares

AI-generated content may be incorrect.**

Correlation Analysis: Income vs. Delinquency Risk

* Correlation coefficient: -0.081
  + This indicates a very weak negative correlation between income and delinquency.
  + In simple terms: higher income is slightly associated with lower delinquency risk, but the relationship is not strong.
* Visualization Insight:
  + The boxplot shows that the median income for delinquent customers is slightly lower than for non-delinquent ones.
  + However, there's a large overlap in income ranges between both groups.

Summary:

While higher income may slightly reduce the risk of delinquency, income alone is not a strong predictor of whether a customer will default. Other factors likely play a more significant role.

**b. Prompt 4b:** identify the top 3 risk factors associated with delinquency.

Python code suggestion for top 3 risk factors.txt

**top\_3\_features**

**Result**

credit\_utilization 0.098392

age 0.096207

income 0.092940

**Top 3 Risk Factors Associated with Delinquency**

Based on a Random Forest model, the **top 3 predictors** of loan delinquency are:

**Credit Utilization** – Customers using a high portion of their available credit are more likely to default.

**Age** – Younger or older borrowers may pose higher risk depending on their financial behavior.

**Income** – Lower income tends to be weakly associated with higher delinquency, though it's not the strongest factor.

These features show the highest importance in predicting whether a customer becomes delinquent.

**5. AI & GenAI Usage**

Generative AI tools were leveraged to expedite data summarization, identify patterns, and support feature encoding.

**Example AI prompts used:**

* “Summarize key patterns, outliers, and missing values in this dataset. Highlight any fields that might present problems for modeling delinquency.”
* “Suggest an imputation strategy for missing income values based on industry best practices.”
* “Analyze the correlation between customer income and delinquency risk, summarizing key findings in simple terms.”
* “identify the top 3 risk factors associated with delinquency.”

These AI-assisted insights improved data preprocessing efficiency and model readiness.

**6. Conclusion & Next Steps**

The EDA process confirmed the dataset's integrity and identified critical features linked to delinquency risk. The next steps include:

* Feature engineering to create composite risk scores.
* Splitting the data into training and test sets.
* Training and tuning predictive models such as Random Forest and evaluating performance.
* Deploying the best-performing model for credit risk prediction.

Further analysis of outliers and domain consultation is recommended to refine the model.