Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

The primary purpose of this dataset is to enable the development of machine learning models that can predict **delinquent behavior**—a critical task for credit institutions aiming to reduce risk exposure, manage lending operations, and enhance decision-making in credit approvals. To **accurately classify or score customers** based on their likelihood of defaulting. Ultimately, this supports:

- Improved customer segmentation,
- Risk-based pricing strategies,
- More informed credit and lending decisions.

2. Dataset Overview

The **Delinquency Prediction Dataset** is designed for financial risk analysis, particularly aimed at predicting the likelihood of a customer becoming delinquent on their credit or loan obligations. This dataset comprises information from 500 customers, with each entry representing various personal, financial, and behavioral features relevant to creditworthiness assessment.

Anomalies & Inconsistencies

- No extreme outliers were detected based on 3-standard deviation filtering.
- No duplicate records were explicitly found during the initial analysis (assumed from shape).

Key dataset attributes:

- Number of records: 891
- Key variables: Income, Credit_Score, Debt_to_Income_ratio, Loan_Balance, Delinquent_Accounts
- Data types:
  - Numerical : Income and Credit_score
  - Categorical : Credit_Card_Type and Location

3. Missing Data Analysis

Key missing data findings:

- Variables with missing values:
  - Income: 39 missing values
  - Loan_Balance: 29 missing values
  - Credit_Score: 2 missing values
- Missing data treatment:
  - Credit_Score (2 missing values) :

    - **Recommended Treatment**: **Imputation (Mean or Median)**
    - **Reason**: Very few values are missing (<1%), and Credit_Score is critical for modeling delinquency. Simple imputation won't distort the data.

  - Loan_Balance (29 missing values) :
    - **Recommended Treatment**: **Median Imputation**
    - **Reason:** Loan balances can have outliers; median is more robust than mean. The number of missing values is moderate and manageable.
  - **Income (39 missing values) :**
    - **Recommended Treatment: Model-Based Imputation or KNN Imputation**
    - **Reason: Income is important for financial analysis, and imputing based on similar records (using KNN or regression) will preserve relationships better than dropping or filling with a constant.**

4. Key Findings and Risk Indicators

An exploratory analysis of the dataset reveals several key trends that may indicate risk factors for delinquency.

Key findings:

- Correlations observed between key variables: Using correlation analysis, three variables were found to have the strongest (though relatively weak) associations with the target variable **Delinquent_Account**:

  - **Income** showed a positive correlation with delinquency ($r \approx 0.045$), suggesting that income level—possibly when unstable or mismatched with expenses—could influence delinquent behavior.

- **Credit_Score** was also positively correlated (r ≈ 0.035), indicating that lower credit scores may slightly increase the likelihood of delinquency.

- **Debt-to-Income Ratio** (r ≈ 0.034) highlights how financial strain, reflected in high debt relative to income, can be a risk factor for missed payments.

No significant outliers were detected, which suggests the dataset is relatively clean. These patterns suggest that **financial stress indicators** such as low income, high debt load, and poor credit ratings play a modest but measurable role in predicting delinquency. These insights will help inform the selection and engineering of features in predictive modeling efforts.

5. AI & GenAI Usage

To address missing values in the **Income** variable, Generative AI tools were used to recommend imputation strategies aligned with industry best practices.

- **For small to moderate levels of missing data** (as observed in this dataset), **median imputation** is preferred over mean, as it is more robust to outliers and skewed distributions—both common in income data.
- Alternatively, for improved accuracy, **model-based imputation techniques** such as **K-Nearest Neighbors (KNN)** or **regression imputation** could be used. These methods utilize related features like credit score, debt-to-income ratio, and employment status to estimate missing income values more precisely.

6. Conclusion & Next Steps

- Key Findings :
    - The dataset consists of 891 records with a mix of numerical and categorical variables relevant to delinquency prediction.
    - Missing values were identified in key fields: **Income (39)**, **Loan_Balance (29)**, and **Credit_Score (2)**. These were treated using appropriate imputation strategies.
    - Correlation analysis identified **Income**, **Credit_Score**, and **Debt-to-Income Ratio** as the top three indicators potentially influencing delinquency.
    - No significant outliers or duplicate records were found, indicating good data quality.

- Generative AI tools provided efficient insights on data trends, missing value treatments, and feature importance, enhancing the overall analysis process.

- Next Steps :
    - **Implement a predictive model** (e.g., logistic regression, random forest) to assess delinquency risk using identified key variables.
    - **Conduct feature engineering** to extract more predictive insights, especially from categorical variables.
    - **Visualize key relationships** through Power BI or matplotlib/seaborn to support decision-making.
    - **Use insights for strategic planning**—such as adjusting lending policies, targeting high-risk profiles, or improving customer segmentation.