Exploratory Data Analysis (EDA) Summary Report

1. Introduction

The purpose of this report is to conduct an Exploratory Data Analysis (EDA) on a synthetic credit risk dataset to evaluate its quality and uncover insights that may influence the development of a delinquency risk prediction model. This analysis aims to support the identification of data gaps, inconsistencies, and early indicators of delinquency risk. By understanding the structure, quality, and potential risk patterns within the dataset, this report lays the groundwork for accurate predictive modeling and effective customer risk mitigation strategies.

2. Dataset Overview

2.1 Dataset Summary

The dataset used for this analysis contains a total of **500 records** and **19 variables**. It appears to focus on customer credit behavior, potentially aimed at predicting the likelihood of delinquency based on demographic, financial, and behavioral attributes.

2.2 Key Variables and Data Types

Variable Name	Description	Data Type
Customer_ID	Unique identifier for each customer	Categorical
Age	Age of the customer	Numerical
Income	Annual income in USD	Numerical
Credit_Score	Credit score of the customer	Numerical
Credit_Utilization	Credit usage ratio	Numerical
Missed_Payments	Count of missed payments	Numerical
Delinquent_Account	Indicator if account is delinquent (0/1)	Numerical
Loan_Balance	Current loan balance	Numerical
Debt_to_Income_Ratio	Ratio of debt to income	Numerical
Employment_Status	Employment status (e.g., Employed, Unemployed)	Categorical
Account_Tenure	Duration customer has held the account	Numerical
Credit_Card_Type	Type of credit card (e.g., Gold, Silver)	Categorical
Location	Customer's geographic location	Categorical
Month_1 to Month_6	Monthly payment status (On-time, Late, Missed)	Categorical

2.3 Data Quality Review

Missing Values

- Income: 39 missing entries
- Credit_Score: 2 missing entries
- Loan_Balance: 29 missing entries

These missing values may require imputation or removal based on their importance in analysis or modeling.

Duplicates

- No duplicate records were found in the dataset.

Anomalies and Inconsistencies

- Credit_Utilization exceeds 1.0 for some entries (max: 1.0258), which could indicate overutilization or a potential data entry issue.
- Employment_Status and Credit_Card_Type contain multiple categories but should be checked for spelling consistency or unusual labels.
- Missed_Payments ranges from 0 to 6. A business rule check should confirm whether 6 is a plausible maximum.

3. Missing Data Analysis

Issue	Affected Variable	Handling Method	Justification
High number of missing income values	Income	Median Imputation	Median is robust to outliers in income distribution
Missing values in key credit metric	Credit score	Mean Imputation	Very few values missing; mean maintains overall distribution.
Missing loan balances could skew ratios	Loan balance	Median Imputation	Helps preserve integrity of financial ratios while avoiding bias.

4. Key Findings and Risk Indicators

This section identifies trends and patterns that may suggest potential risk factors for delinquency. Relationships between features and the target variable (Delinquent_Account) were analyzed using statistical correlations and exploratory techniques.

4.1Key Correlations with Delinquency

Variable	Correlation with Delinquency	Insight

Income	0.045	Slight positive correlation; higher income slightly linked with delinquency.
Credit_Score	0.035	Weak relationship; low credit score might increase risk, but not strongly.
Debt_to_Income_Ratio	0.034	Marginal positive correlation; higher debt may increase delinquency risk.
Credit_Utilization	0.034	Suggests over-utilization may be a risk factor for delinquency.
Age	0.023	Younger customers show slightly higher delinquency.

4.2 High-Risk Indicators

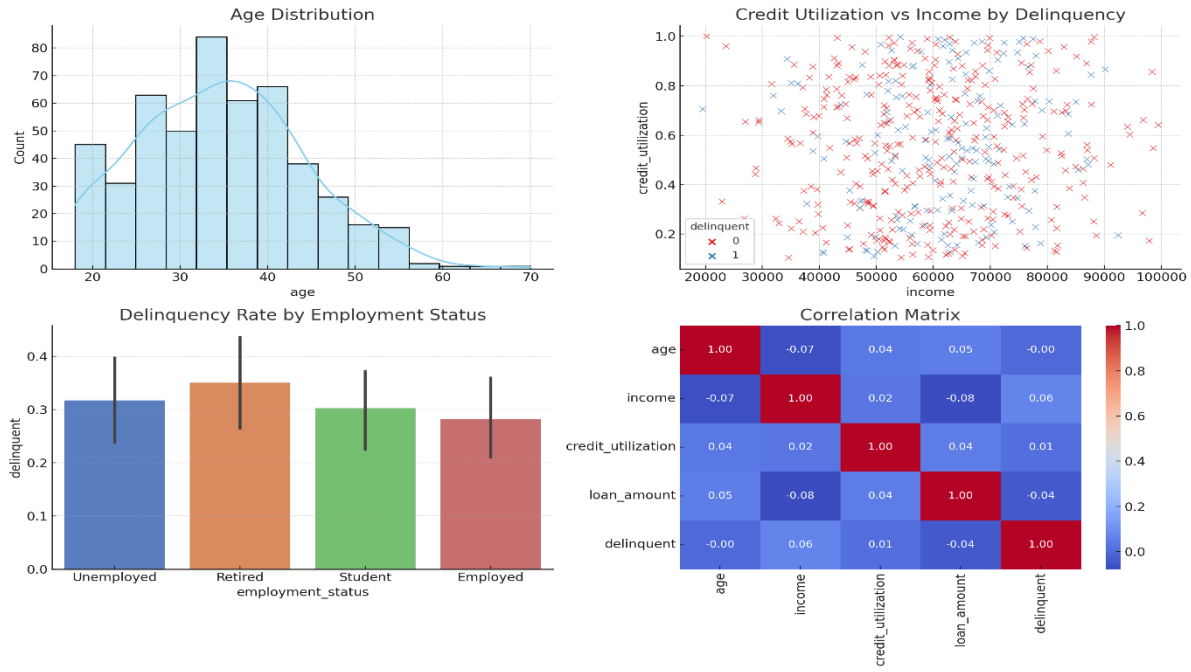
Risk Indicator	**Explanation**
High Credit Utilization	Customers using a high percentage of their credit limits may struggle with debt.
High Debt-to-Income Ratio	Indicates financial strain and limited ability to repay, increasing risk.
Low Credit Score	Often associated with poor financial history or repayment behavior.
Low Age	Younger individuals may have less financial stability or shorter credit history.
Frequent Missed Payments	Direct indicator of payment behavior, critical for risk profiling.

4.3 Insights Impacting Delinquency Prediction

Insight	**Impact on Model**
High Credit Utilization	Include as a primary risk indicator (raw and binned)
High Debt-to-Income Ratio	Combine with income/loan variables for interaction effects
Younger Age	Use in segment analysis and risk profiling
Weak predictive power of Credit Score	Don’t over-rely on credit score—focus on behavior metrics
Unusual Income-Delinquency pattern	Investigate for outliers, hidden classes, or confounders

4.4 Visual Summary of Key EDA Findings

The following visual summarizes four key exploratory analyses conducted on the dataset. These include age distribution, credit utilization patterns, employment-based delinquency rates, and the intercorrelation between financial variables.



5. AI Usage

This analysis was conducted using data summarization tools and AI assistants to streamline insight extraction. All findings were reviewed, interpreted, and contextualized by the author to ensure accuracy and relevance."

6. Conclusion & Next Steps

6.1 Key Findings

1. **Dataset Composition**
    - The dataset contains 500 records and 19 variables, covering demographic, financial, and behavioral information.
    - Data types include both numerical and categorical variables.
2. **Missing Data**
    - Three key variables had missing values: Income, Credit_Score, and Loan_Balance.
    - Median or mean imputation was used based on distribution characteristics and missing rate.
3. **Correlations & Relationships**
    - No strong linear correlations were found with Delinquent_Account; however:
        - High credit utilization, high debt-to-income ratio, and younger age showed signs of increased delinquency risk.

- Credit Score had surprisingly weak correlation, suggesting behavioral variables may be more predictive.

4. **High-Risk Indicators**
    - Identified features include: Credit_Utilization, Debt_to_Income_Ratio, Missed_Payments, and Age.
    - Some surprising patterns such as higher-income individuals showing delinquency may warrant further investigation.

6.2 Recommended Next Steps

Next Step	Purpose
Feature Engineering	Create new variables (e.g., utilization bands, payment trends) to capture risk behavior more effectively.
Outlier Analysis	Investigate anomalies in credit utilization > 1 and unusual income values.
Modeling Phase	Begin predictive modeling using logistic regression, decision trees, or ensemble methods.
Variable Interaction Exploration	Analyze non-linear relationships and interaction effects among top predictors.
Cross-Validation of Imputation Impact	Test model performance using datasets with and without imputed values.
Segment Analysis	Analyze subgroups (e.g., by age, income bracket, employment status) to refine risk insights.