**Exploratory Data Analysis (EDA) Summary**

**1. Introduction**

This report aims 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 inform predictive modeling efforts for credit risk assessment.

**2. Dataset Overview**

The dataset comprises records of customers’ demographic, financial, and payment history data.

**Key dataset attributes:**

* **Number of records:** 436
* **Key variables:**
  + customer\_id: Unique customer identifier
  + age: Customer age (numerical)
  + income: Annual income in USD (numerical)
  + credit\_score: Creditworthiness score (numerical)
  + credit\_utilization: Ratio of used credit to credit limit (numerical)
  + missed\_payments: Number of missed payments (numerical)
  + delinquent\_account: Indicator if account is delinquent (binary)
  + loan\_balance: Outstanding loan amount (numerical)
  + debt\_to\_income\_ratio: Ratio of debt to income (numerical)
  + employment\_status: Employment category encoded numerically
  + account\_tenure: Duration of account in years (numerical)
  + credit\_card\_type: Encoded credit card category (numerical)
  + location: Encoded geographic location (numerical)
  + month\_1 to month\_6: Payment status over six months (encoded categorical)
* **Data types:** Mix of numerical and encoded categorical features.

Initial review confirmed no duplicate records. Column names were normalized to lowercase for consistency.

**3. Missing Data Analysis**

The dataset was checked for missing values and appropriate handling was applied.

**Key missing data findings:**

* **Variables with missing values:** None detected after cleaning.
* **Missing data treatment:** Not applicable (no missing data found).

If missing data were present, imputation strategies such as mean/median for numerical fields or mode for categorical fields would be recommended.

**4. Key Findings and Risk Indicators**

Analysis revealed several relationships between features and delinquency risk.

**Key findings:**

* Negative correlation between credit\_score and missed\_payments, indicating lower credit scores correspond to more missed payments.
* Positive correlation between debt\_to\_income\_ratio and delinquency, suggesting higher debt burdens increase risk.
* Payment status across recent months (month\_1 to month\_6) strongly correlates with delinquency labels.
* No significant anomalies found, but outliers in loan\_balance and income warrant further examination.

**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 in the dataset and identify anomalies.”
* “Suggest an imputation strategy for missing income values based on industry best practices.”
* “Provide encoding methods for categorical payment status features.”

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.