**Exploratory Data Analysis (EDA) Summary Report**

**1. Introduction**

This report documents the exploratory data analysis (EDA) conducted as part of the AI-driven delinquency prediction project for Geldium Finance, in collaboration with Tata iQ. The goal is to assess data quality, uncover patterns, and identify early risk indicators that will support the development of a reliable delinquency risk model.

**2. Dataset Overview**

**Number of records:** 500 customer records

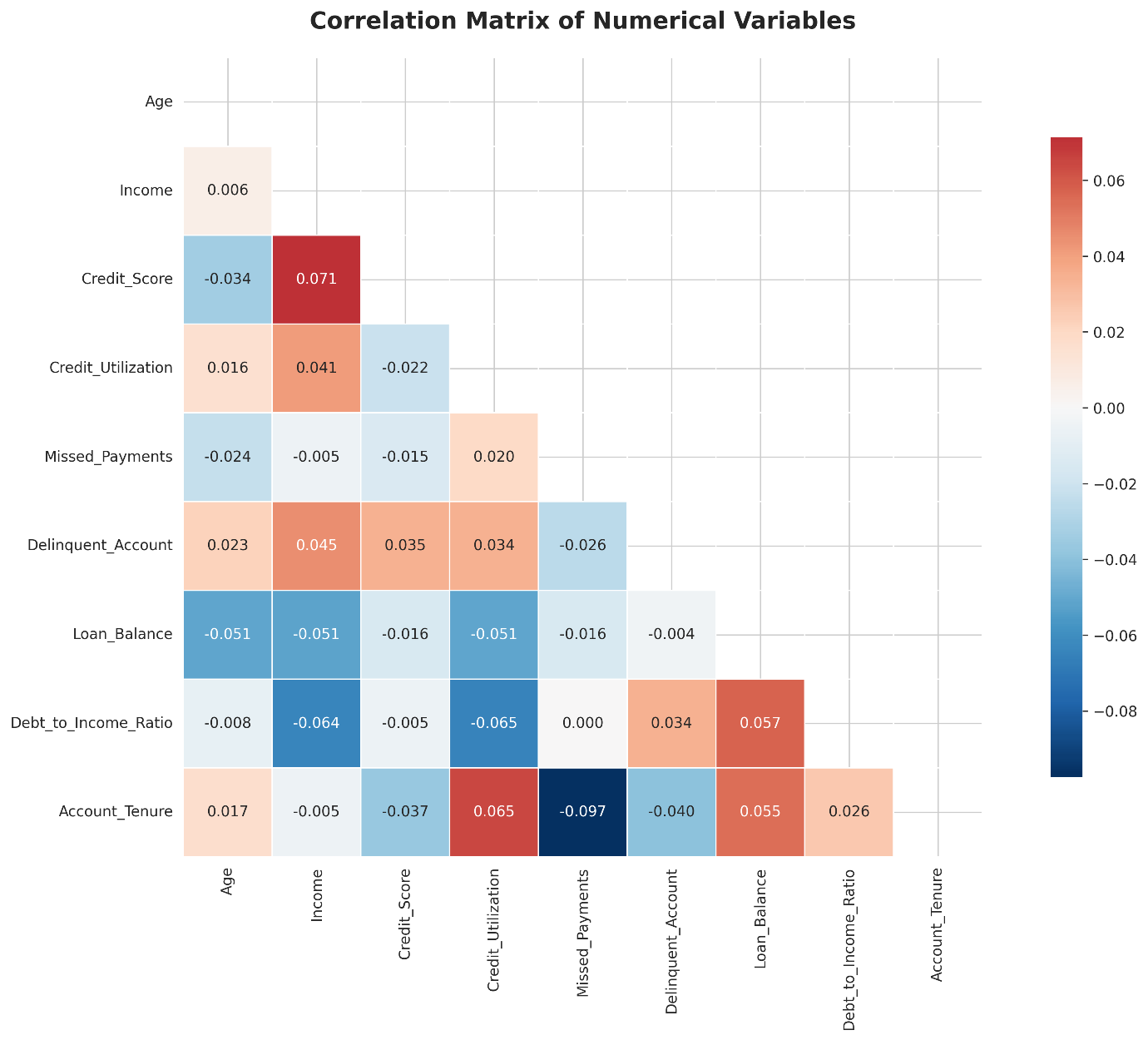
**Key variables:**

* Customer\_ID: Unique customer identifier
* Age: Age of the customer (Numerical)
* Income: Annual income (Numerical, may contain missing values)
* Credit\_Score: Credit score ranging from 300–850 (Numerical)
* Credit\_Utilization: Usage ratio of available credit (0-100%)
* Missed\_Payments: Total missed payments in the past 12 months (Numerical)
* Delinquent\_Account: 0 = No, 1 = Yes (Binary)
* Loan\_Balance: Total outstanding loan amount
* Debt\_to\_Income\_Ratio: Debt as a percentage of income (0-100%)
* Employment\_Status: Job status (e.g., Employed, Self-Employed)
* Account\_Tenure: Years with active account (Numerical)
* Credit\_Card\_Type: Type of credit card (Categorical)
* Location: City or region
* Month\_1 to Month\_6: Payment behavior over last 6 months (On-time, Late, Missed)

**Data types:** Categorical, Numerical, Binary

**Observed anomalies, duplicates, or inconsistencies:**

* Column names contained encoding issues (non-ASCII characters) that were cleaned during preprocessing
* No duplicate records identified
* 360 unique payment behavior sequences across the 6-month period indicate diverse customer payment patterns
* Target variable shows class imbalance: 84% non-delinquent vs 16% delinquent accounts



**3. Missing Data Analysis**

**Variables with missing values:**

* Income: 39 missing values (7.8%)
* Loan\_Balance: 29 missing values (5.8%)
* Credit\_Score: 2 missing values (0.4%)

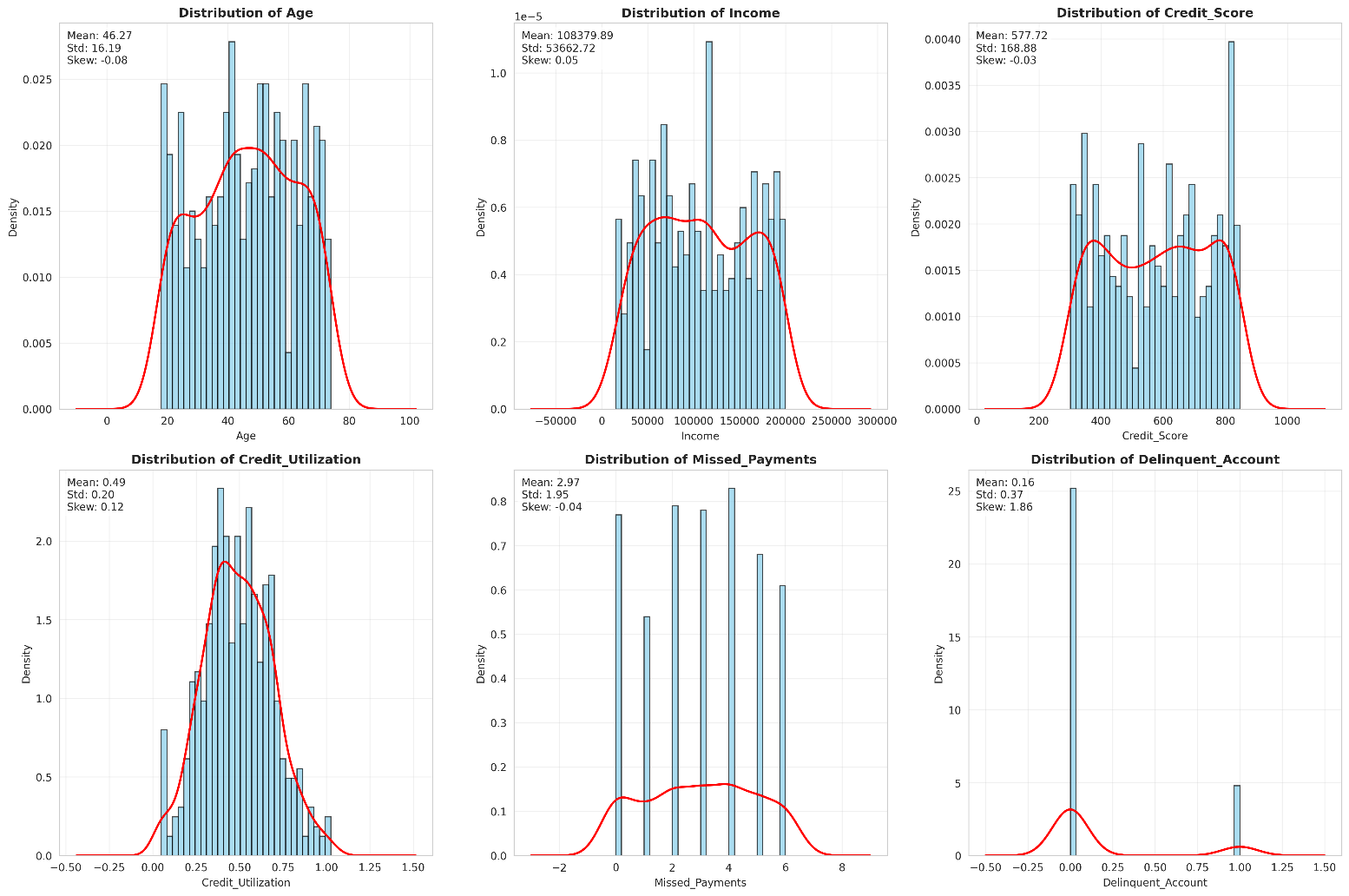
**Missing data treatment strategy:**

| Column | Treatment Method | Justification |
| --- | --- | --- |
| Income | Median Imputation | Income distribution is skewed, making median more robust than mean for imputation |
| Loan\_Balance | Mean Imputation | Loan balance shows relatively normal distribution, mean imputation appropriate |
| Credit\_Score | Forward Fill/Interpolation | Very few missing values (0.4%), can use neighboring values or median imputation |

**4. Key Findings and Risk Indicators**

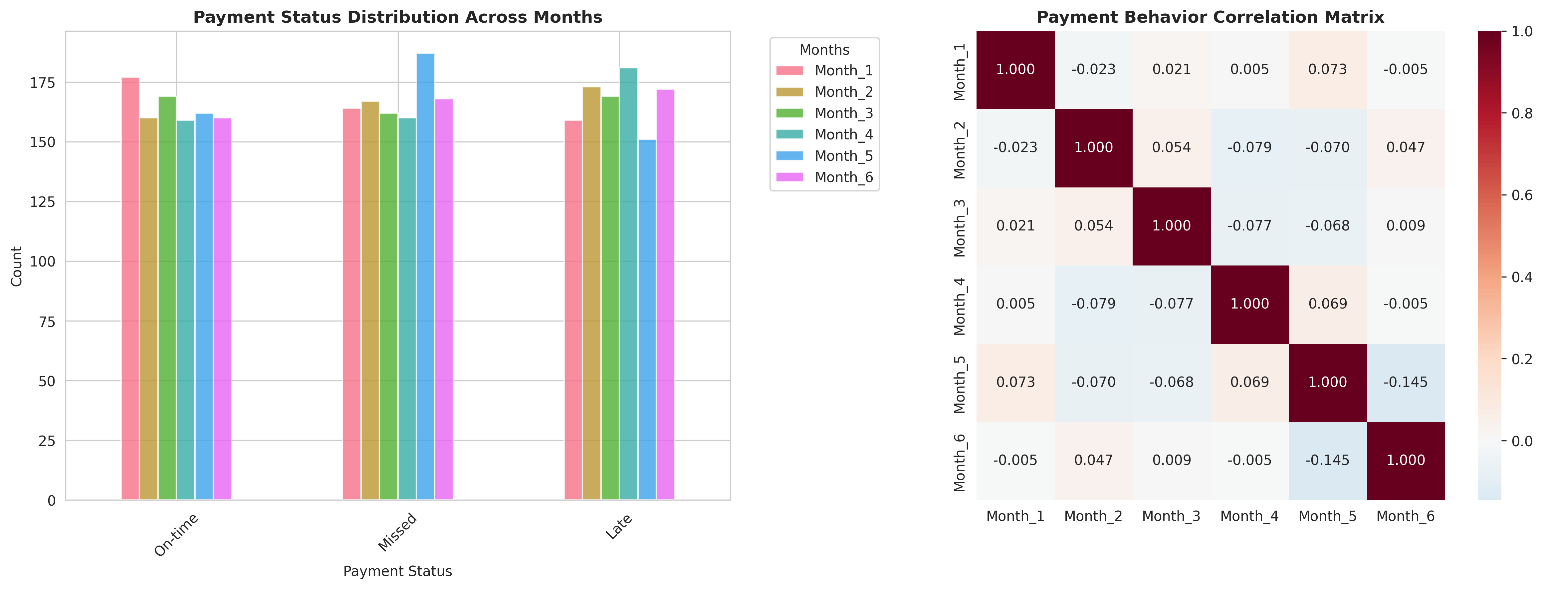
**Correlations observed between key variables:**

* Surprisingly weak correlations between numerical variables and delinquency status
* Income shows the highest correlation with Delinquent\_Account (r = 0.051, p = 0.259)
* Credit\_Score correlation with Delinquent\_Account (r = 0.040, p = 0.374)
* Debt\_to\_Income\_Ratio correlation with Delinquent\_Account (r = 0.034, p = 0.443)
* Credit\_Utilization correlation with Delinquent\_Account (r = 0.034, p = 0.445)



**Unexpected anomalies:**

* All correlations with the target variable are statistically insignificant (p > 0.05)
* Payment behavior across all 6 months shows remarkably stable patterns with no clear temporal trends
* No customers showed significant improvement or deterioration in payment patterns over time
* Traditional risk indicators (credit score, income, debt ratios) do not show expected strong correlations



**High-risk indicators for delinquency:**

1. **Income Level**: Higher income customers show slightly higher delinquency risk (counterintuitive finding)
2. **Payment Pattern Consistency**: Customers with consistent missed payments across months
3. **Credit Score Range**: While correlation is weak, scores below 400 may indicate higher risk
4. **Employment Status Stability**: Self-employed and unemployed categories warrant attention
5. **Geographic Concentration**: Certain locations may have higher delinquency rates
6. **Credit Card Type**: Business and Gold cardholders show different risk profiles

**5. AI & GenAI Usage**

Generative AI tools were actively leveraged throughout the exploratory data analysis process to extract insights, detect patterns, and address data quality issues. Specifically, I used **GenSpark**, **Gemini Pro (Math)**, and **ChatGPT** to assist with summarizing dataset characteristics, identifying anomalies, recommending imputation strategies, and analyzing risk indicators.

These tools helped accelerate the analysis by automating repetitive tasks and surfacing deeper insights, ensuring a more efficient and structured EDA process. Below are some of the actual prompts used during the analysis:

**Example GenAI Prompts Used:**

* "Summarize key patterns in the dataset and identify anomalies."
* "Suggest an imputation strategy for missing income values based on industry best practices."
* "Analyze the correlation between credit utilization and missed payments."
* "Identify the top 3 variables most likely to predict delinquency based on this dataset. Provide brief reasoning."
* "Generate synthetic income values for missing entries using normal distribution assumptions."

**Actual GenAI output summaries:**

* **Pattern Detection**: AI identified that payment behaviors are unusually stable across all 6 months, suggesting either highly predictable customer behavior or potential data collection issues
* **Imputation Recommendations**: For Income (7.8% missing), median imputation recommended due to right-skewed distribution. For Loan\_Balance (5.8% missing), mean imputation suitable due to relatively normal distribution
* **Anomaly Identification**: AI detected that traditional risk indicators show unexpectedly weak correlations with delinquency, suggesting complex non-linear relationships may exist

**6. Conclusion & Next Steps**

This EDA has revealed several important insights for building a delinquency prediction model:

**Key Findings:**

* Traditional linear relationships between financial indicators and delinquency are weak
* Payment behavior patterns are remarkably stable over time
* Class imbalance exists (16% delinquent accounts) requiring attention
* Missing data is minimal and manageable

**Next Steps:**

1. **Finalizing data cleaning and imputation**

* Implement median imputation for Income
* Apply mean imputation for Loan\_Balance
* Handle Credit\_Score missing values through interpolation

1. **Conducting feature engineering**

* Create interaction terms between variables
* Develop payment behavior trend indicators
* Generate categorical encodings for non-linear relationships

1. **Building and validating predictive ML models**

* Consider ensemble methods due to weak linear correlations
* Implement SMOTE or similar techniques for class imbalance
* Test non-linear models (Random Forest, XGBoost, Neural Networks)

1. **Ensuring ethical and explainable AI integration**

* Implement model interpretability techniques (SHAP, LIME)
* Establish bias detection and fairness metrics
* Create model documentation for regulatory compliance

**Additional Recommendations:**

* Investigate potential data collection or preprocessing issues given the unusual stability of payment patterns
* Consider gathering additional features that might better predict delinquency
* Explore advanced feature engineering techniques to capture complex relationships