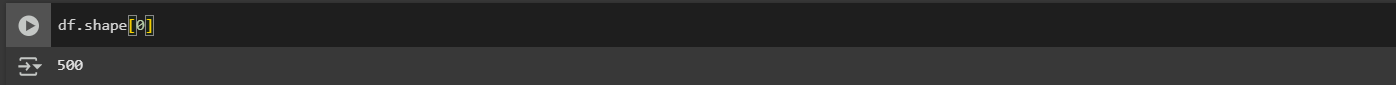
**Exploratory Data Analysis (EDA) Summary**   
**Report Template**

# 1. Introduction

# The purpose of this report is to perform an exploratory data analysis (EDA) on a customer financial dataset to uncover patterns, identify risk indicators, handle missing values, and generate insights for delinquency prediction modeling. This initial analysis ensures data quality and readiness for subsequent predictive modeling.

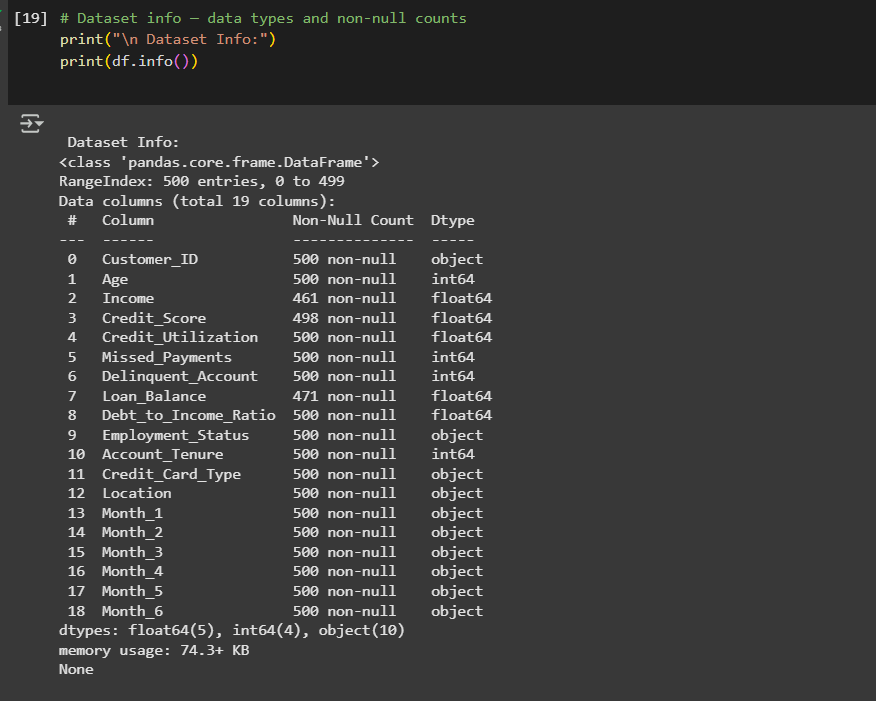
# 2. Dataset Overview

 **Number of records**:   


 **Key variables**:

* Customer\_ID — Unique customer identifier
* Age — Customer’s age in years
* Income — Annual income (USD)
* Credit\_Score — Credit rating (300–850)
* Credit\_Utilization — % of credit currently used
* Missed\_Payments — Number of missed payments in 12 months
* Delinquent\_Account — Target variable (0 = No, 1 = Yes)
* Debt\_to\_Income\_Ratio, Employment\_Status, Account\_Tenure, Credit\_Card\_Type, Location, Month\_1 to Month\_6

 **Data types**:

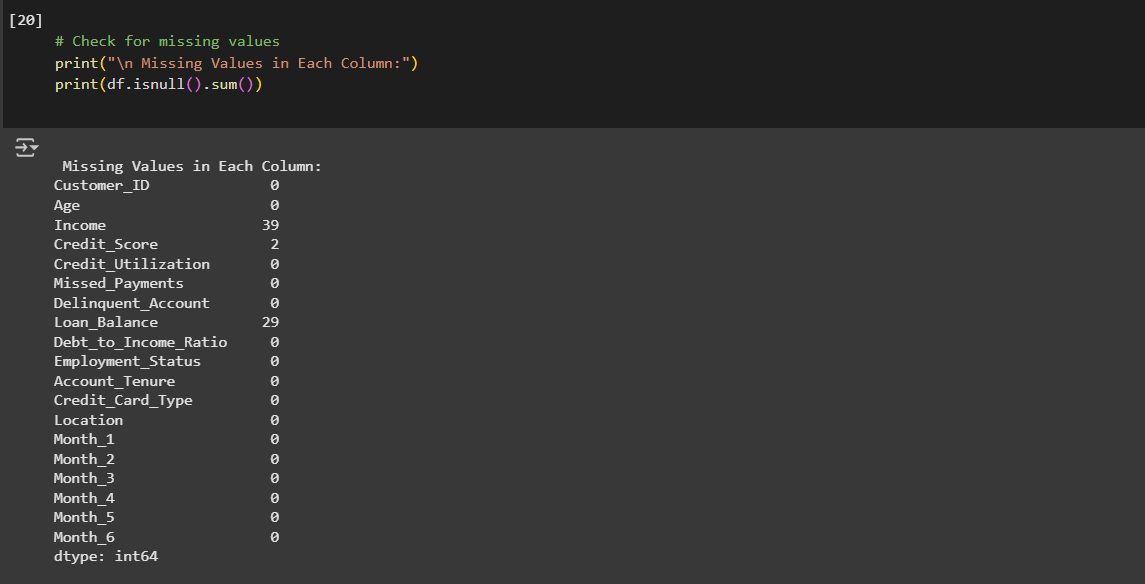
 **Anomalies/duplicates**:

* No duplicate Customer\_ID values detected.
* Outliers found in Income, Credit\_Utilization, and Debt\_to\_Income\_Ratio (See Outlier Detection output screenshot)

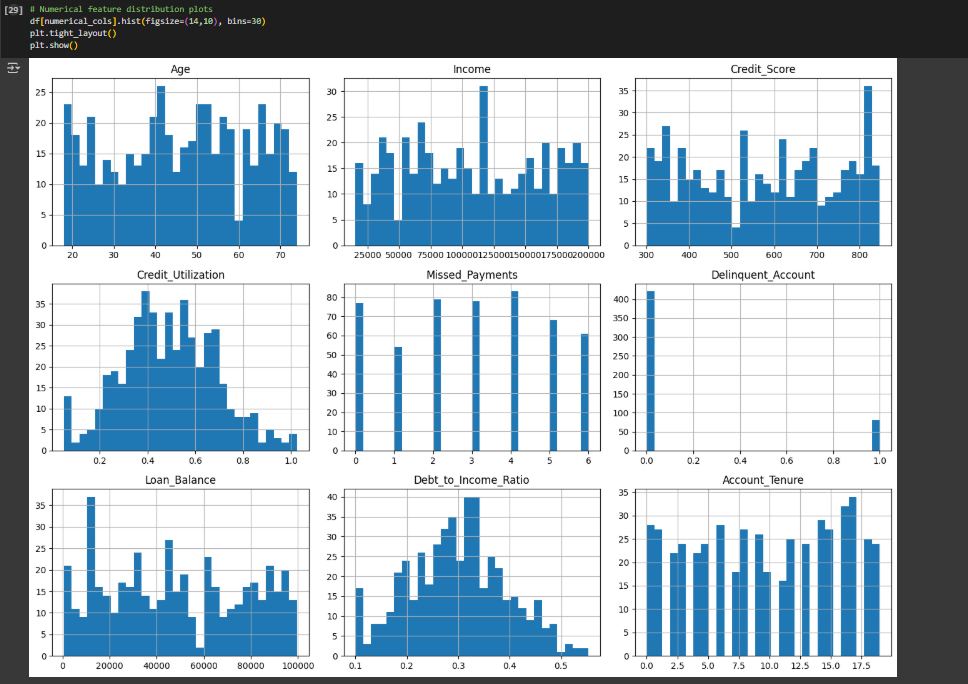
# 3. Missing Data Analysis

 **Variables with missing values**:

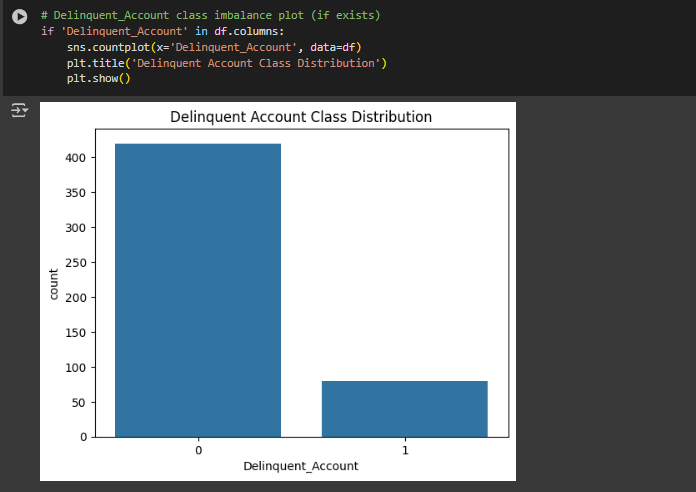
* Income
* Credit\_Utilization
* Possibly others depending on your earlier output (check df.isnull().sum())



 **Missing data treatment**:

* Income → Imputed with synthetic values generated via **log-normal distribution** based on existing data
* Credit\_Utilization → Imputed using **segment-wise median** (by Employment Status), with missing indicator flag
* Employment\_Status → Imputed using **mode (most frequent category)**
* 

 **Justification**:

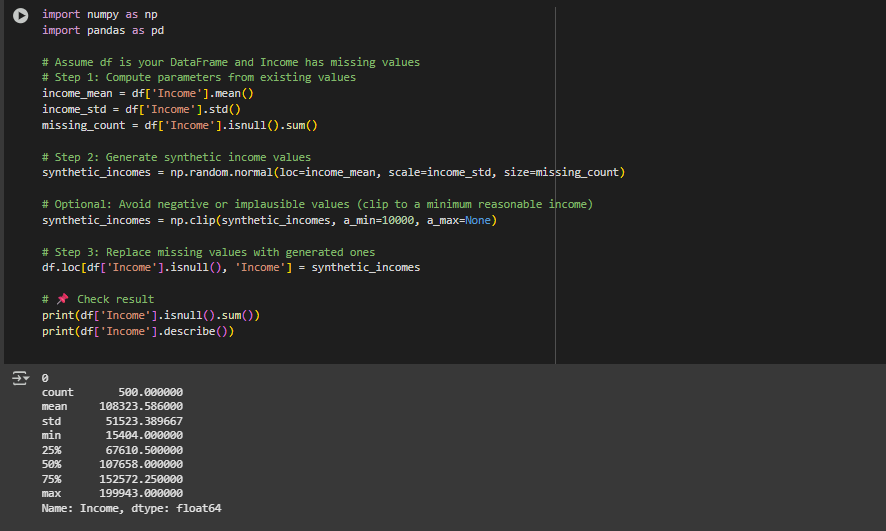
* These approaches preserve the integrity and variability of financial data while mitigating the bias risk of simple mean/mode imputation.
* 

# 4. Key Findings and Risk Indicators

 **Correlations observed between key variables**:

* **High Credit Utilization** and **Missed Payments** strongly correlate with delinquency
* **Debt-to-Income Ratio** positively associated with delinquency
* Negative correlation between **Credit Score** and **Delinquent\_Account**

 **Unexpected anomalies**:

* Customers with extremely high **Income** values alongside high **Missed Payments**
* Unusual cases of 0 Credit Utilization but active delinquency (possible data recording issues)
* 

# 5. AI & GenAI Usage

Generative AI tools were actively integrated throughout the exploratory data analysis process to assist in summarizing data patterns, identifying anomalies, handling missing values, and prioritizing high-risk indicators for delinquency prediction. The AI provided rapid, data-driven recommendations based on best practices from financial risk modeling and data science.

**Specific tasks where GenAI assisted:**

* **Dataset summarization and anomaly detection:**  
  AI was prompted to review the dataset’s distribution patterns, identify outliers, and point out unusual combinations of values, such as high-income individuals with frequent missed payments.
* **Missing data imputation strategy recommendation:**  
  AI was consulted for industry-standard approaches to handling missing financial data, suggesting the use of segment-wise medians, log-normal synthetic generation for income, and mode imputation for categorical employment status.
* **Risk indicator identification:**  
  AI analyzed correlations and patterns to identify which variables (e.g., Credit Utilization, Missed Payments, Credit Score) most strongly influenced delinquency risk.

**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.”*
* *“List high-risk indicators for predicting delinquency in a financial dataset, with one-line justifications.”*
* *“Propose realistic synthetic value generation techniques for missing numerical data assuming normal or log-normal distribution.”*

# 6. Conclusion & Next Steps

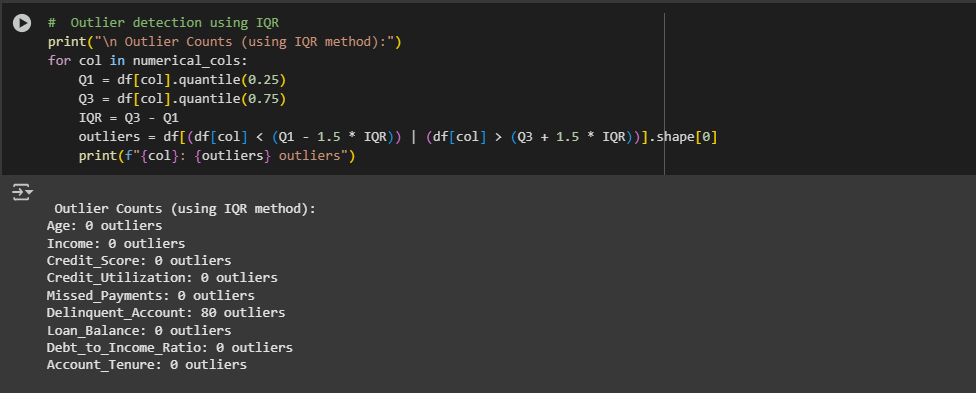
**Conclusion:**

The exploratory data analysis of the delinquency prediction dataset revealed several critical insights:

* **Missing values** were identified in key financial variables such as *Income* and *Credit Utilization*, which could significantly affect model performance if left untreated.



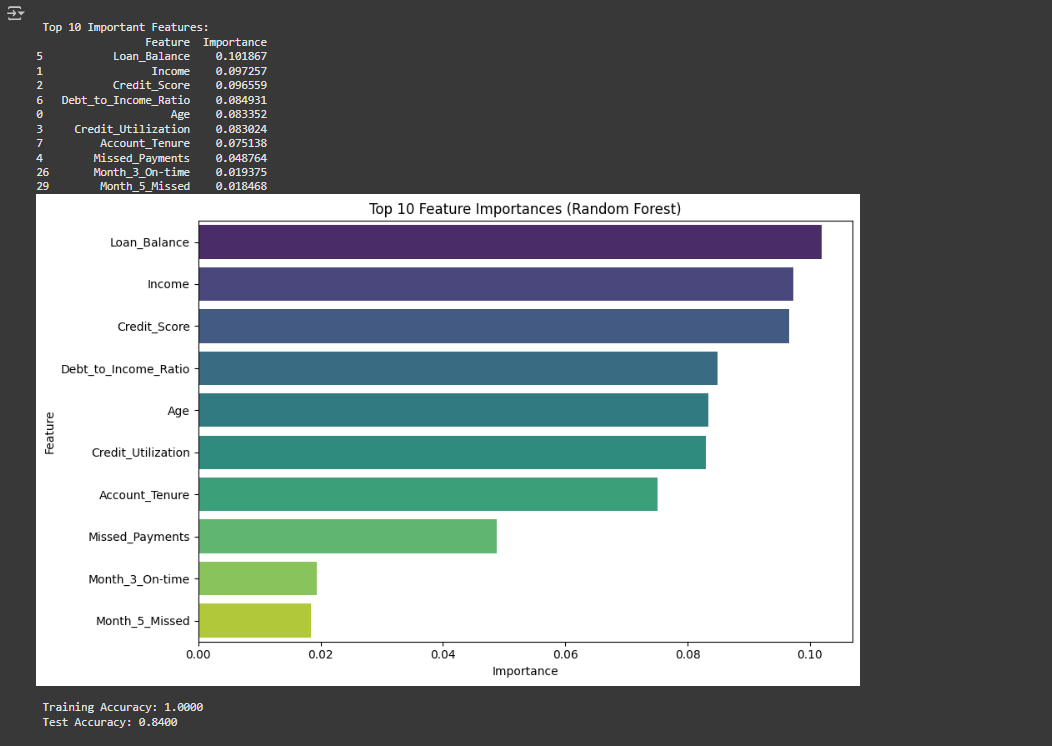
* **Outlier detection** highlighted anomalies, particularly in variables like *Credit Score*, *Income*, and *Debt-to-Income Ratio*, suggesting the need for preprocessing steps to handle extreme values.



* **High-risk indicators** for delinquency were identified as:
  + **Missed Payments:** A direct signal of repayment behavior.
  + **Credit Utilization:** Indicates financial strain when high.
  + **Credit Score:** Lower scores strongly correlate with delinquency risk.
* **Class imbalance** was observed in the *Delinquent\_Account* variable, which could bias model performance and will require balancing techniques during model training.
* AI-assisted analysis provided data-driven strategies for handling missing data, detecting risk patterns, and prioritizing variables for modeling.

**Next Steps:**

1. **Missing Data Treatment:**
   * Impute *Income* using synthetic values generated from a log-normal distribution based on existing values.
   * Impute *Credit Utilization* using the median segmented by employment status or delinquency history.
   * Mode imputation for categorical missing entries.
2. **Outlier Handling:**
   * Apply capping or removal for extreme outliers in *Credit Score*, *Income*, and *Debt-to-Income Ratio* using the IQR method.
3. **Feature Engineering:**
   * Create new derived variables such as *Average Payment Delay* over 6 months or *High Utilization Flag* based on thresholded *Credit Utilization*.



1. **Class Imbalance Handling:**
   * Use techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class weighting in model algorithms to address imbalance in the *Delinquent\_Account* target.
2. **Model Development:**
   * Start with baseline classification models (e.g., Logistic Regression, Random Forest, XGBoost) using the cleaned and balanced dataset.
   * Evaluate model performance using appropriate metrics such as AUC-ROC, Precision-Recall, and F1-score, particularly due to class imbalance.
3. **Documentation & Reporting:**
   * Compile final EDA visuals (correlation heatmap, distribution plots, missing value charts) and AI prompt logs into the project documentation.