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

**1. Introduction**

This report presents the exploratory data analysis for the credit delinquency prediction project. The goal is to understand the dataset’s structure, quality, and key feature patterns to inform downstream predictive modeling and risk assessment for Geldium Finance.

**2. Dataset Overview**

This section summarizes the dataset, including the number of records, key variables, and data types. It also highlights anomalies and inconsistencies observed during the initial review.

* Number of records: 500
* Key variables:
  + Customer\_ID: Unique identifier (categorical)
  + Age: Customer age in years (numerical)
  + Income: Annual income in USD (numerical, missing values present)
  + Credit\_Score: Creditworthiness measure (numerical)
  + Credit\_Utilization: Percentage of available credit in use (numerical)
  + Missed\_Payments: Count of missed payments in last 12 months (numerical)
  + Delinquent\_Account: Delinquency indicator target (binary)
  + Loan\_Balance: Outstanding loan amount (numerical, missing values present)
  + Debt\_to\_Income\_Ratio: Total debt as % of income (numerical)
  + Employment\_Status: Employment category (categorical, inconsistent entries)
  + Account\_Tenure: Years as customer (numerical)
  + Credit\_Card\_Type: Card tier (categorical)
  + Location: Customer city/region (categorical)
  + Month\_1 to Month\_6: Monthly payment history (categorical with values On-time, Late, Missed)
* Data types: Mix of categorical, numerical, binary, and ordinal data
* Anomalies: Some credit utilization values > 1; inconsistent Employment\_Status naming conventions.

**3. Missing Data Analysis**

Identifying and addressing missing data is critical for model accuracy. This section outlines missingness and our treatment strategy.

* Variables with missing values: Income (~39 missing), Loan\_Balance (~29 missing), Employment\_Status inconsistencies
* Missing data treatment:
  + Mean imputation for Income, Credit\_Score, and Loan\_Balance to maintain dataset completeness.
  + Standardize Employment\_Status entries before one-hot encoding.
  + Created synthetic indicators from available payment history for further robustness.
  + Applied Encodings on Categorical data and added new columns, dropped old ones.

**4. Key Findings and Risk Indicators**

This section identifies trends and correlations important for delinquency prediction.

* Correlations observed:
  + Higher Missed\_Payments and Late payments strongly associated with Delinquent\_Account = 1
  + Credit\_Utilization and Debt\_to\_Income\_Ratio positively correlated with delinquency risk
  + Older Account\_Tenure tends to be inversely related to delinquency
  + Employment\_Status and Credit\_Card\_Type categories show varied delinquency rates
* Unexpected anomalies:
  + Occasional credit utilization values slightly above 1 warrant correction or capping
  + Loan\_Balance shows high variability with extreme values needing review
  + Employment\_Status recorded inconsistently across data points.

**5. AI & GenAI Usage**

Generative AI tools aided in summarizing data, imputing missing values, and detecting patterns efficiently.

Mostly I used GenAI for my clearing my questions, doubts and for boilerplate codes.  
Following are some scenarios where I used GenAI:  
- While loading provided .xlsx, I got an unsual error, so to fix it I copy pasted the error, but at last I had to manually convert the .xlsx to .csv to fix it.

- Prompt :“Attached Dataset in csv format and its feature description , analyze it, and give me observations”\

- Prompt: “What does this cell output of df.describe() tells?  
Teach me how to make observations from mean, median , mode, quantiles etc”

- Prompt : Will machine learning model take bool values and during Standardscaler? Or I have to convert it to binary?:

- Prompt : What other visualizations I can do?  
How can I make descisions, what visualization I should do or not?

- Prompt: Give me visualization code for important visualizations for this dataset

I took these code snippets and made own changes

# 6. Conclusion & Next Steps

The dataset provides rich behavioral and financial attributes relevant for delinquency risk prediction but required careful preprocessing to handle missing and inconsistent data. Key predictive features include payment history, credit utilization, and employment status. Next steps involve:

* Implementing imputation and encoding strategies as outlined (Done already, at this point of the summary)
* Feature engineering on payment trends and categorical variables (Done already, at this point of the summary)
* Find suitable ML model (Descision Tree / Random Forest)
* Build and validate predictive model.
* Continuously revisiting EDA post-preprocessing to validate improvements and uncover deeper insights