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

# 1. Introduction

**Purpose:** The purpose of this report is to analyze the customer dataset provided by Geldium Finance to identify key risk factors for credit card delinquency.  
**Goals:** This report summarizes missing data, dataset characteristics, and recommendations for data preprocessing to support AI/ML model development.

# 2. Dataset Overview

The dataset consists of 500 records and 19 columns. It contains customer demographics, credit history, and payment behavior features relevant to predicting delinquency risks.

**Key dataset attributes:**

- Number of records: 500

- Key variables: Customer\_ID, Age, Income, Credit\_Score, Credit\_Utilization, Missed\_Payments, Delinquent\_Account, Loan\_Balance, Employment\_Status, Account\_Tenure, Credit\_Card\_Type, Location, and monthly payment history (Month\_1 to Month\_6)

- Data types: Numeric (int/float) and Categorical (object)

# 3. Missing Data Analysis

Identifying and addressing missing data is critical to ensuring model accuracy. This section outlines missing values in the dataset, the approach taken to handle them, and justifications for the chosen method.

**Key missing data findings:**

- Variables with missing values: Income (39 missing), Credit\_Score (2 missing), Loan\_Balance (29 missing)

**- Missing data treatment: Imputation**

* **Income** – Filled with **median** because income distribution is skewed, and median is robust to outliers.
* **Credit\_Score** – Filled with **mean** since only 2 values are missing and the distribution is approximately normal.
* **Loan\_Balance** – Filled with **median** as loan balances often have outliers, making median more reliable.

**Justification:**

* The percentage of missing values is less than 10% for each column.
* Imputation ensures minimal data loss and preserves dataset size, unlike deletion.

# 4. Key Findings and Risk Indicators

This section identifies trends and patterns that may indicate risk factors for delinquency. Feature relationships and statistical correlations are explored to uncover insights relevant to predictive modeling.

**Key findings:**

**- Correlations observed between key variables:**

* **Income vs. Delinquent\_Account** → **Negative correlation**  
   Customers with **higher income** are **less likely** to have delinquent accounts.
* **Credit\_Score vs. Delinquent\_Account** → **Negative correlation**  
   Customers with **higher credit scores** usually have **lower delinquency risk**.
* **Credit\_Utilization vs. Delinquent\_Account** → **Positive correlation**  
   Customers using a **higher percentage of their credit limit** are **more likely** to default.
* **Missed\_Payments vs. Delinquent\_Account** → **Strong positive correlation**  
   More missed payments strongly indicate higher delinquency risk.
* **Debt\_to\_Income\_Ratio vs. Delinquent\_Account** → **Positive correlation**  
   Customers with **higher debt compared to income** tend to have a higher risk of delinquency.
* **Key Insight:**  
   Delinquency risk increases when income and credit score are low, credit utilization and debt ratio are high, and past missed payments are more frequent.

**- Unexpected anomalies:**

* A few customers with **high income and high credit score** still have **delinquent accounts**, which is unusual and may indicate **behavioral factors** (e.g., overspending despite capacity to pay).
* Some customers with **very low credit utilization** still missed payments, which might suggest **irregular payment habits or negligence**.
* A few records may have **income values that seem too high or too low** compared to the rest of the dataset – these could be **outliers** that need further validation.
* There are instances where **customers with short account tenure (new accounts)** already have missed payments, suggesting **early-stage risk behavior**.

# 5. AI & GenAI Usage

I have used ChatGPT, free version to summarize the dataset, impute missing data, and detect patterns. As it's a free version it couldn't make more advanced data analysis. I worked with this file using pandas and dataframes. I gave df.info() and df.head() output data to ChatGPT. First, I gave it “My Role” and “My Goal” which makes any model to understand the requirements in a best possible way.

**AI prompts used:**

-’hi codey, just started working on TCS internship project given to you’- codey is the nickname given by me to ChatGPT.’

-’this is the excel file, Identify missing values in this dataset and recommend the best imputation strategy based on industry best practices.

- ‘codey, do you find any Correlations observed between key variables in the dataset...summarize the findings’

-’Unexpected anomalies: Highlight data points requiring further investigation’

-‘what is the conclusion and next steps codey? can you Summarize key findings and outline the recommended next steps based on the geldium's need for a more structured, data-driven approach to identifying at-risk customers.’

**- Unexpected anomalies:**

* A few customers with **high income and high credit score** still have **delinquent accounts**, which is unusual and may indicate **behavioral factors** (e.g., overspending despite capacity to pay).
* Some customers with **very low credit utilization** still missed payments, which might suggest **irregular payment habits or negligence**.
* A few records may have **income values that seem too high or too low** compared to the rest of the dataset – these could be **outliers** that need further validation.
* There are instances where **customers with short account tenure (new accounts)** already have missed payments, suggesting **early-stage risk behavior**.

# 6. Conclusion & Next Steps

**Conclusion:**  
 The analysis highlights that **income, credit score, credit utilization, missed payments, and debt-to-income ratio** are significant predictors of delinquency risk. Customers with **low income, low credit score, high credit utilization, and a history of missed payments** are at **higher risk of delinquency**.

A small number of **unexpected anomalies** were also found, such as **high-income customers who are still delinquent**, which indicates that behavioral factors also play a role.

**Next Steps:**

1. **Data Quality Improvement** – Validate extreme income/loan balance values and ensure consistency in missing value imputation.
2. **Feature Engineering** – Create derived features such as average monthly payment behavior or risk scores to improve prediction accuracy.
3. **Model Development** – Build a predictive model (e.g., Logistic Regression, Random Forest, or XGBoost) to classify customers into risk categories.
4. **Explainability & Fairness** – Use SHAP/LIME to explain model outputs and ensure ethical, transparent decision-making.
5. **Recommendation Framework** – Develop a **tiered intervention strategy** for at-risk customers:
   1. **High risk:** Early payment reminders, flexible payment plans
   2. **Medium risk:** Targeted financial education and incentives
   3. **Low risk:** Routine monitoring

By following these steps, Geldium Finance can move toward a **structured, data-driven approach** to proactively identify and manage **at-risk customers**, reducing credit card delinquency rates effectively.