**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

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

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# 1. Introduction:

# ---> This report presents an exploratory data analysis (EDA) of Geldium’s customer dataset. The goal is to identify missing values, data quality issues, and risk indicators that can impact delinquency prediction models for Tata iQ.

# 2. Dataset Overview:

# ****Number of records:**** Approximately 500

**Key variables:** Income, Credit\_Score, Credit\_Utilization, Missed\_Payments, Loan\_Balance, Debt\_to\_Income\_Ratio, Delinquent\_Account, and monthly Payment History (Month\_1 to Month\_6)

Data types:

- Numerical: Age, Income, Credit\_Score, etc.

- Categorical: Employment\_Status, Credit\_Card\_Type

- Binary: Delinquent\_Account (0 = No, 1 = Yes)

# Missing Data Analysis:

**Income:** 39 missing values (7.8%) → Imputed using **median**

**Credit\_Score:** 2 missing values (0.4%) → Imputed using **mean**

**Loan\_Balance:** 29 missing values (5.8%) → Imputed using **regression model** based on Income and Credit\_Score

These methods were chosen based on feature importance, distribution shape, and best practices in credit risk analytics.

# Key Findings and Risk Indicators:

# Top correlated features with delinquency:

- Income, Debt\_to\_Income\_Ratio, Credit\_Utilization (slightly positive correlation)

- Account\_Tenure (negative correlation → older accounts = less risk)

Anomalies Detected:

- Credit\_Utilization values >1.0 (e.g., 1.0258)

- Account\_Tenure = 0 in some entries

- Inconsistent Employment\_Status values: 'EMP', 'employed', etc.

# AI & GenAI Usage:

# Generative AI tools (ChatGPT, Gemini) were used for :

- Detecting missing values and anomalies

- Suggesting imputation strategies

- Highlighting possible high-risk indicators

Prompts used:

- “Summarize key patterns, outliers, and missing values in this dataset.”

- “Suggest an imputation strategy for missing income values.”

- “Identify top 3 variables likely to predict delinquency with reasoning.”

# Conclusion & Next Steps:

The dataset is now cleaned and ready for modeling. Key indicators and data quality issues were addressed.

Next steps:

- Feature engineering

- Predictive model building

- Performance evaluation using metrics like **AUC-ROC**, **Precision**, and **Recall**