Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

The purpose of this report is to conduct an Exploratory Data Analysis (EDA) on Geldium's customer dataset to support Tata iQ's analytics team in refining the company's delinquency risk model. The insights derived here will inform data cleaning, feature engineering, and modeling strategies to help identify high-risk customers and improve intervention approaches.

2. Dataset Overview

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

Key dataset attributes:

- Number of records: 500
- Key variables:
  - `Customer_ID`: Unique identifier for each customer
  - `Age`: Age of the customer
  - `Income`: Annual income
  - `Credit_Score`: Traditional credit score
  - `Credit_Utilization`: Ratio of credit used to total available credit
  - `Missed_Payments`: Count of missed payments
  - `Delinquent_Account`: Target variable (0 = No, 1 = Yes)
  - `Loan_Balance`, `Debt_to_Income_Ratio`, `Employment_Status`, `Account_Tenure`, `Credit_Card_Type`, `Location`
  - `Month_1` to `Month_6`: Monthly repayment status (e.g., "Late", "Missed", "On-time")
- Data types:
  - Numerical: Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Loan_Balance, Debt_to_Income_Ratio, Account_Tenure
  - Categorical: Employment_Status, Credit_Card_Type, Location, Monthly repayment status

Notable anomalies:

- Inconsistent entries for `Employment_Status` (e.g., “EMP”, “employed”, “Employed”)
- Several records missing `Income`, `Loan_Balance`, and `Credit_Score`
- Values of `Credit_Utilization` greater than 1, which is logically invalid
- `Debt_to_Income_Ratio` with extremely low values close to 0.

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:

📌 *Notable Missing or Inconsistent Data*

- `Income`: 39 missing columns
  Missing for **Customer_IDs** `[CUST0041, CUST0043 etc]`
- `Loan_Balance`: 29 missing columns
  Missing for [CUST0009, `CUST0024`, `CUST0026`, `CUST0029`, `CUST0046 etc]`
- `Credit_score: 2 missing columns`
  `Missing for [CUST0116, CUST0379]`
- `Employment_Status`: Inconsistent values like `"EMP"` vs `"employed"` vs `"Employed"` – these need standardization.
- `Credit_Card_Type`: Has multiple categories (e.g., *Student*, *Standard*, *Gold*, *Platinum*, *Business*) — some might be sparse.
- `Delinquent_Account`: Binary (0 or 1) — this is the target variable for modeling.

Missing data treatment:

Variable	Handling Method	Justification
`Income`	Median Imputation	Income is right-skewed; median avoids distortion
`Loan_Balance`	Regression/Median Imputation	Estimated using related financial variables
`Credit_Score`	Mean/Cluster Avg	Based on similar demographic and financial groups
`Employment_Status`	Standardization	Normalized to unify inconsistent category labels

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:

- **Missed_Payments** and **Delinquent_Account**: Strong positive relationship
- **Credit_Utilization > 0.6**: Appears frequently among delinquent accounts
- **Debt_to_Income_Ratio > 0.4**: Suggests risk of default
- **Monthly repayment patterns**: Repeated "Missed" or "Late" flags correlate with delinquency

Unexpected anomalies:

- Customers with high **Income** and **Credit_Score** still marked delinquent (suggesting behavioral risk factors)
- Customers with **Account_Tenure = 0** yet recorded monthly repayment statuses
- Values of **Credit_Utilization > 1.0**, which indicate possible data entry or extraction error

High-risk indicators:

- **Missed_Payments > 3**
- **Credit_Utilization between 0.6 and 1.0**
- **Debt_to_Income_Ratio above 0.4**
- Repetitive "Missed" or "Late" statuses in **Month_1 to Month_6.**

5. AI & GenAI Usage

Generative AI (ChatGPT) was utilized to:

- Identify missing data issues
- Suggest appropriate imputation and standardization techniques
- Detect statistical and behavioral patterns indicative of financial risk

Example AI prompts used:

- “Identify the most predictive variables of delinquency based on repayment behavior and credit attributes.”
- “Recommend imputation techniques for missing financial variables in a credit dataset.”
- “Summarize patterns of delinquency based on demographic and financial behavior.”

6. Conclusion & Next Steps

Summary:

- EDA confirms that high `Credit_Utilization`, frequent `Missed_Payments`, and high `Debt_to_Income_Ratio` are key indicators of delinquency.
- Data contains inconsistencies in employment status and missing entries in critical financial fields.
- Certain customers flagged delinquent despite favorable financial metrics, suggesting the need to include behavioral features.

Next steps:

- Finalize imputation and category standardization
- Normalize and encode features for model input
- Conduct feature importance analysis using correlation and decision trees
- Build and evaluate predictive models to classify delinquent vs. non-delinquent accounts.