Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

The purpose of this report is to conduct a comprehensive Exploratory Data Analysis (EDA) on Geldium's delinquency risk dataset. This analysis aims to assess the dataset's structure, completeness, and quality, and to uncover early indicators of credit risk. By identifying missing values, anomalies, and key patterns, the report will support Tata iQ's analytics team and Geldium's decision-makers in refining their predictive models and enhancing intervention strategies. Ultimately, the goal is to ensure that future risk assessments are built on accurate, fair, and actionable data insights.

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.

Dataset Summary: Delinquency Prediction

🔢 **Number of Records:**

- **500 records** (each representing a unique customer)

📌 **Key Variables:**

Column Name	**Description**
Customer_ID	Unique identifier for each customer
Age	Customer's age (Numerical)
Income	Annual income in USD (Numerical)
Credit_Score	Credit score rating (Numerical)
Credit_Utilization	Ratio of credit used to total available credit (Numerical)
Missed_Payments	Count of missed payments over the last 6 months (Numerical)
Delinquent_Account	Binary flag indicating delinquency status (0 = No, 1 = Yes)
Loan_Balance	Outstanding loan balance (Numerical)
Debt_to_Income_Ratio	Ratio of debt to income (Numerical)
Employment_Status	Employment category (Categorical: Employed, Unemployed, Self-

Column Name	**Description**
	employed, etc.)
Account_Tenure	Duration of account ownership in years (Numerical)
Credit_Card_Type	Type of credit card held (Categorical: Standard, Gold, Platinum, etc.)
Location	City of residence (Categorical)
Month_1 to Month_6	Monthly payment status (Categorical: On-time, Late, Missed)

🧮 **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 payment statuses
- **Binary**: Delinquent_Account

⚠️ **Initial Observations:**

- **Anomalies**:
    - Some Credit_Utilization values exceed 1.0 (e.g., >100%), which may indicate data entry errors or unusual financial behavior.
    - A few Income values are extremely low or high, suggesting potential outliers.
- **Duplicates**:
    - No duplicate Customer_ID values detected.
- **Inconsistencies**:
    - Mixed formatting in Employment_Status (e.g., “EMP”, “employed”, “Employed”) could affect grouping.
    - Some Loan_Balance entries appear to be missing or zero, which may need further validation.

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 or Inconsistent Values:**

- **Income**: A few entries are either missing or contain implausibly low values (e.g., under $1,000).
- **Loan_Balance**: Several records show zero or missing values, which may not reflect actual financial status.
- **Credit_Utilization**: Some values exceed 1.0 (i.e., >100%), suggesting either data entry errors or unusual financial behavior.
- **Employment_Status**: Inconsistent formatting (e.g., "EMP", "employed", "Employed") may affect categorical grouping.
- **Monthly Payment Columns (Month_1 to Month_6)**: Some entries are blank or contain unexpected labels.

🛠️ **Missing Data Treatment Strategy**

Variable	Treatment Method	Justification
Income	Synthetic Generation	Used realistic values based on distribution of similar age and employment groups.
Loan_Balance	Imputation (Median)	Median preserves central tendency and avoids skew from extreme values.
Credit_Utilization	Removal of outliers	Values >1.0 removed to maintain model integrity and reflect realistic usage.
Employment_Status	Standardization	Cleaned and normalized labels to ensure consistent categorical encoding.
Monthly Payments	Imputation (Mode)	Most frequent label used to fill missing entries, preserving behavioral patterns.

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**

- **Credit Utilization vs. Delinquency** Customers with credit utilization rates above 60% show a significantly higher likelihood of delinquency. This suggests that overextension of credit is a strong risk factor.

- **Missed Payments vs. Delinquent Account Status** A clear positive correlation exists: customers with 4 or more missed payments in the past 6 months are disproportionately represented in the delinquent group.
- **Debt-to-Income Ratio vs. Delinquency** Ratios above 0.4 are associated with increased delinquency risk, indicating financial strain relative to income.
- **Employment Status and Income Stability** Unemployed and self-employed individuals tend to have more missed payments and higher credit utilization, suggesting income volatility may contribute to risk.
- **Account Tenure** Accounts less than 2 years old show higher delinquency rates, possibly due to limited credit history or immature financial behavior.

⚠️ Unexpected Anomalies

- **Credit Utilization > 1.0** Several records show utilization rates exceeding 100%, which is not feasible under standard credit models and may indicate data entry errors or misreported balances.
- **Zero or Missing Loan Balances** Some customers with missed payments have zero loan balances, which contradicts expected financial behavior and may require validation.
- **Inconsistent Employment Labels** Variants like “EMP”, “employed”, and “Employed” appear across records, potentially affecting categorical analysis unless standardized.
- **Delinquent Accounts with Low Missed Payments** A few customers are flagged as delinquent despite having only 1–2 missed payments, suggesting either early-stage risk or misclassification.

5. AI & GenAI Usage

🔍 Dataset Summarization & Structure

- *“Provide a summary of the dataset, including column types, value ranges, and missing data.”*
- *“Highlight any categorical variables with inconsistent formatting or unexpected values.”*
- *“Detect duplicate records or customer IDs and assess their impact on analysis.”*

🧹 Missing Data & Imputation

- *“Identify all columns with missing values and recommend appropriate imputation methods.”*
- *“Generate synthetic income values for missing entries using distribution patterns from similar customers.”*
- *“Evaluate whether missing loan balance values can be imputed or should be excluded.”*
- *“Classify missing data as MCAR, MAR, or MNAR and suggest handling strategies.”*

📊 Pattern Detection & Risk Indicators

- *“Analyze the relationship between credit utilization and delinquency risk.”*
- *“Identify top predictors of delinquency based on correlation and behavioral trends.”*
- *“Detect customers with high missed payments but low loan balances—flag for review.”*
- *“Cluster customers based on payment behavior and highlight high-risk segments.”*

📈 Behavioral Trends & Feature Relationships

- *“Summarize trends in monthly payment status and identify customers with escalating risk.”*
- *“Compare delinquency rates across employment types and income brackets.”*
- *“Analyze how account tenure influences delinquency likelihood.”*

⚖️ Fairness & Compliance Checks

- *“Assess whether synthetic data generation introduces bias in delinquency predictions.”*
- *“Evaluate fairness across demographic segments (e.g., age, location) in model inputs.”*

6. Conclusion & Next Steps

🔍 Summary of Key Findings

- The dataset contains **500 customer records** with a mix of numerical and categorical variables relevant to credit risk.
- **Missing data** was identified in key fields such as Income, Loan_Balance, and Monthly Payment Status. These were addressed using a combination of **median imputation**, **synthetic data generation**, and **categorical standardization**.
- Strong correlations were observed between **credit utilization**, **missed payments**, and **delinquency status**, confirming their importance as predictive features.
- **Unexpected anomalies**—such as utilization rates over 100% and inconsistent employment labels—were flagged for further review.
- GenAI tools were effectively used to summarize the dataset, suggest imputation strategies, and surface behavioral patterns.

📌 Recommended Next Steps

1. **Validate Synthetic Data** Ensure that generated income and behavioral values align with real-world distributions to avoid bias in modeling.
2. **Feature Engineering** Create derived variables such as average monthly payment status or payment trend scores to enhance predictive power.

3. **Model Readiness Check** Confirm that all variables are clean, encoded, and normalized before feeding into machine learning models.

4. **Bias & Fairness Audit** Conduct fairness checks across demographic groups (e.g., age, employment status) to ensure ethical risk predictions.

5. **Begin Predictive Modeling** Use the cleaned dataset to train and evaluate models for delinquency prediction, prioritizing high-risk indicators identified during EDA.