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

# Introduction

The purpose of this practice is to understand the giving data at hand and find the missing variables using ai and gen ai tools .

Use proper prompts to help the gen ai tool and find the missing variables , anomalies and risk factors.

Assessment for key variable and finding a correlation between them will help is further understating the data at hand and derive various methods to get unbiased data and help the research team in getting to conclusions as to factors might lead to early detection of delinquency and reach out the customers and assess the situation.

# 2. Dataset Overview

Key dataset attributes:

- Number of records: 500

- Key variables: 1. Missed payments (numerical ) ranges from 0 to 6 ,

2. Credit Score (Numerical) customer with low credit score are likely to be delinquent , while high scores are likely to be non delinquent.

3. Credit Utilization(numerical) :- this is the ratio of credit used to credit utilization as higher credit utilization indicates financial strain as customer are using a large portion of there credit increasing delinquency.

4. Debt to income ratio :- a lower income will increase a chance of debt burden and which may lead to payment difficulties and delinquency.

5. Payment History :- here we look into the month 1 and 6 where we typically focus on the recent payment history eg month 5 and month 6 and see as the payment behavior is late or on time and decided if the as to which customer will fall in the delinquency bracket and non delinquency bracket.

6. Employment status :- here people who are employed are likely to fall in the delinquency bracket as they have a steady income where as customer who are retired and unemployed who have fixed or very minimum income are most likely to be inconsistent in maintaining the credit score and eventually leading to missed payment which is a early prediction to high delinquency.

7.:- Loan balance as high the numbers are of the customer as higher loan balance will increase a financial strain raising delinquency risk and lower balance will decrease the chances non delinquent customer.

Variables of Secondary Importance

Age , income, account tenure , credit card type and location

- Data types: the following data set has Numerical column, payment history is between month 0 and 6 and categorical columns .

# 3. Missing Data Analysis

Key missing data findings:

1 Income (Numerical) Several records are missing the income category, we will use the =COUNTBLANK (C2:C502) formulae to find out the number of empty variable input for the given data set , There are 40 missing records on the variable or column

2. credit score (Numerical):- here we have 3 missing records for this variable

3. loan balance: - We have 30 missing values on this column

No missing values observed in Age, Credit Utilization, Missed Payments, Delinquent Account, Debt to Income Ratio, Account Tenure, Employment Status, Credit Card Type, Location, or Month\_1 to Month\_6 based on the sample.

- Missing data treatment: [Deletion, Imputation, Synthetic Data, etc.]

1 Income :- we will proceed with the missing values for income with imputation by median across the data set . the median is robust to skewness and outliers as it will help preserving the data without introducing any biasness

We will not delete the records as we will be loosing out significant data and the deciding factors will be missed out

Imputation with mean is disregarded as the outliers with high income with not give accurate data.

While mode is unsuitable for continuous data .

2. Loan Balance:- this as well we use imputation by median as it is highly skewed with records having 0 balance to 99005 hence the median is appropriate for skewed distribution

Mean imputation will be disregarded due the skewness for the data and synthetic data generation adds complexity without benefits for this level of missingness

3. Credit score : Imputation with median of the credit score will be used to fill the missing values . we can also the delete the records as that accounts to 1 % of the data set which willhave minimal impact on the data set , mean imputation can be considered but the meidan is preferred as there is an outlier present

# 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:

1. missed payments and delinquent account .As suggested a positive correlation is expected between missed payments and delinquent account. Customer with higher missed payment will fall in the delinquent account

This is a stringer variable related to delinquency as customer who have missed more than 4 missed payments as prominent to a subject for delinquency.

2. credit score and delinquent account . customer with less credit score will have increased delinquency rate

3. credit utilization and delinquent account :- customer with high credit utilization will have high delinquency rate which is a risk indicator

4. Employment status and delinquent account :- unemployed status will also have high delinquency rate compared to employed or self employed . the risk indicator for unemployed or retired have higher delinquency rates as they limited or fixed income.

5. loan balance and delinquent account :- this relation is less clear as they are dependent on other factors as well such as income to debt ratio, loan balance.

- Unexpected anomalies:

In the data set we see credit utilization is higher for a specific customer which is more than 1 and don’t fall into the delinquency bracket . this could be due to wrong entry or instead its in different data type such as percentage.

Missing key data in the variables :- as missingness may not be random for example a customer who is unemployed and does not fall into delinquent account such cases will be masked using imputation with median.

6. inconsistent employee status : as observed in the column the values recorded for employed are as emp, EMP or Employed hence a standardization is required for miscounting which will directly effect delinquency rate.

Finally low debt to income ratio is also considered as an anomaly hence it needs to be checked with loan balance and conform if it’s a data entry issue or it reflect true low debt. Also customer fall in the delinquent bracket even though they have low debt.

# 5. AI & GenAI Usage

Data Overview and Initial Exploration, Data Cleaning and Preprocessing, Univariate Analysis, Correlation Analysis, Key Risk Indicators and Patterns and Missing Data Analysis, Anomaly Detection

# 6. Conclusion & Next Steps

Conclusion: Missed Payments, Credit Score, Credit Utilization, Payment History, Employment Status drive delinquency. Missing data (~4% Income, ~3.2% Loan Balance) imputed with median. Anomalies (Credit Utilization >1) capped.

Next Steps: Investigate anomalies, test missingness patterns, engineer features (non-on-time payments), build predictive model (logistic regression/random forest), validate with statistical tests.