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

1. ****Introduction****

This report summarizes the exploratory data analysis (EDA) conducted on the Delinquency\_prediction\_dataset.csv. The primary purpose is to understand the dataset's structure, identify key variables, assess data quality (including missing values), and uncover significant patterns and risk indicators related to customer delinquency. The insights gained will serve as a foundation for subsequent predictive modeling efforts.

1. **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**:** The dataset contains 500 records.
* **Customer\_ID:** Unique identifier for each customer.
* **Age:** Age of the customer (Numerical).
* **Income**: Customer's annual income (Numerical).
* **Credit\_Score**: Customer's credit score (Numerical).
* **Credit\_Utilization:** Ratio of credit used to available credit (Numerical).
* **Missed\_Payments:** Number of missed payments by the customer (Numerical).
* **Delinquent\_Account:** Binary flag indicating if the account is delinquent (0 = No, 1 = Yes) (Categorical/Target).
* **Loan\_Balance**: Outstanding loan balance (Numerical).
* **Debt\_to\_Income\_Ratio:** Ratio of total debt to total income (Numerical).
* **Employment\_Status:** Customer's employment status (Categorical).
* **Account\_Tenure:** Duration of the account in months (Numerical).
* **Credit\_Card\_Type:** Type of credit card held by the customer (Categorical).
* **Location:** Geographic location of the customer (Categorical).
* **Month\_1 - Month\_6:** Payment status for each of the last six months ('On-time', 'Late', 'Missed') (Categorical).
* **Data types:** The dataset comprises both numerical (e.g., Age, Income, Credit\_Score, Loan\_Balance) and categorical variables (e.g., Delinquent\_Account, Employment\_Status, Credit\_Card\_Type, Location, Month\_X).
* **Anomalies/Inconsistencies:** Missing values were identified in several key columns. No other significant anomalies or duplicates were explicitly detected during this initial review.

1. **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.

* **Variables with missing values:**
  + Income: 39 missing values.
  + Credit\_Score: 2 missing values.
  + Loan\_Balance: 29 missing values.
* **Missing data treatment:**
  + For the purpose of initial correlation analysis, the missing values in Income were imputed using the **median** of the column to ensure calculations could proceed.
  + For a comprehensive approach and based on industry best practices for numerical financial data, the following imputation strategies are recommended:
    - Credit\_Score **(2 missing values):** **Median Imputation** is suitable due to the very low number of missing values and its robustness to potential outliers.
    - Income **(39 missing values) and** Loan\_Balance **(29 missing values):** **Multiple Imputation by Chained Equations (MICE)** is the preferred strategy. MICE accounts for the uncertainty in imputed values and preserves relationships between variables, which is crucial for complex financial data that may have skewed distributions or interdependencies. Alternatively, **Regression Imputation** or **K-Nearest Neighbors (KNN) Imputation** could be considered as robust alternatives to simple imputation.

1. **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.

* **Trends in Late Payments:** The number of 'Late' payments remained relatively consistent across the six observed months:
  + Month 1: 159
  + Month 2: 173
  + Month 3: 169
  + Month 4: 181
  + Month 5: 151
  + Month 6: 172 This indicates that a significant portion of customers consistently exhibit late payment behavior, with Month 4 showing the highest count of late payments.
* **Correlations observed between key variables:**
  + **Income and Delinquency Risk:** A very weak positive correlation of **0.044** was observed between customer income (after median imputation) and delinquency risk. This suggests that income alone is not a strong predictor of delinquency in this dataset.
  + **Missed Payments and Delinquency Risk:** A very weak negative correlation of **-0.026** was observed between the number of missed payments and delinquency risk. This counter-intuitive finding suggests that the Delinquent\_Account flag might represent a specific or severe form of delinquency not solely tied to the number of missed payments, or implies complex non-linear relationships.
* **Top 3 Risk Factors Associated with Delinquency:** The analysis identified the following as the top risk factors, based on their higher observed delinquency rates:
  1. **Credit Card Type: Business Card Holders:** Accounts associated with 'Business' credit cards showed the highest delinquency rate at **21.30%**, indicating a higher risk.
  2. **Location: Los Angeles:** Customers residing in 'Los Angeles' had a notably higher delinquency rate of **19.63%**, suggesting regional economic or demographic influences on delinquency.
  3. **Employment Status: Unemployed:** As expected, customers with an 'Unemployed' status exhibited a higher delinquency rate of **19.35%**, highlighting the significant impact of employment on financial stability.
* **Unexpected anomalies:**
  + The very weak correlation between Missed\_Payments and Delinquent\_Account, and the slightly higher average Income and Credit\_Score for delinquent accounts compared to non-delinquent ones, were unexpected. These suggest that the definition or criteria for 'Delinquent\_Account' might be nuanced, or that other interacting factors are at play, warranting further investigation into these relationships.

1. **AI & GenAI Usage**

Generative AI tools were used to summarize the dataset, impute missing data, and detect patterns. This section documents AI-generated insights and the prompts used to obtain results.

* **AI-generated insights:**
  + Summarized key patterns in various columns, including numerical distributions and categorical frequencies.
  + Identified specific columns with missing data and quantified the extent of missingness.
  + Recommended advanced imputation strategies (e.g., MICE) for missing numerical data based on industry best practices.
  + Analyzed the correlation between Income and Delinquent\_Account.
  + Identified trends in monthly late payments.
  + Identified and quantified the top 3 risk factors associated with delinquency.
* **Example AI prompts used:**
  + 'Analyze this dataset and provide a summary of key columns, including common patterns and missing values.'
  + 'Identify missing values in this dataset and recommend the best imputation strategy based on industry best practices.'
  + 'Analyze the correlation between customer income and delinquency risk, summarizing key findings in simple terms.'
  + 'Analyze trends in late payments and identify the top 3 risk factors associated with delinquency.'

1. **Conclusion & Next Steps**

This EDA revealed critical insights into the dataset and potential risk factors for delinquency. While Income shows a very weak direct correlation with delinquency, other factors like Credit\_Card\_Type (specifically 'Business' cards), Location ('Los Angeles'), and Employment\_Status ('Unemployed') emerged as stronger indicators of higher delinquency risk. The consistent pattern of late payments across months also highlights an ongoing area of concern.

**Recommended Next Steps:**

* **Comprehensive Missing Data Imputation:** Implement the recommended MICE imputation strategy for Income and Loan\_Balance to create a more robust and complete dataset for modeling.
* **Feature Engineering:** Explore creating new features from existing data, such as aggregated payment behavior over the six months (e.g., total late/missed payments, consecutive late/missed payments) which might provide stronger predictive power than individual monthly flags.
* **Further Risk Factor Investigation:** Conduct deeper dives into the identified top risk factors to understand underlying reasons for increased delinquency in those segments. This might involve statistical tests (e.g., chi-square for categorical associations) and visualization.
* **Model Building and Evaluation:** Proceed with developing predictive models (e.g., logistic regression, decision trees, gradient boosting) to forecast delinquency, incorporating the insights and cleaned data from this EDA.
* **Anomaly Investigation:** Further investigate the unexpected weak correlation findings to ensure a complete understanding of the Delinquent\_Account definition and its relationship with other variables.