Exploratory Data Analysis Report

Initial Data Quality Observations

Notable missing or inconsistent data:

- **Missing Data:** The dataset has missing values in `Income` (7.8%), `Loan_Balance` (5.8%), and `Credit_Score` (0.4%). While the missing values for `Credit_Score` are minimal, `Income` and `Loan_Balance` will require a thoughtful imputation strategy.
- **Inconsistent Data:** The `Employment_Status` column had inconsistent entries (`EMP`, `Self-employed`, `Unemployed`) which have been standardized for consistency.

Key anomalies:

- Contrary to expectation, the average `Credit_Score` for delinquent accounts (591) is slightly higher than for non-delinquent accounts (575). This suggests that `Credit_Score` alone is not a straightforward predictor of delinquency.
- The `Credit_Utilization` and `Debt_to_Income_Ratio` for delinquent accounts are only marginally higher than for non-delinquent accounts, indicating a weak direct correlation.

Early indicators of delinquency risk:

- **Credit Utilization:** A higher ratio of credit utilization may be a weak but relevant indicator of financial stress.
- **Payment History:** The month-by-month payment history is a more direct and reliable indicator of a customer's payment behavior than the summary `Missed_Payments` count.

Based on the initial review, the dataset exhibits moderate data quality issues, primarily missing values in the Income and Loan_Balance fields. A notable anomaly is the weak correlation between traditional risk indicators like Credit_Score and actual delinquency, with delinquent accounts having a slightly higher average credit score. This suggests that a simple predictive model based solely on these variables may not be effective. The monthly payment history columns, however, appear to be a more direct and reliable source for identifying delinquency risk.

Missing Data and Data Quality Issues

Issue	Handling Method	Justification
Missing values in `Credit_Score` (0.4%)	Median Imputation	The percentage of missing data is very low, making a simple imp
Missing values in `Income` (7.8%)	Median Imputation	The median is a robust measure that is not sensitive to outliers,
Missing values in `Loan_Balance` (5.8%)	Median Imputation	Similar to Income, using the median for imputation is a simple ar

High-Risk Indicators and Insights

- **Credit Utilization:** A high credit utilization ratio indicates that a customer is using a large portion of their available credit, which is a key sign of potential financial distress and a common precursor to missed payments.

- **Debt-to-Income Ratio:** A higher DTI suggests a customer's debt obligations are significant relative to their income. This reduces their financial flexibility and increases the likelihood of them being unable to meet payment deadlines.
- **Individual Payment History:** The presence of 'Late' or 'Missed' statuses in the `Month_1` through `Month_6` columns are the most direct indicators of payment issues and are crucial for predicting future delinquency. The pattern of these payments (e.g., an increasing frequency of missed payments) can be a powerful predictive signal.
- **Unexpected Finding:** The unexpected positive correlation between average Credit Score and delinquency requires further investigation. This suggests that simple credit score-based models might be misleading, and a more sophisticated model that considers other factors is likely necessary.