Exploratory Data Analysis (EDA) Summary Report
Project: Geldium Credit Card Delinquency Risk Prediction
Prepared for: Tata iQ Analytics Team
Prepared by: Deep
Date: 17 JULY 2025

1. Key Patterns and Anomalies Detected

- Missing Data:
  - payment_history: ~8% missing values. Critical for modeling delinquency.
  - income: ~12% missing. Possibly MNAR (not reported by high-risk individuals).
  - credit_utilization: Outliers above 100%, indicating over-limit usage.
  - employment_status: Formatting inconsistencies (e.g., "Unemployed", "unemp").
- Outliers and Anomalies:
  - age: Contains entries < 18 and > 85, which are likely invalid.
  - debt_to_income_ratio: A few entries above 2 (200%)—implausible and likely erroneous.
  - recent_credit_activity: Unusually high activity in some records—potential duplicates or system error.
- Data Type/Format Issues:
  - Categorical data like employment_status and payment_history needs standardization.

2. Summary of Missing Values and Handling Strategy

| Column | Handling Method | Justification |
|---|---|---|
| payment_history | Imputation (mode) | Crucial for delinquency; mode preserves distribution for categorical data. |
| income | Imputation (median) | Median is robust to outliers and preserves income skew. |
| credit_utilization | Remove extreme outliers | Values >100% are illogical; filter at 95th percentile. |
| employment_status | Standardize categories | Normalize spelling/ capitalization for consistency. |

3. Risk Indicators That May Impact Delinquency Predictions

- High Credit Utilization (>70%)
  - Strongly correlates with delinquency_flag = 1; indicates financial stress.
- Low or Missing Income
  - Often seen alongside high credit usage or late payments.
- Recent Surge in Credit Activity
  - New account openings or inquiries linked with higher delinquency rates.
- Irregular Employment History
  - Unemployed or inconsistent employment status correlates with repayment issues.
- Age Extremes

    - Very young (<21) and very old (>70) users may have different risk profiles.

4. Recommendations
   - Standardize all categorical features and handle missing values based on domain-informed imputation.
   - Remove extreme outliers or cap them using domain knowledge (e.g., 99th percentile).
   - Consider synthetic data generation for sensitive or missing fields (e.g., income) but validate distribution.
   - Ensure fairness in model training by monitoring bias with respect to sensitive attributes (e.g., age, income).
5. Next Steps
   - Proceed with feature engineering based on risk indicators.
   - Validate data quality post-cleaning.
   - Begin model selection and training phase.

End of Report