Talentlabs - MyMahir CADA programme

Mohd Azri bin Mohd Ridzuan Nyanged

loan risk analysis

Exploratory Data Analysis

Business understanding & overview

Background

The bank seeks to improve credit risk management by identifying early indicators of loan repayment difficulties. Misclassification creates two key risks:

- Approving high-risk clients → defaults and financial losses
- Rejecting creditworthy clients → missed revenue opportunities

Objective

Uncover patterns distinguishing defaulters from successful repayers by analyzing current applications and historical loan behavior.

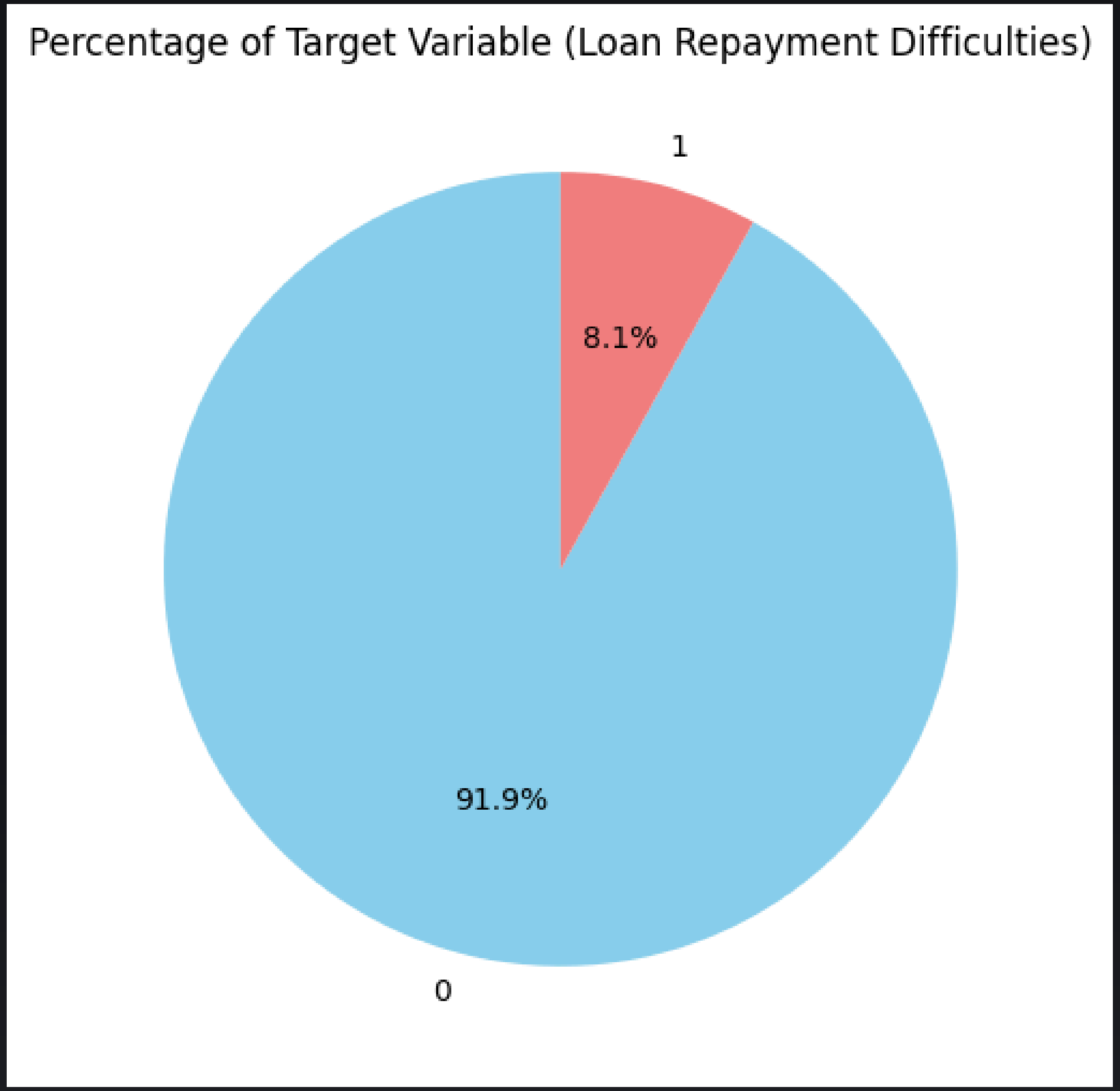
Analysis focus

- Demographic and financial attributes linked to default risk
- Financial stress indicators (loan-to-income, annuity-to-income ratios)
- Historical application patterns as behavioral risk predictors

Understanding the data

- Highly imbalanced data; almost 23:2 ratio.
- Most loans were paid on-time
- Imbalance is expected in lending, however, the minority defaulter group is high-impact and crucial for risk mitigation.

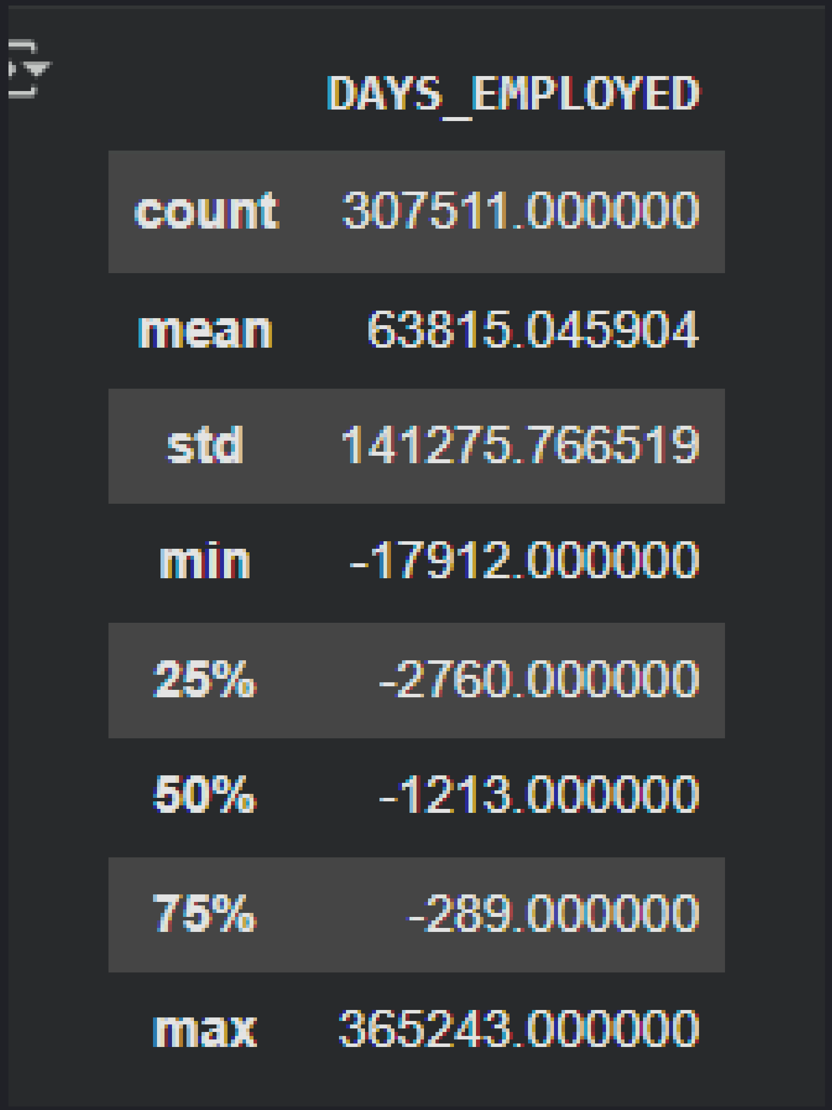
Target variable



Data cleaning

- Removed irrelevant or redundant columns to focus on meaningful features.
- Dropped normalized dwelling variables (too many nulls, low interpretability).
- Dropped document flags (binary indicators, low correlation to risk).
- Excluded masked categorical columns treated as numeric.
- Goal: retain only financial, demographic, and behavioral drivers of default.

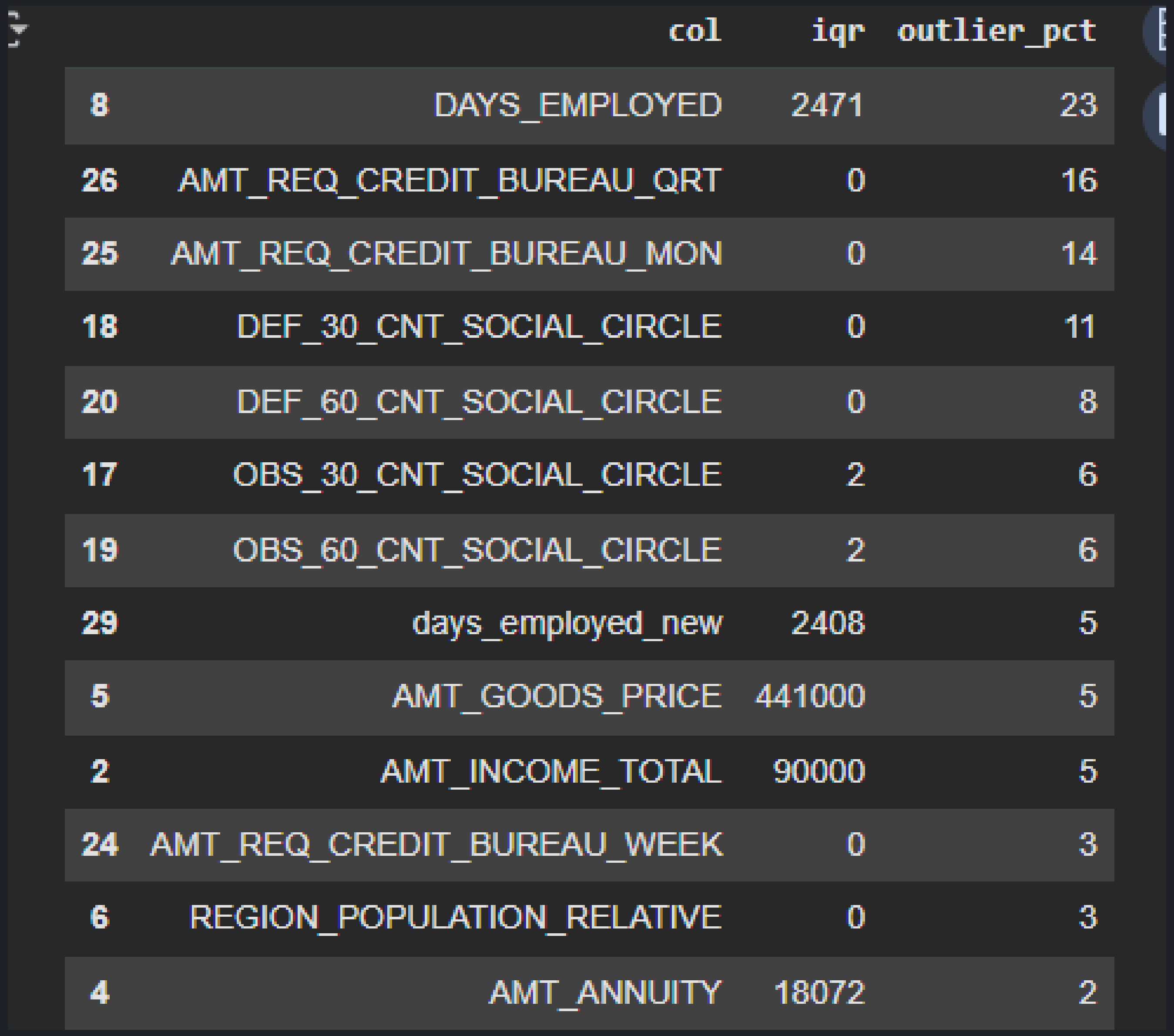
Data cleaning



	DAYS_EMPLOYED
count	307511.000000
mean	63815.045904
std	141275.766519
min	-17912.000000
25%	-2760.000000
50%	-1213.000000
75%	-289.000000
max	365243.000000

- For columns DAYS_EMPLOYED and DAYS_BIRTH, negative values are recorded since it was counted backwards from the day of application.
- This has been reformatted to absolute values to ease analysis.
- In DAYS_EMPLOYED, the max day is 365243 days which roughly equates to 1000 years. This number is a placeholder for unemployed clients, and has been replaced with ‘NaN’.

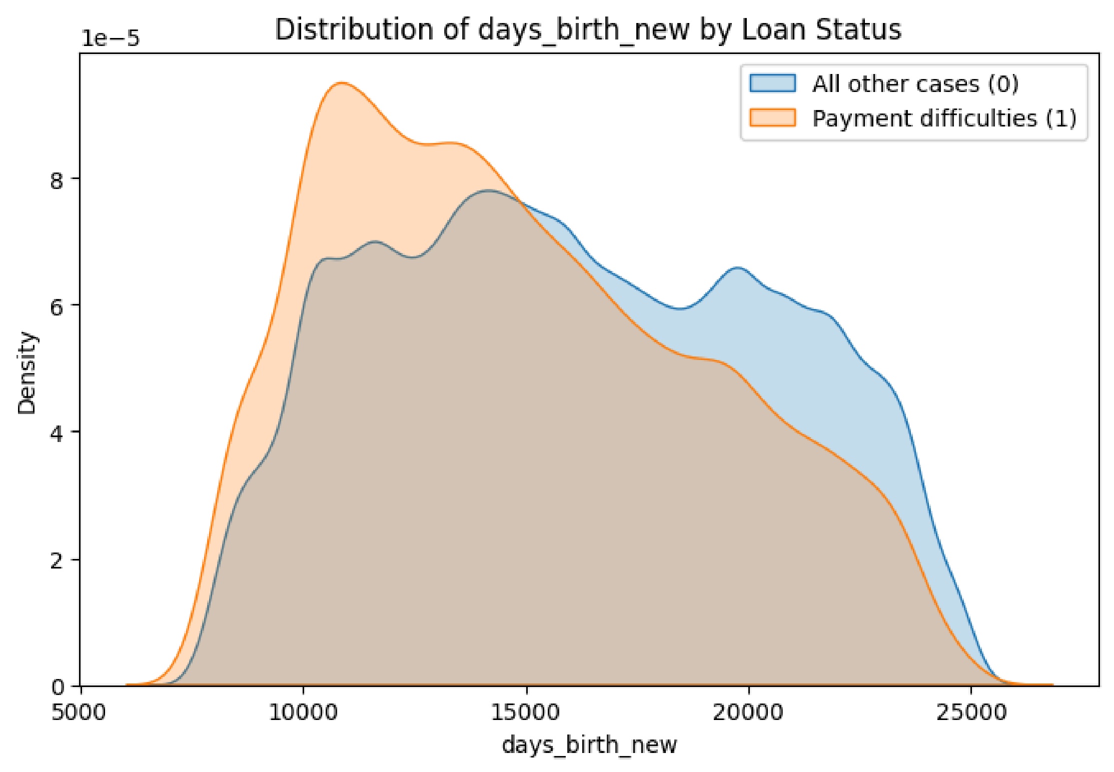
Outlier analysis

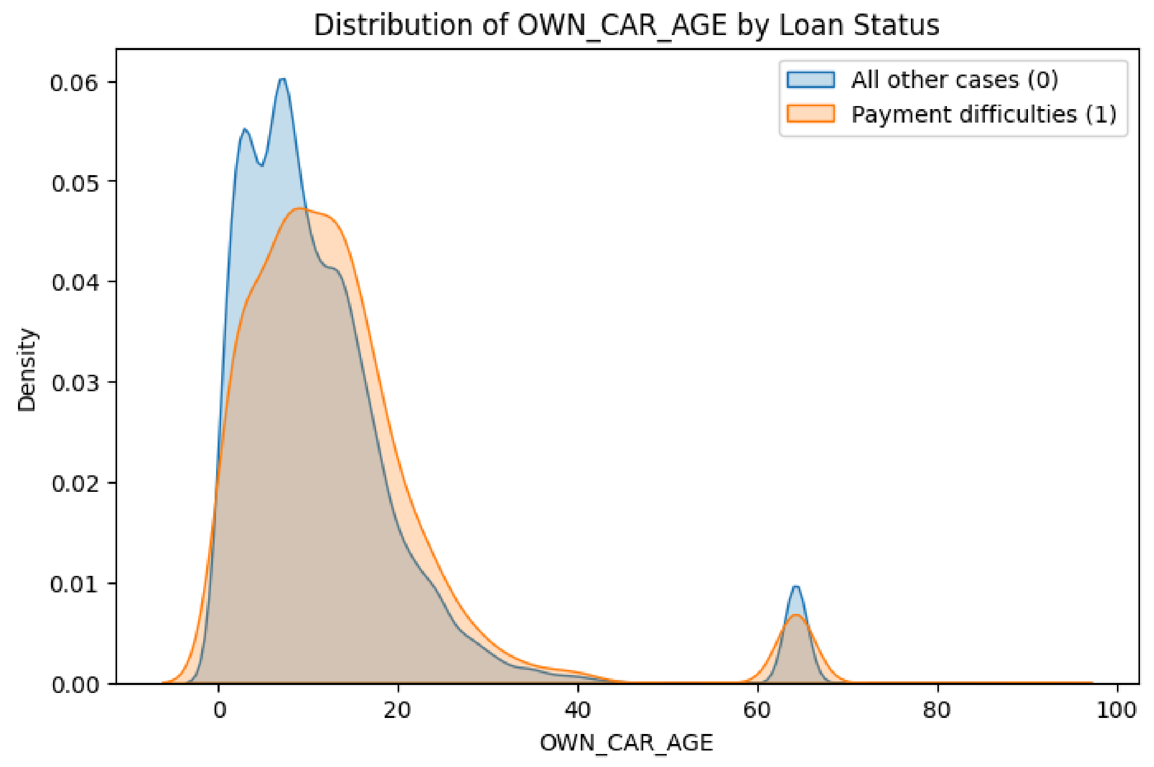


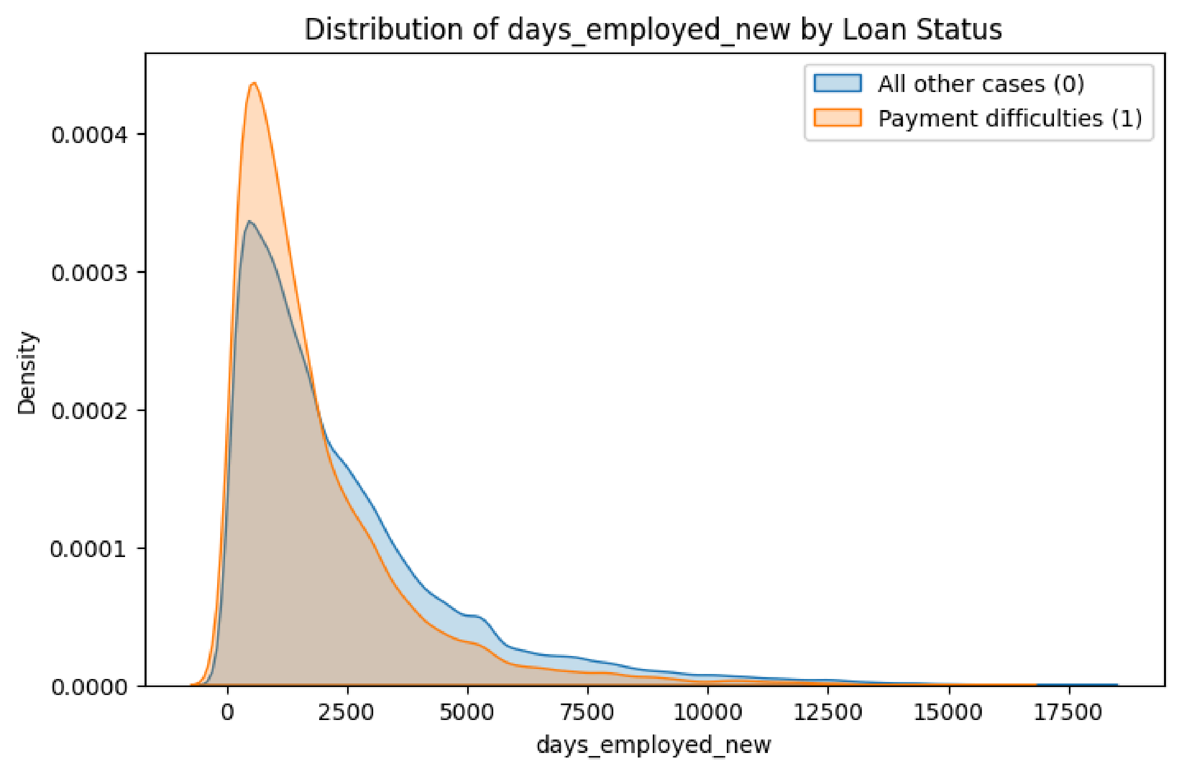
	col	iqr	outlier_pct
8	DAYS_EMPLOYED	2471	23
26	AMT_REQ_CREDIT_BUREAU_QRT	0	16
25	AMT_REQ_CREDIT_BUREAU_MON	0	14
18	DEF_30_CNT_SOCIAL_CIRCLE	0	11
20	DEF_60_CNT_SOCIAL_CIRCLE	0	8
17	OBS_30_CNT_SOCIAL_CIRCLE	2	6
19	OBS_60_CNT_SOCIAL_CIRCLE	2	6
29	days_employed_new	2408	5
5	AMT_GOODS_PRICE	441000	5
2	AMT_INCOME_TOTAL	90000	5
24	AMT_REQ_CREDIT_BUREAU_WEEK	0	3
6	REGION_POPULATION_RELATIVE	0	3
4	AMT_ANNUITY	18072	2

- Used IQR method to identify extreme values in numerical features.
- Outliers were not removed — retained for transparency and full distribution analysis.
- Some variables (e.g., region ratings) were excluded from outlier checks since they are categorical values represented numerically (1–3).
- Purpose: to understand data spread and irregularities, not to clean or impute at this stage.

Distribution analysis


Distribution of days_birth_new by Loan Status

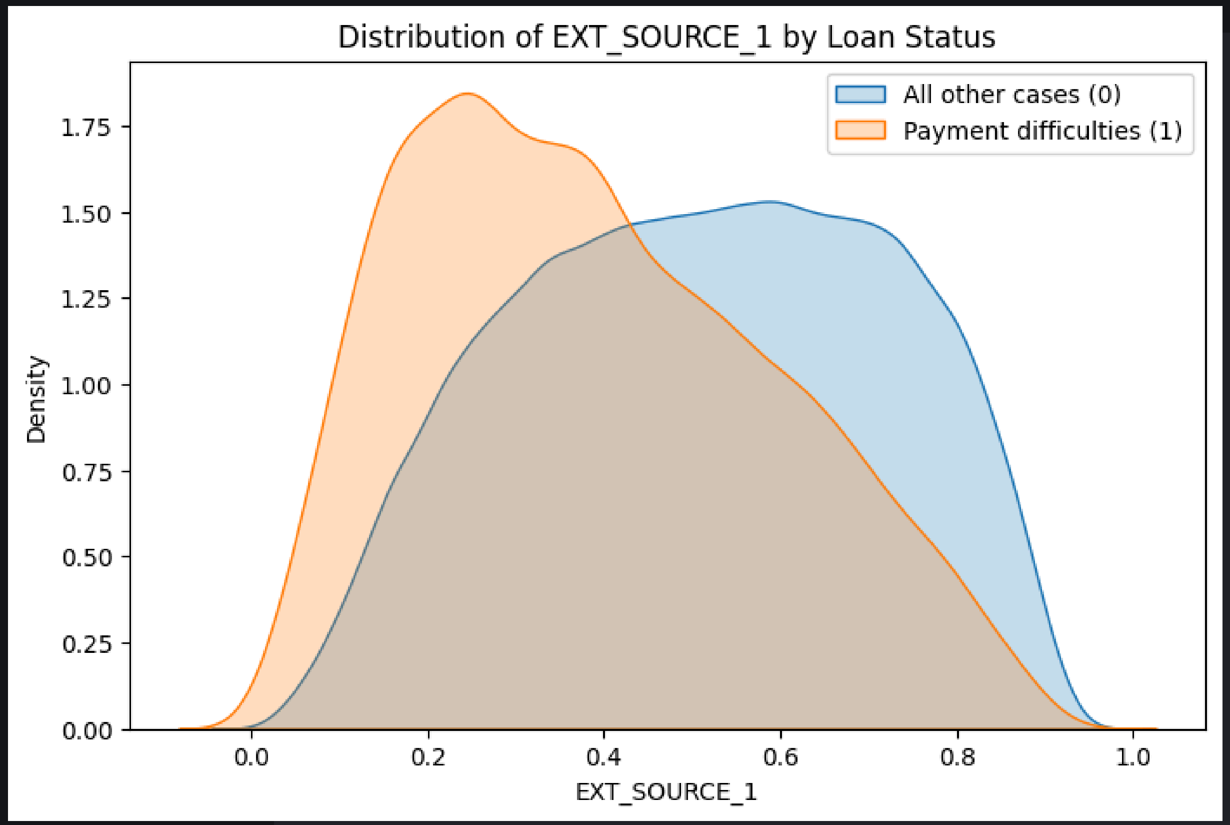

Distribution of OWN_CAR_AGE by Loan Status


Distribution of days_employed_new by Loan Status

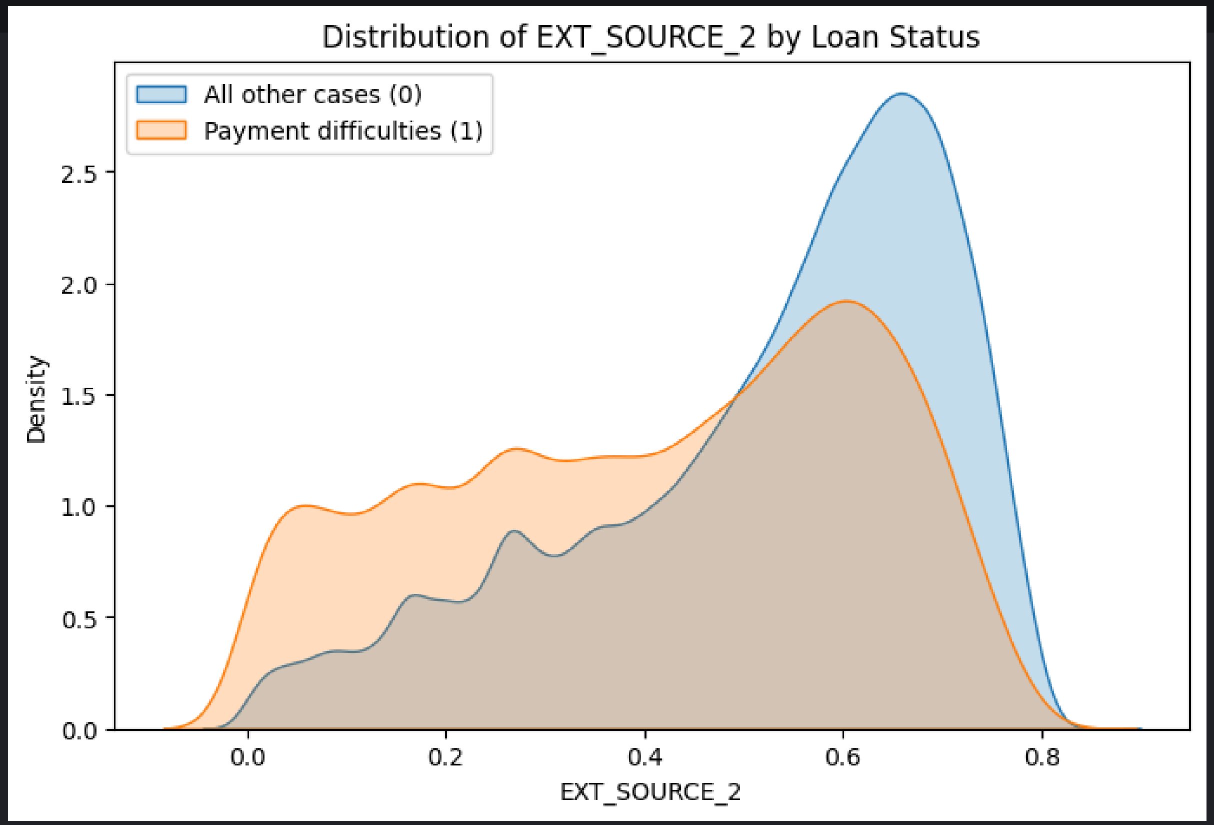
Finding:

- Employment Duration: Shorter tenure (unemployed or early-career) clients show higher default rates → job instability increases repayment risk.
- Age: Younger clients default more often → financial stability improves with age.
- Car Age: Newer cars linked to non-defaulters; defaults rise for cars >10 years old → older car ownership may signal weaker finances.

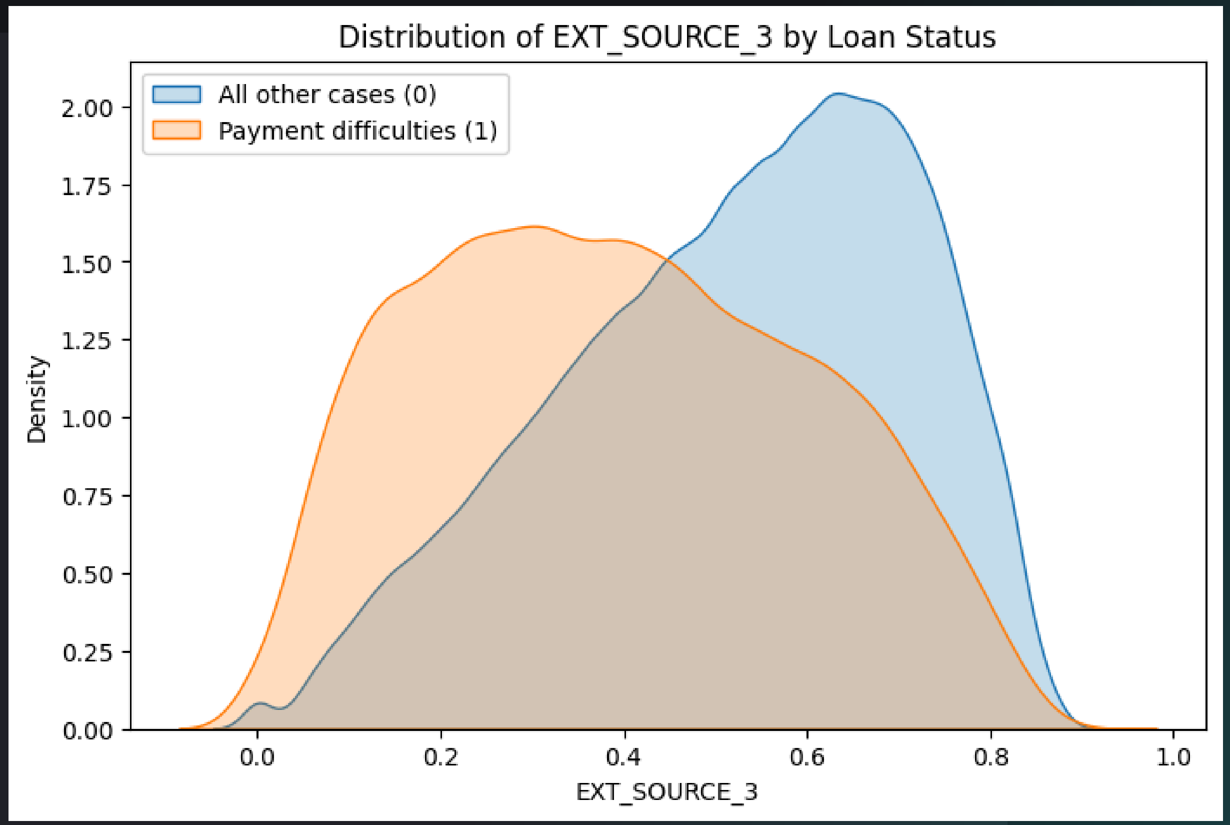
Distribution analysis



Distribution of EXT_SOURCE_1 by Loan Status



Distribution of EXT_SOURCE_2 by Loan Status

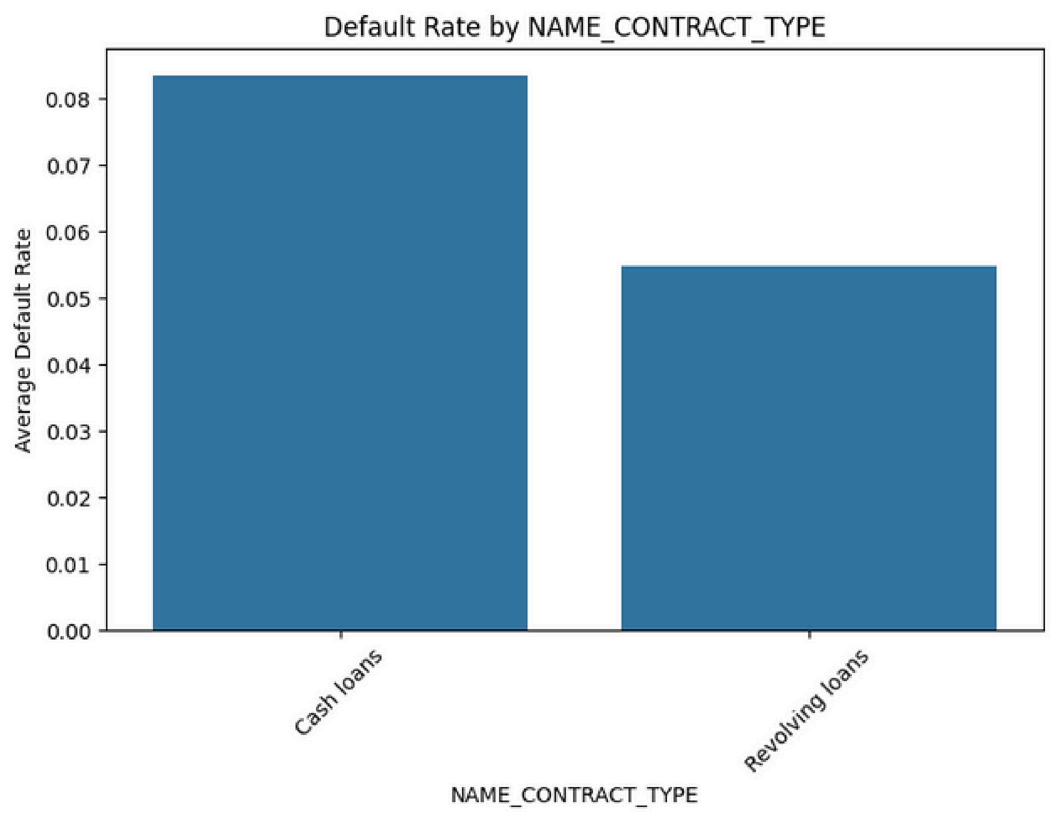


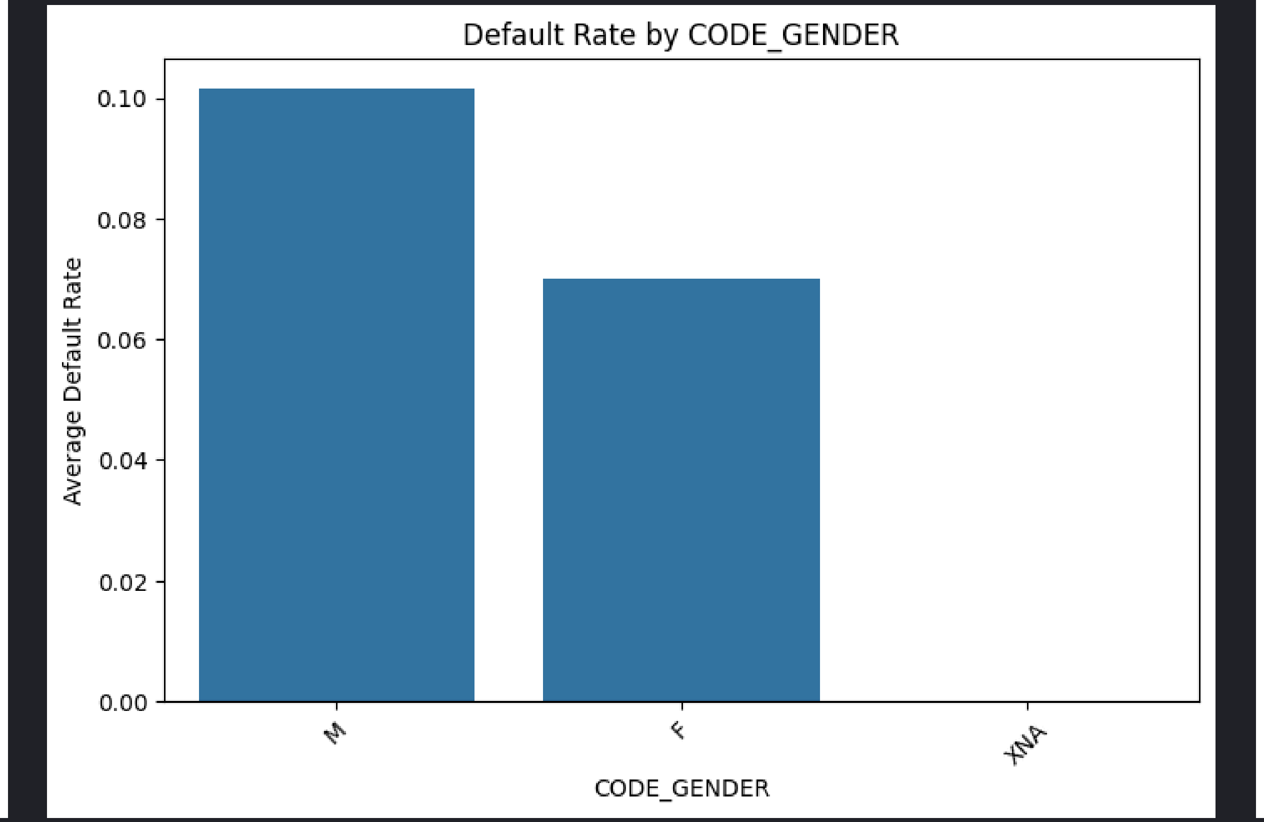
Distribution of EXT_SOURCE_3 by Loan Status

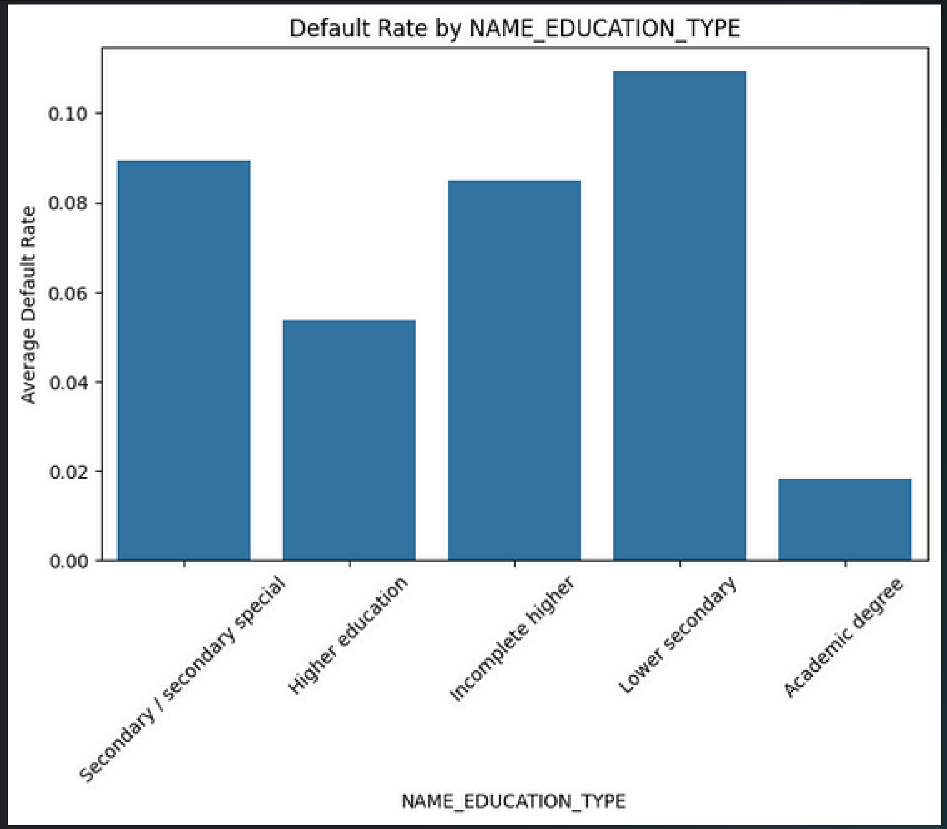
Finding:

All three credit score sources show that clients with lower external credit scores exhibit a noticeably higher likelihood of default, reinforcing that credit score remains one of the most reliable indicators of repayment capacity.

Categorical analysis



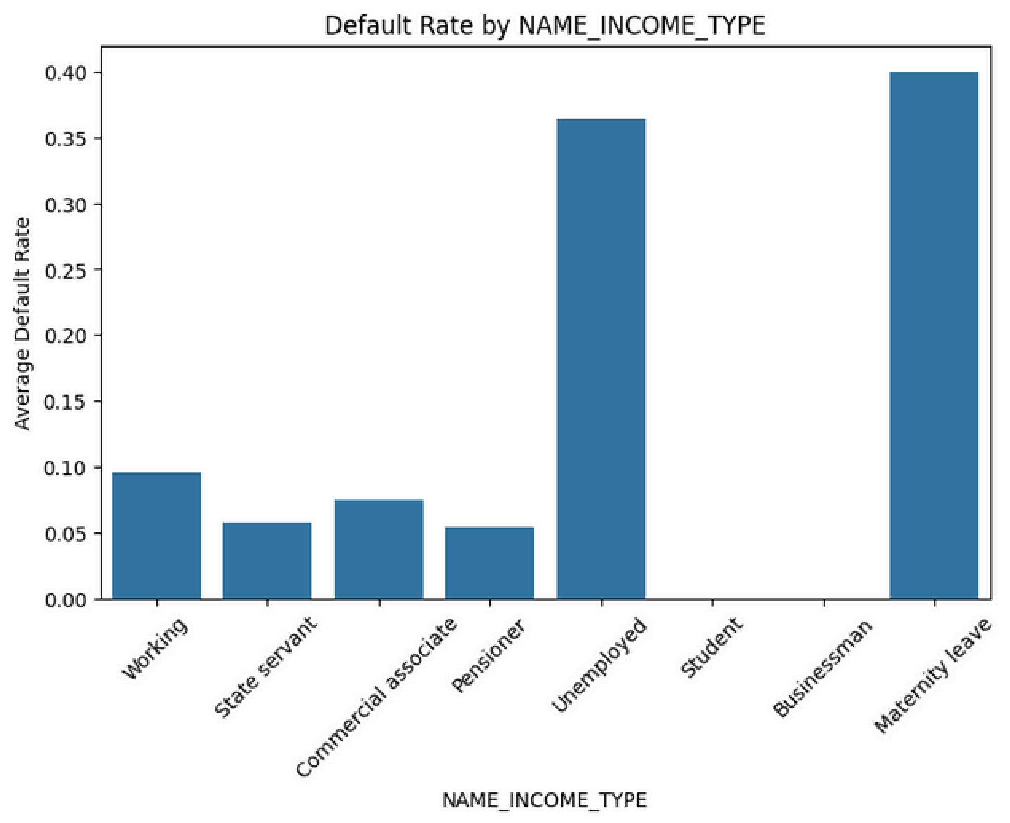


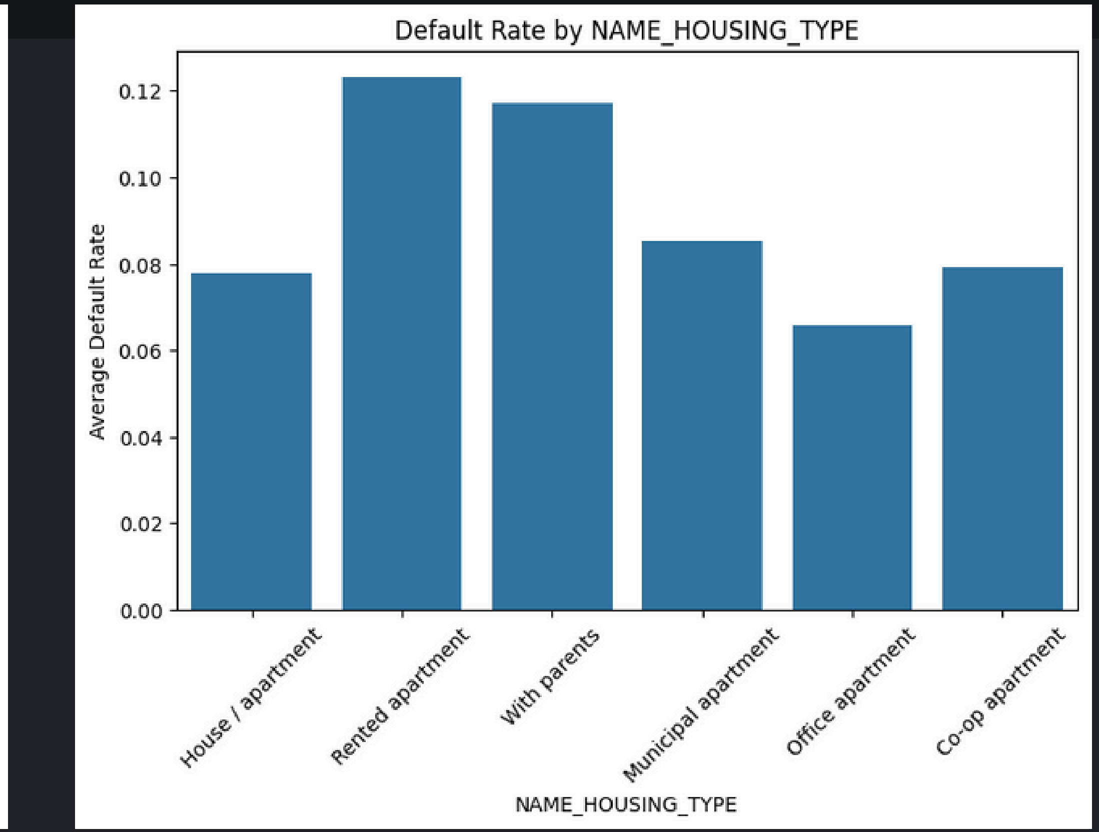


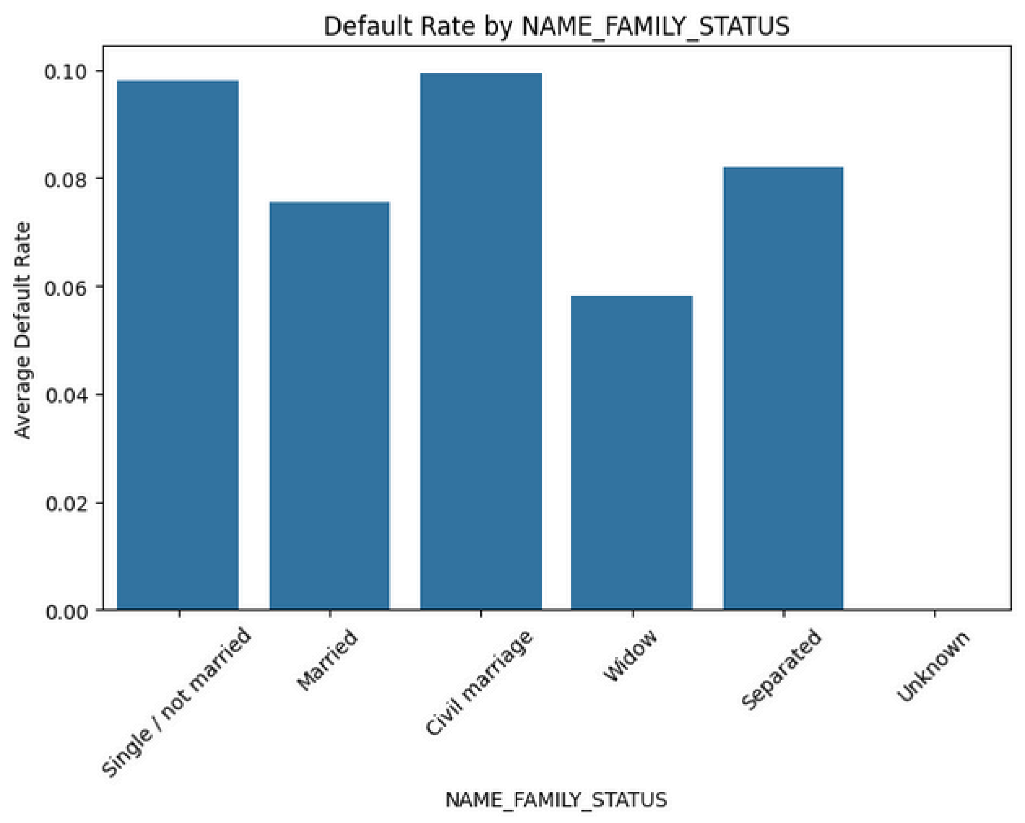
Finding:

Male clients and those with revolving loans show higher repayment difficulties, indicating that gender and credit type influence default risk. Lower education levels correlate with increased repayment issues, highlighting the link between financial stability and education.

Categorical analysis



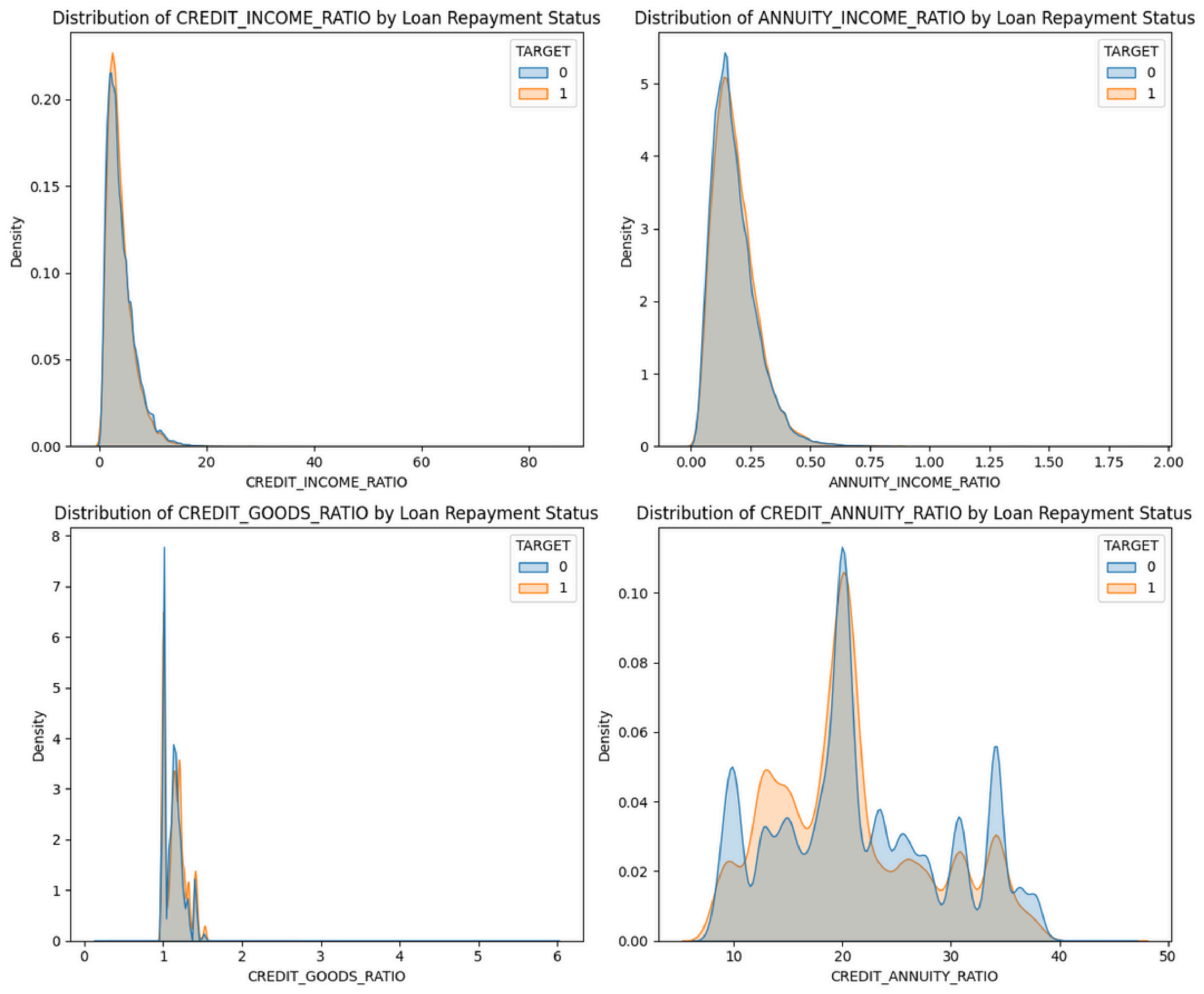




Finding:

Clients on maternity leave or unemployed show the highest default risk, emphasizing the importance of income stability. Single and renters also default more often, while married and homeowners demonstrate stronger financial reliability and repayment discipline.

Ratio analysis

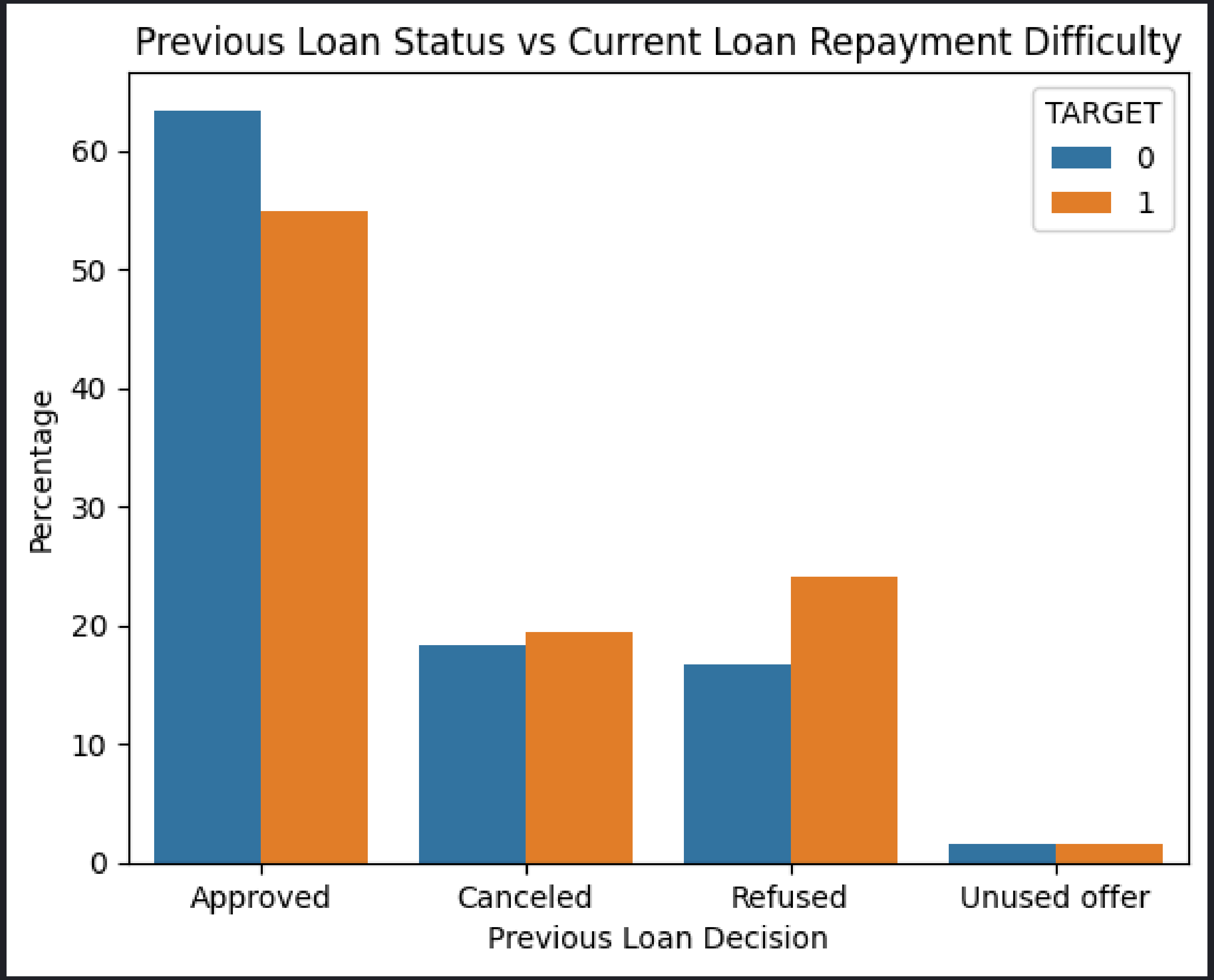


Finding:

- Defaulters recorded slightly higher repayment burden (higher annuity-to-income ratio).
- Credit-to-income and credit-to-goods ratios were similar between groups → loan size alone not a strong risk indicator.
- Suggests repayment pressure, not loan amount, better explains default behavior.
- Note: Combining multiple ratios may yield more reliable indicators of financial stress in future modeling

Past behavior analysis

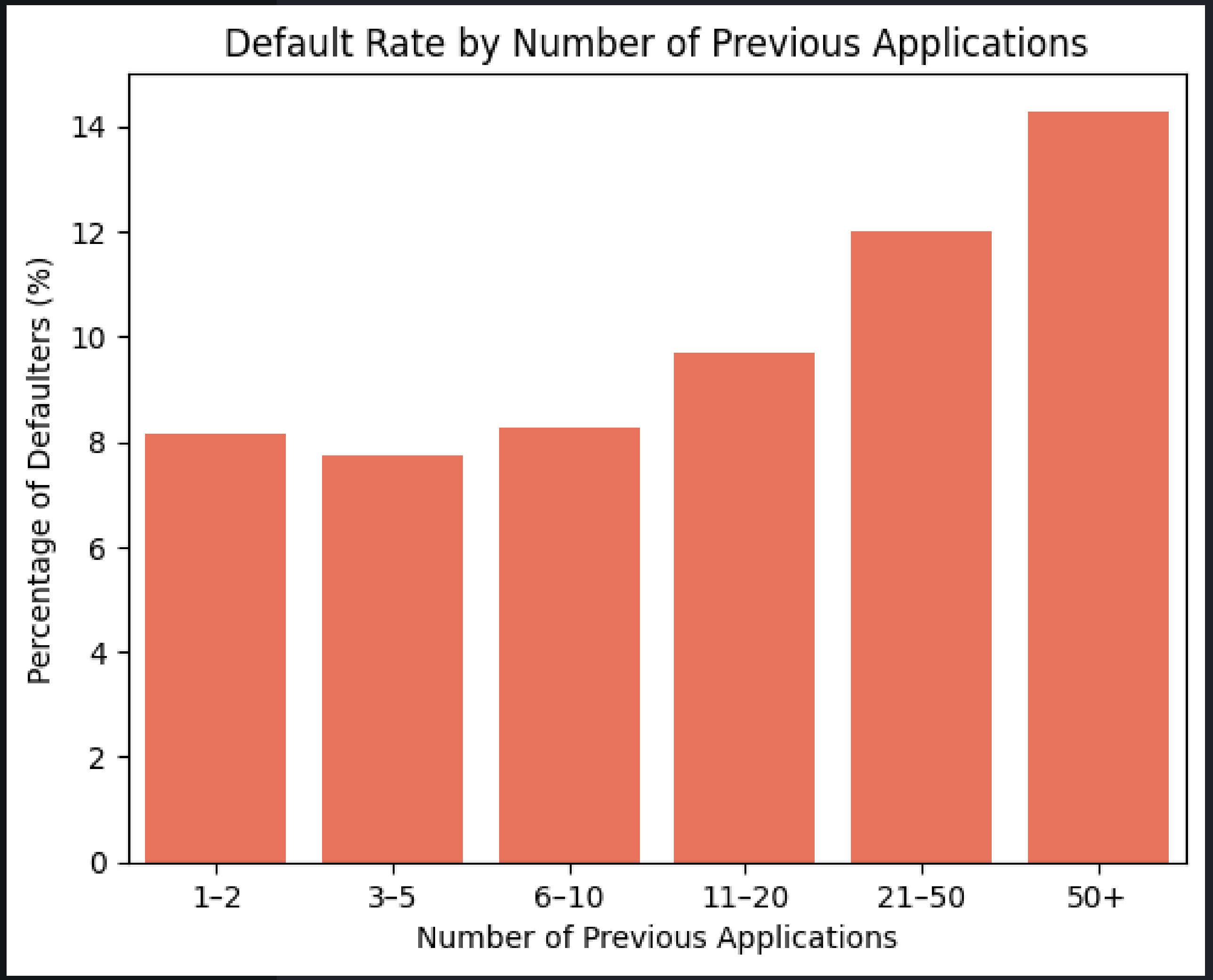
Past loan status analysis



Finding:

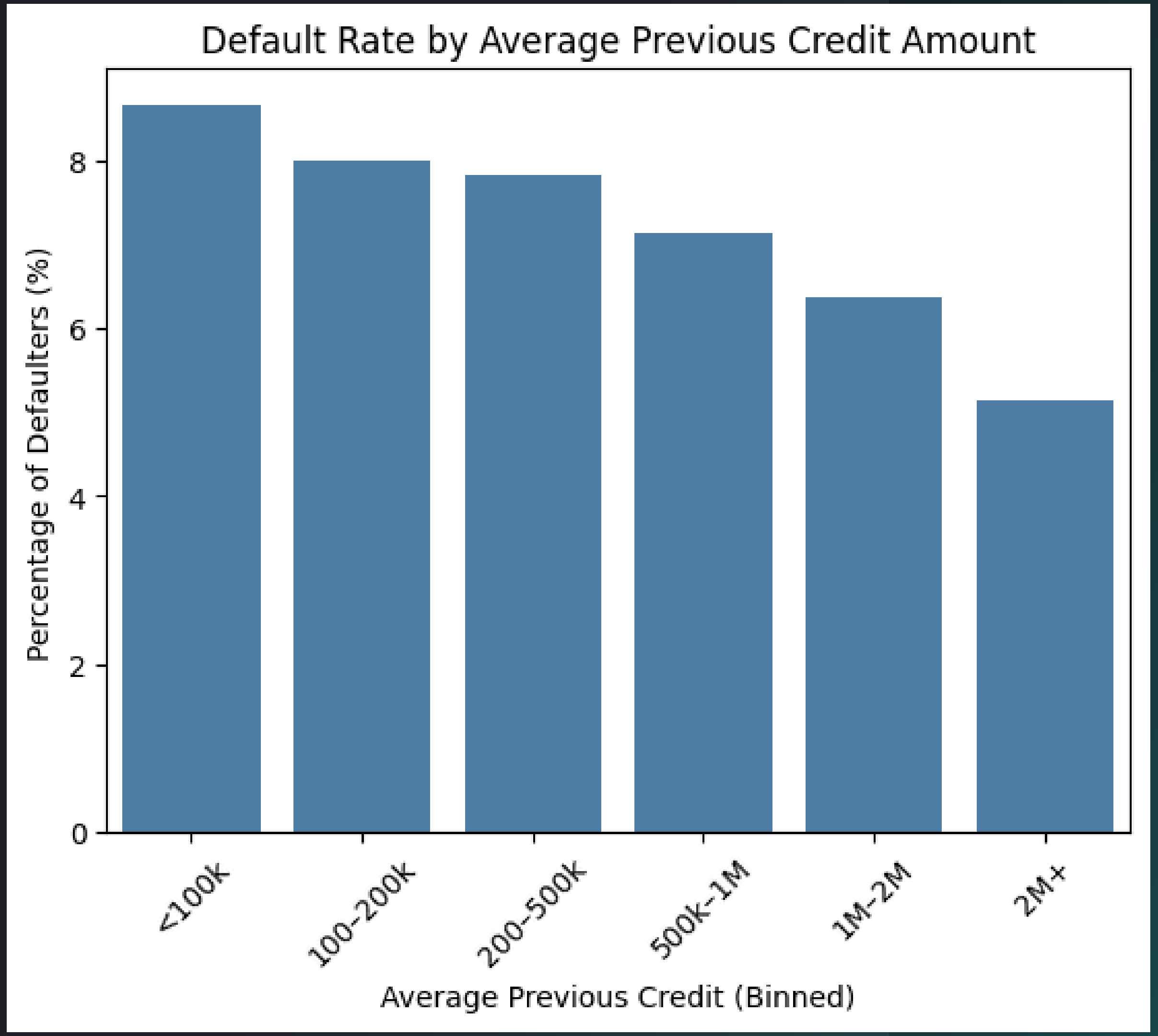
- Clients who defaulted on current loans show a higher share of past cancellations and refusals, and a lower approval rate.
- This pattern indicates inconsistent borrowing behavior and lower lender confidence in these clients' creditworthiness.
- In contrast, clients with a history of approved loans tend to maintain stronger repayment discipline and financial reliability.

Past loan status analysis



Finding:

- Frequent reapplications (21–50 or 50+ past loans) correlate with higher default rates, signaling financial instability or overreliance on credit.
- Clients with smaller past loan amounts (<$100k) also show greater default tendencies, suggesting that lower-value or short-term loans may carry higher risk.
- Highlights the need to monitor borrowing frequency and loan size as early warning indicators in credit assessments.



Final thoughts

Suggested actions

- Integrate past behavior metrics (refusals, cancellations, reapplications) into credit scoring.
- Use financial ratios and employment stability as key approval criteria.
- Provide financial guidance to younger or unstable-income clients.
- Build monitoring dashboards to flag risk trends early.
- Update credit policies and models based on data-driven insights.

Conclusion

- EDA revealed that younger, less stable, and lower-educated clients face higher default risk — financial maturity and employment stability are key drivers of repayment ability.
- Higher repayment burden (annuity-to-income ratio) proved a stronger risk indicator than income or loan size alone.
- Clients with frequent or low-value past loans and more refusals/cancellations showed greater current default tendencies.
- Credit risk is multifactorial — shaped by demographic, financial, and behavioral factors.
- Insights provide a foundation for smarter credit policies, targeted risk management, and future predictive modeling.

Future improvements

- Enrich data with credit bureau and income trend information for deeper financial insight.
- Advance to modeling (e.g., logistic regression, random forest) to predict loan default probabilities.
- Improve data quality via imputation and categorical standardization.
- Segment clients by risk level or loan type to target mitigation strategies.
- Turn insights into action — implement stricter screening and visualize risk metrics via dashboards.

thank you